Benchmarking resting state fMRI connectivity pipelines for classification: Robust accuracy despite processing variability in cross-site eye state prediction

Tatiana Medvedeva*, Irina Knyazeva, Ruslan Masharipov, Alexander Korotkov, Denis Cherednichenko, Maxim Kireev

N. P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences, St. Petersburg, Russia

*Corresponding author: Tatiana Medvedeva
E-mail: tanai2010@mail.ru

Keywords

fMRI, Resting state, Functional connectivity, Denoising, Machine learning

Abstract

The rapid evolution of machine learning (ML) methods has yielded promising results in human brain neuroscience. However, the reproducibility of ML applications in neuroimaging remains limited, challenging the generalizability of inferences to broader populations. In addition to the inherent variability of the brain activity (both in healthy and pathological states), poor reproducibility is further enhanced by inconsistencies in data preprocessing techniques and methods for calculating functional connectivity (FC), which are used as parameters for brain state classification. To systematically assess the impact of abovementioned factors on ML applications to fMRI data, we benchmarked a comprehensive set of FC analysis pipelines for the classification task between fMRI data recorded in two fundamentally different states: eyes open

and eyes closed. In contrast to studies involving heterogeneous clinical populations or using complex cognitive tasks, our controlled experimental design – based on two independent datasets of healthy participants collected in different laboratories – minimizes variability related to a task design or pathological brain states. Classification accuracy and reproducibility were compared for 192 distinct FC analysis pipelines, covering common preprocessing approaches, brain parcellation schemes, and connectivity metrics. Notably, we employed a cross-site validation strategy, training predictive models on data from one site and testing them on the other. Despite the substantial variability in pipeline configurations, we observed consistently high classification accuracy (~80%), confirming that FC-based models can robustly discriminate between well-defined brain states (eye conditions) across different acquisition sites. Best results both in terms of classification accuracy and stability were observed using Pearson correlation and tangent correlation as FC, Brainnetome as atlas, and confound regression strategies based on the CompCor method. These findings highlight the resilience of rs-fMRI FC-derived characteristics to methodological variation and support their utility in the discovery of biomarkers, particularly in settings that involve stable and reproducible brain states.

Introduction

Despite recent advances in the application of machine learning (ML) to classify brain states, conditions, and patient groups using resting-state fMRI (rs-fMRI) functional connectivity (FC) parameters, reproducibility remains a major challenge. This limitation significantly prevents the development of reliable neuromarkers, largely due to methodological variability introduced by diverse data processing pipelines.

From rs-fMRI data preprocessing steps to choices in brain anatomical parcellation and functional connectivity metrics, decisions made during fMRI analysis can strongly influence the results obtained using ML models (Noble et al., 2019). This variability not only complicates the comparison of findings across studies, but also undermines confidence in the generalizability of reported results, limiting progress in understanding brain function using ML approaches, and developing reliable ML-based neuroimaging biomarkers for diagnostic purposes. Therefore, achieving high consistency and reproducibility in research utilizing rs-fMRI functional connectivity for ML-based classification is challenging due to a combination of physiological and technical factors. (Peterson et al., 2023) One of the most problematic factors is head movement during fMRI scanning, which can seriously corrupt the results of a study. (Friston et al., 1996) Additionally, physiological sources of variability include heart rate and respiration, state-dependent changes, which also introduce noise into the fMRI signal. (Murphy et al., 2013)

Recent efforts have highlighted the need for systematic evaluations of how methodological decisions influence the conclusions drawn from fMRI data. Methodological changes may be introduced in all steps of the analysis with limitless number of choices, starting with raw MRI data preprocessing and confound regression of BOLD signal (Ciric et al., 2017; Parkes et al., 2018; Satterthwaite et al., 2013), ending with the choice of functional connectivity measures (Z.-Q. Liu et al., 2025; Luppi et al., 2024) and machine learning models (Dadi et al., 2019; Khosla et al., 2018).

Another source of variation in such studies is the choice of an fMRI dataset, according to which fMRI processing pipelines are tested. Some of the most investigated resting state datasets are ABIDE (Di Martino et al., 2014), HCP (Van Essen et al., 2013), ADNI (Mueller et al., 2005), UK Biobank (Mansour et al., 2023). In most studies evaluating classification performance of ML methods using rs-fMRI data, experimental designs typically assume inter-subject comparisons.

They often involve contrasting patients to controls (Eslami et al., 2019; Riaz et al., 2020; Segato et al., 2020), or distinct clinical populations (Qin et al., 2022). However, such designs inherently lack ground truth and introduce considerable variability due to individual differences that are difficult to control. This uncontrolled inter-subject variability can significantly compromise the interpretability and reproducibility of classification results, obscuring the true impact of methodological choices.

To address this limitation, the present study, which focuses on evaluating the robustness of rs-fMRI-based classification pipelines, adopts a within-subject design. By comparing brain states within the same individuals, we effectively control for inter-subject variability, allowing for a more precise assessment of methodological influences. Specifically, we utilize a well-established experimental paradigm that contrasts the eyes open and fixated (EO-F) and eyes closed (EC) conditions during resting-state — two physiological states known to elicit distinct neural activity patterns, as first demonstrated in classical studies (Berger, 1929; Gusnard & Raichle, 2001; Marx et al., 2004) and consistently reported in subsequent literature (Agcaoglu et al., 2020; Costumero et al., 2020; DeRamus et al., 2021; Han et al., 2023; Patriat et al., 2013; Weng et al., 2020). This binary classification task provides a controlled framework for evaluating classification performance and methodological reproducibility.

Opening and closing the eyes triggers activity changes in many brain areas. (Agcaoglu et al., 2019; Costumero et al., 2020; Gusnard & Raichle, 2001; Marx et al., 2004; Patriat et al., 2013; Weng et al., 2020; Yan et al., 2009) For instance, (Agcaoglu et al., 2019) demonstrated a functional connectivity decrease in visual networks when subjects closed their eyes, presumably due to visual input removal. Conversely, opening the eyes restored this visual network activity, highlighting the brain's quick adjustment to visual changes. It was also found that for auditory and sensorimotor brain networks, there were more functionally connected brain regions during

EC than during EO. Furthermore, the default mode network (DMN), a network of brain regions that are active during resting-state, is reported in numerous studies on differences between EO/EC states (Gusnard & Raichle, 2001; Marx et al., 2004; Yan et al., 2009). The medial prefrontal cortex (mPFC) and the posterior cingulate cortex (PCC) are core regions of the default mode network: these areas are associated with mind wandering and self-referential thinking, cognitive states often occurring when an individual is not focused on tasks dealing with external stimuli. (Mason et al., 2007) By using seed-based between group analysis it was demonstrated that in the EC group primary visual cortex had higher connectivity with the DMN and sensorimotor network, whilst in the EO group higher connectivity was observed with the salience network. (Costumero et al., 2020) However, (Yan et al., 2009) showed that the DMN had stronger connectivity for EO than for EC. Moreover, (Patriat et al., 2013) reported very small differences in functional connectivity maps between EO, EO-F and EC conditions. In (Weng et al., 2020) review the main findings of the EC/EO comparisons were summarized as differences in functional connectivity in the resting state observed in the visual cortex, the sensory cortex and supplementary motor area, the thalamus, frontal eye fields and the visual, sensorimotor, default mode, attention networks as well. Therefore, the EO and EC conditions represent suitable candidates for evaluating the effectiveness of classification methods in distinguishing between different brain states using FC fMRI data.

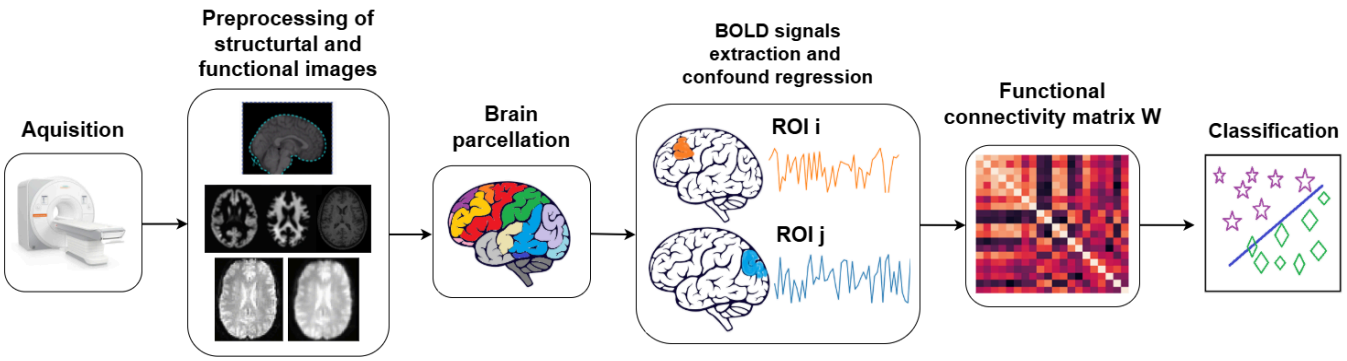

Figure 1
Pipeline of creating a functional connectome based predictive model
1) Data collection stage.
2) Preprocessing of functional and structural images,
which includes image time correction, white and gray matter segmentation,
motion correction, comparison of structural and functional images, image smoothing.
3) Brain parcellation is the grouping of voxels into larger regions of interest.
4) Signal extraction and confound regression – noise associated with
head movement, breathing, blood flow, etc. is removed
from the BOLD signal of the regions of interest.
5) Functional connectivity matrix – each element of the matrix corresponds to
a measure of functional connectivity between the signals of the two regions of interest.
6) The functional connectivity matrix is used to solve a classification problem.

Alongside physiological variability, technical sources of variation further complicate the analysis. Functional connectome-based classification refers to the process of leveraging patterns in functional connectivity data to differentiate among multiple subject states or conditions. (Khosla et al., 2018) Despite advantages of this approach, different analysis workflows may lead to drastically different results. (Dadi et al., 2019) In Figure 1, a common pipeline of creating functional connectome-based classification models is shown. Basically, every step of this pipeline has a great number of parameters to vary:

1. Data acquisition: acquisition parameters variation, effects of scanner and scanned population (e.g., multi-site datasets). (Bari et al., 2019)

2. Functional and structural data minimal preprocessing: variations in raw data processing steps (such as segmentation, motion correction and registration) have a significant impact on results, as shown by (Li et al., 2024).

