

RESEARCH ARTICLE

# Predicting functional networks from region connectivity profiles in task-based versus resting-state fMRI data

Javier Rasero<sup>1,5</sup>, Hannelore Aerts<sup>2</sup>, Marlis Ontivero Ortega<sup>2,3</sup>, Jesus M. Cortes<sup>1,4</sup>, Sebastiano Stramaglia<sup>5,6\*</sup>, Daniele Marinazzo<sup>2</sup>

**1** Biocruces Health Research Institute. Hospital Universitario de Cruces. E-48903, Barakaldo, Spain, **2** Faculty of Psychology and Educational Sciences, Department of Data Analysis, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium, **3** Neuroinformatics Department, Cuban Center for Neuroscience (CNeuro), La Habana, Cuba, **4** Ikerbasque, The Basque Foundation for Science, E-48011, Bilbao, Spain, **5** Dipartimento di Fisica, Università degli Studi "Aldo Moro" Bari, Italy, **6** Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy

\* [sebastiano.stramaglia@ba.infn.it](mailto:sebastiano.stramaglia@ba.infn.it)



**OPEN ACCESS**

**Citation:** Rasero J, Aerts H, Ontivero Ortega M, Cortes JM, Stramaglia S, Marinazzo D (2018) Predicting functional networks from region connectivity profiles in task-based versus resting-state fMRI data. PLoS ONE 13(11): e0207385. <https://doi.org/10.1371/journal.pone.0207385>

**Editor:** Claudio Gallicchio, Università degli Studi di Pisa, ITALY

**Received:** May 17, 2018

**Accepted:** October 29, 2018

**Published:** November 12, 2018

**Copyright:** © 2018 Rasero et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data are owned by the Human Connectome Project (HCP). In order to access data, interested researchers should fill the register form on the web site of HCP <http://www.humanconnectomeproject.org/> and provide a short summary of the research project to be accomplished using the data. Others would be able to access these data in the same manner as the authors. The authors did not have any special access privileges to the data that others would not have. Code for downloading and preprocessing of the data used in this work, and replication of

## Abstract

Intrinsic Connectivity Networks, patterns of correlated activity emerging from “resting-state” BOLD time series, are increasingly being associated with cognitive, clinical, and behavioral aspects, and compared with patterns of activity elicited by specific tasks. We study the reconfiguration of brain networks between task and resting-state conditions by a machine learning approach, to highlight the Intrinsic Connectivity Networks (ICNs) which are more affected by the change of network configurations in task vs. rest. To this end, we use a large cohort of publicly available data in both resting and task-based fMRI paradigms. By applying a battery of different supervised classifiers relying only on task-based measurements, we show that the highest accuracy to predict ICNs is reached with a simple neural network of one hidden layer. In addition, when testing the fitted model on resting state measurements, such architecture yields a performance close to 90% for areas connected to the task performed, which mainly involve the visual and sensorimotor cortex, whilst a relevant decrease of the performance is observed in the other ICNs. On one hand, our results confirm the correspondence of ICNs in both paradigms (task and resting) thus opening a window for future clinical applications to subjects whose participation in a required task cannot be guaranteed. On the other hand it is shown that brain areas not involved in the task display different connectivity patterns in the two paradigms.

## Introduction

Functional magnetic resonance imaging (fMRI) has become a powerful tool to study brain dynamics with relatively fine spatial resolution. One popular paradigm, given the indirect nature of the method, is the block design, alternating task blocks with blocks of passive rest.

results and plots is available at <https://github.com/jrasero/Predicting-icns>.

**Funding:** JR acknowledges financial support from the Minister of Education, Language Policy and Culture (Basque Government) under Doctoral Research Staff Improvement Programme.

**Competing interests:** The authors have declared that no competing interests exist.

Task-associated brain activity can then be inferred by contrasting the level of BOLD signal between task and resting blocks in each voxel.

In a seminal study by Biswal *et al.* [1], it has been shown that spontaneous low frequency fluctuations ( $< 0.1$  Hz) in BOLD signal, present even when subjects are not performing any specific task, also give rise to correlated patterns. In particular, after identification of seed regions in the sensorimotor cortex by a bilateral finger tapping task fMRI protocol, the authors found synchronous fluctuations of BOLD time courses between these seed regions and homologous areas in the opposite hemisphere. This finding demonstrated the existence of a sensorimotor network even at resting state. In addition to the sensorimotor network, several other functional networks that are activated in task designed experiments have since been identified during resting state, such as the dorsal and ventral attention networks, and the fronto-parietal control network [2, 3]. Among these functional networks, the default mode network (DMN) [4], including the posterior cingulate cortex, precuneus and medial prefrontal cortex is perhaps the most ubiquitous and iconic. Regions belonging to this network exhibit a deactivation when cognitive tasks are performed and an increase in internal connectivity during rest [5].

In order to identify these Intrinsic Connectivity Networks (ICNs) emerging from resting-state measurements, the most common methods are seed based analysis and independent component analysis. In seed-based analysis, voxels within a region of interest (“seed region”) are selected and their average BOLD time course is correlated with that of all other voxels in the brain. Voxels showing a correlation with the seed region above a certain threshold are then considered to be functionally related to the region of interest. The main disadvantage of this approach, however, is that it requires *a priori* selection of seed regions. Independent component analysis (ICA), in contrast, is a model-free approach, requiring very few assumptions [6]. This method separates the BOLD time courses of all voxels into different spatial components and ensures maximum statistical independence among them [7]. Since this is a pure data-driven approach, in contrast to the “seed region” scenario, there is no clear link between the components found and the specific brain functions. As a result, component labeling might not be straightforward, especially at the individual subject level. Furthermore, the number of components to be retained is an arbitrary parameter to be provided, that might be fitted by a supplied criterion, such as the minimization of a cost-sensitive function that allows an optimal match of the similarity reached by ICA predictions with respect to the observed dataset.

ICNs in resting state experimental designs can thus be regarded as identifying regions with similar BOLD signal profiles. Therefore, different clustering methods can also be applied to explore the structure of whole-brain BOLD time series, in an attempt to identify functional networks on the basis of resting-state fMRI data (see [8–11] for the use of different unsupervised clustering methods). In order to solve the component labeling problem, supervised learning approaches can also be utilized, where the identification and prediction of these networks is accomplished after fitting the decision boundaries that separate each class by directly supplying the instance labels. In previous studies, a multilayer perceptron was found to be the optimal method to map the topography of ICNs in healthy young control subjects, after which it was applied to a small sample of patients undergoing surgical treatment for intractable epilepsy or tumor resection [12, 13]. The ability to reliably estimate functional networks from resting-state fMRI data would have important clinical implications, for example in neurosurgeons’ pre-surgical planning when patients are unable to cooperate with the task-based paradigm [14]. Furthermore, predictions from models uniquely trained on resting state measurements from healthy people have been demonstrated to robustly match activation profiles of pre-surgical populations that usually suffer from a great deal of variability [15].

The aim of the present study is to extend previous works looking at reconfiguration of brain networks between task and resting-state conditions, and to highlight the brain regions

which are more affected by the change of network configurations in task vs. rest. By selecting the most optimal model, we intend to quantify to which extent one can predict ICNs training in task-based paradigm and predict the ICN patterns emerging purely from resting-state fMRI data. The development of a general framework for ICN identification using supervised machine learning methods could thus shed further light on the overlap between the patterns emerging from both task and resting-state fMRI data.

## Materials and methods

### Subjects

Data used in the preparation of this work were obtained from the MGH-USC Human Connectome Project (HCP) database (<https://ida.loni.usc.edu/login.jsp>) [16]. The HCP project (Principal Investigators: Bruce Rosen, M.D., Ph.D., Martinos Center at Massachusetts General Hospital; Arthur W. Toga, Ph.D., University of California, Los Angeles, Van J. Weeden, MD, Martinos Center at Massachusetts General Hospital) is supported by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH) and the National Institute of Neurological Disorders and Stroke (NINDS). Collectively, the HCP is the result of efforts of co-investigators from the University of California, Los Angeles, Martinos Center for Biomedical Imaging at Massachusetts General Hospital (MGH), Washington University, and the University of Minnesota.

We considered both resting-state and task-based fMRI data from 282 unrelated healthy subjects provided by the s900 release in the Human Connectome Project. All images were reconstructed using algorithm r227. This reconstruction algorithm directly performs the separation of the multi-band multi-slice in k-space, in contrast to a previous reconstruction algorithm (r177), where the separation occurred after transforming the fully acquired sampled data to frequency space along the read-out direction.

