Deriving robust biomarkers from multi-site resting-state data: An Autism-based example

Alexandre Abraham[a,b,*], Michael Milham[e,f], Adriana Di Martino[g], R. Cameron Craddock[e,f], Dimitris Samaras[c,d], Bertrand Thirion[a,b], Gael Varoquaux[a,b]

[a] *Parietal Team, INRIA Saclay-Île-de-France, Saclay, France*
[b] *CEA, DSV, I²BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[c] *Stony Brook University, NY 11794, USA*
[d] *Ecole Centrale, 92290 Châtenay Malabry, France*
[e] *Child Mind Institute, New York, USA*
[f] *Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, USA*
[g] *NYU Langone Medical Center, New York, USA*

Abstract

Resting-state functional Magnetic Resonance Imaging (R-fMRI) holds the promise to reveal functional biomarkers of neuropsychiatric disorders. However, extracting such biomarkers is challenging for complex multi-faceted neuropathologies, such as autism spectrum disorders. Large multi-site datasets increase sample sizes to compensate for this complexity, at the cost of uncontrolled heterogeneity. This heterogeneity raises new challenges, akin to those face in realistic diagnostic applications. Here, we demonstrate the feasibility of inter-site classification of neuropsychiatric status, with an application to the Autism Brain Imaging Data Exchange (ABIDE) database, a large (N=871) multi-site autism dataset. For this purpose, we investigate pipelines that extract the most predictive biomarkers from the data. These R-fMRI pipelines build participant-specific connectomes from functionally-defined brain areas. Connectomes are then compared across participants to learn patterns of connectivity that differentiate typical controls from individuals with autism. We predict this neuropsychiatric status for participants from the same acquisition sites or different, unseen, ones. Good choices of methods for the various steps of the pipeline lead to 67% prediction accuracy on the full ABIDE data, which is significantly better than previously reported results. We perform extensive validation on multiple subsets of the data defined by different inclusion criteria. These enables detailed analysis of the factors contributing to successful connectome-based prediction. First, prediction accuracy improves as we include more subjects, up to the maximum amount of subjects available. Second, the definition of functional brain areas is of paramount importance for biomarker discovery: brain areas extracted from large R-fMRI datasets outperform reference atlases in the classification tasks.

Keywords: data heterogeneity, resting-state fMRI, data pipelines, biomarkers, connectome, autism spectrum disorders

1. Introduction

In psychiatry, as in other fields of medicine, both the standardized observation of signs, as well as the symptom profile are critical for diagnosis. However, compared to other fields of medicine, psychiatry lacks accompanying objective markers that could lead to more refined diagnoses and targeted treatment [1]. Advances in non-invasive brain imaging techniques and analyses (*e.g.* [2, 3]) are showing great promise for uncovering patterns of brain structure and function that can be used as objective measures of mental illness. Such *neurophenotypes* are important for clinical applications such as disease staging, determination of risk prognosis, prediction and monitoring of treatment response, and aid towards diagnosis (*e.g.* [4]).

Among the many imaging techniques available, resting-state fMRI (R-fMRI) is a promising candidate to define functional neurophenotypes [5, 3]. In particular, it is non-invasive and, unlike conventional task-based fMRI, it does not require a constrained experimental setup nor the active and focused participation of the subject. It has been proven to capture interactions between brain regions that may lead to neuropathology diagnostic biomarkers [6]. Numerous studies have linked variations in brain functional architecture measured from R-fMRI to behavioral traits and mental health conditions such as Alzheimer disease (*e.g.* [7, 8], Schizophrenia (*e.g.* [9, 10, 11, 12]), ADHD, autism (*e.g.* [13]) and others (*e.g.* [14]). Extending these findings, predictive modeling approaches have revealed patterns of brain functional connectivity that could serve as biomarkers for classifying depression (*e.g.* [15]), ADHD (*e.g.* [16]), autism (*e.g.* [17]), and even age [18]. This growing number of studies has shown the feasibility of using R-fMRI to identify biomarkers. However questions

*Corresponding author
Email address: abraham.alexandre@gmail.com (Alexandre Abraham)

about the readiness of R-fMRI to detect clinically useful biomarkers remain [13]. In particular, the reproducibility and generalizability of these approaches in research or clinical settings are debatable. Given the modest sample size of most R-fMRI studies, the effect of cross-study differences in data acquisition, image processing, and sampling strategies [19, 20, 21] has not been quantified.

Using larger datasets is commonly cited as a solution to challenges in reproducibility and statistical power [22]. They are considered a prerequisite to R-fMRI-based classifiers for the detection of psychiatric illness. Recent efforts have accelerated the generation of large databases through sharing and aggregating independent data samples[23, 24, 25]. However, a number of concerns must be addressed before accepting the utility of this approach. Most notably, the many potential sources of uncontrolled variation that can exist across studies and sites, which range from MRI acquisition protocols (*e. g.* scanner type, imaging sequence, see [26]), to participant instructions (*e. g.* eyes open vs. closed, see [27]), to recruitment strategies (age-group, IQ-range, level of impairment, treatment history and acceptable comorbidities). Such variation in aggregate samples is often viewed as dissuasive, as its effect on diagnosis and biomarker extraction is unknown. It commonly motivates researchers to limit the number of sites included in their analyses at the cost of sample size.

Cross-validated results obtained from predictive models are more robust to inhomogeneities: they measure model generalizability by applying it to unseen data, *i. e.*, data not used to train the model. In particular, leave-out cross-validation strategies, which remove single individuals (or random subsets), are common in biomarkers studies. However, these strategies do not measure the effect of potential site-specific confounds. In the present study we leverage aggregated R-fMRI samples to address this problem. Instead of leaving out random subsamples as test sets, we left out entire sites to measure performance in the presence of uncontrolled variability.

Beyond challenges due to inter-site data heterogeneity, choices in the functional-connectivity data-processing pipeline further add to the variability of results [28, 27, 29]. While preprocessing procedures are now standard, the different steps of the prediction pipeline vary from one study to another. These entail specifying regions of interest, extracting regional time courses, computing connectivity between regions, and identifying connections that relate to subject's phenotypes [15, 30, 31, 32].

Lack of ground truth for brain functional architecture undermines the validation of R-fMRI data-processing pipelines. The use of functional connectivity for individual prediction suggests a natural figure of merit: prediction accuracy. We contribute quantitative evaluations, to help settling down on a parameter-free pipeline for R-fMRI. Using efficient implementations, we were able to evaluate many pipeline options and select the best method to estimate atlases, extract connectivity matrices, and predict phenotypes.

To demonstrate that pipelines to extract R-fMRI neuro-phenotypes can reliably learn inter-site biomarkers of psychiatric status on inhomogeneous data, we analyzed R-fMRI in the Autism Brain Imaging Data Exchange (ABIDE) [25]. It compiles a dataset of 1112 R-fMRI participants by gathering data from 17 different sites. After preprocessing, we selected 871 to meet quality criteria for MRI and phenotypic information. Our inter-site prediction methodology reproduced conditions found under most clinical settings, by leaving out whole sites and using them as newly seen test sets. To validate the robustness of our approach, we performed nested cross-validation and varied samples per inclusion criteria (*e. g.* sex, age). Finally, to assess losses in predictive power associated with using a heterogeneous aggregate dataset instead of uniformly defined samples, we included a comparison of intra- and inter-site prediction strategies.

2. Material and methods

A connectome is a functional-connectivity matrix between a set brain regions of interest (ROIs). We call such a set of ROIs an atlas, even though some of the methods we consider extract the regions from the data rather than relying on a reference atlas (see Figure 1). We investigate here pipelines that discriminate individuals based on the connection strength of this connectome [33], with a classifier on the edge weights [15].

Specifically, we consider connectome-extraction pipelines composed of four steps: 1) region estimation, 2) time series extraction, 3) matrix estimation, and 4) classification –see Figure 1. We investigate different options for the various steps: brain parcellation method, connectivity matrix estimation, and final classifier.

In order to measure the impact of uncontrolled variability on prediction performance, we designed a diagnosis task where the model classifies participants coming from an acquisition site not seen during training.

