SCIENCE CHINA
Technological Sciences

• Article •

A hybrid method to select morphometric features using tensor completion and F-score rank for gifted children identification

ZHANG Jin[1*], FENG Fan[2*], HAN Tianyi[1], DUAN Feng[2**], SUN Zhe[3], CESAR F. Caiafa[4], JORDI Solé-Casals[2,5,6**]

[1] College of Computer Science, Nankai University, 300071 Tianjin, China
[2] Department of Artificial Intelligence, Nankai University, 300350, China
[3] Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, 351-0198 Saitama, Japan
[4] Instituto Argentino de Radioastronomía—CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina
[5] Department of Psychiatry, University of Cambridge, Cambridge CB2 0SZ, United Kingdom
[6] Data and Signal Processing Group, University of Vic—Central University of Catalonia, 08500 Vic, Catalonia, Spain

Gifted children are able to learn in a more advanced way than others, probably due to neurophysiological differences in the communication efficiency in neural pathways. Topological features contribute to understanding the correlation between the brain structure and intelligence. Despite decades of neuroscience research using MRI, methods based on brain region connectivity patterns are limited by MRI artifacts, which therefore leads to revisiting MRI morphometric features, with the aim of using them to directly identify gifted children instead of using brain connectivity. However, the small, high-dimensional morphometric feature dataset with outliers makes the task of finding good classification models challenging. To this end, a hybrid method is proposed that combines tensor completion and feature selection methods to handle outliers and then select the discriminative features. The proposed method can achieve a classification accuracy of 93.1%, higher than other existing algorithms, which is thus suitable for the small MRI datasets with outliers in supervised classification scenarios.

gifted children identification, morphometric features, tensor completion, feature selection

1 Introduction

Intelligence can be defined as the ability to perceive, comprehend and infer information within a context. Gifted individuals are those with outstanding cognitive abilities and creativity and are said to possess giftedness [1]. They not only have a higher intellectual ability, but also learn faster than the average individual in a quantitatively different way, presumably due to neurophysiological differences [2]. The neurological differences mean that gifted individuals may experience different neural development trajectories than neurotypical individuals during their childhood, leading to greater inter-connectivity between neural pathways [3].

To map the network of anatomically connected regions in an individual human brain, several tools for magnetic resonance imaging (MRI) were developed during decades of neuroscience research, which provided the opportunity to identify gifted children more qualitatively from the perspective of neural network connectivity instead of simply using intelligence quotient (IQ) test scales for quantitative testing. There are two approaches that can test anatomical connectivity: tractography extracted from diffusion-weighted imaging (DWI) [4–6] and morphology patterns calculated from structural covariance network (SCN) [7–9].

Diffusion-weighted tractography can visually reconstruct the trajectory of white matter by tracking the main diffusion directions of water molecules. However, these are disturbed by head movements [10] and the method generates a large number of false-positive connections [11].

*These authors contributed equally to this work
**Corresponding authors (email: duanf@nankai.edu.cn; jordi.sole@uvic.cat)

Structural covariance analysis can construct whole-brain networks in two steps. Firstly, for each brain region in multi-model imaging data, one single morphometric feature is measured. After that, the covariance between each region is estimated, resulting in a single SCN [12, 13]. However, the performance of this method heavily depends on the size of the collected MRI data [14].

Due to the artifacts in the MRI acquisition process and the small size of available samples in the dataset, these two methods are still insufficient to establish the human cortical connectivity network, which is not suitable for small MRI dataset classification.

Thus, we turn our attention to the MRI data itself, and use the morphometric features instead of the brain cortical connectivity to identify gifted children.

On the one hand, due to the limitations of the MRI acquisition device or the scan parameters, the MRI artifact may affect the morphometric feature extraction, resulting in outliers which hinder the performance improvement of the classification model. Recently, research in tensor completion (TC), which is a higher-order extension of matrix completion, has achieved a consistent performance in a variety of real-world applications [15–18]. Given a tensor with incomplete entries, tensor completion methods such as Recent Low-Rank based Tensor Completion (LRTC) [19] and Simultaneous Tensor Decomposition and Completion (STDC) [20] use low-rank factorization to model the available data entries, which are then used for completion to recover the missing values [21–23]. TC methods have already been used in electroencephalography (EEG) to recover missing samples, showing better performances than other simple imputation methods [24, 25].

On the other hand, in machine learning scenarios, typically an enormous number of samples is required to ensure that there are enough samples to learn the rules of a high-dimensional space. The gifted children MRI dataset in this paper contains 2156 features (7 morphometric features with 308 brain regions), which only consists of 29 samples. It is quite difficult to train classification models using such small, high-dimensional datasets. Feature selection methods, such as Principal Components Analysis (PCA) [26] and F-score [27] can reduce the feature dimension, which will be applied to enhance the classification models [28–30].

To address these two problems, here a hybrid method is proposed to select morphometric features and brain regions from the gifted children MRI dataset. Firstly, with a brain parcellation template, the morphometric feature matrix for each sample is extracted. Then, outlier masks for both the gifted and control groups are generated. These two group outlier masks and their morphometric features matrices are used as the input of the STDC algorithm. The whole outlier completed dataset is then split into a training set and a validation set. F-score rank algorithm, the proposed feature selection method, is then applied on the training set to obtain the features mask, which is applied on both the training and validation sets. Multiple machine learning techniques are then used to carry out the classification.

With this hybrid method, the well-trained classification models can achieve a 93.1% accuracy. Hence, the proposed method can be a good alternative to identify morphometric features and brain regions related to giftedness and can be applied to other small, high-dimensional datasets.

The rest of the paper is organized as follows: details on the tensor completion method, the feature selection method and the experiment settings will be outlined in Section 2. Results on various feature selection methods and features of the giftedness will be discussed in Section 3. Finally, a conclusion will be drawn in Section 4.

2 Materials and methods

2.1 Gifted-children MRI dataset

The gifted children MRI dataset [12] used in this paper consists of 29 healthy right-handed male subjects with no history of neurological disorders. Table 1 shows the dataset details. The groups were not significantly different in terms of age. These gifted children not only have a superior IQ, but can also achieve a high performance in various types of tasks, such as spatial, numerical, verbal reasoning, abstracting reasoning, and memory tasks [31].

