Discriminative Structured Feature Engineering for Macroscale Brain Connectomes

Jian Pu, Jun Wang, *Member, IEEE*, Wenwen Yu, Zhuangming Shen, Qian Lv, Kristina Zeljic, Chencheng Zhang, Bomin Sun, Guoxiang Liu, and Zheng Wang*

***Abstract*—Neuroimaging techniques can measure structural and functional brain connectivity with unprecedented detail *in vivo*. This so-called brain connectome can be represented as high dimensional matrices corresponding to edge weights in graphs. After measuring the matrices of two cohorts (i.e., patients and healthy controls), one is often required to formulate computational network models for effective feature engineering to draw discriminative distinctions between the cohorts, as well as estimate the associated statistical significance. We designed a novel method to reveal the intrinsic features of functional matrices of discriminative power for group comparison. More specifically, by encouraging co-selection of edges connected to the same node, we preserved the discriminative edges to maximum extent. To reduce the false positive rate of the extracted discriminative edges, an optimization procedure was developed to evaluate the significance of these edges and remove trivial ones. We validated the proposed method using both synthetic data and real benchmarks, and compared it to ℓ_1 regularized logistic regression, univariate t-test and stability selection. The experimental results clearly showed that the proposed approach outperformed the three competing methods under various settings. In addition to increasing the F-measure of feature selection, our approach captured the endogenous, discriminative connectivity patterns consistent with recent findings in biomedical literature. This data-driven method paves a new avenue of enquiry into the inherent nature of network models for functional brain connectomes.**

***Index Terms*—Classification, feature engineering, macroscale brain connectomes, statistical significance.**

I. Introduction

MACROSCALE brain connectomes [1] refer to the connectivities of distinct brain areas, including both the structural pathways and functional interactions [2]. Compared to microscale brain connectomes that enquire into individual neurons and their synaptic connections, macroscale brain connectomes can measure *in vivo* structural and functional connectivities in animal and human subjects with noninvasive tools. Specifically, functional magnetic resonance imaging (fMRI) can provide superior spatial and temporal resolution compared to computed tomography (CT) and positron emission tomography (PET) by measuring blood oxygenation level-dependent signals to represent functional brain activity. In addition, diffusion magnetic resonance imaging (dMRI) can map the water diffusion process in biological tissues to infer white matter architecture in a voxel-wise manner. Analysis and interpretation of brain connectomes can contribute to the scientific understanding of higher-order cognitive processes (e.g. future learning and cognitive states [3], [4]) and potentially aid in the diagnosis and treatment of neurological disorders [4], [5].

Defining nodes and edges is a prerequisite for constructing a connectivity matrix from functional brain connectome data. Node definition can be based on either anatomical analysis or functional behaviors, so as to satisfy the principles of intra-nodal homogeneity and inter-nodal heterogeneity. Intra-nodal homogeneity underlines the consistency within a single node, while inter-nodal heterogeneity emphasizes the discrepancy. As a consequence, node definition is usually a tradeoff between the homogeneity and heterogeneity of brain regions from a biological perspective. For example, the Automated Anatomical Labeling (AAL) template [6] defined 116 nodes for the human brain according to pure anatomical analysis. A fine-grained strategy [7] could further chart the human brain into 414, 813 and even 1615 regions. From the perspective of exploratory analyses of large data sets, as the number of defined brain nodes increases, efficient dimensionality reduction becomes critical for forming and evaluating hypotheses about macroscale activity structure. Once the brain parcellation scheme is set, the edges of a functional connectivity matrix can be generated by measuring the statistical correlation of any two nodes of the template over time.

By seeking differences in the connectivity matrices between two cohorts (i.e., patients versus healthy controls, or resting-state versus task-specified), researchers aim to reveal the working mechanisms of cognitive processes and identify potential biomarkers for disease diagnoses and therapeutics. Straightforward "bag of edges" analysis considers that all edges are independent and identically distributed [2]. Univariate hypothesis tests e.g., t-tests, F-tests) are therefore performed for each entity of the matrix, and edges with statistically significant values (i.e., $p < 0.05$) are considered critical distinctions. Nevertheless, this analysis is deficient in statistical efficiency

Manuscript received February 14, 2015; revised April 29, 2015 and May 04, 2015; accepted May 05, 2015. Date of publication May 08, 2015; date of current version October 28, 2015. *Asterisk indicates corresponding author.*

J. Pu, W. Yu, Z. Shen, Q. Lv, and K. Zeljic are with Institute of Neuroscience, Key Laboratory of Primate Neurobiology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: jianpu@ion.ac.cn).

J. Wang is with Institute of Data Science and Technology, Alibaba Group, Seattle, WA 98101 USA.

C. Zhang and B. Sun are with Department of Functional Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China.

G. Liu is with Center for Information and Neural Networks(CiNet), National Institute of Information and Communications Technology, Osaka University, Osaka 565-0871, Japan.

*Z. Wang is with Institute of Neuroscience, Key Laboratory of Primate Neurobiology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: zheng.wang@ion.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2015.2431294

owing to the curse of dimensionality (the number of matrix voxels can often reach the scale of 10^4).

One potential issue of "bag of edges" is the negligence of structured information hidden in the connectivity matrix, which can lead to a direct failure in exploring the modularity of brain networks. Consequently, attempts have been made to define network hubs (nodes well positioned to significantly influence to overall network function), particularly in terms of their role in the integration of network modules (closely linked groups of nodes) [8], [9]. Network modules have demonstrated effectiveness in studies of functional network mapping, indicating their potential usefulness in establishing a nuanced theoretical network for neural connectivity. Small-world property (modular structure) [10], [11] and hub distribution [8], [12]–[14] are the most common topological features of brain networks. As such, some graph-based methods treat the connectivity matrix as a global structure and postulate that the distinctions are constituted as one or several connected subgraphs. If subgraphs of the matrix could be well-defined beforehand, model-driven methods e.g., statistical parametric networks (SPNs) [15]) could evaluate the significance of these subgraphs instead of individual edges.

More recently, machine learning techniques have been increasingly employed to identify discriminative features between two cohorts in the biomedical research field [16], [17]. Support vector machines (SVMs) have successfully fulfilled various classification tasks thus far. As SVMs usually deal with vector inputs, the functional matrices are first reformed into feature vectors. The setting of data labels largely depends on diversified experimental purposes, such as test conditions (different tasks, or resting states) [3], clinical diagnoses (patient or healthy control) [18], and even human ages [19]. However, most machine learning approaches lack the capacity to estimate the statistical significance of identified discriminative features and fail to consider the distributed modular structures in brain connectivity matrices. When using the ℓ_2 regularized SVM, there is no specified constraint for feature selection. Therefore, the coefficients of the weight vector need to be cut-off using certain heuristic strategies (e.g. hard thresholding by preserving 70% energy of the weight vector [20]). A judicious option is to use the ℓ_1 regularized SVM instead, which automatically removes the trivial edges with small magnitudes and reserves more discriminative features with larger magnitudes. When the number of training samples is much less than the number of variables, survival count on random subsamples (SCoRS) [21] has been proposed to sub-sample both features and samples to obtain more consistent and reproducible features. To control the rate of false discoveries, a theoretical principle was proposed to select a suitable amount of regularization for specific structure estimation [22]. Leveraging some biology knowledge as prior information, model-driven machine learning methods, such as GraphNet [23], have been introduced to select grouped features simultaneously. Although GraphNet produces desirable structures, the definition of grouped features generally requires advanced knowledge of brain architecture, which is currently insufficient in the research field. Furthermore, a paucity of prior knowledge could mislead network model settings and result in artificial outcomes that hinder the exploration of genuine modular features.

