Towards stability of dynamic FC estimates in neuroimaging and electrophysiology: solutions and limits

Sonsoles Alonso and Diego Vidaurre*

Center for Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, 8000, Denmark

* Corresponding author
Department of Clinical Medicine, Center of Functionally Integrative Neuroscience, Aarhus University, Universitetsbyen 3, 8000 Aarhus, Denmark. dvidaurre@cfin.au.dk

Abstract

Being able to map the spatio-temporal organization of the brain's activity is an important step to further our understanding of the basis of human cognition. This has fuelled the interest in time-varying functional connectivity (FC) methods, which aim to describe the evolution of the statistical coupling between regions across the whole brain. Several methods have been proposed in neuroimaging and electrophysiology to characterize time-varying FC. A frequently unspoken issue of these methods is that their estimations are often unstable across runs of the inference even for the same data; that is, different runs can produce different results. But to draw meaningful relations to behaviour, estimates must be robust and reproducible. Focusing on a generative model of time-varying FC, the Hidden Markov Model (HMM), we proposed two approaches to this problem. First, we consider running the inference multiple times, ranking the runs according to a quantitative measure that weighs the fitness of the data and the simplicity of the model (here, the free energy), and choosing the model with the best score. Second, we introduce a novel method, termed principal components of connectivity analysis (PCCA), that explicitly leverages the variability of the HMM inference by refactoring the different estimates into stable, latent patterns of time-varying FC. We discuss and compare these methods on two separate fMRI and MEG datasets, showing the extent to which, they improve the stability of standard time-varying FC estimations.

Keywords: estimation noise; hidden Markov model; principal component of connectivity analysis; time-varying FC; replicability; reproducibility

1. Introduction

An important aspect of the brain's functional architecture is how the different regions are brought together into functional networks, and how these are dynamically organized at many spatial and temporal scales (Laughlin and Sejnowski, 2003). One of the most widely used metrics to map these functional relationships is functional connectivity (FC), a measure of the statistical dependencies between pairs of brain regions (Friston, 1994). More recently, the exploration of the temporal properties of these interdependencies has revealed meaningful within-session fluctuations in FC, both from functional magnetic resonance imaging (fMRI; Fornito and Bullmore, 2010; Karapanagiotidis et al., 2020; Liégeois et al., 2019; Lurie et al., 2020; Vidaurre et al., 2021, 2018, 2017; Xie et al., 2018).

Unfortunately, partly due to the variety of analytical tools for estimating time-varying FC and their inherent limitations (Dafflon et al., 2022), it is not always easy to compare results across studies. The issue we focus on here is estimation noise. Any method whose inference depends on an optimization procedure, such as the Hidden Markov Model (HMM), is subject to variability across runs of the inference even if we use the same data. That is, estimations can be unstable depending on factors such as the amount of data and the complexity of the model (Vidaurre et al., 2019). This is the case also for simpler approaches, such as Independent Component Analysis (ICA; Beckmann

and Smith, 2004). But to draw meaningful relations to behaviour, we need estimates that are robust and reproducible.

The goal of this study is to introduce two approaches to achieve stability in time-varying FC estimations obtained from the HMM. The first approach, termed best-ranked HMM (BR-HMM) consists of running the model several times and taking the run that best scores with respect to a given quantitative measure of model performance. In the case of the HMM, the natural choice is the free energy, which represents a trade-off between fitness and model complexity based on Bayesian principles. Here, we show that, depending on the characteristics of the data, a large number of runs might be needed to accurately find a robust, reproducible model. To address this issue, we introduce a novel, second approach: principal components of connectivity analysis (PCCA), which necessitates fewer runs to provide reliable patterns of time-varying FC. Briefly, PCCA consists of applying PCA on the state time courses from multiple HMM runs in order to extract the latent temporal structure across HMM runs. The component-specific FC matrices are then computed as linear combinations of the HMM states' FC using principles of Riemannian geometry. We show that the choice of one or the other approach will be determined by the objectives of the study, the available computational resources, and the complexity of the model configuration. We demonstrate the performance of the methods on two separate fMRI and MEG datasets.

