RESEARCH ARTICLE

Feature selection framework for functional connectome fingerprinting

Kendrick Li[1] | Krista Wisner[2] | Gowtham Atluri[1]

[1]Department of EECS, University of Cincinnati, Cincinnati, Ohio

[2]Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana

Correspondence
Gowtham Atluri, Department of EECS, University of Cincinnati, Cincinnati, Ohio, USA.
Email: atlurigm@ucmail.uc.edu

Abstract

The ability to uniquely characterize individual subjects based on their functional connectome (FC) is a key requirement for progress toward *precision psychiatry*. FC fingerprinting is increasingly studied in the neuroimaging community for this purpose, where a variety of approaches have been developed for effective FC fingerprinting. Recent independent studies showed that fingerprinting accuracy suffers at large sample sizes and when coarser parcellations are used for computing the FC. Quantifying this problem and understanding the reasons these factors impact fingerprinting accuracy is crucial to develop more accurate fingerprinting methods for large sample sizes. Part of the challenge in fingerprinting is that FC captures both generic and subject-specific information. A systematic approach for identifying subject-specific FC information is crucial for making progress in addressing the fingerprinting problem. In this study, we addressed three gaps in our understanding of the FC fingerprinting problem. First, we studied the joint effect of sample size and parcellation granularity. Second, we explained the reason for reduced fingerprinting accuracy with increased sample size and reduced parcellation granularity. To this end, we used a clustering quality metric from the data mining community. Third, we developed a general feature selection framework for systematically identifying resting-state functional connectivity (RSFC) elements that capture information to uniquely identify subjects. In sum, we evaluated six different approaches from this framework by quantifying both subject-specific fingerprinting accuracy and the decrease in accuracy with an increase in sample size to identify which approach improved quality metrics the most.

KEYWORDS
fingerprinting, functional connectivity, precision psychiatry

1 | INTRODUCTION

Resting-state functional connectivity (RSFC) studies that estimate connectivity based on blood-oxygen-level-dependent (BOLD) signal measured using functional magnetic resonance imaging (fMRI) have revealed many principles of brain function (Bandettini, 2012; Beckmann, DeLuca, Devlin, & Smith, 2005; Damoiseaux et al., 2006; Fornito & Bullmore, 2015; Fox & Greicius, 2010; Greicius, Krasnow, Reiss, & Menon, 2003). Most of these studies made inferences about RSFC at a group level and showed that such inferences are reliable (Shehzad et al., 2009). While group-level inferences inform us of the generic principles, they obscure principles specific to individual

subjects that are essential for characterizing brain function in health and disease (Dubois & Adolphs, 2016). The increased availability of "dense" (Poldrack, 2017) scans, that is, repeated scans collected from the same individual that resulted in hours of fMRI data from the same subject, ranging from 60 min to 14 hr (Human Connectome Project (HCP) Smith et al., 2013, Midnight Scan Club (MSC), Gordon et al., 2017, MyConnectome, Laumann et al., 2015) present a tremendous opportunity to identify subject-specific functional connectivity and eventually study idiosyncratic properties of normal brain function and disrupted brain function. This has the potential to allow neuroscientists to study the effect of individually targeted medicines and procedures and make progress towards "precision psychiatry" (Poldrack, 2017; Satterthwaite, Xia, & Bassett, 2018).

With the goal of making progress towards precision psychiatry, a problem that has received increased attention in the neuroimaging community is that of *Functional connectome (FC) fingerprinting* (Finn et al., 2015), where the goal is to uniquely identify individual subjects using subject-specific RSFC. Specifically, given a set of *N reference* fMRI scans, one from each of the *N* subjects, and a new *target* fMRI scan from one of the same *N* subjects, the goal is to identify the subject by "matching" RSFC of the target scan with that of the reference scans. As RSFC is used to match the reference and the target scans, we refer to it as a functional fingerprint. This fingerprinting problem has been studied using the above data sets that constitute dense scans from individual subjects (Finn et al., 2015; Miranda-Dominguez et al., 2014).

Early studies on FC fingerprinting reported very high accuracies using a relatively small number of subjects. For instance, Finn et al. (2015), using 126 subjects from HCP, reported a fingerprinting accuracy in the range of 92–94% using whole-brain RSFC and 98–99% using frontoparietal-based RSFC. Waller et al. (2017) reproduced the results of Finn et al. (2015) on the same data set observing around 95% fingerprinting accuracy using frontoparietal-based RSFC. Amico and Goñi (2018), using 100 unrelated subjects from HCP, observed increased accuracy from 94 to 98% by performing the principal component analysis (PCA) on whole-brain RSFC and using the resultant principal components for reconstruction and matching. Xu et al. (2016) studied fingerprinting accuracy using a method involving the boundaries drawn between functional areas delineated using spatial gradients (the approach is discussed elaborately in Wig, Laumann, & Petersen, 2014) and reported a success rate of up to 99% using 30 subjects.

These near 100% success rates may lead one to conclude that fingerprinting is not only a relatively easy problem when dense scans are available but also a solved problem. However, this is far from reality. Waller et al. (2017) studied the generalization of FC fingerprinting at numerous subjects by modeling frontoparietal-RSFC based fingerprinting accuracy using the 900-subjects December 2015, HCP data release. They estimated heavily decreased accuracies of 62 and 42% at 10,000 and 100,000 subjects, respectively. *Understanding the reason for reduced fingerprinting performance at large sample sizes is crucial for addressing the problem effectively*. On the other hand, Peña-Gómez, Avena-Koenigsberger, Sepulcre, and Sporns (2018) showed that finer parcellations used in the computation of RSFC resulted in increased fingerprinting accuracy compared with coarser parcellations (Peña-Gómez et al., 2018).

A variety of methods for performing fingerprinting using RSFC have been developed since Finn et al.'s (2015) seminal work that paved the way for characterizing subject-specific information in functional connectivity matrices and quantifying it. For ease of comparison, we categorized these approaches into two classes: native-RSFC approaches and RSFC-derived approaches. Native-RSFC methods use the RSFC matrix as a whole or select part of it for fingerprinting. Examples include the approach proposed by Finn et al. (2015) which uses the correlation between connectivity values taken from a target RSFC matrix and those from each of the reference RSFC matrices to determine the best match. Finn et al. (2015) also used connectivity values from a predefined sub-network in the RSFC matrix as opposed to using the connectivity values from the entire RSFC matrix. Another approach proposed by Peña-Gómez et al. (2018) uses connectivity from selected regions that are deemed relevant for fingerprinting. In contrast, RSFC-derived approaches construct new features from the RSFC to use for fingerprinting. Examples include PCA based feature construction (Amico & Goñi, 2018), a graph embedding based approach (Abbas et al., 2020), and deep learning-based approaches (Chen & Hu, 2018; Lori, Ramalhosa, Marques, & Alves, 2018; Shojaee, Li, & Atluri, 2019). The native-RSFC methods are the simpler of the two, in that they use a subset of original connectivity values from the RSFC matrix. In contrast, RSFC-derived methods employ complex machine learning models with thousands of parameters (e.g., deep learning models by Shojaee et al. (2019), Chen and Hu (2018), and Lori et al. (2018)) to project the RSFC matrix into a new, unfamiliar embedding "space" that maximizes the variance between subjects. Despite their vastly different complexity, both these classes of approaches reported similar improvements in performance over the baseline approach. This suggests that the RSFC-derived methods are not able to capture any more signal than what is available in the native-RSFC matrices. Moreover, the native-RSFC methods also result in the selection of interpretable features as FC information is computed using correlation of BOLD time series.

The underlying hypothesis in native-RSFC methods is that FC comprises both subject-specific and generic information. To perform fingerprinting accurately an effective delineation of unique subject-specific information is required. Native-RSFC methods are ideally suited for this with the added advantage of capturing interpretable features. *While several native-RSFC methods have been studied independently, a systematic study comparing such methods is needed to determine how to use RSFC elements and procedures most effectively for fingerprinting*.

In this study, we addressed three gaps in our understanding of the FC fingerprinting problem. First, we studied the joint effects of sample size and parcellation granularity on fingerprinting accuracy. Second, we used a novel technique to help explain the reason for reduced fingerprinting accuracy with increased sample size and reduced parcellation granularity. To this end, we used a clustering quality metric from the data mining community. Third, we developed

a general feature selection framework for systematically identifying procedures and RSFC elements that capture information to uniquely identify subjects. We evaluated six different approaches from this framework by quantifying both fingerprinting accuracy and the decrease in accuracy with an increase in sample size. We also evaluated the reproducibility of the selected features in two independent groups of subjects.

2 | METHODS

2.1 | Data and preprocessing

In this study, we used node-time series data from the HCP1200 Parcellation + Time series + Netmats (PTN) release which is part of the March 2017 HCP1200 data release (Smith et al., 2013). Under the HCP protocol, all participants provided written informed consent and the HCP study was approved by the Institutional Review Board at Washington University in St. Louis.

As part of the HCP1200 data release, four resting-state fMRI scans were collected from each subject on 2 days. Two, ∼15 min, scans with a TR of 720 ms were obtained per day, a left-to-right (LR) phase-encoded scan and a right-to-left (RL) phase-encoded scan. Each fMRI scan was then preprocessed using the HCP functional pipeline (Glasser et al., 2013; Smith et al., 2013) which included artifact removal using independent component analysis (ICA) + FMRIB's ICA-based Xnoiseifier (ICA + FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014), and inter-subject registration using the Multimodal Surface Matching algorithm ("MSMAll") (Glasser et al., 2016; Robinson et al., 2014). The processed fMRI scans were in grayordinate space consisting of surface vertices and subcortical voxels.