Emerging evidence suggests that system-level disturbances should be observed in multi-level distributed brain networks, as manifested in various neurological and psychiatric diseases. Therefore, the neuropathological distinctions between patients and healthy controls are likely expressed through communicating components such as groups and hubs. For example, hub alternation has been revealed in schizophrenic [24], [25], comatose [26] and Alzheimer patients [27], and meta-analysis of more than 20,000 subjects and 26 different brain disorders concluded that the hub organization was highly susceptible to maladaption in the course of illness [28]. Thus, topological knowledge about the brain infrastructure can be harnessed as *a priori* information to help design elegant machine learning algorithms. Realizing the complexity of brain networks (diseases) and the urgent demand for statistical assessment of classification, we developed a novel feature engineering method that leverages a mixed norm regularization to automatically identify discriminative network components and perform statistical analysis simultaneously.

Instead of vectorizing the functional matrices, we distinguished the symmetric matrices directly by encouraging the weighted connectivity matrix to be both row and column sparse. Using a logistic empirical loss, the final objective was formulated as an unconstrained non-smooth convex problem, which can be efficiently solved using the alternative Accelerated Proximal Gradient (APG) method. Moreover, we evaluated the statistical significance of the identified features using an optimization procedure rather than a conventional permutation test. We validated our approach based on both synthetic and two real biomedical datasets. For the synthetic data, we examined the power of unraveling correct features when a limited number of features were selected. For real world study, a resting-state functional connectivity dataset was acquired from anesthetized macaque monkeys after injection with a single subanesthetic dose of ketamine. The purpose of this experiment was to identify the altered neural networks after ketamine administration relative to the saline control. From this nonhuman primate dataset, we found that significant features of ketamine action were largely distributed in the limbic system, thalamus, and cingular cortex, consistent with previous studies [29]. The other real world dataset consisted of functional connectivity matrices constructed from clinically diagnosed obsessive-compulsive disorder (OCD) patients and matched healthy controls. The significant findings here included abnormalities of the cortico-striato-thalamo-cortical (CSTC) circuits and the amygdala, which have been reported as major differences between these two cohorts in prior studies [30].

The paper is organized as follows. In Section II, our structural regularized framework is introduced as an optimization problem efficiently solved using the APG method. Furthermore, evaluation of the statistical significance of the data-derived features is described. Section III describes the datasets used for validation. The results are reported in Section IV and conclusions provided in Section V.

II. Methods

In this section, we describe the Discriminative Structured Feature Engineering (DSFE) method in detail. A flow chart of all procedures is illustrated in Fig. 1.

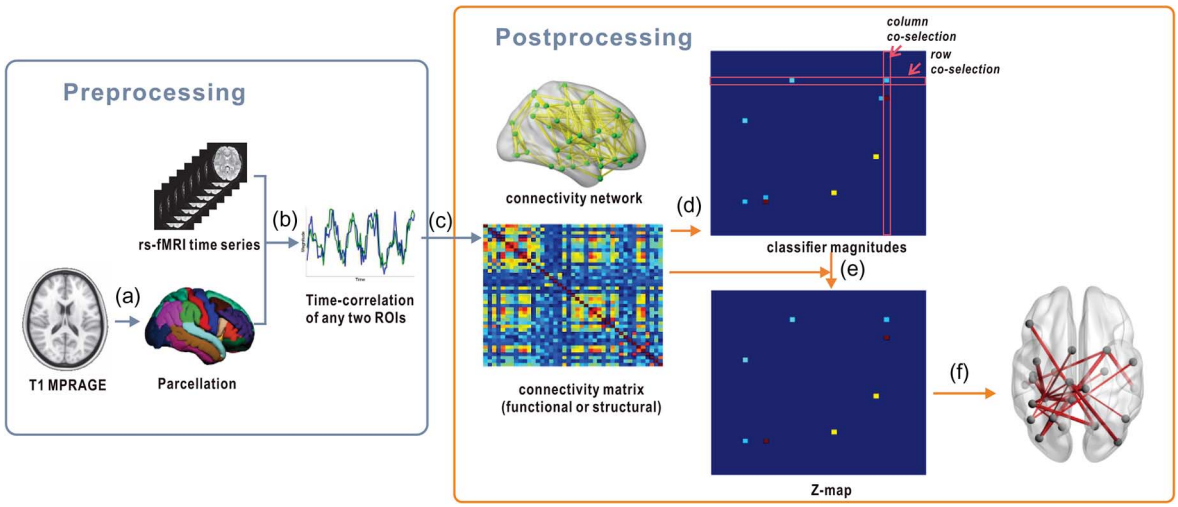

Fig. 1. Flow chart of the preprocessing of fMRI and dMRI data (left blue box) and our DSFE method (right orange box). (a) All individual T1 images (Magnetization Prepared Rapid Acquisition Gradient Echo, MPRAGE) were registered to a template, which defined the nodes of connectivity matrix; (b) Averaged resting state fMRI signal of each node was obtained; (c) Pearson's correlation coefficients between any pair of nodes were computed to form a symmetric functional connectivity matrix; (d,e) Magnitudes of the classifier were learned from the connectivity matrices of the two cohorts and subjected to the statistical analysis; (f) Results obtained by the DSFE method were converted into the z-score map and shown on a 3D brain.

Below we first introduce the notations and define the problem. We then formulate our objective function and derive the optimization procedure, followed by the statistical significance analysis of the identified features. The selection of regularization parameters are presented at the end of this section.

A. Notation and Problem Setup

Assume that we are given a training set with n subjects, each of which is associated with one cohort. The brain connectivity of the i-th subject is represented by a $k \times k$ matrix $\mathbf{R}_i$, with the corresponding label $y_i \in \{0, 1\}$. The proposed DSFE process consists of two key components. First, the bijective operator $g(\cdot)$, denoted as:

$$\mathbf{x}_i = g(\mathbf{R}_i),$$

which performs the feature engineering process to extract a d-dimensional representation $\mathbf{x}_i \in \mathbb{R}^d$ from the connectivity matrix $\mathbf{R}_i$. Second, based on the extracted features $\mathbf{x}_i$, we train a sparse classifier $\mathbf{w} \in \mathbb{R}^d$: to select a small subset of significant feature elements to approximate the true label y_i, such that empirical loss is minimized:

$$\min \mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}),$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ is the aggregation of all available training samples, and $\mathbf{y} = [y_1, \cdots, y_n] \in \{0, 1\}^n$ are the corresponding labels. The empirical loss function $\mathcal{V}(\cdot)$ measures the discrepancy between the predicted label $\hat{y}_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i$ and the given true label y_i. For a classification problem, one often chooses the hinge loss as:

$$\mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \max(0, 1 - y_i \cdot \mathbf{w}^\mathsf{T}\mathbf{x}_i),$$

or the logistic loss as:

$$\mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \cdot \mathbf{w}^\mathsf{T}\mathbf{x}_i}).$$

Hinge loss is popular due to its empirical success in the application of SVMs, while the logistic loss is used in logistic regression. Once the classifier $\mathbf{w}$ is learned, the discriminative edges are obtained by mapping the learned $\mathbf{w}$ back to the connectivity matrix $g^{-1}(\mathbf{w})$.