2.1. Datasets: Preprocessing and Quality Assurance (QA)

We ran our analyses on the ABIDE repository [25], a sample aggregated across 17 independent sites. Since there was no prior coordination between sites, the scan and diagnostic/assessment protocols vary across sites [1]. To easily replicate and extend our work, we rely on a publicly available preprocessed version of this dataset provided by the Preprocessed Connectome Project [2] initiative. We specifically used the data processed with the Configurable Pipeline for the Analysis of Connectomes [34] (C-PAC). Data were selected based on the results of quality visual inspection by three human experts who inspected for largely

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

[2] http://preprocessed-connectomes-project.github.io/abide

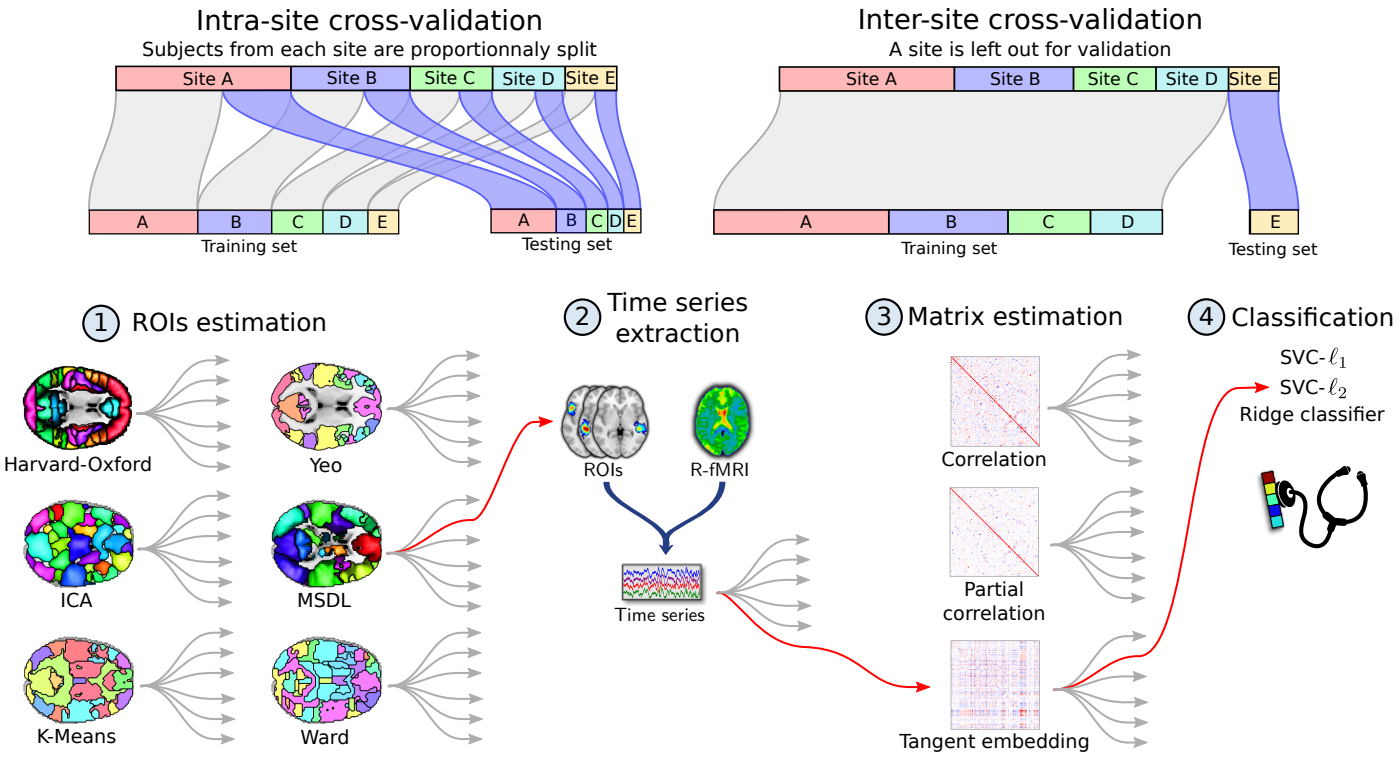

Figure 1: **Functional MRI analysis pipeline.** Cross-validation schemes used to validate the pipeline are presented above. **Intra-site** cross-validation consists of randomly splitting the participants into training and testing sets while preserving the ratio of samples for each site and condition. **Inter-site** cross-validation consists of leaving out participants from an entire site as testing set. In the first step of the pipeline, regions of interest are estimated from the training set. The second step consists of extracting signals of interest from all the participants, which are turned into connectivity features via covariance estimation at the third step. These features are used in the fourth step to perform a supervised learning task and yield an accuracy score. An example of pipeline is highlighted in red. This pipeline is the one that gives best results for inter-site prediction. Each model is decribed in the section relative to material and methods.

incomplete brain coverage, high movement peaks, ghosting and other scanner artifacts. This yielded 871 subjects out of the initial 1112.

Preprocessing included slice-timing correction, image realignment to correct for motion, and intensity normalization. Nuisance regression was performed to remove signal fluctuations induced by head motion, respiration, cardiac pulsation, and scanner drift [35, 36]. Head motion was modeled using 24 regressors derived from the parameters estimated during motion realignment [37], scanner drift was modeled using a quadratic and linear term, and physiological noise was modeled using the 5 principal components with highest variance from a decomposition of white matter and CSF voxel time series (CompCor) [38]. After regression of nuisance signals, fMRI was coregistered on the anatomical image with FSL BBreg, and the results where normalized to MNI space with the non-linear registration from ANTS [39]. Following time series extraction, data were detrended and standardized (dividing by the standard deviation in each voxel).

2.2. Step 1: Region definition

To reduce the dimensionality of the estimated connectomes, and to improve the interpretability of results, connectome nodes were defined from regions of interest (ROIs) rather than single voxels. ROIs can be either selected from a priori atlases or directly estimated from the correlation structure of the data being analyzed. Several choices for both approaches were used and compared to evaluate the impact of ROI definition on prediction accuracy. Note that the ROI were all defined on the training set to avoid possible overfitting.

We considered the following predefined atlases: **Harvard Oxford** [40], a structural atlas computed from 36 individuals' T1 images, **Yeo** [41], a functional atlas composed of 17 networks derived from clustering [42] of functional connectivity on a thousand individuals, and **Craddock** [43], a multiscale functional atlas computed using constrained spectral clustering, to study the impact of the number of regions.

We derived data-driven atlases based on four strategies. We explored two clustering methods: **K-Means**, a technique to cluster fMRI time series [44, 45], which is a top-down approach minimizing cluster-level variance; and **Ward's clustering**, which also minimizes a variance criterion using hierarchical agglomeration. Ward's algorithm admits spatial constraints at no cost and has been extensively used to learn brain parcellations [45]. We also explored two decomposition methods: **Independent component analysis (ICA)** and **multi-subject dictionary**

learning (MSDL). ICA is a widely-used method for extracting brain maps from R-fMRI [46, 47] that maximizes the statistical independence between spatial maps. We specifically used the ICA implementation of [48]. MSDL extracts a group atlas based on dictionary learning and multi-subject modeling, and employs spatial penalization to promote contiguous regions [49]. Unlike MSDL and Ward clustering, ICA and K-Means do not take into account the spatial structure of the features and are not able to enforce a particular spatial constraint on their components. We indirectly provide such structure by applying a Gaussian smoothing on the data [45] with a full-width half maximum varying from 4mm to 12mm. Differences in the number of ROIs in the various atlases present a potential confound for our comparisons. As such, connectomes were constructed using only the 84 largest ROIs from each atlas, following [49][3].

2.3. Step 2: Time-series extraction

The common procedure to extract one representative time series per ROI is to take the average of the voxel signals in each region (for non overlapping ROIs) or the weighted average of non-overlapping regions (for fuzzy maps). For example, ICA and MSDL produce overlapping fuzzy maps. Each voxel's intensity reflects the level of confidence that this voxel belongs to a specific region in the atlas. In this case, we found that extracting time courses using a spatial regression that models each volume of R-fMRI data as a mixture of spatial patterns gives better results (details in Appendix A and Figure A3). For non-overlapping ROIs, this procedure is equivalent to weighted averaging on the ROI maps.