3. BOLD signal denoising strategies: choosing a non-neural signals removal technique among other methods (Ciric et al., 2017; Parkes et al., 2018; Satterthwaite et al., 2013)

4. Brain parcellation: variability of data-driven algorithms (clustering, ICA, etc.) or predefined atlases (structural or functional, with bigger or smaller number of ROIs). (Arslan et al., 2018)

5. Functional connectivity estimation: diversity of algorithms that measure similarity between regions' activation. (Z.-Q. Liu et al., 2025; Peterson et al., 2023)

6. Diversity of Machine learning (ML) models to solve a classification task. (Dadi et al., 2019; Khosla et al., 2018)

Considering all variability sources can be overwhelming due to the numerous complex factors involved at all stages of fMRI data collection and analysis. In this paper, we fixated the last step of the pipeline and used the simplest yet effective classification model – Logistic Regression (Popov et al., 2024), while focusing on benchmarking four critical data processing factors, including: 1) Data acquisition variability – by using 2 datasets acquired in different circumstances; 2) BOLD signal denoising variability – by comparing 12 denoising strategies; 3) Brain parcellation variability – by using 4 atlases; 4) Connectivity estimation variability – by comparing 4 functional connectivity measures.

In total, we evaluated 192 distinct FC pipelines, systematically varying preprocessing strategies, anatomical brain parcellation schemes, and functional connectivity measures. Notably, we adopted a cross-site validation approach, training ML models on data from one research site and testing on another. This approach ensures that our findings reflect true methodological robustness, rather than site-specific artifacts or overfitting. Although large sample sizes are important for developing ML models with high generalizability, in fMRI studies, datasets are often limited in size due to complexity and high cost of data collection. Using two datasets, a smaller

one and a larger one, enables researchers to build predictive models that generalize effectively under different conditions, maximizing the utility of limited but valuable fMRI data.

Furthermore, recent discussions have highlighted the potential benefits of within-participant study designs for enhancing statistical power when neuroimaging data are used. (Marek et al., 2022; Spisak et al., 2023; Tervo-Clemmens et al., 2023) While our primary focus is on the pipeline robustness for state classification, we also investigate whether the structure of the training data – specifically, leveraging repeated measures from fewer subjects versus single measures from more subjects – impacts the performance of our classification models in distinguishing eye states.

Our results demonstrated that despite substantial differences in pipeline configurations, FC-based classification models sustainably achieve high accuracy (about 80% for best pipelines). This finding suggests that functional connectivity features are resilient to many forms of methodological variation when used to capture stable brain state differences. Nevertheless, we identified the most successful pipeline option that allowed us to achieve the highest classification accuracy. Interestingly, our analysis did not reveal a significant difference in model performance between these two experimental designs (within-subject versus between-subject). This suggests that in the terms of state classification (e.g., eyes-open vs. eyes-closed) both within- and between-subject setups can yield comparable predictive performance under appropriate modeling conditions. These results have important implications for the development of reproducible neuroimaging biomarkers, particularly in settings where brain states are expected to be consistent and clearly defined.

Methods

Data Acquisition Variability

Data acquisition (including scanner variability, acquisition parameters and sample bias) is a major source of variation that can significantly impact estimates of functional connectivity. For instance, variations in MR-scanner field strength can lead to differences in signal-to-noise ratio and spatial resolution, affecting the detection of functional connections. Even with identical imaging sequences and parameters, site-specific differences can still occur due to various physical factors. These include field inhomogeneity, configurations of transmit and receive coils, system stability, maintenance routines, scanner drift over time, and other potential variables (Bari et al., 2019). Moreover, we cannot deny the effect of the sample group. In particular, if we take clinical data as an example, disease factors may become confounding along with scanner-specific confounds in situations, when a site is focused on sampling specific types of disorder (Yamashita et al., 2019).

Benchmarking data acquisition variability allows us to assess how changes in acquisition protocols influence the robustness of predictive models. This is particularly crucial for multi-site studies, where harmonizing data across different scanners is an ongoing challenge. Data used in the present study was taken from 2 sources: 1) Beijing: Eyes Open Eyes Closed Study dataset (D. Liu et al., 2013); 2) Research data from the Institute of the Human Brain of the Russian Academy of Sciences (IHB RAS).

BOLD signal denoising variability

Benchmarking denoising strategies is critical for understanding their impact on functional connectivity and avoiding potential confounds in the data. Denoising strategy might either under-correct or over-correct the data, leading to either residual noise contaminating the connectivity estimates or meaningful signals being mistakenly discarded. Estimating variability across different denoising strategies helps researchers find the right balance between preserving functional signals and eliminating noise, ensuring that the connectivity matrices accurately reflect brain network dynamics.

Denoising can significantly alter the connectivity structure, and consequently affect the performance of predictive models. Some techniques are commonly used among researchers (such as regression of motion parameters), others are controversial. For instance, global signal regression (GSR), which is regression of average signal across all voxels in the brain, is among those that are often discussed. The issue is that GSR may introduce poorly explained negative correlations in FC, leading to bias in connectivity estimates (Murphy & Fox, 2017). In the current study, we checked if there were notable differences in strategies that regress global signal out and those that did not.

We also employed CompCor (principal component-based noise correction) in our analysis (Behzadi et al., 2007). Two variants are available: aCompCor, or anatomical CompCor, obtains confound regressors from the principal components of the BOLD signal in white matter (WM) and cerebrospinal fluid (CSF), and tCompCor, or temporal CompCor, takes voxels with high temporal standard deviation.

In some sense, using both CompCor and GSR is excessive, since global signal takes into account time series of WM and CSF (e.g., in the CONN Toolbox they suggest using either CompCor or

GSR. (Whitfield-Gabrieli & Nieto-Castanon, 2012)) In fact, studies that aimed to compare confound regression strategies usually did not apply both GSR and CompCor simultaneously (e.g., (Ciric et al., 2017)). In the present study, we also compared both strategies with simultaneous CompCor and GSR, and only CompCor application to explore their influence on ML classification between analyzed brain states.

In total, we evaluated 12 denoising strategies (Table 1). Being an essential part of fMRI data processing, head motion parameters (HMP) were included in all strategies. The 24 HMP strategy included 6 motion parameters, 6 temporal derivatives, 6 quadratic terms, and 6 quadratic expansions of the derivatives of motion estimates (Friston et al., 1996). Here we also used the 12 HMP strategy, which included 6 motion estimates and their temporal derivatives. We used aCompCor with different number of components: by default we took 10 principal components calculated using a WM and CSF combined anatomical mask, resulting in fixed number of regressors for every subject; and aCompCor50, which included principal components that explain 50% of variance, calculated using a WM and CSF combined anatomical mask (number of regressors varied among subjects). For tCompCor, the number of regressors also differed from subject to subject. Strategies 7-12 included 4 global signal regressors: the global signal, its derivative, its square, and the derivative of its square.

Table 1. Characteristics of used pipelines

Strategy	**Noise correction methods**	**Number of regressors**
1	24 HMP	24
2	aCompCor + 12 HMP	10 + 12
3	aCompCor50 + 12 HMP	*k* + 12
4	aCompCor + 24 HMP	10 + 24

5	aCompCor50 + 24 HMP	*k* + 24
6	tCompCor + aCompCor50 + 24 HMP	*k1* + *k2* + 24
7	24 HMP + GSR	24 + 4
8	aCompCor + 12 HMP + GSR	10 + 12 + 4
9	aCompCor50 + 12 HMP + GSR	*k* + 12 + 4
10	aCompCor + 24 HMP + GSR	10 + 24 + 4
11	aCompCor50 + 24 HMP + GSR	*k* + 24 + 4
12	tCompCor + aCompCor50 + 24 HMP + GSR	*k1* + *k2* + 24 + 4

Notes. HMP, head motion parameters. aCompCor, 10 principal components calculated using a WM and CSF combined anatomical mask. aCompCor50, principal components that explain 50% of variance, calculated using a WM and CSF combined anatomical mask. tCompCor, principal components that are calculated using masks based on temporal variance of voxels. GSR, global signal regressors. k is a varying number of regressors among subjects.

Brain parcellation variability

Parcellation refers to dividing the brain into distinct anatomical regions of interest (ROIs), each representing a node in the brain network. It is a common practice to perform dimensionality reduction, since the dimensionality of fMRI data is quite high. The choice of parcellation scheme (based on brain atlases or data-driven) can lead to significant variability in the resulting connectivity matrices. Parcellation schemes affect the spatial resolution of the brain regions, which can drastically change the connectivity estimates. For instance, using a coarse atlas with larger regions may obscure fine-grained, functionally relevant connections, while overly fine atlases can introduce noise, leading to spurious connections and higher computation costs (Khan & Shang, 2025). Understanding how different atlases influence connectivity estimation helps researchers identify the optimal parcellation for their research questions.

Overall, all methods can be divided into 2 groups: 1) Pre-defined: provide coordinates in normalized brain space and are consistent across studies; 2) Data-driven: calculated (usually

using clustering and linear decomposition algorithms) for specific dataset and cannot be used on other data.

The choice of parcellation approach also impacts the reproducibility of findings across studies. Usage of data-driven parcellations potentially complicates direct comparison across studies utilizing different datasets for parcellation (Badea et al., 2017). Since we analyze two different datasets and in order to make our research more reproducible, we utilized only predefined brain atlases. The four selected brain atlases encompass a range of parcellation strategies, spatial resolutions, and historical usage within the literature.

1. **Automated Anatomical Labeling (AAL)** (Tzourio-Mazoyer et al., 2002) consists of 116 regions encompassing both cortical and subcortical brain structures (we used the SPM12 version available in nilearn). AAL is a widely recognized anatomical brain atlas and serves as a historical standard in neuroimaging research. Its inclusion allows for direct comparison with a large body of existing literature that has utilized this foundational parcellation scheme, based primarily on macroscopic anatomical landmarks.

2. **Schaefer Atlas** (Schaefer et al., 2018) provides a parcellation based on functional connectivity derived from resting-state fMRI data. We selected the 200-region version, which focuses on cortical areas. Including this functional atlas complements the anatomical AAL parcellation by defining regions based on their functional properties and network organization, offering a different perspective on brain division based on distributed connectivity patterns.