With the same scanning procedures and parameters used in [12], all participants were examined on a 3T MRI scanner (Magnetom Trio Tim, Siemens Medical Systems). The raw (anonymised) MRI dataset are available in the Open-Neuro repository (https://openneuro.org/datasets/ds001988).

Table 1 Class Membership Information of Gifted Children Dataset

Group	Control Group (CG)	Gifted Group (GG)
Subject Count	14	15
Average Age	12.53	12.03
Average IQ	122.71	148.80

2.2 Brain region atlas and morphometric features

Each individual brain was parcellated into 308 cortical regions. The parcellation atlas was constructed based on the Desikan-Killiany atlas (68 cortical regions). Each region defined in the Desikan-Killiany atlas was sub-parcellated into spatially contiguous regions by the backtracking algorithm [32] in FreeSurfer, so that the final parcels could be constrained by the original anatomical boundaries. All regions in the new parcellation were approximately equal in size (500 mm^2).

The original feature matrix for each sample consists of 7 morphometric features measured at each of the 308 brain regions. For each brain region, surface and volume-based morphometric features were estimated. Figure 1 shows the

morphometric features (grey matter volume, cortical thickness, surface area, intrinsic curvature, mean curvature, curvature index, and folding index).

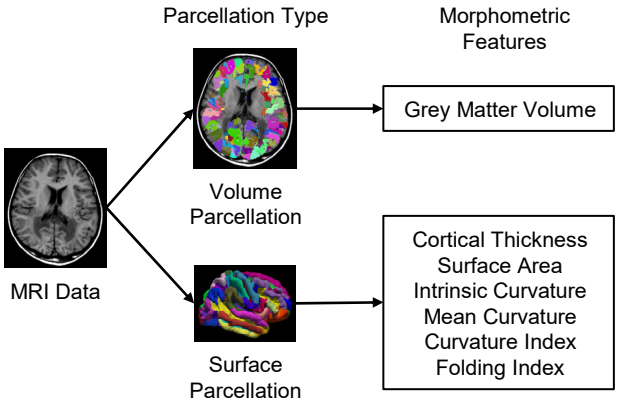

Figure 1 The pipeline for morphometric feature extraction. Each brain was parcellated into 308 regions. Volume and surface parcellation templates were used. All MRI data were mapped to the same cortical parcellation templates to produce a 7 × 308 feature matrix for each subject.

2.3 Outlier detection method

The raw morphometric feature matrix contains outliers due to the acquisition procedure and the applied pre-processing. The interquartile range (IQR) method based on box plots was used to detect outliers, as shown in Figure 2.

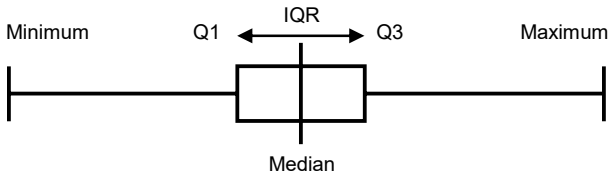

Figure 2 The minimum, maximum and median in the box plot correspond to the min, max and median values in the dataset, respectively. Q_1 and Q_3 are the first and third quartile of the dataset, respectively.

The difference between Q_1 and Q_3 is called the IQR:

$$IQR = Q_3 - Q_1 \tag{1}$$

The decision boundary is set to 1.5 times the IQR. Any morphometric feature value which is smaller than the lower boundary (LB) or larger than the upper boundary (UB) should be considered as an outlier.

$$\begin{aligned} LB &= Q_1 - 1.5 \times IQR \\ UB &= Q_3 + 1.5 \times IQR \end{aligned} \tag{2}$$

2.4 Outlier completion method: STDC

Tensor representation has been widely studied for multiple scenarios where only a subset of entries is missing. Considering an n^{th} tensor $\mathcal{X}$, the Canonical-Polyadic (CP) tensor decomposition [33] can be described as follows:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(k)} \tag{3}$$

where $\boldsymbol{u}_r^{(k)}$ is the decomposed factor along the k^{th} mode, which can be referred to as the k^{th} mode factor.

Given an n^{th} order tensor $\mathcal{X}_0 \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$ and a subset of missing entries $\bar{\Omega}$, the tensor completion methods try to find a tensor $\mathcal{X}$ with its components $\mathcal{Z}, \boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$, such that $\mathcal{X}_0$ and $\mathcal{X}$ have the same observed entries, which can be described as follows:

$$\mathcal{X} = \mathcal{Z} \times \boldsymbol{V}_1^T \times \cdots \times \boldsymbol{V}_n^T \ and \ \Omega(\mathcal{X}) = \Omega(\mathcal{X}_0) \tag{4}$$

where $\mathcal{Z}$ is an n^{th} order tensor of the same size as $\mathcal{X}_0$, and each $\boldsymbol{V}_k$ denote a $I_k \times I_k$ matrix. If $\mathcal{X}$ is a low-rank tensor, then the core tensor $\mathcal{Z}$ is of low rank, or $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$ are a set of low-rank matrices. Such a low-rank property is usually regarded as a global prior in tensor completion.

To find the proper $\mathcal{X}, \mathcal{Z}, \boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$, the STDC method defines an augmented Lagrange function as follows:

$$\begin{aligned} & L(\mathcal{X}, \mathcal{Z}, \boldsymbol{V}_1, \cdots, \boldsymbol{V}_n, \mathcal{Y}, \mu) \\ &= \sum_{k=1}^{n} \alpha_k \left\| \boldsymbol{V}_k \right\|_* + \beta tr((\boldsymbol{V}_1 \otimes \cdots \otimes \boldsymbol{V}_n) \boldsymbol{L} (\boldsymbol{V}_1 \otimes \cdots \otimes \boldsymbol{V}_n)^T) \\ &+ \gamma \left\| \mathcal{Z} \right\|_F^2 + \left\langle \mathcal{Y}, \mathcal{X} - \mathcal{Z} \times \boldsymbol{V}_1^T \times \cdots \times \boldsymbol{V}_n^T \right\rangle \\ &+ \frac{\mu}{2} \left\| \mathcal{X} - \mathcal{Z} \times \boldsymbol{V}_1^T \times \cdots \times \boldsymbol{V}_n^T \right\|_F^2 \end{aligned} \tag{5}$$