After this extraction, we remove at the region level the same nuisance regressors as at the voxel level in the preprocessing. Signals summarizing high-variance voxels (CompCor [38]) are regressed out: the 5 first principal components of the 2% voxels with highest variance. In addition, we extract the signals of ROIs corresponding to physiological or acquisition noise[4] and regress them out to reject non-neural information [33]. Finally, we also regress out primary and secondary head motion artifacts using 24-regressors derived from the parameters estimated during motion realignment [37].

2.4. Step 3: Connectivity matrix estimation

Since scans in the ABIDE dataset do not include enough R-fMRI time points to reliably estimate the true covariance matrix for a given individual, we used the Ledoit-Wolf shrinkage estimator, a parameter-free regularized covariance estimator [50], to estimate relationships between ROI time series (details are given in Appendix B). For pipeline step 3, we studied three different connectivity measures to capture interactions between brain regions: *i)* the correlation itself, *ii)* partial correlations from the inverse covariance matrix [33, 51], and *iii)* the tangent embedding parameterization of the covariance matrix [48, 52]. We obtained one connectivity weight per pair of regions for each subject. At the group level, we ran a nuisance regression across the connectivity coefficients to remove the effects of site, age and gender.

2.5. Step 4: Supervised learning

As ABIDE represents a post-hoc aggregation of data from several different sites, the assessments used to measure autism severity vary somewhat across sites. As a consequence, autism severity scores are not directly quantitatively comparable between sites. However, the diagnosis between ASD and Typical Control (TC) is more reliable and has already been used in the context of classification [53, 13, 54].

Discriminating between ASD and TC individuals is a supervised-learning task. We use the connectivity measures between all pairs of regions, extracted in the previous step, as features to train a classifier across individuals to discriminate ASD patients from TC.

We explore different machine-learning methods[5]. We first rely on the most commonly used approach, the ℓ_2-penalized support vector classification **SVC**. Given that we expect only a few connections to be important for diagnosis, we also include the ℓ_1-penalized sparse **SVC**. Finally, we include **ridge regression**, which also uses an ℓ_2 penality but is faster than the **SVC**. For all models we use the implementation of the scikit-learn library [55][6]. We chose to work only with linear models for their interpretability. In particular, in subsection 3.7 we inspect classifier weights to identify which connections are the most important for prediction.

2.6. Exploring dataset heterogeneity

Previous studies of prediction using the ABIDE sample [13, 54] have included less than 25% of the available data, most likely to limit heterogeneity – not only due to differences in imaging, but also to factors such as age, sex, and handedness. We explored the effect of such choice on the results, by examining data subsets of different heterogeneity (see Table 1). These included: subsample #1 -

[3] For Harvard-Oxford and Yeo reference atlases, selecting 84 ROIs yields slightly better results than using the full version as shown in Figure A2. This is probably due to the fact that these additional regions are smaller and less important regions, hence they induce spurious correlations in the connectivity matrices.

[4] We identify these ROIs using an approach similar to FSL FIX (FMRIB's ICA-based Xnoiseifier). FIX extracts spatial and temporal descriptors from matrix decomposition results and uses a classifier trained on manually labeled components to determine if they correspond to noise.

[5] For the sake of simplicity, we report only the 3 main methods here. However, we also investigated random forests and Gaussian naive Bayes, that gave low prediction performance. Feature selection also led to worse results, both with univariate ANOVA screening and with sparse models such as Lasso or SVC-ℓ_1 –see Figure A4 for more details. Logistic regression gave results similar to SVC.

[6] Software versions: python 2.7, scikit-learn 0.17.1, scipy 0.14.0, numpy 1.9.0, nilearn 0.1.5.

Subsample		Card.	Sites	Selection criteria
1	All subjects	871	16	All subjects after QC
2	Biggest sites	736	11	Remove 6 smallest sites
3	Right handed males	639	14	All subjects without left handed and women
4	Right handed males, 9-18 yo	420	14	Right handed males between 9 and 18 years old
5	Right handed males, 9-18 yo, 3 sites	226	3	Young right handed males from 3 biggest sites

Table 1: **Subsets of ABIDE used in evaluation.** We explore several subsets of ABIDE with different homogeneity. *Card.* stands for the cardinality of the dataset, *i. e.* the number of subjects. More details about these subsets are presented in the appendices –Table A2.

all participants (871 individuals, 727 males and 144 females, 6 to 64 years old) is the full set of individuals that passed QA, subsample #2 - **largest sites** (736 individuals, 613 males and 123 females) 6 sites with less than 30 participants are removed from subsample #1, subsample #3 - **right handed males** (639 individuals) consists of the right-handed males from subsample #1, subsample #4 - **right handed males, 9-18 years** (420 individuals) is the restriction of subsample #3 to participants between 9 and 18 years old; subsample #5 - **right handed males, 9-18 yo, 3 sites** (226 individuals) are the individuals from subsample #4 belonging to the three largest sites (see Table A2 for more details)[7].

2.7. Experiments: prediction on heterogeneous data

ABIDE datasets come from 17 international sites with no prior coordination, which is a typical clinical application setting. In this setting, we want to: *i)* measure the performance of different prediction pipelines, *ii)* use these measures to find the best options for each pipeline step (Figure 1) and, *iii)* extract ASD biomarkers from the best pipeline.

Cross-validation. In order to measure the prediction accuracy of a pipeline, we use 10-fold cross-validation by training it exclusively on training data (including atlas estimation) and predicting on the testing set. Any hyperparameter of the various methods used inside a pipeline is set internally in a nested cross-validation.

We used two validation approaches. **Inter-site prediction** addresses the challenges associated with aggregate datasets by using whole sites as testing sets. This scenario also closely simulates the real world clinical challenge of not being able to sample data from every possible site. Note that this cross-validation can only be applied on a dataset with at least 10 acquisition sites, in order to leave out a different site in each fold. We do not apply it on subsample #5 that has been restricted to 3 sites to reduce site-specific variability. **Intra-site prediction** builds training and testing sets are as homogeneous as possible, reducing site-related variability. It is based on stratified shuffle split cross-validation, which splits participants into training and test sets while preserving the ratio of samples for each site and condition. We used 80% of the dataset for training and the remaining for testing.

Learning curves. A learning curve measures the impact of the number of training samples on the performance of a predictive model. For a given prediction task, the learning curve is constructed by varying the number of samples in the training set with a constant test set. Typically, an increasing curve indicates that the model can still gain additional useful information from more data. It is expected that the performance of a model will plateau at some point.

Summarizing prediction scores for a pipeline choice. To quantify the impact of the different options on the prediction accuracy, we select representative prediction scores for each model. Indeed, for ICA, K-Means, and MSDL, we explore several values for each parameters and thus obtain a range of prediction values. For these methods, we use the 10% best-performing pipelines for post-hoc analysis [8]. We do not rely only on the top-performing pipelines, as they may result from overfitting, nor the complete set of pipelines as some pipelines yield bad scores because of bad parametrization and may artificially lower the performance of these methods.

Impact of pipeline choices on prediction accuracy. In order to determine the best pipeline option at each step, we want to measure the impact of these choices on the prediction scores. In a full-factorial ANOVA, we then fit a linear model to explain prediction accuracy from the choice of pipeline steps, modeled as categorical variables. We report the corresponding effect size and its 95% confidence interval.

Pairwise comparison of pipeline choice. Two options can also be compared by comparing the scores between pairs of pipelines that differ only by that option using a Wilcoxon test that does not rely on Gaussian assumptions.

Extracting biomarkers. For a given pipeline, the final classifier yields a biomarker to diagnose between HC and ASD. To characterize this biomarker, we measure p-values on the classifier weights. We obtain the null distribution of the weights by permuting the prediction labels.

