Local White Matter Architecture Defines Functional Brain Dynamics

Yo Joong Choe*
Kakao
Seongnam-si, South Korea
yj.c@kakaocorp.com

Sivaraman Balakrishnan
Dept. of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, USA
siva@stat.cmu.edu

Aarti Singh
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, PA, USA
aarti@cs.cmu.edu

Jean M. Vettel
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD, USA
jean.m.vettel.civ@mail.mil

Timothy Verstynen
Dept. of Psychology and CNBC
Carnegie Mellon University
Pittsburgh, PA, USA
timothyv@andrew.cmu.edu

Abstract—**Large bundles of myelinated axons, called white matter, anatomically connect disparate brain regions together and compose the structural core of the human connectome. We recently proposed a method of measuring the local integrity along the length of each white matter fascicle, termed the local connectome [1]. If communication efficiency is fundamentally constrained by the integrity along the entire length of a white matter bundle [2], then variability in the functional dynamics of brain networks should be associated with variability in the local connectome. We test this prediction using two statistical approaches that are capable of handling the high dimensionality of data. First, by performing statistical inference on distance-based correlations, we show that similarity in the local connectome between individuals is significantly correlated with similarity in their patterns of functional connectivity. Second, by employing variable selection using sparse canonical correlation analysis and cross-validation, we show that segments of the local connectome are predictive of certain patterns of functional brain dynamics. These results are consistent with the hypothesis that structural variability along axon bundles constrains communication between disparate brain regions.**

Index Terms—**local connectome, structure-function relationship, high-dimensional statistics, canonical correlation analysis**

I. Introduction

The function of macroscopic neural networks is constrained by the integrity of structural connections between disparate regions. This form of long-distance (i.e., centimeters) communication relies on dense bundles of axons that are known as white matter [3]. To prevent degradation of action potentials across long distances, these fiber bundles are supported by the myelin sheath, non-neuronal glial cells that insulate axons and facilitate communication along the fascicle. As a result, the integrity of the myelin sheath is critical for synchronizing information transmission between distal brain areas [2], fostering the ability of these networks to adapt over time [4]. Thus, variability in the myelin sheath, as well as other cellular support mechanisms, would contribute to variability in functional coherence across the circuit.

To study the integrity of structural connectivity, we recently introduced the concept of the local connectome. This is defined as the pattern of fiber systems (i.e., number of fibers, orientation, and size) within a voxel, as well as immediate connectivity between adjacent voxels, that can be quantified using diffusion MRI (dMRI) by measuring the fiber-wise density of microscopic water diffusion within a voxel [1]. The collection of these multi-fiber diffusion density measurements within all white matter voxels is termed the local connectome fingerprint (LCF). The LCF is a high-dimensional feature vector that describes the unique configuration of the structural connectome along the segments of white matter pathways [5]. Thus, the LCF provides a diffusion-informed measure along the fascicles that supports inter-regional communication, rather than determining the start and end positions of a particular fiber bundle.

Since the LCF measures the local integrity along white matter bundles that connect regions across the entire brain, it reflects the overall communication capacity of the brain [2]. Hence, we expect to see that variations in the LCF should also correlate with those in the dynamics of brain networks, measured by connectivity patterns in the resting-state functional MRI (fMRI). To formally validate this intuition, we employ statistical approaches to examine the following hypotheses:

Hypothesis 1 Similarity in the LCF, between individuals, is associated with similarity in their functional connectivity patterns measured with resting-state fMRI.

Hypothesis 2 Variability in specific segments of the LCF is associated with patterns of functional connectivity in specific circuits.

TABLE I: Summary of Abbreviations and Notation

Notation	Definition
LCF	Local connectome fingerprint
FCG	Functional correlation graph
HCP	The Human Connectome Project (dataset)
n	Number of subjects (793)
p	Dimension of LCF vectors (433,386)
q	Dimension of FCG vectors (195,625)
$\mathbf{x}_i$	p-dimensional LCF vector of subject i
$\mathbf{y}_i$	q-dimensional FCG vector of subject i
$d_{\mathcal{X}}$	The scaled Euclidean distance between LCFs (1)
$d_{\mathcal{Y}}$	The Pearson correlation distance between FCGs (2)
$\mathbf{D}^{\mathcal{X}}$	Scaled Euclidean distance matrix between n LCFs
$\mathbf{D}^{\mathcal{Y}}$	Pearson correlation distance matrix between n FCGs
$\mathbf{X}$	$n \times p$ matrix containing the n LCFs as rows
$\mathbf{Y}$	$n \times q$ matrix containing the n FCGs as rows
$\|\cdot\|_1, \|\cdot\|_2$	The ℓ_1- and ℓ_2-norm of a real-valued vector
$\mathbb{S}_n$	The permutation group on $\{1, \ldots, n\}$

Fig. 1: Pipeline for generating LCFs. See [5] for details.

II. MATERIALS AND METHODS

We summarize our abbreviations and notation in Table I.

A. Data Acquisition

1) Participants: We used publicly available dMRI and fMRI data from the S900 (2015) release of the Human Connectome Project (HCP) [6], acquired by Washington University in St. Louis and the University of Minnesota. Out of the 900 participants released, 841 genetically unrelated participants (370 male, ages 22-37, mean age 28.76) had viable dMRI datasets. Among them, $n = 793$ participants had at least one viable resting-state fMRI measurement. Our analysis was restricted to this subsample. All data collection procedures were approved by the institutional review boards at Washington University in St. Louis and the University of Minnesota. The post hoc data analysis was approved as exempt by the institutional review board at Carnegie Mellon University, in accordance with 45 CFR 46.101(b)(4) (IRB Protocol Number: HS14-139).