Since the functional connectivity matrix is usually symmetric, a straightforward method to define the bijection operator $g(\cdot)$ is to vectorize the upper (or lower) triangle part of the connectivity matrix $\mathbf{R}_i$ to form the feature vectors. Given the transformed feature $\mathbf{X}$, a standard linear classifier $\mathbf{w}$ is trained by solving the following regularized quadratic problem:

$$\min_{\mathbf{w}} \mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}) + \lambda \|\mathbf{w}\|_2,$$

where $\|\mathbf{w}\|_2$ is an ℓ_2 regularizer to prevent overfitting, and λ is a regularization parameter to balance the empirical loss and regularization term. Note that solving the minimization problem in Equation II-A often derives dense solutions that do not satisfy the objective of identifying key feature elements. To achieve a sparse selection of discriminative features, one heuristic way is to arbitrarily set a threshold to discard the features with lower amplitude [20]. Although the survived features may be regarded as important with a proper threshold, the classification performance would be very poor. To achieve both sparse solution and higher classification performance, one could alternatively impose an ℓ_1 regularizer to automatically extract the discriminative features, as shown by:

$$\min_{\mathbf{w}} \mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}) + \lambda \|\mathbf{w}\|_1.$$

Minimizing the above ℓ_1 regularized problem leads to sparse solutions, which is considered the key feature of the classification problem. In addition, ℓ_1 regularization based formulation also improves learning performance, especially when limited training samples are available. As proved in [31], ℓ_1 regularized methods require a training set size that grows logarithmically in the number of irrelevant features, while the size of training

samples required for ℓ_2 regularization based approaches grows linearly.

Although the above machine learning techniques can help identify individual discriminative features of two cohorts, their lack of an optimal feature organization may hinder plausible clinical and biological explanations. Therefore, it is advantageous to leverage the functional information of the brain in the formulation. For instance, the GraphNet [32] [23] approach encodes prior associated structures with a positive semi-definite matrix. Combining both the ℓ_1 regularizer and a structure ℓ_2 regularizer, GraphNet can be formulated as follows:

$$\min_{\mathbf{w}} \mathcal{V}(\mathbf{X}, \mathbf{y}, \mathbf{w}) + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \mathbf{w}^{\mathsf{T}} \mathbf{G} \mathbf{w},$$

where matrix $\mathbf{G}$ encodes *a priori* known structure information, and λ_1 and λ_2 are two non-negative regularizer parameters. In [23], the authors directly analyzed the signals of voxels, and defined the structures in the matrix according to the spatial distance of voxels. However, the above model-driven method is highly dependent on the accuracy of prior knowledge about structures or modules in the matrix, which greatly limits its application in complex brain studies such as those investigating neuropsychiatric disease. Hence, it is desirable for a data-driven method to obtain such a structured and discriminative feature set without using advanced information or speculation of the brain mechanisms.

B. Proposed Discriminative Structured Feature Engineering Process

Formulation: As the structures or modules contained in the brain matrix frequently point to the active nodes having multiple connections like the hub [8], [9], [26], [33], the entire symmetric connectivity matrix is vectorized to achieve feature representation as follows:

$$\mathbf{x}_i = g(\mathbf{R}_i) = vec(\mathbf{R}_i),$$

where $vec(\cdot)$ denotes the vectorization operator over a matrix. Instead of posing the regularizer on the structureless vector $\mathbf{w}$ directly, we reconstruct the matrix $\mathbf{W}$ from the coefficient vector $\mathbf{w}$ using:

$$\mathbf{W} = g^{-1}(\mathbf{w}),$$

where $g^{-1}(\cdot)$ is the inverse operator of $g(\cdot)$. Since the matrix $\mathbf{W}$ is symmetric, it is further decomposed into two matrices: a row structured matrix $\mathbf{U}$ and a column structured matrix $\mathbf{V}$ as follows:

$$\mathbf{W} = \frac{1}{2}(\mathbf{U} + \mathbf{V}).$$

We then impose structural penalties $\mathcal{P}$ on these two matrices separately:

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{V}\left(g(\mathbf{U}), g(\mathbf{V}), \mathbf{X}, \mathbf{y}\right) + \lambda \left(\mathcal{P}(\mathbf{U}) + \mathcal{P}(\mathbf{V}^{\mathsf{T}})\right), \qquad (1)$$

and specify the structural penalty $\mathcal{P}(\cdot)$ as a mixed form

$$\mathcal{P}(\mathbf{U}) = \|\mathbf{U}\|_{1,1} + \alpha \|\mathbf{U}\|_{2,1},$$

where the matrix norm $\|\mathbf{U}\|_{1,1} = \sum_i \sum_j |u_{ij}|$, and $\|\mathbf{U}\|_{2,1} = \sum_i (\sum_j u_{ij}^2)^{1/2}$ are the extensions of vector norms. Specifically, the $\ell_{1,1}$ norm over a matrix can be regarded as a natural extension of the ℓ_1 norm over a vector, which encourages the whole matrix to be sparse. For the $\ell_{2,1}$ norm, we first compute the row-wise ℓ_2 norm and then accumulate the ℓ_2 norm using the ℓ_1 norm. Hence, the $\ell_{2,1}$ norm encourages the matrix to be row sparse and leads to the sparse grouping effect. The combination of $\ell_{2,1}$ norm and $\ell_{1,1}$ can be regarded as an improvement of the single $\ell_{2,1}$ norm to sparsify each row vector. Parameter α balances the effects of these two terms and adjusts the sparsity of the row vector.

Optimization: The above procedure of cost optimization in (1) involves two variables, i.e., $\mathbf{U}$ and $\mathbf{V}$. To minimize the cost function and derive optimal solutions, we adopted an alternative optimization strategy as described below.

Given that $\mathbf{V}$ is fixed, the minimization problem with respect to $\mathbf{U}$ is written as

$$\min_{\mathbf{U}} \mathcal{V}\left(g(\mathbf{U}), g(\mathbf{V}), \mathbf{X}, \mathbf{y}\right) + \lambda \mathcal{P}(\mathbf{U}).$$

When the empirical loss function $\mathcal{V}(g(\mathbf{U}), g(\mathbf{V}), \mathbf{X}, \mathbf{y})$ is replaced by the logistic loss, the accelerated proximal gradient method [34] can be conveniently applied to handle the minimization process in order to obtain the optimal $O(1/k^2)$ convergence rate. Briefly speaking, the gradient of the empirical loss is iteratively evaluated and the proximal operator of matrix $\mathbf{U}$ onto structure penalty $\mathcal{P}(\cdot)$ is then optimized:

$$\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{S} - \mathbf{U}\|_2^2 + \lambda_1 \|\mathbf{S}\|_{1,1} + \lambda_2 \|\mathbf{S}\|_{2,1},$$

which can be further divided into $\sqrt{d}$ separated minimizations over vector $\mathbf{s}_{i\cdot}$:

$$\min_{\mathbf{s}_{i\cdot}} \frac{1}{2}\|\mathbf{s}_{i\cdot} - \mathbf{u}_{i\cdot}\|_2^2 + \lambda_1 \|\mathbf{s}_{i\cdot}\|_1 + \lambda_2 \|\mathbf{s}_{i\cdot}\|_2.$$

The optimal solution of the above problem can be computed analytically [35]:

$$\mathbf{s}_{i\cdot} = \begin{cases} \frac{\max\{0, \|\mathbf{h}\|_2 - \lambda_2\}}{\|\mathbf{h}\|_2}\mathbf{h}, & \|\mathbf{h}\|_2 > 0; \\ \mathbf{0}, & \|\mathbf{h}\|_2 = 0; \end{cases}$$

where $\mathbf{h} = sign(\mathbf{u}_{i\cdot}) \cdot \max\{0, |\mathbf{u}_{i\cdot}| - \lambda_1\}$. Similar derivation can be obtained to compute optimal $\mathbf{V}$ when matrix $\mathbf{U}$ is fixed.

