Finding relationships between structural and functional brain networks via connectome fingerprinting

Yo Joong Choe
PhD Student in Statistics and Machine Learning
Carnegie Mellon University
yjchoe@cmu.edu

Advisors: Aarti Singh, Timothy Verstynen, and Sivaraman Balakrishnan

Advanced Data Analysis (ADA) Final Report

Abstract

Neuroscientists are interested in how patterns of anatomical structure and functional activities in our brain are related. Some recent results have found evidences of correlations between structural and functional brain networks, although they rely on using standard connectivity features that fail to account for genetic variations as well as the intricate system of neural pathways that shape our brain. These results are also limited to finding correlations at the whole brain level or within the same region of interest, but not those across specific parts of each brain network that we may only discover through data.

In this work, we address these issues using connectome fingerprinting, a set of techniques that provide charaterizations of the brain network that are unique to each individual. For the structural brain network, we utilize local connectome fingerprinting (LCF), which gives a high-dimensional "fingerprint" that captures the unique patterns of local connections along neural pathways very well. For the functional brain network, we use the resting-state fMRI connectivity matrix, which has been proposed to be a functional version of a connectome fingerprint. Using these characterizations, we attempt to (1) test and give confidence to the correlation between the two high-dimensional modalities and (2) identify local structural connectivity patterns across neural pathways that are highly correlated to specific parts of the functional brain network.

1 Introduction

1.1 Background and motivation

The brain architecture is captured by structural neuroimaging techniques such as diffusion spectrum imaging (DSI) [1], which measures the diffusion of water in brain tissue, while neural activities are captured by functional neuroimaging techniques such as functional magnetic resonance imaging (fMRI) [2], which measures changes in blood flow within the brain. These measurements have been used in a variety of tasks, ranging from medical diagnosis to developmental psychology.

Neuroscientists have long been interested in the notion of *connectivity* in the brain, both in the structural and functional components. Recent results have emphasized its importance in understanding the human brain, as Hagmann in [3] claims that "...the huge brain neuronal communication capacity and computational power critically relies on this subtle and incredibly complex connectivity architecture."

In [4], Honey et al. defines **structural connectivity** as "macroscopic structural linkage, as obtained, for instance, from longrange tract tracing or diffusion imaging tractography" and **functional connectivity** as "the statistical dependence between time series describing the neural dynamics at distinct locations in the brain." Recent results using connectivity measures suggest that it is often these structural and functional connections between different parts of the brain that contain more useful information than the raw signals of water diffusion or blood oxygenation levels [5, 6]. In fact, research has shown that there is a high correlation between the structural and functional connectivity measurements and that the correlation is not uniform across different regions of the brain [7, 8, 9, 10].

In a recent work, Yeh et al. [11] introduced local connectome fingerprints (LCFs), a quantitative measure of the *local* white matter structure based on the density of water diffusion. Figure 1 describes how an LCF is computed from a diffusion MRI (in our case, it comes from DSI scans). These high-dimensional feature vectors were shown to be unique identifiers of individual genetic characteristics, hence the name "fingerprints." Figure 2 from [11] shows that the Euclidean distance between a pair of fingerprints are large if the two come from different individuals but very small if they come from the same person, even when measured at distant time points. It also shows that twins have much more similar fingerprints than strangers, further hinting at the possibility that these structural connectivity features capture the genetic characteristics of each individual. In an identification task, the fingerprints were used to achieve 100% accuracy.

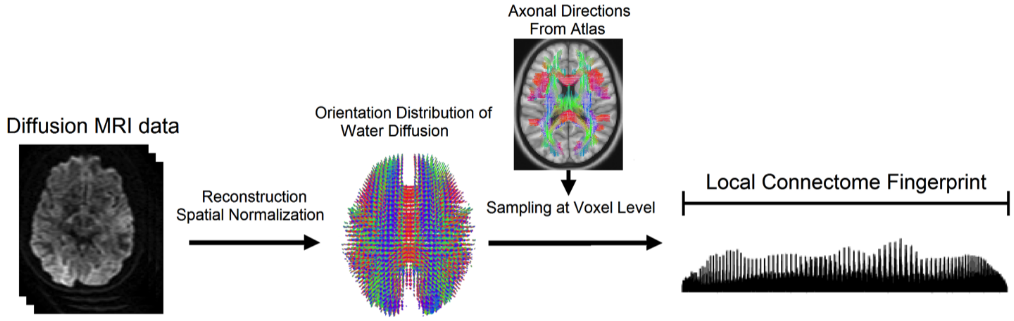
Diffusion MRI data
Reconstruction
Spatial Normalization
Orientation Distribution of Water Diffusion
Axonal Directions From Atlas
Sampling at Voxel Level
Local Connectome Fingerprint

Figure 1: Computing local connectome fingerprints (LCFs) that capture highly specific local white matter structure from diffusion MRI. Figure from [11].

In the functional domain, similar attempts were made to find features from fMRI connectivity data that can distinguish individuals. One recent attempt by Finn et al. in [12] showed that the Pearson correlation matrix between resting-state fMRI time series at pairs of regions achieved 90+% accuracy in identification tasks, showing that the fMRI correlation matrix may contain sufficient information about the individual variability of functional connectivity. Finn et al. defines a functional connectome fingerprint (FCF) as the upper-triangular part of the correlation matrix that gave such high identification accuracy.

Figure 3 summarizes the problem setup.

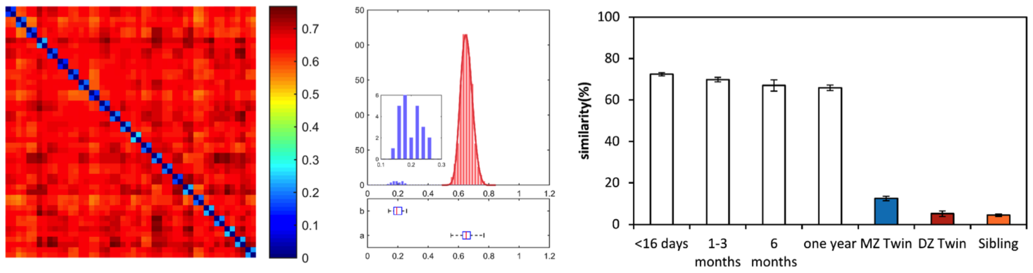

Figure 2: *Left:* Euclidean distances between pairs of local connectome fingerprints. Each subject is represented twice in adjacent rows/columns. The blue blocks on the diagonal indicate that two fingerprints measured from the same subject much closer than two from different subjects. *Middle:* A histogram of within-subject (blue) and between-subject (red) distances. *Right:* Similarity between pairs of fingerprints. The white bars come from the same subject at different time points; the colored bars come from genetically related but different subjects.