2) Diffusion MRI Acquisition: The dMRI data were acquired on a Siemens 3T Skyra scanner using a 2D spin-echo single-shot multiband EPI sequence with a multi-band factor of 3 and monopolar gradient pulse. The spatial resolution was 1.25 mm isotropic (TR = 5500 ms, TE = 89.50 ms). The b-values were 1000, 2000, and 3000 s/mm2 . The total number of diffusion sampling directions was 90 for each of the three shells in addition to 6 b0 images. The total scanning time was approximately 55 minutes.

3) LCF Reconstruction: An outline of the pipeline for generating LCFs is shown in Fig. 1. The dMRI data for each subject was reconstructed in a common stereotaxic space using q-space diffeomorphic reconstruction (QSDR) [7], a nonlinear registration approach that directly reconstructs water diffusion density patterns into a common stereotaxic space at 1 mm resolution. The LCF reconstruction was conducted using DSI Studio (http://dsi-studio.labsolver.org), an open-source diffusion MRI analysis tool for connectome analysis. To compute the LCF, the axonal direction in each voxel was d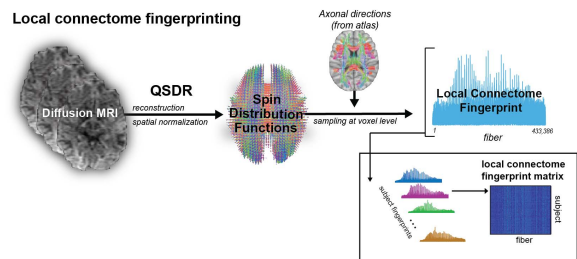erived from the HCP dataset, and all of the data and source code for this analysis are publicly available on the same website.

A spin distribution function (SDF) sampling framework was used to provide a consistent set of directions to sample the magnitude of SDFs along axonal directions in the cerebral white matter. Since each voxel may have more than one axonal direction, multiple measurements were extracted from the SDF for voxels that contained crossing fibers, while a single measurement was extracted for voxels with fibers in a single direction. The appropriate number of density measurements from each voxel was sampled by the left-posterior-superior voxel order and compiled into a sequence of scalar values. Gray matter was excluded using the ICBM-152 white matter mask (MacConnell Brain Imaging Centre, McGill University, Canada). The cerebellum was also excluded due to different slice coverage in cerebellum across participants. Since the density measurement has arbitrary units, the LCF was scaled to make the variance equal to 1 [5]. For each subject $i = 1, \ldots, n$, we denote this high-dimensional LCF of the ith subject, across $p = 433,386$ sampled directions, as $\mathbf{x}_i \in \mathbb{R}^p$. The collection of all n LCFs are compactly represented as a data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ with each LCF as a row vector.

4) Functional MRI Acquisition & Processing: We analyzed the minimally processed resting-state fMRI data acquired as part of the Human Connectome Project (HCP) [8], [6] which used a multi-band gradient echo-planar imaging protocol (see [9] for details on aquisition parameters). The dataset contains volumetric NIFTI data for resting-state fMRI scans (14 minutes each), motion parameters, and physiological data. Only data for the first resting-state scan collected at the A-P phase encoding direction were used for analyses. Using these measurements, we computed the average BOLD (blood-oxygen-level dependent) signals at each of the 626 regions of interest (ROIs) [10] and regressed out the linear effects of the noise terms via ordinary least-squares (OLS). The 16 noise terms include the global signal, 12 motion parameters (6 estimates from a rigid-body transformation to the SBRef image acquired at the start of each scan; 6 temporal derivatives of these estimates), and the top-3 principal component projections of the voxel-level white matter signals (measured at each of 2,258 voxels and 840 seconds). The resulting residual terms were then filtered by a first-order Butterworth bandpass filter

[11] between frequencies 0.08Hz and 0.15Hz.

5) Functional Correlation Graph Construction: For each subject, given a pre-processed time series (840s; 1Hz) at each ROI, we computed the functional correlation graph (FCG), alternatively called the functional connectome fingerprint in [12], by computing the Pearson correlation between time series at every pair of ROIs. For each subject $i = 1, \ldots, n$, we use $\mathbf{y}_i \in \mathbb{R}^q$ to denote the vector of these $q = \binom{626}{2} = 195,625$ Pearson correlations, which we collectively refer to as the ith FCG. The collection of all n FCGs are compactly represented as a data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times q}$ with each FCG as a row vector.

B. Statistical Inference of Distance-based Correlations

Our first goal is to test whether there is a statistically significant relationship between LCFs and FCGs. However, because both the structural and functional feature vectors are high-dimensional, fully multivariate statistical tests of dependence are intractable and uninterpretable. This means that we need to find a way to effectively reduce the dimensionality of each feature vector.

For each pair of subjects, we first compute the pairwise distance between their feature vectors. This gives us one distance matrix between their structural features (LCFs) and another between their functional features (FCGs). Then, we measure the correlation between the resulting pair of structural and functional distance matrices.

Our hypothesis states that if two subjects have similar LCFs, then they are more likely to also have similar FCGs. This hypothesis derives from previous research that found (a) similar LCFs imply genetic similarity [5] and (b) identical FCGs imply that the two graphs most likely come from the same individual [12]. By formally defining a notion of similarity, it is possible to derive distribution-free statistical inference methods that can test whether the two high-dimensional feature vectors are correlated or not. This approach overcomes the high dimensionality while being statistically and theoretically rigorous.