Due to the symmetry of the connectome data $\mathbf{X}$, the constraint $\mathbf{U} = \mathbf{V}^{\mathsf{T}}$ is always satisfied. In other words, the final optimization is actually based on univariate minimization. Therefore, the overall problem is convex and the global optimum is always guaranteed.

C. Evaluation of Statistical Significance

Statistical significance is an integral part of statistical hypothesis testing that helps researchers determine whether a null hypothesis should be rejected or retained. However, the probability of observing significant features has been largely overlooked by the machine learning community. The lack of statistical significance has become problematic for neurobiological and clinical studies particularly because researchers are inclined

to pin down specific or combined identified features (i.e., network hubs or group of connectivities) explicitly as the neural circuit basis of certain brain cognitions or diseases. At present there are two temporary resort options available to circumvent this limitation. One is simply to convert the absolute value of coefficients of each classifier proportionally as a substitute for the wanted confidence level. Hence, an arbitrary threshold is set to cut off the identified features into significant and insignificant categories. Another is to perform permutation tests to examine the significance of individual interested features. To obtain a reliable p-value, a large number of permutation tests are usually applied, which entail considerable computational load in addition to the machine learning process. Nevertheless, permutation tests are conservative in estimating statistical power (i.e., the lowest p-value is 0.001 for $1,000$ permutations). To address this issue, we derived an optimization-based formulation to determine the corresponding statistical significance while attaining the identified features.

The optimal solution is denoted as $\hat{\mathbf{W}}$, and the corresponding vectorized representation as $\hat{\mathbf{w}} = g(\hat{\mathbf{W}})$. As $\hat{\mathbf{w}}$ is the solution of a regularized process, it is predisposed to be a biased estimator. Therefore, we first procure the de-biased solution $\hat{\mathbf{w}}^u$ as [36], [37]:

$$\hat{\mathbf{w}}^u = \hat{\mathbf{w}} + \mathbf{M}\boldsymbol{\theta},$$

where matrix $\mathbf{M}$ is a step length matrix to decorrelate the columns of $\mathbf{X}$, and $\boldsymbol{\theta}$ is the de-biased vector which is the gradient of empirical loss:

$$\boldsymbol{\theta} = \frac{1}{n}\sum_i \frac{y_i \mathbf{x}_i exp(-y_i \mathbf{w}^\mathsf{T}\mathbf{x}_i)}{1 + exp(-y_i \mathbf{w}^\mathsf{T}\mathbf{x}_i)}.$$

Afterwards the de-biased solution $\hat{\mathbf{w}}$ can be approximated by the Gaussian distribution with zero mean $\mathbf{M}\hat{\Sigma}\mathbf{M}$, where $\hat{\Sigma} = (1/n)\sum_i (exp(y_i \mathbf{w}^\mathsf{T}\mathbf{x}_i)/(1+exp(y_i \mathbf{w}^\mathsf{T}\mathbf{x}_i))^2)\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ is the empirical covariance.

To minimize both empirical loss and Gaussian distribution variance, each row vector $\mathbf{m}_i$ of matrix $\mathbf{M}$ can be optimized as [37]:

$$\begin{aligned} &\text{minimize} \quad \mathbf{m}_i \hat{\Sigma} \mathbf{m}_i^\mathsf{T} \\ &\text{subject to} \quad \left\| \hat{\Sigma}\mathbf{m}_i - \mathbf{e}_i \right\|_\infty \le \gamma \end{aligned} \tag{2}$$

where $\mathbf{e}_i$ is a unit vector with i-th element equaling 1. The symbol $\|\cdot\|_\infty$ denotes the maximum element of a vector, and γ is a small positive number. Optimization problem (2) can be efficiently solved here using the standard convex optimization toolbox (i.e., cvx [38]).

For each element of vector $\mathbf{w}$, we denote the null hypothesis as $H_{0,i} : w_i = 0$, and the alternative as $H_{1,i} : w_i \neq 0$. The p-value for this hypothesis test $H_{0,i}$ is determined as follows:

$$p_i = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\hat{w}_i|}{[\mathbf{M}\hat{\Sigma}\mathbf{M}]_{i,i}^{\frac{1}{2}}}\right)\right).$$

D. Selection of Regularization Parameters

The determination of appropriate regularization parameters, i.e., λ and α, is of paramount importance in this proposed method. The parameter λ balances the empirical loss and the structural penalty, while parameter α balances the effect of the $\ell_{1,1}$ and $\ell_{2,1}$ sparsity inducing norms. For parameter λ, an iterative process is utilized to estimate a suitable value until the number of selected features is less than a predefined threshold (discussed in more detail later). This strategy is based on the fact that the number of identified features is proportional to the number of training samples using the ℓ_1 regularization, as discussed in [31]. To balance the effect of the $\ell_{1,1}$ and $\ell_{2,1}$ sparsity inducing norms, α is chosen from $\{2^1, 2^0, 2^{-1}, \cdots, 2^{-10}\}$. The Akaike information criterion (AIC) [39] is then employed to determine a suitable choice of α:

$$AIC = 2k_0 - 2\ln(L),$$

where k_0 is the number of nonzero entries of upper triangular part of $\mathbf{W}$, and L is the likelihood function.

III. DATASETS

We used three datasets to validate the proposed DSFE method. The first was a synthetic dataset, with known ground truth for the discriminative features. The second was the ketamine Dataset, which was a resting-state functional connectivity dataset of macaques after injection with ketamine or saline. The third was an obsessive-compulsive disorder (OCD) dataset, which was a functional connectivity dataset of OCD patients and healthy controls. All MRI images of ketamine and OCD datasets were acquired at the Institute of Neuroscience on a 3T whole-body scanner (Tim Trio; Siemens Healthcare, Erlangen, Germany), with some recorded on an enhanced animal dedicated gradient system (AC88; 80 mT/m maximum gradient strength, 800 $mT \cdot m^{-1} \cdot s^{-1}$ maximum slew rate).

A. Synthetic Dataset

The complex infrastructure of the brain contains various kinds of clustered groups or modules (e.g. primary sensory systems such as visual, somatosensory and motor areas) and scattered groups or association loci (e.g. fronto-parietal task control region) [40]. Artificial data were thus generated to represent two typical scenarios: fully structured data, and partially structured data. The first case is the ideal scenario that fits our structured discriminative feature assumption perfectly, while the second case is more realistic and fits our assumption partially. Although it appears to be an extreme oversimplification of the brain network matrix, it is allowable to begin by such a proof-of-concept experiment with no model assumptions imposed in return.

Each element of the ground truth matrix $\mathbf{W}$ was randomly drawn from a Gaussian distribution $\mathcal{N}(0,1)$ independently. The entries with absolute magnitude below $\sqrt{\log(p/n)}$ were set to zero to exclude the relatively small values, where the dimensionality of data was p and the size of training samples was n. Each element of data sample $\mathbf{x}$ was also drawn from $\mathcal{N}(0,1)$. A linear combination $\tilde{y} = vec(\mathbf{W})' \cdot vec(\mathbf{x})$ was evaluated as the output accordingly. To keep the margin between two classes sufficiently large, we only accepted samples with $|\tilde{y}| < 0.8$. Finally, the binary labels were computed as $y = sign(\tilde{y})$.