2. Materials and Methods

2.1. Data description

2.1.1. Resting-state fMRI HCP data

We used publicly available fMRI data from 100 subjects from the HCP (Van Essen et al., 2013). For each participant, we considered data from 4 resting-state sessions. Please refer to (Van Essen et al., 2012) for full details about the acquisition and preprocessing of the data. In brief, 3T whole-brain fMRI data were acquired with a spatial resolution of 2×2×2 mm and a temporal resolution of 0.72 s. All fMRI processing was performed using FSL (Jenkinson et al., 2012) including minimal high-pass temporal filtering (>2000s FWHM) to remove the linear trends of the data, and artefact removal using ICA+FIX (Griffanti et al., 2014). No low-pass temporal filtering or global signal regression was applied. Group spatial ICA was performed using MELODIC (Beckmann and Smith, 2004) to obtain a parcellation of 25 ICs. These timeseries were directly obtained from the HCP 'PTN' Parcellation+Timeseries+Netmats (specifically, the first 100 subjects from the 'recon2' version). Whereas a higher number of IC parcellations are provided by the HCP, the 25-ICA parcellation was sufficient to map the dynamics of FC as it covers the major functional networks, while keeping the computational cost relatively low.

2.1.2. Resting-state MEG UK-MEG data

The MEG dataset consisted of 5-min resting-state recordings from 10 subjects obtained from the UK MEG Partnership (Hunt et al., 2016) which recruited 77 healthy participants at the University of Nottingham. Resting state MEG data were acquired using a 275-channel CTF MEG system (MISL, Coquitlam, Canada) operating in third-order synthetic gradiometry configuration, at a sampling frequency of 1200 Hz. For MEG coregistration, MRI data collected with a Phillips Achieva 7T was employed. MEG data were downsampled to 250 Hz using an anti-aliasing filter, filtered out frequencies <1 Hz, and source-reconstructed to 42 dipoles (covering whole-brain cortical but not subcortical regions) using a linearly constrained minimum variance scalar beamformer. Thirty-eight of these dipoles were generated using an ICA decomposition on the HCP

resting-state fMRI data, and the remaining correspond to the anterior and posterior precuneus. A weighted mask was employed to project the data to brain space, with each area having its highest value near the centre of gravity. Bad segments were manually eliminated, and spatial leakage was corrected. The current study employed a selection of 10 people from the remaining 55 participants (mean age 26.5 years, maximum age 48 years, minimum age 18 years, 35 men).

2.2. Analysis pipeline

For each dataset, we run (i) the HMM, (ii) BR-HMM and (iii) PCCA. The results of the fMRI and the MEG datasets will be presented using the same order.

2.2.1. Hidden Markov Model

Our analysis begins by estimating time-varying FC patterns using an HMM. The model assumes that time series can be described using a hidden sequence of a finite number of states. In a data-driven way, the HMM characterizes neural timeseries of concatenated data (**Fig. 1A**) using a finite number of K (group-level) states that reoccur across time (**Fig. 1B**). Each state timeseries represents the probability for that state to be active at each time point. Note that the spatial parameters of the HMM are at the group level, while the state time series are specific to each subject. The states are probability distributions within a certain family, and each of them is characterized by a certain set of parameters. To focus on FC, each state was here characterized by a Gaussian distribution with no mean parameter (the mean parameter pinned to zero) and a full covariance matrix (representing pairwise covariation across regions); or, equivalently, as a Wishart distribution. Finally, a K-by-K transition probability matrix is estimated as part of the model, containing the probability of transitioning from one state to another or remaining in the current state. For the MEG dataset, the HMM was performed on the power (frequency filtered into the alpha-band; 8–12 Hz) rather than the raw time series.

The HMM relies on an optimization procedure that starts from a random seed, so it is subject to statistical noise. As a result, different runs of the same model can produce different results. To evaluate how much the results across runs differed from each other, we run the model multiple times (R) and computed the similarity of the estimates across runs. Specifically, the *between-run similarity* was measured for each pair of runs as the statistical dependence between two sets of state timeseries. Because the ordering of the states within a run is arbitrary, the states of a run were first aligned to the states of the other run using the Hungarian algorithm (Munkres, 1957), so that the first state in one run corresponds to the first state in another run, and so forth.

Having quantified the similarity between pairs of runs, it is now possible to explore whether there are groups of runs that are more similar to each other than the rest of the runs; and whether there are more than one distinct group of runs, i.e., more than one HMM solution. To describe the structure of similarities across runs we applied a hierarchical clustering algorithm on the matrix of between-run similarity scores, S. As a distance measure we used $D_{ij}=1-S_{ij}$, where S_{ij} is the similarity score between the ith and j^{th} run. The clustering algorithm starts by regarding each element as a separate cluster; then the closest pair of clusters are iteratively combined into a larger cluster. To measure the distance between two clusters, we used Ward's linkage, which minimizes the variance of the clusters being merged. Dendrograms were used to describe the result of the hierarchical clustering. Each cluster can be seen as a group of runs with the same underlying HMM solution. The size of a cluster indicates how many runs return that solution. The distance between clusters indicates how different the solutions are from each other. Broadly speaking, the higher the number of distant clusters, the higher the number of unique solutions. This aspect is especially important as it determines which of the proposed approaches will be more efficient. We advance that PCCA is

particularly recommended when there are two or more distant solutions that are equally prevalent since PCCA works by integrating information across solutions.