3. **Brainnetome Atlas** (Fan et al., 2016) offers a parcellation that attempts to integrate both anatomical and functional brain connectivity information, resulting in 246 regions (210 cortical and 36 subcortical). Its inclusion allows us to examine results using a parcellation

scheme that incorporates multiple levels of information and provides a medium-level resolution covering both cortical and subcortical brain areas, balancing the purely anatomical and purely functional approaches of AAL and Schaefer, respectively.

4. **HCPex Atlas** (Huang et al., 2022) is a high-resolution volumetric parcellation derived from the surface-based Human Connectome Project multimodal atlas (HCP-MMP v1.0), extended with detailed subcortical regions, totaling 426 parcells. This brain atlas represents a more recent and detailed approach, leveraging high-quality multimodal HCP data to define regions at a finer granularity. Its inclusion allows for the exploration of effects at a higher spatial resolution and provides a parcellation based on advanced mapping techniques.

Connectivity Estimation Variability

Connectivity measures are used to quantify the functional relationships between different regions of the human brain. Variability in connectivity estimation arises from the choice of methods, such as Pearson correlation, tangent correlation, partial correlation. The selection of the appropriate connectivity measure can dramatically impact the outputs of FC-based machine learning models, as these measures capture different aspects of the brain's functional organization (Z.-Q. Liu et al., 2025).

The estimation of functional connectivity should go beyond traditional Pearson correlation, as different measures provide complementary information about human brain network organization:

1. **Pearson Correlation**. This widely-used measure captures linear relationships between time series of different brain regions, it is simple and computationally effective. However, Pearson correlation is limited in its ability to account for indirect interactions between

regions and nonlinear dependencies, which can result in oversimplification of the complex brain networks and overestimation of connectivity (Peterson et al., 2023).

2. **Tangent Correlation**. Tangent correlation extends Pearson correlation by projecting the functional connectivity matrix onto a tangent space, making it more sensitive to individual variability, while preserving commonalities across subjects. This method is especially useful when we need to estimate the differences between subjects more precisely. Moreover, it is robust to noise and inter-subject variability. In predictive modeling, tangent correlation has been shown to improve classification accuracy and sensitivity to individual differences in brain connectivity profiles, providing a more nuanced representation of brain function (Dadi et al., 2019).

3. **Partial Correlation**. Partial correlation accounts for the direct interaction between two brain regions, while controlling for the influence of other regions. This method is essential in connectivity estimation as it reveals the unique relationships between brain regions that are not mediated by other areas, which is crucial for identifying specific pathways that may contribute to behavior or clinical outcomes. Ignoring such direct connections might obscure important predictive information (Peterson et al., 2023).

4. **Regularized Partial Correlation**, or Graphical Lasso. Partial correlation estimation may be extremely overfitted to the data, consequently it is less reliable and stable to retest tests. To lower overfitting, some researchers suggest using regularized partial correlation estimation (Peterson et al., 2023). Graphical lasso is a partial correlation estimation with *L1*-regularization.

Data

IHB RAS data consists of 2 resting state sessions with eyes closed (EC), eyes open with fixation (EO-F) for 84 healthy subjects. Acquisition of MR images was performed with 3T Philips Achieva scanner. EPI scan parameters: 32 axial slices, 120 volumes, repetition time (TR) = 2500 ms, echo time (TE) = 35 ms, flip angle= 90°. Structural images acquisition parameters: 130 axial slices, slice thickness = 0.94 mm, TR = 2500 ms, TE = 3.1 ms, flip angle = 30°.

Beijing data include 48 healthy controls and each subject has 3 rs-fMRI scans. During the first scan, participants were instructed to rest with their eyes closed, whereas the 2nd and 3rd sessions were either with closed or open eyes. Sessions were recorded sequentially. The MR images were acquired using a SIEMENS TRIO 3-Tesla scanner. EPI scan parameters: 33 axial slices, 240 volumes, TR = 2000 ms, TE = 30 ms, flip angle= 90°. T1-weighted MPRAGE parameters: 128 sagittal slices, slice thickness = 1.33 mm, TR = 2530 ms, TE = 3.39 ms, flip angle = 7° (D. Liu et al., 2013).

Raw fMRI data preprocessing for both datasets was performed using fMRIPrep 23.2.0. (Esteban et al., 2019). Anatomical processing included skull stripping (for IHB RAS), spatial normalization according to MNI standard template, and tissue segmentation. For functional data, slice timing correction, motion correction, normalization and co-registration to the corresponding anatomical image were performed. Lastly, data was smoothed using a Gaussian filter with 4 mm kernel. Additionally, confounding time-series (e.g., motion parameters and physiological noise estimates) were extracted for subsequent analyses.

Functional scans tend to have areas with extremely uneven signal that is contaminated by signals of non-neural origin (e.g., breathing in the nasal sinuses area) or caused by MRI scanner characteristics. The elimination of such areas is performed with several steps. Firstly, for every

subject a new brain mask is formed from voxels, the signal level in which is not lower than some part of the global signal (we used 0.6). Then the mean mask for all subjects is formed from the new masks. To binarize the mean mask, we used a threshold to accept voxels that presented in new masks of at least 50% of all subjects.

Further, we used the mean brain mask for all subjects to get the coverage of parcels for each brain atlas. For this, we calculated the number of voxels that fall into each parcel. We rejected parcels that had less than 10% of voxels, according to the mean brain mask. Resulting number of ROI for each atlas: AAL – 106 ROIs of 116, Schaefer – 182 ROIs of 200, Brainnetome – 220 ROIs of 246, HCPex – 369 ROIs of 426.

Next, to extract the BOLD time-series, we used the nilearn[1] library for Python. For each atlas, we extracted time-series with denoising strategies described in Table 1. Confounding time-series used for denoising were calculated in fMRIPrep: framewise displacement (FD), mean relative RMS (root-mean-square) displacement, global signal, head motion parameters, aCompCor and tCompCor estimates. We also added discrete cosines transformation basis regressors to handle low-frequency signal drifts.

We did not discard any subjects due to motion artifacts. The mean framewise displacement for the IHB RAS dataset was 0.154 mm for the EO condition and 0.172 mm for the EC condition. For the Beijing dataset, the corresponding values were 0.094 mm (EO) and 0.106 mm (EC).

[1] https://nilearn.github.io/stable/index.html

Functional connectivity quality metrics

To evaluate and compare all the varieties of pipelines discussed, we used a number of metrics that assess data quality after preprocessing, residual motion artifacts, test-retest reliability and the effectiveness of the classification model.

QC-FC correlations

The Quality Control-Functional Connectivity (QC-FC) metric is a key benchmark used to assess the effectiveness of denoising strategies in resting-state fMRI studies. It provides a direct measure of the relationship between motion-related artifacts and functional connectivity, unlike other quality metrics (e.g., framewise displacement, FD) that focus on general signal quality without directly addressing functional connectivity measures (Ciric et al., 2017; Parkes et al., 2018). QC-FC is calculated by examining the correlation between motion metrics and the functional connectivity strength for each edge in the connectivity matrix across subjects (Power et al., 2015). In the present study, we used the mean relative RMS displacement as a motion metric, which refers to the average frame-to-frame motion of the brain (Jenkinson et al., 2002). High mean relative RMS displacement suggests significant motion artifacts that can corrupt results. Consequently, high QC-FC correlations indicate that motion artifacts are still persistent in the data and influence FC estimates, suggesting a need for further refinement of the denoising process.

Similarity between denoising strategies

To assess similarity between denoising strategies, we calculate Pearson correlation between functional connectivity matrices (vectorized upper triangle) derived using different denoising

strategies. In particular, for each person, we took connectivity matrices in two states and in 12 versions (a total of 24 matrices for one person). Then, within one subject, the correlation between flattened FC matrices is calculated. As a result, we get a correlation matrix, where each element represents the similarity between functional connectivity weights for two states and 12 strategies. Then we get a mean matrix across all subjects. This procedure is repeated for all means of variability separately. Using this metric, we can assess the divergence of the resulting FC weights among selected pipelines.

Reliability measure

The Intraclass Correlation Coefficient (ICC) is an established reliability measure in functional connectivity research, primarily assessing if the variability in a dataset was caused by stable between-subject differences or random and within-subject fluctuations (Shrout & Fleiss, 1979). The ICC score is especially valuable in connectivity studies as it can evaluate consistency at an individual level, which differentiates it from other metrics that often aggregate measures across subjects (Fiecas et al., 2013; Mahadevan et al., 2021).

Depending on the sources of error incorporated in the design, different forms of ICC are available. In FC studies, a common choice is the one-way random effects model ICC(1,1), which assumes subject-level variability while treating each edge independently. This approach calculates ICC based on two primary components: the between-subject mean squares (BMS), representing the variability in edge weights averaged across subjects, and the within-subject mean squares (WMS), which captures edge variability across sessions for each individual. The calculation follows the formula:

$$\mathrm{ICC} = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (k-1) \times \mathrm{WMS}},$$

where k denotes the number of repeated sessions.

One limitation of ICC in FC analyses is its sensitivity to sparse connectivity methods, which often produce numerous edges with near-zero weights for all subjects. Since BMS and WMS for these "null" edges both tend to zero, the ICC values for these edges can be biased downward, clustering around zero. Consequently, using ICC across all edges in a sparse connectivity model may underrepresent reliability. To mitigate this, researchers typically exclude edges that consistently show zero weights and instead focus on those with systematic, non-zero values across subjects. (Peterson et al., 2023) This approach avoids the penalty of sparse models, enabling a more accurate assessment of edge-level reliability, where between-subject variability is present.

Given these considerations, in our study, we applied a masking strategy to focus on edges with meaningful between-subject variability, as suggested by (Peterson et al., 2023). Specifically, we excluded edges with near-zero correlation values, as these typically reflect noise rather than true connectivity. To ensure the ICC reliably captured relevant variability, we retained only edges in the 95th percentile of the group-averaged connectivity matrix for each functional connectivity measure. This approach minimized the influence of null edges, providing a more accurate representation of reliability across meaningful connections in our FC analysis.

Assessment of classification performance

We used logistic regression (LR) to assess classification performance of chosen pipelines. LR is a statistical method commonly used for binary classification tasks, where the goal is to predict one of two possible outcomes. LR estimates the probability that a given input belongs to a particular category by fitting data to a logistic function, producing outputs between 0 and 1. Unlike linear regression, which predicts continuous values, logistic regression uses the logistic (or sigmoid) function to map any real-valued number into a probability. This model is widely used in

fields such as medicine and social sciences for problems like disease prediction and risk assessment, because it is powerful yet quite simple and explainable, which is important in high-risk fields.