The STDC [20] method estimates the decomposed latent factors using partially observed data. It is possible to compute the missing entries from the estimated latent factors. The factor prior, or regularization, is often applied on decomposed components based on low-rank structure [19] in the STDC framework. Patterns can then be obtained in la-tent matrices, yielding the minimum number of rank-one tensors. The STDC [20] can be described as follows:

Algorithm 1. STDC

Input: an incomplete tensor $\mathcal{X}_0 \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$
the parameters $\alpha_1, \cdots, \alpha_n, \beta, \gamma, \mu$

1) Initialize $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$ by
$\boldsymbol{V}_k$: identify matrix
$\mathcal{X}, \mathcal{Z}$: $\mathcal{X}_0$
$\mathcal{Y}$: a tensor with all zero elements

2) Let $t = 1$ and $\mu^1 = \mu$

3) Do
For $t = 1$ to n
Optimize $\boldsymbol{V}_k, \mathcal{X}, \mathcal{Z}$
Update $\mathcal{Y} = \mathcal{Y} + \mu^t (\mathcal{X} - \mathcal{Z} \times \boldsymbol{V}_1^T \times \cdots \times \boldsymbol{V}_n^T)$
$\mu^{t+1} = \rho \mu^t, \rho \in [1.1, 1.2]$
$t = t + 1$
While $\left\| \mathcal{X} - \mathcal{Z} \times \boldsymbol{V}_1^T \times \cdots \times \boldsymbol{V}_n^T \right\|_F^2 > 10^{-4} \left\| \mathcal{X}_0 \right\|_F^2$

Output: the submanifolds $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$
the core tensor $\mathcal{Z}$
the complete tensor $\mathcal{X}$

The raw morphometric feature matrix is treated as the incomplete n^{th} tensor $\mathcal{X}_0$. For both gifted and control groups, their outlier mask matrix is treated as matrix $\boldsymbol{V}_k$.

2.5 Feature selection methods

After tensor completion, the outliers in the feature matrix for both the gifted and control groups will be completed. After splitting the dataset into training sets and validation sets, several feature selection methods, as well as the proposed F-score rank method, are applied only on the training sets to explore which morphometric features and brain regions can achieve a better performance on gifted children identification. A features mask is then generated to indicate the selected morphometric features and brain region indices.

(1) NONE feature selection

For each subject, a feature matrix is defined, which contains 7 morphometric features of 308 brain regions. For each feature matrix, a one-dimensional feature vector can be generated through vectorization. In this case, all the features in the raw feature matrix are used and are put into the classification models. In this paper, this is called the **NONE** feature selection method.

(2) VON feature selection

The von Economo atlas [34] subdivides the cortex into five different categories, according to the laminar structure of the brain cortex and its corresponding functional cortical specializations. Regions located in the primary motor cortex are classified as type 1. Association cortices are referred to as types 2 and 3, while secondary and primary sensory areas are type 4 and type 5, respectively.

For type 2 and type 3 brain cortex regions, there are significant differences in the neurobiological substrate of versatility between the gifted and control groups in terms of the cortical thickness (CT) morphometric features [12]. Thus, it is reasonable to select those types of brain regions as features for the classification system.

In this case, all 7 morphometric features from these corresponding types 2 and 3 brain regions are treated as features, and in this paper, this is called the **VON** feature selection method.

(3) F-score feature selection

Given the training vectors $x_k, k = 1, \cdots, m$, the size of the gifted group p and the size of the control group q, the F-score [27] of the i^{th} feature can be defined as:

$$F(i) = \frac{(p-1)(q-1)(\bar{x}_i^p - \bar{x}_i)^2 + (\bar{x}_i^q - \bar{x}_i)^2}{\sum_{k=1}^{p}(x_{k,i}^p - \bar{x}_i^p)^2 + \sum_{k=1}^{q}(x_{k,i}^q - \bar{x}_i^q)^2} \tag{6}$$

where $\bar{x}_i, \bar{x}_i^p, \bar{x}_i^q$ are the average of the whole, gifted and control groups for the i^{th} feature, respectively, and $\bar{x}_{k,i}^p, \bar{x}_{k,i}^q$ are the i^{th} feature of the k^{th} subject for each group.

The numerator indicates the discrimination between the gifted and control sets, and the denominator indicates the dispersion within each of the two sets.

In this case, a threshold is set to select the top highest features from the morphometric features and brain regions. In this paper, this method will be referred to as the **F-score (FS)** feature selection method.