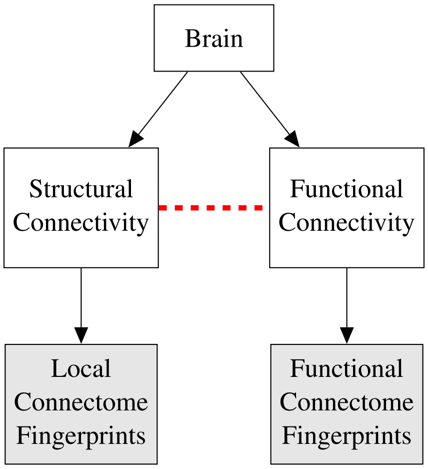
Brain
Structural Connectivity
Functional Connectivity
Local Connectome Fingerprints
Functional Connectome Fingerprints

Figure 3: A graphical description of the problem setup. We obtain two different fingerprint measurements of the brain, and we attempt to establish the link (red dashed line) between structural and functional connectivity using these measurements.

1.2 Objectives

Given a pair of high-dimensional feature vectors representing the structural and functional connectivity patterns of the brain, we are interested in the various forms of correlations between these two sets of features.

The first question to ask is whether such correlation exists at all. Given these high-dimensional feature vectors, it is nontrivial to properly define what it means for the vectors to be correlated and to correctly estimate any form of test statistic. One way to deal with such issue is to use a distance-based approach, where we aggregate over all dimensions by taking the between-subject distance matrix for each set of features. This gives structural and functional distance matrices of size n by n, where n is the number of subjects that is not as big (e.g. 50).

The scientific intuition for this approach is that it is reasonable to believe that subjects with similar structural connectivity patterns also display similar functional connectivity patterns. In particular, we expect that the use of localized structural connectivity features can help us obtain a more accurate measure of the overall correlation. This leads to the first hypothesis that we want to test (in the form of an alternative hypothesis):

Hypothesis 1.1 *Similarity in structural connectivity patterns of human brains is associated with similarity in their functional connectivity patterns.*

While testing Hypothesis 1.1 will provide some insights to the relationship, the measure of correlation aggregated over so many features may not be as useful or intuitive. Thus, the next step would be to find specific subsets of each set of features that are correlated. Such findings can be informative because, once we are able to identify specific subsets of connectivity features, we can validate our results using known scientific results about regional connections across the brain. This leads to our second hypothesis, which we will attempt to validate through scientific knowledge:

Hypothesis 1.2 *Structural connectivity patterns in local regions of the human brain are associated with functional connectivity patterns in other regions of the brain.*

1.3 Notations and assumptions

For the rest of our paper, $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ are the data matrices containing as rows local connectome fingerprints and and functional connectivity features, respectively. n is the sample size, p is the length of a local connectome fingerprint, and q is the length of the functional connectivity features (e.g. the upper-triangle of the fMRI connectivity matrix).

$X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}^q$ represent the ith row of $\mathbf{X}$ and $\mathbf{Y}$ for $i = 1, \ldots, n$. We assume that (X_i, Y_i) are each an i.i.d. sample from an unknown distribution P_{XY} defined on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^p \times \mathbb{R}^q$.

2 Data and Methods

2.1 Datasets

We work with two datasets, each of which contains DSI and fMRI measurements from a pool of neurologically healthy subjects. Table 1 summarizes the two datasets.

2.1.1 CMU-60

The first dataset is the CMU-60 data, which contains measurements from subjects who each responded to 120 trials of the Stroop task [13], in which he or she was asked to ignore the meaning of a printed word and respond with the ink color of the word (red, green, or blue). Subjects performed the task in a magnetic resonance (MR) scanner, and then they each went through a 50-minute, 257-direction DSI scan.

After pre-processing, we have 54 subjects that have a valid local connectome fingerprint (LCF) and a valid resting-state fMRI scan. For each of these subjects, we have a 49,911-dimensional LCF vector extracted from 2mm-resolution DSI scans. For the same set of subjects, we also have a time series of length 210 (sampled at 1Hz) for resting-state fMRI scans at 625 different regions of interest (ROIs), from which we compute the Pearson correlation matrix, i.e. a functional connectome fingerprint (FCF), between time series at every pair of ROIs.

2.1.2 HCP-900

The second dataset comes from the Human Connectome Project [14], which was a large-scale project with the goal of mapping the human connectome "as accurately as possible." While the project itself includes various measurements of brain connectivity and related outcomes, we only take the 1mm-resolution DSI scans and resting-state fMRI scans.

After the same pre-processing procedure as the CMU-60 data, we obtain the local connectome fingerprints and the fMRI scans from 257[1] subjects. Because of the finer resolution in DSI scans, we obtain 433,386-dimensional feature vectors as LCFs. The resting-state fMRI scans are measured on 626 ROIs, and they are measured for a longer period of 840 seconds (again, sampled at 1Hz).

Table 1: A summary of the two datasets.

Dataset		CMU-60	HCP-900
Sample size		**50**	**257**
LCF	Scan	Diffusion spectrum imaging (DSI)	
	Resolution	2mm	1mm
	Dimension	**49,991**	**433,386**
	Uniqueness	100% accuracy in identification tasks[2]	
FCF	Scan	Resting-state functional MRI (fMRI)	
	Duration	210 seconds (1Hz)	840 seconds (1Hz)
	ROI	625	626
	Dimension	**195,000**	**195,625**
	Uniqueness	92+% accuracy in identification tasks[3]	