### Data acquisition and preprocessing

Data were acquired on a customized Siemens 3T “Connectome Skyra” scanner, housed at Washington University in St. Louis, using a standard 32-channel Siemens receive head coil and a “body” transmission coil designed by Siemens specifically (Echo Planar Imaging sequence, Gradient-echo EPI, 1200 volumes, TR = 720 ms, echo time = 33.1 ms, flip angle = 52°, voxel size =  $2 \times 2 \times 2\text{mm}^3$ , field of view =  $208 \times 180\text{mm}^2$ , 72 transversal slices).

Resting-state and task-based fMRI data were collected in two sessions. Each session consisted of two resting-state acquisitions of approximately 15 minutes each, where subjects were instructed to keep their eyes open, followed by task-based fMRI acquisitions of varying durations. For this study, we used the resting-state fMRI data acquired with Left-Right orientation from the first session, and task-based fMRI data with the MOTOR paradigm. In this task, participants were presented with visual cues that asked them to either tap their left or right finger, squeeze their left or right toe, or move their tongue. Each task volume image was obtained from 10 blocks of 12 seconds each, corresponding to 2 tongue movements, 4 hand movements (2 right and 2 left), 4 foot movements (2 right and 2 left), and 3 fixation blocks of 15 seconds. Each block was also preceded by a 3-second cue.

For this study, we used the preprocessed resting-state fMRI data as provided by HCP. In particular, the ICA-FIX pipeline [17] was applied, which cleans the BOLD signal by removing noise components given by the ICA algorithm and that has been proved to increase the quality of the original data [18]. In the case of task data, such a denoised version was not yet available. Therefore, we downloaded the minimally preprocessed data and ICA-FIX processed it through the *hcp\_fix* script that can be found in the HCP repository. This script performs a rigid-body

head motion correction, high-pass temporal filtering, ICA decomposition of the data, FIX identification of the ICA components corresponding to artifacts and elimination of these components from the original data. Furthermore, we used FSL's FNIRT command *applywarp* to transform both paradigm data from MNI space back to subjects' native space through the appropriate matrix transformation that can be found for each subject in the HCP repository and FSL's FLIRT to resample the native structural image at 2mm.

Next, we registered our volume data into the Shen parcellation atlas in the subjects' native space, partitioning each subject's brain in 268 functionally homogeneous and spatially coherent brain regions [19]. Functional connectivity matrices for each subject were then obtained by computing the Pearson correlation between the mean BOLD time course among all pairs of the 268 brain regions. A row (column)  $i$  of a given matrix thus represents the correlation map of the  $i^{\text{th}}$  ROI with all other brain regions. Hence, these correlation maps describe the interactions of a brain node, which define the intrinsic connectivity networks (ICNs) assignment. Specifically, we considered seven cortical ICNs as proposed by Yeo and colleagues [20] (visual (VIS), sensorimotor (SM), dorsal attention (DA), ventral attention (VA), limbic (L), frontoparietal (FP), and default mode network (DMN)), and two more sub-networks for the sake of completeness, comprising subcortical (SUB) and cerebellar regions (CER). The number of regions per network in this atlas is 30, 25, 18, 22, 22, 23, 42, 48 and 38 respectively. ICN assignment to each of the 268 brain regions was performed by overlapping both Shen and Yeo atlas with a minimal 80% threshold. As a consequence, if the number of voxels within a Shen parcel exceeding this threshold belongs to one of Yeo's ICNs, the parcel is assigned to that particular ICN.

Finally, for both task and resting data separately, the 282 resulting individual  $268 \times 268$  Pearson correlation matrices were concatenated together row-wise to obtain  $\mathcal{X}_{\text{task}}$  and  $\mathcal{X}_{\text{rest}}$  super-matrices of  $282 \times 268$  instances and 268 features each. Both these matrices are then later used in the subsequent machine learning analysis. Moreover, since self correlations, *i.e.* values of 1 in the Pearson connectivity matrices, would straightforwardly identify the ICN in a one-hot encoding fashion, we set these features to zero so that we guarantee that interactions between different regions are indeed the responsible for determining the ICN label.

## Classification

Each correlation map was used to identify ICNs by means of four well known classification algorithms in machine learning: Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM) and Random Forest (RF) run under scikit-learn 0.18, an open source Python library that implements a wide range of machine learning tools which include preprocessing, cross-validation and classification algorithms [21]; and a multilayer Neural Network (NN) using Keras in a Tensorflow backend, an API for deep learning written in Python [22].

QDA is a non-linear classifier that stands out for being computationally friendly, inherently multiclass and very simple, with no hyperparameters to tune. In this classifier, the posterior label conditional probability functions  $p(y|x)$  are obtained through a likelihood  $p(x|y)$  taken as multivariate Gaussian distribution but, unlike in Linear Discriminant Analysis, covariances of the classes are not assumed to be equal, which lead to quadratic decision boundaries.

SVM is a maximum-margin classifier, aiming to find a hyperplane that maximizes the distance to the nearest points on each side representing each class (the so-called support vectors). It is specially suited for high dimensional problems, since it can linearly separate any dataset by going to a high dimensional space with the aid of a kernel function.

RF, on the other hand, is an ensemble of decision trees, which split the feature space according to the maximization of the Gini criterion, where each decision tree is trained on a random

subset of examples and each splitting considers only a random subset of features, so that uncorrelated trees are obtained. By doing so, and averaging over all trees, one can reduce the variance and therefore avoid overfitting.

Lastly, multilayer neural networks are able to fit a nonlinear function by passing the examples through its architecture while minimizing a loss function. The basic setup is a first layer with units equal to the number of features, a (few) hidden layer(s) and a final output layer that encodes the label information. The training is performed by updating the links (weights) among layers by back propagation in order to reduce the output error. We have explored several network architectures varying both depth and number of units per layer. Each layer has a *ReLU* activation except for the last one defining the example class which is a *Softmax* function. We used a Stochastic gradient descent optimizer with a learning rate of 0.01, a decay of  $10^{-6}$  and a Nesterov momentum of 0.9. Likewise, in order to prevent overfitting we adopted an early stopping criterion on a 10% of the training dataset and a dropout rate of 0.2, which randomly sets this fraction rate of input units to 0 at each update during training time.

While MLP and RF can inherently map correlation maps into a 9-dimensional space, such an implementation is not as straightforward in SVM which is more suitable for binary classification. In order to address this issue, a common approach is to use a one-versus-all strategy, where one builds as many classifiers as class labels and trains each one taking one label as positive versus the rest being negative cases. Finally, classification of a test example is made by reporting the highest confidence score predicted after applying all classifiers. We also applied such a strategy to RF.

Out-of-sample performance has been estimated on resting fMRI data, acting thus as our unseen **test set**, with the aim of addressing the question of how accurate these predictions are when they are obtained from a classifier fitted on task fMRI, acting here as the **training set**. Before that, we compared and selected the best classifier, for which we used exclusively this training set to compute the global performance of different classifiers in a 5 times repeated 10-fold cross-validation (Setting  $k = 10$  is an optimal choice in terms of the trade-off between variance and bias of the classifier [23]). Each holdout piece of data of each fold within this cross-validation on task fMRI acts therefore as a **validation set**.

Model selection has been carried out in two-steps. We first calculated within the mentioned cross-validation strategy the performance for each algorithm varying their hyperparameters defined on a grid. For the case of neural networks, we trained 20 different architectures corresponding to different number of hidden units and layers. For the case of SVM, we considered a Linear Kernel and Radial Basis Kernel approximation, taking for both seven regularization terms  $C$  within the range  $[10^{-5} - 10^2]$  and three different values of  $\gamma$  between  $10^{-2}$  and 10 for the non-linear kernel case. Both Random Forest cases were searched across the same hyperparameters space, varying the size of trees ensemble and tuning the minimal samples required to split an internal node, which allowed us to control the growth of the trees. QDA does not have any parameter to tune. More details about the hyperparameters can be found in the Supporting Information section.

Second, we compared the algorithms with the hyperparameters giving the best performance, so that it allows us to finally select the classifier with the highest accuracy in its most optimal configuration. We also assessed the significance of the averaged accuracy across the selected algorithms by means of a Friedman test, which is a non-parametric ANOVA version for repeated measures. If this was statistically significant, post-hoc analysis was performed, where the best algorithm was compared to the rest using a Wilcoxon signed-rank test and then Bonferroni corrected for multiple comparisons.