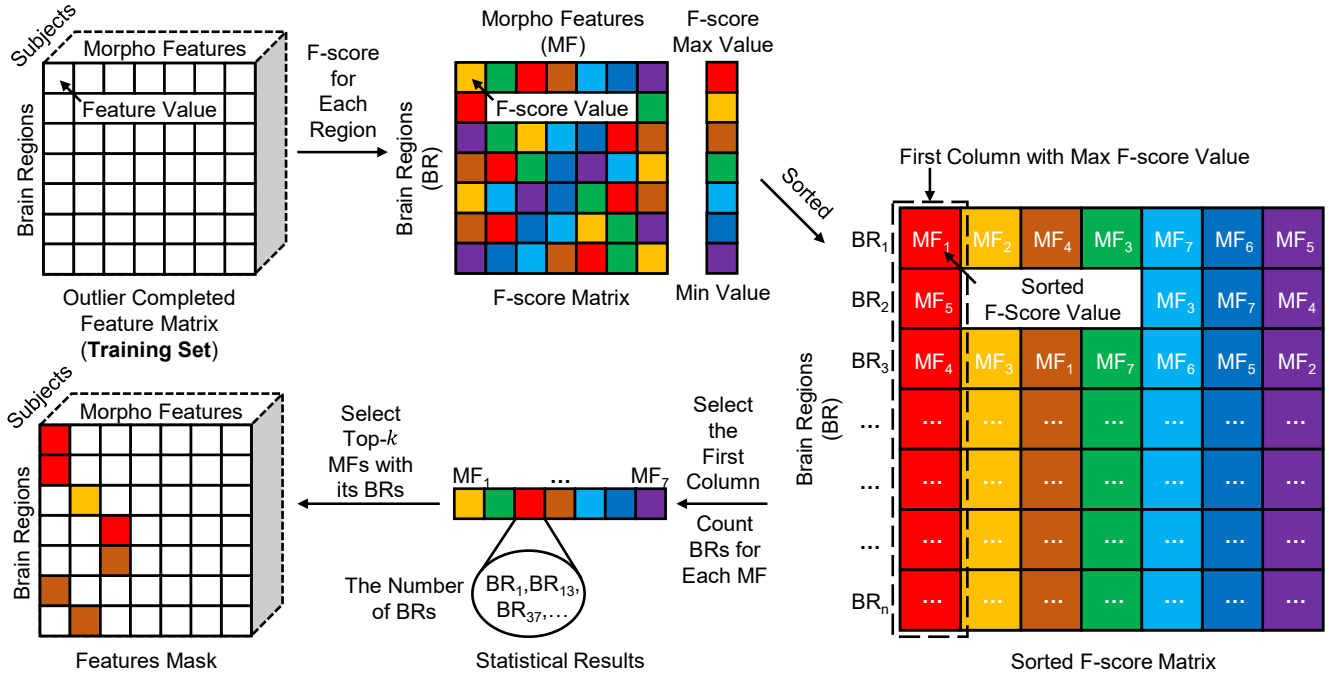

Figure 3 The pipeline of the proposed F-score rank method. Notice that all the feature selection methods are only applied on the training set. The range of the F-score value in the F-score matrix is expressed in different colours (from red to purple). After sorting the F-score value of each row in descending order, only the first column is selected, which is the maximum F-score value with different morphometric feature indexes for each brain region. In this figure, the top 3 morphometric features (red, orange and brown blocks) with their corresponding brain regions are selected after counting the number of brain regions for each morphometric feature. Thus, the brain regions are mutually exclusive for each selected morphometric feature in the final features mask.

(4) F-score rank feature selection

For the original F-score algorithm, the input will be a one-dimensional feature vector in which the morphometric features and brain regions are feature candidates which are all treated as independent items. However, the functions performed by adjacent brain regions may be similar, and the correlation between morphometric features and brain regions may be discarded by this method. Thus, to obtain more interpretable features and to further reduce the number of features, here the F-score rank feature selection method is proposed, which can cluster brain regions according to their correlation with morphometric features.

Figure 3 shows the pipeline of the proposed F-score rank feature selection method. Firstly, the F-score value of 7 morphometric features for each brain region is calculated among the training set, yielding a two-dimensional F-score matrix. For each region, the F-score values are sorted by descending order, where the morphometric feature with the highest F-score value should be put in the first column. The morphometric feature and brain region indexes corresponding to the highest F-score value are considered to have the strongest correlation. Hence, only the first column is picked up and the number of brain regions for each group is counted. In this step, all 308 brain regions will be categorized into 7 groups (morphometric features). Finally, as with the F-score method, a threshold is set to select the morphometric features and brain regions from the statistical results, and a feature mask will be generated to define the feature index for the classification model input.

In this case, a subset of both the morphometric features and the brain regions are selected by categorizing brain regions into morphometric feature groups. In this paper, this method will be referred to as the **F-score rank (FSR)**.

(5) Combined feature selection

Furthermore, the VON and F-score methods were combined to create a joint method. In this case, only type 2 and type 3 regions are considered when calculating the F-score value for morphometric features. In this paper, this method is referred to as the **VFS** feature selection method.

For all the feature selection methods, a feature mask will be generated, which can indicate the selected feature indices. Then, the feature mask is mapped to the training set for model training and to the validation set for evaluation.

2.6 Classification model and cross validation

Since the gifted MRI dataset only contains 29 subjects, traditional machine learning classification methods, such as the K-Nearest Neighbor (KNN) and the support vector machine (SVM), are considered in this paper instead of using deep learning network (DNN) methods.

Considering the size of the dataset, it is hard to execute k-fold validation. Thus, we applied the Leave-One-Out (LOO) cross validation, where the number of folds equals the number of instances in the dataset.

2.7 Evaluation metric

The accuracy (ACC) was calculated, which measures the proportion of correctly classified subjects for each group, to evaluate the performance of outlier completion methods and feature selection methods.

The proposed hybrid method for selecting discriminative features is evaluated in three steps. In the first step, the performance on outlier detection and completion is evaluated. The performance of the classification models before and after the outlier completion was tested without using any feature selection methods. In the second step, the classification performance of the several feature selection methods mentioned in this paper was evaluated. Furthermore, these feature selection methods were tested on both the original feature matrix and the completed feature matrix. In the last step, the number of selected features in the F-score rank method and the most discriminative features are discussed.

2.8 Hyperparameters and experimental settings

The number of features is the most important hyperparameter for each feature selection method. For the NONE feature selection method, all feature candidates are used for classification, so therefore the number of features is 7×308. The VON method only selected brain regions of type 2 (113 regions) and type 3 (71 regions). Based on the experimental results, the F-score method can achieve the best performance when the number of features is set to 200. For the F-score rank method, however, the maximum number of features is 308, when all morphometric features are selected. Choosing the top 3 morphometric features and brain regions within their categories can achieve the best performance, according to the experimental results.

For all the experiments reported in this paper, a Linux server running Ubuntu 16.04 was used. The server contains one Intel Core i7-6700 processor running at 2.6 GHz.

In summary, Figure 4 shows the whole pipeline for selecting discriminative features with our hybrid method.

Firstly, with the help of brain region parcellation, the morphometric features are estimated to form the original feature matrix, which contains many outliers. Using the IQR method, an outlier mask will be generated to indicate the missing parts for the tensor completion method. Both the original feature matrix and its outlier mask will be used as the input of the tensor completion method. Then, the outlier completed matrix will be split into a training and a validation set, and feature selection methods will be executed during the LOO cross-validation. The selected feature indices will be mapped to the training set and the validation set, respectively. Finally, the training set feature vector will be used for training purposes and the validation set feature vector (only in one instance) will be used for evaluation purposes.