As part of the HCP1200 PTN release, these minimally-preprocessed scans were further processed using group-PCA and group spatial-ICA to generate group-based independent components (ICs) at different granularities: 15, 25, 50, 100, 200, and 300. Group ICs were then mapped to each subject's rfMRI data using dual-regression to compute subject-specific time series for each IC. We consider each IC as a "node" in the brain network. This processing was performed only on 1,003 healthy young adults (ages 22–35, 469 Male, 534 Female) that have four resting-state fMRI runs of 1,200 time points each (totaling 4,800), resulting in four sets of node-time series (one for each scan) at 15, 25, 50, 100, 200, 300 parcellation granularities. Additional details about these steps are available in the HCP documentation (2017).

2.2 | FC fingerprinting

In this section, we define the terminology that we used in the rest of the paper and discuss the process used for fingerprinting analysis.

We refer to fMRI scans for which we know which subject they are collected from as "reference" scans. We refer to the new set of scans that are to be matched with reference scans as "target" scans.

Given a set of N reference scans $\{R_1, R_2, ..., R_N\}$ from N different subjects, and a set of target scans $\{T_1, T_2, ..., T_N\}$ from the same set of subjects, the problem of FC fingerprinting is to determine for each target scan T_i the corresponding subject's reference scan R_j by matching their FC. There are two key steps here: (a) computing FC, (b) matching FC.

2.2.1 | Computing FC

We z-score normalized and concatenated LR and RL node time series from each of the 2 days, creating two ∼30 min separate-day node time series for each subject. For computing FC, we computed Pearson correlation between each pair of node time series, generating a pairwise correlation matrix. We computed FC on node time series for the 2 days separately, resulting in two FCs from each subject: FC_{d1} and FC_{d2}. As node-time series data are available for different granularities of parcellation, these two FCs were computed for each granularity. We refer to the set of FCs computed from *day 1* node time series at different granularities as FC_{d1}^{15}, FC_{d1}^{25}, FC_{d1}^{50}, FC_{d1}^{100}, FC_{d1}^{200}, and FC_{d1}^{300} and those from *day 2* node time series as FC_{d2}^{15}, FC_{d2}^{25}, FC_{d2}^{50}, FC_{d2}^{100}, FC_{d2}^{200}, and FC_{d2}^{300}.

2.2.2 | Matching FC

For matching FCs, we used the method described in Finn et al. (2015). Specifically, for each FC computed from a target scan T_i, we computed the Pearson correlation between the vector constructed by taking all upper-triangular values of the target FC matrix with that of each of the reference FCs. The reference FC that showed the highest correlation with the target FC is treated as a match. As this approach uses connectivity values for all pairs of regions, we refer to this approach as the *Full-FC* approach.

The accuracy of fingerprinting is computed as the fraction of subjects for which the target scans were correctly matched with their reference scans. As we have two FCs from each subject (FC_{d1} and FC_{d2}), we computed fingerprinting accuracy in two ways: (a) using FC_{d1} as a reference and FC_{d2} as a target, (b) using FC_{d2} as a reference and FC_{d1} as target. The results from the former and latter cases are labeled as *Day1 Ref; Day2 Tgt* and *Day2 Ref; Day1 Tgt*, respectively.

2.3 | Silhouette coefficient based analysis for FC fingerprinting

2.3.1 | Impact of sample size and parcellation granularity on FC fingerprinting

To study the effect of sample size and parcellation granularity on FC fingerprinting accuracy, we first computed the fingerprinting accuracy for different sample sizes and parcellation granularities. We used sample sizes ranging from 100 to 1,000 subjects in steps of 100 and each parcellation granularity (15, 25, 50, 100, 200, and 300). We repeated

this analysis over 150 randomly sampled sets for each sample size and granularity. To estimate fingerprinting accuracy at very large sample sizes that are currently infeasible to enroll and study for the above scenarios, we first learned linear models using the log-transformed number of subjects and resultant accuracies and then estimated fingerprinting accuracy at 100,000 subjects.

2.3.2 | Silhouette coefficient

The underlying hypothesis in Finn et al.'s (2015) approach is that all FCs that are generated from the same subject reside in close proximity in some high-dimensional space and are also well separated from FCs generated from other subjects. However, this has been challenged by observations in recent studies that found decreased fingerprinting accuracy with an increase in the number of subjects (Waller et al., 2017). Our new hypothesis is that as we increase the number of FCs from different subjects, there is an increase in the chance that an FC, which previously would have correctly matched with an FC from the same subject, may match with a new FC from a different subject.

To test this hypothesis, we used the Silhouette coefficient (Tan, Steinbach, & Kumar, 2006), a commonly used cluster evaluation metric from the data mining community, to determine how well separated subject FCs are from other subject FCs in high-dimensional space. The Silhouette coefficient measures the cohesion and separation of some clustering of data points in a data set (see Figure 1 left). Values are computed for each data point ranging from −1 to 1, where values closer to −1 indicate that the data point is more similar to points that are assigned to other clusters than to points within its assigned cluster, while values closer to 1 indicate that the data point is more similar to points within its assigned cluster than to points that are assigned to other clusters. For the fingerprinting hypothesis, FCs computed for a subject are treated as members of a cluster. In this scenario, negative values indicate FCs more similar to FCs from other subjects than it is to FCs from its own subject. For a more

complete treatment of the Silhouette coefficient, we refer an interested reader to (Tan et al., 2006).

We calculated the Silhouette value for each FC (FC_{d1} and FC_{d2}) by treating the two FCs generated from a subject as members of one cluster, that is, by creating N clusters with two members each, and computed the average Silhouette value over all FCs. We repeated this on 150 randomly sampled sets for sample sizes ranging from 100 to 1,000 subjects in steps of 100 and each parcellation granularity (15, 25, 50, 100, 200, and 300). To directly measure the degree of overlap between subject clusters, that is, the extent of FC cluttering in a high-dimensional space, we defined an "overlap ratio" measure which is computed as the fraction of subjects with a negative Silhouette value for an FC. For example, if four subjects out of 10 contained at least one of their two FCs with a negative Silhouette value, the overlap ratio is 4/10 = 0.4 subjects with overlapping subject spaces. This indicates that 40% of subjects would have misidentified FCs and would reduce fingerprinting accuracy. To quantitatively compare the effect of parcellation granularity on the degree of cluttering as the sample size increased, we used the Silhouette values and overlap ratios and learned logarithmic models, with their corresponding p-values, with respect to the number of subjects.

2.4 | Feature selection framework for improved FC fingerprinting

Subject-level FC contains both generic and subject-specific information. By detecting *elements* of the FC that contain subject-specific information, we can effectively utilize these elements to reduce FC cluttering and improve FC fingerprinting accuracy at larger sample sizes.

We developed a general FC feature selection framework to discover subject-specific information from FC matrices. Our framework computes the amount of subject-specific information present in each feature by accounting for how well segregated each subject's FCs are from other subject's FCs and selects the feature with the most
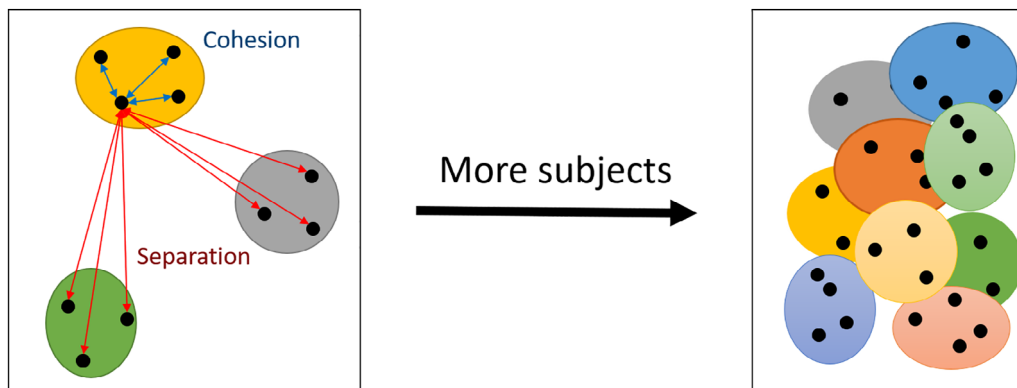

FIGURE 1 *Left*: Representation of FCs in a geometric space that reflects the underlying assumption in FC fingerprinting that FCs from the same subject have high cohesion, that is, are very similar to each other, and FCs from one subject are separated sufficiently from FCs of other subjects. *Right*: A representation that reflects our hypothesis that as we increase the number of subjects the degree of overlap between "subject spaces" increases, thereby reducing separation

subject-specific information. Our FC Feature Selection framework consists of four design parameters: (a) What are treated as features? (b) What FC distance measure is used to quantify the dissimilarity between two FCs? (c) What scoring metric is used to determine the amount of subject-specific information in a feature? (d) What number of features will be selected? A combination of choices for the four design parameters will result in a feature selection approach. The parameter choices along with the resultant feature selection approach spawned from our framework are listed in Table 1.

For each of these four design parameters, the choices we studied in this work are discussed below.

2.4.1 | Features

The first design parameter for the FC Feature Selection framework involves choosing which FC elements are to be used as features. We studied two possible choices: nodes and edges. When nodes are considered, we treated the FC profile of individual nodes as features. A node's FC profile is defined as the set of edges between a node and all other nodes, that is, a row in the FC. As we select nodes on each iteration, we call these approaches *Node selection* (NS) approaches. Alternatively, FC edges can be treated as features. Similarly, as edges are selected on each iteration we call these approaches *Edge selection* (ES) approaches. The process of node selection and edge selection is illustrated in Figure 2.

2.4.2 | FC distance measure

The second design parameter involves selecting a distance measure between FCs using a feature of interest. We used two different measures: standardized Euclidean (SE) distance (δ) and Pearson correlation (ρ).

Standardized Euclidean distance is computed as:

$$\delta_{ij}^k = \sqrt{\left(f_i - f_j\right)S^{-1}\left(f_i - f_j\right)'}$$

For a node k, f_i, and f_j are k^{th} rows in the ith and j^{th} FC matrices, respectively. S is the diagonal matrix whose diagonal elements are the SD of elements in the kth row for all FCs considered. For an edge k, f_i, and f_j are k^{th} edge weight in the ith and j^{th} FC matrices, respectively. S is the SD of edge k for all FCs considered.