In this study, we suppose that the classification performance of machine learning models trained on a given preprocessing pipeline can serve as an indicator of whether the data have been sufficiently preprocessed to enable the model to capture meaningful data patterns. Classification metrics provide quantitative measures for evaluating ML model performance. Among these, accuracy, which is defined as the proportion of correct predictions out of the total number of predictions, is commonly used, particularly when class distributions are balanced.

With relatively small datasets prone to overfitting, our evaluation relied on a cross-site validation approach that incorporated data from both sites during training to improve generalization. Specifically, during one iteration of cross-validation (15 cross-validation iterations in total), we trained LR on the Beijing dataset plus a subset of the IHB RAS dataset . For training subset we took 92 samples from the Beijing dataset, consisting of 46 EO and 46 EC (we discarded second EC session to balance classes), and 50 samples of both states from the IHB RAS dataset (25 subjects were chosen randomly on every cross-validation fold). The generalization performance was then assessed on the remaining 118 samples (59 subjects) from the IHB RAS dataset. This approach ensures that models are not merely memorizing site-specific patterns, thereby reducing overfitting, and improves the reliability of classification metrics by utilizing a larger test set from the “target” site for evaluation.

Impact of intra- and inter-subject information during ML model training

To examine the impact of training data structure, following ongoing discussions about the advantages of within-subject designs for enhancing statistical power (Marek et al., 2022), we

evaluated two approaches to constructing our training sets, using the IHB RAS dataset for training and the Beijing dataset for testing. This “inverted” testing procedure allowed us to examine the impact of leveraging repeated measures (intra-subject information) versus unique subject exposures (inter-subject information) on classification performance.

Repeated Measures Design (Emphasizing Intra-Subject Data): the training set was constructed using data from a subset of 42 subjects from the IHB RAS dataset, with both EO and EC states included for each subject (totaling 84 scans from 42 individuals). 15-fold cross-validation was performed by selecting a new set of 42 subjects on each fold.

Unique Exposures Design (Emphasizing Inter-Subject Data): the training set utilized data from all 84 IHB RAS subjects, but each subject contributed only one state (either EO or EC, balanced across the group; totaling 84 scans from 84 individuals). For cross-validation (15 folds), the assignment of subjects to EO or EC within the training fold was varied.

For both designs, some samples from the Beijing dataset (e.g., 5 subjects, 10 scans, structured equivalently to the respective IHB RAS training design) were added to the training set to facilitate the cross-site model application, with the remainder of the Beijing dataset used for testing.

Results

QC-FC correlations

In the present study, we calculated QC-FC as the Pearson correlation (absolute value) between subjects’ edge weights (edge in the connectivity matrix) and mean relative RMS displacement, as it was described by (Ciric et al., 2017; Peterson et al., 2023). Then we took the mean edge value across subjects to get the overall view on the influence of motion artifacts. The results of the

comparison between the processing strategies analyzed with rs-fMRI data and the FC calculation strategies including both GSR and its absence are presented in Figure 2 for the EC state, since there was no difference in QC-FC scores between the EC and EO states.

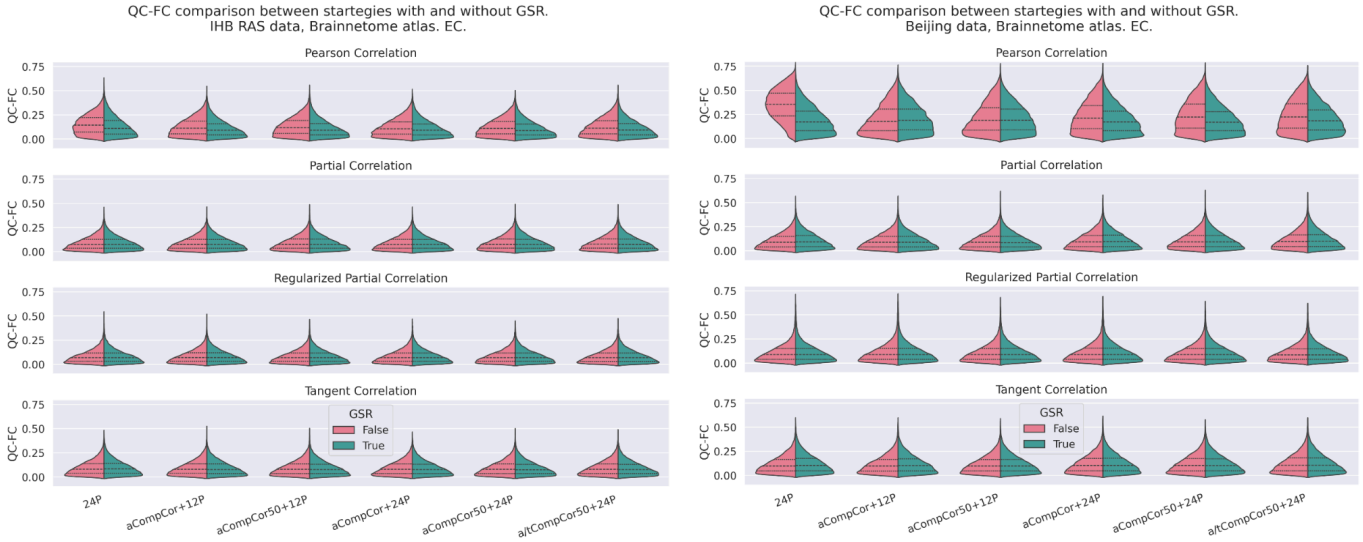

Figure 2
Comparison of QC-FC scores between strategies with (green) and without GSR (pink) for closed eyes. IHB RAS data (left), Beijing data (right). Every violin plot shows the distribution of QC-FC scores for every FC measure (rows) and strategies (columns). The central line within each violin plot represents the median. The upper and lower lines indicate the 25th percentile and the 75th percentile.

Judging by the plot, the difference in QC-FC between strategies is not clearly pronounced, except for Pearson correlation. The median QC-FC for the IHB RAS data with Pearson correlation was 0.12 without GSR, and it was reduced to 0.09 after GSR application . Pearson correlation scores for the Beijing data were 0.23 without GSR and reduced to 0.18 after GSR. It should be noted that the worst results were revealed for Pearson correlation, when only 24 head motion parameters were included for confound regression. These results are in line with previous findings showing Pearson correlation to be the worst in terms of QC-FC scores (Peterson et al., 2023).

Both partial and regularized partial correlation showed better results. The median QC-FC for partial correlation was 0.092 for the Beijing dataset and 0.076 for the IHB RAS dataset. In case of

regularized partial correlation, the median QC-FC was 0.09 for Beijing and 0.07 for the IHB RAS dataset. Speaking of tangent correlation, the median QC-FC for Beijing was 0.1 and 0.08 for the IHB RAS datasets.

In summary, the residual motion in the FC matrices was more pronounced for the Beijing dataset, even though the mean FD was lower for this data and all denoising steps were made similarly for both datasets.

Similarity between denoising strategies

As mentioned earlier, the similarity between denoising strategies was measured as the correlation between FC vectors of different denoising strategies and brain states. As a result, we got a matrix, where each entry is a mean similarity value across subjects. In Figure 3, an example for the IHB RAS dataset is presented for the Brainnetome atlas and all four FC measures.

From the results obtained, it follows that the difference between the EO and EC states in the case of partial correlation was less pronounced than in other connectivity measures, which can affect classification performance. In addition, a partial correlation showed high similarity between denoising strategies within each state. Thus, when using partial correlation as a connectivity measure, the choice of denoising strategy would not have a significant impact. Speaking about Pearson correlation, it can be seen in Figure 3.A that the strategies with GSR are correlated more between each other, than with their non-GSR pairs. The same pattern is not as pronounced for the other measures.