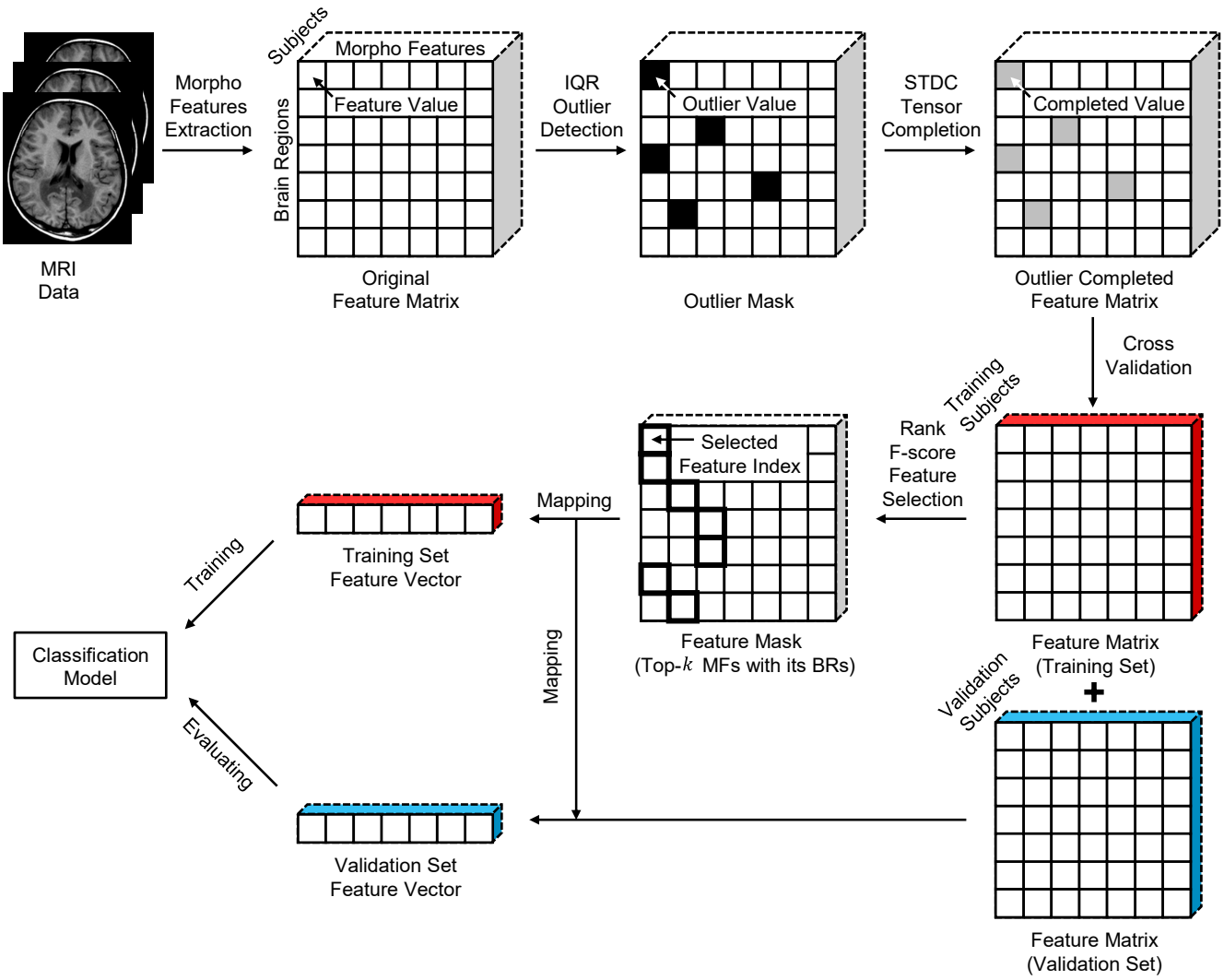

Figure 4 The pipeline of the proposed hybrid method for gifted children identification. The original MRI data is parcellated using a brain atlas, and morphometric features for each brain region are estimated, yielding the original feature matrix. The black blocks in the outlier mask indicate the missing parts detected by the IQR method. With the help of STDC, the completed values are represented by the grey blocks in the outlier completed feature matrix. Before applying the feature selection, the whole feature matrix is split into a training and a validation set. The blocks with a bold border in the feature mask are the selected feature indices, which can map to both the training and the validation sets for further classification.

3 Results and discussion

3.1 Outlier detection and completion

(1) Outlier detection results

The results for outlier detection using IQR methods are presented in Figure 5 (for each morphometric feature) and Figure 6.

Figure 5 shows the number of outliers for each morphometric feature. The last three features have many more outliers than the first four features, which may be caused by the feature estimating method itself. Despite only 4% of the data being outliers, these still have a huge impact on the accuracy of the classification.

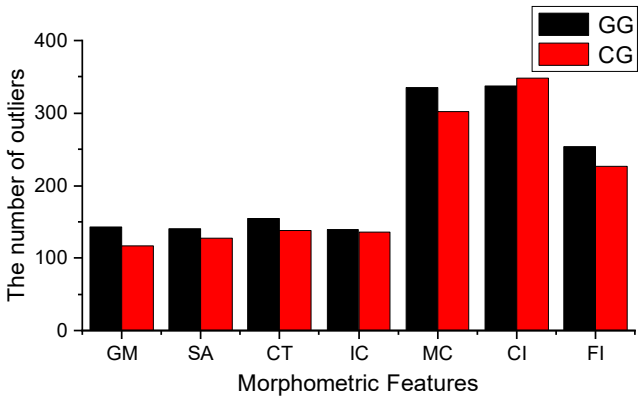

Figure 5 The number of outliers for 7 morphometric features in the gifted and control groups. The last three features have the most outliers.

The number of outliers for each subject was counted, as shown in Figure 6. It can be observed that the 7th subject in the gifted group (GG) and the 14th subject in the control group (CG) have the most outliers among all the subjects, which could make it hard for the classification model to correctly identify these two samples. In addition, all the subjects have outliers, which indicates that it is necessary to detect and fill them properly.