[7] The list of subjects used for each cross-validation fold are available at https://team.inria.fr/parietal/files/2016/04/cv_abide.zip

[8] Note that the criterion used to select the data-point used for the post-hoc analysis, *i. e.* the 10% best performing, is independent from the factors studied in this analysis, as we select the 10% best performing for each value of these factors. Hence, the danger of inflating effects by a selection [56] does not apply to the post hoc analysis.

2.8. Computational aspects

For data-driven brain atlas models (*e. g.* ICA), region-extraction methods entail feature learning on up to 300 GB of data, too big to fit in memory. Systematic study requires many cross-validations and subsamples and thus careful computational optimizations. Hence, for K-Means and ICA, we reduced the length of the time series by applying PCA dimensionality reduction with efficient randomized SVD [57]. MSDL uses algorithmic optimizations [49] to process a large number of subjects in a reasonable time [9] (2 hours for 871 subjects). However, it is still the most costly method because of the number of runs required to find the optimal value for its 3 parameters.

For reproducibility, we rely on the publicly available implementations of the scikit-learn package [55] for all the clustering algorithms, connectivity matrix estimators and predictors.

3. Results

Here we present the most notable trends emerging from our analyses. More details are provided in supplementary materials. First, we compared inter- and intra-site prediction of ASD diagnostic labels while varying the number of subjects in the training set. Second, in a post-hoc analysis, we identified the pipeline choices most relevant for prediction and proposed a good choice of pipeline steps for prediction. Finally, by analyzing the weights of the classifiers, we highlighted functional connections that best discriminate between ASD and TC.

3.1. Overall prediction results

For inter-site prediction, the highest accuracy obtained on the whole dataset is 66.8% (see Table 2) which exceeds previously published ABIDE findings [53, 13, 54] and chance at 53.7%[10]. Importantly, performance increases steadily with training set size, for all subsamples (Figures 2 and A6). This increase implies that optimal classification performance is not reached even for the largest dataset tested.

The main difference between intra-site and inter-site settings is that the variability of inter-site prediction is higher (Table 2). However, this difference disappears with a large enough training set (Figure 2). This is encouraging for clinical applications.

As shown in Figure 3, the best predicting pipeline is MSDL, tangent space embedding and l_2-regularized classifiers (SVC-l_2 and ridge classifier). All effects are observed with $p < 0.01$. Detailed pairwise comparisons (Figure 4) confirm these trends by showing the effect of each option compared to the best ones. Results of intra-site prediction have smaller standard error, confirming higher variability of inter-site prediction results (Figure A5).

Subsample	#5	#4	#3	#2	#1
Inter-site Accuracy		69.7% ±8.9%	65.1% ±5.8%	68.7% ±9.3%	66.8% ±5.4%
Inter-site Specificity		74.0% ±12.9%	69.1% ±7.8%	73.9% ±11.7%	72.3% ±12.1%
Inter-site Sensitivity		65.4% ±13.7%	61.3% ±7.7%	62.8% ±13.6%	61.0% ±17.0%
Intra-site Accuracy	66.6% ±5.4%	65.8% ±5.9%	65.7% ±4.9%	67.9% ±1.9%	66.9% ±2.7%
Intra-site Specificity	78.4% ±9.7%	72.3% ±9.1%	77.5% ±8.1%	76.6% ±5.0%	78.3% ±4.1%
Intra-site Sensitivity	53.2% ±10.9%	58.1% ±9.2%	52.2% ±9.1%	58.1% ±5.0%	53.2% ±5.8%

Table 2: **Average accuracy, specificity and sensitivity scores (and standard deviation) for top performing pipelines.** Accuracy is the fraction of well-classified individuals (chance level is at 53.7%). Specificity is the fraction of well classified healthy individuals among all healthy individuals (percentage of well-classified negatives). Sensitivity is the ratio of ASD individuals among all subjects classified as ASD (percentage of positives that are true positives). Results per atlas are available in the appendices –Table A5.

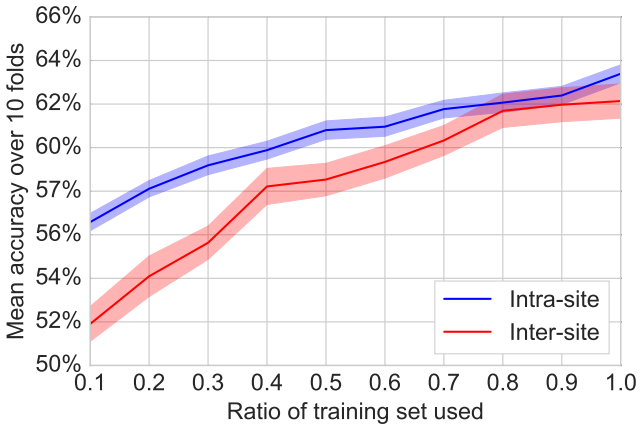

Figure 2: **Learning curve.** Classification results obtained by varying the ratio of the training set using to train the classifier while keeping a fixed testing set. The colored bands represent the standard error of the prediction. A score increasing with the number of subjects, *i. e.* a positive slope indicates that the addition of new subjects improves performance. This curve is an average of the results obtained on several subsamples. Detailed results per subsamples are presented in the appendices –Figure A6

3.2. Effect of the choice of atlas

The choice of atlas appears to have the greatest impact on prediction accuracy (Figure 3). Estimation of an atlas from the data with MSDL, or using reference atlases leads to maximal performance. Of note, for smaller samples, data-driven atlas-extraction methods other than MSDL perform poorly (Figure A8). In this regime, noise

[9] Despite the optimizations, the whole study presented here is very computationally costly. Indeed, the nested cross-validation, the various pipeline options, and the various subsets, would lead to a computation time of 10 years on a computer with 8 cores and 32GB RAM. We used a computer grid to reduce it to two months.

[10] Chance is calculated by taking the highest score obtained by a dummy classifier that predicts random labels following the distribution of the two classes.

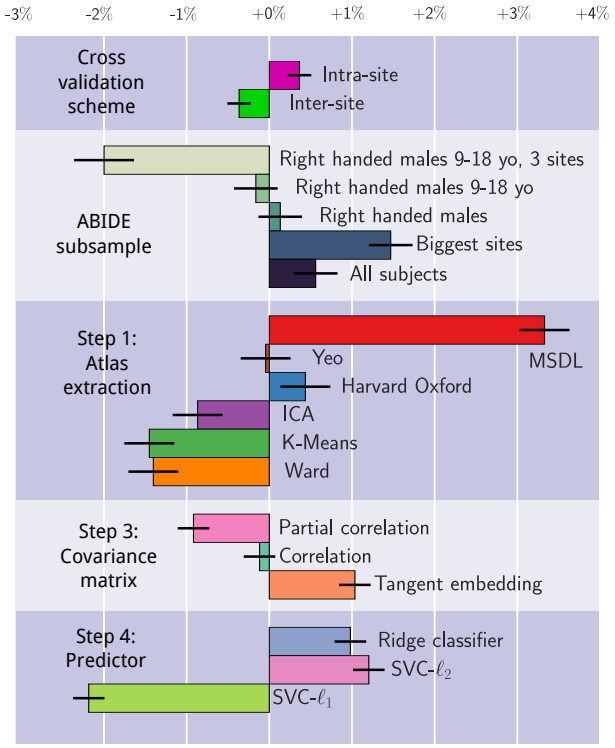

Figure 3: **Impact of pipeline steps on prediction.** Each block of bars represents a step of the pipeline (namely step 1, 3 and 4). We also prepended two steps corresponding to the cross-validation schemes and ABIDE subsamples. Each bar represents the impact of the corresponding option on the prediction accuracy, relative to the mean prediction. This effect is measured via a full-factorial analysis of variance (ANOVA), analyzing the contribution of each step in a linear model. Each step of the pipeline is considered as a categorical variable. Error bars give the 95% confidence interval. Multi Subject Dictionary Learning (MSDL) atlas extraction method gives significantly better results while reference atlases are slightly better than the mean. Among all matrix types, tangent embedding is the best on all ABIDE subsets. Finally, l_2-regularized classifiers perform better than l_1-regularized ones. Results for each subsamples are presented in the appendices –Figure A8.

