Benchmarking functional connectome-based predictive models for resting-state fMRI

Kamalaker Dadi[a,b,*], Mehdi Rahim[a,b], Alexandre Abraham[a,b], Darya Chyzhyk[a,b,c], Michael Milham[c], Bertrand Thirion[a,b], Gaël Varoquaux[a,b], for the Alzheimer's Disease Neuroimaging Initiative[d]

[a] *Parietal project-team, INRIA Saclay-île de France, France*
[b] *CEA/Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[c] *Center for the Developing Brain Child Mind Institute, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, USA*
[d] *One of the dataset used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*

Abstract

Functional connectomes reveal biomarkers of individual psychological or clinical traits. However, there is great variability in the analytic pipelines typically used to derive them from rest-fMRI cohorts. Here, we consider a specific type of studies, using predictive models on the edge weights of functional connectomes, for which we highlight the best modeling choices. We systematically study the prediction performances of models in 6 different cohorts and a total of 2 000 individuals, encompassing neuro-degenerative (Alzheimer's, Post-traumatic stress disorder), neuro-psychiatric (Schizophrenia, Autism), drug impact (Cannabis use) clinical settings and psychological trait (fluid intelligence). The typical prediction procedure from rest-fMRI consists of three main steps: defining brain regions, representing the interactions, and supervised learning. For each step we benchmark typical choices: 8 different ways of defining regions –either pre-defined or generated from the rest-fMRI data– 3 measures to build functional connectomes from the extracted time-series, and 10 classification models to compare functional interactions across subjects. Our benchmarks summarize more than 240 different pipelines and outline modeling choices that show consistent prediction performances in spite of variations in the populations and sites. We find that regions defined from functional data work best; that it is beneficial to capture between-region interactions with tangent-based parametrization of covariances, a midway between correlations and partial correlation; and that simple linear predictors such as a logistic regression give the best prediction. Our work is a step forward to establishing reproducible imaging-based biomarkers for clinical settings.

Keywords: Resting-state fMRI; Functional connectomes; Predictive modeling; Classification; Population study

1. Introduction

Resting-state functional Magnetic Resonance Imaging (rest-fMRI), based on the analysis of brain activity without specific task, has become a tool of choice to probe human brain function in healthy and diseased populations. As it can easily be acquired in many different individuals, rest-fMRI is a promising candidate for markers of brain function (Biswal et al., 2010; Greicius, 2008). This has lead to the rise of large-scale rest-fMRI data collections, such as the human connectome project (Van Essen et al., 2013) or ABIDE (Di Martino et al., 2014). Larger datasets bring increased statistical power (Elliott et al., 2008), and many population-imaging studies use rest-fMRI to relate brain imaging to neuropathologies or other behavior and population phenotypes (Miller et al., 2016; Dubois and Adolphs, 2016). These efforts build biomarkers from rest-fMRI with predictive models (Woo et al., 2017).

A *functional connectome* – characterizing the network structure of the brain (Sporns et al., 2005)– can be extracted from functional interactions in rest-fMRI data (Varoquaux and Craddock, 2013). The weights of the corresponding brain functional connectome are used to characterize individual subjects behavior, cognition, and mental health (Craddock et al., 2009; Richiardi et al., 2010; Milazzo et al., 2014; Smith et al., 2015; Miller et al., 2016; Colclough et al., 2017; Dubois et al., 2018), aging (Liem et al., 2017) as well as brain pathologies (Drysdale et al., 2016; Abraham et al., 2017; Ng et al., 2017).

Machine-learning pipelines are key to turning functional connectomes into biomarkers that *predict* the phenotype of interest (Woo et al., 2017). On rest-fMRI, such a pipeline typically comprises of 3 crucial steps as depicted on Figure 1, linking functional connectomes to the target phenotype (Varoquaux and Craddock, 2013; Craddock et al., 2015). Yet, there exist many variations of this prototypical pipeline, even for classification from edge-weights of brain functional connectomes, as revealed by reviews of

*Corresponding author

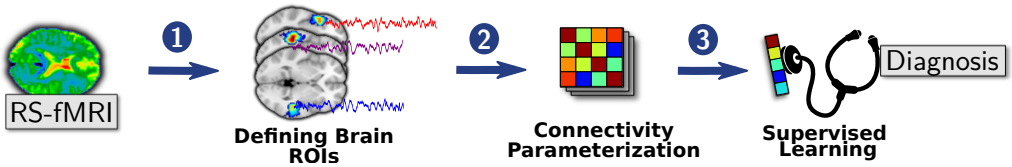

Figure 1: **Functional connectome-prediction pipeline** with three main steps: **1**) definition of brain regions (ROIs) from rest-fMRI images or using already defined reference atlases, **2**) quantifying functional interactions from time series signals extracted from these ROIs and **3**) comparisons of functional interactions across subjects using supervised learning.

the field (Wolfers et al., 2015; Arbabshirani et al., 2017; Brown and Hamarneh, 2016). These various choices have a sizable impact on the accuracy of population studies, and are seldom discussed (Carp, 2012). The cost of such analytical variation is twofold. First, it puts the burden on the practitioner to explore many options and make choices without systematic guidance. Second, methods variations create researchers degrees of freedom (Simmons et al., 2011) that can compromise the measure of the prediction accuracy of biomarkers (Varoquaux, 2017). Guidelines on optimal modeling choices are thus of great value for rest-fMRI biomarker research.

Here, we perform a systematic benchmark of common choices for the different steps of the functional connectome-based classification pipeline. To outline the preferable strategies, we analyze the prediction accuracy across 6 different cohorts, with different clinical questions and one psychological trait, different sample sizes, and prediction problems of different difficulties. While best model choice may vary depending on the prediction task, our benchmarks outline some trends. Specifically, we explore the following analytical choices:

- How should nodes be chosen: via pre-defined atlases, or data-driven approaches? How many nodes are needed for brain-imaging based diagnosis? Should nodes be distributed brain networks or regions of interest (ROIs)?
- How should weights of brain functional connectomes be represented: via correlations, partial correlation, or more complex models capturing the geometry of covariance matrices?
- What classifiers should be used for machine learning on weights of brain functional connectomes? Should linear or non-linear models be preferred? Should sparse or non-sparse models be used? With or without feature selection?

Besides these main questions, we did additional experiments on preprocessing strategies —studying the effect of band-pass filtering and global signal regression— and on covariance estimators, by comparing sparsity-inducing to classical shrinkage.

The paper is organized as follows: we first review current practices and methods used to-date for prediction of psychiatric diseases from weights of brain functional connectomes. Then, we present the different choices that we benchmark for the steps of classification pipelines and describe these methods. Finally, we report our experimental results and the trends that they reveal.

2. Methods: functional connectome-classification pipeline

Figure 1 shows the standard rest-fMRI classification pipeline that we consider.

2.1. A brief review of current practices: functional connectome-based predictive methods

We first survey methods used for prediction studies based on three extensive reviews: Wolfers et al. (2015); Arbabshirani et al. (2017); Brown and Hamarneh (2016). From these reviews, 27 studies used rest-fMRI and gave good classification scores. Below, we briefly outline the choices in the different pipeline step used (see Table A2 in the appendix for the full list).

Definition of brain ROIs. Studies define ROIs to extract signals with a variety of approaches:

- balls[1] of radius varying from $5mm$ to $10mm$ centered at coordinates from the literature (Dosenbach et al., 2010; Power et al., 2011);
- reference anatomical atlases such as AAL (Tzourio-Mazoyer et al., 2002), sulci-based atlases (Perrot et al., 2009; Desikan et al., 2006), or connectivity-based cortical landmarks (Zhu et al., 2013);
- data-driven approaches based on k-means or Ward clustering, as well as Independent Component Analysis (ICA) approaches (Calhoun et al., 2001; Beckmann and Smith, 2004) or dictionary learning (Abraham et al., 2013).

The number of nodes used was typically around 100, but ranged from dozens to several hundreds.

Representation of brain functional connectomes. Studies define functional interactions from second-order statistics –based on signal covariance– using Pearson's correlation or partial correlations estimated mostly either with the maximum-likelihood formula for the covariance or the Ledoit-Wolf shrinkage covariance estimator (Ledoit and Wolf, 2004; Varoquaux and Craddock, 2013; Brier et al., 2015). Partial correlation between nodes is useful to rule

[1] We used the term ball rather a sphere. From a mathematical standpoint, A "ball" is the inside of a sphere.

out indirect effects in the correlation structure, but calls for shrunk estimates (Smith et al., 2011; Varoquaux et al., 2010b). Mathematical arguments have also led to representations tailored to the manifold-structure of covariance matrices (Varoquaux et al., 2010a; Ng et al., 2014; Dodero et al., 2015; Colclough et al., 2017). We benchmark the simplest of these, a *tangent* representation of the manifold which underlies the more complex developments (see Appendix A for a quick introduction to this formalism).

Classifiers used for prediction. Many different classifiers have been used, whether linear or non-linear, sparse or non-sparse, optionally with prior feature selection. See Table A2 for the comprehensive list of classifiers used in these studies.

Finally, beyond the prototypical pipeline exposed above, some studies employ complex-graph network modeling approaches –*e.g.* network modularity or centrality (Rubinov and Sporns, 2011)– (Wolfers et al., 2015; Arbabshirani et al., 2017; Brown and Hamarneh, 2016) These approaches are seldom combined with supervised learning. Indeed, graph-theory metrics capture well global aspects of brain connectivity, but do not lend themselves well to tuning to connections in specific subnetworks (Hallquist and Hillary, 2018). Here, we focus on machine-learning methods that extract discriminant connections; as such we do not study graph-theoretical approaches.

The current practice is very diverse, without standard modeling choices. To open the way toward informed decisions, we explore popular variants of the classic machine-learning pipeline to predict on connectomes. We measure the impact of choices at each step on prediction for diverse targets across multiple datasets. We detail below the specific modeling choices included in our benchmarks.

2.2. Definition of brain regions of interest (ROIs)

For functional connectomes, the hypothesis is that the definition of ROIs should capture well the relevant functional units (Smith et al., 2011). We study both anatomically and functionally defined reference brain atlases, as well as data-driven methods that define ROIs from the data at hand. ROI selection is a difficult choice, as the optimal may vary for different conditions or pathologies.

A selection of pre-defined atlases. We consider four standard atlases, of which two are structural atlases: *i)* **Automated Anatomical Labeling (AAL)** (Tzourio-Mazoyer et al., 2002), a structural atlas with 116 ROIs defined from the anatomy of a reference subject, *ii)* **Harvard Oxford** (Desikan et al., 2006), a probabilistic atlas of anatomical structures, contains of 48 cortical & 11 subcortical ROIs in each hemisphere, ie 118 ROIs in total. We also include two functional atlases: *iii)* **Bootstrap Analysis of Stable Clusters (BASC)** (Bellec et al., 2010), a multi-scale functional atlas built with clustering on rest-fMRI, coming with different $\{36, 64, 122, 197, 325, 444\}$ numbers of ROIs; *iv)* **Power**, a coordinate-based atlas consisting of 264 coordinates used to position balls of $5mm$ radius (Power et al., 2011). For an additional set of benchmarks, on larger data, we use only pre-computed regions. For a pre-computed functional atlas with dictionary learning, we use an atlas[2] computed by Mensch et al. (2016a) with a very scalable sparse dictionary-learning algorithm on the HCP900 dataset (Van Essen et al., 2012). This algorithm, MODL (massive online dictionary learning), solves the ℓ_1 dictionary-learning problem with an algorithm fast on very large datasets that converges to the same solution as standard on-line solvers (Mensch et al., 2018).

A selection of data-driven methods. We consider four popular data-driven methods to extract brain ROIs from intrinsic brain activity (Yeo et al., 2011; Kahnt et al., 2012; Thirion et al., 2014; Calhoun et al., 2001; Beckmann and Smith, 2004; Abraham et al., 2013). We choose to define ROIs using two clustering methods: *i)* **K-Means** (Hastie et al., 2009), and *ii)* hierarchical agglomerative clustering using **Wards algorithm** (Ward, 1963) with spatial connectivity constraints (Michel et al., 2012); and two linear decomposition methods: *iii)* **Canonical Independent Component Analysis (GroupICA or CanICA)** (Varoquaux et al., 2010c), *iv)* **Dictionary Learning - ℓ_1 (DictLearn)** (Mensch et al., 2016b).

Dimension selection in data-driven atlases. For clustering methods, we extract brain atlases with a varying number of ROIs in $dim = \{\mathbf{40, 60, 80, 100, 120, 150, 200, 300}\}$. With linear decomposition methods *i.e.* CanICA and DictLearn, we explore the following number of components: $dim = \{\mathbf{40, 60, 80, 100, 120}\}$[3].