1) Choice of Distance Metrics: In [5], Yeh et al. establish that LCFs are highly specific to each individual. More precisely, they show that the Euclidean distance between any pair of LCFs effectively captures the genetic (and temporal) difference between the two measurements, achieving 100% accuracy across 17,398 leave-one-out identification tasks. Therefore, to quantify individual variability in structural features, we use the Euclidean distance, scaled by the number of features as in [5]:

$$d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \frac{1}{p} \|\mathbf{x} - \mathbf{x}'\|_2 = \frac{1}{p} \sqrt{\sum_{k=1}^{p} (x_k - x'_k)^2} \quad (1)$$

To estimate distance between functional features, we follow the approach that Finn et al. [12] used on FCGs of the Q2-released version of the HCP dataset. They successfully predicted identity with 92.9-94.4% test set accuracy using the Pearson correlation, and the accuracy increased to 98-99% when comparing specific sub-networks (the medial frontal network and the frontoparietal network). Since our goal is to capture individual variability, not maximize prediction accuracy, we use the Pearson correlation distance on the entire FCG:

$$
\begin{aligned}
d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') &= 1 - \rho(\mathbf{y}, \mathbf{y}') \\
&= 1 - \frac{\sum_{l=1}^{q} (y_l - \overline{\mathbf{y}})(y'_l - \overline{\mathbf{y}'})}{\sqrt{\sum_{l=1}^{q} (y_l - \overline{\mathbf{y}})^2} \sqrt{\sum_{l=1}^{q} (y'_l - \overline{\mathbf{y}'})^2}}
\end{aligned} \quad (2)
$$

where ρ denotes the Pearson correlation and $\overline{\mathbf{y}} = \frac{1}{q} \sum_{l=1}^{l} y_l$ denotes the mean of all entries in the vector $\mathbf{y}$. We note that $d_{\mathcal{Y}}$ is not a proper distance metric in the mathematical sense, because it does not satisfy positive definiteness or triangle inequality. It is nevertheless nonnegative, symmetric, and is exactly zero when the two inputs are identical (it is also zero when two inputs are scalar multiples of each other).

Given these choices of metrics, we can represent all such distances on our data compactly in two $n \times n$ distance matrices, $\mathbf{D}^{\mathcal{X}} \in \mathbb{R}^{n \times n}$ and $\mathbf{D}^{\mathcal{Y}} \in \mathbb{R}^{n \times n}$, such that $\mathbf{D}_{ij}^{\mathcal{X}} = d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{D}_{ij}^{\mathcal{Y}} = d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j)$.

2) Setting Up a Valid Hypothesis Test: In general, it is highly nontrivial to set up a proper statistical test comparing distance matrices, because the entries of each distance matrix are not independent from each other. Intuitively, for any pair of subjects i and j, the distance between the ith and jth feature vectors is correlated with the distance between the ith feature vector and any other feature vector. Thus, standard statistical approaches that rely on the i.i.d. assumption will *not* give valid results if they are naïvely applied to distance matrices.

We instead use the distance matrices to construct null and alternative hypotheses and derive proper statistical inference strategies. Given independent copies of random vectors $(\mathbf{x}, \mathbf{y}) \sim P_{\mathcal{X}\mathcal{Y}}$, where $P_{\mathcal{X}\mathcal{Y}}$ is the joint distribution of $\mathbf{x}$ and $\mathbf{y}$, we test

$$H_0 : \mathcal{R}(\mathbf{x}, \mathbf{y}) = 0 \quad \text{and} \quad H_1 : \mathcal{R}(\mathbf{x}, \mathbf{y}) > 0 \quad (3)$$

where

$$
\begin{aligned}
\mathcal{R}(\mathbf{x}, \mathbf{y}) &= \rho(d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'), d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')) \\
&= \frac{\mathsf{Cov}\left(d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'), d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')\right)}{\sqrt{\mathsf{Var}\left(d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')\right)} \sqrt{\mathsf{Var}\left(d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')\right)}}
\end{aligned} \quad (4)
$$

In short, $\mathcal{R}$ is the Pearson correlation between the two random distances, each of which is a function of two independent and identically distributed random variables. In our approach, the null hypothesis states that the Euclidean distance between the LCFs of two subjects is uncorrelated with the correlation distance between their FCGs. The alternative hypothesis states that the two distances are in fact positively correlated. Note that it is natural to consider a one-sided hypothesis here because we know that both distances are likely to increase as two subjects become more genetically distant [5], [12].

While there are no known parametric statistical tests corresponding to (3), we can extend the permutation test of Pearson correlation in standard linear regression to our case. Given the structural and functional distance matrices $\mathbf{D}^{\mathcal{X}} \in$

$\mathbb{R}^{n \times n}$ and $\mathbf{D}^{\mathcal{Y}} \in \mathbb{R}^{n \times n}$, let $\overline{\mathbf{D}^{\mathcal{X}}} = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{D}_{ij}^{\mathcal{X}}$ and $\overline{\mathbf{D}^{\mathcal{Y}}} = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{D}_{ij}^{\mathcal{Y}}$, where $\Sigma_{i<j}$ denotes the double sum $\sum_{i=1}^{n} \sum_{j=i+1}^{n}$. Then, the sample test statistic for (3) is given by

$$\widehat{\mathcal{R}}_n(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i<j} \left(\mathbf{D}_{ij}^{\mathcal{X}} - \overline{\mathbf{D}^{\mathcal{X}}} \right) \left(\mathbf{D}_{ij}^{\mathcal{Y}} - \overline{\mathbf{D}^{\mathcal{Y}}} \right)}{\sqrt{\sum_{i<j} \left(\mathbf{D}_{ij}^{\mathcal{X}} - \overline{\mathbf{D}^{\mathcal{X}}} \right)^2} \sqrt{\sum_{i<j} \left(\mathbf{D}_{ij}^{\mathcal{Y}} - \overline{\mathbf{D}^{\mathcal{Y}}} \right)^2}} \tag{5}$$