Given the nature of our data, where we have 268 observations for each subject, one could argue that these are correlated and therefore training and test datasets would not be totally

independent if observations from the same subject fall in both subsamples. Such an effect is known as double-dipping and could lead to inflated association scores [24]. We thus performed the 10-fold data partitions **based on subjects** so as to guarantee that observations from the same subjects are exclusively either in the training dataset or the test dataset.

Once the best algorithm has been selected using exclusively task data, we sought to generalize its results when tested on a different fMRI paradigm. As a consequence, we followed the same cross-validation scheme as explained before such that for each iteration we used 90% of the subjects' task fMRI data as the training set to make predictions relying on the 10% of the remaining subjects, but now taking their resting fMRI data as test set. As such, one could infer whether intrinsic connectivity networks from a model fitted exclusively on task data could be recovered even when the subject is not able to perform the task.

Results are reported in several ways: (1) global accuracy (*i.e.* proportion of correctly classified instances); (2) a confusion matrix, where each row represents the instances in a predicted class and each column the examples in an actual class; (3) and Receiver operating characteristic (ROC) and Precision-Recall (PR) curves which exhibit model behaviour affected by different thresholds on the decision function predicted for each class. ROC curves represent the true positive rate (TPR) versus false positive rate (FPR), where in binary classification  $TPR = TP/P = \#(\text{pred}==\text{pos} - \text{pos}) / (\# \text{ pos})$  and  $FPR = FP/N = \#(\text{pred}==\text{pos} - \text{neg}) / (\# \text{ neg})$ , where # indicates number of examples. *Precision* is defined as the ratio of positive predictions that are indeed positive and *Recall* as the proportion of positive instances that are classified so.

## Workflow

The different steps taken for the obtainment and use of the matrix of features within the machine learning algorithms analysis is shown in Fig 1, which can be summarised as follows:

1. Formation of the matrix of feature vectors for both task and resting fMRI data.
2. 5 times repeated 10-fold cross-validation on **task** data to select the best predictive model. The data partition is based on subjects so that observations from the same subject do not fall in both training and test data.
3. Results generalisation following the same cross-validation technique explained before but now with the test samples containing only **resting** data whereas the model is still trained on the task data.

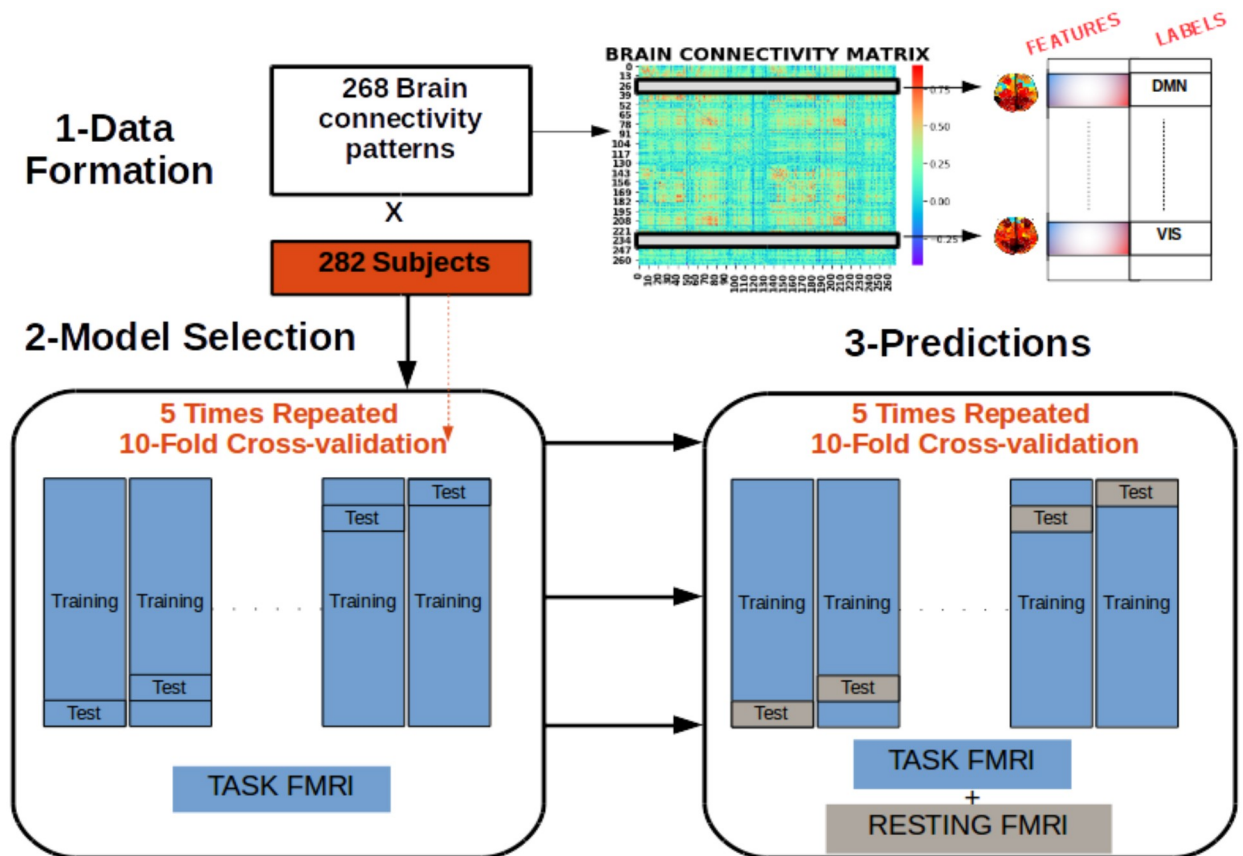
## Computing resources

All the calculations have been performed using the High-performance Computing Cluster of BioCruces Bizkaia. For computing, such a platform consists of 4 Servers Dell PowerEdge R430, each 96 GB RAM DDR4, 2 CPU Intel E5-2660 v3 (10 cores c/u), 2 HDD 1TB hot swap in RAID1, OS CentOS 7.1 and workload manager SLURM; 120 TB are reserved for storage.

## Results

### Performance on task fMRI

Global averaged accuracy after cross-validating the different methods for the most optimal hyperparameter scenario can be seen in Table 1. The performance for the rest of hyperparameters tried across the different folds for the different algorithms are depicted in S1, S2, S3 and S4 Figs, which exhibit the stability of the results w.r.t different shufflings of the data.



**Fig 1. Workflow.** The connectivity matrices for each subject were manipulated to obtain a matrix of features for both task and resting data separately and both these matrices were used subsequently to fit different machine learning algorithm and make predictions.