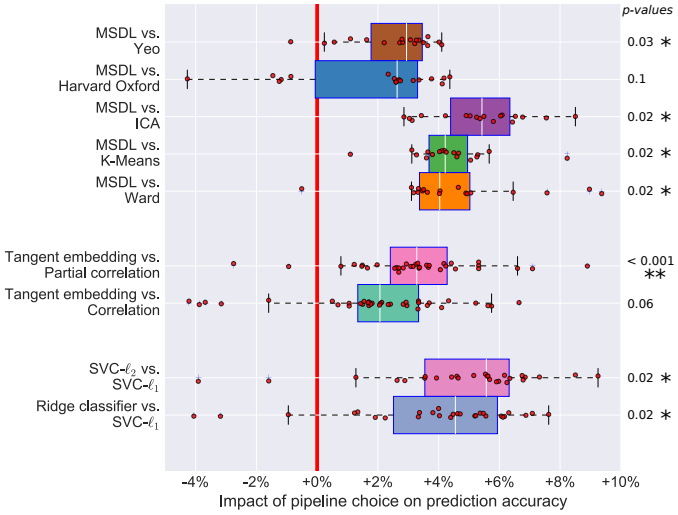

Figure 4: **Comparison of pipeline options.** Each plot compares the classification accuracy scores if one parameter is changed and the others kept fixed. We measured statistical signficance using a Wilcoxon signed rank test and corrected for multiple comparisons. Detailed one-to-one comparison of the scores are shown in the appendices –Figure A5

can be problematic and limit the generalizability of atlases derived from data.

MSDL performs well regardless of sample size, and the performance gain is significant compared to all other strategies apart from the Harvard-Oxford atlas (Figure 4). This is likely attributable to MSDL's strong spatially-structured regularization that increases its robustness to noise.

3.3. Effect of the covariance matrix estimator

While correlation and partial correlation approaches are currently the two dominant approaches to connectome estimation in the current R-fMRI literature, our results highlight the value of tangent-space projection that outperforms the other approaches in all settings (Figure 3 and Figure 4), though the difference is significant only compared with correlation matrices (Figure 4). This gain in prediction accuracy is not without a cost, as the matrices derived cannot be read as correlation matrices. Yet, the differences in connectivity that they capture can be directly interpreted [58].

Among the correlation-based approaches, full correlations outperformed partial correlations, most notably for intra-site prediction. These limitations of partial correlations may reflect estimation challenges[11] with the relatively short time series included in ABIDE datasets (typically 5-6 minutes per participant).

3.4. Effect of the final classifier

The results show that ℓ_2-regularized classifiers perform best in all settings (Figure 3 and Figure 4). This may be due to the fact that a global hypoconnectivity, as previously observed in ASD patients, is not well captured by ℓ_1-regularized classifiers, which induce sparsity. In addition ℓ_2-penalization is rotationally-invariant, which means that it is not sensitive to an arbitrary mixing of features. Such mixing may happen if a functional brain module sits astride several regions. Thus, a possible explanation for the good performance of ℓ_2-penalization is that it does not need an optimal choice of regions, perfectly aligned with the underlying functional modules.

3.5. Effect of the dataset size and heterogeneity

While not a property of the data-processing pipeline itself, a large number of training subjects is the most important factor of success for prediction. Indeed, we observe that prediction results improve with the number of subjects, even for large number of subjects (Figure 3).

[11] Note that we experimented with a variety of more sophisticated covariance estimators, including GraphLasso, though without improved results.

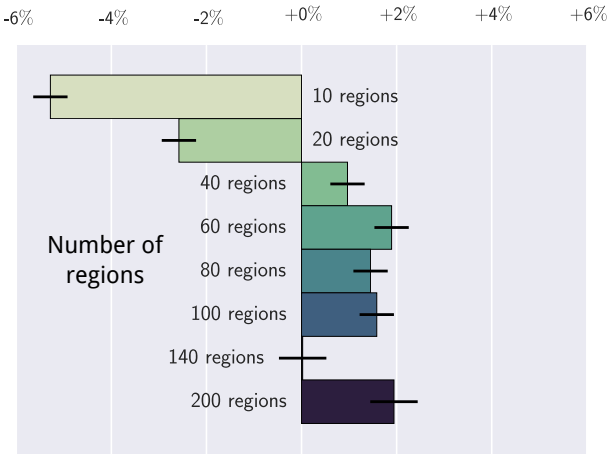

Figure 5: **Impact of the number of regions on prediction**: each bar indicates the impact of the number of regions on the prediction accuracy relative to the mean of the prediction. These values are coefficients in a linear model explaining the best classification scores as function of the number of regions. Error bars give the 95% confidence interval, computed by a full factorial ANOVA. Atlases containing more than 40 ROIs give better results in all settings. Results per subsamples are presented in the appendices –Figure A9.

Despite that increase, we note that prediction on the largest subsample (subsample #1) gives lower scores than prediction on subsample #2, which may be due to the variance introduced by small sites. We also note that the importance of the choice between each option of the pipeline decreases with the number of subjects (Figure A8).

3.6. Effect of the number of regions

In order to compare models of similar complexity, the previous experiments were restricted to 84 regions as in [49]. However, this choice is based on previous studies and may not be the optimal choice for our problem. To validate this aspect of the study, we also varied the number of regions and analyzed its impact on the classification results in Figure 5, similarly to Figure 3. In order to avoid computational cost[12], we explored only two atlases: the data-driven Ward's hierarchical clustering and spectral clustering atlases computed in [43].

Across all settings, we distinguish 3 different regimes. First, below 20 ROIs, performance is bad. Indeed very large regions average out signals of different functional areas. Above 140 ROIs, the results seem to be unstable, though this trend is less clear in intra-site prediction. Finally, the best results are between 40 and 100 ROIs. Our choice of 84 ROIs is within the range of best performing models.

[12] As MSDL is much more computationally expensive than Ward, we have not explored the effect of modifying of the number of regions extracted with MSDL. However, we expect that the optimal number of regions is not finely dependent on the region-extraction method.

3.7. Inspecting discriminative connections

A connection between two regions is considered discriminative if its value helps diagnosing ASD from TC. To understand what is driving the best prediction pipeline, we extract its most discriminative connections as described in section 2.7. Given that the 10-fold cross-validation yields 10 different atlases, we start by building a consensus atlas by taking all the regions with a DICE similarity score above 0.9. We obtain an atlas composed of 37 regions (see Figure 6). These regions and the discriminative connections form the neurophenotypes. Note that discriminative connections should be interpreted with caution. Indeed, our study covers a wide range of age and diagnostic criteria. Additionally, predictive models cannot be interpreted as edge-level tests.

Default Mode Network (Figure 6.a). We observe a lower connectivity between left and right temporo-parietal junctions. Both hypoconnectivity between regions on adolescent and adult patients [59, 60] and hyperconnectivity in children [61] have been reported. Our findings are concordant with [62] that observed lower fronto-parietal connectivity and stronger connectivity between temporal ROIs.

Pareto-insular network (Figure 6.b). We observe interhemispheric hypoconnectivity between left and right anterior insulae, and left and right inferior parietal lobes. Those regions are part of a frontoparietal control network related to learning and memory [63]. Such hypoconnectivity has already been widely evoked in ASD [64, 65] and previously found in the ABIDE dataset [13, 66]. These ROIs also match anatomical regions of increased cortical density in ASD [67].

Semantic ROIs (Figure 6.c). We observe a more complicated pattern in ROIs related to language. Connectivity is stronger between the right supramarginal gyrus and the left Middle Temporal Gyrus (MTG) while a lower connectivity is observed between the right MTG and the left temporo-parietal junction. Differences of laterality in language networks implicating the middle temporal gyri have already been observed [68]. Although this symptom is often considered as decorrelated from ASD, we find that these regions are relevant for diagnosis.