For each data-driven method, we learn brain ROIs on the training set only, to avoid possible overfit (Abraham et al., 2017). In a cross-validation loop, for each split, we define the brain ROIs on a training set and use the atlases to learn connectivity patterns for prediction. We also applied additional Gaussian smoothing of $6mm$ on preprocessed rest-fMRI datasets for all data-driven methods prior to learning brain ROIs to enhance the region extraction step.

Nodes formed local regions or distributed networks?. Current practices in functional connectomics includes defining nodes as local **regions** of the brain (Shirer et al., 2012;

[2] Pre-computed sparse dictionaries with the MODL approach of Mensch et al. (2016a) are available from `https://team.inria.fr/parietal/files/2018/10/MODL_rois.zip`

[3] We also investigated higher dimensionality (150, 200 and 300) on some of the datasets, but could not do a systematic study above 300 because of high computational costs. These preliminary results showed no improvements in prediction accuracy compared to lower dimensionalities. It should be noted that the resulting components typically encompass several brain regions, which explains the dimension difference.

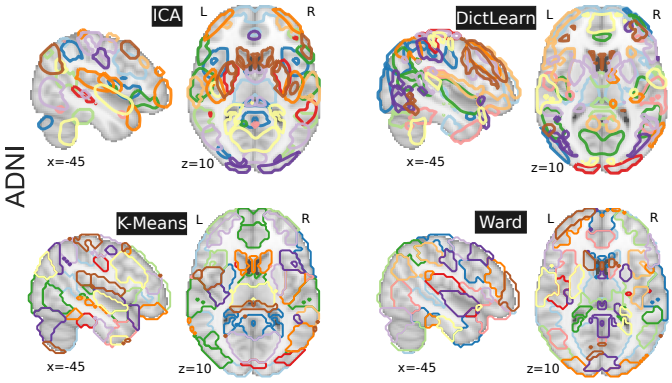

Figure 2: **Brain regions extracted with ICA, DictLearn, KMeans, and Ward** For ICA and dictionary learning, the dimensionality is of 80 and 60 resting-state networks – which are then broken up into more regions – yielding 150 regions, and 120 for KMeans and Ward clustering. Colors are arbitrary.

Craddock et al., 2012), or as full distributed functional **networks** that may include several regions (Smith et al., 2015; Yeo et al., 2011). We consider both approaches: using the distributed networks, or breaking them up in regions with a segmentation step to separate out regions (Abraham et al., 2014a). For example, a bi-hemispheric brain networks is separated into regions, one in each hemisphere.

We use a Random-Walker based extraction of regions from the brain networks obtained by CanICA and DictLearn as proposed in Abraham et al. (2014a). By contrast, for K-Means and BASC, we simply break out clusters in their connected components. During this procedure, we remove spurious regions of size $< 1500mm^3$. Figure 2 shows an example of the set of brain regions obtained from the various data-driven methods on the ADNI rest-fMRI data.

2.3. Connectivity parametrization

We extract representative time series for each node. For signal extraction, we explored several denoising strategies to account for non-neural artifacts: with or without low-pass filtering or global signal mean regression (details in Appendix B). To estimate functional connectomes efficiently, we use the Ledoit-Wolf regularized shrinkage estimator (Ledoit and Wolf, 2004; Varoquaux and Craddock, 2013; Brier et al., 2015), which gives a closed form expression for the shrinkage parameter. This estimator yields well-conditioned estimators despite the variation in length of time series across rest-fMRI datasets. We also explored non-regularized and sparse estimator for the covariance (see Appendix H.2). With this covariance structure, we study three different parametrizations of functional interactions: **full correlation**, **partial correlation** (Smith et al., 2011; Varoquaux and Craddock, 2013) and the **tangent space of covariance matrices**. The latter is less frequently used but has solid mathematical foundations and a variety of groups have reported good decoding performances with this framework (Varoquaux et al., 2010a; Barachant et al., 2013; Ng et al., 2014; Dodero et al., 2015; Qiu et al., 2015; Rahim et al., 2017; Wong et al., 2018). We compared two variants, using as a reference point the Euclidean mean (Varoquaux et al., 2010a) or the geometric mean (Ng et al., 2014); in both cases we rely on Nilearn implementation (Abraham et al., 2014b). Note that computing partial correlation or tangent space require inverting covariance matrices, hence these must be well conditioned. Non regularized covariance estimation is thus not useable for these parametrizations.

For each parametrization, we vectorize the functional connectome, using the lower triangular part of the connectomes matrix for classification. Additionally, on the ACPI dataset, we considered the ADHD status of the subjects as a variable of non interest and regressed it out in this second-level analysis, as we were interested in predicting the consumption of Marijuana.

2.4. Supervised learning: Classifiers

The final step of our pipeline predicts a binary phenotypic status from connectivity features extracted from previous step. We consider several linear and non-linear classifiers for prediction *i.e.* both sparse and non-sparse methods. For non-linear methods, we consider **Nearest Neighbors (K-NN)** (Cover and Hart, 1967) with K=1 and Euclidean distance metric, **Gaussian Naïve Bayes (GNB)** and **Random Forests Classifier (RF)** (Breiman, 2001). For linear classifiers we consider sparse ℓ_1 regularization[4] for **Support Vector Classification (SVC)**, and **Logistic Regression** (Hastie et al., 2009). For non-sparse linear classifiers –*i.e.* ℓ_2 regularization– we consider **Ridge classification, SVC, Logistic regression**. For SVC, we also considered 10% feature screening with univariate ANOVA. With regards to the regularization parameter (*eg* soft margin parameter in SVC), we use the default $C = 1$ or $\alpha = 1$, which has been found to be a good default (Varoquaux et al., 2017).

3. Experimental study

To benchmark the various predictive-modeling choices, we apply the functional connectome-classification pipeline on five publicly-available rest-fMRI datasets. We study prediction from functional connectomes of various clinical outcomes –neuro-degenerative and neuro-psychiatric disorders, drug abuse impact, fluid intelligence. We focus on binary classification problems, predicting a phenotypic target between two groups. We use the following datasets, summarized in Table 1:

1. **COBRE**, Center for Biomedical Research Excellence[5], comprising rest-fMRI data to study schizophrenia and bipolar disorder (Calhoun et al., 2012). We

[4] We also included **Lasso** as another choice of classifier in the pipeline. We observed significantly low prediction performance.

[5] cobre.mrn.org

focus on predicting schizophrenia diagnosis versus normal control.

2. **ADNI**, the Alzheimer's Disease Neuroimaging Initiative[6] database studies neuro-degenerative diseases (Mueller et al., 2005). We focus on using rest-fMRI to discriminate individuals with Mild Cognitive Impairment (MCI) from individuals diagnosed with Alzheimer's Disease (AD).
3. **ADNIDOD**, funded by the US Department of Defense (DoD) to study brain aging in Vietnam War Veterans[7], includes rest-fMRI data of individuals with post-traumatic stress disorders (PTSD) or brain traumatic injuries. We focus on discriminating PTSD condition from normal controls.
4. **ACPI**, Addiction Connectome Preprocessed Initiative[8], a longitudinal study to investigate the effect of cannabis use among adults with a childhood diagnosis of ADHD. In particular we use readily-preprocessed rest-fMRI data from Multimodal treatment study of Attention Deficit Hyperactivity Disorder (MTA). We attempt to discriminate whether individuals have consumed marijuana or not.
5. **ABIDE**, Autism Brain Imaging Data Exchange database investigates the neural basis of autism (Di Martino et al., 2014). We use the data from Preprocessed Connectome Project (Craddock et al., 2013) to discriminate individuals from Autism Spectrum Disorder from normal controls.
6. **HCP**,[9] Human Connectome Project contains imaging and behavioral data of healthy subjects (Van Essen et al., 2013). We use preprocessed rest-fMRI data from HCP900 release (Van Essen et al., 2012) to discriminate individuals from high IQ and low IQ. We used HCP rs-fMRI datasets to probe a different setting: data with longer acquisitions. Due to the data size, we limit the benchmarks here to pre-computed atlases.

3.1. rest-fMRI data processing: softwares and related

Data preprocessing. We preprocess COBRE, ADNI, and ADNIDOD. We use a standard protocol that includes: motion correction, fMRI co-registration to T1-MRI, normalization to the MNI template using SPM12[10], Gaussian spatial smoothing ($FWHM = 5mm$). The SPM based preprocessing pipeline is implemented through pypreprocess[11]- Python scripts relying on Nipype interface (Gorgolewski et al., 2011). All subjects were visually inspected and excluded from the analysis if they have severe scanner artifacts or head movements with amplitude larger than $2mm$. Since pre-processed rest-fMRI subjects from ABIDE and ACPI are available, we choose images preprocessed using C-PAC pipeline (Craddock et al.), without global signal regression. For ACPI, we choose linearly registered images using (Advanced Normalization Tools) ANTS and without motion scrubbing and no global signal regression. For already available preprocessed rest-fMRI subjects, we select the protocols such that it matches with the standard protocol we use. We have not done any additional preprocessing steps on ABIDE and ACPI.

Dataset	Prediction task	Groups
COBRE	Schizophrenia vs Control	65/77
ADNI	AD vs MCI	40/96
ADNIDOD	PTSD vs Control	89/78
ACPI	Marijuana use vs Control	62/64
ABIDE	Autism vs Control	402/464
HCP	High IQ vs Low IQ	213/230

Table 1: **Datasets and prediction tasks**, as well as the number of subjects in each group. COBRE - 142 subjects, ADNI - 136 subjects, ADNIDOD - 167 subjects, ACPI - 126 subjects, ABIDE - 866 subjects, HCP - 443. IQ represents fluid intelligence and we found number of subjects responded to IQ task are 788. The acquisition parameters of each dataset are summarized in Table A3.

Exclusion criteria. We not only exclude subjects based on visual inspection of preprocessed data, but also subjects that do not fall into binary classification groups, *eg* we removed subjects who had both bipolar disorder and schizoaffective groups from COBRE samples. For HCP, we select the subjects with single session and phase encoding in a left-to-right (LR) direction. Out of these selected subjects, we discriminate the low IQ from the high IQ individuals, where the data are split in 3 according to quantiles 0.333 and 0.666, and the subjects in the middle group are excluded to make the prediction easier in a binary classification setup (see Table 1 for numbers of subjects included in the analysis).

Cross validation and error measure. We perform cross-validation (CV) by randomly shuffling and splitting each dataset over 100 folds, forming two sets of subjects: 75% for training the classifier and learning brain atlases with data-driven models and the remaining 25% for testing on unseen data (Varoquaux et al., 2017). We create *stratified* folds, preserving the ratio of samples between groups. For each split, we measure the Area Under the Curve (AUC) from the Receiver Operating Characteristics (ROC) curve: 1 is a perfect prediction and .5 is chance. The final prediction scores in AUC (> 120k scores) are used to measure the impact of various choices in our prediction pipeline outlined below in results section.

Computations and implementation. Our experimental study consists of more than 240 types of pipelines (8 at-

[6] www.adni-info.org
[7] www.adni-info.org/DOD.html
[8] http://fcon_1000.projects.nitrc.org/indi/ACPI/html/
[9] We perform some additional experiments on the Human Connectome Project (HCP) data, to assess that our experimental results still hold when using high-quality datasets like *HCP*.
[10] www.fil.ion.ucl.ac.uk/spm/
[11] https://github.com/neurospin/pypreprocess

lases × 3 connectivity measures × 10 classifiers, plus some variants such as 3 filtering options and 3 covariance estimator options). These pipelines were run on each of 5 datasets for 100 CV folds. As a result, there are more than 500 000 pipeline fits, from the raw data to the supervised step, a heavy computational load. Technically, we rely on efficient implementations open-source scientific computing packages using Python 2.7: Nilearn v0.3 (Abraham et al., 2014b) to define brain atlases, extract representative time-series and timeseries confounds regression, and build connectivity measures. All machine-learning methods used for prediction *i.e.*, classifiers and cross-validation are implemented with scikit-learn v0.18.1 (Pedregosa et al., 2011). For visualization, we rely on Nilearn for brain-related figures while matplotlib is used (Hunter, 2007) for generating other figures.

4. Results: benchmarks of pipeline choices

We now outline which modeling choices have an important impact on predicting over diverse phenotypes from all rest-fMRI datasets.

We report in Table 2 the AUC scores obtained for all rest-fMRI datasets. The scores reported in the table are simplified to the optimal choice selection at each step in the pipeline which showed significant impact. These optimal choice of steps are discussed in following sections.