<https://doi.org/10.1371/journal.pone.0207385.g001>

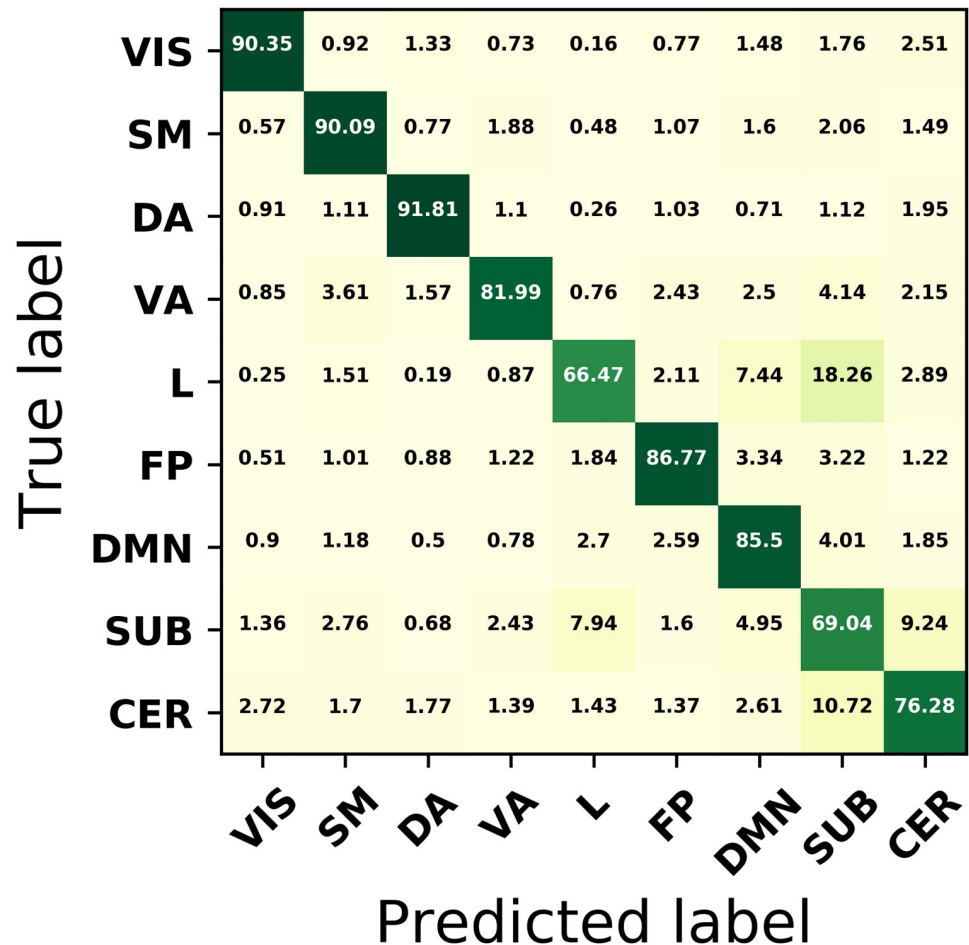
First, statistically significant differences across the six classifiers in their most optimal configuration were found ( $p = 7.25 \cdot 10^{-52}$ ). Second, the neural network classifier by far outperformed both Support Vector Machine ( $p = 3.77 \cdot 10^{-9}$  for both kernels) and Random Forest algorithms ( $p = 3.77 \cdot 10^{-9}$  in the multiclass case and  $p = 3.76 \cdot 10^{-9}$  in the OVR scheme), popular choices in literature when dealing with machine learning problems. In addition, neural networks also improved the decent results of about 77% provided by Quadratic Discriminant Analysis classifier ( $p = 4.02 \cdot 10^{-9}$ ), as a consequence of the added complexity. Moreover, within the set of neural network models, there seems to be a tendency of better performance as the neural network becomes deeper, as well as with increasing number of hidden units, which

**Table 1. Classification accuracy scores.** Basic statistics of the algorithms' performance across the  $5 \times 10$  folds for the most optimal hyperparameter configuration in parenthesis.

Classifier	Mean (%)	Var (%)	Min (%)	Max (%)
NN (512, 256, 128, 64)	80.90	0.01	78.24	83.10
RF MULTI (750 trees, 10 min_samples)	67.21	0.01	65.00	69.32
RF OVR (750 trees, 2 min_samples)	71.62	0.01	69.04	73.63
SVM linear (C = 0.01)	57.74	0.01	55.62	59.80
SVM RBF (C = 1, $\gamma = 0.01$ )	54.23	0.01	51.53	56.58
QDA	76.89	0.02	74.63	79.37

<https://doi.org/10.1371/journal.pone.0207385.t001>

# Task-Training Task-Test (%)



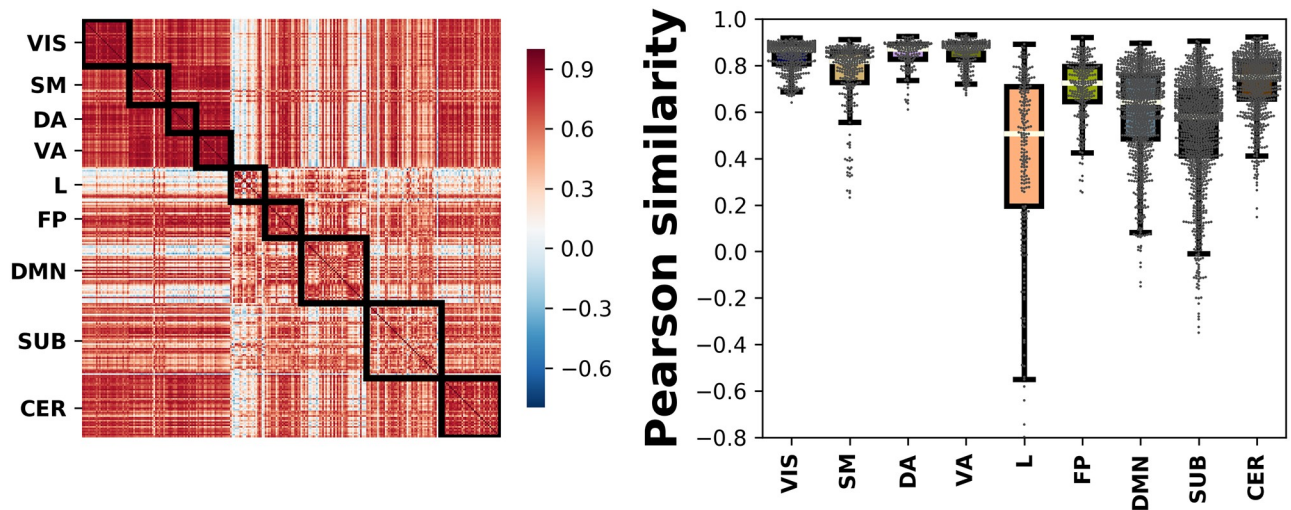
**Fig 2. Confusion matrix for the best model in task fMRI.** Confusion matrix using exclusively task fMRI data for the selection of the best classification model, which in this case corresponds to a NN with 4 hidden layer with 512, 256, 128 and 64 units.

<https://doi.org/10.1371/journal.pone.0207385.g002>

exhibits the demand for a higher number of parameters to define the decision boundaries separating each class. Notably, a decreasing complexity in architecture seems to capture better the intrinsic structure of the data, with a neural network with four hidden layers of 512, 256, 128 and 64 units, respectively, providing the best performance with a global accuracy of ~ 81%. In the following, we will concentrate on the results given by this model.

Since we are dealing with a multi-class problem, it is important to calculate the classification performance of each ICN. This can be accomplished by means of the confusion matrix, that is depicted in Fig 2 for our best model. As one can see, all ICNs exhibit a good performance, specially for those networks whose activation is usually demanded during our task performance such as motor, visual and dorsal attention areas, achieving all rates above 90%. The limbic system however stands out of this behaviour, since its accuracy drops to approximately 66%, with a remarkable 18% of examples misclassified as subcortical regions.





**Fig 3. Pearson similarity between brain regions.** Representation of correlations amongst the 268 pattern connectivities from the subjects-averaged connectivity matrix. On the left, the full correlation matrix, ordered according to their group label, whose members' interactions are outlined by a bold rectangle. On the right, the intra-group correlation distribution corresponding to the upper off-diagonal entries in each bold rectangle.

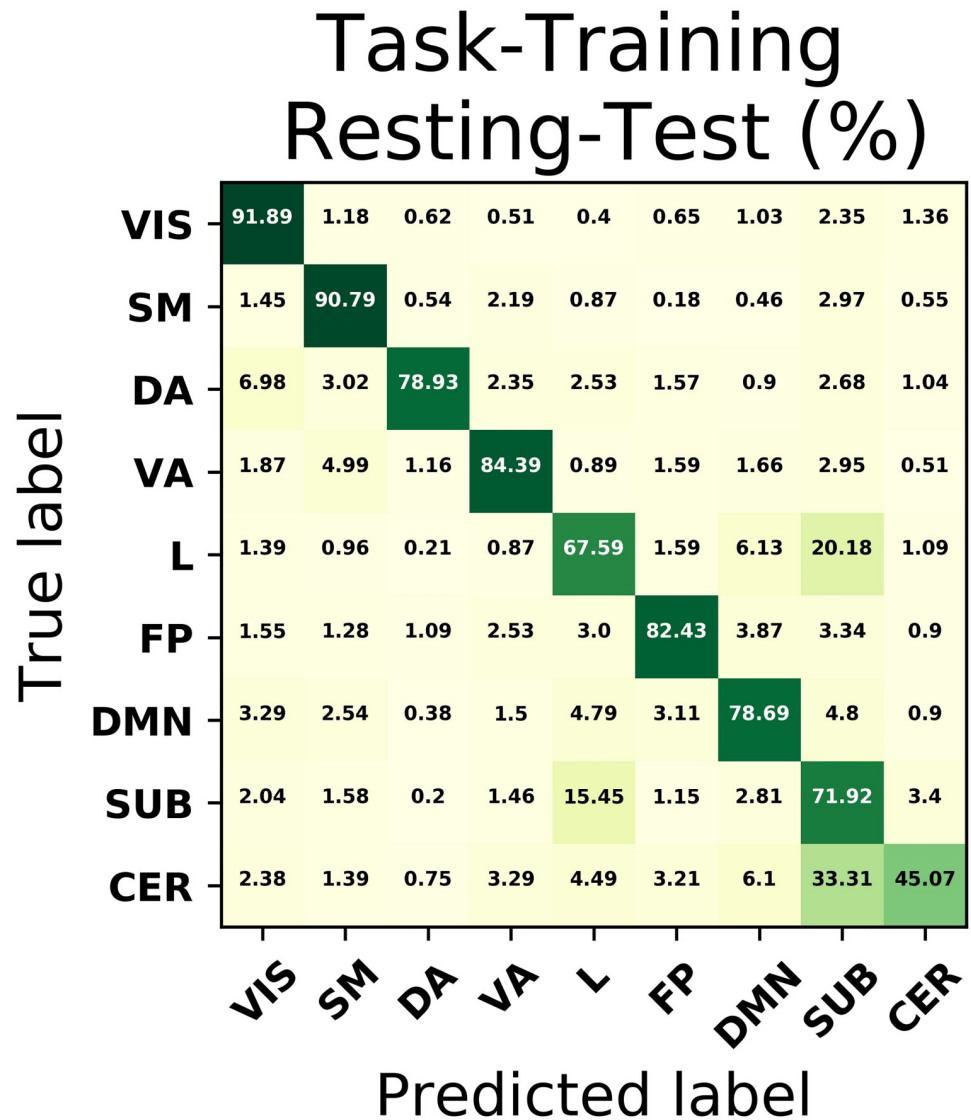
<https://doi.org/10.1371/journal.pone.0207385.g003>

