Medical Image Analysis 63 (2020) 101662

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Self-weighted adaptive structure learning for ASD diagnosis via multi-template multi-center representation

Fanglin Huang[a], Ee-Leng Tan[b], Peng Yang[a], Shan Huang[a], Le Ou-Yang[c], Jiuwen Cao[d], Tianfu Wang[a,*], Baiying Lei[a,*]

[a] *National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China*
[b] *School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore*
[c] *Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, College of Information Engineering, Shenzhen University, Shenzhen 518060, China*
[d] *Artificial Intelligence Institute, Hangzhou Dianzi University, Zhejiang 310010, China*

ARTICLE INFO

Article history:
Received 23 September 2019
Revised 13 January 2020
Accepted 31 January 2020
Available online 1 February 2020

Keywords:
Autism spectrum disorder
Pearson's correlation (PC) -based sparse low-rank representation
Multi-template multi-center
Self-weighted adaptive structure learning

ABSTRACT

As a kind of neurodevelopmental disease, autism spectrum disorder (ASD) can cause severe social, communication, interaction, and behavioral challenges. To date, many imaging-based machine learning techniques have been proposed to address ASD diagnosis issues. However, most of these techniques are restricted to a single template or dataset from one imaging center. In this paper, we propose a novel multi-template multi-center ensemble classification scheme for automatic ASD diagnosis. Specifically, based on different pre-defined templates, we construct multiple functional connectivity (FC) brain networks for each subject based on our proposed Pearson's correlation-based sparse low-rank representation. After extracting features from these FC networks, informative features to learn optimal similarity matrix are then selected by our self-weighted adaptive structure learning (SASL) model. For each template, the SASL method automatically assigns an optimal weight learned from the structural information without additional weights and parameters. Finally, an ensemble strategy based on the multi- template multi-center representations is applied to derive the final diagnosis results. Extensive experiments are conducted on the publicly available Autism Brain Imaging Data Exchange (ABIDE) database to demonstrate the efficacy of our proposed method. Experimental results verify that our proposed method boosts ASD diagnosis performance and outperforms state-of-the-art methods.

1. Introduction

Autism spectrum disorder (ASD) is a group of fast-growing and highly heterogeneous neural developmental syndrome of largely unknown etiology, defined by impairments in social functioning, deficits in communication, and unusual behaviors and interests (Lord et al., 2000). According to a survey conducted by the Centers for Disease Control and Prevention (CDC), there were about 1 in 59 children diagnosed as ASD (Baio et al., 2018). It indicates that ASD is a considerable health issue and developing an effective method for timely diagnosis is urgently needed. As ASD is found to be related to the functional connectivity disruptions (Shi et al., 2013), resting-state functional magnetic resonance imaging (rs-fMRI) has been a commonly used method to reveal ASD's mechanism in recent years. Specifically, represented as the correlation of blood-oxygen-level dependent (BOLD) signal (time series) in different brain regions, functional connectivity (FC) can offer more stable and sensitive biomarkers for brain disease diagnosis, such as depression (Craddock, 2010), Alzheimer's disease (AD) (Suk et al., 2016), and ASD (Khan et al., 2013).

Recently, numerous methods have been proposed for FC network construction, based on rs-fMRI data. Pearson's correlation (PC) coefficient is the simplest way of estimating FC between different regions (Smith et al., 2013). However, PC only models the simple linear relationship without involving biological mechanisms of brain diseases which can improve diagnosis results. Lee et al. (2011) proposed a method named sparse representation (SR), which incorporated sparsity prior while constructing FC networks. Since brain networks are not only sparse but also modularized (Sporns, 2011), Qiao et al. (2016) proposed the sparse low-rank representation (SLR), which encoded modularity structure by

* Corresponding authors.
E-mail addresses: tfwang@szu.edu.cn (T. Wang), leiby@szu.edu.cn (B. Lei).

https://doi.org/10.1016/j.media.2020.101662
1361-8415/© 2020 Elsevier B.V. All rights reserved.

adding a rank constraint on the basis of SR. However, the performance achieved by PC-based scheme is obviously better than that of SR-based scheme in ASD diagnosis (Li et al., 2017). Hence, to effectively fit data and encode prior information, we propose a novel method named PC-based sparse low-rank representation (PSLR) to construct the FC networks with physiological significance.

Although there are numerous machine learning techniques proposed for ASD diagnosis based on feature representations extracted from FC networks (Huang et al., 2018; Wang et al., 2018a; Zhao et al., 2018), these methods adopt only a single pre-defined template. Feature representations derived from one single template is unable to reveal the group differences between different populations (ASD patients and normal controls (NC)) comprehensively. By contrast, multi-template based methods, which extract multiple feature sets for a subject based on different templates, can provide distinct yet complementary information to identify different disease status. It thus leads to more promising identification performance (Jin et al., 2015; Koikkalainen et al., 2011; Leporé et al., 2008; Liu et al., 2015; Min et al., 2014). For example, Liu et al. (2015) proposed to identify AD patients from NC and progressive mild cognitive impairment (p-MCI) patients from stable MCI (s-MCI) patients by using features representations derived from multiple atlases (i.e., templates). After the feature selection process of view-centralized multi-atlas, better diagnosis results were obtained. Jin et al. (2015) proposed to segment the publicly available infant Automatic Anatomical Labeling (AAL) atlas with 90 cerebral regions of interest (ROIs) into 203 and 403 sub-ROIs. Based on these ROIs, each subject can get three different scales of FC networks and thus get three different sets of feature representations. By concatenating the multi-template based feature representations of each subject, more promising ASD identification results were achieved.

Combining feature representations derived from multiple templates can effectively enhance diagnosis performance. However, a lot of previous works are limited to a single dataset obtained from a single center of Autism Brain Imaging Data Exchange (ABIDE) database. In practice, ASD images can be obtained from different imaging centers. As different centers use different imaging devices, imaging parameters, and have different age ranges and gender ratios of the subjects, it is unreasonable to obtain a diagnosis model from an imaging center and then apply it directly to another imaging center. Also, there are many methods proposed (Abraham et al., 2017; Chen et al., 2015; Dvornek et al., 2017; Jun et al., 2019; Heinsfeld et al; Kam et al., 2017; Nielsen et al., 2013; Plitt et al., 2015; Zhuang et al., 2017), which gathered data from multiple centers to generate a bigger dataset for diagnosis jointly or directly use the entire ABIDE dataset. Although such method expands the dataset, the heterogeneity across different centers still remains. To handle this issue, some studies are focused on finding relationships between different image centers to further improve diagnosis performance for each center (Wang et al., 2017; Wang et al., 2018b). For example, Wang et al. (2017) developed a multi-center multi-modality learning framework for ASD diagnosis based on the center-center and modality-modality relation. However, the current multi-center based methods involve only one template for classification. Therefore, developing an effective diagnosis method based on multi-template and multi-center data is highly desirable.

Regarding ASD diagnosis, most previous approaches simply concatenate (Min et al., 2014) or average (Koikkalainen et al., 2011; Leporé et al., 2008) multiple sets of features and then perform feature selection and classification algorithms on the new features directly. They overlook the underlying structural information in multi-template data, e.g., local manifold structure. Actually, integrating structural information into the model learning process can further identify the disease-related features, which can then boost the classification results (Jie et al., 2015). However, most studies (Jie et al., 2015; Lei et al., 2018; Liu et al., 2016b) exist some limitations while incorporating structural information. First, those approaches perform the following procedure based on the constructed similarity matrix that derived from raw data and remains unchanged for subsequent processing. However, original data always contains noise, which may reduce the accuracy of the similarity matrix and further disrupt the local manifold structure. Second, the similarity matrix obtained by the conventional methods only has one connected component. The number of connected components of the similarity matrix should be the same as the number of class labels in the ideal state (Nie et al., 2016). Note that the size of similarity matrix is $N \times N$, where N is number of subjects, and the FC network per person is $r \times r$, where r is the number of ROI. In addition, most approaches combine local structures in different templates via additional weight parameters, which make the model selection difficult.

To explore the structural information existing in multi-template data without additional tuning weight parameters, we propose a novel self-weighted adaptive structure learning method for ASD diagnosis based on multi-template multi-center ensemble scheme (SASL-E). To the best of our knowledge, no previous works adaptively explored the structure information in multi-template data to facilitate ASD classification. We evaluate the proposed SASL-E method on the ABIDE database. The extensive experiments imply that our proposed method is feasible for ASD diagnosis. The three main contributions of this article are:

- A network construction method is proposed to generate multiple FC networks for each subject. With the sparse and low-rank regularizers, we can encode modularity prior.
- A classification model based on multi-template data is proposed for each center. Based on the multi-task learning framework, the most discriminative features in each template can be selected by introducing the structural information.
- An ensemble strategy is further utilized for classification models corresponding to the image centers, the final ASD diagnosis results can be achieved from the multi-template multi-center data.

The rest of the article is arranged as follows. Section 2 reviews some network building and manifold learning methods. Section 3 provides information about the imaging dataset, including subjects and data preprocessing. The proposed method for ASD diagnosis and its optimization process is detailed in Sections 4 and 5, respectively. Section 6 analyzes the experimental results. Our discussion is provided in the next section followed by the conclusion section.

2. Related work

2.1. FC network construction

The widely used scheme to model FC network is based on correlation (full correlation or partial correlation) statistics, which tends to perform better than that based on higher-order statistics (Smith et al., 2011). The typical correlation-based methods include PC and SR, etc. PC models FC network by measuring the relationship between different ROIs. Due to its simplicity, high computational efficiency and statistical robustness, it becomes the most popular approach. However, it only models full correlation involving confusion effects from other ROIs. Conversely, partial correlation can mitigate this issue by regressing out the potential confounding influence. However, during the process of estimating partial correlation, the inverse operation of the covariance matrix is involved, which tends to be ill-posed, especially when the dimension of time series is fewer than the number of ROIs. To address this issue, the regularization terms are generally introduced to the

Table 1
The demographic information and acquisition parameters of data.

Center	NYU	UM_1	UCLA_1	Yale
Male/female	136/35	59/23	49/6	34/14
Age	15.35±6.59	13.89±2.88	13.57 ± 2.34	12.87 ± 2.93
Patient/control	73/98	36/46	28/27	22/26
Devices(model)	Siemens (Magnetom Allerga)	GE (Signa)	Siemens (Magnetom TrioTim)	Siemens (Magnetom TrioTim)
Voxel size(mm^3)	3.0 × 3.0 × 4.0	3.438 × 3.438 × 3.0	3.0 × 3.0 × 4.0	3.4 × 3.4 × 4.0
Flip angle(deg)	90	90	90	60
TR(ms)	2000	2000	3000	2000
TE(ms)	15	30	28	25
Bandwidth(Hz/Px)	3906	NA	2442	2520

corresponding mathematical model, through which we can not only obtain a stable solution, but also incorporate prior information, such as sparsity (Lee et al., 2011; Rosa et al., 2015), group-sparsity (Wee et al., 2014), low-rank (Liu et al., 2013), and modularity (Qiao et al., 2016).

2.2. Manifold learning

According to the concept of manifold learning, there is always a low- dimensional manifold which can represent the structure of high-dimensional data (Nie et al., 2016). As some high- dimensional data produces dimensional redundancy, manifold learning that can map data from high-dimensional to low-dimensional has triggered abundant concern in various applications, especially in feature learning (Cai et al., 2010; Hou et al., 2014; Jie et al., 2015; Lei et al., 2018; Liu et al., 2016b; Nie et al., 2010; Nie et al., 2016). For example, by combining MRI and positron emission tomography (PET) data, Jie et al. (2015) developed the manifold regularized multi-task feature selection model (M2TFS) to identify discriminative AD-related features. As the manifold regularizer involves graph Laplacians, it is crucial to learn the similarity matrices. However, Jie et al. (2015) generated them in each modality via 0-1 weighting before feature learning. Therefore, the similarity matrices are untrustworthy. In addition, Liu et al. (2016b) proposed the relationship induced multi-template learning (RIML) for automatic diagnosis of AD and MCI by modeling the template-template and subject-subject relations. However, similar to M2TFS, it also constructed similarity matrices by the conventional method (heat kernel weight) before feature learning, in which one connected component was involved. In another work, Lei et al. (2018) proposed an adaptive ensemble manifold learning (AEML) for neuroimaging retrieval. However, it kept different features of multi-source data with unequal informative powers via a series of non-negative weights. As mentioned previously, these methods have at least two limitations when constructing the similarity matrix, i.e., untrustworthy similarity matrix and inappropriate neighbor allocation. These shortcomings lead to the untrustworthy similarity matrix and eventually a suboptimal result. Moreover, the combination of multiple graphs is unsatisfactory in these works.