Impact of methodological choices. We study the prediction score of each pipeline *relative to the mean across pipelines* on each fold. This relative measure discards the variance in scores due to folds or datasets. From these relative prediction scores, we study the impact of the choice of each step in the prediction pipeline: choice of classifiers, connectivity parametrizations, and brain atlases. This is a multifactorial set of choices and there are two points of view on the impact of a choice for a given step. First, the impact of the choice for one step may be considered when the other steps are optimal, or close to optimal. Second, the impact of one step may be considered for all other choices for the other steps –marginally on the choice of other steps. Empirically, the two scenario lead to similar conclusions. In the following figures, we study the first situation, focusing on "good choices": given a choice for one step, we report data for top third highest performing scores (quantiles 0.666) for the choices in the other steps. Appendix C gives results for all scores, hence studying one choice, marginally upon the others.

Accuracy	COBRE	ADNIDOD	ADNI	ABIDE	ACPI
5^{th} percentile	75.5%	69.9%	57.8%	66%	42.5%
Median	86.2%	79.5%	72.5%	71.1%	55.4%
95^{th} percentile	95%	90.6%	84.5%	75.6%	68.7%

Table 2: **5th percentile, median and 95th percentile of accuracy scores in AUC over cross-validation folds (n = 100) for all five rest-fMRI datasets.** Accuracy scores reported correspond to optimal choices in functional connectivity prediction pipeline: brain regions defined with regions using DictLearn, connectivity matrices parametrized by their tangent-space representation, and an ℓ_2-regularized logistic regression as a classifier, as discussed below. Best prediction is achieved with schizophrenia vs control discrimination task on COBRE dataset at 86.2% (median).

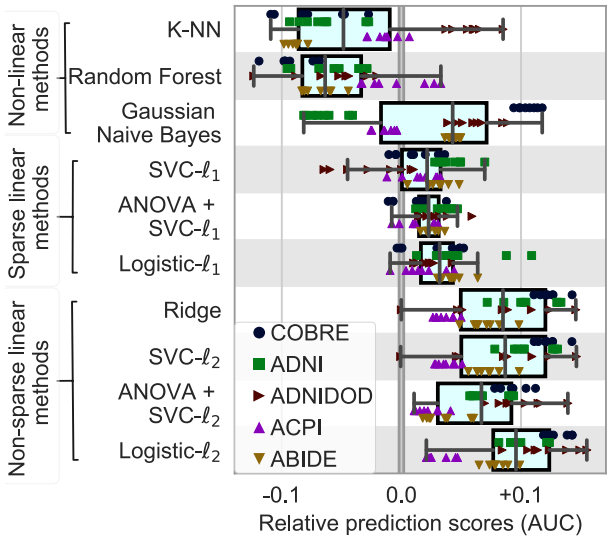

Figure 3: **Impact of classifier choices on prediction accuracy**, for all rest-fMRI datasets and all folds. For each classifier choice, only the top third highest performing scores are represented when varying the modeling choices for other steps in the pipeline: brain-region definition and connectivity parametrization. Figure A1 gives all the data points, not limited to good choices in the overall pipeline. Overall, ℓ_2-regularized linear classifiers perform better, with a slight lead for ℓ_2 logistic regression. The box plot gives the distribution across folds (n=100) and datasets (denoted by markers) of prediction score for a given choice (classifier) relative to the mean across all choices (regions-definition and connectivity parametrizations, classifiers). The box displays the median and quartiles, while the whiskers give the 5^{th} and 95^{th} percentiles.

4.1. Choice of classifier

Figure 3 summarizes the performances of classifiers on prediction scores for all rest-fMRI datasets. The results display a certain amount of variance across folds and datasets (*i.e.,* prediction targets). However, they show that non-sparse (ℓ_2-regularized) linear classifiers perform better, with a slight lead for logistic-ℓ_2. Using non-linear classifiers does not appear useful; neither does sparsity. The results in Figure 3 are conditional on a good choice for the other steps of the pipeline. The marginal performances of the different choices of classifiers –i.e. cconsidering all other choices in the pipeline– are shown in Figure A1. They show similar trends, leading to preferring ℓ_2-regularized linear classifiers.

4.2. Choice of connectivity parameterization

Figure 4 summarizes the impact of covariance matrix parametrization on the relative prediction scores for all rest-fMRI datasets. Tangent-space parametrization tends

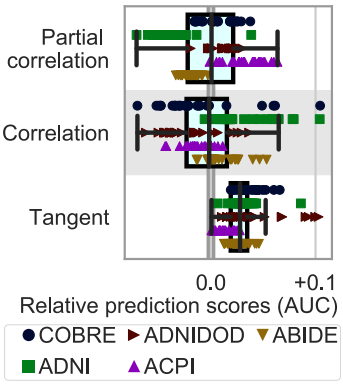

Figure 4: **Impact of connectivity parameterization on prediction accuracy**, for all rest-fMRI datasets and folds. For each parametrization choice, only the top third highest performing scores are represented when varying the modeling choices for other steps in the pipeline: brain-region definition and classifier. Figure A2 gives all the data points, not limited to good choices in the overall pipeline. Prediction using tangent space based connectivity parameterization displays higher accuracy with relatively lower variance than using full or partial correlation. The box displays the median and quartiles, while the whiskers give the 5^{th} and 95^{th} percentiles.

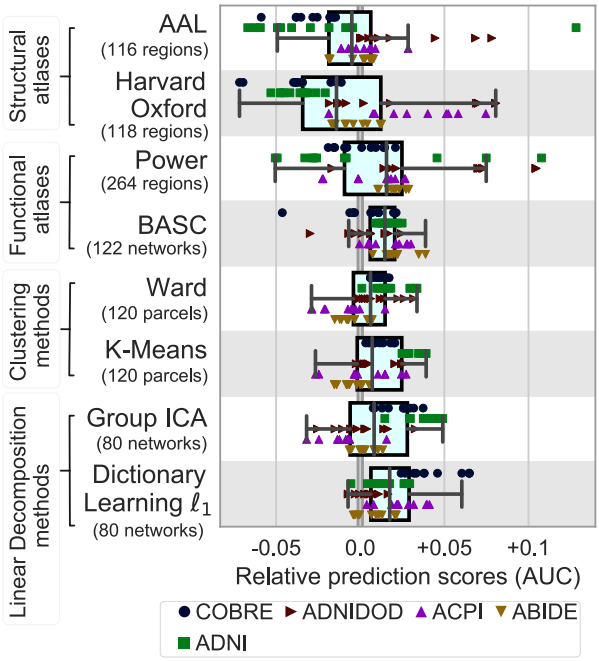

Figure 5: **Impact of region-definition method on prediction accuracy**, for all rest-fMRI datasets and folds. For each region-definition choice, only the top third highest performing scores are represented when varying the modeling choices for other steps in the pipeline: classifier and connectivity parametrization. Figure A3 gives all the data points, not limited to good choices in the overall pipeline. Learning atlases from rest-fMRI data tends the prediction for all tasks. By contrast anatomical atlases perform poorly over diverse tasks. The box displays the median and quartiles, while the whiskers give the 5^{th} and 95^{th} percentiles.

to outperform full correlations or partial correlations. Indeed, it performs better on average, but also has less variance across datasets (prediction targets) or folds. Results are similar for simpler variant of the tangent-space parametrization relying on a simple Euclidean mean rather than the full geometric (Riemannian) –see Appendix A for more details. While scores in Figure 4 are conditional on a good choice for other pipeline parameters, Figure A2 gives results marginal to all choices. In both settings, connectivity matrices built with tangent space parametrization give an improvement compared to full or partial correlations.

4.3. Choice of regions definition method

To find the preferred approaches to define brain regions, we proceed in two steps. First, for each method, we find the dimensionality that gives the best prediction. This holds for the BASC atlas, that comes in various dimensionalities, and for data-driven region-definition methods, for which we vary the dimensionality. Second, we study the prediction accuracy for each approach at the optimal dimensionality.

Best approach. Figure 5 summarizes the relative prediction performance of all choices of region-definition methods. While the systematic effects are small compared to the variance over the folds and the datasets, the general trend is that regions defined from functional data lead to better prediction than regions defined from anatomy. Using ℓ_1 dictionary learning to define regions from rest-fMRI data appears to be the best method, closely followed by ICA, which is also based on a linear decomposition model. Interestingly, BASC, an atlas pre-defined on unrelated rest-fMRI datasets using data-driven clustering technique, performs almost as well as the best regions-extraction method applied to the rest-fMRI data of interest. Unlike other pre-defined atlases, like Harvard Oxford or AAL, that lack some crucial functional regions. The BASC atlas (Bellec et al., 2010) is readily available online, and is thus easy to apply to data. Figure 5 shows the impact of region-definition approach conditional on good choices in the other steps of the pipeline, however studying the impact of region-definition independently of other choices (Figure A3). Both comparisons highlight that defining regions from functional data gives the best-performing pipelines, and that linear-decomposition methods are to be preferred.

Optimal dimensionality. The choice of the best dimensionality for each approach paints a less clear picture (Figure 6): a range of dimensionalities lead for good prediction for each method[12]. We find that there is a very soft optimum: prediction reaches a plateau as the number of extracted networks increases, and then slowly decreases for some methods. To favor the most parsimonious model, in this paper we choose to work at the lower end of the

[12]Note that, these curves are shown for the optimal choices found above: an ℓ_2-penalized logistic regression as a classifier, and tangent-space parametrization to clarify the interpretation.

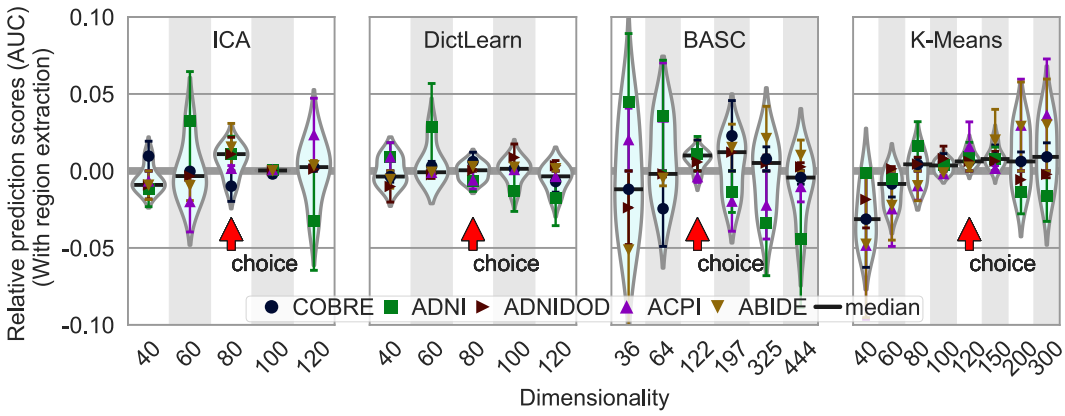

Figure 6: **Impact of the number of regions in atlases on prediction accuracy**. The figure shows the distribution of the relative accuracy AUC scores across methods on the five rest-fMRI datasets, as a function of the number of regions. Horizontal bars (black) represent the median of the relative scores for the given number of regions. The chosen dimensionality for each method is indicated by a red arrow and was selected as the one with lowest variance in the error, and a median above zero.

plateau (red arrow on Figure 6): simpler models for better stability and statistical control. While this choice is not clear cut, the curves also suggest that, in a reasonable range, it does not have a large impact on prediction accuracy. Note that the dimensionality here corresponds to the number of networks, these are then broken up into separate regions. We find that the typical number of regions at the optimal is around 150 (Appendix E).

Localized regions or distributed networks. Nodes of the functional connectomes may be defined from localized regions, or the distributed networks that naturally arise from approaches such as ICA or dictionary learning. The choice of one over the other has little impact over prediction, though there is slight, non significant, benefit to using regions (Figure A5).

4.4. Larger datasets and pre-computed atlases

To investigate the consistency of analytics choices for higher-quality datasets, we perform extra benchmarks including the HCP data. As this data comprises much longer time-series, we restrict our analysis to pre-computed atlases, that alleviate computational costs. We share the resulting time-series and scripts to reproduce our analysis[13].

Figure 7 summarizes the impact of method choice on the prediction accuracy for all six different cohorts. This experiment outline similar tradeoffs as the others: functional atlas pre-computed with dictionary learning (here MODL, from Mensch et al. (2016a)), tangent-space parametrization, and ℓ_2-regularized classifiers are preferable. This experiment is not as systematic as the other, as a very large dataset like HCP would require much more computing power to study region extraction[14]. Yet, even for region-definition methods, it outlines similar trends than when tuning the regions to the data at hand.

4.5. Filtering, global signal, and covariance estimation

When extracting functional signal for connectivity modeling, there are many options to reject confounds, including temporal filtering or global signal mean regression (Fox et al., 2009; Murphy et al., 2009; Power et al., 2012). Also, to extract a structure reflecting well brain connectivity, it has been show that careful covariance estimation is useful and that sparse inverse covariance methods perform well (Smith et al., 2011; Varoquaux et al., 2010b). For both of these steps, our experiments for predictive modeling applications do not reveal clear preferences.