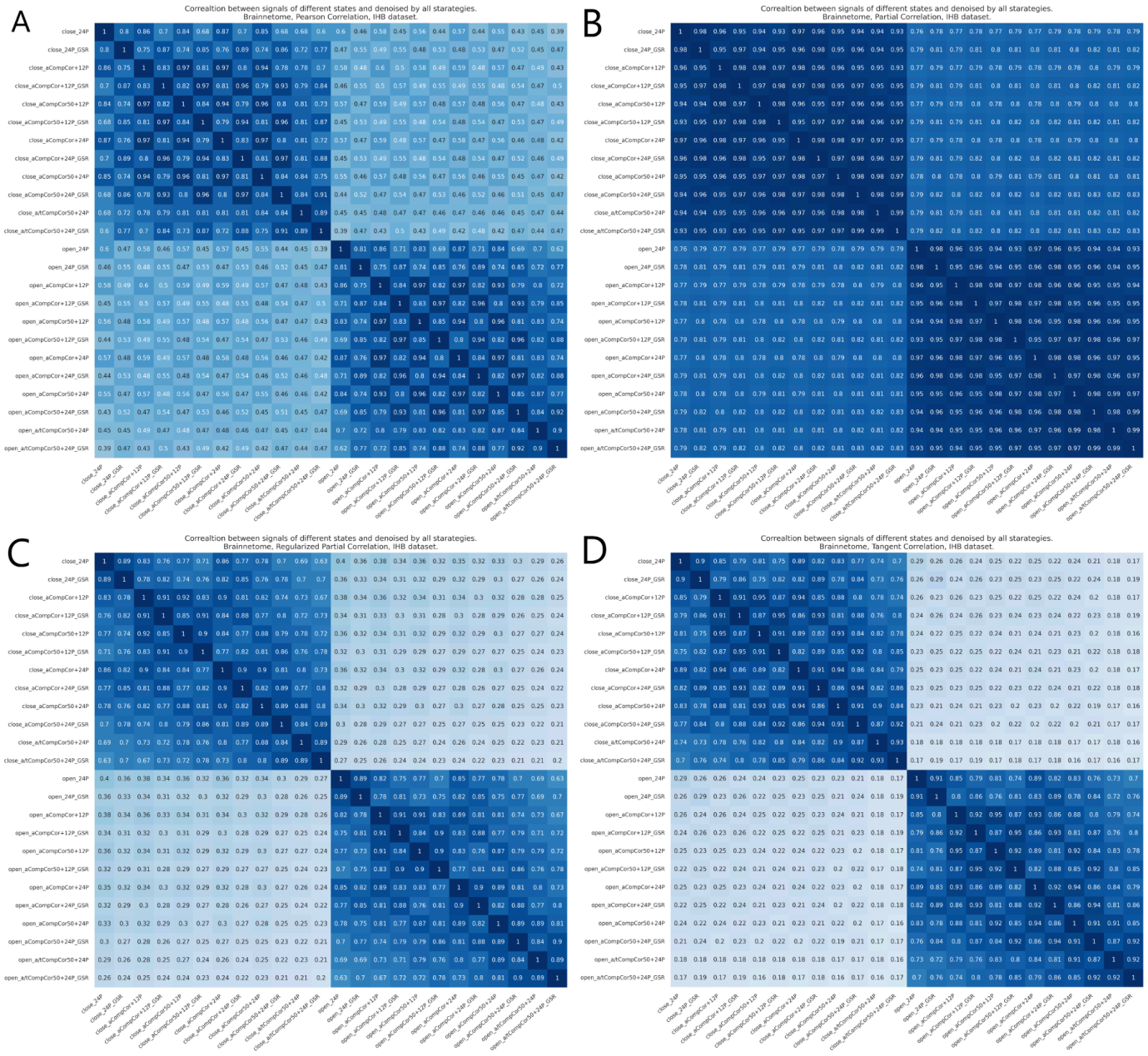

Figure 3

Similarity between signals of both EO and EC states denoised using different strategies for the IHB RAS data, the Brainnetome atlas. Each heatmap shows similarity for one FC measure. The color range is from white (0, low similarity) to dark blue (1, high similarity). Mean values across subjects are presented.

Intraclass Correlation Coefficient

Since ICC should be calculated using a dataset that has both multiple scanning sessions and repeated conditions (scanning with EC occurred twice), we calculated it only for the Beijing

dataset. The Beijing data had 3 sequentially recorded sessions: first session was only with closed eyes, second and third were alternately closed or open. Therefore, we calculated ICC between first session EC and second session EC, first session EC and third session EC.

Given that (Peterson et al., 2023) hypothesized that the average ICC scores for sparse FC measures tend to zero, because most values of such FC matrices are close to zero and vary from subject to subject only due to noise, we checked if there were differences in ICC scores for pipelines with masking (thresholding) of near-zero entries and without it. The comparison of ICC scores between the masked and unmasked FC matrices is shown in Figure 4.

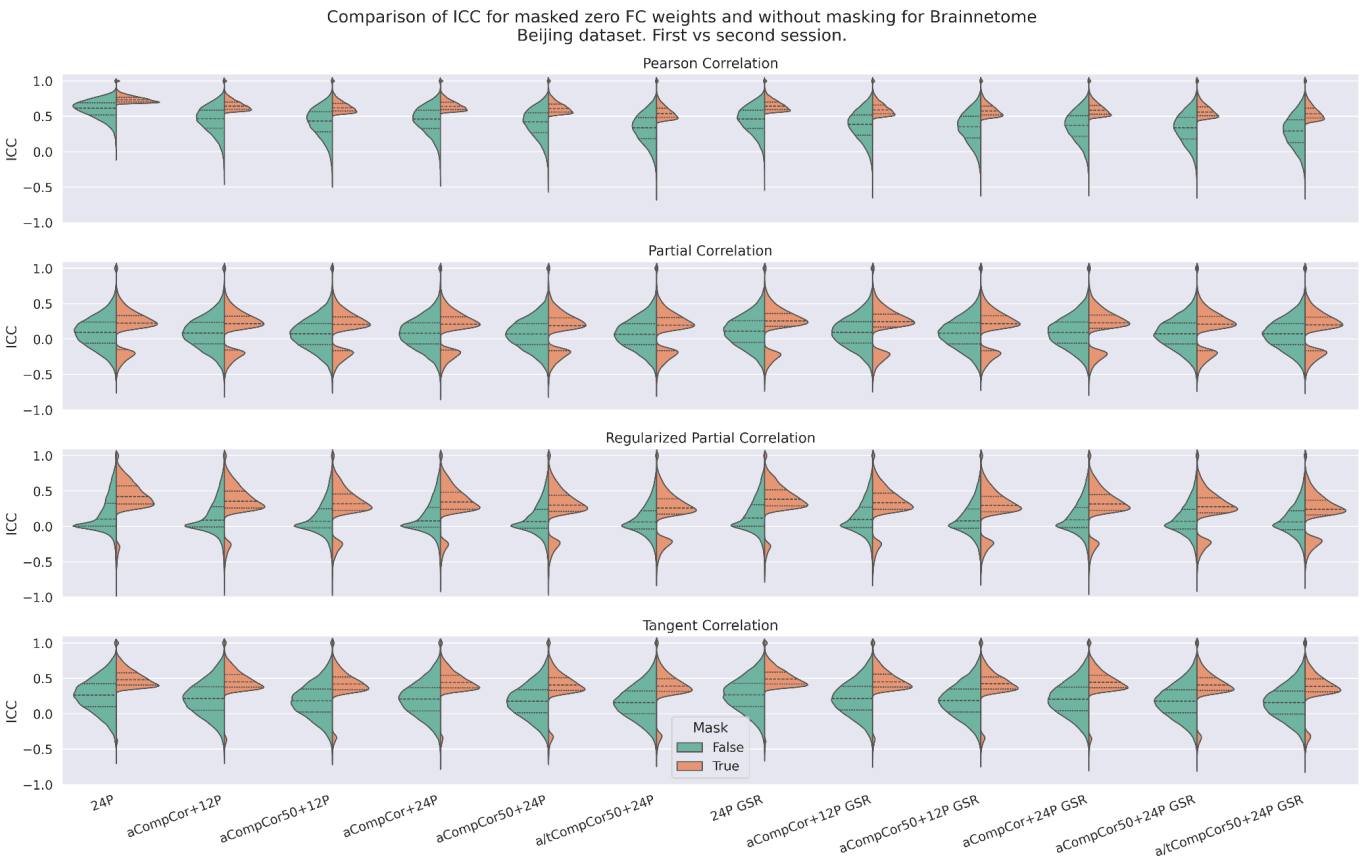

Figure 4
Comparison of ICC scores with (orange) and without masking (green) of near-zero FC weights. The Beijing data, EC first and second session. The Brainnetome atlas taken for visualization. Every violin plot shows the distribution of ICC scores for every FC measure (rows) and strategies (columns). The central line within each violin plot represents the median. The upper and lower lines indicate the 25th percentile and the 75th percentile.

There is no need in applying masking onto Pearson correlation, but we give it for comparison.

It should be noted that even after applying a masking procedure, a great number of ICC scores remained negative. Based on the ICC formula used in our analysis, negative values occur when the variability within the subject exceeds the variability between the subjects. Compared to tangent correlation, which yielded fewer negative ICC scores after masking, both partial and regularized partial correlation appeared less effective in capturing within-subject variability and more susceptible to noise.

Selecting best pipeline for brain state classification

As previously noted, we used a multi-site dataset to train a logistic regression model for a binary classification task – eyes closed versus eyes open with balanced class distributions. The results are presented in Figure 5. During cross-validation, the IHB RAS samples were randomly assigned to the training or testing set in each fold.

As clearly seen in Figure 5, the classification accuracy strongly depends on the functional connectivity measure. Thus, partial and regularized partial correlation show the worst classification accuracy for all atlases and strategies (0.62±0.04 for partial correlation, 0.64±0.06 for regularized partial correlation, mean accuracy score among all strategies and atlases). Although for both regularized partial correlation and tangent correlation, the difference between EO and EC states was clear (see Figure 3.C and Figure 3.D: similarity scores within EC state are much higher, than similarity scores between EC and EO), tangent correlation showed higher accuracy scores for all possible variants of the pipeline (0.77±0.04 , mean accuracy score among all strategies and atlases).

For Pearson correlation, the addition of GSR was associated with changes in classification accuracy. However, this effect was not consistent between atlases. Specifically, adding GSR improved accuracy for the AAL (0.82±0.01 vs. 0.70±0.04, GSR vs. no-GSR, respectively – here and further), Brainnetome (0.82±0.01 vs. 0.74±0.04) and HCPex (0.81±0.02 vs. 0.77±0.02) atlases, and reduced it for Schaefer200 (0.79±0.03 vs. 0.86±0.02). A similar pattern was observed in Figure 3.A, where signal similarity was higher for GSR-denoised data compared to their counterparts without GSR.