Given (5), a permutation test can be devised by randomly shuffling one of the feature vectors (say $\mathbf{x}$, without loss of generality) among the n subjects. This corresponds to permuting the rows of the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Mathematically, for a random permutation $\sigma \in \mathbb{S}_n$ of n elements, the empirical distribution of *permuted* correlations

$$\widehat{\mathcal{R}}_{n,\sigma}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i<j} \left(\mathbf{D}_{\sigma(i),\sigma(j)}^{\mathcal{X}} - \overline{\mathbf{D}^{\mathcal{X}}} \right) \left(\mathbf{D}_{ij}^{\mathcal{Y}} - \overline{\mathbf{D}^{\mathcal{Y}}} \right)}{\sqrt{\sum_{i<j} \left(\mathbf{D}_{ij}^{\mathcal{X}} - \overline{\mathbf{D}^{\mathcal{X}}} \right)^2} \sqrt{\sum_{i<j} \left(\mathbf{D}_{ij}^{\mathcal{Y}} - \overline{\mathbf{D}^{\mathcal{Y}}} \right)^2}} \tag{6}$$

estimates the null distribution of $\mathcal{R}(\mathbf{x}, \mathbf{y})$. If the sample correlation (5) deviates from this null distribution significantly, then we can reject the null hypothesis of the test in (3).

This test can be viewed as a variant of the Mantel test [13], which jointly permutes both features among the n subjects to test the same statistic. Yet, because our version does not permute the feature dimension, it does not introduce unintended bias coming from spatial correlations [14].

Note that a nonzero correlation will imply statistical dependence, but not the other way around. When we take $d_{\mathcal{Y}}$ to be the Euclidean distance instead of the correlation distance, however, we obtain distance correlation (dCor) [15], where a zero value implies statistical independence. We will consider the statistical test (3) both when $d_{\mathcal{Y}}$ is the correlation distance and when $d_{\mathcal{Y}}$ is the Euclidean distance. In the latter case, we use the unbiased version of the statistic that leads to a t-test [16].[1]

3) Constructing a Valid Confidence Interval with Subsampling: The permutation test is nonparametric, but it does not readily yield confidence intervals unless a stronger assumption (and tedious computation) is made [17], [18]. Subsampling [19] is an alternative approach to statistical inference that makes less assumptions and gives a confidence interval as its outcome. It estimates the true distribution of $\mathcal{R}(\mathbf{x}, \mathbf{y})$ by computing the empirical version of the statistic many times on different random subsamples of the full data.

Subsampling notably differs from the more standard bootstrapping because it samples *without* replacement and only samples a fraction of the n data points. The first difference is crucial in our scenario, because any duplicate sample from

bootstrapping will zero out entries of $\mathbf{D}^{\mathcal{X}}$ and $\mathbf{D}^{\mathcal{Y}}$ and thus lead to a biased (higher) estimate of $\mathcal{R}(\mathbf{x}, \mathbf{y})$.

C. High-Dimensional Canonical Correlation Analysis with Cross-Validation

While statistical inference of the distance-based correlation will provide some insights to the structure-function relationship, this measure of correlation aggregated over so many features may not be as intuitive or informative. In search of more detailed and interpretable relationships between the two sets of features, we attempt to find small subsets of the LCF that are predictive of small subsets of the FCG *on a held-out set*.

1) Canonical Correlation Analysis: For a pair of random vectors, canonical correlation analysis (CCA) [20] finds a pair of linear transformations ("alignments") onto the same Euclidean space such that the projections are the most correlated. Assuming *centered* data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times q}$, CCA solves the following biconvex constrained optimization problem:

$$\begin{aligned} \underset{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q}{\text{maximize}} \quad & \langle \mathbf{Xu}, \mathbf{Yv} \rangle \\ \text{subject to} \quad & \|\mathbf{Xu}\|_2^2 \leq 1 \\ & \|\mathbf{Yv}\|_2^2 \leq 1 \end{aligned} \tag{7}$$

The objective is often written alternatively as $\mathbf{u}^T \widehat{\Sigma}_{\mathbf{xy}} \mathbf{v}$, up to a $\frac{1}{n}$ constant, where $\widehat{\Sigma}_{\mathbf{xy}} = \frac{1}{n} \mathbf{X}^T \mathbf{Y}$ is the empirical cross-covariance matrix. When the columns of $\mathbf{X}$ and $\mathbf{Y}$ are further standardized, the solution to this biconvex problem is given by the left and right singular vectors of the empirical cross-covariance matrix $\widehat{\Sigma}_{\mathbf{xy}}$ that correspond to its largest singular value.

Intuitively, CCA captures the directions in $\mathcal{X}$ and $\mathcal{Y}$ that explain the largest cross-correlation. If we assume that $\mathbf{X}$ and $\mathbf{Y}$ indeed have some correlation structures, then CCA will find the linear transformations that recover such structures.

2) Sparse CCA: In high dimensions, i.e. when the data dimensions p and q are large compared to the sample size n, the estimate $\widehat{\Sigma}_{\mathbf{xy}}$ of the true cross-covariance is no longer consistent unless more structural assumptions are made [21], [22]. It is also considered a more difficult problem than sparse PCA [23], [24], which itself is considered challenging due to the poor behavior of the sample covariance matrix as an estimate [25]. To obtain a reliable estimate of the high-dimensional cross-correlation structure, we assume that there are interesting low-dimensional correlation structures between subsets of the structural and functional features. This allows us to focus on a sparse subset of each set of features that are the most correlated to one another.