3. Dataset and preprocessing

3.1. Subjects

We conduct a series of experiments on the ABIDE[1] database, which comprise 17 international imaging centers (sites) and involves 1112 ASD patients and NC. In our experiment, we choose to assess the diagnosis performances of our proposed method utilizing the rs-fMRI imaging data obtained from four imaging sites, that is, New York University (NYU) Langone Medical Center, University of California, Los Angeles: Sample 1 (UCLA_1), University of Michigan: Sample 1 (UM_1), and Yale Child Study Center (YALE). Due to their relatively large sample size, we thus believe that the result analysis and performance comparison in our experiment will be more reasonable and acceptable. The demographic information and acquisition parameters of the subjects used in our experiments are detailed in Table 1.

3.2. Data preprocessing

In our experiments, we use the Data Processing Assistant for Resting-State fMRI (DPARSF) (Chao-Gan, 2014) to preprocess the rs-fMRI image data. Specifically, for each subject, we first discard the first 10 rs-fMRI volumes. Then, slice timing correction is conducted followed by head motion correction. Next, we normalize all rs-fMRI images to the Montreal Neurological Institute (MNI) space with the resolution of 3 × 3 × 3 mm^3.We further perform the Nuisance variable regression. The rs-fMRI images are partitioned into 116, 160, and 200 ROIs according to different pre-defined templates (i.e., AAL (Tzourio-Mazoyer et al., 2002), Dosenbach 160 (Dos 160) (Dosenbach et al., 2010), and Craddock 200 (CC 200) (Craddock et al., 2012). For each ROI, band-pass filtering of 0.005–0.1 Hz is utilized. Scrubbing is then conducted to the filtered time points. The volumes equal to or greater than 0.5 mm displacement are removed. Moreover, the volumes with excessive motion are also removed. Noted that we exclude subjects with fewer than 3 volumes remained after scrubbing, Fig. 1.

3.3. Notations

In this paper, boldface uppercase letters are used to denote matrices, boldface lowercase letters denote vectors, and normal italic letters denote scalars, respectively. For a matrix $\mathbf{X} = [\mathrm{x}_{i,j}]$, we use $\mathbf{x}^i$ and $\mathbf{x}_j$ to denote its i-th row and j-th column, respectively. Also, we use $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$, $\|\mathbf{X}\|_1 = \sum_i |\mathbf{x}^i|$, and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$ to denote Frobenius norm, ℓ_1 and $\ell_{2,\,1}$-norm of a matrix $\mathbf{X}$, respectively. Furthermore, we denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $Tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively. The other notations used in this article are summarized in Table 2.

3.4. FC network construction

For the FC network construction, the proposed method of PSLR explicitly considers the prior structure of brain connectivity pattern. Fig. 2 shows the schematic overview of network construction. By warping a certain pre-defined template to the former rs-fMRI space, the whole brain can be divided into r ROIs, from which we can extract average time series. Supposing that $\mathbf{t}_i \in \mathbb{R}^t$ is time series of i-th ROI, the series of the whole brain of a subject can be denoted as $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \cdots \mathbf{t}_r] \in \mathbb{R}^{t \times r}$. To extract biomarkers, we need to construct the FC network and the simplest way to get it is PC,

[1] See http://fcon_1000.projects.nitrc.org/indi/abide/ for specific information

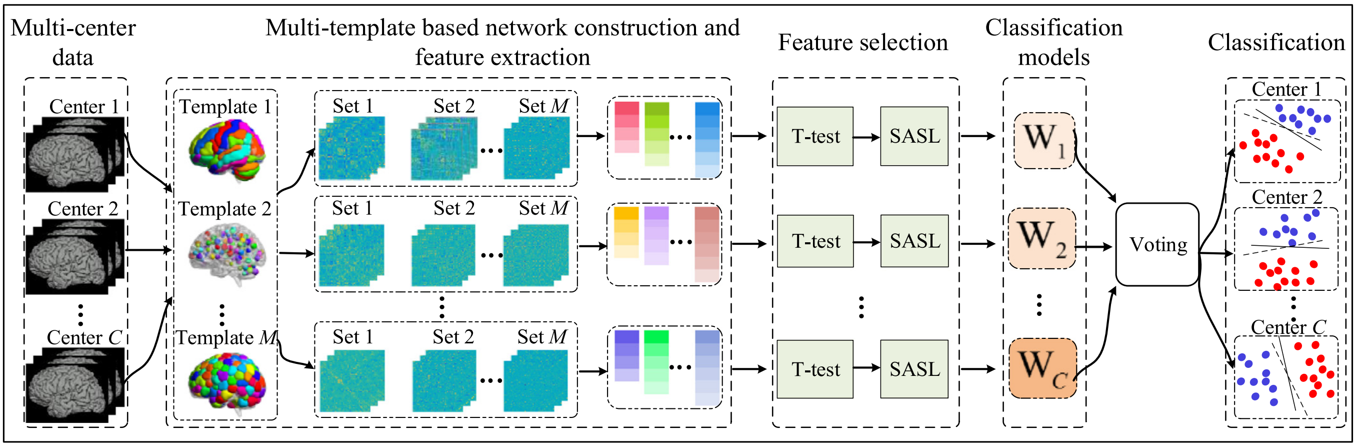


Fig. 1. Overview of the proposed framework for ASD classification. First, we build multiple FC networks for each subject based on different pre-defined templates and extract FC feature representations from them to generate feature matrces. Second, We perform a self-weighted adaptive structure learning method to obtain the corresponding classification models. Finally, we use an esemble strategy to classify the samples into different groups.

Table 2
The notations in this paper.

Notation	Size	Description
t		The number of time series
r		The number of ROIs
m		The m-th template
N		The number of training subjects
d		Feature dimension
M		The number of templates
C		The number of imaging centers
c		The number of classes
θ_k		The step size of gradient descent
λ_1, λ_2, γ, μ_m, α, β, λ		Regularization parameters
σ		Scale parameter in heat kernel weighting
$\mathbf{T}$	$t \times r$	Time series
$\mathbf{A}$	$r \times r$	Functional brain network
$\mathbf{X}^m$	$N \times d$	FC feature matrix in the m-th template
$\mathbf{W}$	$d \times M$	Weight coefficients matrix
$\mathbf{y}$	$N \times 1$	Label vector
$\mathbf{S}$	$N \times N$	Similarity matrix
$\mathbf{D}$	$N \times N$	Degree matrix
$\mathbf{L}_s$	$N \times N$	Laplacian matrix
$\mathbf{F}$	$N \times c$	Class indicator matrix
$\mathbf{U}$	$r \times r$	Unitary matrices $\mathbf{U}$ and $\mathbf{V}$ produced by singular value decomposition
$\mathbf{V}$	$r \times r$	

which is defined as

$$a_{ij} = \frac{\left(\mathbf{t}_i - \bar{\mathbf{t}}_i\right)^T \left(\mathbf{t}_j - \bar{\mathbf{t}}_\mathbf{j}\right)}{\sqrt{\left(\mathbf{t}_i - \bar{\mathbf{t}}_i\right)^T \left(\mathbf{t}_i - \bar{\mathbf{t}}_i\right)}\sqrt{\left(\mathbf{t}_j - \bar{\mathbf{t}}_j\right)^T \left(\mathbf{t}_j - \bar{\mathbf{t}}_\mathbf{j}\right)}}, \tag{1}$$

where a_{ij} denotes the FC between i-th ROI and j-th ROI, all entries in vector $\bar{\mathbf{t}}_i$ are the mean of vector $\mathbf{t}_i$. Supposing $\mathbf{t}_i$ is centralized by $\mathbf{t}_i - \bar{\mathbf{t}}_i$ and further normalized by $\sqrt{(\mathbf{t}_i - \bar{\mathbf{t}}_i)^T(\mathbf{t}_i - \bar{\mathbf{t}}_i)}$, Eq. (1) is simplified as $a_{ij} = \mathbf{t}_i{}^T\mathbf{t}_j$ or $\mathbf{A} = \mathbf{T}^T\mathbf{T}$. It is the solution of the following regression problem

$$\min_{\mathbf{A}} \sum\nolimits_{i,j=1}^{r} \left\|\mathbf{t}_i - a_{ij}\mathbf{t}_j\right\|^2, \tag{2}$$

where $\mathbf{A}$ represents the edge weight matrix of a FC network. The matrix form of Eq. (2) is

$$\min_{\mathbf{A}} \left\|\mathbf{A} - \mathbf{T}^T\mathbf{T}\right\|_F^2. \tag{3}$$

Since the brain is organized with modular structure (Sporns, 2011), we can construct a novel FC network by incorporating modularity prior on the basis of PC scheme, which is defined as

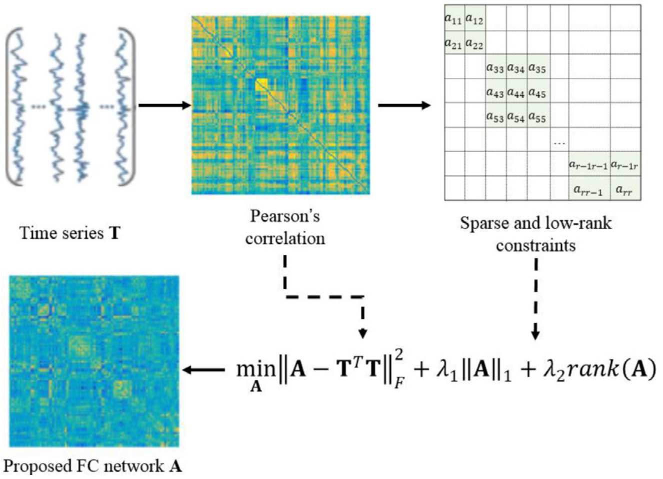

a_{11} a_{12}
a_{21} a_{22}
a_{33} a_{34} a_{35}
a_{43} a_{44} a_{45}
a_{53} a_{54} a_{55}
...
a_{r-1r-1} a_{r-1r}
a_{rr-1} a_{rr}
Time series $\mathbf{T}$
Pearson's correlation
Sparse and low-rank constraints
$\min_{\mathbf{A}} \|\mathbf{A} - \mathbf{T}^T\mathbf{T}\|_F^2 + \lambda_1\|\mathbf{A}\|_1 + \lambda_2 rank(\mathbf{A})$
Proposed FC network $\mathbf{A}$

Fig. 2. Schematic overview of the proposed PSLR framework for functional brain network construction. Given the whole brain Time series X, we encode modularity prior by adding sparse and low-rank constants on the basis of PC scheme.

$$\min_{\mathbf{A}} \left\|\mathbf{A} - \mathbf{T}^T\mathbf{T}\right\|_F^2 + \lambda_1\|\mathbf{A}\|_1 + \lambda_2 rank(\mathbf{A}), \tag{4}$$

where $\|\mathbf{A} - \mathbf{T}^T\mathbf{T}\|_F^2$ is used to fit the input data, $\|\mathbf{A}\|_1$ is used for ensuring sparsity of the edge weight matrix, and rank($\mathbf{A}$) is the rank function of $\mathbf{A}$. The above sparse regularizer and rank constraint together can model the modularity structure using λ_1 and λ_2 as control parameters. However, the rank function is non-convex. To tackle it, we relax it to trace norm$\|\mathbf{A}\|_*$. Thus, the objective function then becomes

$$\min_{\mathbf{A}} \left\|\mathbf{A} - \mathbf{T}^T\mathbf{T}\right\|_F^2 + \lambda_1\|\mathbf{A}\|_1 + \lambda_2\|\mathbf{A}\|_*. \tag{5}$$

Here, we argue that the proposed method has two characteristics: (1) Compared to the SR-based method involving the ill-posed issue, it is statistically more robust; (2) Compared to the PC, it introduces the biological prior and makes the constructed networks have physiological significance.