2.2 Methods

2.2.1 Distance-based analysis of correlations across all regions

In order to verify Hypothesis 1.1, we estimate the correlation between the distance in structural features (i.e. LCFs) and the distance in functional features. Identification task results from [11] and [12] suggest that the natural choice of distance metric in each set of features would be the *(scaled) Euclidean distance* for LCFs and *correlation distance*, which is defined as 1 minus the Pearson correlation, for FCFs. This gives two n by n distance matrices, which we denote as $D^X = [d_X(X_i, X_j)]_{i,j=1}^n$ and $D^Y = [d_Y(Y_i, Y_j)]_{i,j=1}^n$, where in this case d_X is the scaled Euclidean distance and d_Y is the correlation distance:

$$\underset{\text{(Structural)}}{\begin{bmatrix} & & \\ & d_X(X_i, X_j) & \\ & & \end{bmatrix}} \quad \underset{\text{(Functional)}}{\begin{bmatrix} & & \\ & d_Y(Y_i, Y_j) & \\ & & \end{bmatrix}}$$

In general, it is a nontrivial problem to set up a proper mathematical hypothesis corresponding to Hypothesis 1.1, because the entries of each distance matrix are not i.i.d. – in fact, it has $\binom{n}{2}$ entries with only n degrees of freedom. Thus, standard approaches to test or give confidence intervals to the Pearson or Spearman correlation will not work. We can construct the null and alternative hypotheses from the statement of our hypothesis as follows: given independent copies $(X, Y), (X', Y') \sim P_{XY}$, where P_{XY} is the true joint distribution of (X, Y), we test

$$H_0 : \mathcal{R}\left(d_X(X, X'), d_Y(Y, Y')\right) = 0 \quad \text{vs.} \quad H_1 : \mathcal{R}\left(d_X(X, X'), d_Y(Y, Y')\right) > 0 \tag{1}$$

where $\mathcal{R}$ is some form of correlation between the two random variables representing distances between independent pairs of copies in $\mathbb{R}^p$, the space where X, X' resides, and in $\mathbb{R}^q$, the space where Y, Y' resides. We also choose a one-sided alternative when appropriate (e.g. in permutation test with $\mathcal{R}$ being the linear correlation) since we do not expect this correlation to be negative. If two copies are measured on the same subject, we know that $d_X(X, X') \approx 0$ and also $d_Y(Y, Y')$ is very small. As a result, similarity in structural features should monotonically be related to similarity in functional features, if there is any such relationship.

[1] There are 842 LCFs readily available, but only 257 of the corresponding fMRI scans are available as of now. The rest of fMRI scans are still undergoing pre-processing steps.

When $\mathcal{R}$ is the Pearson correlation and both d_X and d_Y are Euclidean distances (or their powers of $\gamma \in [0, 2]$, we recover **distance correlation (dCor)** [15, 16]. There are two appealing properties of this version of correlation. First, the test of whether dCor is zero or not is equivalent to the test of independence between X and Y, i.e. $\mathcal{R}\left(d_X(X, X'), d_Y(Y, Y')\right) = 0$ if and only if X is independent of Y. Second, dCor has a known limiting distribution which allows for asymptotically exact testing, although it is also known that the test loses power in high dimensions (i.e. harder to reject) [17]. Our first result will come from applying the t-test of independence using dCor, as described in [16].

The main issue with using a dCor test is that the choice of distances d_X and d_Y are limited to certain powers of the Euclidean distance. In our case, we prefer d_Y to be a correlation-based distance, since the data Y are themselves Pearson correlations based on time series. Previous identification tasks from [11] and [12] also suggest that using the Euclidean distance for local connectome fingerprints and the correlation distance for fMRI connectivity matrices properly retains the individual genetic representations from each set of features. Therefore, we extend this test to the cases where d_X is the (scaled) Euclidean distance and d_Y is the correlation distance (1 minus the Pearson correlation). We also consider $\mathcal{R}$ to be either the Pearson correlation or the Spearman correlation, depending on results of our exploratory analysis.

The main problem with these extensions is that, indeed, we lose the two theoretical properties of the Euclidean distance correlation. Therefore, rejecting such tests no longer implies that the two sets of features are independent (although this is a stronger statement than what we are trying to answer in Hypothesis 1.1), so that we need to be careful about our interpretation of the results. Also, we do not have a closed-form limiting distribution for exact testing, so we need to resort to other methods, in particular permutation-based tests and subsampling-based confidence intervals.

In summary, the three methods we consider for testing are:

1. Distance correlation (dCor, [15, 16]) t-test of independence,
2. Permutation-based test of correlation, and
3. Subsampling-based confidence interval construction for correlation.

For dCor, we can follow the results from [16], which shows that given the de-biased estimate of the statistic $\mathcal{R}_n^*$ the statistic

$$\sqrt{\nu - 1} \cdot \frac{\mathcal{R}_n^*}{\sqrt{1 - \left(\mathcal{R}_n^*\right)^2}}$$

follows a Student's t-distribution with $(\nu - 1)$ degrees of freedom, where $\nu = \frac{n(n-3)}{2}$, as $p, q \to \infty$.

For the **permutation test**, in which we take d_Y to be the correlation distance, we extend the idea of a permutation test for the Pearson correlation in univariate regression, in which case samples from each feature are randomly permuted to compute a random correlation under the null hypothesis. In our high-dimensional case, we also randomly permute the sample in each feature of X and of Y to obtain a random correlation under H_0. We repeat the permutation multiple times to construct an approximate null distribution of $\mathcal{R}$, and compute the quantile of the true correlation for our dataset to obtain a p-value. While this method does not perfectly simulate random data from the null distribution, it is a natural proxy for the null to see how unlikely it is to observe the correlation we obtain on the actual dataset.

Subsampling is another method that can be applied to our distance-based analysis. The idea is to construct a confidence interval using several subsamples (without replacement) of the original sample, which in our case is the set of subjects. We prefer subsampling over bootstrapping because the distance matrix for each bootstrap sample (with replacement) will have zeros between every pair of duplicate subjects. We also know that subsampling is valid because we can in fact show that the correlation between distance matrices has a limiting distribution. To see this, we first note that the numerator of our statistic is is a U-statistic. By letting $Z_i = (X_i, Y_i)$ for each subject i, we see that

$$r(D^X, D^Y) = \frac{1}{\binom{n}{2}} \sum_{i<j} d_X(X_i, X_j) d_Y(Y_i, Y_j) = \frac{1}{\binom{n}{2}} \sum_{i<j} u(Z_i, Z_j)$$

where $u(Z_i, Z_j) = d_X(X_i, X_j)d_Y(Y_i, Y_j)$ for each pair $i < j$. This implies that the standard central limit theorem for U-statistics holds and we can estimate confidence intervals using subsampling. Since the denominator is the product of $r(D^X, D^X)$ and $r(D^Y, D^Y)$, we can conclude that the limiting distribution exists by the continuous mapping theorem and the multivariate delta method.