A popular approach to finding sparse subsets of features is to use ℓ^1-regularization. In our case, we add an ℓ_1-penalty to the alignment vectors in (7):

$$\|\mathbf{u}\|_1 \leq c_1 \quad \text{and} \quad \|\mathbf{v}\|_1 \leq c_2 \tag{8}$$

where $c_1, c_2 > 0$ are sparsity parameters. The ℓ_1-penalty, most commonly used in the Lasso [26], performs variable selection

---

[1] We implement the unbiased dCor t-test [16] by (substantially) modifying the MATLAB implementation found in http://mathworks.com/matlabcentral/fileexchange/39905-distance-correlation.

by forcing some of the entries to be precisely zero when the sparsity parameters are sufficiently small. A penalized version of CCA that combines (7) and (8) has been called sparse CCA in the literature, and an alternating convex optimization algorithm can be used to find a sparse solution [27], [28].

Yet, the ℓ_1-penalty alone is not sufficient for effective variable selection in our setting. One reason is that both the LCF and the FCG naturally contain interesting correlation structures within their entries, while ℓ_1-regularization tends to select only one entry from a correlated group [29]. Another reason is that ℓ_1-penalized CCA from (7) and (8) is not strictly biconvex in high dimensions, so that the optimization problem can be unstable. Both of these issues can be alleviated by further including an ℓ_2-penalty:

$$\|\mathbf{u}\|_2 \leq d_1 \quad \text{and} \quad \|\mathbf{v}\|_2 \leq d_2 \qquad (9)$$

with constants $d_1, d_2 > 0$.[2] The resulting optimization problem can be viewed as the elastic net [29] applied to CCA. It is now a strictly biconvex problem, and we can find a feasible solution efficiently by alternately applying existing convex optimization solvers. We note that, in general, there is no known algorithm for this biconvex problem that guarantees a globally optimal solution [27]. For our analysis, we use the MATLAB implementation from [30] that is freely available online.[3]

3) k-Fold Cross-Validation: For sparse CCA, we use k-fold cross-validation to find the set of sparsity parameters that give the highest canonical correlations between subsets of the LCFs and the FCGs.

Specifically, using $k = 5$, we first split the n subjects into training and test sets with the ratio of 5 to 1. Then, we randomly partition the training set (size $\lceil 5/6 \rceil n$) into 5 equally sized subsamples, fit sparse CCA with each candidate set of sparsity parameters to 4 of the subsamples, and use the fitted alignment vectors $\mathbf{u}$ and $\mathbf{v}$ to align the feature vectors from the unused subsample (i.e. the validation set). The resulting canonical correlation on the validation set can be viewed as an estimate of canonical correlation on unseen data. By leaving out each of the 5 subsamples in the previous step, we obtain 5 such estimates of the canonical correlation, and the average of these 5 estimates can be used to validate the performance of the candidate set of sparsity parameters. After these steps, we choose the set of sparsity parameters that give the largest average canonical correlation on the validation set.

The resulting alignment vectors can transform unseen feature vectors coming from the same distribution as our dataset, so that the LCFs are the most correlated to the FCGs in the transformed space. The final performance of these alignment vectors is measured by the correlation between the alignments of the test set, which was unused throughout the cross-validation steps.
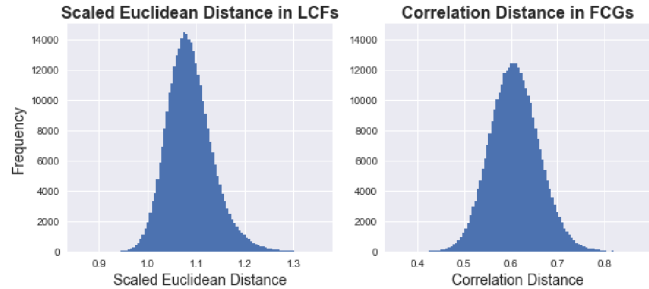

Fig. 2: Pairwise distances between pairs of subjects' LCFs (left) and FCGs (right).

TABLE II: Summary of Statistical Inference Results ($n = 793$)

Method	Correlation	Result Type	Result
Permutation (6)	0.120	*p*-value	**< 0.001*****
dCor *t*-test [16]	0.252	*p*-value	**< 0.001*****
Subsampling	0.120	95% conf. int.	**(0.098, 0.141)+**

III. RESULTS

A. Exploratory Analysis

We first present exploratory analysis results for the inter-subject distances in LCFs and FCGs. Fig. 2 shows that the feature distances between different subjects appear substantially distant from zero. This in part reproduces the results from [5] and [12], in which it is shown that the distances between different individuals are significantly greater than those between the same subjects. This justifies our choice of distances (1) and (2) for the permutation test as well as the subsampling-based confidence interval.

B. Statistical Inference

In Table II, we summarize our results from the permutation test, the dCor *t*-test, and the subsampling-based confidence interval. Significance levels are marked with * ($p < .05$), ** ($p < .01$), and *** ($p < .001$). Significant confidence intervals are marked with $^+$. We used $100,000$ random permutations for the hypothesis tests as well as $100,000$ subsamples for the confidence interval construction. Subsampling ratio was chosen as 0.135, following the procedure in [31].

Using a significance level of $\alpha = 0.05$, we find from the permutation test that there is indeed a statistically significant correlation between the Euclidean distances in LCFs and the correlation distances in FCGs. The dCor *t*-test of independence confirms that the two sets of features are statistically dependent, despite the fact that the test makes strong assumptions. Further, because the 95% confidence interval does not include zero, we conclude that the correlation between LCF distances and FCG distances is statistically significant.