Note that we cannot compute correlation when individual edges are treated as candidates for selection. As we test each edge individually, we will have to compute the correlation between two scalar values, which is not feasible.

2.4.3 | Scoring metric

The third design parameter involves selecting a metric to capture the amount of subject-specific information available in each feature. We investigated two cost functions: the average cross-session
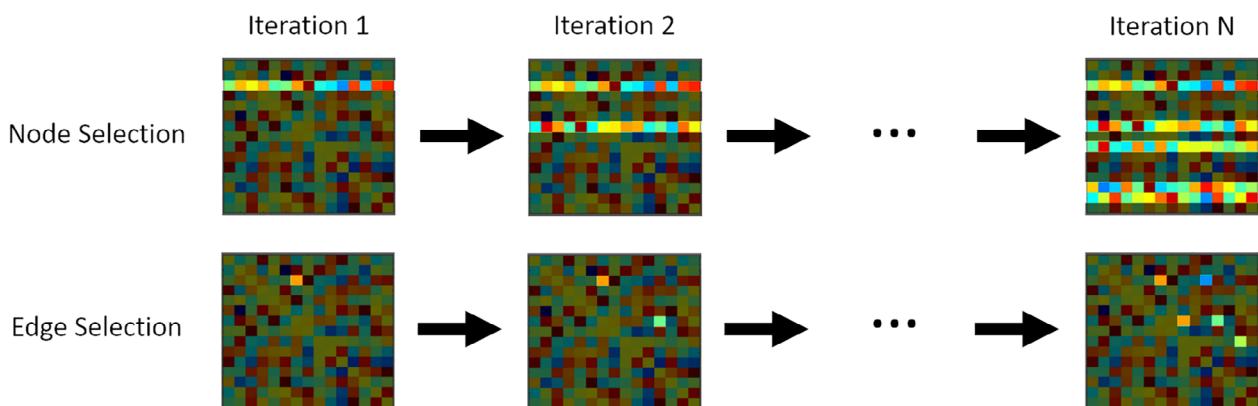
TABLE 1 FC feature selection approaches that resulted from the different parameter choices in our feature selection framework

Name	Feature definition	Feature score	Measure of FC distance	Stop criteria
NS_ACSC_ρ	Node profile	ACSC	Correlation	Validation test
NS_RSC_ρ	Node profile	RSC	Correlation	Validation test
NS_ACSC_δ	Node profile	ACSC	Standardized Euclidean	Validation test
NS_RSC_δ	Node profile	RSC	Standardized Euclidean	Validation test
ES_ACSC_δ	Edge	ACSC	Standardized Euclidean	Validation test
ES_RSC_δ	Edge	RSC	Standardized Euclidean	Validation test

FIGURE 2 Feature definition for FC Feature Selection Framework. Node selection (NS) involves selecting individual rows or columns from the FC matrix, whereas Edge selection (ES) involves selecting individual elements in the FC matrix. The highlighted elements indicate the selected features

cost (ACSC), inspired by the cost function introduced in Peña-Gómez et al.'s (2018), and the rank-sum cost (RSC), introduced by Airan et al. (2016). ACSC and RSC are illustrated in Figure 3 along with the underlying matrices that are used for computing these costs.

ACSC is computed using the cross-session distance matrix. Cross-session distance matrix D captures the distance between FCs from two separate sessions that are computed using an FC distance measure and a feature of interest. Specifically, an element D_{ij} captures the distance between the feature value in the ith subject's FC from the first session and that of the jth subject's FC from the second session. ACSC is then computed as the average off-diagonal distance minus the average diagonal distance in the cross-session distance matrix. A high ACSC indicates low same-subject cross-session distance and high different-subject cross-session distance for the feature in question, suggesting that the feature is effective for FC fingerprinting.

RSC is computed using a rank matrix. A rank matrix R is a $2N \times 2N$ matrix where the columns and rows indicate the FCs computed for N subjects using scans from each of the two sessions. Each row in the rank matrix captures the rank of the distances from FC_i to each of the remaining $2N - 1$ FCs. For the row FC_i, the rank of FC_i in the columns is given as 0 and so the highest value in each row is $2N - 1$. RSC is computed as the sum of all same-subject across-session ranks, that is, $\sum_{i,j} R_{ij}$ where $i \neq j$ and FC_i and FC_j are FCs for the same subject computed using scans from two different sessions. A low RSC indicates that the feature in question captures more subject-specific information as it results in low relative distance

between same-subject cross-session FCs when compared to all other FCs.

It is worth noting the difference between the distance matrix used by ACSC and the rank matrix used by RSC. In addition to the apparent difference in what they capture, that is, distance versus rank, there is also the difference in the extent of information available in these matrices. Specifically, RSC ranks FCs both within the session and between sessions, while the distance matrix only captures between-session distances. This provides RSC an additional advantage of accounting for within-session similarities, in addition to between session-similarities, which enables RSC to capture features that yield better-fingerprinting performance as demonstrated in our results.

2.4.4 | Number of features

The final design parameter involves determining the total number of features to select. This can be done by choosing a user-specified number of features or by choosing the number of features in a manner that maximizes fingerprinting performance. In our experiments, we scored features on training samples and then chose the number of features to use by evaluating the extracted set of features on validation samples. Specifically, after scoring features using the training set, we measure the average Silhouette value of the FCs in the validation set using the top $x\%$ scoring features for a wide range of values x. The $x\%$ top-scoring feature set with the highest average Silhouette value is then chosen to compute the FC fingerprinting accuracy on the test data set.
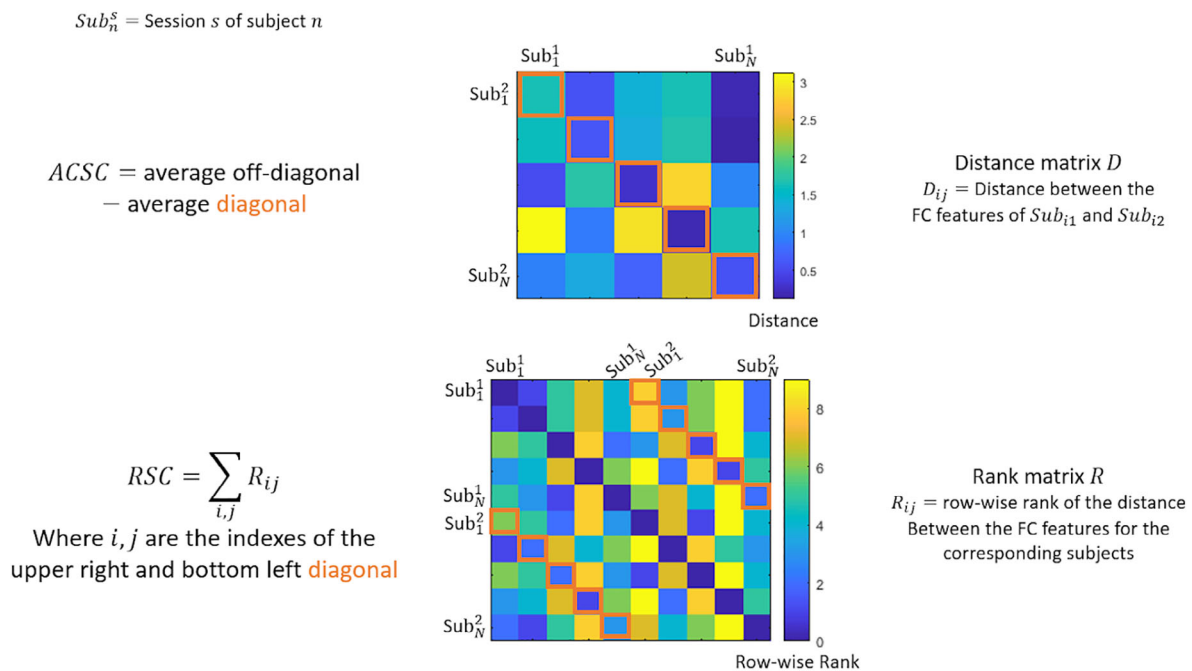

FIGURE 3 Illustration of ACSC and RSC cost functions used for scoring. This involves computing D and R matrices where the highlighted elements in orange color are used to calculate ACSC and RSC, respectively

2.4.5 | Approaches

The six different FC Feature Selection approaches we defined based on the above design parameters are: NS_ACSC_ρ, NS_RSC_ρ, NS_ACSC_δ, NS_RSC_δ, ES_ACSC_δ, and ES_RSC_δ. Table 1 shows the different parameter choices involved in defining these approaches. A MATLAB implementation of these approaches is made available on our github page: https://github.com/STDSLab/Feature-Selection-for-fMRI-Fingerprinting

2.5 | Comparative evaluation and Silhouette coefficient based analysis of FC feature selection approaches

To determine which FC Feature Selection approach, listed in Table 1, is most effective in improving FC fingerprinting accuracy over the Full-FC approach, we performed a comparative evaluation of all these approaches including the Full-FC approach.

This comparative evaluation was performed using a train-validation-test method, in line with our approach to determine the number of features. Specifically, from the FC^{300} data set, we created three randomly selected nonoverlapping cohorts "training set," "validation set," and "test set." Hundred subjects were selected for both the training and validation sets, leaving 803 subjects in the test set. For each approach, we computed scores for each feature based on their selected design parameters (features, FC distance measure, scoring metric). After scoring each feature, we extract top x feature sets over a range of cutoff points. The effectiveness of each feature set was then measured by computing the average subject Silhouette value on the validation set where only the features belonging to the feature set were used when computing the Silhouette values for each FC. The feature set yielding the highest average Silhouette value on the validation set was selected as the optimal feature set (features found at the optimal stopping point for identifying unique subject information). The Silhouette value was used to determine the optimal feature set rather than the FC fingerprinting accuracy as the Silhouette value is more sensitive to differences in feature sets. For example, cutoff points at 10 and 11% would result in different Silhouette values but may result in the same fingerprinting accuracy. By using Silhouette values, we would be able to select a definitive optimal cutoff point when both cutoffs may tie when using fingerprinting accuracy. Furthermore, as we showed in our results Silhouette coefficient reflected overall FC fingerprinting accuracy. We then randomly selected sets of 100 to 800 subjects in steps of 100 from the test set. The optimal feature set for each approach was then used to perform FC fingerprinting to compute the fingerprinting accuracy, average subject Silhouette value, and overlap ratio for each sample size. To account for sampling bias on the training set, this evaluation was repeated over 150 randomly sampled sets. These performance metrics are also computed for the Full-FC approach using randomly selected sets of 100 to 100 subjects in steps of 100 from the test set. Note that training and validation sets are not used in the context of the Full-FC approach.