Different time-series filtering approaches (band-pass or global-signal regression) make no visible differences on prediction accuracy (Figure A8). A likely reason is that the supervised step can learn a predictor that is independent of the corresponding noise in the signal.

With regards to covariance estimation, we also investigate the empirical covariance (maximum likelihood estimator) and sparse inverse covariance (Appendix H). The empirical covariance can only be used to compute correlations –as partial correlation or tangent parametrization require an invertible covariance matrix– in which case it performs similarly as the Ledoit-Wolf estimator (see Figure A9). Sparse inverse covariance performs as well or worse than the Ledoit-Wolf estimator. This latter estimator is easier to use, as it is faster and does not require setting a regularization parameter. Learning discriminant connectivity patterns across conditions does not seem to require the same regularization –sparsity– as identifying the brain connectivity structure.

[13] github.com/KamalakerDadi/benchmark_rsfMRI_prediction

[14] To ensure a correct nested cross-validation and avoid circularity (overfitting), data-driven region-extraction methods must be run on each fold, hence several hundred time for each pipeline configuration.

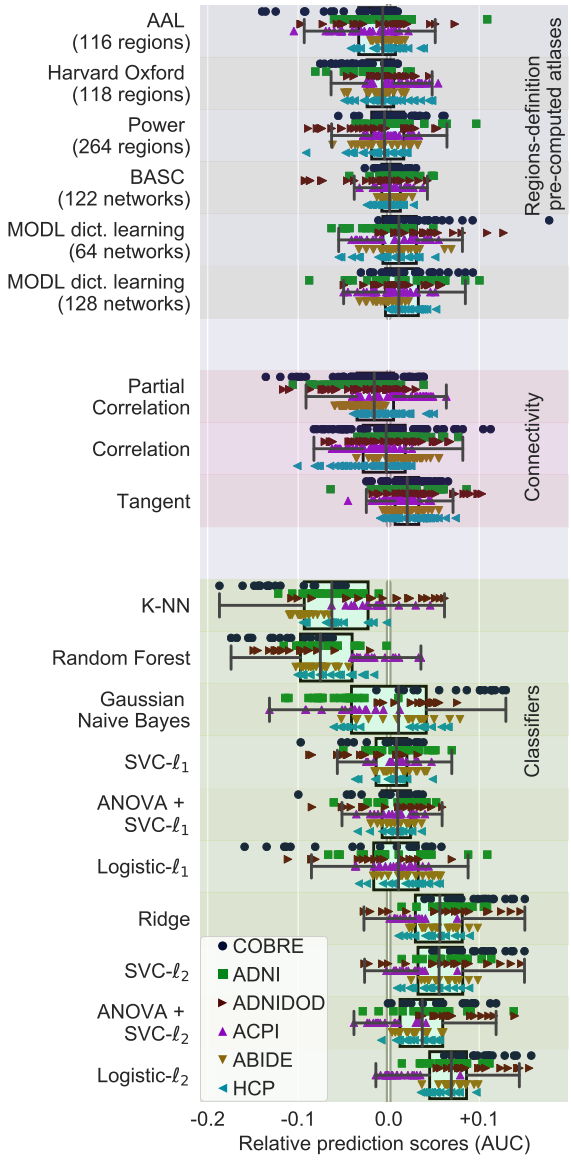

Figure 7: **Pipelining choices with precomputed regions, across six datasets**: Marginal distribution of relative prediction scores, using only pre-computed atlases for regions definition, where MODL is a parcellation built using a form of Online dictionary learning. Restricting to pre-computed regions and adding a different dataset (HCP) gives results consistent with Figure 5, 4, and 3: best choices are regions defined functionally, with decomposition methods (MODL) followed by clustering methods (BASC), tangent-space parametrization of connectivity, and ℓ_2-regularized logistic regression. The box displays the median and quartiles, while the whiskers give the 5^{th} and 95^{th} percentiles. Table A1 reports the corresponding absolute scores.

5. Discussion

An increasing amount of studies use predictive models on functional connectomes, for instance in population-imaging settings to relate brain activity to psychological traits or to build biomarkers of pathologies. While the basic steps of a pipeline are fairly universal –definition of brain regions, construction of an interaction matrix, and supervised learning– studies in the literature show many methodological variants (Table A2). Recommendations on methods that perform well can increase practitioner's productivity and limit vibration effects that risk undermining the reliability of biomarkers (Varoquaux, 2017). A challenge to such recommendations is the heterogeneity of prediction settings, for instance across different acquisition centers or clinical questions.

Here, we investigate methodological choices across 6 databases covering different clinical questions and behavioral task. We systematically compare commonly used functional connectome-based prediction methods. We find that some trends emerge, despite a large variance due to variability across subjects –visible across the folds– and across cohorts and clinical questions. Non-sparse linear models, such as logistic regression, appear as a good default choice of classifier. The lack of success of sparse approaches suggests that the discriminant signal is distributed across the functional connectome for the tasks we study. The tangent-space parametrization of functional connectomes brings improvements to prediction accuracy. With regards to nodes of the functional connectomes, defining them from rest-fMRI data gives slight benefits in prediction. Linear decomposition methods, such as dictionary learning or ICA, are good approaches to define these nodes from the rest-fMRI data at hand. Unlike clustering methods base on "hard" assignment, they provide a soft assignment to regions, enabling to capture a form of uncertainty in the definition of regions. Alternatively, the MODL[15] (Mensch et al., 2016a) or BASC (Bellec et al., 2010) atlases provide good readily-available nodes that simplify the process and alleviate computational cost. The good analytic performance of pre-computed atlases is promising and calls for further study. Establishing standard atlases brings significant computational benefits, as the definition of regions and the extraction of signal is the most computation-intensive part of the pipeline –in particular when performed inside a nested cross-validation loop. We found that using around 100 networks (corresponding to 150 regions) was sufficient for good prediction, though for many region-definition approaches a finer resolution did not hurt average prediction accuracy but only increased variance.

Overall, these results are consistent with the practice of the field. Preliminary comparisons in Abraham et al. (2017) on a single cohort revealed similar trends though ICA had performed poorly while here, with more systematic benchmarking, it appears to be a good solution. ICA has been used to define functional parcellations or nodes of functional connectomes by many groups (Kiviniemi et al., 2009; Rashid et al., 2014; Smith et al., 2015; Miller et al., 2016). More generally, it is well recognized that the nodes should be defined to match functional networks (Smith et al., 2011). Logistic regression, or the closely-related

[15] https://team.inria.fr/parietal/files/2018/10/MODL_rois.zip

SVM, is the go-to classifier for many. Tangent-space parametrization of the connectivity matrix is more exotic, probably due to the mathematical complexity of its original presentation. However, it is gaining traction outside of methods studies (Colclough et al., 2017; Ng et al., 2017) and is simple to implement, as summarized in Appendix A.

To enable comparison across different cohorts, we focused on 2-class classification problems. However, the results in terms of regions definition and connectivity parametrization should extend to other supervised learning settings, such as regression –*e.g.* for age prediction (Liem et al., 2017)– multi-output approaches as with Canonical Correlation Analysis popular in large-scale population imaging settings (Smith et al., 2015; Miller et al., 2016) for dimensional approaches to psychology.

Limitations and Challenges. The main limitation of our study is probably that we had to make choices and focus on the most popular methods. Indeed, to study systematically methods avoiding overfit requires computational-intense nested cross-validation (where the nesting is required to set the methods' internal parameters). In particular, we did not investigate Total-Variation constrained dictionary learning (TV-MSDL, Abraham et al. (2013)). This approach defines regions by imposing spatial structure in a linear-decomposition model. In a previous study, we found it promising (Abraham et al., 2017), but it entailed too large of a computational cost for this multi-cohort study. Another important class of methods that this study did not investigate are biomarkers based on graph-theoretical approaches. Indeed, we benchmarked variants of a specific pipeline –region definition, followed by construction of a connectivity matrix, and supervised learning on it. Graph-theoretical approaches are an additional step to add to this pipeline. A full study of all options with this additional step would result in a combinatorial explosion of pipelines and prohibitive computational costs. We hope that the good choices of regions for edge-level models outlined in this study is also a good one for graph-theoretical approaches and that further studies can focus on exploring only a subset of the options covered here.

With evolving techniques, characteristics of data change, and optimal choices may evolve. However, the consistency of results on HCP suggest that our conclusions apply to high-quality datasets using state of the art techniques. A potential concern is the low accuracy for markers of drug abuse in subjects from ACPI datasets. Nevertheless, our pipelines achieved similar accuracy as reported in a previous study on the same data (Meszlényi et al., 2016). Finally, the analysis performed here can only outline trends across datasets. Indeed, the study does not establish that a pipeline choice strictly dominates others (see notes on statistical analysis in Appendix J), but it gives expected improvements. In term of expected improvement, the choice of classifier is the most important, as going from a poor to a good choice can improve the AUC by more than .1. Both choice of region and choice of parametrization bring smaller expected improvements.

Step	**Recommendation**
1: region extraction	Functional regions, *eg* Dictionary learning or ICA
2: connectivity matrix	Tangent-space embedding
3: supervised learning	Non-sparse linear model, *eg* logistic regression or SVM

Table 3: **Recommendations for rest-fMRI based prediction pipeline.**

6. Conclusion

Predictive models on rest-fMRI bring the promise of robust and reliable biomarkers: given new brain imaging data, they should give accurate predictions of clinics or behavior (Woo et al., 2017). The framework of the functional connectomes grounds well the analysis of rest-fMRI; yet instantiating it still calls for many arbitrary choices.

Our study reveals trends that can provide good defaults to practitioners, summarized on Table 3: regions defined from functional data, for instance with ICA or dictionary learning as in the pre-computed MODL atlas, representing connectivity with the tangent embedding of covariance matrices, and using a non-sparse linear model, such as a logistic regression. In particular, good defaults can limit the combinatorial explosion of analytic pipelines, which decreases the computational cost of running a study and makes its conclusion more robust statistically. Yet, as it is well known in machine learning (Wolpert, 1996), there cannot be a one-size-fits-all solution to data analysis: optimal choices will differ on datasets with very different properties from the datasets studied here.

Acknowledgments

This work is funded by NiConnect project (ANR-11-BINF-0004_NiConnect) and the CATI project. We also thank to the open-source data community and pre-processed data initiatives for giving access to rest-fMRI datasets. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate;

Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

7. References

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

Abraham, A., Dohmatob, E., Thirion, B., Samaras, D., Varoquaux, G., 2014a. Region segmentation for sparse decompositions: better brain parcellations from rest fMRI. Frontiers in neuroinformatics 8.

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

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014b. Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics 8.

Anderson, A., Douglas, P.K., Kerr, W.T., Haynes, V.S., Yuille, A.L., Xie, J., Wu, Y.N., Brown, J.A., Cohen, M.S., 2014. Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. NeuroImage 102, 207–219.

Arbabshirani, M.R., Kiehl, K.A., Pearlson, G.D., Calhoun, V.D., 2013. Classification of schizophrenia patients based on resting-state functional network connectivity. Frontiers in Neuroscience 7.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. NeuroImage 145, 137–165.

Barachant, A., Bonnet, S., Congedo, M., Jutten, C., 2013. Classification of covariance matrices using a riemannian-based kernel for bci applications. Neurocomputing 112, 172 – 178.

Bassett, D.S., Nelson, B.G., Mueller, B.A., Camchong, J., Lim, K.O., 2012. Altered resting state complexity in schizophrenia. NeuroImage 59, 2196–2207.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. Trans Med Im 23, 137.

Behzadi, Y., Restom, K., Liau, J., Liu, T., 2007. A component based noise correction method (compcor) for BOLD and perfusion based fMRI. Neuroimage 37, 90.

Bellec, P., Rosa-Neto, P., Lyttelton, O., Benali, H., Evans, A., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51, 1126.

Biswal, B., Mennes, M., Zuo, X., Gohel, S., Kelly, C., Smith, S., Beckmann, C., et al., 2010. Toward discovery science of human brain function. Proc Ntl Acad Sci 107, 4734.

Breiman, L., 2001. Random forests. Machine Learning 45, 5.

Brier, M.R., Mitra, A., McCarthy, J.E., Ances, B.M., Snyder, A.Z., 2015. Partial covariance based functional connectivity computation using ledoit–wolf covariance regularization. NeuroImage 121, 29–38.

Brown, C.J., Hamarneh, G., 2016. Machine learning on human connectome data from mri. arXiv:1611.08699 .

Calhoun, V., Sui, J., Kiehl, K., Turner, J., Allen, E., Pearlson, G., 2012. Exploring the psychosis functional connectome: Aberrant intrinsic networks in schizophrenia and bipolar disorder. Frontiers in Psychiatry .