B. Ketamine Dataset

Resting-state functional MRI data were collected from nine anesthetized monkeys 18 h after ketamine or saline administration. The purpose of this animal study was to investigate the sustained antidepressant effect of a single subanesthetic dose of ketamine, with details described elsewhere [41]. There were a total of 76 valid imaging datasets for statistical comparison between the two cohorts (39 for the ketamine condition and 37 for the saline condition). All experimental procedures for non-human primate subjects in this study were approved by the Institute of Neuroscience Animal Care and Use Committee and by the Shanghai Institutes for Biological Sciences Biomedical Research Ethics Committee, and conformed to the National Institutes of Health guidelines for the humane care and use of laboratory animals.

The parameters for T1-weighted imaging were: $TR = 2,500$ ms, $TE = 3$ ms, $TI = 1,100$ ms, $voxelsize = 0.5 \times 0.5 \times 0.5$ mm^3. The parameters for resting-state fMRI were set as: $TR = 2,000$ ms, $TE = 29$ ms, $voxelsize = 1.5 \times 1.5 \times 2.5$ mm^3. Each hemisphere was divided into 39 ROIs for cortical areas using the F99 macaque standard surface template [42] and eight sub-cortical areas (amygdala, hippocampus, thalamus, hypothalamus, caudate, putamen, palladium, and nucleus accumbens) from the INIA 19 template [43]. Thus, for each subject, we constructed a 94×94 matrix with each entry representing the Pearson correlation coefficient of the averaged time courses of two ROIs, which reflected the functional synchronization of distinct brain areas.

C. Obsessive-Compulsive Disorder (OCD) Dataset

The OCD dataset consisted of 37 OCD patients, and 37 healthy controls. The objective of this human study was to examine the neural deficits of OCD patients relative to healthy people and deepen our understanding of the neurophysiological and neuropathological mechanisms of OCD. The detailed biological findings are described elsewhere [44]. All subjects were adults, with an average age of 32.2 ± 8.1 for OCD patients and 31.1 ± 8.5 for healthy controls. In addition, subjects with brain trauma, previous neurosurgerical procedures, drug addiction, or nervous system diseases were excluded from the study. The research was approved by the Ethics Committee of Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University. After being provided with a complete description of the study, participants gave written informed consent.

The parameters for T1-weighted imaging were: $TR = 2,300$ ms, $TE = 3$ ms, $TI = 1,000$ ms, $voxelsize = 1 \times 1 \times 1$ mm^3. The parameters for resting-state fMRI were set as: $TR = 2,000$, $TE = 30$ ms, $voxelsize = 2 \times 2 \times 3$ mm^3. We used FreeSufer software (http://surfer.nmr.mgh.harvard.edu/) to register and segment each hemisphere into 41 ROIs including 34 gyral-based cortical parcels using the Desikan-Killiany atlas [45] and 7 subcortical segments (amygdala, hippocampus, thalamus, caudate, putamen, palladium, and nucleus accumbens). In the end, for each subject, we constructed a 82×82 symmetric matrix with each entry representing the Pearson correlation coefficients between the time series between pairs of ROIs.

IV. Results

In this section, we first compare the proposed DSFE method with univariate t-test and ℓ_1 regularized logistic regression on the synthetic dataset. Both the proposed DSFE method and two competing methods were repeated 20 times to average out the random effects of data generation. We then verify our method on two real functional connectome datasets. The prediction performance is validated using the leave-one-out strategy. Significant and identified alternations are discussed.

As mentioned in Section II-D, the coefficient λ was determined to satisfy the reconstruction constraints, while parameter α was determined using the Akaike information criterion. In this experiment, λ was initialized at 0.02 and doubled for each iteration. γ was set at 0.05, though values within [0.01, 0.05] would also produce similar results. In addition, we shrank the set of identified features by zeroing those elements with magnitudes of less than $\sqrt{\log(p/n)}$. The code of our DSFE is available at http://www.ion.ac.cn/laboratories/int.asp?id=79, and two real connectome datasets are also available on request.

A. Application to Synthetic Data

We first compared the proposed DSFE method using three competing approaches: the univariate t-test, ℓ_1 regularized logistic regression and stability selection [22]. The univariate t-test is regarded as a basic baseline method. ℓ_1 regularized logistic regression is considered as a baseline to verify the effectiveness of our structural penalty, and stability selection is a representative of randomized methods. We used the F-measure, which has been widely used in information retrieval, to compare the overall performance of false positive rate and false negative rate:

$$\mathrm{F-measure} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

To evaluate the power of the proposed DSFE, univariate t-test and ℓ_1 regularized logistic regression, we plotted the F-measure when $1, 2, \cdots, 15$ features were selected. The number of ground truth features is designed to be between 25 and 100. For the univariate t-test, the features were selected according to their significance level. For ℓ_1 regularized logistic regression, the regularization parameter was automatically tuned using a binary search such that the number of nonzero features of the learned model matched the desired number. For stability selection, we randomly selected half of the total samples for 100 tests. For the proposed DSFE, parameter α and λ were tuned automatically. More specifically, α was initialized as 0.05, after which λ was tuned using the same binary search. If the above step didn't converge, α was reset randomly and λ was tuned.

The mean F-measure for 20 random trials is shown in Fig. 2. For both settings of the ground truth (i.e., fully structured and partially structured), the proposed DSFE method (red solid line) consistently outperforms the univariate t-test, ℓ_1 regularized logistic regression and stability selection. Note that the F-measure improvement of our method over ℓ_1 regularized logistic regression is almost the same size as the improvement of ℓ_1 regularized logistic regression over the univariate t-test.

We also note that there is a significant drop of F-measure when more features are selected, which means the smaller the feature set we identify, the more reliable the features would be.

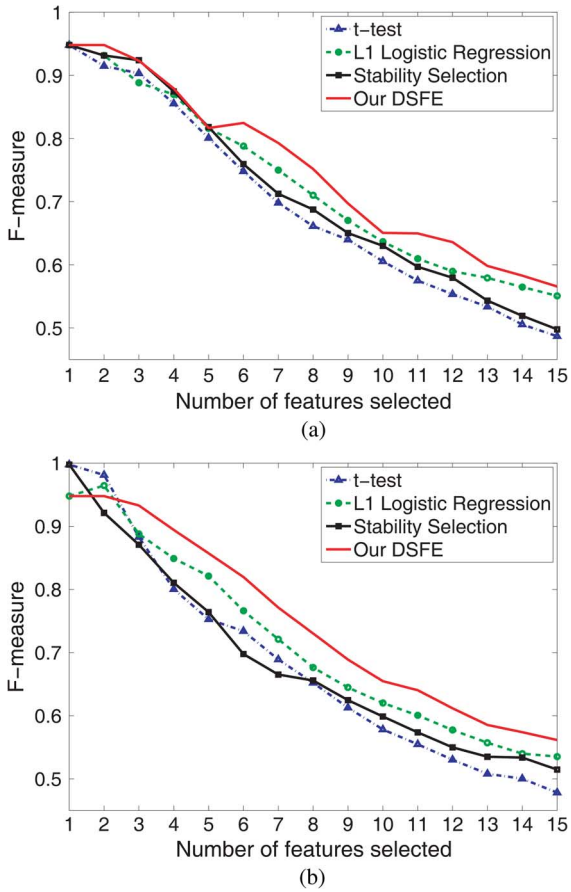

Fig. 2. Comparison of F-measure of univariate t-test, ℓ_1 logistic regression, stability selection and the proposed DSFE method by selecting various numbers of features. The setting of ground truth model: (a) Fully structured; (b) Partially structured.