Using empirical performance metrics computed for different fingerprinting methods, we learned linear models between a logarithm of the number of subjects ($ln(N)$) and performance metrics (fingerprinting accuracy, average Silhouette value, and overlap ratio) to extrapolate the fingerprinting performance for a large number of subjects.

2.6 | Evaluating the reproducibility of RSFC elements selected by the most effective approach

To investigate the reproducibility of the FC features selected by the most effective feature selection approach, we split the data arbitrarily into two cohorts and compared the features selected from different cohorts. Specifically, from the FC^{300} data set two nonoverlapping, 500 subject cohorts A and B were randomly selected and the corresponding scores for each feature were computed per cohort. The cohort A scores were then compared to the corresponding scores from cohort B for each feature. In addition, we tested the statistical significance of reproducibility by computing a hypergeometric distribution-based p-value (Balakrishnan & Nevzorov, 2004) for the amount of overlap in the selected features in cohorts A and B. Specifically, from a set of N features when x number of features are selected using feature selection from cohort A and y number of features are selected from cohort B, the hypergeometric distribution is used to estimate the probability of finding an overlap of f features or more between x and y due to random chance. Low probabilities indicate that finding f or greater intersecting features in between features selected in cohorts A and B is very unlikely by random chance, suggesting that the features found using the most effective approach are consistent across subject cohorts.

3 | RESULTS

3.1 | Silhouette coefficient based analysis for FC fingerprinting

We computed RSFC-based fingerprinting accuracy by varying the number of subjects and the parcellation granularity independently. For each combination, we measured the fingerprinting accuracy for 150 bootstrap samples of the data. This allowed us to robustly model the effects of these two parameters on the fingerprinting accuracy and to extrapolate accuracies at sample sizes beyond what is currently feasible. The average bootstrap sample fingerprinting accuracies for all combinations are shown in Figure 3a for scenario *Day1 Ref; Day2 Tgt*. We observed from Figure 3a that an increase in sample size resulted in a decrease in fingerprinting accuracy while an increase in parcellation granularity resulted in an increase in increased fingerprinting accuracy. Specifically, at a parcellation granularity of 300, fingerprinting accuracy decreased from 94.05 to 89.64%, a 4.41% decrease, as the sample size increased from 100 to 1,000 subjects. On the other hand, for a sample size of 100, fingerprinting accuracy

increased from 63.48 to 94.05%, a 30.57% increase, when parcellation granularity increased from 15 to 300.

To shed light on fingerprinting performance at sample sizes beyond those available in the HCP data set, we modeled the effect of sample size on fingerprinting accuracy at a parcellation granularity of 300. To learn our model, we used the fingerprinting accuracies measured for each bootstrap sample across different sample sizes. These data are shown as boxplots for each sample size in Figure 3b. We observed a similar decrease in accuracy with an increase in sample sizes, with respect to our previous observations, for both scenarios *Day1 Ref; Day2 Tgt* and *Day2 Ref; Day1 Tgt*, with the former scenario showing marginally higher accuracies. We learned a model to predict accuracy based on the logarithm of a number of subjects *n*. The resultant models are shown as blue (*Day1 Ref; Day2 Tgt*) and red (*Day2 Ref; Day1 Tgt*) dotted lines in Figure 3b. Specifically, the learned models are:

$$Day1\,Ref; Day2\,Tgt: a = 1 - 0.0145 ln(n) \quad p-value = 0$$

$$Day2\,Ref; Day1\,Tgt: a = 1 - 0.0166 ln(n) \quad p-value = 0$$

These models quantify the fingerprinting accuracy decrease rate with respect to the sample size as the coefficient of *ln(n)*. The low *p*-value of these models indicate that the accuracy is predicted significantly better compared to a degenerative model with only an intercept term (i.e., without the *ln(n)* term). Using these models, we estimated the fingerprinting accuracy for 100,000 subjects, using scenario *Day2 Ref; Day1 Tgt*, to be 80.89%. This suggests that for a large number of subjects, FC fingerprinting is unable to match the high accuracies observed for smaller data sets.

To explain the reason for the decrease in FC fingerprinting accuracy with an increase in sample size, we hypothesized that an increase in sample size is accompanied by an increase in subject FC overlap, that is, a decrease in the segregation of RSFCs in the high-dimensional space where FCs reside. This decrease in the segregation of RSFCs further explains the decrease in inaccuracy caused by the increase in sample size. To investigate this, we used the Silhouette coefficient and overlap ratio to quantify the segregation of RSFCs at different sample sizes and parcellation granularities. As per our hypothesis, the Silhouette coefficient is expected to decrease and the overlap ratio is expected to increase with an increase in sample size. We measured the average subject Silhouette value and overlap ratio for 150 bootstrap samples of the data for different number of subjects and different parcellation granularities. These measurements are shown in Figure 3c,d as boxplots for each sample size with respect to scenario *Day1 Ref; Day2 Tgt*. We observed that as the sample size increased, Silhouette values decreased and overlap ratios increased. At a parcellation granularity of 300, the average subject Silhouette value computed (mean of averages over 150 bootstrap samples) decreased from 0.329 to 0.273 as the sample size increased from 100 to 1,000 subjects. For the same increase in sample size, an increase in the overlap ratio (mean of overlap ratios over 150 bootstrap samples) was also observed from 0.071 to 0.121. These observations suggest that with

an increase in sample size, the corresponding decrease of segregation results in decreased fingerprinting accuracy.

Additionally, we observed that the parcellation granularity plays a role in the amount of decrease and increase for Silhouette values and overlap ratios, respectively. When parcellation granularity decreased from 300 to 15, we observed a 2.134 times decrease in the Silhouette value and a 2.580 times increase in overlap ratio when the sample size increased from 100 to 1,000 subjects. Specifically, for this increase in the number of subjects at a parcellation granularity of 15, the average subject Silhouette value decreased from 0.136 to 0.0165 and the overlap ratio increased from 0.399 to 0.528. To quantitatively compare the effect of parcellation granularity on the rate of change for the Silhouette value and overlap ratio as the sample size increased, we learned a model to predict the average subject Silhouette value *s* and overlap ratio *r* based on the logarithm of the number of subjects *n*. To provide an unbiased comparison between the rate of change for different parcellation granularities, we set a constant intercept value for all Silhouette value models and all overlap ratio models. The models for the average subject Silhouette value are shown in Figure 3c as blue (*FC^15*), red (*FC^50*), and yellow (*FC^300*) dotted lines. The overlap ratio models are shown similarly in Figure 3d. The Silhouette models are:

$$FC^{15}: s = 0.4 - 0.0561 ln(n) \quad p-value = 0$$

$$FC^{50}: s = 0.4 - 0.0328 ln(n) \quad p-value = 0$$

$$FC^{300}: s = 0.4 - 0.0176 ln(n) \quad p-value = 0$$

The corresponding overlap ratio models are:

$$FC^{15}: r = 0.0722 ln(n) + 0.04 \quad p-value = 0$$

$$FC^{50}: r = 0.0408 ln(n) + 0.04 \quad p-value = 0$$

$$FC^{300}: r = 0.0103 ln(n) + 0.04 \quad p-value = 0$$

The coefficient for the *ln(n)* term quantifies the rate of change for the Silhouette value and overlap ratio with respect to the sample size. Our models suggest that finer parcellation granularities exhibit a slower rate of decrease for Silhouette values and a slower rate of increase for overlap ratios with an increase in the sample size. Specifically, as the granularity increased from 15 to 300 the rate of decrease for the Silhouette value reduced from 0.0561 to 0.0176. Over the same increase in granularity, we observed that the rate of increase for the overlap ratio was reduced from 0.0722 to 0.0103. This reduction in the rate of change for the Silhouette value and overlap ratio suggests that finer parcellations produce RSFCs which segregate better as the number of subjects increase.

To summarize our observations: (a) We observed that as the number of subjects increased FC fingerprinting accuracy decreased, and as the parcellation granularity increased FC fingerprinting accuracy increased. (b) Using the Silhouette value and overlap ratio, we

observed that as the number of subjects increased RSFC segregation in high-dimensional space decreased. This supported our hypotheses that the decrease in fingerprinting accuracy is caused by a reduction in RSFC segregation due to an increase in the number of subjects. (c) We observed that as the parcellation granularity increased, the rate of decrease for the Silhouette vale and the rate of increase for the overlap ratio are both reduced. This suggested that finer parcellations produce better segregating RSFCs with an increase in the number of subjects. We repeated this analysis on motion-censored data and made the same observations as shown in section 1.2 of the Supplementary data.

As finer parcellations reduce segregation and improve fingerprinting accuracy, we observed that smaller regions are better able to capture subject-specific information. There can be multiple explanations for relatively low fingerprinting accuracy at coarse parcellations. The finer regions involved in variable connectivity across subjects but reliable connectivity within each subject could be merged with neighboring regions that may be either unreliable within each subject (e.g., due to

periodic changes in connectivity, Handwerker, Roopchansingh, Gonzalez-Castillo, & Bandettini, 2012) or merged with neighboring regions that may be less variable across subjects. Another reason could be due to finer regions that may be merged with neighboring regions that may have a different signal-to-noise ratio (Liu, 2016) that results in unreliable edges at the coarse level. Effective FC fingerprinting thereby requires not just a fine parcellation of the brain, but intelligent methods to identify and select those small regions with subject-specific information over those regions with generic information.