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

Carp, J., 2012. On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. Frontiers in neuroscience 6.

Chen, G., Ward, B.D., Xie, C., Li, W., Wu, Z., Jones, J.L., Franczak, M., Antuono, P., Li, S.J., 2011. 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, 213–221.

Cheng, W., Ji, X., Zhang, J., Feng, J., 2012. Individual classification of ADHD patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. Frontiers in Systems Neuroscience 6.

Colclough, G.L., Smith, S.M., Nichols, T.E., Winkler, A.M., Sotiropoulos, S.N., Glasser, M.F., Van Essen, D.C., Woolrich, M.W., 2017. The heritability of multi-modal connectivity in human brain activity. eLife 6, e20178.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13, 21–27.

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., Yan, C., Bellec, P., 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives ¡br /¿. Frontiers in Neuroinformatics .

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S.S., Yan, C., Li, Q., Lurie, D., Vogelstein, J., Burns, R., Colcombe, S., Mennes, M., Kelly, C., Di Martino, A., Castellanos, F.X., Milham, M., . Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). Frontiers in Neuroinformatics .

Craddock, R.C., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S., 2009. Disease state prediction from resting state functional connectivity. Magnetic Resonance in Medicine 62, 1619.

Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human brain mapping 33, 1914.

Craddock, R.C., Tungaraza, R.L., Milham, M.P., 2015. Connectomics and new approaches for analyzing human brain functional connectivity. GigaScience 4, 13.

Desikan, R., S., Ségonne, F., Fischl, B., Quinn, B., T., Dickerson, B., C., Blacker, D., Buckner, R., L., Dale, A., M., Maguire, R., P., Hyman, B., T., Albert, M., S., Killiany, R., J., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968.

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., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659–667.

Dodero, L., Minh, H.Q., Biagio, M.S., Murino, V., Sona, D., 2015. Kernel-based classification for brain connectivity graphs on the riemannian manifold of positive definite matrices, in: International Symposium on Biomedical Imaging (ISBI), IEEE.

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

Drysdale, A.T., Grosenick, L., et al., 2016. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine .

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fmri. Trends in cognitive sciences 20, 425–443.

Dubois, J.C., Galdi, P., Paul, L.K., Adolphs, R., 2018. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. bioRxiv , 257865.

Elliott, P., Peakman, T.C., et al., 2008. The UK biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. International Journal of Epidemiology 37, 234–244.

Fei, F., Jie, B., Zhang, D., 2014. Frequent and discriminative subnetwork mining for mild cognitive impairment classification. Brain Connectivity 4, 347–360.

Fletcher, P., Joshi, S., 2007. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87, 250.

Fox, M.D., Zhang, D., Snyder, A.Z., Raichle, M.E., 2009. The global signal and observed anticorrelated resting state brain networks. Journal of neurophysiology 101, 3270–3283.

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

Gellerup, D., 2016. Discriminating Parkinsons Disease Using Functional Connectivity and Brain Network Analysis. Ph.D. thesis. University of Texas - Arlington.

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform 5, 13.

Greicius, M., 2008. Resting-state functional connectivity in neuropsychiatric disorders. Current opinion in neurology 21, 424.

Guo, H., Cao, X., Liu, Z., Li, H., Chen, J., Zhang, K., 2012. Machine learning classifier using abnormal brain network topological metrics in major depressive disorder. NeuroReport 23, 1006–1011.

Hallquist, M.N., Hillary, F.G., 2018. Graph theory approaches to functional network organization in brain disorders: A critique for a brave new small-world. bioRxiv , 243741.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer.

Hunter, J.D., 2007. Matplotlib: A 2d graphics environment. Computing In Science & Engineering 9, 90–95.

Iidaka, T., 2015. Resting state functional magnetic resonance imaging and neural network classified autism and control. Cortex 63, 55–67.

Jie, B., Shen, D., Zhang, D., 2014. Brain connectivity hypernetwork for MCI classification, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Springer International Publishing, pp. 724–732.

Kahnt, T., Chang, L.J., Park, S.Q., Heinzle, J., Haynes, J.D., 2012. Connectivity-based parcellation of the human orbitofrontal cortex. Journal of Neuroscience , 6240–6250.

Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., 2015. Identifying patients with alzheimer's disease using resting-state fMRI and graph theory. Clinical Neurophysiology 126, 2132–2141.

Kiviniemi, V., Starck, T., Remes, J., Long, X., Nikkinen, J., Haapea, M., Veijola, J., et al., 2009. Functional segmentation of the brain cortex using high model order group PICA. Hum Brain Map 30, 3865.

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

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., et al., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188.

Mazumder, R., Hastie, T., 2012. The graphical lasso: New insights and alternatives. Electronic journal of statistics 6, 2125.

Mensch, A., Mairal, J., Thirion, B., Varoquaux, G., 2016a. Dictionary Learning for Massive Matrix Factorization, in: International Conference on Machine Learning, pp. 1737–1746.

Mensch, A., Mairal, J., Thirion, B., Varoquaux, G., 2018. Stochastic subsampling for factorizing huge matrices. IEEE Transactions on Signal Processing 66, 113–128.

Mensch, A., Varoquaux, G., Thirion, B., 2016b. Compressed Online Dictionary Learning for Fast Resting-State fMRI Decomposition, in: Proc. ISBI, p. 1282.

Meszlényi, R., Peska, L., Gál, V., Vidnyánszky, Z., Buza, K., 2016. A model for classification based on the functional connectivity pattern dynamics of the brain, in: 2016 Third European Network Intelligence Conference (ENIC), pp. 203–208.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B., 2012. A supervised clustering approach for fMRI-based inference of brain states. Pattern Recognition 45, 2041.

Milazzo, A.C., Ng, B., Jiang, H., Shirer, W., Varoquaux, G., Poline, J.B., Thirion, B., Greicius, M.D., 2014. Identification of mood-relevant brain connections using a continuous, subject-driven rumination paradigm. Cereb. Cortex 26, 933–942.

Miller, K.L., Alfaro-Almagro, F., et al., 2016. Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nature Neuroscience .

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The alzheimers disease neuroimaging initiative. Neuroimaging Clinics of North America 15, 869.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? Neuroimage 44, 893–905.

Ng, B., Dressler, M., Varoquaux, G., Poline, J.B., Greicius, M., Thirion, B., 2014. Transport on Riemannian Manifold for Functional Connectivity-based Classification, in: MICCAI - 17th International Conference on Medical Image Computing and Computer Assisted Intervention.

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.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825.

Pennec, X., Fillard, P., Ayache, N., 2006. A Riemannian framework for tensor computing. Int J Comp Vision 66, 41.

Perrot, M., Riviere, D., Tucholka, A., Mangin, J.F., 2009. Joint Bayesian Cortical Sulci Recognition and Spatial Normalization, in: IPMI.

Power, J., Cohen, A., Nelson, S., Wig, G., Barnes, K., Church, J., Vogel, A., Laumann, T., Miezin, F., Schlaggar, B., Petersen, S., 2011. Functional network organization of the human brain. Neuron 72, 665–678.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. Neuroimage 59, 2142–2154.

Pruett, J.R., Kandala, S., Hoertel, S., Snyder, A.Z., Elison, J.T., Nishino, T., Feczko, E., Dosenbach, N.U., Nardos, B., Power, J.D., Adeyemo, B., Botteron, K.N., McKinstry, R.C., Evans, A.C., Hazlett, H.C., Dager, S.R., Paterson, S., Schultz, R.T., Collins, D.L., Fonov, V.S., Styner, M., Gerig, G., Das, S., Kostopoulos, P., Constantino, J.N., Estes, A.M., Petersen, S.E., Schlaggar, B.L., Piven, J., 2015. Accurate age classification of 6 and 12 month-old infants based on resting-state functional connectivity magnetic resonance imaging data. Developmental Cognitive Neuroscience 12, 123–133.

Qiu, A., Lee, A., Tan, M., Chung, M.K., 2015. Manifold learning on brain functional networks in aging. Medical Image Analysis 20, 52 – 60.

Rahim, M., Thirion, B., Varoquaux, G., 2017. Population-shrinkage of covariance to estimate better brain functional connectivity, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 460–468.

Rashid, B., Arbabshirani, M.R., Damaraju, E., Cetin, M.S., Miller, R., Pearlson, G.D., Calhoun, V.D., 2016. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. NeuroImage 134, 645–657.

Rashid, B., Damaraju, E., Pearlson, G.D., Calhoun, V.D., 2014. Dynamic connectivity states estimated from resting fMRI identify differences among schizophrenia, bipolar disorder, and healthy control subjects. Frontiers in human neuroscience 8, 897.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D., 2010. Decoding brain states from fMRI connectivity graphs. NeuroImage .

Rosa, M.J., Portugal, L., Hahn, T., Fallgatter, A.J., Garrido, M.I., Shawe-Taylor, J., Mourao-Miranda, J., 2015. Sparse network-based models for patient classification using fMRI. NeuroImage 105.

Rubinov, M., Sporns, O., 2011. Weight-conserving characterization of complex functional brain networks. NeuroImage 56, 2068.

Shen, H., Wang, L., Liu, Y., Hu, D., 2010. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. NeuroImage 49, 3110.

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

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science 22, 1359.

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

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E., Glasser, M.F., Ugurbil, K., Barch, D.M., Van Essen, D.C., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nature neuroscience 18, 1565–1567.

Sporns, O., Tononi, G., Kotter, R., 2005. The human connectome: a structural description of the human brain. PLoS Comput Biol 1, e42.

Thirion, B., Varoquaux, G., Dohmatob, E., Poline, J., 2014. Which fMRI clustering gives good brain parcellations? Name: Frontiers in Neuroscience 8, 167.

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

Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Della Penna, S., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., Yacoub, E., 2012. The human connectome project: A data acquisition perspective. NeuroImage 62, 2222–2231.

Van Essen, D.C., Smith, et al., 2013. The wu-minn human connectome project: an overview. Neuroimage 80, 62–79.

Vanderweyen, D., , Munsell, B.C., Mintzer, J.E., Mintzer, O., Gajadhar, A., Zhu, X., Wu, G., Joseph, J., 2015. Identifying abnormal network alterations common to traumatic brain injury and alzheimer's disease patients using functional connectome data, in: Machine Learning in Medical Imaging. Springer International Publishing, pp. 229–237.

Varoquaux, G., 2017. Cross-validation failure: small sample sizes lead to large error bars. NeuroImage .

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B., 2010a. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, in: MICCAI.

Varoquaux, G., Craddock, R.C., 2013. Learning and comparing functional connectomes across subjects. NeuroImage 80, 405.

Varoquaux, G., Gramfort, A., Poline, J.B., Thirion, B., 2010b. Brain covariance selection: better individual functional connectivity models using population prior, in: NIPS.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage 145, 166–179.

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

Wang, L., Fei, F., Jie, B., Zhang, D., 2014. Combining multiple network features for mild cognitive impairment classification, in: 2014 IEEE International Conference on Data Mining Workshop.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236.

Wee, C.Y., Yap, P.T., Shen, D., 2016. Diagnosis of autism spectrum disorders using temporally distinct resting-state functional connectivity networks. CNS Neuroscience & Therapeutics 22, 212–219.

Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neuroscience & Biobehavioral Reviews 57, 328.

Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural computation 8, 1341–1390.

Wong, E., Anderson, J.S., Zielinski, B.A., Fletcher, P.T., 2018. Riemannian regression and classification models of brain networks applied to autism, in: International Workshop on Connectomics in Neuroimaging, Springer. pp. 78–87.

Woo, C.W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience 20, 365–377.

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

Zhu, D., Li, K., Terry, D.P., Puente, A.N., Wang, L., Shen, D., Miller, L.S., Liu, T., 2013. Connectome-scale assessments of structural and functional connectivity in MCI. Human Brain Mapping 35, 2911–2923.

Appendix A. Computing the covariance tangent-space

Most of the methods that we study are readily-available in several computing environments, including Matlab and Python with a variety of well-maintained implementations. However, the only library that provides the tangent-space parametrization of covariance matrices is the Nilearn Python library [16]. To facilitate reproducing our analysis in different environments, we describe here how to compute this parametrization with a few simple formulas. The computation is made of two step: First a group average covariance matrix $\mathbf{\Sigma}^\star$ is computed from the covariances of the training subjects: $\{\mathbf{\Sigma}_i, i \in \mathcal{T}rain\}$. Second, it is used to transform covariance matrices, in the train set or the test set.

Computing the group average. As with any analysis based on covariance or correlation matrices, it is preferable to compute individual covariances from time series with an estimator that ensures well-conditioned matrices. The Ledoit and Wolf (2004) estimator is a good default choice (Varoquaux and Craddock, 2013; Brier et al., 2015).