This difference in performance by each ICN might be addressed by looking at the similarity among the ICNs in more detail. In Fig 3, we show the Pearson cross-correlation matrix between pattern connectivities of the 268 nodes from the subjects-averaged connectivity matrix on the left, whereas on the right we depict the dispersion of the off-diagonal terms within each class (the bold boxes on the correlation matrix) in order to account for the intra-group similarity specifically. As it can be noticed, examples from the limbic and subcortical networks are more distant from each other, leading to an increase of similarity variance, which explains why the classifier struggles to build a decision function that distinguishes them efficiently. Likewise, the between-networks terms in the correlation matrix (those out of the black boxes) reflect the vicinity of subcortical regions to other networks and it also shows that, while having higher intra-group similarity, examples belonging to the cerebellar network are more misclassified than those of DMN and FP, as a consequence of a smaller inter-group distance with other networks, especially with subcortical regions.

### Generalisation to resting-state fMRI data

Next, we investigated how well the best neural network model generalised the results found so far when testing is performed on resting-state fMRI data from a model which was **only** trained using task-based data. In other words, we examined whether intrinsic connectivity networks can be successfully predicted in experiments where no active collaboration of the subjects is required.

Fig 4 shows the mean confusion matrix in a 5 times repeated 10-fold cross-validation based on subjects where the training has been performed after using the task-based fMRI dataset and testing on resting-state fMRI data. Results demonstrate a similar performance pattern compared to that of the previous section, with high prediction rates in both VIS and SM network, of approximately 92% and 91% of sensitivity (also known as recall) respectively. Similarly, quick visual identification within the studied task to conditionally perform a specific movement encodes a bottom-up process which demands activation of ventral attention areas. Such areas are also well predicted by our fitted model with a 84% sensitivity. In contrast, dorsal attention areas attached to top-down stimuli drastically decrease their performance to 79%,

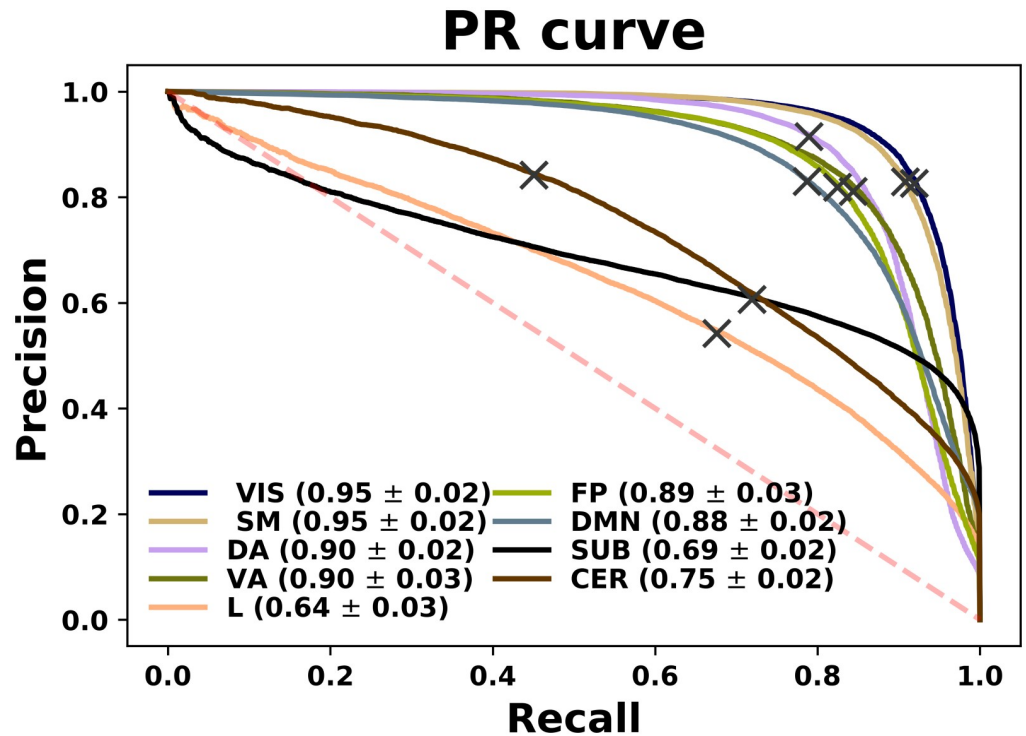


**Fig 4. Confusion matrix using resting data as test.** Confusion matrix for the best classification model, which corresponded to a neural network with four hidden layer of 512, 256, 128 and 64 units, but now using task-based fMRI data as training set and resting-state fMRI data as testing set.

<https://doi.org/10.1371/journal.pone.0207385.g004>

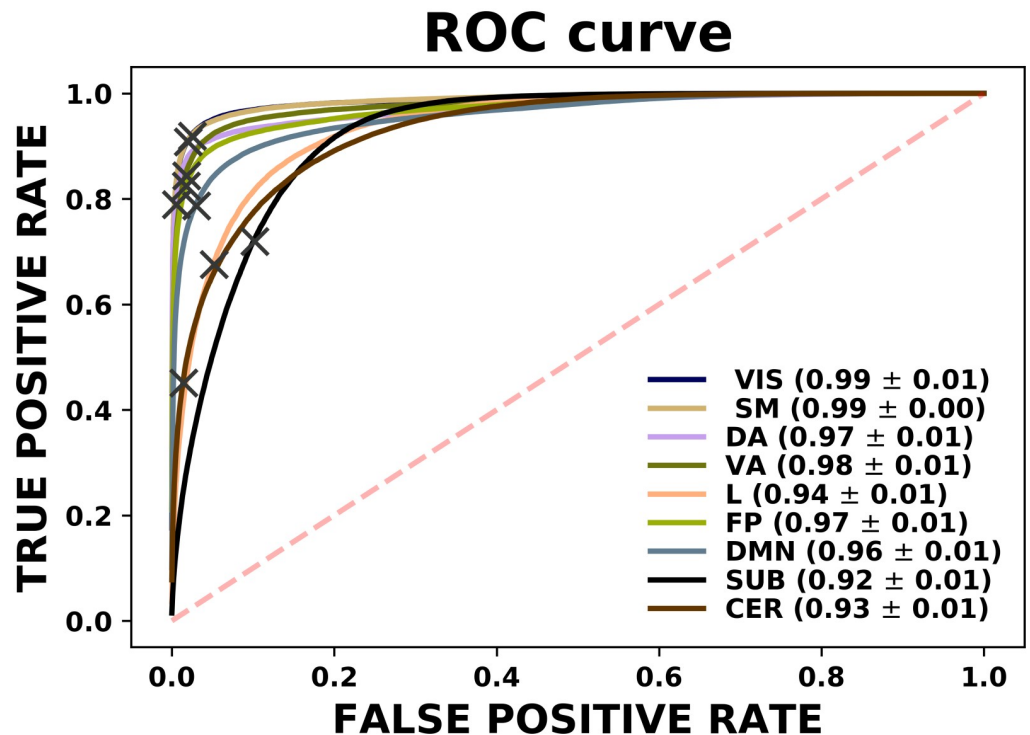
with a moderate proportion being misclassified into the visual network, possibly reflecting some characteristic of the task paradigm in study. On the other hand, DMN regions whose activity negatively correlates with that of the other networks during rest also exhibit a decent performance of about 79%, which suggests that these areas maintain an intrinsic correlation that allows them to be optimally fitted by a model even during task. Finally, limbic and cerebellum systems involved in learning, memory and behaviour suffer from variability and hence are poorly replicated by our model.