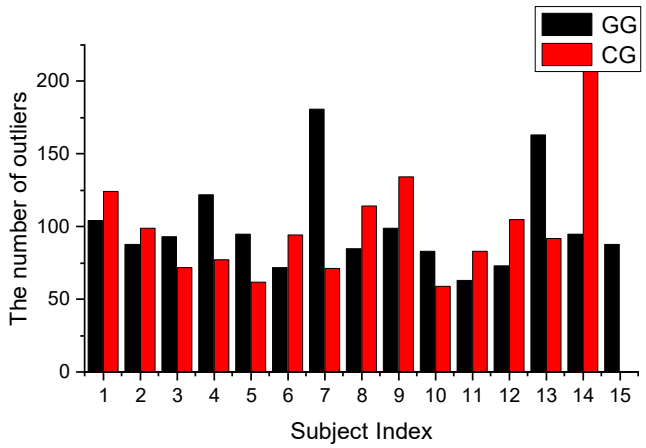

Figure 6 The number of outliers for each subject in the gifted and control groups. All subjects have more or less outliers.

(2) Outlier completion with STDC

After handling the outliers, another outlier detection was conducted to evaluate the performance of the STDC method on outlier elimination, as presented in Figure 7.

In this step, the STDC was iteratively applied 3 times on the STDC output completed matrix. For each epoch, the previous completed matrix and its outlier mask served as the new input for the STDC method.

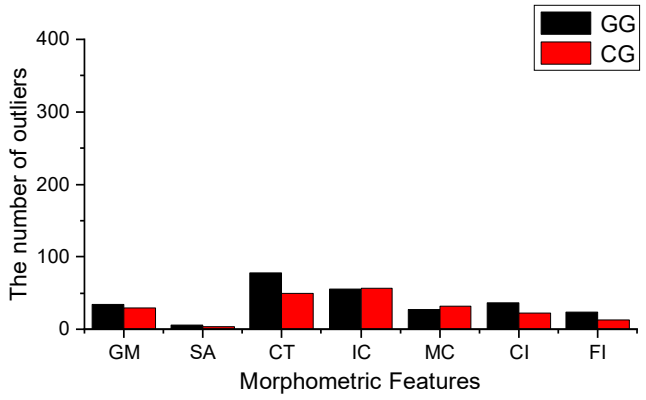

Figure 7 The number of outliers for 7 morphometric features after iteratively applying the STDC method 3 times. Compared with Figure 5, the number of outliers significantly decreases.

Compared with Figure 5, Figure 7 shows that the number of outliers is significantly decreased with the STDC method. This superior performance of the method can be seen as the number of outliers is dropped by an average of 3 times.

Moreover, the classification performance on the completed feature matrix was tested using the median value filling method, LRTC, and STDC by SVM and KNN.

The ACC results in Table 2 show that, after handling the outliers properly, the model can achieve a higher accuracy. It can be seen that the tensor completion algorithms such as LRTC and STDC are more suitable to handle the outliers.

Table 2 ACC for outlier completion methods with SVM and KNN. Notice that feature selection methods are not involved in this phase.

Outlier Completion Method	SVM	KNN
Fill Median Value	0.502	0.622
LRTC	0.766	0.787
STDC	0.793	0.852

3.2 Feature selection method performances

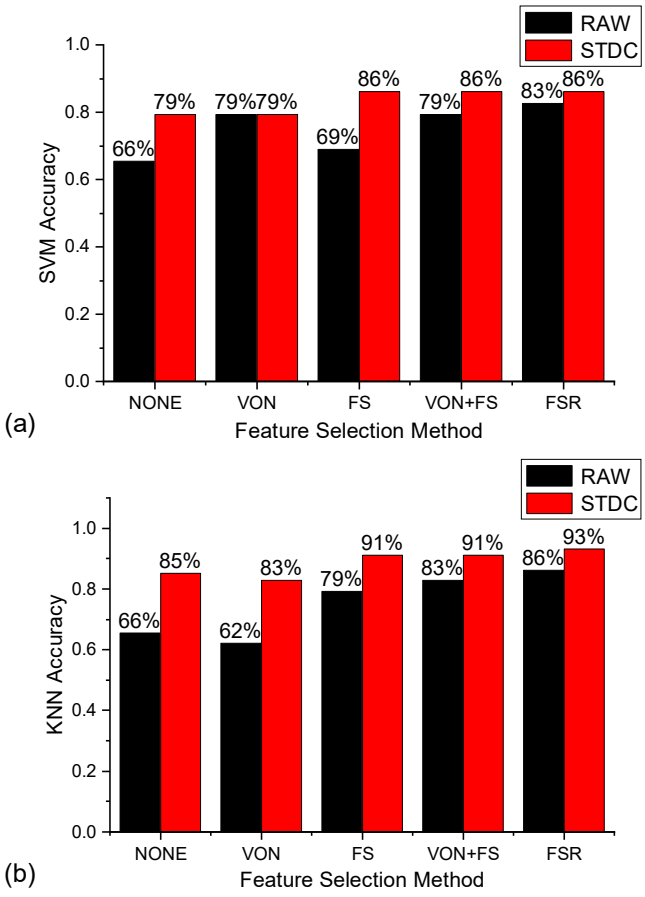

Figure 8 ACC for feature selection methods applied to the original morphometric feature matrix (black bar) and to the completed morphometric feature matrix (red bar), with SVM (a) and KNN (b).

In this step, all the previously mentioned feature selection methods and our proposed F-score rank method were evaluated. These methods were also applied on the original morphometric feature matrix to remove the impact of STDC and thus depict a clearer evaluation of the methods.

Figure 8 shows the accuracy of SVM and KNN for each feature selection method applied to the original feature matrix and to the outlier completed feature matrix.

With the help of STDC, the outlier completed feature matrix can achieve a higher ACC than the original feature matrix, which indicates the outlier completion is also crucial for subsequent feature selection.

For all of feature selection methods, the proposed F-score rank method achieved the highest accuracy (93%) even without the STDC method being applied (86%).