4. Discussion

We studied pipelines that extract neurophenotypes from aggregate R-fMRI datasets through the following steps: 1) region definition from R-fMRI, 2) extraction of regions activity time series 3) estimation of functional interactions between regions, and 4) construction of a discriminant model for brain-based diagnostic classification. The proposed pipelines can be built with the Nilearn neuroimaging-analysis software and the atlas computed

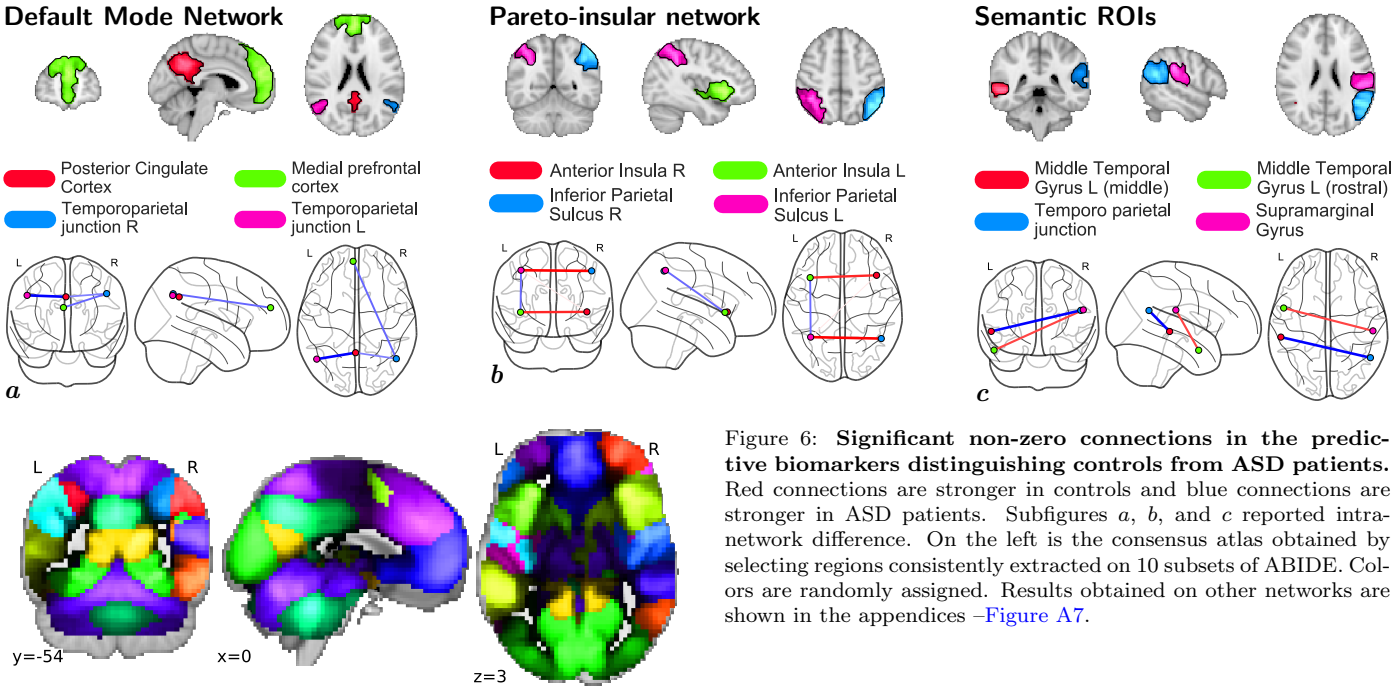

Figure 6: **Significant non-zero connections in the predictive biomarkers distinguishing controls from ASD patients.** Red connections are stronger in controls and blue connections are stronger in ASD patients. Subfigures *a*, *b*, and *c* reported intra-network difference. On the left is the consensus atlas obtained by selecting regions consistently extracted on 10 subsets of ABIDE. Colors are randomly assigned. Results obtained on other networks are shown in the appendices –Figure A7.

with MSDL is available for download [13]. We have demonstrated that they can successfully predict diagnostic status across new acquisition sites in a real-world situation, a large Autism database. This validation with a leave-one-site-out cross-validation strategy reduces the risk of overfitting data from a single site, as opposed to the commonly used leave-one-subject-out validation approach [53]. It enabled us to measure the impact of methodological choices on prediction. In the ABIDE dataset, this approach classified the label of autism versus control with an accuracy of 67% – a rate superior to prior work using the larger ABIDE sample.

The heterogeneity of weakly-controlled datasets formed by aggregation of multiple sites poses great challenges to develop brain-based classifiers for psychiatric illnesses. Among those most commonly cited are: *i)* the many sources of uncontrolled variation that can arise across studies and sites (*e. g.* scanner type, pulse sequence, recruitment strategies, sample composition), *ii)* the potential for developing classifiers that are reflective of only the largest sites included, and *iii)* the danger of extracting biomarkers that will be useful only on sites included in the training sample. Counter to these concerns, our results demonstrated the feasibility of using weakly-controlled heterogeneous datasets to identify imaging biomarkers that are robust to differences among sites. Indeed, our prediction accuracy of 67% is the highest reported to-date on the whole ABIDE dataset (compared to 60% in [53]), though better prediction has been reported on curated more homogeneous subsets [13, 54]. Importantly, we found that increasing sample size helped tackling heterogeneity as inter-site prediction accuracy approached that of intra-site prediction. This undoubtedly increases the perceived value that can be derived from aggregate datasets, such as the ABIDE and the ADHD-200 [16].

Our results suggest directions to improve prediction accuracy. First, more data is needed to optimize the classifier. Indeed, our experiments show that with a training set of 690 participants aggregated across multiple sites (80% of 871 available subjects), the pipelines have not reached their optimal performance. Second, a broader diversity of data will be needed to more definitively assess prediction for sites not included in the training set. Both of these needs motivate more data sharing.

Methodologically, our systematic comparison of choices in the pipeline steps can give indications to establish standard processing pipelines. Given that processing pipelines differ greatly between studies [28], comparing data-analysis methods is challenging. On the ABIDE dataset, we found that a good choice of processing steps can strongly improve prediction accuracy. In particular, the choice of regions is most important. We found that extracting these regions with MSDL gives best results. However, the benefit of this method is not always clear cut – on small datasets, references atlases are also good performers. For the rest of the pipeline, tangent embedding and ℓ_2-regularized classifier are overall the best choice to maximize prediction accuracy.

While the primary goals of this study were methodo-

[13] http://team.inria.fr/parietal/files/2015/07/MSDL_ABIDE.zip

logical, it is worth noting that biomarkers identified were largely consistent with the current understanding of the neural basis of ASD. The most informative features to predict ASD concentrated in the intrinsic functional connectivity of three main functional systems: the default-mode, parieto-insular, and language networks, previously involved in ASD [69, 70, 71]. Specifically, decreased homotopic connectivity between temporo-parietal junction, insula and inferior parietal lobes appeared to characterize ASD. Decreases in homotopic functional connectivity have been previously reported in R-fMRI studies with moderately [64, 72] or large samples such as ABIDE [25]. Here our findings point toward a set of inter-hemispheric connections involving heteromodal association cortex subserving social cognition (*i. e.* temporo-parietal junction [73]), and cognitive control (anterior insula and right inferior parietal cortex [74]), commonly affected in ASD [64, 65, 13, 66]. Limitations of our biomarkers may arise from the heterogeneity of ASD as a neuropsychiatric disorder [75].

Beyond forming a neural correlate of the disorder, predictive biomarkers can help define diagnosis subgroups – the next logical endeavor in classification studies [76].