The results obtained so far predict one *and only one* class for each instance, corresponding to the class assigned with the highest probability. Instead, we can directly look into this class probability prediction for each example and see how the model performs when changing the threshold of class assignment. We show this in Figs 5 and 6 through the ROC curves and the Precision-Recall curves. We can see that both representations describe the performance of the



**Fig 5. ROC curves for each class separately.** The areas under these curves can be found in the legend located on the right side. Grey crosses display the model with the specific threshold yielding the results shown before. Those curves above the red dashed line represent exhibit a discriminating power.

<https://doi.org/10.1371/journal.pone.0207385.g005>

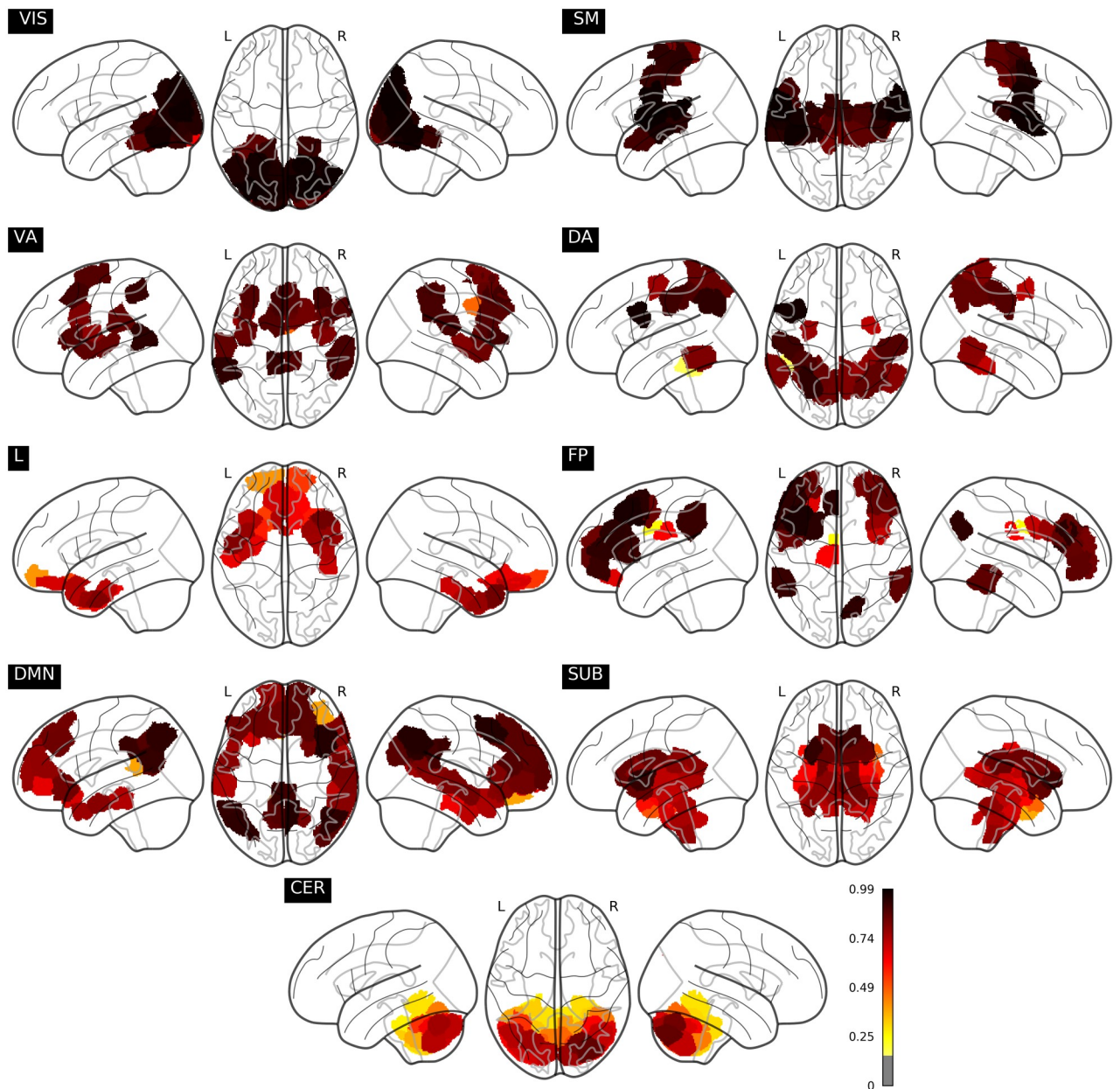


**Fig 6. PR curves for each class separately.** The areas under these curves can be found in the legend located on the right side. Grey crosses display the model with the specific threshold yielding the results shown before. Those curves above the red dashed line exhibit discriminating power.

<https://doi.org/10.1371/journal.pone.0207385.g006>

model when generalised to different settings, with most of the networks occupying large areas in both spaces and with a clear prominence from areas belonging to the SM and VIS networks. In addition, had we solely relied on results provided by the confusion matrix of Fig 4, which represents a particular point of these curves, we would not have appreciated for example that the poorest performance of the cerebellar network depended on a suboptimal threshold and could be improved beyond the performance of the limbic and subcortical system.

Finally, it is reasonable to assume that performance from correlation patterns of different regions depend on the location of these latter in the brain. In order to address this issue, we show in Fig 7 the mean classification rate of each region across the different folds of the cross-



**Fig 7. Classification accuracy of each node.** The best model is used to average each node's accuracy performance across all subjects contained in the resting fMRI test set in each fold of the cross-validation scheme.

<https://doi.org/10.1371/journal.pone.0207385.g007>

validation procedure. This figure shows that, for instance, regions in the cerebellum display rather polarised results. Areas strictly in the right hemisphere within the posterior of the crus cerebellum I/II and lobes VI, VIIb and VIII have a successful accuracy greater than 80%, whereas anterior areas, lobes in the left hemisphere along with the vermis exhibit poorer performance than chance, being mainly mislabelled as subcortical regions. The cerebellum takes part in motor and attention tasks, where there is for example a clear asymmetry in the foot, hand, and tongue movement activation maps for intrinsic functional connectivity (left hemisphere) and task functional connectivity (right hemisphere) [25]. On the other hand, being mainly involved in emotional and long-term cognitive functions, the limbic network is *a priori* the most detached system in relation with the task under exam here. Nonetheless, limbic regions located in the inferior and middle temporal gyrus and pole have decent classifying power, specially in the left hemisphere, whereas accuracy in orbital parts of the left frontal gyrus and rectus drastically drops below 50%. In addition, default mode network and frontoparietal regions, which both yield a decent performance, exhibit interesting features. In general, they both have regions with prediction accuracies well above 90%, with higher incidence on the middle and posterior cingulum in both hemispheres and the left precentral and middle frontal gyrus, and poorly classifying parts depicted in yellow that correspond to the frontal gyrus orbital part, the olfactory cortex, rectus, hippocampus, calcarine and precuneus in the left hemisphere for default mode networks nodes; and the anterior and middle cingulum for both networks.

## Discussion

We have fitted different models on the whole-brain functional connectivity patterns obtained from the Pearson correlation matrices from a large population of healthy subjects, and quantified how well they could predict Intrinsic Connectivity Networks. We obtained the highest accuracy with a neural network of four hidden layers and 512, 256, 128 and 64 intermediate units each. Given the models trained, this result suggests that the complexity of the target function plays an important role with a demand of going deep in architecture, but also that mapping to decreasing numbers of intermediate units seems to better capture the inherent properties of the data, in accordance with the behavior of other well known machine learning algorithms such as autoencoders and convolutional neural networks. This behaviour could be even more evident if we considered a representation at the voxel level due to the increase of dimensionality, and therefore more complex models would likely be required. Moreover, well-established algorithms such as Random Forest (RF) and Support Vector Machines (SVM) were not capable of reaching a performance similar to the one obtained by neural networks. Two hypotheses might be formulated to explain this phenomenon. First, RF splits the space taking a subset of features, so this model fails to capture the optimal splitting out of the correlation maps since one would expect the whole pattern to be important. Second, both SVM and RF in general find it more challenging to build the decision boundary in high dimensional multi-class settings. On the other hand, the benefits of neural networks, that are well known for their capability to model accurately any arbitrary function [26], allow to improve the decent performance provided by the other attempted classifier QDA, suggesting that observations can be to some extent approximated by a multivariate Gaussian function.