Strictly speaking, the group average should be computed according to the geometry of covariance matrices (Varoquaux et al., 2010a; Pennec et al., 2006). This is a Frechet mean, which is computed by minimizing a cost function for instance using algorithm 3 of Fletcher and Joshi (2007). A simpler approach relies on using the Euclidean mean, which we found to give almost the same predictive performance. In this case, the formula of the mean is the standard one:

$$\text{Euclidean mean}: \mathbf{\Sigma}_\star = \frac{1}{n_{train}} \sum_{i \in \mathcal{T}rain} \mathbf{\Sigma}_i \qquad \text{(A.1)}$$

Transforming covariance matrices. Given the group reference covariance matrix $\mathbf{\Sigma}_\star$, covariance matrices are transformed in the tangent-space representation by whitening them as follows (Varoquaux et al., 2010a). Computations are easily written with eigenvalues decompositions[17]: given a subject's covariance matrix $\mathbf{\Sigma}_i$,

1. Compute the whitened matrix $\tilde{\mathbf{\Sigma}}_i = \mathbf{\Sigma}_\star^{-1/2} \mathbf{\Sigma}_i \mathbf{\Sigma}_\star^{-1/2}$:

$$\tilde{\mathbf{\Sigma}}_i \leftarrow \mathbf{U}^T \mathbf{\Delta}^{-\frac{1}{2}} \mathbf{U}\, \mathbf{\Sigma}_i\, \mathbf{U}^T \mathbf{\Delta}^{-\frac{1}{2}} \mathbf{U} \qquad \text{(A.2)}$$

   where $\mathbf{U}^T \mathbf{\Delta}\, \mathbf{U} = \mathbf{\Sigma}_\star$ by eigen-value decomposition, and operations on the diagonal matrix $\mathbf{\Delta}$ are element-wise operation applied to the diagonal.

2. Compute the matrix logarithm $\mathbf{logm}\, \tilde{\mathbf{\Sigma}}_i$:

$$\mathbf{logm}(\tilde{\mathbf{\Sigma}}_i) = \tilde{\mathbf{U}}^T \log(\tilde{\mathbf{\Delta}}_\mathbf{i})\, \tilde{\mathbf{U}} \qquad \text{(A.3)}$$

   where $\tilde{\mathbf{\Sigma}}_i = \tilde{\mathbf{U}}^T \tilde{\mathbf{\Delta}}_i\, \tilde{\mathbf{U}}$ and the logarithm is applied to the diagonal elements of $\tilde{\mathbf{\Delta}}_i$.

Finally, the resulting matrix is turned to a vector and its entries are used as a features for the classifier.

The motivation from these transformations arises from the fact that covariance matrices –or correlations matrices– form a specific manifold of the $\mathbb{R}^{p \times p}$ matrices. Their structure is broken by standard additive arithmetic's: the difference of two covariances may create a matrix that does not correspond to the covariance matrix of a signal. Optimal statistical analysis calls for following the structure of the manifold (Pennec et al., 2006). The tangent-space parametrization is a simple way to approximate this structure by Euclidean geometry, in which standard additions and subtractions can be used (Varoquaux et al., 2010a).

With regards to statistical analysis, the structure of covariance matrices appears as constraints, or dependencies, between the coefficients of the matrix. As a result, these coefficients alone form a poor representation for second-level statistical analysis. The tangent-space approximation yields a parametrization of the problem in which features are *i.i.d.* (Varoquaux et al., 2010a). Such a parametrization is optimal for statistical learning.

Appendix B. Time-series signals extraction

In this appendix, we give more details on time-series extraction, to complement subsection 2.3. After defining brain ROIs, we extract a representative time-series for each ROI in each subject. For atlases composed of non-overlapping ROIs as can be seen in Figure 2 (bottom row), we simply compute the weighted average of the fMRI time series signals over all voxels within that specific region. For fuzzy overlapping ROIs, such as the atlases driven by CanICA and DictLearn as shown in Figure 2 (top row), we use ordinary least squares regression to unmix the signal in each voxel as the additive decomposition of signals over several overlapping ROIs. This is the same procedure as in (Abraham et al., 2017). Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be the subject-specific signals, written as p voxels by n timepoints, and $\mathbf{V} \in \mathbb{R}^{k \times p}$ the atlas of k maps supported on p voxels. We estimate $\mathbf{U} \in \mathbb{R}^{n \times k}$, the set of time series for each ROI, using:

[16] http://nilearn.github.io/

[17] All covariance matrices are symmetric definite positive, and well-conditioned if estimated with the Ledoit-Wolf approach

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}\|^2$$

At the signal-extraction level, we regress out confounds or non-neural information (Varoquaux and Craddock, 2013). As confounding time-series we use: 10 CompCor (Behzadi et al., 2007) on the whole brain and 6 motion related. We remove motion-related signal only for COBRE, ADNI and ADNIDOD as they are provided as raw data. We have not done any additional preprocessing steps on already preprocessed public datasets like ABIDE[18], ACPI[19]. The signal of each region is also then normalized, detrended and bandpass-filtered between 0.01 and 0.1Hz. All these steps are done with Nilearn v0.3.

Investigating filtering choices. At signal extraction level, we perform additional experiments to assess the impact of filtering strategies (low-pass filtering, global signal regression) on prediction accuracy. See Appendix F for more details. Overall, we observe no significant differences between filtering strategies —low-pass filter and no global signal regression, low-pass filter and global signal regression, no low-pass filter and no global signal regression. See Appendix Appendix F for complete comparisons.

Appendix C. Comparing each step marginally on the others

The figures in the main part of the paper summarize the impact of one modeling choice in the pipeline conditionally on nearly-optimal choices for the other steps. Here we compare the modeling choices at each step marginally on all other choices, *i.e.* considering all results and including well performing and poorly performing pipelines. This approach studies each step of the method independently from the other steps. The results are overall similar to performing the conditional analysis, however the variance is larger, as the plots pool together pipelines that perform well and pipelines that perform poorly.

Appendix C.1. Step:3 Choice of classifiers

Figure A1 shows the relative impact of classifier choices on prediction accuracy. All ℓ_2 regularized classifiers are performing markedly better than other considered classifiers, with a logistic regression as the best performer.

Appendix C.2. Step 2: Choices of connectivity parametrizations

Figure A2 shows the relative impact of connectivity parametrization on prediction accuracy when considering all choices for the other pipeline steps. The tangent-space parametrization performs better than correlation or partial correlation and gives less variance.

Appendix C.3. Step:1 Choices of region-definition methods

Figure A3 shows the relative impact of region-definition choices on prediction accuracy for all choices in remaining steps. Atlases which are functionally derived lead to good performances. Linear-decomposition methods, appear as the best choice for region-definition methods, in particular, Online Dictionary Learning.

[18] http://preprocessed-connectomes-project.org/abide/
[19] http://fcon_1000.projects.nitrc.org/indi/ACPI/html/

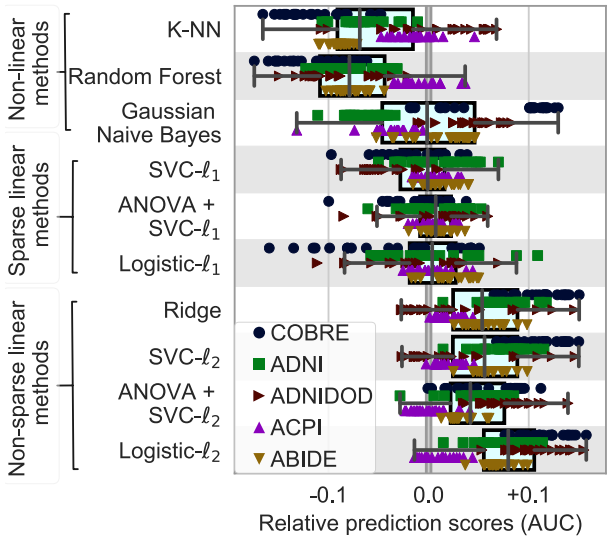

Figure A1: **Impact of classifiers on prediction accuracy**: Marginal distribution of the relative prediction accuracy of all classification choices for all rest-fMRI datasets. The results are obtained covering all the choices for the remaining steps *i.e.,* atlases and connectivity parametrizations. Non-sparse linear models perform well and ℓ_2-regularized logistic regression appears as the best choice.

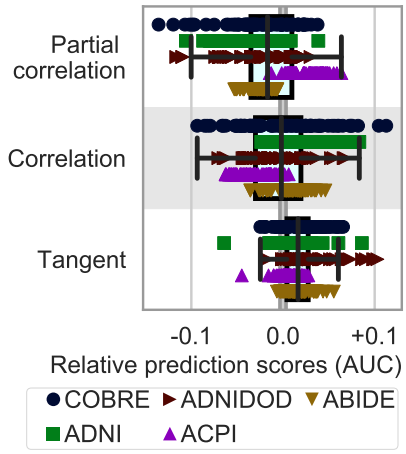

Figure A2: **Impact of connectivity parametrization on prediction accuracy**: Marginal distribution of relative prediction accuracies for all rest-fMRI datasets, considering all pipelining options. Tangent space parametrization displays the highest accuracy and smallest variation across all datasets and folds.

Appendix D. Additional experiments on region-definition methods

Here we give additional results related to step 1 of the pipeline: defining nodes, formed of brain regions or brain networks. An important parameter to chose is the optimal dimensionality *dim i.e.*, how many networks are needed to predict from the rest-fMRI images. Another choice is whether these networks should be broken up into simply connected regions – with a region-extraction step– or whether distributed networks can be readily used. These two parameters may be important in the comparison of brain-region definition, in particular for data-driven approaches such as linear decomposition or clus-

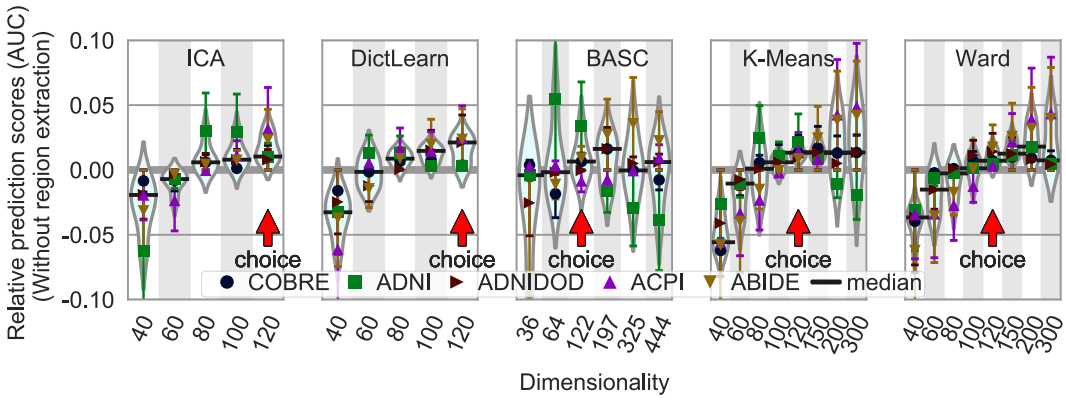

Figure A4: **Impact of the number of regions in data-driven atlases on prediction accuracy**. For each method, the distribution of relative prediction accuracy (AUC) is displayed as a function of the number of regions, across rest-fMRI datasets. The horizontal bar (black) represents the median of the relatively mean scores for each dimensionality. The whiskers of each data point represent the 95% confidence interval for a given dataset across the folds. The optimal choices (red arrow) are selected as the minimal variance score with a median close to the maximum. The prediction scores are obtained for the optimal pipeline setting, involving tangent-space parameterization and logistic regression-ℓ_2 classifier.

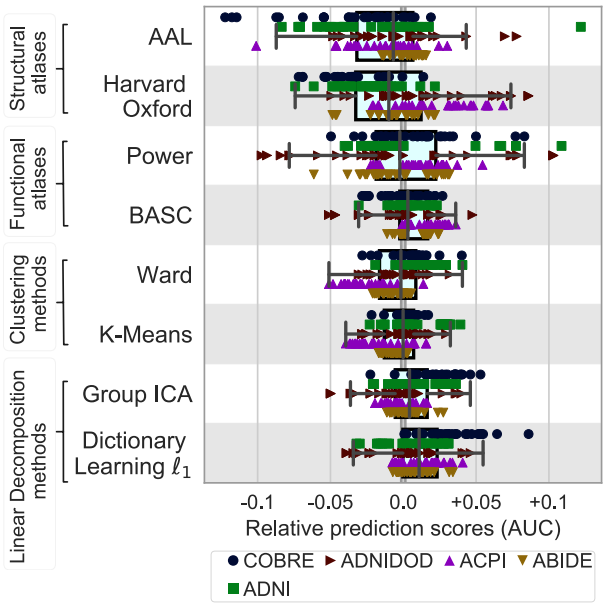

Figure A3: **Impact of regions-definition method on prediction accuracy**: Marginal distribution of relative prediction accuracy per region-definition approach across all rest-fMRI datasets. This is obtained while considering all pipelining options uniformly in all other steps. Among all pre-defined atlases, BASC is best. In data-driven based atlases, the best choice is Online dictionary learning.

tering methods. In our study, we found that atlases learned using linear decomposition methods give a good prediction.