3.5. Self-weighted adaptive structure learning

Each sample can get corresponding number of FC networks of different sizes according to multiple different templates via Eq. (5).

For each functional connectivity weight matrix $\mathbf{A}$, we consider each entry (i.e., FC between two ROIs) in it as a feature. Since the matrix is symmetrical (diagonal values were always set to zeroes), we only need to accept the elements above diagonal values. Then we reshape these features into a vector. After this processing performed for each subject, we can obtain multiple FC feature matrices corresponding to FC networks of different sizes (corresponding to different templates). Prior to training, we conduct a feature selection process for the following reasons. First, the number of features increases with the square of the number of ROIs, from a 116-ROIs network to a 200-ROIs network, which increases the feature dimensions substantially. Since not all of these features are relevant to ASD diagnosis, feature selection can reduce feature dimension and boost the generalization performance. Second, the computation time will be significantly reduced after feature selection. Here, we adopt a t-test to select preliminary sets of most relevant features, which greatly mitigate the computation burden for the subsequent steps.

3.5.1. *Multi-task sparse learning*

Supposing that there are M templates and thus M supervised learning tasks. Denote $\mathbf{X}^m = [\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_N^m]^T \in \mathbb{R}^{N\times d}$ as the FC feature matrix for the mth template (i.e., the m-th task), where $\mathbf{x}_i^m \in \mathbb{R}^d$ is the feature vector of ith subject of the m-th template, d is the feature dimension and N is the number of training subjects. In a similar way, we denote $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^N$ as the associated label vector, where $y_i \in \{-1, +1\}$ is the associated class label (i.e., NC or ASD patient) of ith subject. Obviously, different templates from the same subject have the same class label. By feature-to-label mapping with $\mathbf{w}^m \in \mathbb{R}^d$ for mth task, the linear multi-task learning model can be expressed as

$$\min_{\mathbf{W}} \frac{1}{2}\sum\nolimits_{m=1}^{M} \|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{2,1}, \tag{6}$$

where $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^M] \in \mathbb{R}^{d\times M}$ is the weight coefficient matrix combined by the weight vectors of M templates. $\|\mathbf{W}\|_{2,1}$ is applied to penalize all entries in the same row of $\mathbf{W}$, through which a small set of common features shared across different tasks can be selected. The sparsity control parameter γ is applied to adjust the relative contributions of the two items in Eq. (6). A larger γ leads to select fewer features, while a smaller γ causes to select more features for classification.

Note that the different feature sets for each training sample are derived from different templates and therefore correspond to different ROI pairs. Hence, the $\ell_{2,1}$-norm regularizer in Eq. (6), which guides to select the common features across different tasks, is inappropriate for our case. To force sparsity of matrix $\mathbf{W}$ and identify the most discriminative features corresponding to each template, we rewrite the above linear multi-task feature learning model as

$$\min_{\mathbf{W}} \frac{1}{2}\sum\nolimits_{m=1}^{M} \|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1}, \tag{7}$$

where ℓ_1-norm is imposed to control sparsity of matrix $\mathbf{W}$. Different from $\ell_{2,1}$-norm which forces some rows of $\mathbf{W}$ to be zeros, $\ell_{1,1}$-norm forces some entries of $\mathbf{W}$ to be zero. It helps identify informative features specific to different templates.

3.5.2. *Self-weighted adaptive structure learning*

Let $\mathbf{S} \in \mathbb{R}^{N\times N}$ denote similarity matrix, and s_{ij} denote the $(i\ j)$-entry of $\mathbf{S}$. We assume that each sample $\mathbf{x}_i$ is linked to all other samples by probability s_{ij}, and such probability can be seen as the similarity between them. Since the distance between subjects is inversely related to their similarity, we can obtain the similarity matrix by solving the following model:

$$\begin{aligned}&\min_{\mathbf{S}} \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} \left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,\ s.t.\\ &\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1,\end{aligned} \tag{8}$$

where $\mathbf{s}_i$ is the similarity vector for the i-th subject, α is the turning parameter, $\alpha\|\mathbf{S}\|_F^2$ is imposed to avoid the trivial solution, i.e., the similarity between the nearest samples is 1.

However, the similarity matrix $\mathbf{S}$ obtained by Eq. (8) only contains one connected component. To ensure that the similarity matrix has the proper number of connected components, we first calculate the corresponding Laplacian matrix $\mathbf{L}_s$ ($\mathbf{L}_s = \mathbf{D} - (\mathbf{S}^T + \mathbf{S})/2$), where $\mathbf{D}$ ($d_{ii} = \sum_j (s_{ij} + s_{ji})/2$) is the degree matrix. According to the property of the Laplacian matrix, we can get that the similarity matrix has c connected components when $\text{rank}(\mathbf{L}_s) = n - c$, c is the number of label category. Accordingly, we add the rank constraint into Eq. (8) as follows:

$$\begin{aligned}&\min_{\mathbf{S}} \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} \left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,\ s.t.\\ &\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \text{rank}(\mathbf{L}_s) = n - c.\end{aligned} \tag{9}$$

It adaptively allocates neighborhoods for each data point, that is, the similarity between data points will change. Accordingly, $\mathbf{S}$ is updated until it contains the exactly connected component.

We now extend the above adaptive similarity learning to the multi-template case. Since the data derived from different templates are collected from the same subject, we deem that these data should follow the same intrinsic distribution. Therefore, the similarities of these subjects in diverse templates should be identical. Accordingly, we obtain the shared similarity matrix by solving the following model:

$$\begin{aligned}&\min_{\mathbf{S},\mu_m} \sum\nolimits_{m=1}^{M} (\mu_m)^{\varphi} \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} \left\|\mathbf{x}_i^m-\mathbf{x}_j^m\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,\\ &s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \text{rank}(\mathbf{L}_s) = n - c, 0 \le \mu_m \le 1,\\ &\mu_m^T 1 = 1.\end{aligned} \tag{10}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots \mu_M]$ is the weight vector for each template, φ is the non-negative scalar, which is applied to keep the distribution of weights smooth.

It is known that if the features of two subjects are very similar, the distance between them in the label space should be also small. In the above multi-task sparse learning model, we map the data from the initial feature space to response space via a linear mapping function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$. Although it preserves the relationship between labels and subjects, these interdependencies between samples are neglected. To address this shortcoming and improve the performance of our model, we introduce a new regularization item

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{S},\mu_m} \sum\nolimits_{m=1}^{M} (\mu_m)^{\varphi} \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} \left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,\\ &s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \text{rank}(\mathbf{L}_s) = n - c, 0 \le \mu_m \le 1,\\ &\mu_m^T 1 = 1.\end{aligned} \tag{11}$$

with the above regularizer, we can keep the local neighboring structures of the same-category data during projecting.

By incorporating the above regularization terms, the following objective function can be obtained

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{S},\mu_m} \frac{1}{2}\sum\nolimits_{m=1}^{M} \|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1}\\ &+\beta\sum\nolimits_{m=1}^{M} (\mu_m)^{\varphi} \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} \left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,\\ &s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \text{rank}(\mathbf{L}_s) = n - c, 0 \le \mu_m \le 1,\\ &\mu_m^T 1 = 1.\end{aligned} \tag{12}$$

where β denotes the positive parameter.

To remove the undesirable parameter φ, we develop a novel self-weighted multi-template learning method with adaptive neighbors, which is defined as

$$\min_{\mathbf{W},\mathbf{S}} \frac{1}{2}\sum_{m=1}^{M} \|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1}$$

$$+\beta\sum_{m=1}^{M}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij}} + \alpha\|\mathbf{S}\|_F^2,$$

$$s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \mathrm{rank}(\mathbf{L}_s) = n - c. \tag{13}$$

the weight of m-th template can be represent as $\frac{1}{2\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\|(\mathbf{w}^m)^T\mathbf{x}_i{}^m-(\mathbf{w}^m)^{\boldsymbol{T}}\mathbf{x}_j{}^m\|_2^2 s_{ij}}}$. We will introduce its solution process in the optimization section.

3.6. *Ensemble learning*

On the one hand, to better utilize multiple sets of informative features derived from different templates, we adopt an ensemble strategy. Specifically, after feature selection, we obtain M subsets of feature corresponding to M templates. We can then develop M classifiers separately via linear function $\mathbf{y} = \mathbf{X}^m\mathbf{w}^m$. Then, we employ the majority voting strategy to balance outputs of M different classifiers for the class label of a new testing subject. On the other hand, we can get four weight matrices corresponding to four datasets, after performing SASL method. Each weight matrix is just like a doctor. Furthermore, to take advantage of these weight matrices, we also perform a voting strategy to get the final classification results for each center.

4. Optimization

4.1. *The optimization of FC network*

Because of the existence of ℓ_1-norm regularizer and trace norm regularizer, the objective function in Eq. (5) is convex yet non-differentiable. To deal with this problem, we employ the proximal method (Combettes and Pesquet, 2011) for its efficiency and simplicity. Specifically, we first process the data-fitting term $f(\mathbf{T},\mathbf{A}) = \|\mathbf{A}-\mathbf{T}^T\mathbf{T}\|_F^2$. Because its gradient is $\nabla f(\mathbf{T},\mathbf{A}) = 2(\mathbf{A}-\mathbf{T}^T\mathbf{T})$, the gradient descent step is

$$\mathbf{A}_k = \mathbf{A}_{k-1} - \theta_k f(\mathbf{T},\mathbf{A}_{k-1}), \tag{14}$$

where θ_k stands for step size.

The proximal operator of $\lambda_1\|\mathbf{A}\|_1$ can be expressed as the following soft thresholding operation on $\mathbf{A}$:

$$prox_{\lambda_1\|\cdot\|_1}(\mathbf{A}) = \left[sgn\left(a_{ij}\right)\times max\left\{\left|a_{ij}\right|-\lambda_1, 0\right\}\right]_{r\times r}. \tag{15}$$

In a similar way, the proximal operator of $\lambda_2\|\mathbf{A}\|_*$ can be expressed as the following shrinkage operation on the singular values of $\mathbf{A}$:

$$prox_{\lambda_2\|\cdot\|_*}(\mathbf{A}) = \mathbf{U}diag(max\{\sigma_1-\lambda_2, 0\},\cdots, max\{\sigma_r-\lambda_2, 0\})\mathbf{V}^T, \tag{16}$$

where $\mathbf{U}diag(\sigma_1, \cdots, \sigma_r)\mathbf{V}^T$ is the singular value decomposition of FC matrix $\mathbf{A}$.

To prevent the present $\mathbf{A}_k$ falling outside of the "feasible region" regularized by $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_*$, we orderly force proximal operation on $\mathbf{A}_k$ by $prox_{\lambda_1\|\cdot\|_1}(\mathbf{A})$ and $prox_{\lambda_1\|\cdot\|_*}(\mathbf{A})$ given in Eqs. (15) and (16), respectively. The algorithm to solve Eq. (5) is summarized in Algorithm 1.