Resting state can be regarded as an intrinsic alternate resonance between different brain areas connected at large scale, in contrast to the more isolated activation of the corresponding brain areas during task. Even though this difference in paradigm, the underlying interactions at the neural level are the same, so one would expect both to match to some degree. We have adopted a machine learning approach to prove this. Following the findings of a previous study

[27], it is remarkable how accurately intrinsic connectivity networks can be identified using only task data, which reflects the high degree of correspondence between both modalities. Not only could visual and sensorimotor regions be clearly represented by both paradigms, as expected, but also other functional networks turned out to have correlation maps which lie within well defined decision boundaries. For example, time series of default mode network regions, though having lower participation during cognitive tasks, when cross-correlated with the average time series of all other brain regions, demonstrated that their integration and participation make them differentiable (in the end, Pearson correlation is unaltered by a change of scale). Owing to their well known strong internal connectivity, one expects characteristics of such networks to be susceptible to hold across different paradigms. In contrast, the limbic system showed the worst performance, with limbic regions often miscategorised as belonging to the subcortical network, reflecting the lack of organization due to its weak internal connectivity. This was caused by some areas of both networks showing a similar pattern of activity, forcing the classifier to lean towards the subcortical network, the majority class over the limbic group.

The connection of task-evoked experiments with resting-state becomes even clearer if one looks at the results presented in Figs 4, 5 and 6, where VIS and SM networks in particular (which are mainly involved in the MOTOR task), are recovered exclusively from resting data with great accuracy. This apparent correspondence between both modalities is being thoroughly investigated. In a seminal work, Smith *et al.* [27] demonstrated the existence of a high similarity between the ICA-analysis results extracted independently from a resting-state fMRI dataset consisting of 36 healthy subjects and activation maps from BrainMap database of functional imaging studies across nearly 30,000 human subjects. This suggests that the brain at rest is continuously active so as to be a composition of all the inherent possible tasks, such that a model trained based on a specific task will emerge naturally as one of its components when tested on resting state. On the other hand, in [28] an exhaustive study of this was carried out showing that the same networks that mostly discriminate individuals were also most predictive of cognitive behaviour. In [29] prediction of activation maps by resting-state fMRI were overlapped with maps used to fit a wide range of task-based models and demonstrated that individual differences in brain response are inherently linked to the brain itself rather than to a specific manifestation given a certain task. Nonetheless, our work rather embraces the spirit of [13] in terms of the methodology used. We have aimed at successfully building the decision surfaces that separate the different ICNs by performing an exhaustive machine learning algorithm search in a large cohort of subjects. It is also worthwhile to stress that individual task-rest correspondence and between-subject reproducibility of patterns are different and possibly orthogonal problems, and that the cognitive relevance of these networks is mainly expressed on the subject-specific level [30].

Regarding the results obtained, one might be tempted to take the outcome of the ROC curves in Fig 5 as outstanding, where all the classes exhibit a ROC area greater than 0.9. However, as noted in [31], plotting ROC curves separately for each label might not be appropriate in multi-class settings where negative examples far exceed the number of positive instances. In this case  $FP \ll TN$ , so changes in the number of false positives hardly affect the False Positive Rate, which continues to hold small and therefore the area obtained is usually large. As a consequence, it is better to use Precision-Recall curves, which concentrate only on the positive class and therefore do not suffer from these issues. As it can be seen in Fig 6, the areas found go more in concordance with the findings discussed previously, where the VIS and SM networks seem to be identified most accurately. In addition, we can see two interesting features from carefully inspecting these curves. Firstly, having both VIS and SM systems the largest areas, their observed thresholds (depicted in the figures as grey crosses) lie on a slightly non-

optimal point on their respective curve. If we look at their performance in the diagonal of the confusion matrix of Fig 4, which is nothing less than the Recall measure, we therefore notice that this is accompanied by a loss of precision. Secondly, even though CER network seemed to behave extremely poor when testing the model on resting fMRI, we can see that this is an effect of the threshold used. In fact, when we vary this quantity, we can see that this group performs much better than the limbic and subcortical systems. Finally, it is also remarkable how well areas in DMN can be predicted from a model that was trained exclusively on task data, during which its activation would not be expected. One possible explanation might be that the functional connectivity within this network has a precise mapping with its structural connectivity [32] and therefore, the integration in DMN might preserve across different tasks.

## Limitations

The interpretation of the results of the present study should be seen in the light of the following limitations. These are mainly related to the construction of the matrix of features and the definition of the labels to be classified.

First, the use of the correlations pattern of a given node to the whole brain as feature that determines the ICN of that node should be treated carefully. In particular, setting all the terms but the self-correlation in the feature vector to zero, the resulting vector would directly provide the ICN label of the node with a 100% of accuracy since each node vector would be unique and orthogonal to the rest and therefore classification would not be needed whatsoever. However, our hypothesis is that ICN assignment should be predicted by the whole pattern connectivity of the node. As a consequence, and given that we are limited by the fixed dimensions of the feature vectors (these can not change across subjects), we set self-correlations values to zero so as to diminish their effect and allow the rest terms to truly determine the label of each example.

Second, the labels used to fit the model are not fully and uniquely defined since they are threshold dependent when matching Shen with Yeo's atlas. As a result, this spatial variability might somewhat blur the connectivity maps and therefore reduce the performance.

Third, for this study we have not considered performing global signal regression (GSR) when preprocessing the data. Such a step is still controversial among the neuroimaging community given the increase of negative correlations and exhibits the fact that there is not a single "right" way to process resting state data [33].

Finally, it is worthwhile mentioning a subtle technical limitation regarding hypothesis testing adopted when comparing between the different algorithms. It is well known that the usual 10-fold cross-validation strategy violates the independence assumption of a paired t-test and as a consequence, some authors have suggested adopting a 5x2 CV to alleviate this problem [34]. However, such a pessimistic biased strategy lacks power and replicability, which has led other authors to propose to use multiple runs of 10-fold cross-validation where their results amongst pairs of classifiers are testing using a paired t-test with fewer degrees of freedom, but this application is limited to binary data [35]. In our case, non-parametric testing along with multiple runs of 10-fold cross-validation tries to approximate such a situation.

## Conclusion

Successfully delineating intrinsic connectivity networks of the brain is of key importance to fully understand its behaviour and the patterns emerging during evoked activity (and vice-versa). Multivariate methods in machine learning turn out to be promising techniques in successfully identifying ICN networks from both task and resting-state fMRI data. As usual, the selection of the optimal algorithm is a crucial step. Obviously the optimal algorithm depends on the complexity of the problem. In our case, we have seen that a simple neural network with

just one layer yields the best results. Whether a higher resolution or different parcellation might require to go deeper calls for future investigation.

As mentioned, the accurate prediction of the different baseline functions of the brain even when the model has been fitted using a different protocol (in our case a visual-tapping task) might be relevant for important future applications in clinical neuroscience. Localization of functions that are questionable due to the incapability of subjects to perform specific tasks could easily arise in this proposed scenario and could therefore help clinicians isolate the areas demanding a correct treatment for subject recovery. Moreover, future studies using the strategy followed in this work can address how brain regions in patients who have undergone neurosurgery change their connectivity pattern and therefore specialise in performing new tasks.

## Supporting information

**S1 Fig. Neural networks behaviour with different hyperparameters.** Global accuracy of a neural network with different layers (the number of components of each vector on the y-axis) and hidden units (the number within each component of the vectors on the y-axis) using a 5 times repeated 10-Fold Cross-validation exclusively on task fMRI.

(TIFF)

**S2 Fig. Random Forest behaviour with different hyperparameters.** Global accuracy of a Random Forest, both in its multi-class and one-versus-rest version, varying the number of trees in the ensemble (“trees” parameter on the y-axis) and the minimum of sample in order to stop dividing each internal node (“min\_samp” parameter on the y-axis) using a 5 times repeated 10-Fold Cross-validation exclusively on task fMRI.