Appendix D.1. Varying dimensionality without region extraction

While the results in the main part of the manuscript are presented after regions extraction, Figure A4 studies the optimal dimensionality without region extraction. The optimal choices are not far from what we have observed with region extraction (Figure 6). As shown with a red arrow, the dimensions shown good prediction impact are: ICA and DictLearn - 80 networks, K-Means and Ward - 120 clusters, 122 networks with BASC. As dimensionality goes higher, we observe that the variance in prediction accuracy increases.

Appendix D.2. Defining nodes with regions or networks

One the optimal dimensionality chosen with and without region extraction, we compare for each method whether region-definition with regions or networks gives best prediction. Figure A5 summarizes the results but with no obvious clear-cut conclusions. The figure shows the distribution of differences in prediction scores between regions-based approaches and network-based approaches. There is a very slight tendency to favor region-based approaches, but the trend is not significant.

Appendix E. Number of regions extracted for different network-definition methods

Data-driven method tend to naturally extract networks rather than regions: ICA and dictionary learning give distributed networks, while KMeans gives clusters made of different connect components. Only Ward clustering readily gives ROIs as it has connectivity constraints.

We use a region-extraction procedure to go from networks to regions (Abraham et al., 2014a). In this appendix, we study the distribution of the number of regions extracted from networks obtained with different approaches for increasing dimensionality. Figure A6 summarizes this distribution each rest-fMRI dataset. The computational cost is too high to learn

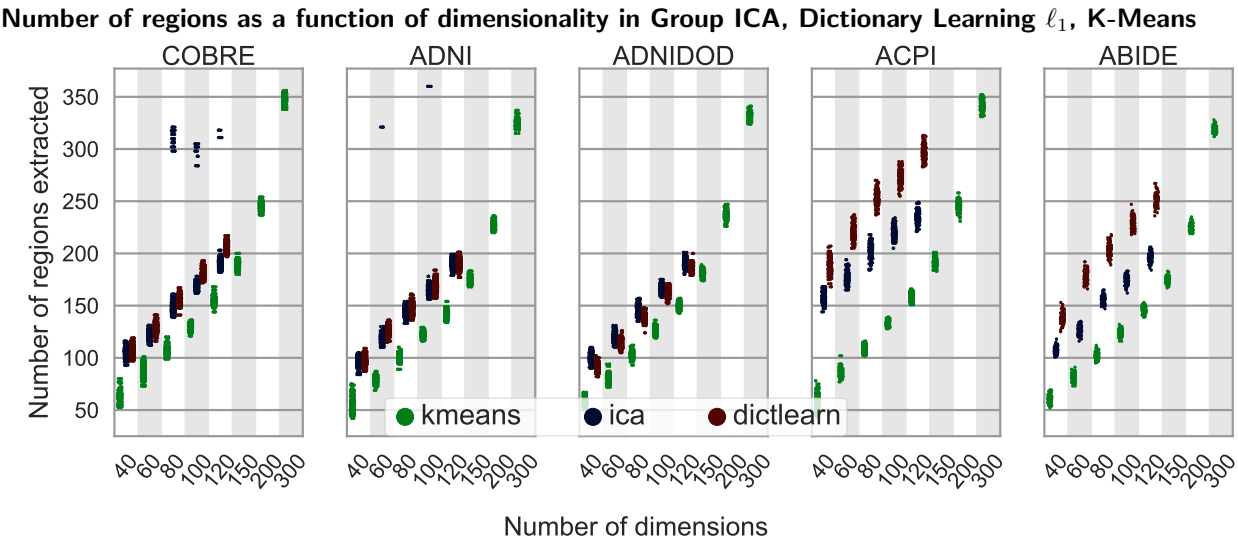

Figure A6: **Distribution of number of regions extracted on brain maps given the increase in number of dimensions.**

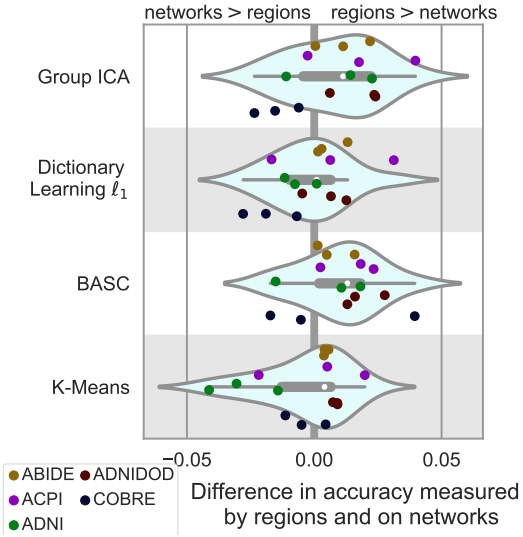

Figure A5: **Impact of regions- vs network-based representation on prediction accuracy**: Each data point represents the difference in relatively mean prediction scores between regions –i.e. with extraction of connected components– and networks –without such an extraction. Each time, the optimal dimensionality for the corresponding option is used. Points on the left side indicate that the network representation is better, while points on the right side indicate superior performance of the region-based representation. Results are shown for each rest-fMRI dataset. Regions-based representations appear better suited, but this effect is not significant. Note that Ward clustering does not appear in this figure as it extracts connected regions.

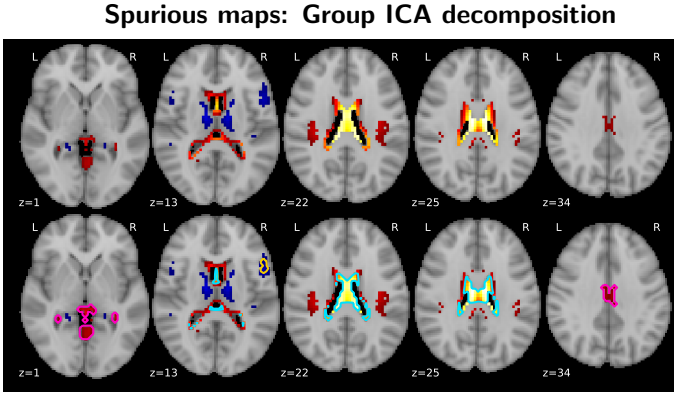

Figure A7: **Spurious map obtained using Group ICA method on COBRE dataset (top) and regions extracted from this spurious map outlined with contours (bottom)**: This shows a simple example of the degeneracy of the distribution of regions reaching 300 even for low dimensionality such as *dim* = 80. Full distribution of regions on various atlases are shown on Figure A6 for comparison.

spatial maps for each split using Group ICA and Online Dictionary Learning. Hence we show outcomes for dimensions up to 120. For each method, the distribution of the number of regions regularly increases with increasing dimensionality: the number of regions is roughly proportional to the number of networks.

For the optimal choice of dimensionality (*dim* = 80 for Group ICA and DictLearn and *dim* = 120 for K-Means as studied on Figure 6), the average number of regions lies on average around 150.

Region extraction from ICA network sometimes displays an ill-controlled behavior, for instance for *dim* = 80, 100, 120 on COBRE, where a small number of folds lead to 300 or more regions. We believe that these high number of regions are extracted from an noisy ICA map with little structure as shown in Figure A7. Dictionary learning, which has a criteria on sparsity of of the maps, does not create such unstructured maps, and therefore does not suffer from the same problem.

Appendix F. Experiments on filtering time-series

The results in the main part of the paper (eg Figure 3), (Figure 4, (Figure 5) as well as the marginal distribution figures shown on Appendix C, are established with low-pass filter of the time series and without global signal regression.

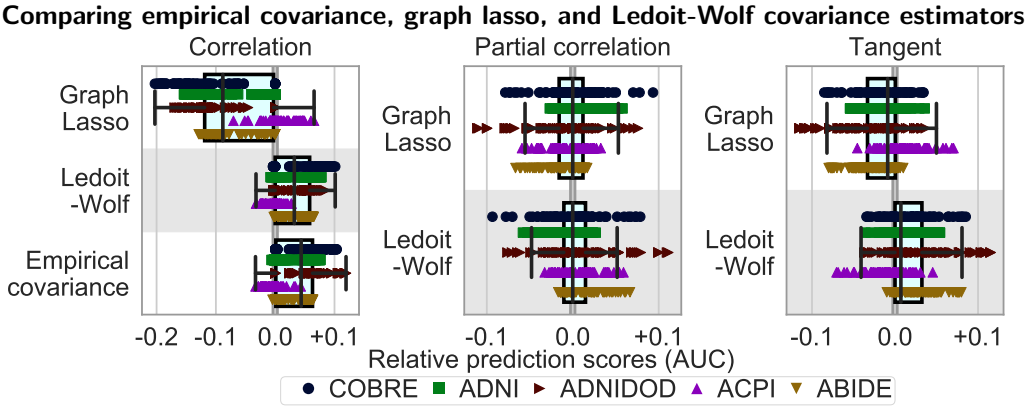

Figure A9: **Impact of covariance estimator, for the different connectivity parametrizations**: Marginal distribution of relative prediction scores per connectivity method across covariance structures for all rest-fMRI datasets. This is obtained by considering the good choices in dimensionality as studied in the main figures. Overall, using empirical covariance instead of the Ledoit-Wolf shrinkage give huge difference in prediction perfomances; note that partial correlation and tangent parametrization require the use of shrinkage. Using graph lasso tend to perform lower compared with other estimators.

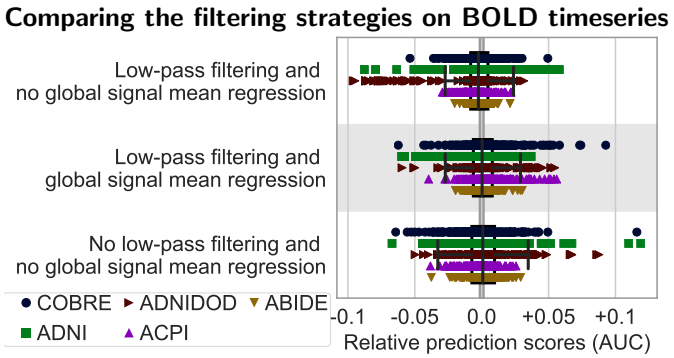

Figure A8: **Temporal signal filtering strategies on prediction accuracy (AUC) for five rs-fMRI datasets**: Distribution of relative prediction scores showed no big differences across three filtering strategies: lowpass filter and no global signal regression, lowpass filter and global signal regression, no lowpass filter and no global signal regression.

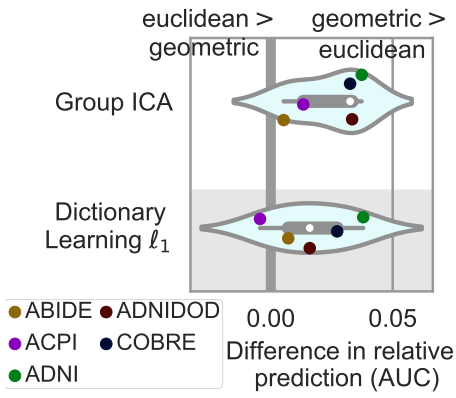

Figure A10: **Comparison between geometric & Euclidean distance metrics on tangent space parametrization of covariances**: Distribution of difference in prediction score (AUC) between these 2 metrics outlined for the dictionary learning atlas. Geometric distance based connectivity parametrization yields higher accuracy than Euclidean distance based parametrization.

Here, we investigate the impact of filtering strategies on prediction accuracy. We compare three filtering schemes: low-pass filtering and no global signal regression, low-pass filtering and global signal regression, no low-pass filtering and no global signal regression. Figure A8 shows the outcome for all of these filtering combinations: there is no significant differences on prediction accuracy across all rs-fMRI datasets.

Appendix G. Tangent-space parametrization: Euclidean versus Geometric mean

We also compare two variants of tangent-space parametrization: using the geometric (Frechet) mean and using the simple Euclidean mean for $\mathbf{\Sigma}^\star$ (Appendix A). Figure A10 highlights the difference between the relative prediction scores for each dataset. We compare between the optimal choices in brain atlas methods *i.e.* GroupICA and DictLearn. Geometric mean based parametrization gives a slightly better prediction accuracy than Euclidean mean based parametrization. Euclidean mean gives a reduced computational cost, but the Geometric mean is better justified in theory.