B. Application to Monkey Brain Connectome

The DSFE approach was applied to identify the discriminative features in the functional connectivity matrices of the ketamine and saline groups. We first analyzed how the number of identified features affects the performance of prediction. The analysis was based on the leave-one-out strategy, i.e., in each round, one was used for testing, and the others were used for training. Five settings of the number of identified features were considered: 1–20, 21–100, 101–200, 201–1000 and the baseline using all 4371 features. For the first four nontrivial settings, parameter λ was automatically tuned using binary search to satisfy such constraints, while α was chosen using the AIC criterion described in Section II-D. After that, linear SVMs with default settings were trained using these identified features. The final prediction performance is shown in Fig. 3. Notice that the AUC score (0.8344) is significant when a parsimonious model is trained (using 1–20 identified features). It implies that a small subset of brain connections is powerful enough to reflect the effect of a ketamine injection. We also note that the AUC score using 201–1000 features (0.9446) is significantly higher than the baseline using all 4171 features (0.9161), which collaborated the effectiveness of the proposed DSFE.

Recall the discussion in Section IV-A, the F-measure may decrease with the increment of the number of identified features.

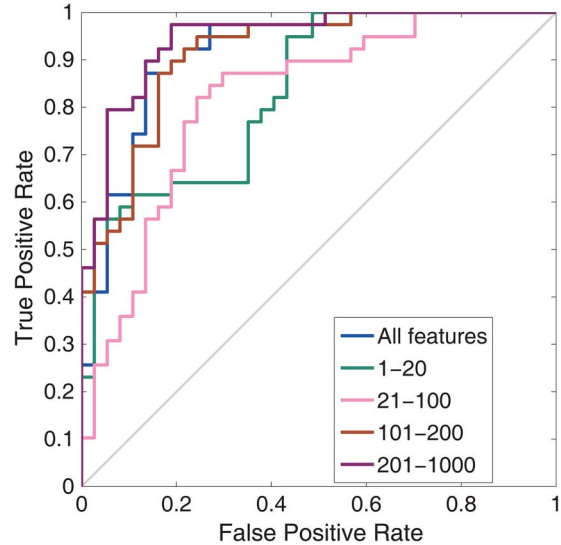

Fig. 3. ROC curves in the ketamine dataset using all connections and a selected number of connections identified by DSFE. Areas under the ROC curves (AUC) for all connections, 1–20, 21–100, 101–200 and 201–1000 are: 0.9161, 0.8344, 0.8177, 0.9120 and 0.9446, respectively.

TABLE I
IDENTIFIED DISCRIMINATIVE FUNCTIONAL CONNECTIVITY BETWEEN KETAMINE AND SALINE GROUPS. ALTERED CONNECTIVITIES ARE LISTED IN DESCENDING ORDER OF STATISTICAL SIGNIFICANCE. CONNECTIONS ABOVE THE DASHED LINE WERE STATISTICALLY SIGNIFICANT ($p < 0.05$)

Brain Regions	Brain Regions	Classifier Weights	Z-score
Ia.L	HC.L	1.01	7.62
Amyg.L	PFCoi.R	0.47	5.81
S1.R	HC.R	0.48	4.94
CCs.L	PMCdl.R	1.58	4.48
HT.R	PCi.R	0.57	4.47
CCs.L	PCi.R	0.51	4.17
Tha.L	CCa.L	1.11	3.98
S1.R	G.R	0.48	2.91
PMCdl.L	CCs.L	1.10	1.57
Ip.L	PMCdl.R	0.48	1.30

Hence, to balance both the number of identified features and prediction performance, we regarded 101–200 as an appropriate range, and the expected number of identified features was set to 150.

Table I lists the identified connections between brain regions in descending order of Z-score, with statistically significant features after Bonferroni correction shown above the dashed line ($p < 0.05$). It should be noted that the orders of classifier weights and Z-score are not relevant to each other because the Z-score is not only affected by the weights of the classifier, but also by second-order characteristics of the data. Thus, it is not appropriate to use classifier weights directly as a substitute for statistical hypothesis tests in some biological applications. To better visualize the identified features of the functional matrices, the alterations in functional connectivity were illustrated from different angles of view in a reconstructed 3D monkey brain (Fig. 4).

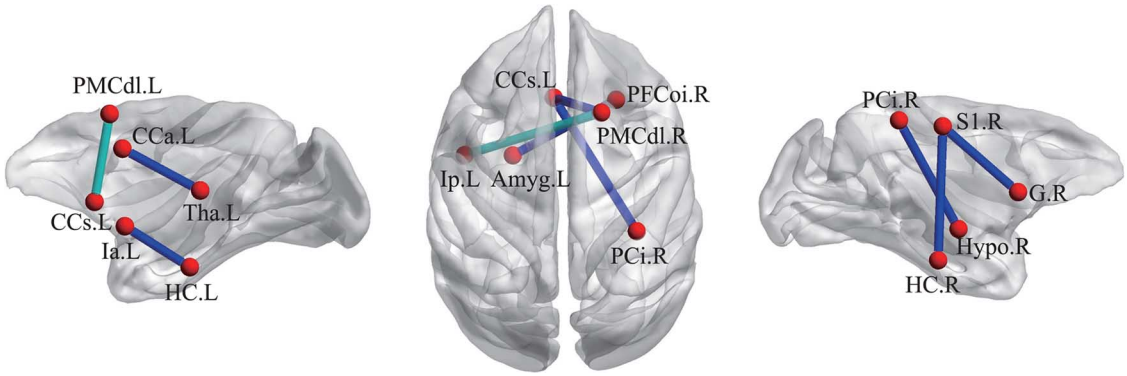

Fig. 4. Illustration of functional connectivity significantly altered by ketamine intake compared to saline control. Red dots indicate brain nodes (ROIs) and dark blue lines and cyan lines indicate the identified significant and identified insignificant connections, respectively.

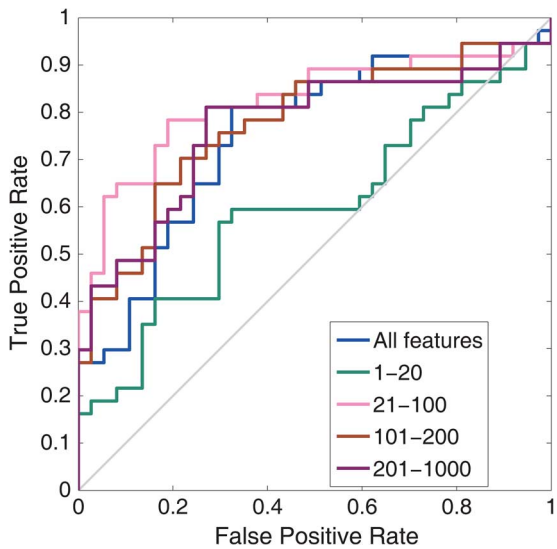

Fig. 5. ROC curves in the OCD dataset using all connections and a selected number of connections identified by DSFE. AUC for all connections, 1–20, 21–100, 101–200 and 201–1000 are: 0.7524, 0.5975, 0.8218, 0.7757 and 0.7692, respectively.