3.2 | Comparative evaluation of FC feature selection approaches

We performed a comparative evaluation of the FC Feature selection approaches, including the baseline Full-FC approach. The average fingerprinting accuracies for these approaches are shown in Figure 4a as a function of a number of subjects. All feature selection approaches
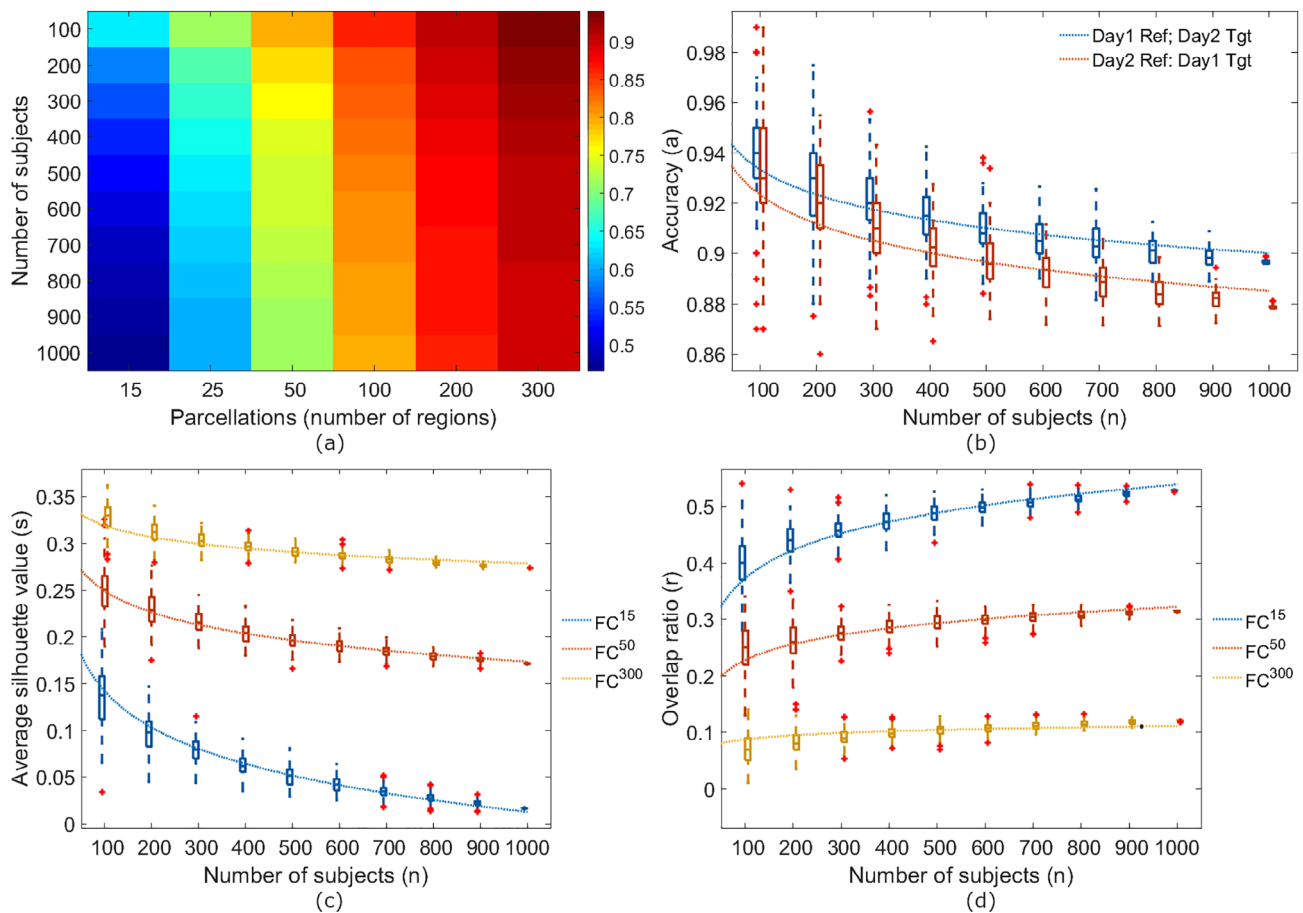

FIGURE 4 *Top*: Impact of sample size and parcellation granularity on Accuracy. (a) Average fingerprinting accuracy as a function of sample size and parcellation granularity for scenario *Day1 Ref; Day2 Tgt*. (b) Box plots of fingerprinting accuracies computed for 150 random runs using different sample sizes on the FC^{300} data set for scenarios *Day1 Ref; Day2 Tgt* (blue) and *Day2 Ref; Day1 Tgt* (red). Curves in blue and red show models learned for the two scenarios. *Bottom*: Silhouette coefficient analysis explaining the effect of sample size and parcellation granularity on fingerprinting accuracy. Box plots in (c) and (d) show Average subject Silhouette values (higher value indicates more unique) and overlap ratios (higher value indicates less unique), respectively, computed for 150 random runs using different sample sizes on the FC^{15} (blue), FC^{50} (red), and FC^{300} (yellow) data sets for scenario *Day1 Ref; Day2 Tgt*. Curves in blue, red, and yellow show models learned on the three data sets

performed better than the Full-FC approach, which used all edges in the FC matrix for fingerprinting. Of the feature selection approaches, ES approaches (ES_ACSC_δ and ES_RSC_δ) performed better than NS approaches. With regards to the scoring metric, ES approaches with ACSC resulted in marginally better accuracy than that of RSC. This is also the case for NS approaches that use correlation as a distance measure, ACSC performed better than RSC. In contrast, for NS approaches that use standardized Euclidean distance, RSC performed better than ACSC. We repeated this analysis on motion-censored data and made the same observations as shown in section 1.3 of the Supplementary data.

To investigate the potential of these approaches at improving FC fingerprinting accuracy a at larger sample sizes n, we learned linear models for each approach, including the Full-FC approach, on the log-transformed number of subjects and then estimated accuracies at 100,000 subjects. The resultant models are shown as curves in Figure 4a. The learned models are as follows:

$$Full-FC : a = 1 - 0.0145\ln(n)\ \ p-value = 0$$

$$NS\_ACSC\_\rho : a = 1 - 0.0081\ln(n)\ p-value = 0$$

$$NS\_RSC\_\rho : a = 1 - 0.0092\ln(n)\ \ p-value = 0$$

$$NS\_ACSC\_\delta : a = 1 - 0.0123\ln(n)\ p-value = 0$$

$$NS\_RSC\_\delta : a = 1 - 0.0105\ln(n)\ \ p-value = 0$$

$$ES\_ACSC\_\delta : a = 1 - 0.0023\ln(n)\ p-value = 0$$

$$ES\_RSC\_\delta : a = 1 - 0.0024\ln(n)\ p-value = 0$$

These models enabled us to quantitatively compare the accuracy decay with respect to increase in the sample size. Specifically, the coefficient of the $ln(n)$ the term indicates the rate of decrease in the accuracy of the approach as sample size increases. These models show the smallest reduction in accuracy decay for ES approaches, followed by NS approaches and the Full-FC approach. While marginal differences in decay are observed between ACSC and RSC based

models for both ES and NS approaches when standardized Euclidean distance is used, the NS_ACSC_ρ approach showed smaller decay compared with the NS_RSC_ρ approach. Using the above models, we estimated the fingerprinting accuracy estimates for 100,000 subjects and observed that the ES_ACSC_δ approach improved accuracy by 14.04% (estimated 97.35%), the NS_ACSC_ρ approach improved accuracy by 7.36% (estimated 90.67%), and the NS_RSC_δ approach improved accuracy by 4.60% (estimated 87.91%) with respect to the Full-FC approach (estimated 83.31%).

The average accuracy for 800 subjects along with the average number of selected features for the scenario *Day1 Ref; Day2 Tgt* on the test set across all 150 runs are shown in Table 2. While NS approaches select nodes and ES approaches select edges, to be able to compare the number of selected features we reported the number of edges selected for each of the feature selection approaches, given edges are involved in the calculations of each (see Section 2.4). From Table 2 it can be observed that the feature selection approach which performed the best (ES_ACSC_δ) made use of only 259 edges from the FC matrix, whereas the Full-FC approach which performed the worst made use of all 44,850 edges. This observation that ES approaches using a small set of edges results in the highest accuracy is supportive of our hypothesis that the FC matrix contains both subject-specific information and population-specific information. It is also worth noting that NS approaches generally select many more edges than ES approaches. For example, the NC_RSC_ρ approach which selected the least number of edges (4754) on average used 18 times more edges than ES_ACSC_δ. This is because NS approaches are forced to select all the edges in the node profile, a problem that ES approaches are free from.

3.3 | Silhouette coefficient based analysis to explain the utility of FC feature selection approaches

We studied the impact of feature selection approaches on the average subject Silhouette value for each approach at different sample sizes. These results are shown in Figure 4b. ES approaches resulted in the highest average Silhouette value and NS approaches resulted in a better average Silhouette value compared to the Full-FC approach. With

Name	Accuracy	# features	Silhouette value	Overlap ratio
Full-FC approach	90.06%	44,850	0.2794	0.1151
NS ACSC ρ	94.16%	14,271.4 (28.6)	0.3476	0.0745
NS RSC ρ	93.45%	4,754.1 (15.9)	0.3687	0.0825
NS ACSC δ	91.58%	4,903.6 (16.4)	0.3561	0.1008
NS RSC δ	92.72%	5,232.5 (17.5)	0.3637	0.0936
ES ACSC δ	98.17%	259	0.5439	0.0244
ES RSC δ	98.06%	248	0.5420	0.0248

TABLE 2 Comparative analysis of the FC feature selection approaches

Note: Average accuracy, silhouette value, and overlap ratio are reported for 150 runs for 800 subjects for scenario *Day1 Ref; Day2 Tgt*. ES_ACSC_δ outperformed other approaches, including the Full-FC approach.

regards to the scoring metric, ES approaches using ACSC resulted in a marginally better average Silhouette value than that of RSC. However, in the case of NS approaches, RSC resulted in a better average Silhouette value than ACSC. This is particularly because RSC takes into account both within-session and between-session FC distances, whereas ACSC takes only the latter into account.