Appendix H. Covariance estimators: unregularized, ℓ_1-regularized

Figure A9 compares different covariance estimators for the three connectivity parametrizations. .

Appendix H.1. Empirical covariance: unregularized

We considered empirical covariance as our choice of unregularized covariance estimator. Using empirical covariance in the pipelines did not improve the predictions. We found little difference when compared with Ledoit-Wolf using full correlation as can be seen from Figure A9.

The empirical covariances are ill-conditioned matrices and not invertible, hence they cannot be used for partial correlations or tangent parametrization. Ledoit-Wolf shrinkage or sparse inverse covariance estimators overcome such limitations.

Dataset	Accuracy		
	5^{th} percentile	Median	95^{th} percentile
HCP	62.3%	67.8%	74.9%
COBRE	72.4%	83.9%	91.9%
ACPI	39.8%	53.7%	68.7%
ADNI	57.4%	72.2%	85.0%
ADNIDOD	69.9%	80.2%	88.0%
ABIDE	64.5%	69.7%	75.3%

Table A1: **5th percentile, median and 95th percentile of accuracy scores in AUC over cross-validation folds (n = 100) for all six rest-fMRI datasets including HCP.** Accuracy scores reported correspond to optimal choices in functional connectivity prediction pipeline as shown on Figure 7: precomputed atlas defined using massive online dictionary learning (MODL), connectivity matrices parametrized by their tangent-space representation, and an ℓ_2-regularized logistic regression as a classifier.

Appendix H.2. GraphLasso: ℓ_1-regularized estimator

Sparse inverse covariance is effective at recovering brain connectivity (Smith et al., 2011; Varoquaux et al., 2010b). We use the graphical lasso (Friedman et al., 2008) to estimate sparse covariance matrices, in order to study **full correlation**, **partial correlation** and **tangent space parametrization** in comparison with ℓ_2-regularized covariance estimator (Ledoit-Wolf). Such estimator requires the choice of a regularization parameter that sets the amount of sparsity. A good regularization parameter typically depends on the amount of time points available and the number of nodes. For each dataset and choice of regions, we used an inner-loop optimization to set regularization parameter: we test two parameters values, 0.5 and 0.2, where 0 corresponds to a non regularized covariance and 1 or above to fully sparse covariances. For all datasets, 0.5 was the best trade-off as using 0.2 gave ill-conditioned results. Note that parameter selection for the graphical lasso on a wide and varied dataset is challenging as the graphical lasso runs into convergence problems when covering a variety of regularization parameters on covariance matrices with different properties. This happens very seldom, but with 100 folds on 1 500 subjects, and a dozen different atlases, the problem makes automatic parameter selection difficult. The convergence problem is well understood theoretically: the algorithm is a primal-dual algorithm and errors can accumulate between the primal and the dual solution (Mazumder and Hastie, 2012).

Figure A9 shows the impact of graph lasso on connectivity parametrizations. Overall, we observed no improvements in the prediction results based on graph lasso with respect to Ledoit-Wolf.

Appendix I. A review of current practices in functional connectome-based classification

Table A2 summarizes about the list of methods used for prediction studies in diverse psychiatric diseases.

Appendix J. A note of statistical analysis of cross-validation

Cross-validation cannot easily be used to reject null hypotheses when comparing analytic choices as the multiple per-fold values that it gives are not independent (as discussed in appendices of Varoquaux (2017)). Rather, they are resampling estimates, and therefore give a posterior on the prediction accuracy: the prediction accuracy is a property of the model fitted on the data at hand –see sec 8.4 Hastie et al. (2009) for the link between resampling and Bayesian statistics. Comparing distribution of cross-validation accuracy thus cannot establish frequentist p-values –that the differences observed between models are not due to chance– but it can give posterior predictive distribution, and therefore the expected improvement of one model compared to another on new data.

Reference	Clinical question & Accuracy	#Subjects	Functional matrix	# Nodes (type of nodes)	Classifier
(Nielsen et al., 2013)	ASD 60%	964	Pearson's correlation	7266 (coordinates)	SVM
(Abraham et al., 2017)	ASD 67%	811	Covariance: Full & Partial correlation, Tangent-space parametrization	84 (data-driven)	Gaussian Naive Bayes, Random Forests, Ridge* Lasso, SVM-ℓ_1 & ℓ_2*, ANOVA + SVM-ℓ_2
(Iidaka, 2015)	ASD 90%	640	Pearson's correlation	90 (anatomical)	Kernel Discriminant analysis (Neural network)
(Dodero et al., 2015)	ASD 60.76%	94	graph Laplacian (Riemannian manifold)	264 (coordinates)	Kernel SVM-ℓ_2
(Wee et al., 2016)	ASD 71%	92	Pearson's correlation	116 (anatomical)	SVM-ℓ_2 & Lasso SVM-ℓ_2 & Lasso
(Anderson et al., 2014)	ADHD 67%	730	Network modularity & centrality	90 (data-driven)	C4.5 decision trees
(Cheng et al., 2012)	ADHD 76%	730	Pearson's: Partial & full correlations	90 (anatomical)	Kernel SVM-ℓ_2
(Rashid et al., 2016)	Schizo, bipolar 59.12%	273	Covariance: Full correlation	100 (data-driven)	SVM
(Bassett et al., 2012)	Schizo 75%	58	Network modularity & centrality	90 (anatomical)	SVM-ℓ_2
(Arbabshirani et al., 2013)	Schizo 96%	56	Pearson's correlation	9 (data-driven)	Bayes, Fisher, Logistic Perceptron, SVM, K-nearest neighbor*, Gaussian Naive & Quadratic Bayes, Binary decision trees, Radial Basis Function-SVM
(Shen et al., 2010)	Schizo 92%	52	Pearson's correlation	116 (anatomical)	C-means
(Guo et al., 2012)	MDD 79%	76	Network modularity & centrality	90 (anatomical)	Radial Basis Function-SVM* Neural Network* C4.5 decision trees Linear Discriminant Analysis Logistic Regression
(Craddock et al., 2009)	MDD 95%	40	Pearson's correlation	15 (coordinates)	SVM-ℓ_1
(Rosa et al., 2015)	MDD 85%	38	Inverse covariance	137 (pre-defined)	SVM-ℓ_1
(Gellerup, 2016)	PD 84%	45	Pearson's correlation	264 (coordinates)	SVM-ℓ_2
(Khazaee et al., 2015)	AD/MCI/NC 88%	168	Network modularity & centrality	90 (pre-defined)	SVM-ℓ_2
(Vanderweyen et al., 2015)	AD/TBI/NC 82%	69	Partial correlations	264 (coordinates)	SVM-ℓ_2 & Lasso
(Chen et al., 2011)	AD 87%	55	Pearson's correlation	116 (anatomical)	Linear Discriminant analysis
(Fei et al., 2014)	MCI 97%	37	Frequent sub-network mining (gSpan)	116 (anatomical)	Graph-kernel
(Jie et al., 2014)	MCI 95%	37	Hyper-network graph	116 (anatomical)	Multi-kernel SVM-ℓ_2
(Wang et al., 2014)	MCI 97%	37	Local cluster coefficient + Sub-network (gSpan)	116 (anatomical)	Multi-kernel SVM-ℓ_2
(Zhu et al., 2013)	MCI 96%	28	Pearson's correlation	358 (coordinates)	SVM-ℓ_2
(Dosenbach et al., 2010)	Age groups 91%	122	Pearson's correlation	160 (meta-analyses)	SVM-ℓ_2 (regression)
(Pruett et al., 2015)	Clinical risk 81%	128	Pearson's correlation	230 (meta-analyses)	SVM-ℓ_2
(Qiu et al., 2015)	Age $r = 0.59$	178	Inverse Covariance (Riemannian manifold)	80	Linear regression ℓ_2
(Ng et al., 2014)	Before/After motor learning 98%	51	Pearson's correlation (Riemannian manifold)	78 (data-driven)	SVM-ℓ_2
(Colclough et al., 2017)	Heritability	820	Partial & full correlations (Riemannian manifold)	39 (data-driven)	–

Table A2: **A comprehensive list of functional connectome-based prediction studies on psychiatric diseases**. This table demonstrates the variants of methods in the prediction pipeline for various clinical questions. SVM - Support Vector Machines, NN - Neural Network, ANOVA - Analysis of Variance. * - denotes well performed classifiers respective to their current study.

Dataset	Acquisition type	Slice thickness (mm)	FoV (mm)	Voxel size (mm)	Matrix size	TR ($msec$)	TE ($msec$)	Flip angle (°)	Number of volumes†
COBRE	T2*-weighted gradient-echo EPI	3.5	240	$3.75 \times 3.75 \times 4.55$	64×64	2000	29	75	150
ADNI	T2*-weighted gradient-echo EPI	3.3	240	$3.31 \times 3.31 \times 3.31$	64×64	3000	30	80	135
ADNIDOD	T2*-weighted gradient-echo EPI	3.3	240	$3.28 \times 3.28 \times 3.3$	64×64	2900	30	90	160
ACPI (MTA)	T2*-weighted	1.20	256	$1.0 \times 1.0 \times 1.2$	256×256	2170	4.33	7	180
ABIDE									
Caltech	T2*-weighted single-shot EPI	3.5	224	$3.5 \times 3.5 \times 3.5$	64×64	2000	30	75	146
CMU_a	T2*-weighted	3.0	192	$3.0 \times 3.0 \times 3.0$	64×64	2000	30	73	236
CMU_b	T2*-weighted	3.0	192	$3.0 \times 3.0 \times 3.0$	64×64	2000	30	73	316
KKI	T2*-weighted	3.0	256	$3.0 \times 3.0 \times 3.0$	84×84	2500	30	75	152
MaxMun	T2*-weighted gradient-echo EPI	4.0	192	$3.0 \times 3.0 \times 4.0$	64×64	3000	30	80	116
NYU	T2*-weighted	4.0	240	$3.0 \times 3.0 \times 4.0$	80×80	2000	15	90	176
Olin	T2*-weighted	4.0	220	$3.4 \times 3.4 \times 4.0$	64×64	1500	27	60	206
OHSU	T2*-weighted	3.8	240	$3.8 \times 3.8 \times 3.8$	64×64	2500	30	90	78
SDSU	T2*-weighted gradient-echo EPI	3.4	220	$3.4 \times 3.4 \times 3.4$	64×64	2000	30	90	176
SBL	T2*-weighted	2.72	220	$2.75 \times 2.75 \times 2.72$	80×80	2200	30	80	196
Stanford	T2*-weighted	4.5	200	$3.125 \times 3.125 \times 4.5$	64×64	2000	30	80	236
Trinity	T2*-weighted	3.5	240	$3.0 \times 3.0 \times 3.5$	80×80	2000	28	90	146
UCLA_1	T2*-weighted	4.0	192	$3.0 \times 3.0 \times 4.0$	64×64	3000	28	90	116
UCLA_2	T2*-weighted	4.0	192	$3.0 \times 3.0 \times 4.0$	64×64	3000	28	90	116
Leuven_1	T2*-weighted	4.0	230	$3.59 \times 3.59 \times 4.0$	64×64	1667	33	90	246
Leuven_2	T2*-weighted	4.0	230	$3.59 \times 3.59 \times 4.0$	64×64	1667	33	90	246
UM_1	T2*-weighted	3.0	220	$3.438 \times 3.438 \times 3.0$	64×64	2000	30	90	296
UM_2	T2*-weighted	3.0	220	$3.438 \times 3.438 \times 3.0$	64×64	2000	30	90	296
Pitt	T2*-weighted	4.0	200	$3.1 \times 3.1 \times 4.0$	64×64	1500	25	70	196
USM	T2*-weighted	3.0	220	$3.4 \times 3.4 \times 3.0$	64×64	2000	28	90	236
Yale	T2*-weighted	4.0	220	$3.4 \times 3.4 \times 4.0$	64×64	2000	25	60	196
HCP	T2*-weighted gradient-echo EPI	2.0	208	$2.0 \times 2.0 \times 2.0$	104×90	720	33.1	52	1200

Table A3: **Parameters used for the acquisition of rs-fMRI datasets**. HCP - Human Connectome Project, MTA - Multimodal Treatment of Attention Deficit Hyperactivity Disorder - the acquisition site of ACPI datasets, CMU - Carnegie Mellon University, KKI - Kennedy Krieger Institute, MaxMun - Ludwig Maximilians University Munich, NYU - New York University Langone Medical Center, OHSU - Oregon Health and Science University, SDSU - San Diego State University, SBL - Social Brain lab, UCLA - University of California, Los Angeles, UM - University of Michigan, Pitt - University of Pittsburgh, USM - University of Utah School of Medicine, EPI - Echo planar imaging, TR - Repetition time, TE - Echo time, FoV - Field of View, †- the number of volumes reported are what have been included in the analysis pipelines.