Algorithm 1 Proposed FC network.

```
Input: T
Output: A
Initialize: A;
While not converge
   1. $\mathbf{A} \leftarrow \mathbf{A} - \theta(2\mathbf{A} - 2\mathbf{T}^T\mathbf{T})$;
   2. $\mathbf{A} \leftarrow prox_{\lambda_1\|\cdot\|_1}(\mathbf{A}) = [sgn(a_{ij})\times max\{|a_{ij}|-\lambda_1, 0\}]_{r\times r}$
   3. $\mathbf{A} \leftarrow prox_{\lambda_2\|\cdot\|_1}(\mathbf{A}) = \mathbf{U}diag(max\{\sigma_1-\lambda_2, 0\},\cdots, max\{\sigma_r-\lambda_2, 0\})\mathbf{V}^T$
End
```

4.2. *The optimization of SASL model*

Because of the rank constraint $\mathrm{rank}(\mathbf{L}_s) = n - c$ relies on similarity matrix $\mathbf{S}$, problem (13) is difficult to address. To handle this issue, let $\sigma_i(\mathbf{L}_s)$ represent the ith smallest eigenvalue of $\mathbf{L}_s$. Then we can easily get $\sigma_i(\mathbf{L}_s) \ge 0$ due to the positive semi-definite nature of $\mathbf{L}_s$. Thus, $\mathrm{rank}(\mathbf{L}_s) = n - c$ is eequivalent to $\sum_{i=1}^{c}\sigma_i(\mathbf{L}_s) = 0$. In the light of Ky Fan's Theorem (Fan, 1949), we have

$$\sum_{i=1}^{c}\sigma_i(\mathbf{L}_s) = \min_{\mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} Tr\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right), \tag{17}$$

where $\mathbf{F} = [f_1, \ldots f_N]$ is the class indicator matrix. Therefore, Eq. (13) can be rewritten as

$$\min_{\mathbf{W},\mathbf{S},\mathbf{F}} \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1} + 2\lambda Tr\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right)$$

$$+\beta\sum_{m=1}^{M}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij}} + \alpha\|\mathbf{S}\|_F^2,$$

$$s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{18}$$

Obviously, when parameter λ is large enough, $Tr(\mathbf{F}^T\mathbf{L}_S\mathbf{F})$ will infinitely approach to zero and thus $\sum_{i=1}^{c}\sigma_i(\mathbf{L}_s) = 0$ holds.

To address the optimization problem of Eq. (18), we compute the derivatives of the square root, and obtain the following formulations:

$$\min_{\mathbf{W},\mathbf{S},\mathbf{F}} \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1} + 2\lambda Tr\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right)$$

$$+\beta\sum_{m=1}^{M}\mu_m\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij} + \alpha\|\mathbf{S}\|_F^2,$$

$$s.t.\forall i,\ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1, \mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \tag{19}$$

and

$$\mu_m = \frac{1}{2\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|(\mathbf{w}^m)^T\mathbf{x}_i{}^m - (\mathbf{w}^m)^{\boldsymbol{T}}\mathbf{x}_j{}^m\right\|_2^2 s_{ij}}}, \tag{20}$$

where the value of μ_m is dependent on the variables $\mathbf{S}$ and $\mathbf{W}$. If μ_m is stationary, Eq. (19) can be optimized by sequentially computing the derivative for the variables $\mathbf{W}$ and $\mathbf{S}$. Therefore, we adopt an alternative way to optimize Eq. (18).

4.2.1. *Fix **S**, **F** and μ_m, Update **W***

By fixing $\mathbf{S}$, $\mathbf{F}$ and μ_m, Eq. (19) can be transformed into the following problem

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1}$$

$$+\beta\sum_{m=1}^{M}\mu_m\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|(\mathbf{w}^m)^T\left(\mathbf{x}_i^m-\mathbf{x}_j^m\right)\right\|_2^2 s_{ij}. \tag{21}$$

According to an equation in spectral analysis, $\sum_{i,j}\|f_i - f_j\|_2^2 s_{ij} = 2Tr(\mathbf{F}^T\mathbf{L}_s\mathbf{F})$, we can rewrite Eq. (21) as

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{y}-\mathbf{X}^m\mathbf{w}^m\|_2^2 + \gamma\|\mathbf{W}\|_{1,1}$$

$$+\beta\sum_{m=1}^{M}\mu_m(\mathbf{X}^m\mathbf{w}^m)^T\mathbf{L}_s(\mathbf{X}^m\mathbf{w}^m), \tag{22}$$

$\mathbf{W}$ can be optimized by referring to tool box MALSAR1.1 (Zhou et al., 2011a).

Algorithm 2 Self-weighted adaptive structure learning

```
Input: Multi-template data $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, \cdots \mathbf{X}^M\}$, $\mathbf{X}^m \in \mathbb{R}^{N\times d}$, response vector $\mathbf{y}$, number of classes $c$, parameters $\gamma$ and $\beta$.
Output: Weight coefficient matrix $\mathbf{W}$
Initialize: The weight for each template $\mu_m = \frac{1}{M}$, then each row $\mathbf{s}_i$ of $\mathbf{s}$ can be initialized by solving the following problem
$\min_{\mathbf{s}_i^T 1=1, 0\le s_{ij}\le 1} \sum_{j=1}^{N} (\frac{1}{\mu_m} \sum_{m=1}^{M} \|\mathbf{x}_i^m - \mathbf{x}_j^m\|_2^2 s_{ij} + \alpha s_{ij}^2)$,
while not converge
Update $\mu_m$ by using Eq. (20);
Update $\mathbf{W}$ by solving problem (22);
Calculate $\mathbf{L}_s = \mathbf{D} - (\mathbf{S}^T + \mathbf{S})/2$, where the degree matrix $\mathbf{D}$ is a diagonal matrix and the $i$-th element is defined as $d_{ii} = \sum_j (s_{ij} + s_{ji})/2$.
Update $\mathbf{F}$ via solving problem (23), the optimal solution $\mathbf{F}$ is formed by $c$ eigenvectors specific to the $c$ smallest eigenvalues of $\mathbf{L}_S$.
Update $\mathbf{S}$, each $\mathbf{s}_i$ is calculate by solving problem (27) individually.
end
```

4.2.2. Fix ***S*** and ***W***, Update ***F*** and μ_m

When **S** and **W** are fixed, we can easily calculate the value of μ_m by Eq. (20). Eq. (19) can be reformulated as

$$\min_{\mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} Tr\left(\mathbf{F}^T \mathbf{L}_S \mathbf{F}\right). \tag{23}$$

The optimal solution **F** is composed by the c eigenvectors specific to the c smallest eigenvalues of $\mathbf{L}_S$.

4.2.3. Fix ***W, F*** and μ_m, Update ***S***

When **W, F,** and μ_m are fixed, Eq. (19) can be reformulated as:

$$\min_{\mathbf{S}} \beta \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{N} \mu_m \left\| (\mathbf{w}^m)^T \left(\mathbf{x}_i^m - \mathbf{x}_j^m\right) \right\|_2^2 s_{ij}$$
$$+\alpha \|\mathbf{S}\|_F^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| f_i - f_j \right\|_2^2 s_{ij},$$
$$s.t. \forall i, \ \ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1. \tag{24}$$

For convenience, we define $d_{ij}^x = \sum_{m=1}^{M} \mu_m \| (\mathbf{w}^m)^T (\mathbf{x}_i^m - \mathbf{x}_j^m) \|_2^2$ and $d_{ij}^f = \| f_i - f_j \|_2^2$. Then we have

$$\min_{\mathbf{S}} \beta \sum_{i=1}^{N} \sum_{j=1}^{N} \left(d_{ij}^x s_{ij} + \alpha s_{ij}^2 + \lambda d_{ij}^f s_{ij} \right)$$
$$s.t. \forall i, \ \ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1. \tag{25}$$

For each sample, it is noteworthy that the similarity vector is independent, thus we can address the following problem individually for each sample

$$\min_{\mathbf{s}_i} \sum_{j=1}^{N} \left(\beta d_{ij}^x s_{ij} + \alpha s_{ij}^2 + \lambda d_{ij}^f s_{ij} \right) \ s.t. \forall i, \ \ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1. \tag{26}$$

We denote vector $\mathbf{d}_i \in \mathbb{R}^d$ with $d_{ij} = \beta d_{ij}^x + \lambda d_{ij}^f$. Eq. (24) is equivalent to

$$\min_{\mathbf{s}_i} \left\| \mathbf{s}_i + \frac{1}{2\alpha} \mathbf{d}_i \right\|_2^2 \ s.t. \forall i, \ \ \mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1. \tag{27}$$

The value of parameter α can be determined via the number of adaptive neighbors. The corresponding algorithm can be acquired in the literature (Nie et al., 2017). Adaptive neighbors mean that the k nearest neighbors to any sample $\mathbf{x}_i$ is changed in every iteration since the weighted distance d_{ij} is updated in every iteration. The above algorithm is summarized in Algorithm 2.

In Eq. (19), when Algorithm 2 converges, the form of Laplacian regularizer can be seen as the linear combination of manifold in terms of multiple templates, and μ_m is the corresponding weight. As Eq. (19) is derived from Eq. (13), we can confirm that it is entirely a self-weighted process. Moreover, according to Eq. (19), if the m-th template is more suited for diagnosis, the associated weight μ_m will be larger. By contrast, a weaker template will be assigned a smaller weight. This implies that our self-weighted multi-template learning model is effective.

5. Experiments and results

5.1. Experimental setup

In our experiments, the ASD identification problem is treated as a binary classification problem. We tag normal controls as "−1" and ASD patients as "+1". A 10-fold cross-validation is adopted to evaluate the performance of the proposed method. Particularly, the subjects of each class from each imaging center are partitioned into ten disjoint folds randomly. Nine subsets are then picked up from each imaging center to generate the training set. The remaining subset is utilized for testing. Moreover, to avoid the biased result entailed by the fold selection, we repeat the 10-fold cross-validation process ten times and report the average of the results.

5.2. Evaluation metrics

To estimate the diagnosis performances of all the methods mentioned in this article quantitatively, we utilize the metrics of accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic (ROC) curve (AUC). Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively, then we can define the ACC, SEN, SPE, PPV, and NPV as:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{28}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{29}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{30}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{31}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{32}$$

5.3. Comparison methods

To verify the efficacy of our proposed model based on multi-center multi-template feature representations, we conduct a series of experiments. Table 3 sums up the approaches under comparisons. In the comparisons with state-of-the-art methods section, we report the results of our proposed SASL-E and comparison methods. Specifically, to show the superiority of using multi-center representation through an ensemble strategy, we consider the following two strategies: 1) single-center- based method, denoted as -S, 2) our proposed ensemble-based method, denoted as -E. More specifically, in the single-center- based method, the data collected from each center are treated separately. In the ensemble-based

Table 3
Summary of the methods for comparison.

Method	Description	
CSVC	C-SVC in the library for support vector machines (LIBSVM). For this method, linear kernel is applied.	
LeastL	SASL with $\alpha = 0, \beta = 0$, such that local manifold regularization is disabled.	
Binary	The manifold regularizer is combined in our multi-task framework. The similarity matrix in it is obtained via 0-1 weighting, i.e., if the sample $\mathbf{x}_i^m$ and sample $\mathbf{x}_j^m$ in the m-th template have the same label, then the similarity $\mathbf{s}_{ij}^m$ between them is set to 1, otherwise 0.	
kNN	The manifold regularizer is combined in our multi-task framework. The similarity matrix in it is obtained via heat kernel weighting. $s_{ij}^m = e^{\frac{\|\mathbf{x}_i^m - \mathbf{x}_j^{m2}\|}{\sigma}}$, σ is a scale parameter.	
LogisticL	The multiple task learning algorithm in the package of MALSAR1.1 with the function Logistic_Lasso().	
SASL	The proposed SASL method. For each template, 100 inter-region functional connection features are used.	
116 ROIs	Run SASL with feature representations derived from AAL.	
160 ROIs	Run SASL with feature representations derived from Dos 160.	
200 ROIs	Run SASL with feature representations derived from CC 200.	
PC	$\|\mathbf{A} - \mathbf{T}^{\mathbf{T}}\mathbf{T}\|_{\mathrm{F}}^2$	
SR	$\|\mathbf{T} - \mathbf{TA}\|_{\mathrm{F}}^2$	$\lambda_1\|\mathbf{A}\|_1$
SLR	$\|\mathbf{T} - \mathbf{TA}\|_{\mathrm{F}}^2$	$\lambda_1\|\mathbf{A}\|_1 + \lambda_2\|\mathbf{A}\|_*$
PSLR	$\|\mathbf{A} - \mathbf{T}^{\mathbf{T}}\mathbf{T}\|_{\mathrm{F}}^2$	$\lambda_1\|\mathbf{A}\|_1 + \lambda_2\|\mathbf{A}\|_*$

Table 4
Classification results of NYU center.