There are two reasons we prefer subsampling instead of the more standard procedure of bootstrapping. First, we know that subsampling holds under much weaker assumptions (that the statistic has a limiting distribution). Second, we expect that bootstrapping the subjects will actually fail in our case, because each bootstrap sample contains multiple copies of the same feature vector, resulting in a distance matrix containing many zeros. We will empirically demonstrate this phenomenon in our results section.

While there are no known standard method of choosing the size of each subsample or the subsampling ratio, there are methods that are originally developed for choosing m in the m-out-of-n bootstrap procedure (sampling m examples with replacement). In our case, we follow Bickel and Sakov's procedure in [18], which chooses the size that minimizes the sup-norm difference in the corresponding bootstrap distribution and that of the next candidate size.

For all of the above methods, we do note that the functional connectivity features are not as reliable as local connectome fingerprints. This is mainly because the functional features are taken to be a flattened version of a correlation matrix, resulting in undesired correlation structures between features. Thus, we will also consider various alternative functional network measures from [6] as our features, and use a corresponding distance metric to compute the analogous distance matrix. We will show our results for the pair of structural and functional fingerprints as well as for the structural LCFs and one of the functional network features.

In all of our tests, we choose the significance level to be $\alpha = .05$.

2.2.2 Finding local and group-wise correlations in high-dimensions

Once we have an evidence that there is some correlation between the structural and functional connectivity features, we proceed with finding correlations between subsets of the features that represent specific regions of the brain. This task will address the validity of Hypothesis 1.2.

While it is highly nontrivial to make inference involving a subset of high-dimensional features that are chosen using a selection technique, we can make predictive claims that involve clever selection of variables, especially given such a high-dimensional setting. In our particular case, we are mainly interested in the structure of the cross-covariance (or similarly the cross-correlation) matrix between the two sets of features, denoted as Σ_{XY}.

The first attempt is to apply **canonical correlation analysis (CCA)**, which finds a pair of linear transformations that map the two sets of features into the same Euclidean space, such that the projections are the most correlated.

Specifically, given two sets of features $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, CCA finds the "alignment" vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ that maximizes the canonical correlation (CC):

$$\mathrm{CC}(X, Y) = \mathrm{cor}(Xu, Yv) = \frac{1}{n} u^T X^T Y v = u^T \Sigma_{XY} v$$

such that $\|Xu\|_2^2 \leq 1$ and $\|Yv\|_2^2 \leq 1$.

In the low-dimensional case, CCA is mathematically equivalent to a principal component analysis (PCA) on the cross-covariance matrix Σ_{XY}, as the CC objective suggests. This also suggests that, in high-dimensional settings, just like with PCA, the estimate of the covariance by doing a singular value decomposition will be inconsistent, even though the algorithms for computing them may be mathematically equivalent.

In the high-dimensional case, we attempt to perform CCA in the following ways:

1. *Sparse CCA* [19, 20, 21]: apply an ℓ^1 penalty to the alignment vectors. This has a convenient effect of selecting subsets of variables from each set of features that give a high cross-correlation between them.
2. *Dimensionality reduction followed by CCA*: first apply a dimensionality reduction technique such as low-rank PCA and then run a low-dimensional CCA.

We will show our results in both cases.

3 Results

3.1 Correlation of distances

3.1.1 Exploratory analysis

We first present exploratory analysis results for the distances. Figures 4 and 5 in part reproduce the results from [11] and [12], in which they show that the distances between different individuals are substantially greater than that between the same subjects. This justifies our choice of distances for the permutation-based test and the subsampling-based confidence interval.

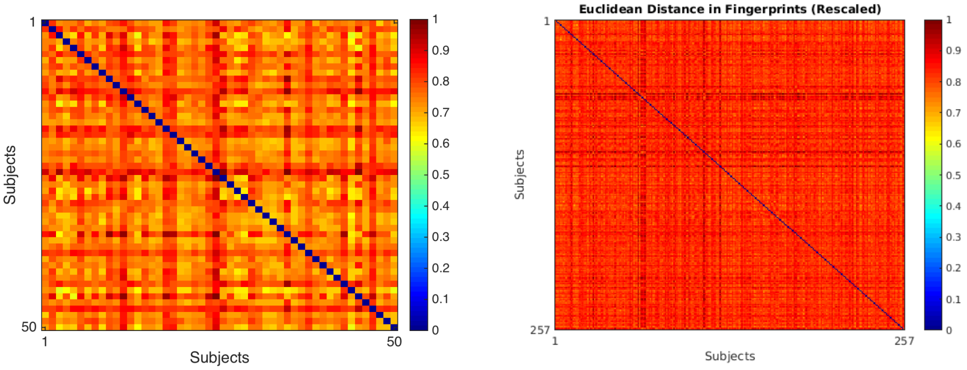

Figure 4: Scaled Euclidean distances between pairs of subjects' local connectome fingerprints (LCFs) for all pairs of subjects. *Left:* CMU-60 ($n = 50$). *Right:* HCP-900 ($n = 257$).

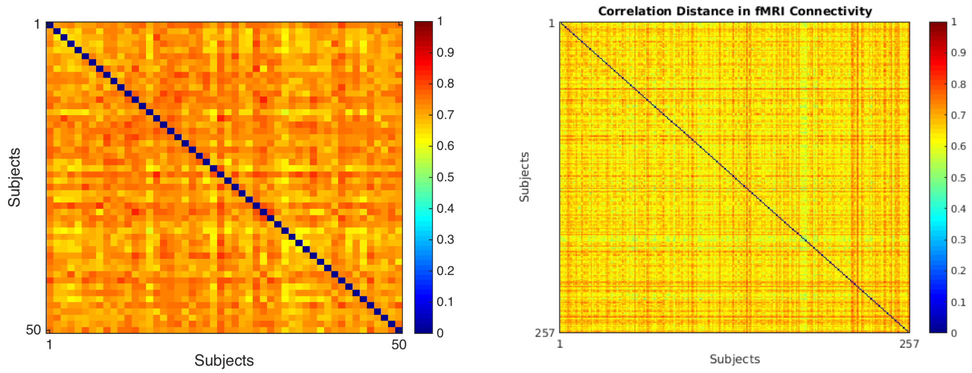

Figure 5: Correlation distances between pairs of subjects' functional connectome fingerprints (FCFs, i.e. fMRI correlation matrices) for all pairs of subjects. *Left:* CMU-60 ($n = 50$). *Right:* HCP-900 ($n = 257$).