The number of features selected using feature selection approaches are reported along with the average Silhouette values and overlap ratios in Table 2. The approach that resulted in the highest average Silhouette value and least overlap ratio was the ES_ACSC_δ approach (259 edges), the Full-FC approach (44,850 edges) resulted in the least performance. ES approaches which use a small set of edges resulted in the best segregation, that is, an overlap ratio of 0.02 as opposed to 0.12 for Full-FC approach. This observation is identical to that of Section 3.2 and is supportive of our hypothesis that the FC matrix contains both subject-specific information and the population-specific information and that the edge selection approaches are able to identify the edges that represent subject-specific information.

As with the accuracy results, we computed linear models on the log-transformed number of subjects (n) for average subject Silhouette value (s) to investigate these measures at larger sample sizes. The resultant models found for s are shown in Figure 4b. Specifically, the learned models are:

$$Full-FC : s = -0.0238ln(n) + 0.439 \, p-value = 0$$

$$NS\_ACSC\_\rho : s = -0.0312ln(n) + 0.553 \, p-value = 0$$

$$NS\_RSC\_\rho : s = -0.0317ln(n) + 0.578 \, p-value = 0$$

$$NS\_ACSC\_\delta : s = -0.0349ln(n) + 0.585 \, p-value = 0$$

$$NS\_RSC\_\delta : s = -0.0326ln(n) + 0.579 \, p-value = 0$$

$$ES\_ACSC\_\delta : s = -0.0343ln(n) + 0.773 \, p-value = 0$$

$$ES\_RSC\_\delta : s = -0.0351ln(n) + 0.776 \, p-value = 0$$

These models quantitatively describe the change in average subject Silhouette value as the sample size increased. We observed that the main contributor for the better performance of feature selection approaches is high average subject Silhouette values, where ES_ACSC_δ increased the offset value by 0.334 for the Full-FC approach, from 0.439 to 0.773. Although the average subject Silhouette value decay with an increase in the sample size for feature selection approaches are larger for Full-FC, this increase is marginal compared to the offset value within the range of realistic sample sizes.

3.4 | Evaluating the reproducibility of FC elements selected by the ES_ACSC_δ approach

We evaluated the reproducibility of the FC features selected by the ES_ACSC_δ approach by investigating the extent of overlap between features selected from two nonoverlapping subject cohorts. The high overlap of selected features indicates that the selected features are reproducible in capturing subject-specific information. When we selected the top 259 edges from two nonoverlapping, 500-subject cohorts A and B (chosen from the FC^{300} data set), we discovered an overlap of 185 features selected. To measure the significance of this overlap, we computed a hypergeometric distribution p-value. With a 185 overlap from a 259 selection, the p-value is 0. This p-value indicates that an overlap of this magnitude is statistically significant, suggesting that the ES_ACSC_δ approach consistently reproduced subject-specific edges effective for FC fingerprinting.

We visualized this consistency by plotting edge scores from each cohort against the other in Figure 5a, where the orange lines indicate
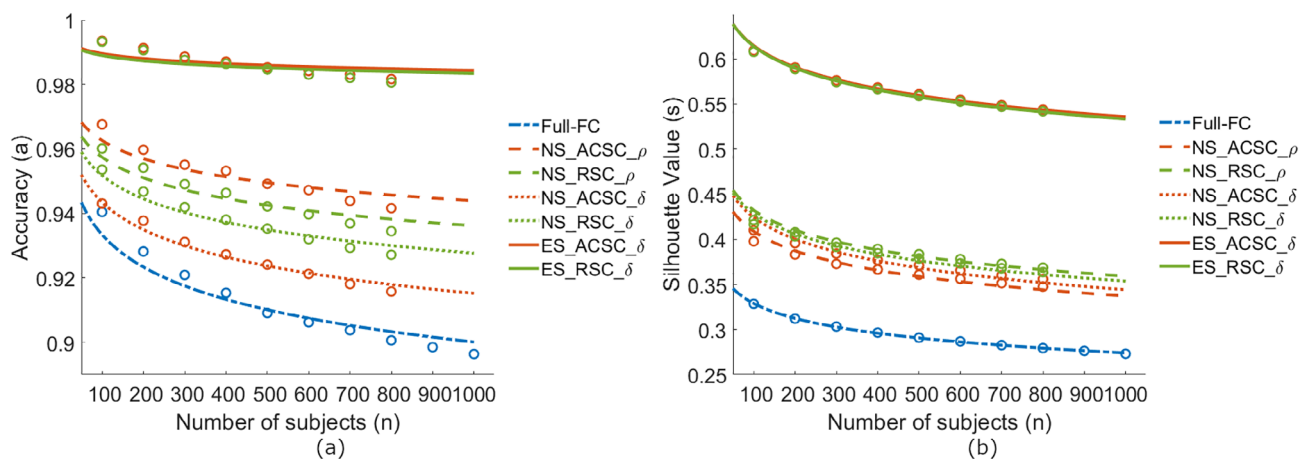

FIGURE 5 Performance of FC Feature Selection approaches and the Full-FC approach. Box plots showing: (a) fingerprinting accuracy and (b) average subject Silhouette value for 150 random runs computed across a range of sample sizes for each FC Feature Selection approach and the Full-FC approach. Data used was from the FC^{300} data set for scenario *Day1 Ref; Day2 Tgt*. Curves show the models learned to capture the impact of sample sizes on the three measures. The accuracies for ES approaches are much higher than others

the cutoff point for the top 259 edges selected from each cohort. Orange points indicate the 185 edges which are consistently discovered in both cohorts. Not only do we observe fairly consistent edge scores in the top 259 edges, but we also observed consistent scores across all edges with lower-scoring edges exhibiting marginally lower consistency compared with the higher scoring edges. To investigate the spatial relationships between the 259 best scoring edges in the FC, we plotted a heatmap where these edges are colored based on their scores, shown in Figure 5b. We observe that these edges mostly lie among a small set of nodes located in the upper left corner of the FC matrix.

The edges selected by ES_ACSC_δ are shown in Figure 5c for cohort A. The general connectivity of the selected edges is between three major areas: frontal cortex, parietotemporal region, and subcortical regions.

The superior performance of ES approaches is due to the selection of edges with high between-session variance, but low within-

subject differences, as those are the most effective for FC fingerprinting. Nodes contain a mixture of high-variance and low-variance edges and hence NS approaches are inferior to ES approaches. To demonstrate this, we compared the variance of edges selected by ES_ACSC_δ to those present in each node, by plotting the density of edge variance for edges in each node along with the variance of edges selected by ES_ACSC_δ (shown in Figure 6d. We observed that all nodes contain a large fraction of low variance edges when compared with the edges selected by ES_ACSC_δ. These results underscore the tendency of ES approaches to capture edges that exhibit high variance across subjects.

We also used a rank-sum based approach to demonstrate this. For each edge in the FC, we computed the variance of its connectivity value across Day 1 FCs (1,003 scans). We then ranked each edge in the decreasing order of variance, that is, the edge with the highest variance will be rank 1. Using the best 259 edges selected by the
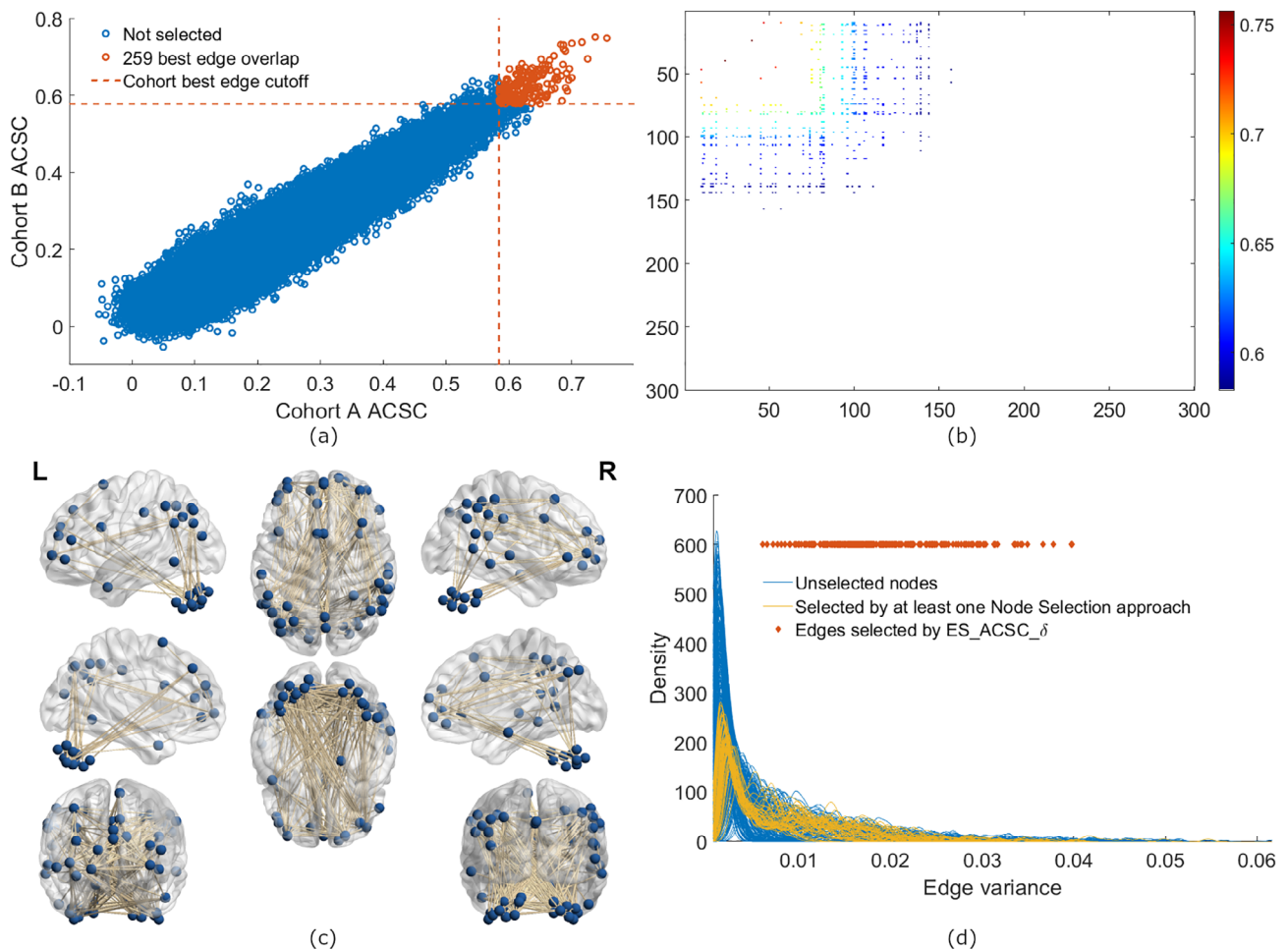