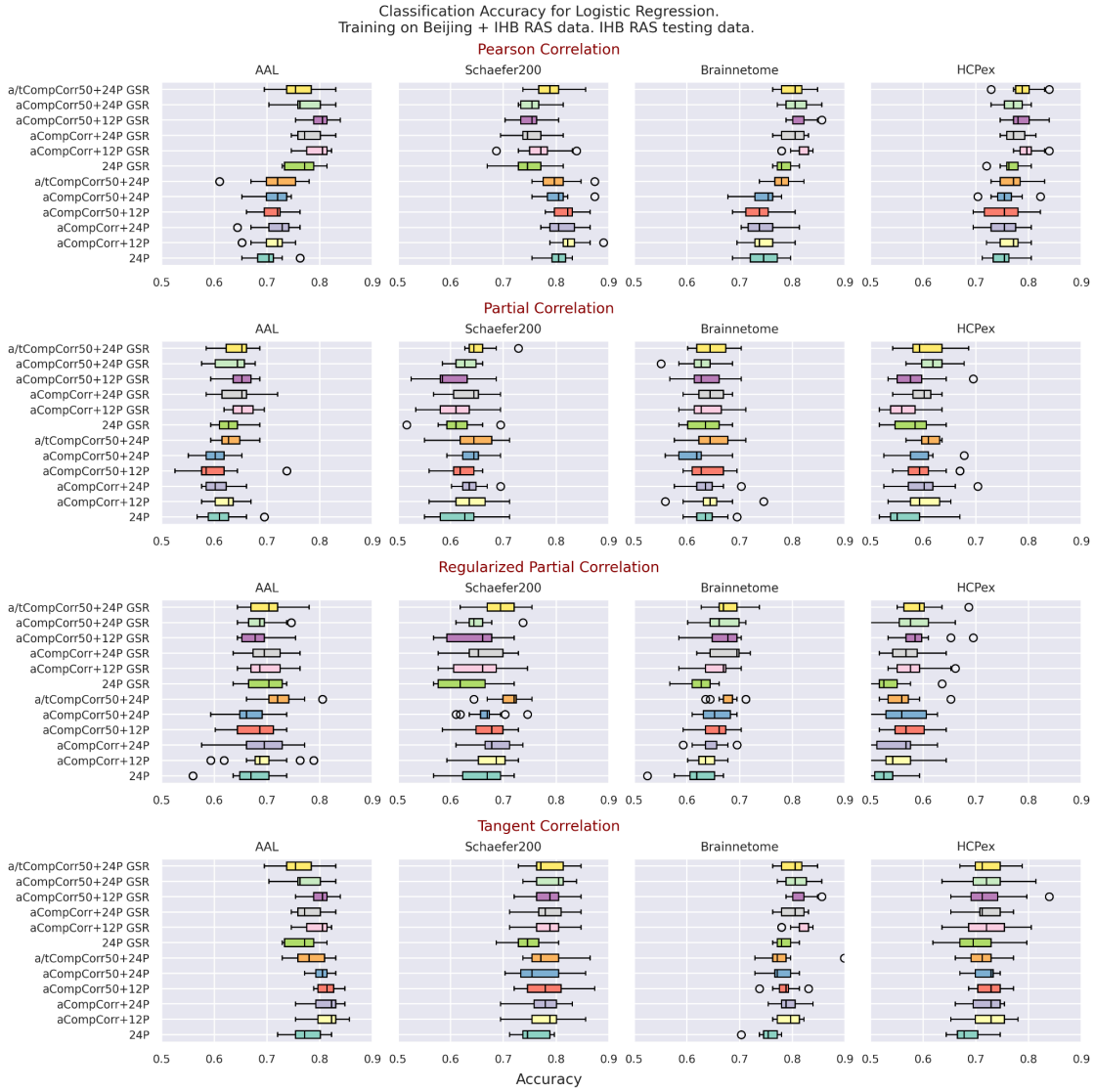

Figure 5

Classification accuracy for logistic regression models using different connectivity estimation methods and parcellations. The models were trained using the Beijing data and a subset of IHB RAS

data and tested on IHB RAS data. Each subplot shows accuracy distributions across 15 cross-validation folds for combinations of denoising strategies, parcellation schemes, and connectivity estimation methods. Boxplots illustrate the median, interquartile range, and outliers. Overall, performance varies depending on the choice of preprocessing, parcellation, and connectivity metric.

Regarding the impact of the atlas choice, it is important to consider that logistic regression (as well as other machine learning models) tends to perform worse with high-dimensional input data, a phenomenon commonly referred to as the “curse of dimensionality”. Although PCA was applied to reduce input dimensionality, the number of training samples remained relatively low compared to the input vector size, which may have contributed to convergence problems. Taking this into account, the average classification accuracies across all functional connectivity measures and denoising strategies were as follows: AAL – 0.713 ± 0.07, Schaefer200 – 0.714 ± 0.08, Brainnetome – 0.715 ± 0.08, and HCPex – 0.660 ± 0.09.

The number of brain regions is particularly critical for partial and regularized partial correlation methods, as the computation of partial correlation involves matrix inversion and the solution of an optimization problem. Regularized partial correlation introduces an additional regularization term, which is also based on optimization. As a result, the performance of regularized partial correlation was especially dependent on the choice of brain atlas. The average classification accuracies obtained with logistic regression were as follows: AAL – 0.69 ± 0.04, Schaefer200 – 0.67 ± 0.05, Brainnetome – 0.65 ± 0.03, and HCPex – 0.56 ± 0.04.

The high classification accuracy was demonstrated for the tangent correlation and the Pearson correlation. At the same time, the use of tangent correlation allowed us to obtain a more stable result with respect to both the noise reduction strategy and the atlas selection (Figure 6.2 shows the standard deviation of accuracy in a 15-folds cross-validation). The highest accuracy was

observed in combinations with Brainnetome and Schaefer200 atlases, in comparison to the larger HCPex atlas and the rougher AAL.

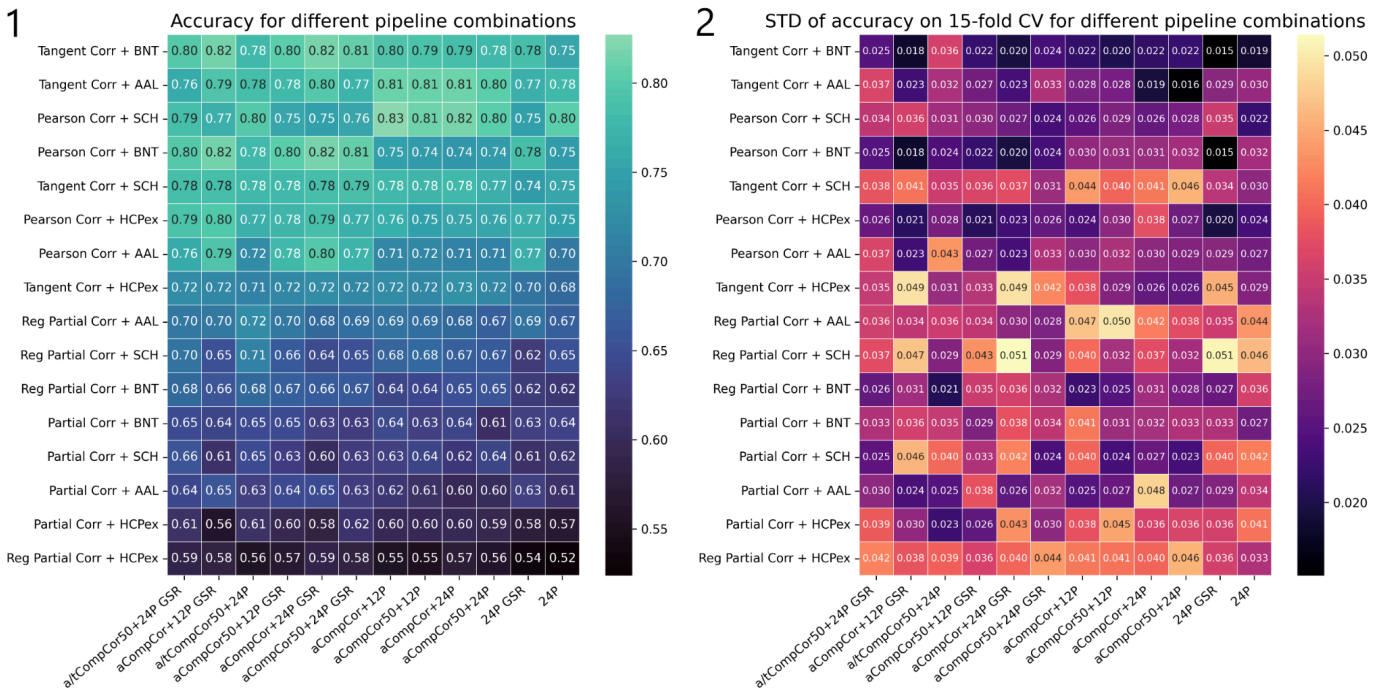

Figure 6

1) Figure shows mean accuracy on 15 CV folds for all 192 pipeline combinations.

2) STD of accuracy on 15-folds.

Both rows and columns are sorted by mean accuracy.

In summary, the best and most reproducible results were achieved with a pipeline that included the Brainnetome atlas, the tangent correlation, and the aCompCor + 12P + GSR denoising strategy with mean accuracy 0.82 ± 0.018 (mean ± SD). Analysis of denoising strategies revealed that the addition of GSR seems to positively affect classification performance.

When employing the “inverted” testing design to evaluate the influence of repeated (intra-subject) versus unique (inter-subject) measurements on model performance, we observed comparable classification accuracies across both training strategies. These findings suggest that, in the context of our classification task, the inter-subject design itself does not appear to be a limiting factor in model generalizability. In addition, during employing these experiments, we found that

the overall classification performance was better compared to the abovementioned results, even partial correlation reached 0.9 accuracy here (Figure 7).

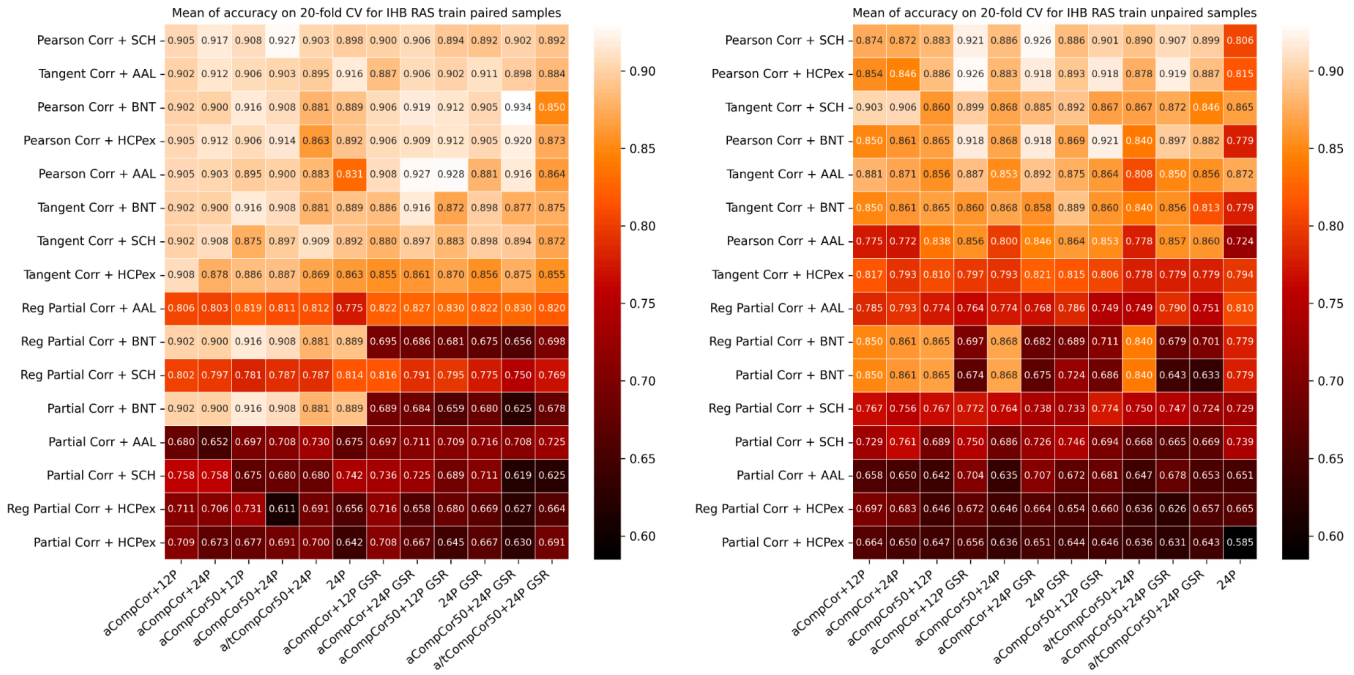

Figure 7
Mean accuracy on 20-fold cross validation for all 192 pipeline combinations.
Paired and unpaired training data (IHB RAS).
Rows and columns are sorted by mean accuracy.