Next, we plot the structural and functional pairwise distances in a scatterplot to explore the overall trend. Figure 6 suggests that there is a positive trend between the pairwise distances in the structural and functional features.

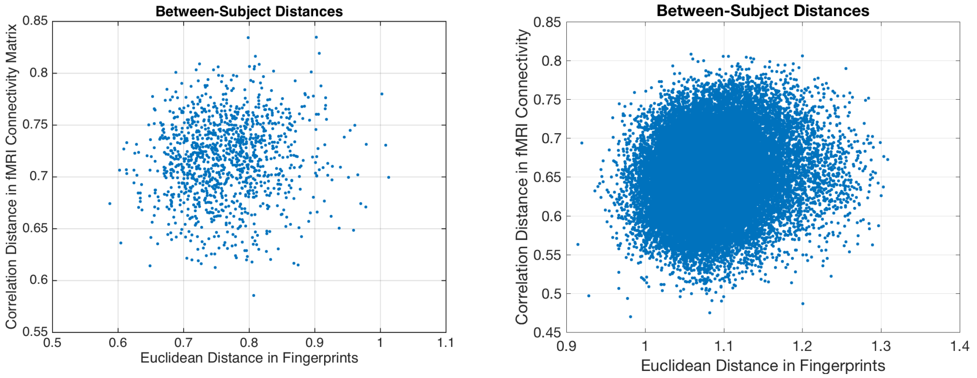

Figure 6: Scatterplots of all pairwise distances in local connectome fingerprints (x-axis) and in functional connectome fingerprints (y-axis). *Left:* CMU-60 ($n = 50$). *Right:* HCP-900 ($n = 257$).

3.1.2 Test results and confidence intervals

Motivated by these results, we now present our results on the dCor t-test, permutation test, and subsampling confidence interval. Significance levels are marked with † ($p < .1$), * ($p < .05$), ** ($p < .01$), and *** ($p < .001$). Significant confidence intervals are marked with $^+$.

Table 2: CMU-60, $n = 50$

Method	Correlation	Result Type	Result
Permutation test	0.088	(one-sided) p-value	0.309
Subsampling	0.088	95% confidence interval	(0.013, 0.192)$^+$
dCor t-test	0.038	one-sided p-value	0.480

Table 3: HCP-900, $n = 257$

Method	Correlation	Result Type	Result
Permutation test	0.152	one-sided p-value	$< 0.001^{***}$
Subsampling	0.152	95% confidence interval	(0.112, 0.192)$^+$
dCor t-test	0.239	one-sided p-value	$< 0.001^{***}$

Figure 7 visualizes the result from our permutation tests. The justification for taking a one-sided p-value is that it is only reasonable to expect that similar structural connectivity features are correlated to similar functional features rather than disssimilar functional features, since copies of features from the same subjects give a distance of zero in both domains.

For the HCP-900 dataset, we find from the permutation test and subsampling results that there is indeed a significant correlation between the Euclidean distances in local connectome fingerprints and the correlation distances in functional connectome fingerprints. The dCor t-test of independence confirms that the two sets of fingerprints are statistically dependent, despite the fact that the test makes strong assumptions. The analogous results on the CMU-60 dataset are weaker, but we expect this to happen due to the small sample size.

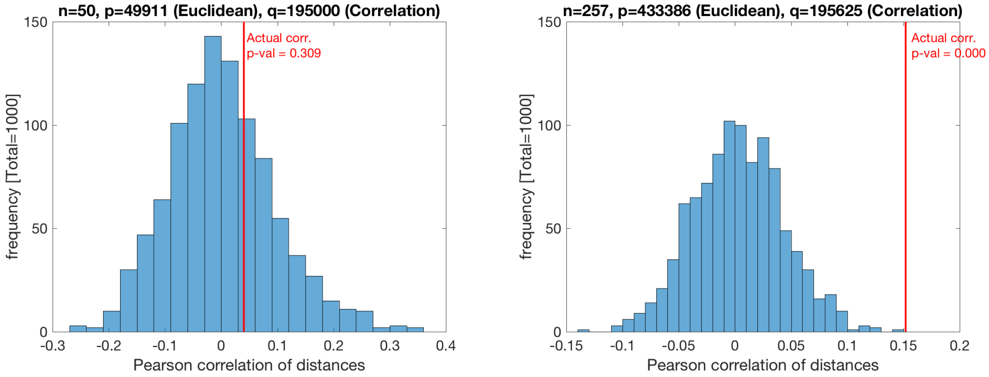

Figure 7: Simulated null distribution using 1,000 random permutations. Red vertical line indicates the correlation on the actual dataset and the p-values are the proportion of random correlations on the right-side tail of the red vertical line. *Left:* CMU-60 dataset, $n = 50$. *Right:* HCP-900 dataset, $n = 257$.

Figure 8 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 matrix always contain many zeros, leading to an unusually high correlation than 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.

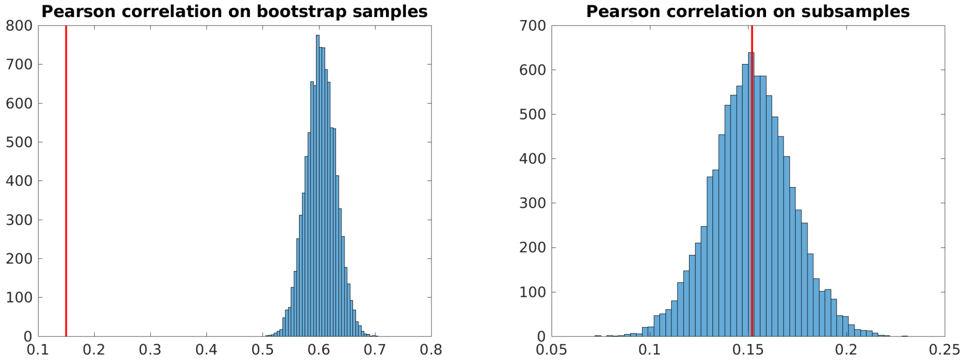

Figure 8: Histogram of linear correlations for 10,000 bootstrap samples (left) and 10,000 subsamples (right) of the HCP-900 ($n = 257$) dataset. Red vertical line indicates the true correlation (0.152). Red vertical line indicates the actual correlation computed on the full data.