Method	ACC	SEN	SPE	PPV	NPV	AUC	p-value
CSVC	70.41±9.14	60.39±14.40	77.81±13.48	69.26±15.05	72.89±8.22	68.10±12.30	<1e-4
LeastL-S	72.40±9.14	62.54±14.70	79.78±13.53	72.09±14.68	74.48±8.52	69.98±10.57	<1e-4
LogisticL-S	73.06±7.80	60.82±16.02	82.11±11.63	74.04±13.93	74.54±7.94	71.02±11.02	<1e-4
Binary-S	74.99±9.09	64.27±17.76	83.03±11.82	75.24±15.94	76.65±9.25	73.12±12.30	<1e-4
kNN-S	71.33±9.71	68.46±16.89	73.53±14.06	67.03±14.12	76.79±10.27	67.03±13.81	<1e-4
SASL-S	77.43±8.50	64.41±17.67	87.13±10.59	80.71±14.57	77.66±9.09	76.33±11.78	<1e-4
LeastL-E	74.04±9.21	60.07±18.19	84.58±12.20	76.22±17.46	74.75±9.20	71.95±12.78	<1e-4
LogisticL-E	74.28±10.17	75.39±15.89	73.39±15.51	69.87±13.86	81.19±11.03	73.66±12.37	<1e-4
Binary -E	76.24±7.75	62.55±16.98	86.47±10.57	79.73±13.86	76.58±8.93	73.34±10.80	<1e-4
k-NN-E	70.87±8.69	66.14±17.61	74.46±13.34	67.62±12.68	75.82±10.04	67.35±12.50	<1e-4
SASL-E	77.63±8.95	65.54±19.79	86.67±11.23	80.06±15.86	78.41±10.07	76.67±11.72	

method, we apply a voting strategy for the classification models of all centers to get the final classification result of each center. To show the effectiveness of feature selection, we then compare our method with C-Support Vector Machine (CSVC) (Chang and Lin, 2012), which conduct classification directly without feature selection. In addition, one of the major contributions of this work is to incorporate the local manifold structure of each template space via self-weighted adaptive learning. Accordingly, we compare our method to the following three strategies: 1) We set our proposed SASL method by $\alpha = 0, \beta = 0$ to disable the local manifold regularization, and we denote it as LeastL; 2) We combine the manifold regularizer mentioned in (Jie et al., 2015) in our multi-task framework, where the similarity matrix is obtained via the method 0-1 weighting, We denote this method as binary; 3) Similarly, referring to (Liu et al., 2016b), we incorporate the local manifold structure of each template space, where the similarity matrix is obtained via the method of heat kernel weighting. We denote this method as kNN. We also compare the proposed method with other commonly used multi-task learning algorithms, including the function Logistic_Lasso() in MALSAR1.1 package (Zhou et al., 2011b), we denote it as LogisticL. Finally, to reflect the advantages of our method using discriminative information acquired from different centers, we also add a group of comparisons, i.e., C4-based method. In the C4-based method, the data from four centers (NYU, UCLA_1, UM_1, and YALE) are gathered together as the dataset for classification.

In the comparisons with different templates section, to investigate the contribution of using multi-template feature representations under the multi-task framework, we derive a single template multi-center classification from SASL to show the performances with respect to different inputs (feature representations are collected based on different templates: AAL, Dos 160 and CC 200). For ease of description, we denote these comparison methods as 116 ROIs, 160 ROIs, and 200 ROIs. In the comparisons of different FC networks section, to investigate the contribution of our proposed PSLR network, we compare it to other algorithms, including PC (Smith et al., 2013), SR (Lee et al., 2011) and SLR (Qiao et al., 2016). The objective functions are listed in Table 3.

5.4. Comparisons with State-of-the-Art methods

In this sub-section, we compare our proposed SASL-E method with other competing methods, the overall classification performances of each method on four imaging centers are reported in Tables 4–7. The numbers in the table marked in bold indicate the best results. From the four tables, we observe our proposed SASL-E algorithm achieves mean classification accuracies of 77.63%, 82.73%, 78.11%, and 89.13% on NYU, UCLA_1, UM_1, and YALE, respectively, which are generally superior to its rivals on the same centers. Generally, compared with single-center-based methods, corresponding ensemble-based methods can achieve better results. This observation shows the advantage of ensemble learning. CSVC-S conducts CSVC on the dataset obtained from a single center without feature selection. Comparing its performances with other single-center - based method with feature selection (LeastL-S, LogisticL-S, Binary-S, kNN-S, SASL-S), we observe that feature selection improves classification results. Different from other methods aforementioned, SASL integrates the local manifold structure of multiple templates via self-weighted adaptive structure learning and obtains better results. The experimental results verify the advantages of joint feature selection and multi-template adaptive structure learning. Con-

Table 5
Classification results of UCLA_1 center.

Method	ACC	SEN	SPE	PPV	NPV	AUC	*p*-value
CSVC	70.40±16.98	72.33±25.21	68.17±30.44	72.67±24.84	69.92±26.55	62.39±27.02	<1e-4
LeastL-S	73.47±16.40	81.67±24.56	64.33±31.34	72.18±22.32	75.77±31.38	68.83±24.77	<1e-4
LogisticL-S	73.68±16.44	88.83±19.69	58.67±31.65	71.45±19.55	77.00±34.41	69.61±23.39	<1e-4
Binary-S	77.52±17.15	96.50±10.67	58.33±31.83	73.33±17.85	84.58±33.57	75.42±22.13	<1e-4
kNN-S	69.07±17.84	76.67±23.69	61.33±30.23	70.57±21.10	69.37±30.61	60.44±24.55	<1e-4
SASL-S	80.75±15.46	88.83±16.76	72.83±27.99	80.72±18.13	85.17±13.59	77.53±22.15	<1e-3
LeastL-E	77.38±16.96	82.83±23.74	71.67±26.54	77.40±21.51	82.87±23.44	73.14±23.43	<1e-4
LogisticL-E	76.84±15.84	83.00±21.97	70.33±26.76	77.20±20.36	82.65±22.90	74.08±24.00	<1e-4
Binary-E	80.55±17.03	88.67±19.23	72.50±27.36	79.53±19.17	85.33±24.93	76.86±23.93	<1e-4
kNN-E	78.60±16.67	81.17±24.00	76.00±25.44	80.83±21.22	83.58±20.28	72.56±23.76	<1e-4
SASL-E	82.73±15.59	84.67±22.81	81.50±22.33	84.32±19.44	86.95±15.54	78.67±20.94	

Table 6
Classification results of UM_1 center.

Method	ACC	SEN	SPE	PPV	NPV	AUC	*p*-value
CSVC-S	70.99±13.45	56.42±23.33	82.70±18.66	73.68±23.17	72.34±13.63	66.79±18.39	<1e-4
LeastL-S	73.08±13.46	59.08±25.02	84.00±17.14	75.88±25.96	74.18±13.37	70.13±19.79	<1e-4
LogisticL-S	72.95±14.04	59.92±23.74	83.20±16.63	75.30±23.27	74.65±15.52	69.36±21.36	<1e-4
Binary-S	75.17±13.29	61.92±25.47	85.60±16.13	80.20±21.14	75.89±14.67	71.97±17.65	<1e-4
kNN-S	73.31±13.53	70.17±24.71	75.65±19.84	71.50±23.02	78.67±15.45	67.34±19.76	<1e-4
SASL-S	77.03±11.84	62.33±23.16	88.40±14.21	84.00±20.33	76.47±12.68	72.46±18.15	<1e-4
LeastL-E	75.02±11.81	59.17±28.34	87.40±17.03	81.67±24.45	76.07±14.60	71.45±18.72	<1e-4
LogisticL-E	74.33±12.25	52.17±24.46	91.65±13.26	80.72±28.53	72.12±11.61	72.10±16.47	<1e-4
Binary -E	76.74±14.31	63.08±25.47	87.70±15.03	81.43±24.13	76.59±14.20	72.01±21.61	<1e-3
kNN-E	75.47±12.57	60.25±28.23	87.10±17.81	77.87±28.48	76.58±14.16	70.00±17.19	<1e-4
SASL-E	78.11±12.68	68.67±27.88	85.50±16.15	82.18±20.70	80.60±15.38	75.44±18.80	

Table 7
Classification results of YALE center.

Method	ACC	SEN	SPE	PPV	NPV	AUC	*p*-value
CSVC	79.37±17.52	72.50±30.56	86.00±20.88	80.67±28.55	81.53±19.70	74.53±25.23	<1e-4
LeastL-S	84.97±14.66	82.83±27.06	86.17±21.20	84.83±24.11	88.37±18.22	82.08±22.83	<1e-4
LogisticL-S	82.70±16.63	82.33±25.82	83.50±24.10	82.58±24.93	86.68±19.47	80.42±22.59	<1e-4
Binary-S	86.20±13.86	86.00±22.69	87.00±28.89	87.00±19.40	90.25±15.45	83.75±19.24	<1e-4
kNN-S	85.58±14.95	82.00±26.66	88.83±17.74	86.00±23.12	88.47±16.45	83.00±22.15	<1e-4
SASL-S	87.52±14.42	84.83±24.28	89.83±17.87	90.42±18.02	90.18±15.70	84.22±20.61	<1e-3
LeastL-E	86.22±16.86	86.67±23.09	85.36±23.21	87.02±21.24	90.60±16.54	84.25±23.39	<1e-4
LogisticL-E	86.53±15.83	89.33±20.72	84.00±22.21	85.33±20.86	91.85±15.81	84.89±20.44	<1e-4
Binary-E	87.08±14.54	89.17±20.29	85.17±20.91	87.00±18.21	92.58±14.31	85.44±20.85	0.0029
k-NN-E	84.18±16.08	83.67±25.95	85.00±20.85	83.50±23.92	88.85±17.27	79.25±25.43	<1e-4
SASL-E	89.13±12.52	91.00±17.48	87.67±19.34	89.35±16.30	93.33±12.92	87.33±17.04	

versely, LeastL and LogisticL ignore the structure information exists in different templates. Although Binary and kNN incorporate the local structure information, the neighbor assignment of its similarity matrix is not optimal. By contrast, SASL treats feature selection and local structure learning simultaneously and the similarity matrix learnt by rank constraint has a reasonable neighbor allocation. Thus, our proposed method outperforms its rivals. We also conduct a two-sample *t*-test on the classification accuracy realized by the competing methods and our method, and the *p*-values are reported in the following corresponding tables. The resulting *p*-values further demonstrate that our method is statistically better than the comparison method with respect to ASD diagnosis.

In addition, to further verify the effectiveness of our approach, we also evaluate our SASL method and other competing methods on a multi-center dataset (C4 dataset). Table 8 reports the corresponding classification results. We find that our method is also superior to other methods in this evaluation, yet the results of each method are generally lower than that in the above four tables. The main reason is that there is data heterogeneity across different centers.

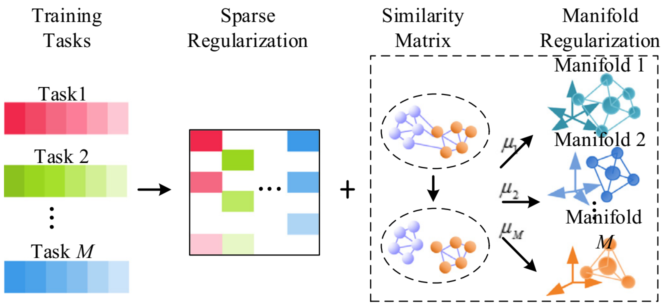


Fig. 3. Illustration of the self-weighted adaptive structure learning model.

5.5. Comparisons with different templates

In this sub-section, we conduct a series of experiments to demonstrate the advantage of using multi-template feature representations for ASD diagnosis, Fig. 3.