(TIFF)

**S3 Fig. Support Vector Machine behaviour with different hyperparameters.** Global accuracy of a Support Vector Machine with a linear kernel varying the regularization coefficient (“C” parameter on the y-axis) using a 5 times repeated 10-Fold Cross-validation exclusively on task fMRI.

(TIFF)

**S4 Fig. Support Vector Machine behaviour with different hyperparameters.** Global accuracy of a Support Vector Machine with a Radial Basis Kernel approximation varying the regularization coefficient (“C” parameter on the y-axis) and gamma kernel coefficient ( $\gamma$  on the y-axis) using a 5 Times repeated 10-Fold Cross-validation exclusively on task fMRI.

(TIFF)

## Acknowledgments

Data collection and sharing for this project was provided by the Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California.

## Author Contributions

**Conceptualization:** Javier Rasero, Hannelore Aerts, Marlis Ontivero Ortega, Jesus M. Cortes, Sebastiano Stramaglia, Daniele Marinazzo.



**Data curation:** Javier Rasero.

**Methodology:** Hannelore Aerts, Marlis Ontivero Ortega, Sebastiano Stramaglia, Daniele Marinazzo.

**Software:** Javier Rasero.

**Supervision:** Sebastiano Stramaglia, Daniele Marinazzo.

**Validation:** Javier Rasero.

**Visualization:** Javier Rasero.

**Writing – original draft:** Javier Rasero, Hannelore Aerts, Marlis Ontivero Ortega, Jesus M. Cortes, Sebastiano Stramaglia, Daniele Marinazzo.

## References

1. Biswal B, Yetkin Z, F H, M V, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*. 1995; 34(4):537–541. <https://doi.org/10.1002/mrm.1910340409>
2. Fox MD, Corbetta M, Snyder AZ, Vincent JL, Raichle ME. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(26):10046–10051. <https://doi.org/10.1073/pnas.0604187103> PMID: 16788060
3. Vincent JL, Kahn I, Snyder AZ, Raichle ME, Buckner RL. Evidence for a Frontoparietal Control System Revealed by Intrinsic Functional Connectivity. *Journal of Neurophysiology*. 2008; 100(6):3328–3342. <https://doi.org/10.1152/jn.90355.2008> PMID: 18799601
4. Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*. 2001; 98(2):676–682. <https://doi.org/10.1073/pnas.98.2.676> PMID: 11209064
5. Gusnard DA, Raichle ME. Searching for a baseline: Functional imaging and the resting human brain. *Nat Rev Neurosci*. 2001; 2(10):685–694. <https://doi.org/10.1038/35094500> PMID: 11584306
6. Beckmann CF, DeLuca M, Devlin JT, Smith SM. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2005; 360(1457):1001–1013. <https://doi.org/10.1098/rstb.2005.1634>
7. Damoiseaux JS, Rombouts SAR, Barkhof F, Scheltens P, Stam CJ, Smith SM, et al. Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(37):13848–13853. <https://doi.org/10.1073/pnas.0601417103> PMID: 16945915
8. Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K. Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magnetic Resonance Imaging*. 2002; 20(4):305–317. [https://doi.org/10.1016/S0730-725X\(02\)00503-9](https://doi.org/10.1016/S0730-725X(02)00503-9) PMID: 12165349
9. Lee MH, Hacker CD, Snyder AZ, Corbetta M, Zhang D, Leuthardt EC, et al. Clustering of Resting State Networks. *PLoS ONE*. 2012; 7:1–12.
10. Bellec P, Rosa-Neto P, Lyttelton OC, Benali H, Evans AC. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage*. 2010; 51(3):1126–39. <https://doi.org/10.1016/j.neuroimage.2010.02.082> PMID: 20226257
11. van den Heuvel M, Mandl R, Hulshoff Pol H. Normalized Cut Group Clustering of Resting-State fMRI Data. *PLOS ONE*. 2008; 3(4):1–11. <https://doi.org/10.1371/journal.pone.0002001>
12. Hacker CD, Laumann TO, Szrama NP, Baldassarre A, Snyder AZ, Leuthardt EC, et al. Resting State Network Estimation in Individual Subjects. *Neuroimage*. 2013; 15:616–633. <https://doi.org/10.1016/j.neuroimage.2013.05.108> PMID: 23735260
13. Mitchell TJ, Hacker CD, Breshears JD, Szrama NP, Sharma M, Bundy DT, et al. A Novel Data-Driven Approach to Preoperative Mapping of Functional Cortex Using Resting-State Functional Magnetic Resonance Imaging. *Neurosurgery*. 2013; 73(6):969–983. <https://doi.org/10.1227/NEU.000000000000141> PMID: 24264234
14. Lee MH, Miller-Thomas MM, Benziger TL, Marcus DS, Hacker CD, Leuthardt EC, et al. Clinical Resting-state fMRI in the Preoperative Setting. *Topics in Magnetic Resonance Imaging*. 2016; 25:11–17. <https://doi.org/10.1097/RMR.0000000000000075> PMID: 26848556

15. Jones P, O V, L N, Adcock JE, Stacey R, Jbabdi S. Resting connectivity predicts task activation in pre-surgical populations. *NeuroImage: Clinical*. 2017; 13:378–385. <https://doi.org/10.1016/j.nicl.2016.12.028>
16. Essen DCV, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. *NeuroImage*. 2013; 80:62–79. <https://doi.org/10.1016/j.neuroimage.2013.05.041>. PMID: 23684880
17. Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL ftWMHC. The Minimal Preprocessing Pipelines for the Human Connectome Project. *Neuroimage*. 2013; 80:105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127> PMID: 23668970
18. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*. 2014; 90:449–468. <https://doi.org/10.1016/j.neuroimage.2013.11.046>. PMID: 24389422
19. Shen X, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage*. 2013; p. 403–415. <https://doi.org/10.1016/j.neuroimage.2013.05.081> PMID: 23747961
20. Yeo T, T B, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*. 2011; 106(3):1125–1165. <https://doi.org/10.1152/jn.00338.2011> PMID: 21653723
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011; 12:2825–2830.
22. Chollet F, et al. Keras: The Python Deep Learning library. 2015;.
23. Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2. IJCAI'95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143. Available from: <http://dl.acm.org/citation.cfm?id=1643031.1643047>.
24. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience—the dangers of double dipping. *Nat Neurosci*. 2009; 12(5):535–540. <https://doi.org/10.1038/nn.2303> PMID: 19396166
25. Buckner RL. The Cerebellum and Cognitive Function: 25 Years of Insight from Anatomy and Neuroimaging. *Neuron*. 2013; 80(3):807–815. <https://doi.org/10.1016/j.neuron.2013.10.044> PMID: 24183029
26. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? *Journal of Statistical Physics*. 2017; 168(6):1223–1247.
27. Smith A, M S, Fox PT, Miller KL, Glahn DC, Fox PM, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106(31):13040–13045. <https://doi.org/10.1073/pnas.0905267106> PMID: 19620724
28. Finn ES, Shen X, Scheinost D, et al. Functional connectome fingerprinting: Identifying individuals based on patterns of brain connectivity. *Nature neuroscience*. 2015; 18(11):1664–1671. <https://doi.org/10.1038/nn.4135> PMID: 26457551
29. Tavor I, Jones OP, Mars RB, Smith SM, Behrens TE, Jbabdi S. Task-free MRI predicts individual differences in brain activity during task performance. *Science*. 2016; 352:216–220. <https://doi.org/10.1126/science.aad8127> PMID: 27124457
30. Kong R, Li J, Sun N, Sabuncu MR, Liu H, Schaefer A, et al. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality and Emotion. *Cerebral Cortex*; 2018. bhy123. <https://doi.org/10.1093/cercor/bhy123>
31. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In: *Proceedings of the 23rd International Conference on Machine Learning*; 2006. p. 233–240.
32. Greicius MD, Supekar K, Menon V, Dougherty RF. Resting-State Functional Connectivity Reflects Structural Connectivity in the Default Mode Network. *Cerebral Cortex (New York, NY)*. 2009; 19(1):72–78.
33. Murphy K, Fox MD. Towards a consensus regarding global signal regression for resting state functional connectivity MRI, *NeuroImage*; 2017. 154: 169–173 <https://doi.org/10.1016/j.neuroimage.2016.11.052> PMID: 27888059
34. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms *Neural Computation*; 1998. 10:7, 1895–1923 <https://doi.org/10.1162/089976698300017197> PMID: 9744903
35. Bouckaert RR. Choosing Between Two Learning Algorithms Based on Calibrated Tests *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*; 2003. 8:51–58