3.1.3 Results on functional network features

In Tables 4 and 5, we further give a distance-based correlation analysis between various functional connectivity distances and the Euclidean distance between structural connectome fingerprints. The list of functional features is taken from the seminal paper [6]. We also included core-periphery networks [22] and its maximized coreness objective. Note that this method as well as community-based methods, both of which measure the modularity of functional networks, give significant correlations with the structural distances. For measures on binary networks, these correlations were thresholded at the value 0.5. The default choice of distance between two subjects' functional connectivity measures is the Euclidean distance (.), unless noted otherwise.

While these results are preliminary and only limited to the CMU-60 dataset at this point, we suspect that certain sets of features such as mean node degree and community structures can help us improve upon our correlation results further.

Table 4: Permutation-based tests and subsampling-based confidence intervals for **linear** correlation using functional network features for the CMU-60 dataset ($n = 50$). Each functional feature was tested with the Euclidean distance in structural connectome fingerprints, which have length 49,911.

Functional Features (Length)	Distance	Linear Corr.	p-value	95% Conf. Int.
fMRI Connectivity Matrix (195K)	**Correlation**	**0.088**	**0.309**	**(0.013, 0.192)**$^{+}$
Mean Degree (625)	.	0.010	0.931	(-0.030, 0.064)
Mean Strength (625)	.	-0.016	0.879	(-0.058, 0.038)
Clustering Coefficients (625)	.	-0.019	0.863	(-0.057, 0.032)
Mean Path Length (1)	.	0.021	0.809	(-0.023, 0.078)
Small-Worldness (1)	.	-0.037	0.688	(-0.081, 0.014)
Connection Density (1)	.	0.039	0.659	(-0.005, 0.097)
Betweenness Centrality (625)	.	-0.106	0.338	(-0.167, -0.048)
PageRank Centrality (625)	.	-0.040	0.691	(-0.098, 0.027)
Core-Periphery Partition (625)	Hamming	0.033	0.587	(-0.012, 0.076)
Maximum Coreness (1)	.	0.168	0.040*	(0.086, 0.232)$^{+}$
Community Detection (625)	Hamming	0.171	0.002**	(0.093, 0.219)$^{+}$
Community-based Modularity (1)	.	0.049	0.504	(-0.008, 0.099)

Table 5: Permutation-based tests and subsampling-based confidence intervals for **rank** correlation using functional network features for the CMU-60 dataset ($n = 50$). Each functional feature was tested with the Euclidean distance in structural connectome fingerprints, which have length 49,911.

Functional Features (Length)	Distance	Rank Corr.	p-value	95% Conf. Int.
fMRI Connectivity Matrix (195K)	**Correlation**	**0.094**	**0.134**	**(0.021, 0.151)**$^{+}$
Mean Degree (625)	.	0.092	0.362	(0.029, 0.162)$^{+}$
Mean Strength (625)	.	0.052	0.598	(-0.011, 0.121)
Clustering Coefficients (625)	.	0.054	0.583	(-0.007, 0.120)
Mean Path Length (1)	.	0.058	0.452	(0.010, 0.111)$^{+}$
Small-Worldness (1)	.	0.012	0.893	(-0.040, 0.067)
Connection Density (1)	.	0.086	0.268	(0.038, 0.141)$^{+}$
Betweenness Centrality (625)	.	-0.130	0.241	(-0.194, -0.069)
PageRank Centrality (625)	.	-0.008	0.933	(-0.076, 0.061)
Core-Periphery Partition (625)	Hamming	0.033	0.579	(-0.012, 0.079)
Maximum Coreness (1)	.	0.142	0.059†	(0.066, 0.205)$^{+}$
Community Partition (625)	Hamming	0.167	0.002**	(0.098, 0.212)$^{+}$
Community-based Modularity (1)	.	0.037	0.576	(-0.013, 0.081)

3.2 CCA-based results

Sparse CCA is run by code from [21], which adds an elastic net ($\ell^1 + \ell^2$) penalty by default. In the following results, the ℓ^1-regularization parameter is chosen differently or is cross-validated, while the ℓ^2-regularization parameter is fixed to 1 for simplicity. Non-sparse CCA that follows PCA does not apply any penalty to the original CCA problem.

3.2.1 Exploratory analysis

Given pairs of high-dimensional features that we will project into lower-dimensional subspaces, we visualize the decaying patterns of the singular values. Overall, while the first few principal components appear to contain spikes, it is not as obvious in general whether PCA will be effective given the flat decay.

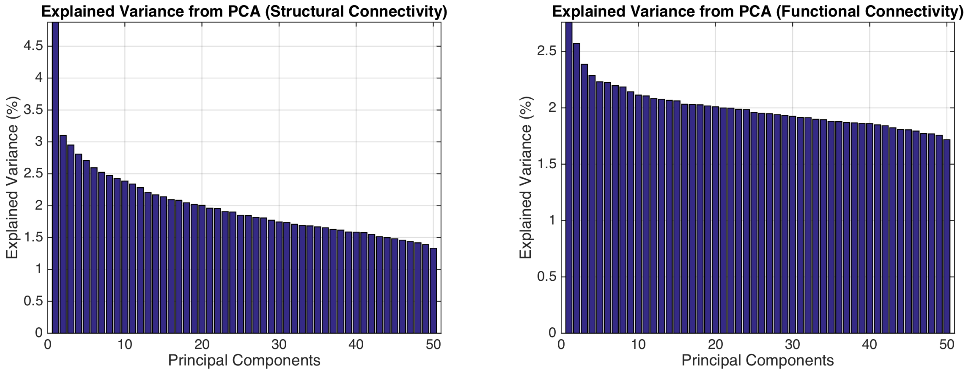

Figure 9: Explained variance (% of singular values) from PCA on structural (left) and functional (right) connectivity patterns for the CMU-60 dataset.