The discriminative connections extracted from the whole brain functional connectivity matrix in response to ketamine intake exhibited a synergistic regulation of the cortico-limbic-striatal circuits, mostly targeting the limbic system (anterior insula, hippocampus, amygdala, thalamus, hypothalamus and palladium), cingular cortex (subgenual and anterior sections), orbitoinferior prefrontal cortex, and somatosensory system. Our findings are consistent with present knowledge of the pharmacological effects of ketamine on the brain [29] and complementary to previous studies on human subjects [46]–[48].

C. Application to Human Brain Connectome

Lastly, we tested the DSFE method to compare the brain connectome of obsessive-compulsive disorder (OCD) patients from that of healthy controls (HC). The procedure we used on this dataset is identical to the one we used on the Ketamine dataset. As shown in Fig. 5, the AUC score (0.5975) is close to chance when a parsimonious model is trained (using 1–20 identified features). We also note that the AUC score using 21–100 feature significantly outperforms the baseline method using all features. Therefore, we set the expected number of features to 60.

TABLE II
IDENTIFIED DISCRIMINATIVE FUNCTIONAL CONNECTIVITY BETWEEN OCD PATIENTS AND HEALTHY CONTROLS. ALTERED CONNECTIVITIES ARE LISTED IN DESCENDING ORDER OF STATISTICAL SIGNIFICANCE. CONNECTIONS ABOVE THE DASHED LINE WERE STATISTICALLY SIGNIFICANT ($p < 0.05$)

Brain Regions	Brain Regions	Classifier Weights	Z-score
Amyg.R	PARH.L	0.91	2.27
PCUN.L	CAC.R	0.37	2.00
POPE.R	IP.L	0.35	1.66
RAC.L	PTRI.R	1.08	1.52
GP.L	INS.R	0.38	1.30
PORB.L	BSTS.L	0.38	1.22
Tha.R	SP.R	0.35	1.18
PTRI.L	BSTS.L	0.90	1.18
RMF.L	IP.L	0.43	1.14
GP.L	FUS.L	0.46	0.95
RAC.R	PTRI.R	0.71	0.87
Tha.R	BSTS.R	0.45	0.86

All identified features are summarized in Table II. Although most of the selected features are insignificant, the significant alterations of the amygdala and the anterior cingulate cortex are consistent with the results of [30], [49]–[51]. All identified alterations of functional connectivity are also illustrated from different angles of view in a reconstructed 3D human brain in Fig. 6. These results confirm previous reports regarding the abnormalities of neural circuits in OCD patients, which are attributed to the cortico-striato-thalamo-cortical (CSTC) circuits [30]: the orbitofrontal cortex, the anterior cingulate cortex, the inferior frontal gyrus, and the thalamus.

V. CONCLUSION

We proposed a novel data-driven method, namely discriminative structured feature engineering (DSFE), to identify structured discriminative features for group comparison by exploring a logistic regression with mixed norm regularization. To validate its routine use in biomedical application and clinical practice, we integrated statistical analysis to rigorously evaluate the reliability of the identified features. Our DSFE method and associated statistical analysis were formulated as convex optimization problems that can be solved efficiently. Simulation experiments on the synthetic data showed that DSFE is more pow-

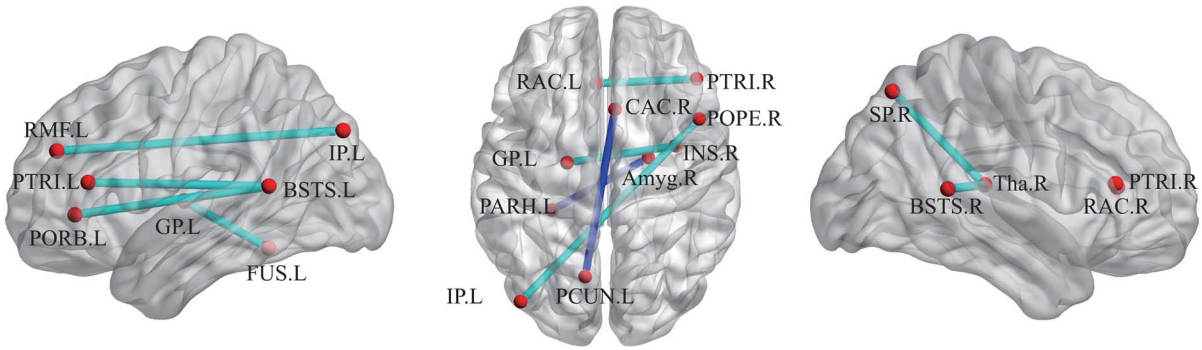

Fig. 6. Illustration of functional connectivity significantly altered by OCD patients compared to healthy control. Red dots indicate brain nodes (ROIs) and dark blue lines and cyan lines indicate the identified significant and identified insignificant connections, respectively.

erful in identifying the correct features than the traditional univariate t-test and traditional logistic regression. Functional brain connectome data results of both human and nonhuman primate analysis further demonstrated that DSFE efficiently deciphered the network model-based matrix with high dimensionality and distinguished the brains of healthy controls from subjects under pharmacological therapy or with an array of neurological and psychiatric disorders.

Despite substantial growth in the quantity of brain connectome data, from the microscale of individual synaptic connections between neurons to the macroscale of brain regions and interregional pathways, multivariate exploratory analyses of large datasets for understanding complex biological systems have not yet received sufficient attention. Beyond classification with SVMs commonly used in the biomedicine, we attempted for the first time to exploit the intrinsic nature of brain architecture (small-world organization) and identify discriminative structured features with statistical power. Our approach paves a new avenue for tapping into functional brain matrices and for obtaining biologically and clinically informative results in exploratory analysis that may aid in the development of clinically useful prognoses and therapeutics in the future.

Acknowledgment

The authors would like to thank the anonymous reviewers and associate editor for their constructive comments and suggestions. The authors also thank Drs. Ravi Menon, Lawrence Wald, John Gore, Franz Schmitt, Renate Jerecic, Thomas Benner, Kecheng Liu, Ignacio Vallines, and Hui Liu for their generous help and contribution to the construction of our custom-tuned gradient-insert (AC88) 3T MRI facility on nonhuman primate subjects.