Table 8
Classification results of C4 dataset.

Method	ACC	SEN	SPE	PPV	NPV	AUC	*p*-value
CSVC-C4	69.01±6.68	64.13±9.90	72.92±10.08	66.48±9.06	71.76±6.07	69.24±8.24	<1e-4
LeastL-C4	70.05±6.50	63.54±11.32	75.35±9.48	68.31±9.41	72.29±6.39	70.23±8.22	<1e-4
LogisticL-C4	69.74±6.38	62.82±13.85	75.37±11.13	68.56±10.25	72.25±7.12	69.41±8.34	<1e-4
Binary-C4	70.99±6.49	63.27±10.48	77.20±9.54	69.93±9.25	72.51±5.76	70.64±7.82	<1e-4
kNN-C4	71.00±6.50	65.64±11.44	75.35±10.03	69.08±8.84	73.53±6.66	70.42±7.83	<1e-4
SASL-C4	71.77±7.90	65.94±10.43	76.44±40.88	70.31±10.83	73.71±7.00	71.81±9.09	

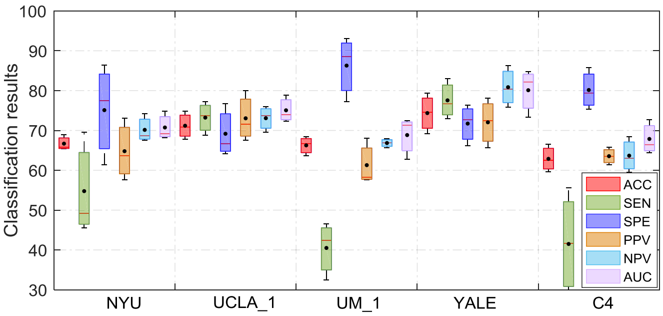


Fig. 4. Distributions of ACC, SEN, SPE, PPV, NPV, and AUC achieved on 116 ROIs, 160 ROIs, and 200 ROIs.

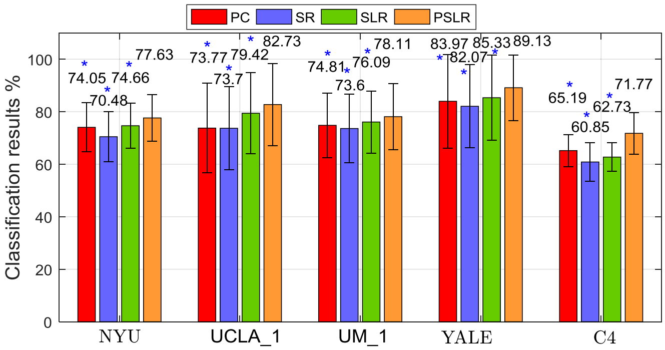


Fig. 6. Classification accuracy of different FC networks. The black lines represent standard deviation. The blue asterisks represent the *p*-values less than 1e−4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

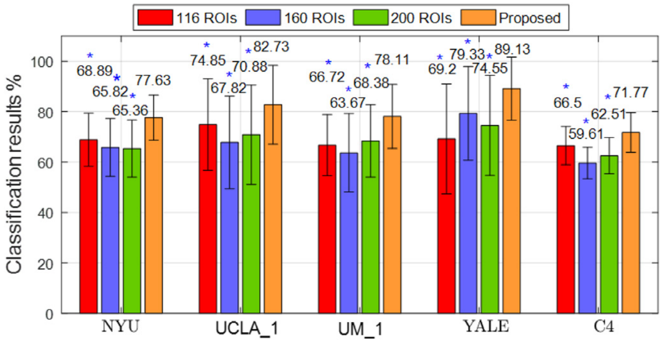


Fig. 5. Classification accuracy of different templates. The black lines represent standard deviation. The blue asterisks represent the *p*-values less than 1e−4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 draws the box plots for each dataset. We observe the classification results utilizing different single template are large divergent, in spite of different datasets. The reason for different classification results using different templates is that the disease-related features obtained from a certain template may be more discriminative than those achieved by other templates.

Fig. 5 shows the ACC values achieved via different feature inputs in different datasets. From Fig. 5, we find that the multi-template feature representations achieve consistently higher accuracy for all different datasets. For NYU, UCLA_1, UM_1 and YALE centers, our multi-template method achieves an accuracy of 77.63%, 82.73%, 78.11%, and 89.13%, respectively, while the best results on the feature sets derived from individual template is 68.89% (AAL template only), 74.85% (AAL template only), 68.38% (CC 200 template only), 79.33% (Dos 160 template only), of the corresponding centers, respectively. Similarly, our multi-template method achieves an accuracy of 71.77% of C4 dataset, while the best results of single-template methods are 66.50% (AAL template only). We use blue asterisks to represent the *p*-values less than 1e-4. These resulting *p*-values further verify that our multi-template learning is effective for ASD diagnosis, which implies different templates can provide the complementary information needed to classify the images accurately.

5.6. Comparisons with different FC networks

In this sub-section, to assess the benefits of our proposed FC network, we also apply our method on the feature representations extracted from different FC networks. As shown in Fig. 6, our proposed FC network always achieves the highest classification accuracy, which means that modeling function brain network reasonably can provide more discriminative biomarkers and thus further improve the ASD diagnosis performance. Compared to the SR network, we find the SLR network achieves the higher classification accuracy for all datasets. It indicates that the SLR incorporating the modularity prior by imposing sparse and low-rank regularization constraint can construct more accurate functional brain networks. However, compared to our proposed PSLR network, the classification performance of the SLR network is still relatively poor. The reason is that the effectiveness of data fitting term used by SLR is relatively poor in capturing the inverse structure of data in ASD classification. The resulting *p*-values further verify that our PSLR method is statistically superior to the comparison methods on the task of ASD diagnosis.

5.7. Comparisons with existing methods

In this sub-section, to demonstrate the effectiveness of our proposed method in differentiating ASD patients from NC, we also compare the performance obtained by our SASL-E method on the ABIDE database with that of several recent state-of-the-art methods reported in the literature. Since factors such as dataset size, template types and number, FC network types and classifiers vary considerably between our study and previous ones, it is unreasonable to make direct comparisons. The quantitative indicators can reflect the performance of different methods from the side. Table 9 summarizes the details of each method (i.e., Center, Template, Network, and Classifier) and its highest performance (i.e., ACC, SEN and SPE).

From Table 9, we find that the Hidden Markov Models with l_1-norm Support Vector Machines (HMM+l_1-SVM) adopted by

Table 9
Algorithm comparisons with existing studies.

Method	Center	Template	Network	Classifier	ACC(%)	SEN(%)	SPE(%)
Kam et al. (2017)	NYU+UM (263)	Single	PC	DRBM	67.42	58.33	75.00
Jun et al. (2019)	NYU+UM (292)	Single	HMM	l_1-SVM	75.86	83.33	70.59
Plitt et al. (2015)	NYU+USM+UCLA1 (178)	Single	Fisher transformed PC	l_2LR	71.35	70.33	72.41
Nielsen et al. (2013)	16 Sites[b] (964)	Single	Fisher transformed PC	leave-one-out	60.00	62.00	58.00
Chen et al. (2015)	17 Sites[a] (252)	Single	Fisher transformed PC	RFE-SVM	66.00	60.00	72.00
Abraham et al. (2017)	17 Sites[a] (871)	Single (MSDL)	Tangent embedding	SVC-l_2	66.90	78.30	53.20
Zhuang et al. (2019)	17 Sites[a] (1035)	Single	PC	InvNe	71	-	-
Dvornek et al. (2017)	17 Sites[a] (1100)	Single	-	RNN-LSTM	66.80	-	-
Heinsfeld et al. (2017)	17 Sites[a] (1005)	Single	PC	DNN	70.00	74.00	63.00
Wang et al. (2018a)	NYU+UM_1 (279)	Single	PC+HOFC	Sparse-MVTC-E	72.60	79.00	64.10
Wang et al. (2017)	NYU+UM_1+YALE+STANFORD (259)	Single	PC	M3CC	76.51	78.46	74.69
Wang et al. (2018b)	NYU+UM+USM+UCLA+Leuven(468)	Single	PC	MCLRR+KNN	66.10	70.21	66.43
Proposed	UM_1+NYU+YALE+UCLA_1 (356)	Multiple	PSLR	SASL-E	89.13	91.00	87.67

DRBM: Discriminative Restricted Boltzmann Machine; HMM: Hidden Markov Models; l_1-SVM: l_1-norm Support Vector Machines; l_2LR: l_2-regularized Logistic Regression; RFE-SVM: Recursive Feature Elimination with Support Vector Machine; MSDL: Multi-Subject Dictionary Learning; SVC-l_2: the l_2-penalized support vector classification; InvNe: Invertible Networks; RNN-LSTM: Recurrent Neural Network with Long Short-Term Memory; DNN: Deep Neural Network; HOFC: High-order Functional Connectivity; Sparse-MVTC-E: Sparse Multi-View Task-Centralized Ensemble Classification; M3CC: Multi-Modality Multi-Center Classification; MCLRR: Multi-Center Low-rank Representation Learning; kNN: k-nearest neighbor;

[a] California Institute of Technology (Caltech), Carnegie Mellon University (CMU), Kennedy Krieger Institute (KKI), Ludwig Maximilians University Munich (MaxMum), NYU Langone Medical Center (NYU), Olin, Institute of Living at Hartford Hospital (Olin), Oregon Health and Science University (OHSU), San Diego State University (SDSU), Social Brain Lab BCN NIC UMC Groningen and Netherlands Institute for Neurosciences (SBL), Stanford University (Stanford), Trinity Centre for Health Sciences (Trinity), University of California Los Angeles (UCLA), University of Leuven (Leuven), University of Michigan (UM), University of Pittsburgh School of Medicine (Pitt), University of Utah School of Medicine (USM), Yale Child Stud (Yale).

[b] Caltech, KKI, MaxMum, NYU, Olin, OHSU, SDSU, SBL, Stanford, Trinity, UCLA, Leuven, UM, Pitt, USM, Yale.

Jun et al. (2019) achieved the best results among the multi-center methods (data from multiple centers are combined directly), with an accuracy of 75.86%, a sensitivity of 83.33%, and a specificity of 70.59%. Our method achieves much better performance with a margin of 13.27% in accuracy, 7.67% in sensitivity and 17.08% in specificity. Although deep learning based methods trigger widespread attention, our method shows improvements by 22.33% and 19.13% compared to Recurrent Neural Network with Long-Short Term Memory (RNN-LSTM) (Dvornek et al., 2017) and Deep Neural Network (DNN) (Heinsfeld et al., 2017). It is worth noting that Wang et al. (2017), proposed the Multi-Modality Multi-Center Classification (M3CC) method using multi-task learning framework to address the heterogeneity of data between different centers, and achieved an accuracy of 76.51%, a sensitivity of 78.46%, and a specificity of 74.69%. By contrast, our SASL-E method improves the performance by 12.62%, 12.54% and 12.98% in accuracy, sensitivity, and specificity, respectively. For the same purpose, Wang et al. (2018b) proposed a Multi-Center Low-rank Representation Learning with the *k*-nearest neighbor (MCLRR+KNN) method, but its performance is significantly lower than our method. Additionally, to reveal the relationship of ASD with sex and age, Wang et al. (2018a) proposed a Sparse Multi-view Task-Centralized Ensemble learning (Sparse-MVTC-E) method and achieved an accuracy of 72.60%, a sensitivity of79.00%, and a specificity of 64.10%. Compared to our method, Sparse-MVTC-E shows inferior performance with a margin of 16.53% in accuracy, 12.00% in sensitivity and 23.57% in specificity.