As a psychiatric study, the present work has a number of limitations. First, although composed of samples from geographically distinct sites, the representativeness of the ABIDE sample has not been established. As such, the features identified may be biased and necessitate further replication; though their consistency with prior findings alleviates this concern to some degree. Further work calls for characterizing this prediction pipeline on more participants (*e. g.* ABIDE II) and different pathologies (*e. g.* ADHD-200 dataset). Second, while the present work primarily relied on prediction accuracy to assess the classifiers derived, a variety of alternative measures exist (*e. g.* sensitivity, specificity, J-statistic). This decision was largely motivated to facilitate comparison of our findings with those from prior studies using ABIDE, which most frequently used prediction accuracy. Depending on the specific application for a classifier, the prediction metric to optimize may vary. For example screening tools would need to favor sensitivity over accuracy or specificity; in contrast with medical diagnostic tests that require specificity over sensitivity or accuracy [4]. The versatility of the pipeline studied allows us to maximize either of these scores depending on the experimenter's needs. Finally, while the present work focused on the ABIDE dataset, it is likely that its findings regarding optimal pipeline decisions will carry over to other aggregate samples, as well as more homogeneous samples. With that said, future applications will be required to verify this point.

In sum, we experimented with a large number of different pipelines and parametrizations on the task of predicting ASD diagnosis on the ABIDE dataset. From an extensive study of the results, we have found that R-fMRI from a large amount of participants sampled from many different sites could lead to functional-connectivity biomarkers of ASD that are robust, including to inter-site variability. Their predictive power markedly increases with the number of participants included. This increase holds promise for optimal biomarkers forming good probes of disorders.

Acknowledgments. We acknowledge funding from the NiConnect project and the SUBSample project from the DIGITEO Institute, France. The effort of Adriana Di Martino was partially supported by the NIMH grant 1R21MH107045-01A1. Finally, we thank the anonymous reviewers for their feedback, as well as all sites and investigators who have worked to share their data through ABIDE.

References

[1] S. Kapur, A. G. Phillips, T. R. Insel, Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it&quest, Molecular psychiatry 17 (12) (2012) 1174–1179.

[2] R. C. Craddock, S. Jbabdi, C.-G. Yan, J. T. Vogelstein, F. X. Castellanos, A. Di Martino, C. Kelly, K. Heberlein, S. Colcombe, M. P. Milham, Imaging human connectomes at the macroscale, Nature methods 10 (6) (2013) 524–539.

[3] D. C. Van Essen, K. Ugurbil, The future of the human connectome, Neuroimage 62 (2) (2012) 1299–1310.

[4] F. X. Castellanos, A. Di Martino, R. C. Craddock, A. D. Mehta, M. P. Milham, Clinical applications of the functional connectome, Neuroimage 80 (2013) 527–540.

[5] A. C. Kelly, L. Q. Uddin, B. B. Biswal, F. X. Castellanos, M. P. Milham, Competition between functional brain networks mediates behavioral variability, Neuroimage 39 (1) (2008) 527–537.

[6] M. Greicius, Resting-state functional connectivity in neuropsychiatric disorders, Current opinion in neurology 21 (2008) 424.

[7] M. Greicius, G. Srivastava, A. Reiss, V. Menon, Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI, Proceedings of the National Academy of Sciences 101 (2004) 4637.

[8] G. Chen, B. D. Ward, C. Xie, W. Li, Z. Wu, J. L. Jones, M. Franczak, P. Antuono, S.-J. Li, Classification of alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional mr imaging, Radiology 259 (1) (2011) 213–221.

[9] A. Garrity, G. Pearlson, K. McKiernan, D. Lloyd, K. Kiehl, V. Calhoun, Aberrant" default mode" functional connectivity in schizophrenia, Am J Psychiatry 164.

[10] Y. Zhou, M. Liang, L. Tian, K. Wang, Y. Hao, H. Liu, Z. Liu, T. Jiang, Functional disintegration in paranoid schizophrenia using resting-state fMRI, Schizophrenia research 97 (1) (2007) 194–205.

[11] M. J. Jafri, G. D. Pearlson, M. Stevens, V. D. Calhoun, A method for functional network connectivity among spatially independent resting-state components in schizophrenia, Neuroimage 39 (4) (2008) 1666–1681.

[12] V. D. Calhoun, J. Sui, K. Kiehl, J. Turner, E. Allen, G. Pearlson, Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder, Frontiers in psychiatry 2.

[13] M. Plitt, K. A. Barnes, A. Martin, Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards, NeuroImage: Clinical.

[14] J. S. Anderson, J. A. Nielsen, M. A. Ferguson, M. C. Burback, E. T. Cox, L. Dai, G. Gerig, J. O. Edgin, J. R. Korenberg, Abnormal brain synchrony in down syndrome, NeuroImage: Clinical.

[15] R. Craddock, P. Holtzheimer III, X. Hu, H. Mayberg, Disease state prediction from resting state functional connectivity, Magnetic resonance in Medicine 62 (2009) 1619.

[16] A.-. Consortium, et al., The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience, Frontiers in systems neuroscience 6.

[17] J. S. Anderson, J. A. Nielsen, A. L. Froehlich, M. B. DuBray, T. J. Druzgal, A. N. Cariello, J. R. Cooperrider, B. A. Zielinski, C. Ravichandran, P. T. Fletcher, et al., Functional connectivity magnetic resonance imaging classification of autism, Brain 134 (12) (2011) 3742–3754.

[18] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, et al., Prediction of individual brain maturity using fMRI, Science 329 (5997) (2010) 1358–1361.

[19] J. E. Desmond, G. H. Glover, Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses, Journal of neuroscience methods 118 (2) (2002) 115–128.

[20] K. Murphy, H. Garavan, An empirical investigation into the number of subjects required for an event-related fmri study, Neuroimage 22 (2) (2004) 879–885.

[21] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, J.-B. Poline, Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses, Neuroimage 35 (1) (2007) 105–120.

[22] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, M. R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nature Reviews Neuroscience 14 (5) (2013) 365–376.

[23] D. A. Fair, J. T. Nigg, S. Iyer, D. Bathula, K. L. Mills, N. U. Dosenbach, B. L. Schlaggar, M. Mennes, D. Gutman, S. Bangaru, et al., Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data, Frontiers in systems neuroscience 6.

[24] M. Mennes, B. B. Biswal, F. X. Castellanos, M. P. Milham, Making data sharing work: The fcp/indi experience, Neuroimage 82 (2013) 683–691.

[25] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Molecular psychiatry 19 (6) (2014) 659–667.

[26] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2008) 432.

[27] C.-G. Yan, R. C. Craddock, X.-N. Zuo, Y.-F. Zang, M. P. Milham, Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes, Neuroimage 80 (2013) 246–262.

[28] J. Carp, The secret lives of experiments: methods reporting in the fMRI literature, Neuroimage 63 (1) (2012) 289–300.

[29] W. R. Shirer, H. Jiang, C. M. Price, B. Ng, M. D. Greicius, Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination, NeuroImage.

[30] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, D. Van De Ville, Decoding brain states from fMRI connectivity graphs, NeuroImage.

[31] W. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, M. Greicius, Decoding subject-driven cognitive states with whole-brain connectivity patterns, Cerebral Cortex 22 (2012) 158.

[32] S. B. Eickhoff, B. Thirion, G. Varoquaux, D. Bzdok, Connectivity-based parcellation: Critique and implications, Human brain mapping 36 (12) (2015) 4771–4792.

[33] G. Varoquaux, C. Craddock, Learning and comparing functional connectomes across subjects, NeuroImage 80 (2013) 405.

[34] C. Craddock, S. Sikka, B. Cheung, R. Khanuja, S. Ghosh, C. Yan, Q. Li, D. Lurie, J. Vogelstein, R. Burns, et al., Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac), Front Neuroinform 42.

[35] T. E. Lund, M. D. Nørgaard, E. Rostrup, J. B. Rowe, O. B. Paulson, Motion or activity: their role in intra-and inter-subject variation in fMRI, Neuroimage 26 (3) (2005) 960–964.

[36] M. Fox, A. Snyder, J. Vincent, M. Corbetta, D. Van Essen, M. Raichle, The human brain is intrinsically organized into dynamic, anticorrelated functional networks, Proc Ntl Acad Sci 102 (2005) 9673.

[37] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. Frith, R. S. J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, Hum Brain Mapp (1995) 189.

[38] Y. Behzadi, K. Restom, J. Liau, T. Liu, A component based noise correction method (CompCor) for bold and perfusion based fMRI, Neuroimage 37 (2007) 90.