References

[1] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: A structural description of the human brain," *PLoS Comput. Biol.*, vol. 1, no. 1, p. e42, 2005.
[2] R. C. Craddock *et al.*, "Imaging human connectomes at the macroscale," *Nature Methods*, vol. 10, no. 6, pp. 524–539, 2013.
[3] N. B. Turk-Browne, "Functional interactions as big data in the human brain," *Science*, vol. 342, no. 6158, pp. 580–584, 2013.
[4] J. E. Gabrieli, S. Ghosh, and S. Whitfield-Gabrieli, "Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience," *Neuron*, vol. 85, no. 1, pp. 11–26, 2015.
[5] F. X. Castellanos, A. D. Martino, R. C. Craddock, A. D. Mehta, and M. P. Milham, "Clinical applications of the functional connectome," *NeuroImage*, vol. 80, no. 1, pp. 527–540, 2013.
[6] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
[7] C. Echtermeyer *et al.*, "Integrating temporal and spatial scales: Human structural network motifs across age and region-of-interest size," *Front. Neuroinform.*, vol. 5, no. 10, pp. 1–14, 2011.
[8] O. Sporns, C. J. Honey, and R. Kötter, "Identification and classification of hubs in brain networks," *PLoS ONE*, vol. 2, no. 10, p. e1049, 2007.
[9] J. D. Power, B. L. Schlaggar, C. N. Lessov-Schlaggar, and S. E. Petersen, "Evidence for hubs in human functional brain networks," *Neuron*, vol. 79, no. 4, pp. 798–813, 2013.
[10] E. T. Bullmore and D. S. Bassett, "Brain graphs: Graphical models of the human brain connectome.," *Annu. Rev. Clin. Psychol.*, vol. 7, no. 1, pp. 113–140, 2011.
[11] Y.-T. Chang, D. Pantazis, and R. M. Leahy, "To cut or not to cut? assessing the modular structure of brain networks," *NeuroImage*, vol. 91, no. 1, pp. 99–108, 2014.
[12] M. A. de Reus and M. P. van den Heuvel, "Rich club organization and intermodule communication in the cat connectome.," *J. Neurosci.*, vol. 33, no. 32, pp. 12929–12939, 2013.
[13] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *J. Neurosci.*, vol. 31, no. 44, pp. 15,775–15,786, 2011.
[14] M. P. van den Heuvel and O. Sporns, "Network hubs in the human brain," *Trends Cognitive Sci.*, vol. 17, no. 12, pp. 683–696, 2013.
[15] D. E. Meskaldji *et al.*, "Adaptive strategy for the statistical analysis of connectomes," *PLoS ONE*, vol. 6, no. 8, p. e23009, 2011.
[16] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Mller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, 2011.
[17] J. Richiardi, S. Achard, H. Bunke, and D. V. D. Ville, "Machine learning with brain graphs: Predictive modeling approaches for functional imaging in systems neuroscience," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 58–70, May 2013.
[18] L. L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu, "Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis," *Brain*, vol. 135, no. 5, pp. 1498–1507, 2012.
[19] N. Dosenbach *et al.*, "Prediction of individual brain maturity using fMRI," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.
[20] F. Li *et al.*, "Multivariate pattern analysis of DTI reveals differential white matter in individuals with obsessive-compulsive disorder," *Hum Brain Mapp.*, vol. 35, no. 6, pp. 2643–2651, 2013.
[21] J. M. Rondina *et al.*, "SCoRS—A method based on stability for feature selection and mapping in neuroimaging," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 85–98, Jan. 2014.
[22] N. Meinshausen and P. Bühlmann, "Stability selection," *J. R. Stat. Soc. Ser. B*, vol. 72, no. 4, pp. 417–473, 2010.
[23] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with GraphNet.," *NeuroImage*, vol. 72, no. 1, pp. 304–321, 2013.
[24] M. P. van den Heuvel and A. Fornito, "Brain networks in schizophrenia.," *Neuropsychol. Rev.*, vol. 24, no. 1, pp. 32–48, 2014.

[25] M. Rubinov and E. Bullmore, "Schizophrenia and abnormal brain network hubs.," *Dialogues Clin. Neurosci.*, vol. 15, no. 3, pp. 339–349, 2013.
[26] S. Achard *et al.*, "Hubs of brain functional networks are radically reorganized in comatose patients.," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 50, pp. 20,608–20,613, 2012.
[27] X. Miao, X. Wu, R. Li, K. Chen, and L. Yao, "Altered connectivity pattern of hubs in default-mode network with Alzheimer's disease: An Granger causality modeling approach.," *PLoS One*, vol. 6, no. 10, p. e25546, 2011.
[28] N. A. Crossley *et al.*, "The hubs of the human connectome are generally implicated in the anatomy of brain disorders.," *Brain*, vol. 137, no. 8, pp. 2382–2395, 2014.
[29] B. Sinner and B. Graf, "Ketamine," in *Modern Anesthetics*, J. Schttler and H. Schwilden, Eds. Berlin, Germany: Springer, 2008, vol. 182, Handbook Exp. Pharmacol., pp. 313–333.
[30] M. R. Milad and S. L. Rauch, "Obsessive-compulsive disorder: Beyond segregated cortico-striatal pathways," *Trends Cognitive Sci.*, vol. 16, no. 1, pp. 43–51, 2012.
[31] A. Ng, "Feature selection, l1 vs. l2 regularization, rotational invariance," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 78–86.
[32] M. Slawski, W. zu Castell, and G. Tutz, "Feature selection guided by structural information," *Ann. Appl. Stat.*, vol. 4, no. 2, pp. 1056–1080, 2010.
[33] M. P. van den Heuvel and O. Sporns, "Network hubs in the human brain," *Trends Cognitive Sci.*, vol. 17, no. 12, pp. 683–696, 2013.
[34] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[35] P. Sprechmann, I. Ramłrez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, 2011.
[36] A. Javanmard and A. Montanari, "Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6522–6554, Oct. 2013.
[37] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional statistical models," in *Adv. Neural Inf. Process. Syst.*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Red Hook, NY: Curran Associates, Inc., 2013, pp. 1187–1195.
[38] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming version 2.1," 2014 [Online]. Available: http://cvxr.com/cvx
[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Statistics. New York: Springer, 2001.
[40] J. Power *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.
[41] Q. Lv *et al.*, "Large-scale persistent network reconfiguration induced by ketamine in anesthetized monkeys: Relevance to mood disorders," *Biol. Psychiatry*, to be published.
[42] D. C. V. Essen *et al.*, "An integrated software suite for surface-based analyses of cerebral cortex," *J. Am. Med. Inform. Assoc.*, vol. 8, no. 5, pp. 443–459, 2001.
[43] T. Rohlfing *et al.*, "The INIA19 template and NeuroMaps atlas for primate brain image parcellation and spatial normalization," *Front. Neuroinf.*, vol. 6, no. 27, pp. 1–15, 2012.
[44] Q. Lv *et al.*, "Structural abnormalities of habitual and goal-directed systems in refractory obsessive-compulsive disorder," *Biol. Psychiatry*.
[45] R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.
[46] F. Musso *et al.*, "Ketamine effects on brain function: Simultaneous fMRI/EEG during a visual oddball task," *NeuroImage*, vol. 58, no. 2, pp. 508–525, 2011.
[47] M. Scheidegger *et al.*, "Ketamine decreases resting state functional network connectivity in healthy subjects: Implications for antidepressant drug action," *PLoS ONE*, vol. 7, no. 9, p. e44799, 2012.
[48] M. Niesters *et al.*, "Effect of subanesthetic ketamine on intrinsic functional brain connectivity: A placebo-controlled functional magnetic resonance imaging study in healthy male volunteers," *Anesthesiology*, vol. 117, no. 4, pp. 868–877, 2012.
[49] O. A. van den Heuvel *et al.*, "Amygdala activity in obsessive-compulsive disorder with contamination fear: A study with oxygen-15 water positron emission tomography," *Psychiatry Res., Neuroimag.*, vol. 132, no. 3, pp. 225–237, 2004.
[50] D. Simon, N. Adler, C. Kaufmann, and N. Kathmann, "Amygdala hyperactivation during symptom provocation in obsessive-compulsive disorder and its modulation by distraction," *NeuroImage, Clin.*, vol. 4, no. 1, pp. 549–557, 2014.
[51] K. D. Fitzgerald *et al.*, "Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder," *Biol. Psychiatry*, vol. 57, no. 3, pp. 287–294, 2005.