FIGURE 6 *Top*: Reproducibility of features found by the approach with the highest fingerprinting accuracy, the ES_ACSC_δ approach. (a) Comparison between edgewise ACSC values computed from 500 subject cohorts *A* and *B* selected from the FC^{300} data set. Orange circles show the 259 edges selected for both cohorts *A* and *B*. (b) Edges in the RSFC selected in cohort *A* colored by effectiveness (edges to the left of the dashed vertical line in a). *Bottom*: (c) Edges selected using the ES_ACSC_δ approach in cohort *A*. (d) Edge variance density for all 300 nodes. Each line shows the edge variance density of a node where higher densities indicate that more edges in the node have the specified edge variance in the Day 1 scans. Yellow lines indicate nodes which were selected by at least one Node Selection approach. Diamonds in orange indicate the edge variances of the edges selected by the ES_ACSC_δ approach

ES_ACSC_ δ approach (as reported in Table 2), we computed the sum of ranks for this set of edges. We also randomly picked 1,000 sets of 259 edges and computed the rank sums for each set of edges. The rank-sum of the 259 edges selected by ES_ACSC_ δ is smaller than all the 1,000 rank sums and so the p-value is <.001. These results are shown in Supplementary Figure S8. Our results indicated that the variance of the edges selected by ES_ACSC_ δ has significantly high overall variance compared to what is expected by random chance.

4 | DISCUSSION

In this work, we showed that an increase in sample size resulted in a decrease in fingerprinting accuracy, while an increase in parcellation granularity increased fingerprinting accuracy. We discovered the reason behind these observations using Silhouette coefficient, a metric used in data mining literature to assess clustering quality. We showed that as the number of subjects increase, the high-dimensional space is more cluttered, resulting in the high similarity between FCs from different subjects. We introduced a generic feature selection framework for improving fingerprinting accuracy that is similar to *feature selection*, a technique used in classification problems in data mining to identify the most informative features. We evaluated different instances of our framework to identify the most effective approach. Edge selection approaches, in particular ES_ACSC_δ, were observed to be the most effective in identifying subject-specific characteristics, despite comprising a relatively small number of edges. We showed using a Silhouette metric based analysis how the selected edges result in better segregation of FCs from the same subject. Our results lastly indicate that the edges selected using ES_ACSC_δ are highly reproducible and substantially different from edges (and associated edges) selected using other methods we evaluated.

4.1 | Impact of sample size and parcellation granularity

Our observations that an increase in sample size resulted in a decrease in fingerprinting accuracy while the increase in parcellation granularity resulted in an increase in fingerprinting accuracy, are in line with existing literature. Waller et al. (2017), reported decreased FC fingerprinting accuracy with increasing sample sizes on the HCP 900 data set. Peña-Gómez et al. (2018) and Airan et al. (2016) reported increased fingerprinting accuracy as the parcellation granularity increased. As part of their work, Peña-Gómez et al. (2018) used a random parcellation while Airan et al. (2016) used both a random parcellation and a spectral clustering-based parcellation (Craddock, James, Holtzheimer III, Hu, & Mayberg, 2012). In contrast, we used the ICA-based parcellation generated using the HCP 1200 data set. These observations collectively suggest that an increase in fingerprinting accuracy is mainly due to finer parcellation granularity and much less due to the method used for parcellation.

4.2 | Using Silhouette coefficient to explain the impact of sample size on fingerprinting

One of the major contributions of this work is in shedding light on the reason behind the change in fingerprinting accuracy with change in the number of subjects and parcellation granularity, where we used the Silhouette coefficient that is often used to assess the quality of clustering of data points by measuring the cohesion of points within clusters and separation of clusters. We treated FCs from the same subject as a cluster, and to measure the degree of overlap among subject clusters we defined an "overlap ratio" metric based on the number of subjects with negative Silhouette coefficients. We observed low Silhouette values and high overlap ratios when the number of subjects increased and the reverse when parcellation granularity was increased. This observation suggests that FCs in high-dimensional space became more cluttered as the sample size increased. Furthermore, we found that finer parcellations not only reduced FC cluttering, but also exhibited reduced cluttering as the number of subjects increased. This suggests that subject-specific functional connectivity which exists among smaller ROIs is lost as they are agglomerated with adjacent ROIs to form bigger ROIs and that analyses using the smaller ROIs are resistant to negative effects of cluttering in larger sample sizes.

4.3 | Introducing a feature selection framework and its evaluation

Another contribution of this work is the proposed feature selection framework to alleviate the cluttering problem. This framework consists of four design choices: (a) Which FC elements are defined as features, for example, nodes or edges? (b) What cost function is used to score features to determine optimal features for selection? (c) What measure is used to compute the distance for our cost function? (d) What stopping criteria are used to determine the number of final features to select? We tested a variety of possible instantiations of our framework to provide insight into which feature, cost function, and distance measure are most effective in improving FC fingerprinting accuracy. To this end, we compared two features, node profiles (node selection (NS)) and edges (Edge Selection (ES)); two cost functions, Average Cross-Session Cost (ACSC) and Rank Sum Cost (RSC); and two distance measures, Standardized Euclidean (SE) and Pearson correlation (correlation) in our comparative analysis.

This direction of feature selection has been earlier explored to a small extent and in an ad hoc manner. For example, the approach proposed by Peña-Gómez et al. (2018) can be seen as an instantiation of our framework. Specifically, it is a NS approach that uses ACSC scoring metric. The metrics of differential power discussed in Finn et al. (2015) and intraclass correlation used by Amico and Goñi (2018) that are proposed to quantify subject-specific information in nodes and edges can serve as cost functions in our framework. However, the current study systematically examined and compared a wider range of methods to guide future investigations. Our comparative

evaluation resulted in three key observations: (a) ES approaches are the most effective in improving fingerprinting accuracy and maintaining high fingerprinting accuracy at large sample sizes, (b) When comparing node profiles, correlation-based approaches are more effective than SE-based approaches, and (c) While the effect of cost functions ACSC and RSC for analysis depends on the distance measurement used, the difference in accuracy is marginal. As ES approaches are more effective, this suggests that subject-specific information lies in specific region-region interactions, that is, edges, rather than for the entire region's connectivity profile. The fact that correlation-based approaches are more effective than SE-based approaches indicates that the relative edge strength is more reliable between scans from the same individual than the actual value of the edge weights for fingerprinting.

4.4 | Importance of modeling accuracy decay with the increase in sample size

For each of the feature, selection approaches studied in this work we modeled the rate of decrease in accuracy with the increase in sample size. We observed that that ES approaches exhibit the least rate of decrease, in addition to achieving the highest accuracy. While recently published fingerprinting approaches (Abbas et al., 2020; Shojaee et al., 2019; Venkatesh, Jaja, & Pessoa, 2020) report higher accuracy compared to baseline approaches, they rarely account for performance degeneracy with the increase in sample size. We argue that fingerprinting problem is susceptible to the "cluttering" issue we demonstrated in this paper and to make progress towards precision psychiatry it is also important to study the "rate of decrease" inaccuracy with the increase in sample size.

4.5 | Selected edges and their reproducibility

The edges selected using the best approach (ES_ACSC_δ) in our framework exist between three major areas: frontal cortex, parietotemporal region, and the subcortical regions. While the role of subcortical regions in the context of fingerprinting has not been explored in the literature, the edges between the frontal and parietotemporal regions have been found to play an important role in multiple studies. Specifically, using a network-centric analysis Finn et al. (2015) observed that the edges in the frontoparietal network were the most effective for fingerprinting. Using a PCA-based approach, Amico and Goñi (2018) also reported high identifiability from regions in the frontal–parietal and default mode networks. Airan et al. (2016) observed that the prefrontal cortex and parietotemporal regions are among the regions that provided the most individual differentiation. Peña-Gómez et al. (2018) found that voxels in the frontal–parietal network and default-mode network were selected the most for fingerprinting, with some voxels from the dorsal-attention and salience networks. Miranda-Dominguez et al. (2014) observed that the most variable connections in their individual-based models lie

in the frontal–parietal cortices. Using a dynamic-FC-based fingerprinting approach, Liu, Liao, Xia, and He (2018) found the default-mode, dorsal-attention, and frontoparietal regions contributed the most to fingerprinting. Thus, our findings fit with the broader literature and highlight a need to further consider subcortical regions in future studies.

By performing reproducibility analysis on the best approach (ES_ACSC_δ) in our framework, we observed that the selected edges are significantly reproduced in independent cohorts. This result suggests that the selected edges are independent of the samples considered for feature selection yet consistent, which is an important measurement property that has been ignored in recent studies that proposed sophisticated fingerprinting approaches that are based on deep learning (Chen & Hu, 2018; Shojaee et al., 2019) and graph embedding (Abbas et al., 2020). Not studying the reproducibility of the resultant models leaves the question of generalizability of the learned fingerprinting models to other cohorts unanswered.