[39] B. B. Avants, N. Tustison, G. Song, Advanced normalization tools (ants), Insight J 2 (2009) 1–35.

[40] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (2006) 968.

[41] B. Yeo, F. Krienen, J. Sepulcre, M. Sabuncu, et al., The organization of the human cerebral cortex estimated by intrinsic functional connectivity, J Neurophysio 106 (2011) 1125.

[42] D. Lashkari, E. Vul, N. Kanwisher, P. Golland, Discovering structure in the space of fMRI selectivity profiles, NeuroImage 50 (2010) 1085.

[43] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, H. S. Mayberg, A whole brain fMRI atlas generated via spatially constrained spectral clustering, Human brain mapping 33 (8) (2012) 1914–1928.

[44] C. Goutte, P. Toft, E. Rostrup, F. Å. Nielsen, L. K. Hansen, On clustering fMRI time series, NeuroImage 9 (3) (1999) 298–310.

[45] B. Thirion, G. Varoquaux, E. Dohmatob, J.-B. Poline, Which fMRI clustering gives good brain parcellations?, Frontiers in neuroscience 8.

[46] V. D. Calhoun, T. Adali, G. D. Pearlson, J. J. Pekar, A method for making group inferences from fMRI data using independent component analysis., Hum Brain Mapp 14 (2001) 140.

[47] C. F. Beckmann, S. M. Smith, Probabilistic independent component analysis for functional magnetic resonance imaging, Trans Med Im 23 (2004) 137–152.

[48] G. Varoquaux, S. Sadaghiani, P. Pinel, A. Kleinschmidt, J. B. Poline, B. Thirion, A group model for stable multi-subject ICA on fMRI datasets, NeuroImage 51 (2010) 288.

[49] A. Abraham, E. Dohmatob, B. Thirion, D. Samaras, G. Varoquaux, Extracting brain regions from rest fMRI with Total-Variation constrained dictionary learning, in: MICCAI, 2013, p. 607.

[50] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. 88 (2004) 365.

[51] S. Smith, K. Miller, G. Salimi-Khorshidi, M. Webster, C. Beckmann, T. Nichols, J. Ramsey, M. Woolrich, Network modelling methods for fMRI, Neuroimage 54 (2011) 875.

[52] B. Ng, M. Dressler, G. Varoquaux, J. B. Poline, M. Greicius, B. Thirion, Transport on riemannian manifold for functional connectivity-based classification, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer International Publishing, 2014, pp. 405–412.

[53] J. A. Nielsen, B. A. Zielinski, P. T. Fletcher, A. L. Alexander, N. Lange, E. D. Bigler, J. E. Lainhart, J. S. Anderson, Multisite functional connectivity MRI classification of autism: ABIDE results, Frontiers in human neuroscience 7.

[54] C. P. Chen, C. L. Keown, A. Jahedi, A. Nair, M. E. Pflieger, B. A. Bailey, R.-A. Müller, Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism, NeuroImage: Clinical.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825.

[56] E. Vul, C. Harris, P. Winkielman, H. Pashler, Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition, Perspectives on psychological science 4 (3) (2009) 274–290.

[57] P. Martinsson, V. Rokhlin, M. Tygert, A randomized algorithm for the decomposition of matrices, Applied and Computational Harmonic Analysis 30 (2011) 47.

[58] G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, B. Thirion, Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, in: MICCAI, 2010, pp. 200–208.

[59] V. L. Cherkassky, R. K. Kana, T. A. Keller, M. A. Just, Functional connectivity in a baseline resting-state network in autism, Neuroreport 17 (16) (2006) 1687–1690.

[60] D. P. Kennedy, E. Redcay, E. Courchesne, Failing to deactivate: resting functional abnormalities in autism, Proceedings of the National Academy of Sciences 103 (21) (2006) 8275–8280.

[61] K. Supekar, L. Q. Uddin, A. Khouzam, J. Phillips, W. D. Gaillard, L. E. Kenworthy, B. E. Yerys, C. J. Vaidya, V. Menon, Brain hyperconnectivity in children with autism and its links to social deficits, Cell reports 5 (3) (2013) 738–747.

[62] C. S. Monk, S. J. Peltier, J. L. Wiggins, S.-J. Weng, M. Carrasco, S. Risi, C. Lord, Abnormalities of intrinsic functional connectivity in autism spectrum disorders, Neuroimage 47 (2) (2009) 764–772.

[63] T. Iidaka, A. Matsumoto, J. Nogawa, Y. Yamamoto, N. Sadato, Frontoparietal network involved in successful retrieval from episodic memory. spatial and temporal analyses using fMRI and ERP, Cerebral Cortex 16 (9) (2006) 1349–1360.

[64] J. S. Anderson, T. J. Druzgal, A. Froehlich, M. B. DuBray, N. Lange, A. L. Alexander, T. Abildskov, J. A. Nielsen, A. N. Cariello, J. R. Cooperrider, et al., Decreased interhemispheric functional connectivity in autism, Cerebral cortex (2010) bhq190.

[65] A. Di Martino, C. Kelly, R. Grzadzinski, X.-N. Zuo, M. Mennes, M. A. Mairena, C. Lord, F. X. Castellanos, M. P. Milham, Aberrant striatal functional connectivity in children with autism, Biological psychiatry 69 (9) (2011) 847–856.

[66] S. Ray, M. Miller, S. Karalunas, C. Robertson, D. S. Grayson, R. P. Cary, E. Hawkey, J. G. Painter, D. Kriz, E. Fombonne, et al., Structural and functional connectivity of the human brain in autism spectrum disorders and attention-deficit/hyperactivity disorder: A rich club-organization study, Human brain mapping 35 (12) (2014) 6032–6048.

[67] S. Haar, S. Berman, M. Behrmann, I. Dinstein, Anatomical abnormalities in autism?, Cerebral Cortex (2014) bhu242.

[68] N. M. Kleinhans, R.-A. Müller, D. N. Cohen, E. Courchesne, Atypical functional lateralization of language in autism spectrum disorders, Brain research 1221 (2008) 115–125.

[69] M. E. Vissers, M. X. Cohen, H. M. Geurts, Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links, Neuroscience & Biobehavioral Reviews 36 (1) (2012) 604–625.

[70] P. Rane, D. Cochran, S. M. Hodge, C. Haselgrove, D. N. Kennedy, J. A. Frazier, Connectivity in autism: A review of mri connectivity studies, Harvard review of psychiatry 23 (4) (2015) 223–244.

[71] L. M. Hernandez, J. D. Rudie, S. A. Green, S. Bookheimer, M. Dapretto, Neural signatures of autism spectrum disorders: insights into brain network dynamics, Neuropsychopharmacology 40 (1) (2015) 171–189.

[72] I. Dinstein, K. Pierce, L. Eyler, S. Solso, R. Malach, M. Behrmann, E. Courchesne, Disrupted neural synchronization in toddlers with autism, Neuron 70 (6) (2011) 1218–1225.

[73] R. Saxe, N. Kanwisher, People thinking about thinking people: the role of the temporo-parietal junction in theory of mind, Neuroimage 19 (4) (2003) 1835–1842.

[74] V. Menon, L. Q. Uddin, Saliency, switching, attention and control: a network model of insula function, Brain Structure and Function 214 (5-6) (2010) 655–667.

[75] D. H. Geschwind, P. Levitt, Autism spectrum disorders: developmental disconnection syndromes, Current opinion in neurobiology 17 (1) (2007) 103–111.

[76] E. Loth, W. Spooren, L. M. Ham, M. B. Isaac, C. Auriche-Benichou, T. Banaschewski, S. Baron-Cohen, K. Broich, S. Bölte, T. Bourgeron, et al., Identification and validation of biomarkers for autism spectrum disorders, Nature Reviews Drug Discovery 15 (1) (2016) 70–73.

[77] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, LIBLINEAR: A library for large linear classification, The Journal of Machine Learning Research 9 (2008) 1871.

[78] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, S. M. Smith, Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers, Neuroimage 90 (2014) 449–468.