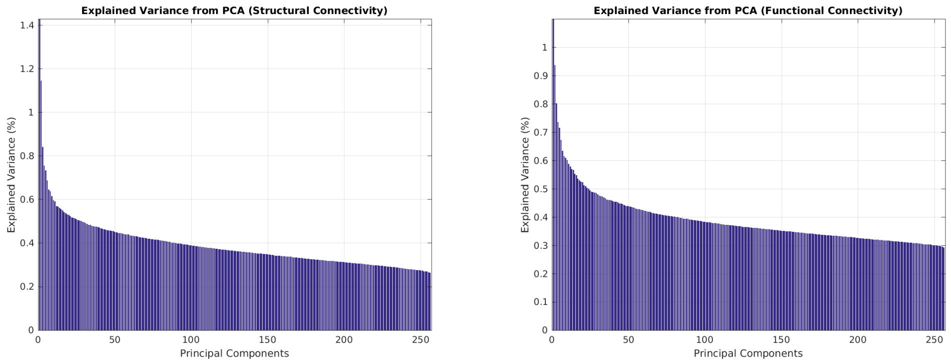

Figure 10: Explained variance (% of singular values) from PCA on structural (left) and functional (right) connectivity patterns for the HCP-900 dataset.

3.2.2 Cross-validation

We present results for both sparse CCA and for PCA then non-sparse CCA using a 5-fold cross-validation. For sparse CCA, the ℓ^1-regularization parameter is cross-validated. For PCA+CCA, the number of principal components is cross-validated.

The corresponding results have high variance, due to the limited sample size in the CMU-60 dataset. While the consistency of results on the HCP-900 dataset appear to be slightly better, we still observe a lot of variability. This variability is due to both (1) the variance coming from the optimization procedure, where the problem is ill-conditioned and thus can yield many different solutions, and (2) the variance coming from the lack of statistical consistency in the high-dimensional setting.

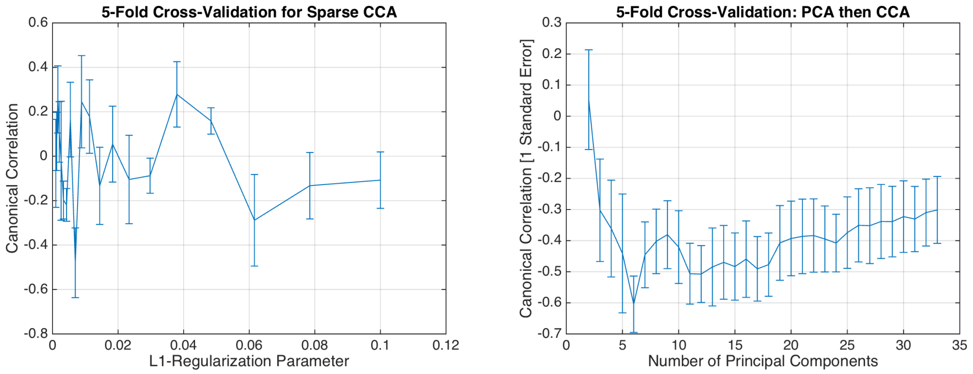

Figure 11: 5-fold cross-validation for the canonical correlation in sparse CCA (left) and non-sparse CCA after PCA (right) on the CMU-60 ($n = 50$) dataset. Error bars represent 1 standard error. Note that, due to the small sample size, the shape of these plots tend to change quite a bit depending on how the 5 folds are split. For PCA then CCA, The negative canonical correlation is not an interesting result, because the objective function (canonical correlation) is supposed to be maximized, not minimized.

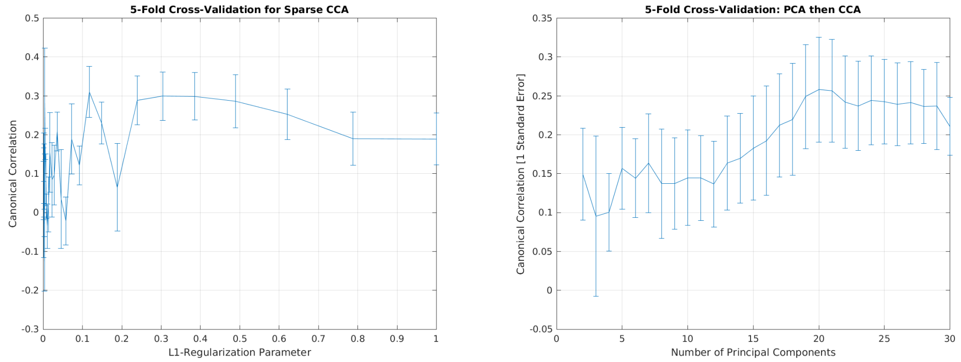

Figure 12: 5-fold cross-validation for the canonical correlation in sparse CCA (left) and non-sparse CCA after PCA (right) on the HCP-900 ($n = 257$) dataset. Error bars represent 1 standard error. The shape of these plots can still change a bit over different runs, but the results appear more reasonable as we have more subjects than CMU-60.

3.2.3 CCA projections with cross-validated parameters

Using these cross-validated parameters, we now compute the corresponding canonical projections and plot them in the Euclidean space. Since the objective of CCA is to maximize the correlation between these projected points from the training set, we expect to see a linearly increasing pattern in the training data. The generalization performance we want to see is that the same projections applied to the testing data also follows the similar pattern.

In Figure 13, for the CMU-60 dataset, it is not obvious how to interpret these patterns given such a small amount of data in each of the 5 folds and in the test set. Indeed, there is a lot of variability in the resulting patterns over multiple runs, as we are splitting the small dataset into 5 folds and a test set.

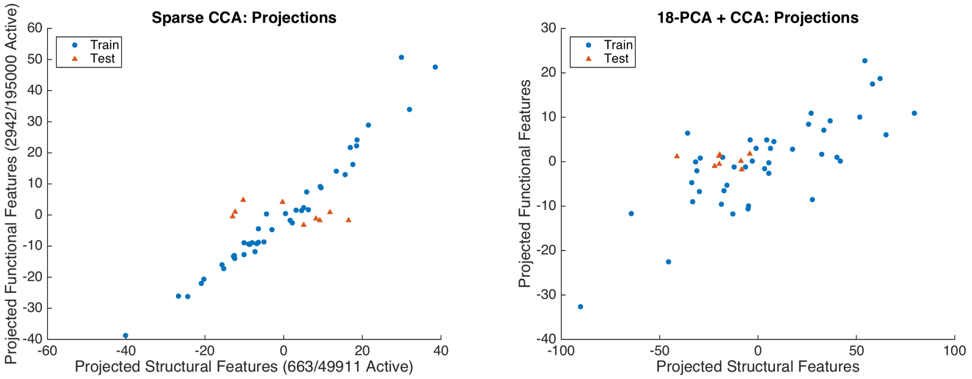

Figure 13: Sparse CCA (left) and PCA-then-CCA (right) projections using 5-fold cross-validated parameters on the CMU-60 dataset. Given the small number of active features and the small sample size, we suspect that the variance of our results can be very high.