4.6 | Limitations and future work

One limitation of our work is that we used data from related individuals in the HCP. However, note that a similar analysis we performed on unrelated individuals in the HCP resulted in similar observations (Li & Atluri, 2018). Another limitation is that our framework follows a greedy selection which may result in suboptimal selection of features, because this framework cannot find a globally optimal solution due to the *greedy* nature of the search. Other optimization strategies such as genetic algorithms can be explored to overcome this limitation. In this study we did not investigate the impact of collection parameters such as scan duration and scan quality on the fingerprinting performance (Airan et al., 2016; Amico & Goñi, 2018; Finn et al., 2015; Horien et al., 2018). We also did not study the role of brain parcellation on the performance of our feature selection approaches.

A direction that could be explored as part of future work is to study the impact of time between scans on the degree of cluttering. MyConnetcome data set (Poldrack, 2021) that provides 14 hr of resting-state data collected from one subject is ideal for these investigations. Another promising direction is to extend this feature selection work to dynamic FCs (dFC). There has been much interest in modeling FC by taking into account the transient nature of the edge connectivity (Allen et al., 2014; Preti, Bolton, & Van De Ville, 2017). Using the transient nature of the edges as a way to uniquely profile individual subjects and selecting edges that are effective for this purpose is a direction that needs to be explored.

5 | CONCLUSION

In this study we demonstrated the reason for the reduction in fingerprinting performance with increase in sample size using Silhouette coefficient that is commonly used in the data mining community. We proposed a generic feature selection framework for identifying

features from the FC that are highly suited to uniquely identify subjects. We evaluated six different approaches from this framework and showed that ES-based approaches are superior to NS-based approaches in both fingerprinting accuracies and the rate of decrease of this accuracy with increase in sample size. We also observed that the edges selected using the best approach are independent of the cohort from which the data are collected, yet consistent across cohorts. The current findings are hoped to provide clarity and guidance to continue to improve the measurement qualities of future FC fingerprinting investigations.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Gowtham Atluri https://orcid.org/0000-0001-5619-6688

REFERENCES

Abbas, K., Amico, E., Svaldi, D. O., Tipnis, U., Duong-Tran, D. A., Liu, M., … Goñi, J. (2020). GEFF: Graph embedding for functional fingerprinting. *NeuroImage*, *221*, 117181. https://doi.org/10.1016/j.neuroimage.2020.117181

Airan, R. D., Vogelstein, J. T., Pillai, J. J., Caffo, B., Pekar, J. J., & Sair, H. I. (2016). Factors affecting characterization and localization of inter-individual differences in functional connectivity using MRI. *Human Brain Mapping*, *37*(5), 1986–1997. https://doi.org/10.1002/hbm.23150

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, *24*(3), 663–676. https://doi.org/10.1093/cercor/bhs352

Amico, E., & Goñi, J. (2018). The quest for identifiability in human functional connectomes. *Scientific Reports*, *8*(1), 8254. https://doi.org/10.1038/s41598-018-25089-1

Balakrishnan, N., & Nevzorov, V. B. (2004). *A primer on statistical distributions*. USA: John Wiley & Sons.

Bandettini, P. A. (2012). Twenty years of functional MRI: The science and the stories. *NeuroImage*, *62*(2), 575–588. https://doi.org/10.1016/j.neuroimage.2012.04.026

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *360*(1457), 1001–1013. https://doi.org/10.1098/rstb.2005.1634

Chen, S., & Hu, X. (2018). Individual identification using the functional brain fingerprint detected by the recurrent neural network. *Brain Connectivity*, *8*(4), 197–204. https://doi.org/10.1089/brain.2017.0561

Craddock, R. C., James, G. A., Holtzheimer, P. E., III, 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. https://doi.org/10.1002/hbm.21333

Damoiseaux, J. S., Rombouts, S. A., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences*, *103*(37), 13848–13853. https://doi.org/10.1073/pnas.0601417103

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, *20*(6), 425–443. https://doi.org/10.1016/j.tics.2016.03.014

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., … Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

Fornito, A., & Bullmore, E. T. (2015). Reconciling abnormalities of brain network structure and function in schizophrenia. *Current Opinion in Neurobiology*, *30*, 44–50. https://doi.org/10.1016/j.conb.2014.08.006

Fox, M. D., & Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience*, *4*, 19. https://doi.org/10.3389/fnsys.2010.00019

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. https://doi.org/10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … Jenkinson, M. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., … Hampton, J. M. (2017). Precision functional mapping of individual human brains. *Neuron*, *95*(4), 791–807. https://doi.org/10.1016/j.neuron.2017.07.011

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, *100*(1), 253–258. https://doi.org/10.1073/pnas.0135058100

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., … Smith, S. M. (2014). ICA-based artefact and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, *95*, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034

Handwerker, D. A., Roopchansingh, V., Gonzalez-Castillo, J., & Bandettini, P. A. (2012). Periodic changes in fMRI connectivity. *NeuroImage*, *63*(3), 1712–1719. https://doi.org/10.1016/j.neuroimage.2012.06.078

HCP Documentation. (2017). Retrieved from https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP1200-DenseConnectome+PTN+Appendix-July2017.pdf

Horien, C., Noble, S., Finn, E. S., Shen, X., Scheinost, D., & Constable, R. T. (2018). Considering factors affecting the connectome-based identification process: Comment on Waller et al. *NeuroImage*, *169*, 172–175. https://doi.org/10.1016/j.neuroimage.2017.12.045

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M. Y., … Schlaggar, B. L. (2015). Functional system and areal organization of a highly sampled individual human brain. *Neuron*, *87*(3), 657–670. https://doi.org/10.1016/j.neuron.2015.06.037

Li, K., & Atluri, G. (2018). Towards effective functional connectome fingerprinting. In *International workshop on Connectomics in neuroimaging* (pp. 107–116). Cham: Springer. https://doi.org/10.1007/978-3-030-00755-3_12

Liu, J., Liao, X., Xia, M., & He, Y. (2018). Chrronectome fingerprinting: Identifying individuals and predicting higher cognitive functions using

dynamic brain connectivity patterns. *Human Brain Mapping, 39*(2), 902–915. https://doi.org/10.1002/hbm.23890

Liu, T. T. (2016). Noise contributions to the fMRI signal: An overview. *NeuroImage, 143*, 141–151. https://doi.org/10.1016/j.neuroimage.2016.09.008

Lori, N. F., Ramalhosa, I., Marques, P., & Alves, V. (2018). Deep learning based pipeline for fingerprinting using brain functional MRI connectivity data. *Procedia Computer Science, 141*, 539–544. https://doi.org/10.1016/j.procs.2018.10.129

Miranda-Dominguez, O., Mills, B. D., Carpenter, S. D., Grant, K. A., Kroenke, C. D., Nigg, J. T., & Fair, D. A. (2014). Connectotyping: Model based fingerprinting of the functional connectome. *PLoS One, 9*(11), e111048. https://doi.org/10.1371/journal.pone.0111048

Peña-Gómez, C., Avena-Koenigsberger, A., Sepulcre, J., & Sporns, O. (2018). Spatiotemporal network markers of individual variability in the human functional connectome. *Cerebral Cortex, 28*, 1–13. https://doi.org/10.1093/cercor/bhx170

Poldrack, R. (2021). Diving into the deep end: A personal reflection on the MyConnectome Study. *Current Opinion in Behavioral Sciences 40*, 1–4.

Poldrack, R. A. (2017). Precision neuroscience: Dense sampling of individual brains. *Neuron, 95*(4), 727–729. https://doi.org/10.1016/j.neuron.2017.08.002

Preti, M. G., Bolton, T. A., & Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage, 160*, 41–54. https://doi.org/10.1016/j.neuroimage.2016.12.061

Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., ... Jenkinson, M. (2014). MSM: A new flexible framework for multimodal surface matching. *NeuroImage, 100*, 414–426. https://doi.org/10.1016/j.neuroimage.2014.05.069

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith, S. M. (2014). Automatic Denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage, 90*, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Satterthwaite, T. D., Xia, C. H., & Bassett, D. S. (2018). Personalized neuroscience: Common and individual-specific features in functional brain networks. *Neuron, 98*(2), 243–245. https://doi.org/10.1016/j.neuron.2018.04.007

Shehzad, Z., Kelly, A. C., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., ... Petkova, E. (2009). The resting brain: Unconstrained yet reliable. *Cerebral Cortex, 19*(10), 2209–2229. https://doi.org/10.1093/cercor/bhn256

Shojaee, A., Li, K., & Atluri, G. (2019). A machine learning framework for accurate functional connectome fingerprinting and an application of a siamese network. In *International workshop on connectomics in neuroimaging* (pp. 83–94). Cham: Springer. https://doi.org/10.1007/978-3-030-32391-2_9

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., ... Kelly, M. (2013). Resting-state fMRI in the human connectome project. *NeuroImage, 80*, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. London, England: Pearson Education.

Venkatesh, M., Jaja, J., & Pessoa, L. (2020). Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *NeuroImage, 207*, 116398. https://doi.org/10.1016/j.neuroimage.2019.116398

Waller, L., Walter, H., Kruschwitz, J. D., Reuter, L., Müller, S., Erk, S., & Veer, I. M. (2017). Evaluating the replicability, specificity, and generalizability of connectome fingerprints. *NeuroImage, 158*, 371–377. https://doi.org/10.1016/j.neuroimage.2017.07.016

Wig, G. S., Laumann, T. O., & Petersen, S. E. (2014). An approach for parcellating human cortical areas using resting-state correlations. *NeuroImage, 93*, 276–291. https://doi.org/10.1016/j.neuroimage.2013.07.035

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. https://doi.org/10.1093/cercor/bhw241

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Li K, Wisner K, Atluri G. Feature selection framework for functional connectome fingerprinting. *Hum Brain Mapp*. 2021;42:3717–3732. https://doi.org/10.1002/hbm.25379