5.8. Discriminative features

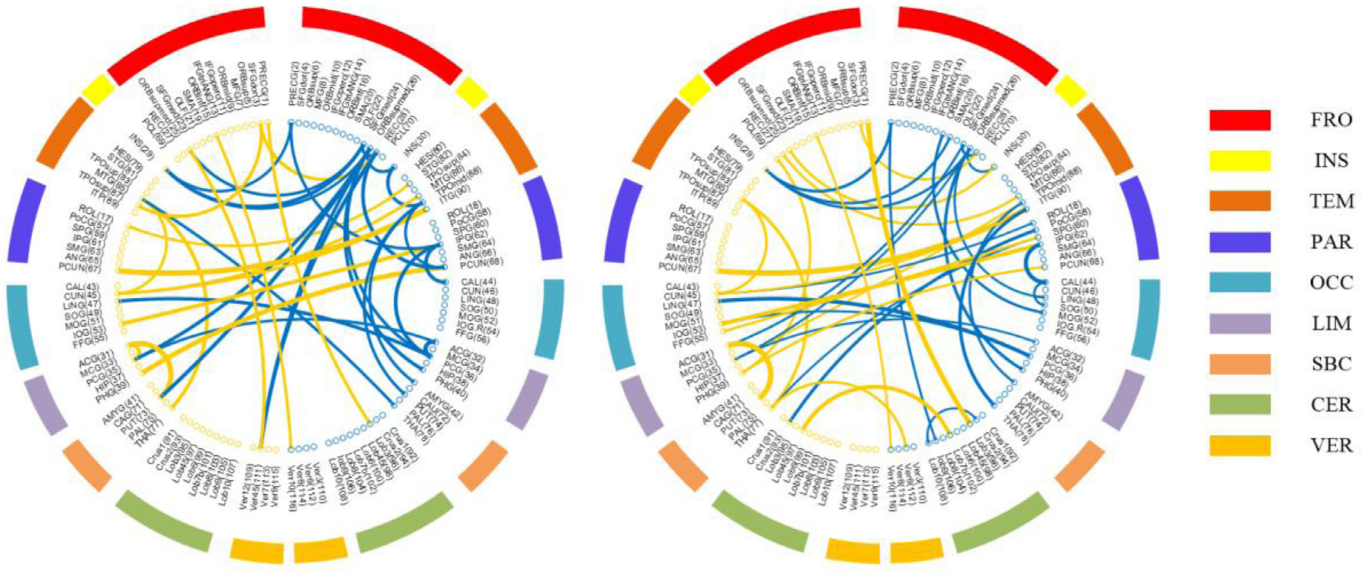
In this sub-section, the most informative FC features selected to differentiate the ASD patients from NC are reported. By analyzing the FC features determined by the proposed SASL-E method at each fold of cross-validation, although the selected features at different centers are not exactly the same, the features closely related to ASD are always selected in all centers and assigned to larger coefficients. For quantitative analysis, the top 40 commonly selected FC features over all folds of cross-validation are summarized in Table 10. The numbers in the second and third columns denote the indices that correspond to the brain regions in the AAL template. Fig. 7 visualizes the top 40 commonly selected discriminative FC features shared by multiple imaging centers. Fig. 8 further visualizes the top 5, 10, and 20 commonly selected discriminative FC features and corresponding brain regions. The thickness of each line in Figs. 7 and 8 imply the total weight of the feature across all the centers.

Using Table 10 and Figs. 7 and 8, we summarize some characteristics of the discriminative FCs regarding their hemispheric distributions and attributions. First, the connections help to accurately diagnose ASD that do not only exist in the left or right hemisphere individually but also across two hemispheres.

Second, although the number of functional connections is similar in the left hemisphere and the right hemisphere, the weights of connections in the left hemisphere are relatively larger, which is consistent with the results reported by Chandana et al. (2005). Third, the selected connections contain multiple cortical regions and subcortical structures related to ASD. For example, the left precuneus and left anterior cingulate gyrus exist in the two pairs of functional connections with the largest weight have found to be associated with ASD pathology. Urbain et al. (2015) pointed out that compared to NC children, the left precuneus of children with ASD shows more activity in the 2-back working memory task. According to Mundy (2003), impairment in the development of the anterior cingulate gyrus may lead to socio-cognitive deficits in ASD. In addition, region of the left insula was reported highly related to the emotional and communicative deficits of ASD (Urbain et al., 2016). The left superior frontal gyrus is found highly associated with social recognition ability (Andrews-Hanna et al., 2014). Our findings demonstrate that vermis 6, vermis 10 and vermis 45 in the human cerebellum show obvious contributions to ASD identification. In addition to participating in fine motor function, the cerebellum plays an important role in advanced cognitive functions including language (Hampson and Blatt, 2015). Several other discriminative subcortical regions, including the bilateral thalamus and left putamen, are also found to be abnormal in former ASD study (Cerliani et al., 2015). These observations help us understand the important biomarker information associated with ASD, and thus allowing us to further comprehend the pathology of ASD.

Table 10
Top 40 commonly selected fc features over all folds of cross-validations

	ROI 1	ROI 2		ROI 1	ROI 2
1	Precuneus_L 67	Temporal_Inf_R 90	21	Frontal_Sup_L 3	Rectus_R 28
2	Cingulum_Ant_L 31	Putamen_L 73	22	Precentral_L 1	Cerebelum_3_R 96
3	ParaHippocampal_L 39	Precuneus_R 68	23	Frontal_Mid_Orb_R 26	Cingulum_Post_R 36
4	Cingulum_Ant_R 32	Cingulum_Post_R 36	24	Frontal_Sup_Medial_R 24	Frontal_Mid_Orb_L 25
5	Occipital_Inf_L 53	Parietal_Inf_R 62	25	Cuneus_L 45	Temporal_Mid_R 86
6	Rectus_R 28	Vermis_4_5 111	26	Thalamus_R 78	Temporal_Mid_L 85
7	Frontal_Mid_Orb_R 26	Putamen_L 73	27	Frontal_Mid_Orb_R 26	Insula_R 30
8	Frontal_Sup_Medial_R 24	Insula_L 29	28	Putamen_R 74	Temporal_Pole_Mid_R 88
9	Cingulum_Post_R 36	Parietal_Inf_R 62	29	Precentral_R 2	Parietal_Inf_R 62
10	Rectus_R 28	Precuneus_R 68	30	Cuneus_L 45	Temporal_Mid_L 85
11	Thalamus_L 77	Temporal_Mid_L 85	31	Amygdala_R 42	Temporal_Pole_Mid_R 88
12	Frontal_Mid_Orb_R 26	Cingulum_Post_L 35	32	Cuneus_L 45	Temporal_Inf_R 90
13	Cingulum_Ant_L 31	Cingulum_Post_L 35	33	Frontal_Inf_Orb_L 15	Vermis_10 116
14	Cingulum_Post_R 36	Lingual_L 47	34	Precentral_R 2	Insula_L 29
15	Frontal_Mid_Orb_R 26	ParaHippocampal_R 40	35	Frontal_Sup_Medial_R 24	Rectus_R 28
16	Calcarine_R 44	Parietal_Inf_R 62	36	Precuneus_R 68	Temporal_Pole_Mid 88
17	Frontal_Sup_L 3	Temporal_Pole_Mid_L 87	37	Cingulum_Ant_R 32	Cingulum_Post_L 35
18	Frontal_Mid_Orb_R 26	Vermis_4_5 111	38	Rolandic_Oper_L 17	Temporal_Sup_R 82
19	Insula_R 30	Temporal_Pole_Mid_R 88	39	Frontal_Mid_Orb_R 26	Insula_L 29
20	Frontal_Mid_Orb_L 25	Vermis_4_5 111	40	Frontal_Sup_Medial_L 23	Cingulum_Post_R 36



(a) (b)

Fig. 7. Common discriminative rs-fMRI connections selected by our SASL method from (a) four imaging centers via ensemble strategy. (b) C4 dataset. The blue and yellow bubbles represent the right and left hemisphere region. The total weight of the common selected feature across multiple centers is shown via thickness of each line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We also report the most informative FC features selected by SASL conducted on C4 datasets. The results are also shown in Figs. 7 and 8. We find that the distribution of FC and corresponding ROIs are similar. It further indicates our SASL method can identify the most ASD-related biomarkers.

6. Discussions

6.1. FC network

The sparse and low-rank regularization constraints are able to model the modularity prior using λ_1 and λ_2 as control parameters. To analyze the effect of these two parameters, λ_1 and λ_2 in Eq. (5) are both set in the range of $\{2^{(-5)}, 2^{(-4)}, 2^{(-3)}, 2^{(-2)}, 2^{(-1)}, 1, 2^{(1)}, 2^{(2)}, 2^{(3)}, 2^{(4)}, 2^{(5)}\}$. Fig. 9 shows ACC values corresponding to different parametric combinations of the proposed PSLR method. From this figure, we find that 1) the choice of regularization parameters has an impact on the classification results, which means that the classification performances are sensitive to the regularization parameters; 2) the modularity prior can help improve classification performance. Specifically, we achieve the mean accuracy of 71.77% on C4 dataset with sparsity regularized parameters $\lambda_1 = 2^1$and low-rank regularized parameters $\lambda_2 = 2^{-5}$.

6.2. Limitations

We conducted a series of experiments on four imaging centers. Although appealing classification performance is achieved, there are still some limitations.

First, we use the sparse and low-rank constraints to estimate modular FC for each subject one by one. Nevertheless, we ignore

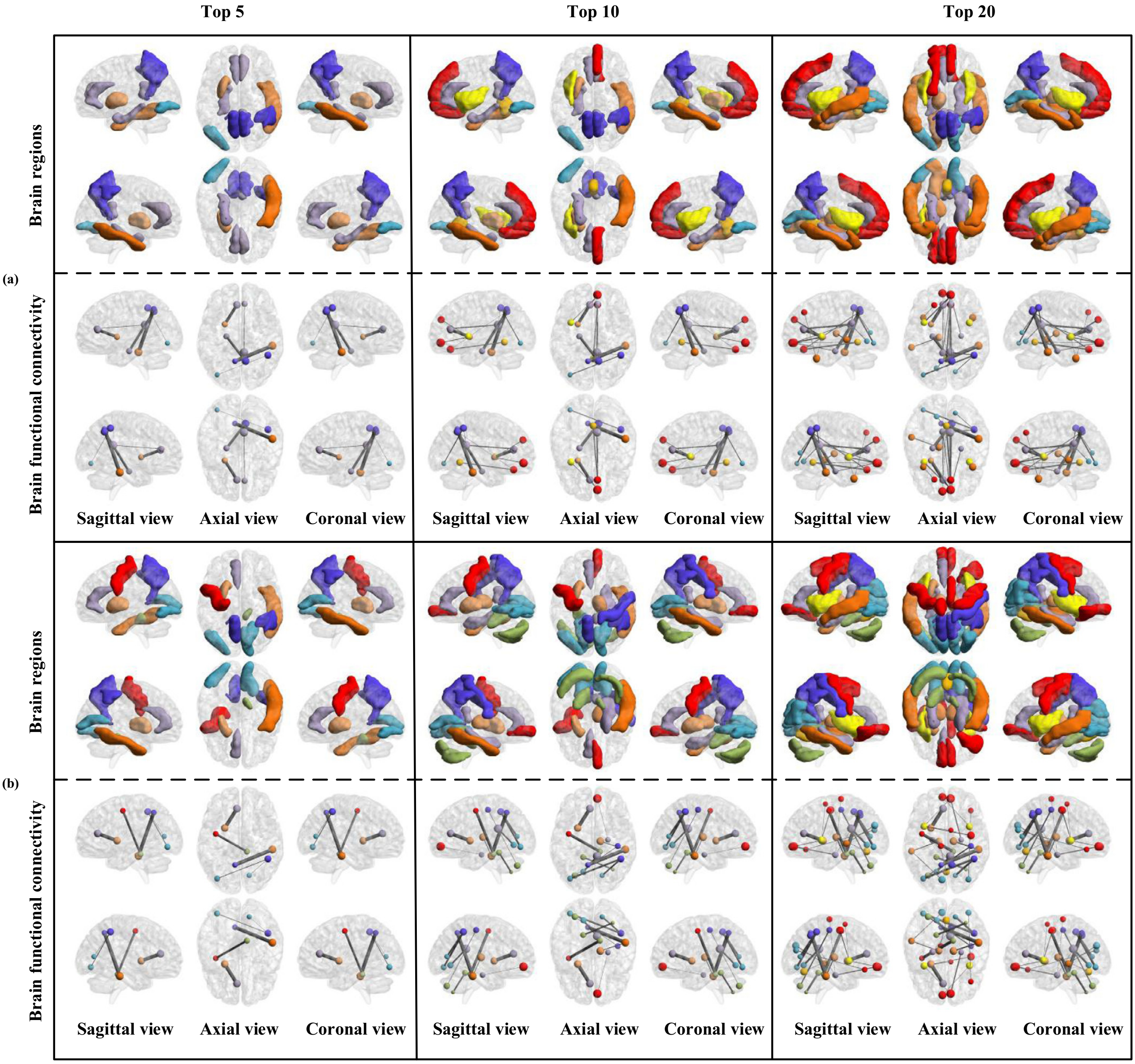


Fig. 8. Common discriminative rs-fMRI connections and corresponding brain regions selected by our SASL (a) four imaging centers via ensemble strategy. (b) C4 dataset. For subfigure, the top and bottom sub-figure represent brain regions and brain functional connectivity, respectively. The size of bubble represents the number of FC it involves.

the group information, which tends to be shared by the similar structures FCs of different subjects. Therefore, in future work, we may need to integrate this information in the future work.