Each of these results indicate that the similarity in the local connectome between individuals is significantly correlated to the similarity in their functional connectivity patterns. Specifically, our results show that if two individuals have similar
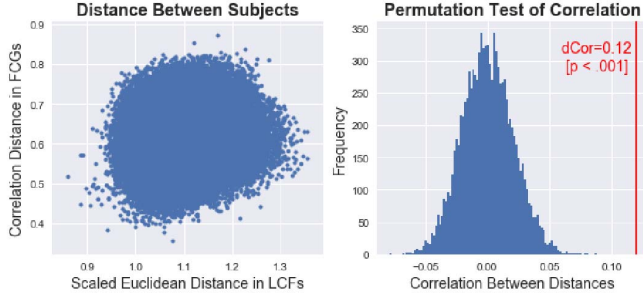
---

[2]For simplicity, we fix these constants to be 1 in our analysis.
[3]http://people.stern.nyu.edu/xchen3/Code/groupCCA.zip

Fig. 3: (Left) Scatterplot of all pairwise distances between the LCFs (x-axis) and between the FCGs (y-axis). (Right) Simulated null distribution using 10,000 random permutations of subjects (6). Red vertical line indicates the correlation on the actual dataset (5). The p-value is the proportion of random correlations that fall on the right side of the red vertical line.
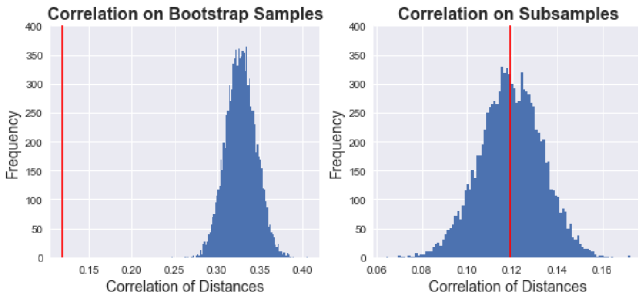

Fig. 4: Histogram of linear correlations for 10,000 bootstrap samples (left) and 10,000 subsamples (right) of the HCP ($n = 793$) dataset. Red vertical line indicates the correlation on the actual dataset (0.120).

local white matter architectures, they are also more likely to have similar functional brain dynamics.

Note that the correlation value for permutation test and subsampling are indeed identical, because they both compute exactly (5). The value in dCor t-test [16] differs, however, not only because the distance metric is changed to the Euclidean distances but because the test uses an unbiased estimate of the (Euclidean distance-based) statistic. While conceptually similar, the two computed values are estimates of different statistics and thus cannot be compared directly.

Fig. 3 visualizes the result from our permutation tests. On the left, we plot the structural and functional pairwise distances in a scatterplot to explore the overall trend. The scatterplot suggests that there is a positive trend between the pairwise distances in the structural and functional features. On the right, the permutation test shows that the correlation on real data is on the far-right tail of the correlation on simulated null data, suggesting that there is a statistically significant positive correlation between the structural and functional pairwise distances.

Nonparametric estimates of the correlation give analogous results (Spearman's ρ: 0.112, Kendall's τ: 0.075). This is not
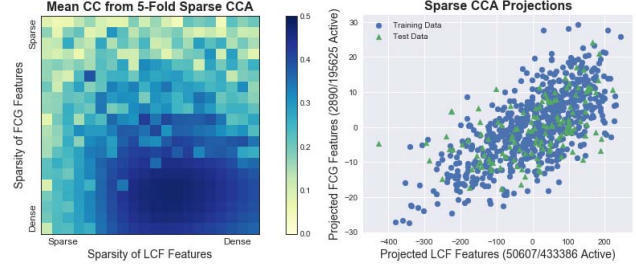

Fig. 5: Cross-validated sparse CCA projections on the HCP dataset. (Left) 5-fold cross-validation plot on a validation set. (Right) 2D Projections of the training and test set using the best parameter found. Canonical correlations: 0.689 (train), 0.515 (test).

surprising, given that the 2D scatterplot in Fig. 3 does not display an obvious nonlinear trend.

Fig. 4 justifies our use of subsampling instead of bootstrapping for our confidence intervals. As we described earlier, because each bootstrap sample contains multiple copies of the same subject, the resulting structural and functional distance matrices always contain many zeros, leading to a spuriously high correlation compared to the truth. The plots show that the bootstrap distribution fails to capture the actual correlation and is significantly biased upwards, while subsampling does not have this issue because it samples from the data without replacement.

C. Sparse CCA

For sparse CCA, we select a pair of sparsity (ℓ_1) parameters from a 2D grid, one for LCFs and another for FCGs, that yields the maximum canonical correlation on the validation set. Our cross-validation plot in the left panel of Fig. 5 shows that there is a contiguous region of sparsity levels in both structural and functional features where the canonical correlation on the validation set is maximized. Using the optimal regularization parameters, we find that sparse CCA selects 50,607 (11.7%) LCF features and 2,890 (1.48%) FCG features to give a canonical correlation of 0.689 (train) and 0.515 (test).[4]

The cross-validated sparse CCA projections of the training data as well as the test data are plotted in Fig. 5. Since the objective of CCA is to maximize the correlation between these projected points, we expect to see a linearly increasing pattern in the projected space. The right panel of Fig. 5 demonstrates this expectation: the projections of the training data and the test data exhibit similar linearly increasing patterns with a similar degree of variation. This implies that the alignment vectors we found can generalize well to unseen data in terms of correlation in the linearly projected space.