To investigate the contributions of tensor completion and feature selection to the classification accuracy, experiments were run under three different scenarios: original feature matrix without feature selection methods; outlier completed feature matrix without feature selection methods; original feature matrix with the F-score rank method.

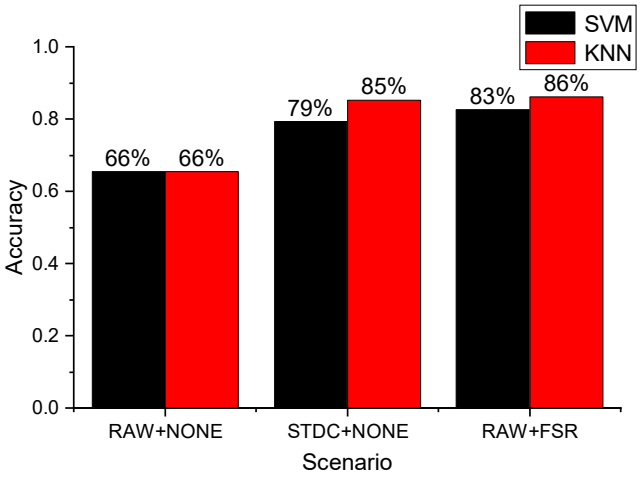

Figure 9 The accuracy of SVM (black) and KNN (red) under the different scenarios. The leftmost columns are the baseline, the middle columns are used to check the impact of the STDC, and the rightmost columns are used to evaluate the influence of the F-score rank method.

It can be observed that compared with using both STDC and F-score rank in Figure 8, using only one of these methods will result in a decrease in accuracy (as seen in Figure 9). In addition, compared with the baseline, the F-score rank feature selection method can achieve a slightly better accuracy than tensor completion, as shown in Figure 9. Thus, it is necessary to use both methods when facing a small dataset with outliers.

3.3 The discriminative features and brain regions

After outlier completion, the F-score rank method is applied on the completed feature matrix. During this phase, the statistical result vector (Figure 3) was collected, and the distribution of the number of brain regions among morphometric features (Figure 10) and the von Economo atlas (Figure 11) was examined.

Among these morphometric features, Surface Area, Cortical Thickness and Gray Matter Volume are the most discriminatory. The number of regions that belong to these three features account for 80% of all features.

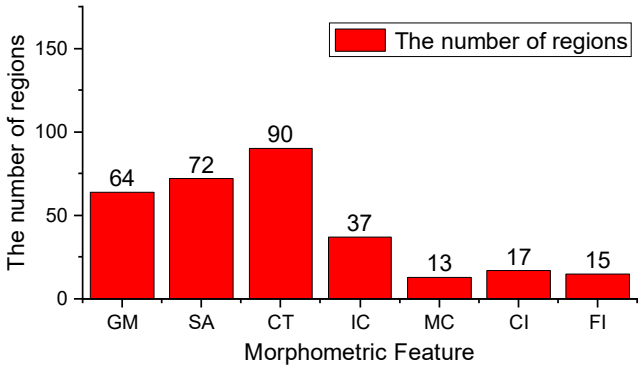

Figure 10 Distribution of the number of brain regions for the training set among 7 morphometric features during the execution of the F-score rank method, which indicates that GM, SA and CT are most discriminative features.

Having selected the SA, CT and GM morphometric features, the distribution of the number of brain regions among the von Economo categories was also analysed. It was found that for all of the 226 selected brain regions, most of them belong to types 1, 2, 3 and 4.

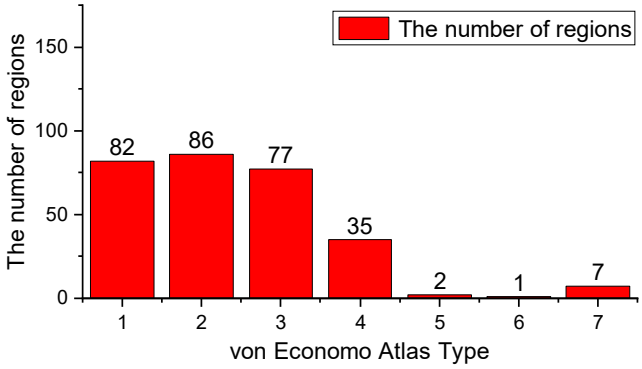

Figure 11 The distribution of the number of brain regions for the training set among 7 von Economo atlas types after choosing the top 3 most discriminative morphometric features. Regions that belong to types 1, 2, 3 and 4 are more discriminative than others.

4 Conclusions

Due to the uncertainty associated with MRI artifacts and heavy acquisition costs, it is not easy to identify gifted children from small, high-dimensional MRI datasets.

In this paper, a hybrid method was proposed to identify gifted children. The novelty of the proposed method is that all the morphometric features and brain regions are explored in depth. The experimental results suggest that the combination of both methods can achieve a better performance than using only one method. Additionally, this hybrid method can be easily expanded to other similar scenarios.

This work was supported by the National Key R&D Program of China (No. 2017YFE0129700), the National Natural Science Foundation of China (Key Program) (No. 11932013), the National Natural Science Foundation of China (No. 61673224), the Tianjin Natural Science Foundation for Distinguished Young Scholars (No. 18JCJQJC46100), and the Tianjin Science and Technology Plan Project (No. 18ZXJMTG00260). J.S-C. work is also based upon work from COST Action CA18106, supported by COST (European Cooperation in Science and Technology).