Second, the proposed method is only based on feature representations derived from rs-fMRI data, while feature representations can also be obtained from multiple sources. For example, Liu et al. (2016a) proposed to differentiate AD/MCI patients from healthy controls using three types of data resources (i.e., MRI, PET and cerebrospinal fluid biomarkers). In addition, Zhu et al. (2015) proposed a relational regularization feature learning method for AD diagnosis. To make full use of complementary information of multiple data sources, the feature matrix is constructed by concatenating MRI and PET features at each row. In another work, Wang et al. (2017) realized automatic diagnosis of ASD with two types of feature representations, namely, regional morphological features from T1-weighted MRI and FC features extracted from rs-fMRI. Multi-source features contain a wealth of information, which is beneficial to identify brain disease. Therefore, multi-source feature learning will be our feature learning directions.

Third, age and gender are important factors in the development and progression of ASD. Wiggins et al. (2012) discovered that patients with ASD differ from NC with respect to their age-related changes in the functional connectivity. Alaerts et al. (2016) also observed females and males with ASD have inconsistent connectivity patterns. Specifically, males are mainly in hypo-connectivity patterns, yet females are mainly in hyper-connectivity patterns. Unfortunately, most current methods including our method neglect the age-related changes of brain functional connectivity that occur during development and fail to consider the sex divergences of both ASD patients and NC. Therefore, the effects of age and gender on the network structure and the classification performance are worth to investigate in future work.

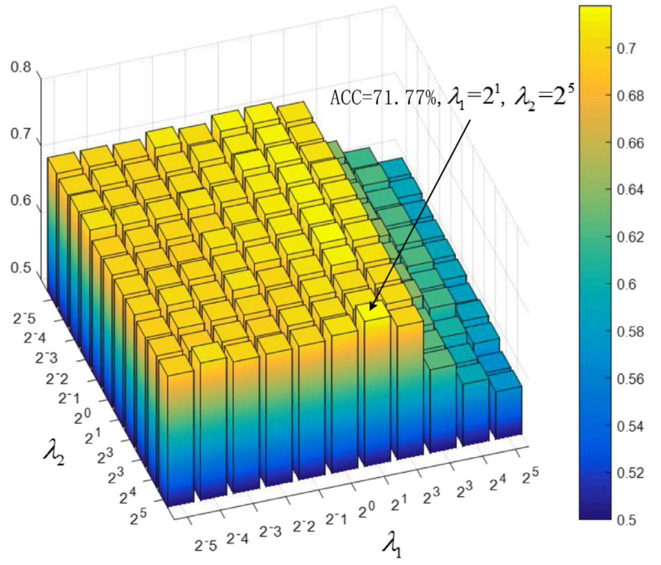


Fig. 9. Classification accuracy based on the networks estimated by our proposed PSLR method with different regularized parameters. The experiment is conducted on C4 dataset.

7. Conclusions

We propose a novel self-weighted adaptive structure learning based multi-template multi-center feature representations. It learns the local neighboring structure via an adaptive process for each template space and combine them together to enhance ASD diagnosis performance. We assess the diagnosis performance of the proposed method on four datasets from the ABIDE database. Our method achieves the best performance among all methods under comparison, including single-center based methods, single-template based methods and different FC network construct methods. Also, it is suggested that the fusion of FC features from multiple centers and multiple templates can offer more informative and discriminative biomarkers for aiding brain disease diagnosis. To evaluate our proposed FC evaluation method, we also compare it with other commonly used feature representation methods. The results on four datasets confirmed the efficacy of our proposed method.

Declaration of Competing Interest

The authors have no financial and personal conflicts of interest.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China (nos. U1902209, U1902209, 61871274, and 61801305), Guangdong Pearl River Talents Plan (2016ZT06S220), Shenzhen Peacock Plan (nos. KQTD2016053112051497 and KQTD2015033016104926), and Shenzhen Key Basic Research Project (nos. JCYJ20180507184647636, JCYJ20170818142347251, JCYJ20170818094109846 and JCYJ20170413152804728).

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. NeuroImage 147, 736–745.

Alaerts, K., Swinnen, S.P., Wenderoth, N., 2016. Sex differences in autism: a resting-state fMRI investigation of functional brain connectivity in males and females. Soc. Cogn. Affect. Neurosci. 11, 1002–1016.

Andrews-Hanna, J.R., Smallwood, J., Spreng, R.N., 2014. The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. Ann. New York Acad. Sci. 1316, 29–52.

Cerliani, L., Mennes, M., Thomas, R.M., Martino, A.D., Thioux, M., Keysers, C., 2015. Increased functional connectivity between subcortical and cortical resting-state networks in autism spectrum disorder. JAMA Psychiatry 72, 767–777.

Chandana, S.R., Behen, M.E., Juhász, C., Muzik, O., Rothermel, R.D., Mangner, T.J., Chakraborty, P.K., Chugani, H.T., Chugani, D.C., 2005. Significance of abnormalities in developmental trajectory and asymmetry of cortical serotonin synthesis in autism. Int. J. Dev. Neurosci. 23, 171–182.

Chang, C.-C., Lin, C.-J., (2012). LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie. ntu. edu. tw/~ cjlin/libsvm.

Chao-Gan, Y., (2014). Data Processing Assistant for Resting-State fMRI (DPARSF).

Chen, C.P., Keown, C.L., Jahedi, A., Nair, A., Pflieger, M.E., Bailey, B.A., Müller, R.-A., 2015. Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism. NeuroImage 8, 238–245.

Combettes, P.L., Pesquet, J.-C., 2011. Proximal splitting methods in signal processing. In: Fixed-point Algorithms for Inverse Problems in Science and Engineering. Springer, pp. 185–212.

Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum. Brain Mapp. 33, 1914–1928.

Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett Jr., J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.

Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S., 2017. Identifying autism from resting-state fMRI using long short-term memory networks. In: Proc. MLMI, Quebec City, Quebec, Canada. Springer, pp. 362–370.

Fan, K., 1949. On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc. Natl. Acad. Sci. USA 35, 652–655.

Hampson, D.R., Blatt, G.J., 2015. Autism spectrum disorders and neuropathology of the cerebellum. Front. Neurosci. 9, 420.

Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F., 2017. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage 17, 16–23.

Jie, B., Zhang, D., Cheng, B., Shen, D., Initiative, A.s.D.N., 2015. Manifold regularized multitask feature learning for multimodality disease classification. Hum. Brain Mapp. 36, 489–507.

Jun, E., Kang, E., Choi, J., Suk, H.-I., 2019. Modeling regional dynamics in low-frequency fluctuation and its application to Autism spectrum disorder diagnosis. NeuroImage 184, 669–686.

Kam, T.E., Suk, H.I., Lee, S.W., 2017. Multiple functional networks modeling for autism spectrum disorder diagnosis. Hum. Brain Mapp. 38, 5804–5821.

Lee, H., Lee, D.S., Kang, H., Kim, B.-N., Moo, K.C., 2011. Sparse brain network recovery under compressed sensing. IEEE Trans. Med. Imaging 30, 1154–1165.

Liu, M., Zhang, D., Adeli, E., Shen, D., 2016a. Inherent structure based multi-view learning with multi-template feature representation for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. 63, 1473–1482.

Liu, M., Zhang, D., Shen, D., 2016b. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging 35, 1463–1474.

Mundy, P., 2003. Annotation: The neural basis of social impairments in autism: The role of the dorsal medial-frontal cortex and anterior cingulate system. J. Child. Psychol. Psychiatry 44, 793–809.

Nie, F., Cai, G., Li, X., 2017. Multi-View clustering and semi-supervised classification with adaptive neighbours. In: Proc. AAAI, San Francisco, California, USA, pp. 2408–2414.

Nielsen, J.A., Zielinski, B.A., Fletcher, P.T., Alexander, A.L., Lange, N., Bigler, E.D., Lainhart, J.E., Anderson, J.S., 2013. Multisite functional connectivity MRI classification of autism: ABIDE results. Front. Hum. Neurosci. 7, 599.

Plitt, M., Barnes, K.A., Martin, A., 2015. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. NeuroImage 7, 359–366.

Qiao, L., Zhang, H., Kim, M., Teng, S., Zhang, L., Shen, D., 2016. Estimating functional brain networks by incorporating a modularity prior. NeuroImage 141, 399–407.

Smith, S.M., Diego, V., Beckmann, C.F., Glasser, M.F., Mark, J., Miller, K.L., Nichols, T.E., Robinson, E.C., Gholamreza, S.K., Woolrich, M.W., 2013. Functional connectomics from resting-state fMRI. Trends Cogn. Sci. 17, 666–682.

Sporns, O., 2011. Brain Network Disease. MIT press.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

Urbain, C., Vogan, V.M., Ye, A.X., Pang, E.W., Doesburg, S.M., Taylor, M.J., 2016. Desynchronization of fronto-temporal networks during working memory processing in autism. Hum. Brain Mapp. 37, 153–164.

Urbain, C.M., Pang, E.W., Taylor, M.J., 2015. Atypical spatiotemporal signatures of working memory brain processes in autism. Transl. Psychiatry 5, e617.

Wang, J., Wang, Q., Peng, J., Nie, D., Zhao, F., Kim, M., Zhang, H., Wee, C.Y., Wang, S., Shen, D., 2017. Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study. Hum. Brain Mapp. 38, 3081–3097.

Wang, J., Wang, Q., Zhang, H., Chen, J., Wang, S., Shen, D., 2018. Sparse multiview task-centralized ensemble learning for ASD diagnosis based on age-and Ssex-related functional connectivity patterns. IEEE Trans. Cybern. 1–14.

Wang, M., Zhang, D., Huang, J., Shen, D., Liu, M., 2018b. Low-rank representation for multi-center autism spectrum disorder identification. In: Proc. MICCAI. Springer, pp. 647–654.

Wiggins, J.L., Bedoyan, J.K., Peltier, S.J., Ashinoff, S., Carrasco, M., Weng, S.-J., Welsh, R.C., Martin, D.M., Monk, C.S., 2012. The impact of serotonin transporter (5-HTTLPR) genotype on the development of resting-state functional connectivity in children and adolescents: a preliminary report. Neuroimage 59, 2760–2770.

Zhou, J., Chen, J., Ye, J., 2011a. Malsar: Multi-task learning via structural regularization. Arizona State University.

Zhu, X., Suk, H.I., Wang, L., Lee, S.W., Shen, D., 2015. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Med. Image Anal. 38, 205–214.

Zhuang, J., Dvornek, N., Li, X., Ventola, P, S. Duncan, J., 2017. Invertible network for classification and biomarker selection for ASD. In: Proc. MICCAI, Shenzhen, Guangdong, China, pp. 700–708.