Note that the projections still have relatively high variance across the linear trend. This variability is likely due to both

---

[4]As described in Section II-C3, the final training canonical correlation is computed on all 5 folds (size $\lceil 5/6 \rceil n$) of the training set. The test canonical correlation is computed on the held-out test set (size $\lfloor 1/6 \rfloor n$), which is unused during cross-validation.

the variance coming from the optimization problem, which is ill-conditioned and thus contains many local optima, and the variance coming from the lack of statistical consistency in the high-dimensional setting. Indeed, the optimal number of variables chosen by cross-validation (50,607 and 2,890) is still greater than the number of subjects (793).

In Fig. 6, we visualize the LCF and FCG features selected by sparse CCA using the optimal sparsity parameters. In both modalities, sparse CCA focuses on connectivity patterns in specific regions of the brain. In particular, within the high-dimensional LCF space, the algorithm points to contiguous local pathways of the white matter structure. Our results show that this specific set of local white matter pathways are highly predictive of the lateral dynamics of functional connectivity between the left and right hemispheres. The structure-function association is observed between the core white matter pathways that regulate both intracortical and cortical-subcortical communication, including the corpus callosum, thalamic radiations, corticospinal, and corona radiata pathways, and the resting state functional activity in a diversity of cortical and subcortical nodes. This suggests that the structure-function relationship is strongest in the large major communication fascicles that are critical for global brain network communication.

D. Canonically Correlated Subcluster Pairs

In order to see if there is substructure in the structure-function relationships identified by sparse CCA, we decompose the canonical correlations into smaller subclusters of both the LCF and FCG entries. In Fig. 7, we show the three most canonically correlated pairs of LCF and FCG subclusters, which are computed by a simple agglomerative clustering (complete-linkage, same distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ respectively) of the selected LCF and FCG features into 5 subclusters each. We compute the canonical correlation between each pair of subclusters without additional regularization terms, as in (7).

Here we see even further specificity in the structure-function relationship. For example, variability in the centrum semiovale (Fig. 7, left), should predict functional dynamics of intrahemispheric and interhemispheric cortical networks. This pattern largely holds in the corresponding functional networks. In contrast, a small cluster along the inferior longitudinal fasiculus (Fig. 7, middle), a major means of communication along the ventral visual stream, correlates with primarily ventral visual pathway functional dynamics, as well as communication between dorsal and ventral visual streams. Finally, variability in the internal capsule (Fig. 7, right), a major means of communication between cortex and subcortical areas, correlates with primarily functional dynamics between cortical and subcortical nodes. Thus, the specificity of the structure-function relationships identified in this subclustering analysis is consistent with *a priori* predictions derived from the neuroanatomical literature.

IV. DISCUSSION

In this paper, we show how variability in local white matter architecture is associated with global patterns of functional
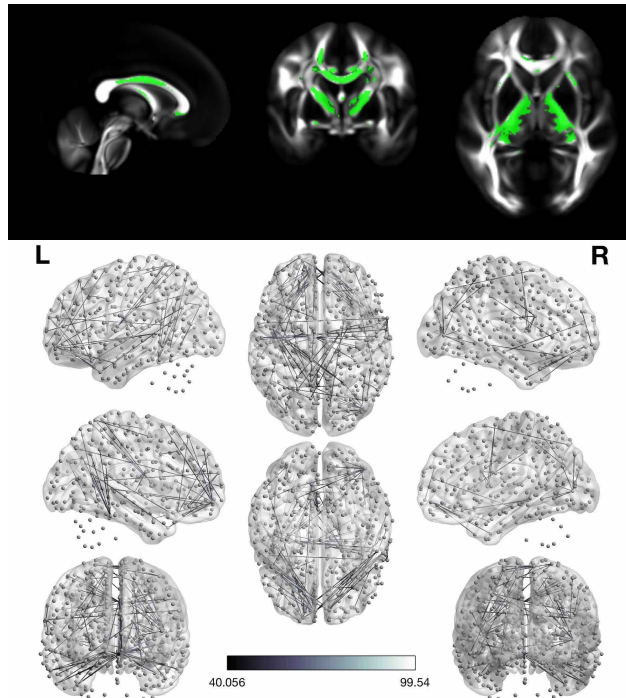

Fig. 6: Visualization of cross-validated CCA projections for LCFs in MRIcron (top) and FCGs in BrainNet (bottom). We select the features that gave the best testing canonical correlation in cross-validation (best parameter, best fold).

brain dynamics. Using distance-based correlations, we found a small, but significant effect whereby individuals with more similar local white matter architecture tended to also be more similar in their functional connectome. Using sparse CCA approaches, we were able to show that individual variability in white matter architecture along major brain fascicles correlated with individual differences in functional dynamics within the specific class of brain networks that would be predicted by existing neuroanatomical knowledge. Thus, in conjunction with the constraints of global end-to-end structural connectivity [10], our results highlight how variability in the local white matter systems also impacts global brain communication.

REFERENCES

[1] F.-C. Yeh, D. Badre, and T. Verstynen, "Connectometry: A statistical approach harnessing the analytical potential of the local connectome," *Neuroimage*, vol. 125, pp. 162–171, 2016.

[2] S. Pajevic, P. J. Basser, and R. D. Fields, "Role of myelin plasticity in oscillations and synchrony of neuronal activity," *Neuroscience*, vol. 276, pp. 135–147, 2014.

[3] R. E. Passingham, K. E. Stephan, and R. Kötter, "The anatomical basis of functional localization in the cortex," *Nature Reviews Neuroscience*, vol. 3, no. 8, p. 606, 2002.

[4] R. D. Fields, "A new mechanism of nervous system plasticity: activity-dependent myelination," *Nature Reviews Neuroscience*, vol. 16, no. 12, p. 756, 2015.