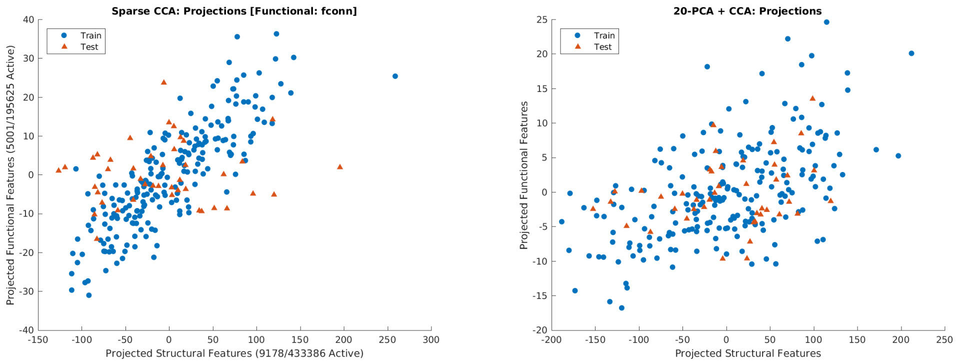

Figure 14: Sparse CCA (left) and PCA-then-CCA (right) projections using 5-fold cross-validated parameters on the CMU-60 dataset. Given the small number of active features and the small sample size, we suspect that the variance of our results can be very high.

In Figure 14, for the HCP-900 dataset, we see that the PCA-then-CCA approach seems to obtain a pair of projections that can generalize to the test set, although the overall pattern is more spread out than sparse CCA. In sparse CCA, we see that the projections do not seem to generalize well to the test set, and this is due to a potential overfitting by choosing too specific subsets of each set of features, even if the sparsity parameter is chosen by cross-validation. In fact, the set of selected features vary a lot across multiple runs, suggesting that we do not see sparsistency in our results for these datasets. A future work is to understand which features get selected over multiple runs and see if different features from a set of correlated features gets selected across different runs. This would imply that we need to take into account the underlying clustering patterns of each set of features.

4 Conclusions

In this report, we first show that there is a significant correlation between the distances in local connectome fingerprints and in functional connectome fingerprints (fMRI correlation matrix). This reaffirms previous results in the literature that demonstrates the structure-function relationship in the human connectome. At the same time, our results also show that the novel representations of local connectome features can be useful in explaining such relationship. We note that the results were not highly significant with some of the tests we performed on the CMU-60 dataset, and we suspect that this is most likely due to the small sample size ($n = 50$) of the dataset. On the HCP-900 dataset ($n = 257$), we obtain statistical significance on all of our tests, including the distance correlation (dCor) t-test of independence. On the other hand, our attempts using some of the functional network features gave mixed results, indicating further investigations on interesting network measures need to be done.

Secondly, we identified certain sub-connections in the local connectome and subgraphs of the functional networks that can be highly correlated, using sparse canonical correlation analysis and dimensionality reduction techniques. While our results on the CMU-60 dataset had very high variances for us to get to any meaningful conclusion, our results on the HCP-900 dataset suggest that approaches such as sparse CCA and PCA-then-CCA can be effective in identifying these sub-connections. In particular we obtained projections that can generalize to unseen data, in the sense that the projections align structural and functional connectivity features that highlights the correlation between the two. A future work is to account for correlated structures within each modality to give consistent results from sparse CCA and to validate our results with existing scientific knowledge by looking at which features get selected consistently.

References

[1] Van J Wedeen, Patric Hagmann, Wen-Yih Isaac Tseng, Timothy G Reese, and Robert M Weisskoff. Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic Resonance in Medicine*, 54(6):1377–1386, 2005.

[2] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[3] P Hagmann. *From diffusion MRI to brain connectomics. 2005*. PhD thesis, PhD Thesis (Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland).

[4] Christopher J Honey, Jean-Philippe Thivierge, and Olaf Sporns. Can structure predict function in the human brain? *Neuroimage*, 52(3):766–776, 2010.

[5] Michael D Greicius, Ben Krasnow, Allan L Reiss, and Vinod Menon. 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, 2003.

[6] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[7] Michael D Greicius, Kaustubh Supekar, Vinod Menon, and Robert F Dougherty. Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cerebral cortex*, 19(1):72–78, 2009.

[8] CJ Honey, O Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

[9] Timothy D Verstynen. The organization and dynamics of corticostriatal pathways link the medial orbitofrontal cortex to future behavioral responses. *Journal of neurophysiology*, 112(10):2457–2469, 2014.

[10] Sashank J. Reddi. Understanding the relationship between functional and structural connectivity of brain networks. *Data Analysis Project Report, Machine Learning Department, Carnegie Mellon University*, 2015.

[11] Fang-Cheng Yeh, Jean Vettel, Aarti Singh, Barnabas Poczos, Scott Grafton, Kirk Erickson, Wen-Yih Isaac Tseng, and Timothy Verstynen. Quantifying differences and similarities in whole-brain white matter architecture using local connectome fingerprints. *PLoS Computational Biology*, 12(11):e1005203, 2016.

[12] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 2015.

[13] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[14] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, WU-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[15] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[16] Gábor J Székely and Maria L Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[17] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

[18] Peter J Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985, 2008.

[19] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[20] Sivaraman Balakrishnan, Kriti Puniyani, and John Lafferty. Sparse additive functional and kernel cca. *arXiv preprint arXiv:1206.4669*, 2012.

[21] Xi Chen, Han Liu, and Jaime G Carbonell. Structured sparse canonical correlation analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[22] M Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. Core-periphery structure in networks. *SIAM Journal on Applied mathematics*, 74(1):167–190, 2014.