Discussions

In the present study, we systematically benchmarked 192 different rs-fMRI processing pipelines on an eyes-open versus eyes-closed classification task, leveraging a cross-site validation approach. Our primary finding is the remarkable consistency of high classification accuracy across a wide array of pipeline configurations and between two independently acquired datasets. This suggests that to differentiate physiologically distinct human brain states associated with eye condition, the core information embedded in the rs-fMRI FC patterns is largely robust to considerable methodological variability.

The choice of an EO/EC paradigm was strategic, as it represents a well-defined physiological manipulation known to induce large-scale changes in brain activity and functional connectivity (Agcaoglu et al., 2019; Han et al., 2023; Weng et al., 2020; Yan et al., 2009). The pronounced nature of these state differences likely contributes to the observed robustness, as the strong signal-to-noise ratio of the underlying biological effect may overshadow more subtle variations introduced by different processing approaches. This contrasts with studies involving more nuanced cognitive tasks or heterogeneous clinical populations, where pipeline choices might exert a more critical influence on outcomes.

Despite the overall high accuracy, our results revealed performance differences attributable to specific methodological choices. Among the compared functional connectivity measures, the most superior and stable classification performance was revealed for the tangent correlation and Pearson correlations. Tangent correlation, in particular, demonstrated high accuracy and reliability with low variability between brain atlases and denoising strategies compared, aligning with previous reports of its effectiveness in brain-state prediction (Dadi et al., 2019; Ng et al., 2017) and fingerprinting (Abbas et al., 2023). However, this method is reported less frequently than the more popular partial correlation or Pearson correlations (about 9000 papers in the last 5 years that mention Pearson correlation as their method versus about 1100 mentioning tangent correlation, based on a search on Semantic Scholar).

In the present study, partial correlation and regularized partial correlation consistently yielded lower classification accuracies. This poorer performance could be attributed to several factors, including: estimation instability and “curse of dimensionality”, especially with a higher number of ROIs compared to samples; sensitivity to noise, which might be amplified when conditioning out other regions; and the regularization step potentially being too aggressive or difficult to optimize. On the other hand, both partial and regularized partial correlation showed better reliability,

compared to Pearson correlation, which does not contradict conclusions by (Peterson et al., 2023). Nevertheless, the computational complexity associated with these partial correlation methods, in addition to their suboptimal performance in this specific classification task, suggests that simpler correlation measures might be more preferable for similar high-contrast state classifications.

We observed that brain parcellations like Brainnetome and Schaefer200 atlases often led to slightly higher and more stable accuracies. These atlases represent a balance, with Brainnetome integrating anatomical and functional information (Fan et al., 2016) and Schaefer200 being functionally defined (Schaefer et al., 2018), both at resolutions that may optimally capture large-scale network reconfigurations between the EO and EC states, with functional atlases of finer granularity (especially Schaefer) generally outperforming others in reproducibility and classification performance. The anatomically-based AAL might be too coarse for certain nuances, whereas the high-resolution HCPex atlas could potentially introduce more noise or computational burden for the logistic regression model, especially with connectivity measures sensitive to dimensionality. These results align with previous findings, where the authors discussed the impact of brain atlas size on the results of the study (Khan & Shang, 2025; Wu et al., 2025). The authors agreed that more detailed brain atlases provided better connectivity features, but required more computational resources. Accordingly, coarser brain atlases showed less accurate results and did not require long computations. In this sense, both Schaefer200 and Brainnetome are an optimal solution for this particular task.

Denoising strategies based on the inclusion of GSR mainly had a positive impact on classification accuracy. This might be because GSR can help normalize signal intensity between subjects and sessions. SSR centers FC coefficients around zero, and effectively reduces widespread physiological artifacts (Golestani & Chen, 2024; Murphy & Fox, 2017). Some authors report that

GSR has little or no effect on classification performance (Shi et al., 2021; Sotero et al., 2023). Although the QC-FC correlations (Figure 2) indicated that some motion-related artifacts persisted, particularly with Pearson correlation and more basic denoising (e.g., 24 HMP), this did not dramatically affect the ability to classify EO/EC brain states.

Cross-site validation demonstrated generalizability between the datasets used. Models trained on data from one site (plus a fraction of the second) successfully generalized to unseen data from the other site. Intriguingly, models trained predominantly on the IHB RAS dataset and tested on the Beijing dataset demonstrated even higher overall performance compared to training on the Beijing data and testing on the IHB RAS data, with even partial correlation achieving high accuracy in some configurations. This asymmetry could be due to various factors, including differences in sample sizes available for the primary training pool, intrinsic data quality variations not fully mitigated by raw data preprocessing, or subtle scanner-specific signal characteristics that might make one dataset a more “universal” training set for this particular task. For example, Beijing data were acquired specifically for the EO/EC task, while IHB RAS data were collected from various task-based studies).

Limitations: The primary limitation of the present study is the high contrast nature of the EO/EC states. The robustness observed here may not be extended to experimental tasks with subtler neural signatures or studies aiming to detect small effect sizes in clinical cohorts. Secondly, we employed a single classification algorithm (Logistic Regression); while chosen for its simplicity and interpretability, other more complex models might interact differently with pipeline choices. Furthermore, due to the enormous number of possible pipeline combinations, we tested only a small part of the possible options, not including other popular denoising methods, (e.g., ICA-AROMA, scrubbing) (Ciric et al., 2017), parcellation methods (other brain atlases and data-driven parcellations) (Wu et al., 2025), and FC measures (Z.-Q. Liu et al., 2025). Finally, our

study focused on healthy participants and pipeline effects in patient populations could differ substantially from healthy cohort.

Implications and Future Directions: Our findings bolster confidence in the use of FC-based classification for brain states that are physiologically well distinguished and reproducible. The resilience to pipeline variability and cross-site differences is encouraging for multi-site studies and biomarker development, suggesting that meaningful signals can be extracted even without perfect pipeline harmonization, especially for strong biological effects. Future research should extend this benchmarking approach to more complex cognitive paradigms, clinical data, and investigate the interaction between pipeline choices and more sophisticated machine learning architectures. Further exploration into the asymmetrical cross-site generalization performance could also yield insights into optimizing multi-site study designs.

Conclusion

This study systematically benchmarked the impact of 192 distinct rs-fMRI processing pipelines on the classification of resting states with open eyes versus closed eyes, utilizing a rigorous cross-site validation strategy. Our central finding is the remarkable robustness of FC-based predictive models: despite substantial variability in BOLD signal denoising strategies, brain parcellation schemes, and connectivity estimation metrics, high classification accuracy was consistently achieved across both acquisition sites. Our results demonstrate successful cross-site generalization, highlighting the potential of functional connectivity measures as robust markers even when data are collected in different settings.

In summary, our study shows that FC-based classification reliably differentiates fundamental brain states in resting conditions with open versus closed eyes in diverse processing pipelines

and acquisition settings. Despite the robustness of the results regarding the pipelines used, we found that Pearson correlation is deservedly considered the most popular FC measure, but its effectiveness strongly depends on the quality of the data denoising. In this sense, tangent correlation can be a good compromise, combining stability to noise and relatively low computational complexity. Speaking of the brain atlas choice, the number of ROIs should be correlated with the amount of training data. In our case, the best performing brain atlases are Schaefer200 and Brainnetome with 144 samples in the training dataset.

Research funding

This study was performed within the state assignment of the Ministry of Education and Science of Russian Federation (IHB RAS budget project № 125030703331-8)

References

Abbas, K., Liu, M., Wang, M., Duong-Tran, D., Tipnis, U., Amico, E., Kaplan, A. D., Dzemidzic, M., Kareken, D., Ances, B. M., Harezlak, J., & Goñi, J. (2023). Tangent functional connectomes uncover more unique phenotypic traits. *iScience*, *26*(9), 107624. https://doi.org/10.1016/j.isci.2023.107624

Agcaoglu, O., Wilson, T. W., Wang, Y.-P., Stephen, J., & Calhoun, V. D. (2019). Resting state connectivity differences in eyes open versus eyes closed conditions. *Human Brain Mapping*, *40*(8), 2488–2498. https://doi.org/10.1002/hbm.24539

Agcaoglu, O., Wilson, T. W., Wang, Y.-P., Stephen, J. M., & Calhoun, V. D. (2020). Dynamic Resting-State Connectivity Differences in Eyes Open Versus Eyes Closed Conditions. *Brain Connectivity*, *10*(9), 504–519. https://doi.org/10.1089/brain.2020.0768

Arslan, S., Ktena, S. I., Makropoulos, A., Robinson, E. C., Rueckert, D., & Parisot, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral

cortex. *NeuroImage*, *170*, 5–30. https://doi.org/10.1016/j.neuroimage.2017.04.014

Badea, L., Onu, M., Wu, T., Roceanu, A., & Bajenaru, O. (2017). Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLOS ONE*, *12*(11), e0188196. https://doi.org/10.1371/journal.pone.0188196

Bari, S., Amico, E., Vike, N., Talavage, T. M., & Goñi, J. (2019). Uncovering multi-site identifiability based on resting-state functional connectomes. *NeuroImage*, *202*, 115967. https://doi.org/10.1016/j.neuroimage.2019.06.045

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*(1), 527–570. https://doi.org/10.1007/BF01797193

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Costumero, V., Bueichekú, E., Adrián-Ventura, J., & Ávila, C. (2020). Opening or closing eyes at rest modulates the functional connectivity of V1 with default and salience networks. *Scientific Reports*, *10*, 9137. https://doi.org/10.1038/s41598-020-66100-y