[5] F.-C. Yeh, J. Vettel, A. Singh, B. Poczos, S. Grafton, K. Erickson, W.-Y. I. Tseng, and T. Verstynen, "Quantifying differences and similarities in whole-brain white matter architecture using local connectome fin-
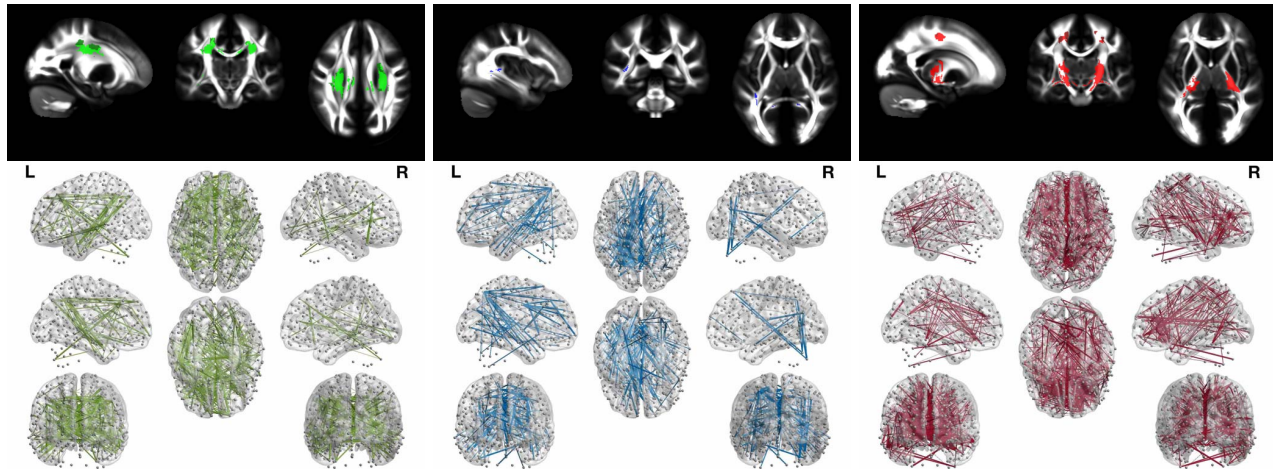
Fig. 7: Canonically correlated subcluster pairs between the selected LCF and FCG features.

gerprints," *PLoS Computational Biology*, vol. 12, no. 11, p. e1005203, 2016.

[6] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[7] F.-C. Yeh and W.-Y. I. Tseng, "Ntu-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction," *Neuroimage*, vol. 58, no. 1, pp. 91–99, 2011.

[8] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni *et al.*, "The minimal preprocessing pipelines for the human connectome project," *Neuroimage*, vol. 80, pp. 105–124, 2013.

[9] S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Uğurbil, "Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri," *Magnetic Resonance in Medicine*, vol. 63, no. 5, pp. 1144–1153, 2010.

[10] A. M. Hermundstad, D. S. Bassett, K. S. Brown, E. M. Aminoff, D. Clewett, S. Freeman, A. Frithsen, A. Johnson, C. M. Tipper, M. B. Miller *et al.*, "Structural foundations of resting-state and task-based functional connectivity in the human brain," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 6169–6174, 2013.

[11] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[12] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature neuroscience*, 2015.

[13] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[14] G. Guillot and F. Rousset, "Dismantling the mantel tests," *Methods in Ecology and Evolution*, vol. 4, no. 4, pp. 336–344, 2013.

[15] G. J. Székely, M. L. Rizzo, N. K. Bakirov *et al.*, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[16] G. J. Székely and M. L. Rizzo, "The distance correlation t-test of independence in high dimension," *Journal of Multivariate Analysis*, vol. 117, pp. 193–213, 2013.

[17] R. John and J. Robinson, "Significance levels and confidence intervals for permutation tests," *Journal of Statistical Computation and Simulation*, vol. 16, no. 3-4, pp. 161–173, 1983.

[18] P. H. Garthwaite, "Confidence intervals from randomization tests," *Biometrics*, pp. 1387–1393, 1996.

[19] D. N. Politis, J. P. Romano, and M. Wolf, "Subsampling springer series in statistics," 1999.

[20] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[21] Z. Bao, J. Hu, G. Pan, and W. Zhou, "Canonical correlation coefficients of high-dimensional normal vectors: finite rank case," *arXiv preprint arXiv:1407.7194*, 2014.

[22] I. M. Johnstone, "Multivariate analysis and jacobi ensembles: Largest eigenvalue, tracy–widom limits and rates of convergence," *Annals of statistics*, vol. 36, no. 6, p. 2638, 2008.

[23] M. Chen, C. Gao, Z. Ren, and H. H. Zhou, "Sparse cca via precision adjusted iterative thresholding," *arXiv preprint arXiv:1311.6186*, 2013.

[24] C. Gao, Z. Ma, Z. Ren, H. H. Zhou *et al.*, "Minimax estimation in sparse canonical correlation analysis," *The Annals of Statistics*, vol. 43, no. 5, pp. 2168–2197, 2015.

[25] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of statistics*, pp. 295–327, 2001.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[27] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, p. kxp008, 2009.

[28] E. Parkhomenko, D. Tritchler, and J. Beyene, "Genome-wide sparse canonical correlation of gene expression with genotypes," in *BMC proceedings*, vol. 1, no. 1. BioMed Central, 2007, p. S119.

[29] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[30] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis." in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[31] P. J. Bickel and A. Sakov, "On the choice of m in the m out of n bootstrap and confidence bounds for extrema," *Statistica Sinica*, pp. 967–985, 2008.