1. Navas-Sánchez FJ, Carmona S, Alemán-Gómez Y, et al (2016) Cortical morphometry in frontoparietal and default mode networks in math-gifted adolescents. Human Brain Mapping 37:. https://doi.org/10.1002/hbm.23143
2. Gross MUM (2006) Exceptionally gifted children: Long-term outcomes of academic acceleration and nonacceleration. Journal for the Education of the Gifted 29:. https://doi.org/10.4219/jeg-2006-247
3. Navas-Sánchez FJ, Alemán-Gómez Y, Sánchez-Gonzalez J, et al (2014) White matter microstructure correlates of mathematical giftedness and intelligence quotient. Human Brain Mapping 35:. https://doi.org/10.1002/hbm.22355
4. Hales PW, d'Arco F, Cooper J, et al (2019) Arterial spin labelling and diffusion-weighted imaging in paediatric brain tumours. NeuroImage: Clinical 22:. https://doi.org/10.1016/j.nicl.2019.101696
5. Raja R, Rosenberg G, Caprihan A (2019) Review of diffusion MRI studies in chronic white matter diseases. Neuroscience Letters 694
6. Assaf Y, Johansen-Berg H, Thiebaut de Schotten M (2019) The role of diffusion MRI in neuroscience. NMR in Biomedicine 32
7. Yun JY, Boedhoe PSW, Vriend C, et al (2020) Brain structural covariance networks in obsessive-compulsive disorder: A graph analysis from the ENIGMA consortium. Brain 143:. https://doi.org/10.1093/brain/awaa001
8. Qi T, Schaadt G, Cafiero R, et al (2019) The emergence of long-range language network structural covariance and language abilities. NeuroImage 191:. https://doi.org/10.1016/j.neuroimage.2019.02.014
9. DuPre E, Spreng RN (2017) Structural covariance networks across the life span, from 6 to 94 years of age. Network Neuroscience 1:. https://doi.org/10.1162/netn_a_00016
10. Walker L, Gozzi M, Lenroot R, et al (2012) Diffusion tensor imaging in young children with autism: Biological effects and potential confounds. Biological Psychiatry 72:. https://doi.org/10.1016/j.biopsych.2012.08.001
11. Maier-Hein K, Neher P, Houde J-C, et al (2016) Tractography-based connectomes are dominated by false-positive connections. bioRxiv. https://doi.org/10.1101/084137
12. Solé-Casals J, Serra-Grabulosa JM, Romero-Garcia R, et al (2019) Structural brain network of gifted children has a more integrated and versatile topology. Brain Structure and Function 224:. https://doi.org/10.1007/s00429-019-01914-9
13. Bethlehem RAI, Romero-Garcia R, Mak E, et al (2017) Structural covariance networks in children with autism or ADHD. Cerebral Cortex 27:. https://doi.org/10.1093/cercor/bhx135
14. Seidlitz J, Váša F, Shinn M, et al (2018) Morphometric Similarity Networks Detect Microscale Cortical Organization and Predict Inter-Individual Cognitive Variation. Neuron 97:. https://doi.org/10.1016/j.neuron.2017.11.039
15. Yang JH, Zhao X le, Ji TY, et al (2020) Low-rank tensor train for tensor robust principal component analysis. Applied Mathematics and Computation 367:. https://doi.org/10.1016/j.amc.2019.124783
16. Zhao X le, Xu WH, Jiang TX, et al (2020) Deep plug-and-play prior for low-rank tensor completion. Neurocomputing 400:. https://doi.org/10.1016/j.neucom.2020.03.018
17. Huang H, Liu Y, Liu J, Zhu C (2020) Provable tensor ring completion. Signal Processing 171:. https://doi.org/10.1016/j.sigpro.2020.107486
18. Lacroix T, Obozinski G, Usunier N (2020) Tensor decompositions for temporal knowledge base completion. arXiv
19. Lu C, Peng X, Wei Y (2019) Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
20. Chen YL, Hsu CT, Liao HYM (2014) Simultaneous tensor decomposition and completion using factor priors. IEEE Transactions on Pattern Analysis and Machine Intelligence 36:. https://doi.org/10.1109/TPAMI.2013.164
21. Balažević I, Allen C, Hospedales TM (2020) Tucker: Tensor factorization for knowledge graph completion. In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference
22. Najafi M, He L, Yu PS (2019) Outlier-robust multi-aspect streaming tensor completion and factorization. In: IJCAI International Joint Conference on Artificial Intelligence
23. Ko CY, Batselier K, Daniel L, et al (2020) Fast and Accurate Tensor Completion with Total Variation Regularized Tensor Trains. IEEE Transactions on Image Processing 29:. https://doi.org/10.1109/TIP.2020.2995061
24. Solé-Casals J, Caiafa CF, Zhao Q, Cichocki A (2018) Brain-Computer Interface with Corrupted EEG Data: a Tensor Completion Approach. Cognitive Computation 10:. https://doi.org/10.1007/s12559-018-9574-9
25. Feng D, Hao J, Zhenglu Y, et al (2021) On the Robustness of EEG Tensor Completion Methods . SCIENCE CHINA Technological Sciences
26. Gárate-Escamila AK, Hajjam El Hassani A, Andrès E (2020) Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked 19:. https://doi.org/10.1016/j.imu.2020.100330
27. Chen YW, Lin CJ (2006) Combining SVMs with various feature selection strategies. Studies in Fuzziness and Soft Computing 207:. https://doi.org/10.1007/978-3-540-35488-8_13
28. Tsagris M, Lagani V, Tsamardinos I (2018) Feature selection for high-dimensional temporal data. BMC Bioinformatics 19:. https://doi.org/10.1186/s12859-018-2023-7
29. Ang JC, Mirzal A, Haron H, Hamed HNA (2016) Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics 13:. https://doi.org/10.1109/TCBB.2015.2478454
30. Rouhi A, Nezamabadi-Pour H (2020) Feature selection in high-dimensional data. In: Advances in Intelligent Systems and Computing
31. Limiñana Gras RM, Bordoy M, Ballesta GJ, Berna JC (2010) Creativity, intelectual abilities and response styles: Implications for academic performance in the secondary school. Anales de Psicología/Annals of Psychology 26:
32. Desikan RS, Ségonne F, Fischl B, et al (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31:. https://doi.org/10.1016/j.neuroimage.2006.01.021
33. Li Z, Sergin ND, Yan H, et al (2019) Tensor completion for weakly-dependent data on graph for metro passenger flow prediction. arXiv
34. van den Heuvel MP, Scholtens LH, Barrett LF, et al (2015) Bridging cytoarchitectonics and connectomics in human cerebral cortex. Journal of Neuroscience 35:. https://doi.org/10.1523/JNEUROSCI.2630-15.2015