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, *192*, 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062

DeRamus, T., Faghiri, A., Iraji, A., Agcaoglu, O., Vergara, V., Fu, Z., Silva, R., Gazula, H., Stephen, J., Wilson, T. W., Wang, Y.-P., & Calhoun, V. (2021). Modular and state-relevant functional network connectivity in high-frequency eyes open vs eyes closed resting fMRI data. *Journal of Neuroscience Methods*, *358*, 109202. https://doi.org/10.1016/j.jneumeth.2021.109202

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., & Saeed, F. (2019). ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Frontiers in Neuroinformatics*, *13*. https://www.frontiersin.org/articles/10.3389/fninf.2019.00070

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A., & Feng, D. (2013). Quantifying temporal correlations: A test–retest evaluation of functional connectivity in resting-state fMRI. *NeuroImage*, *65*, 231–241. https://doi.org/10.1016/j.neuroimage.2012.09.052

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., & Turner, R. (1996). Movement-Related effects in fMRI time-series. *Magnetic Resonance in Medicine*, *35*(3), 346–355. https://doi.org/10.1002/mrm.1910350312

Golestani, A. M., & Chen, J. J. (2024). Comparing data-driven physiological denoising approaches for resting-state fMRI: Implications for the study of aging. *Frontiers in Neuroscience*, *18*. https://doi.org/10.3389/fnins.2024.1223230

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience*, *2*(10), Article 10. https://doi.org/10.1038/35094500

Han, J., Zhou, L., Wu, H., Huang, Y., Qiu, M., Huang, L., Lee, C., Lane, T. J., & Qin, P. (2023). Eyes-Open and Eyes-Closed Resting State Network Connectivity Differences. *Brain Sciences*, *13*(1), Article 1. https://doi.org/10.3390/brainsci13010122

Huang, C.-C., Rolls, E. T., Feng, J., & Lin, C.-P. (2022). An extended Human Connectome Project multimodal parcellation atlas of the human cortex and subcortical areas. *Brain Structure and Function*, *227*(3), 763–778. https://doi.org/10.1007/s00429-021-02421-6

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Khan, N. A., & Shang, X. (2025). A short investigation of the effect of the selection of human brain atlases on the performance of ASD's classification models. *Frontiers in Neuroscience*, *19*, 1497881. https://doi.org/10.3389/fnins.2025.1497881

Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., & Sabuncu, M. R. (2018). *Machine learning in resting-state fMRI analysis*. http://arxiv.org/abs/1812.11477

Li, X., Bianchini Esper, N., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Sólon Heinsfeld, A., Adebimpe, A., Vogelstein, J. T., Yan, C.-G., Esteban, O., Poldrack, R. A., Craddock, C., Fair, D., Satterthwaite, T., Kiar, G., & Milham, M. P. (2024). Moving beyond processing- and analysis-related variation in resting-state functional brain imaging. *Nature Human Behaviour*, *8*(10), 2003–2017. https://doi.org/10.1038/s41562-024-01942-4

Liu, D., Dong, Z., Zuo, X., Wang, J., & Zang, Y. (2013). Eyes-open/eyes-closed dataset sharing for reproducibility evaluation of resting state fMRI data analysis methods. *Neuroinformatics*, *11*(4), 469–476. https://doi.org/10.1007/s12021-013-9187-0

Liu, Z.-Q., Luppi, A. I., Hansen, J. Y., Tian, Y. E., Zalesky, A., Yeo, B. T. T., Fulcher, B. D., & Misic, B. (2025). Benchmarking methods for mapping functional connectivity in the brain. *Nature Methods*, *22*(7), 1593–1602. https://doi.org/10.1038/s41592-025-02704-4

Luppi, A. I., Gellersen, H. M., Liu, Z.-Q., Peattie, A. R. D., Manktelow, A. E., Adapa, R., Owen, A. M., Naci, L., Menon, D. K., Dimitriadis, S. I., & Stamatakis, E. A. (2024). Systematic evaluation of fMRI data-processing pipelines for consistent functional connectomics. *Nature Communications*, *15*(1), 4745. https://doi.org/10.1038/s41467-024-48781-5

Mahadevan, A. S., Tooley, U. A., Bertolero, M. A., Mackey, A. P., & Bassett, D. S. (2021). Evaluating the sensitivity of functional connectivity measures to motion artifact in resting-state fMRI data. *NeuroImage*, *241*, 118408. https://doi.org/10.1016/j.neuroimage.2021.118408

Mansour, L. S., Biase, M. A. D., Smith, R. E., Zalesky, A., & Seguin, C. (2023). *Connectomes for 40,000 UK Biobank participants: A multi-modal, multi-scale brain network resource* (p. 2023.03.10.532036). bioRxiv. https://doi.org/10.1101/2023.03.10.532036

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Marx, E., Deutschländer, A., Stephan, T., Dieterich, M., Wiesmann, M., & Brandt, T. (2004). Eyes open and eyes closed as rest conditions: Impact on brain activation patterns. *NeuroImage*, *21*(4), 1818–1824. https://doi.org/10.1016/j.neuroimage.2003.12.026

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering Minds: The Default Network and Stimulus-Independent Thought. *Science*, *315*(5810), 393–395. https://doi.org/10.1126/science.1131295

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, *15*(4), 869–877. https://doi.org/10.1016/j.nic.2005.09.008

Murphy, K., Birn, R. M., & Bandettini, P. A. (2013). Resting-state fMRI confounds and cleanup. *NeuroImage*, *80*, 349–359. https://doi.org/10.1016/j.neuroimage.2013.04.001

Murphy, K., & Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage*, *154*, 169–173. https://doi.org/10.1016/j.neuroimage.2016.11.052

Ng, B., Varoquaux, G., Poline, J. B., Thirion, B., Greicius, M. D., & Poston, K. L. (2017). Distinct alterations in Parkinson's medication-state and disease-state connectivity. *NeuroImage: Clinical*, *16*, 575–585. https://doi.org/10.1016/j.nicl.2017.09.004

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, *203*, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157

Parkes, L., Fulcher, B., Yücel, M., & Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, *171*, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Patriat, R., Molloy, E. K., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., Prabhakaran, V., & Birn, R. M. (2013). The effect of resting condition on resting-state fMRI reliability and consistency: A comparison between resting with eyes open, closed, and fixated. *NeuroImage*, *78*, 463–473. https://doi.org/10.1016/j.neuroimage.2013.04.013

Peterson, K. L., Sanchez-Romero, R., Mill, R. D., & Cole, M. W. (2023). *Regularized partial correlation provides reliable functional connectivity estimates while correcting for widespread confounding* (p. 2023.09.16.558065). bioRxiv. https://doi.org/10.1101/2023.09.16.558065

Popov, P., Mahmood, U., Fu, Z., Yang, C., Calhoun, V., & Plis, S. (2024). A simple but tough-to-beat baseline for fMRI time-series classification. *NeuroImage*, *303*, 120909. https://doi.org/10.1016/j.neuroimage.2024.120909

Power, J. D., Schlaggar, B. L., & Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage*, *0*, 536–551. https://doi.org/10.1016/j.neuroimage.2014.10.044

Qin, K., Lei, D., Pinaya, W. H. L., Pan, N., Li, W., Zhu, Z., Sweeney, J. A., Mechelli, A., & Gong, Q. (2022). Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites. *eBioMedicine*, *78*. https://doi.org/10.1016/j.ebiom.2022.103977

Riaz, A., Asad, M., Alonso, E., & Slabaugh, G. (2020). DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. *Journal of Neuroscience Methods*, *335*, 108506. https://doi.org/10.1016/j.jneumeth.2019.108506

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of

resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Segato, A., Marzullo, A., Calimeri, F., & De Momi, E. (2020). Artificial intelligence for brain diseases: A systematic review. *APL Bioengineering*, *4*(4), 041503. https://doi.org/10.1063/5.0011697

Shi, D., Li, Y., Zhang, H., Yao, X., Wang, S., Wang, G., & Ren, K. (2021). Machine Learning of Schizophrenia Detection with Structural and Functional Neuroimaging. *Disease Markers*, *2021*, 1–12. https://doi.org/10.1155/2021/9963824

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Sotero, R. C., Sanchez-Bornot, J. M., Shaharabi-Farahani, I., & Iturria-Medina, Y. (2023). Examining the Impact of fMRI Preprocessing Steps on Machine Learning-Based Classification of Autism Spectrum Disorder. *2023 the 7th International Conference on Medical and Health Informatics (ICMHI)*, 19–24. https://doi.org/10.1145/3608298.3608302

Spisak, T., Bingel, U., & Wager, T. D. (2023). Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*(7951), E4–E7. https://doi.org/10.1038/s41586-023-05745-x

Tervo-Clemmens, B., Marek, S., Chauvin, R. J., Van, A. N., Kay, B. P., Laumann, T. O., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., & Dosenbach, N. U. F. (2023). Reply to: Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*(7951), E8–E12. https://doi.org/10.1038/s41586-023-05746-w

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*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn

HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Weng, Y., Liu, X., Hu, H., Huang, H., Zheng, S., Chen, Q., Song, J., Cao, B., Wang, J., Wang, S., & Huang, R. (2020). Open eyes and closed eyes elicit different temporal properties of brain functional networks. *NeuroImage*, *222*, 117230. https://doi.org/10.1016/j.neuroimage.2020.117230

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, *2*(3), 125–141. https://doi.org/10.1089/brain.2012.0073

Wu, X., Liang, C., Bustillo, J., Kochunov, P., Wen, X., Sui, J., Jiang, R., Yang, X., Fu, Z., Zhang, D., Calhoun, V. D., & Qi, S. (2025). The Impact of Atlas Parcellation on Functional Connectivity Analysis Across Six Psychiatric Disorders. *Human Brain Mapping*, *46*(5), e70206. https://doi.org/10.1002/hbm.70206

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology*, *17*(4), e3000042. https://doi.org/10.1371/journal.pbio.3000042

Yan, C., Liu, D., He, Y., Zou, Q., Zhu, C., Zuo, X., Long, X., & Zang, Y. (2009). Spontaneous Brain Activity in the Default Mode Network Is Sensitive to Different Resting-State Conditions with Limited Cognitive Load. *PLOS ONE*, *4*(5), e5743. https://doi.org/10.1371/journal.pone.0005743