While the between-run similarity provides a measure of the overall correspondence between two sets of states, we are often interested in the specific correspondence of the states: for example, in hypothesis-driven research with specific FC patterns of interest in mind. To explore the stability of the individual states, we first computed the *between-state correlation* for every pair of states, i.e., Pearson correlation between two state timeseries. The total number of states $M=K\times R$. Then, we assessed the structure of similarities across states by applying Ward's hierarchical clustering on the between-state correlation matrix, P. As a distance measure, we used $D_{ij}=1-P_{ij}$, where P_{ij} represents the correlation coefficient between the ith and j^{th} state. As a result, highly correlated states will be clustered together. Each cluster can be seen as a group of states with the same underlying FC pattern. The size of each cluster indicates how often that pattern is found across runs.

2.2.2. Best-ranked Hidden Markov Model (BR-HMM)

The BR-HMM approach is based on running the model multiple times and using the free energy values of these runs to choose the best run —i.e., the one with the lowest free energy. The free energy in the HMM inference is a Bayesian measure of a trade-off between model fitness to the data and complexity (given by how the posteriors differ from the priors). The equations for the computation of the free energy in the context of the HMM can be found elsewhere (Vidaurre et al., 2016).

With a sufficient number of runs, we found that the run with minimal free energy is considerably stable, i.e., if we run BR-HMM multiple times, the best run is always the same. The question is then how many HMM runs are sufficient to reach that stable minimal value of the free energy. We compare the number of runs against the lowest free energy value across runs to find out.

We explored this effect statistically by quantifying the stability of the estimations from the BR runs across N repetitions of the entire BR-HMM procedure. In each repetition, the HMM was run R times and the lowest free energy run was selected as the BR run of that repetition. Again here, for each pair of, in this case, BR runs, we computed the between-run similarity and the between-state correlation to measure the correspondence across estimates. Finally, we examined the relationship between stability and the number of runs, from R=10 to R=300. We will demonstrate that the more complex the landscape of HMM solutions is, the more runs are required to find the most prominent solution.

2.2.3. Principal Components of Connectivity Analysis (PCCA)

The logic of PCCA is to take advantage of the HMM inference variability by refactoring the different estimations into stable, latent patterns of time-varying FC. As shown in **Fig. 1B** and **1D**, we apply PCA to the concatenated set of K state timeseries across R runs. By doing so, we reduce the dimensionality of the data from $M=K\times R$ states to a small set of orthogonal principal components (PC) that capture the dominant fluctuations in FC, ordered from more to less explained variance. The novelty here is that PCCA captures the principal components, not of signal covariance, but changes in covariance —and therefore of changes in FC.

In detail, the state timeseries generated over multiple runs (R) were concatenated into a matrix (Y) with as many rows as frames in the dataset (time × subjects) and M columns ($R\times K$ states). PCA was then performed by eigendecomposing the gram matrix $Y Y^T$, which returns a set of M eigenvectors and eigenvalues according to the equation:

$$YY^T = U\Lambda U^T \quad (1)$$

U contained the (orthonormal) eigenvectors on its columns and Λ contained the corresponding eigenvalues in its diagonal. These eigenvectors, or PCs, represented patterns of variation across all time-varying FC states, while the eigenvalues described the contribution of each PC in terms of explained variance. The PCs (columns of U) were arranged according to their relative contributions to Y, from higher to lower eigenvalues.

To estimate the spatial patterns of time-varying FC associated with each PC, we computed one covariance matrix for each PC (denoted as Q_i). Q_i was obtained as a linear combination of the covariance matrices of the original states (C_m), using the corresponding PCA coefficients (w_m) as weights. However, a linear combination of the covariance matrix is only guaranteed to be a proper covariance matrix if the mixing weights are all positive. Since PCA coefficients can also be negative, this poses a problem. To preserve the covariance topology, we first project each state covariance into the tangent space (Bellaïche, 1996)—a Euclidean space where arbitrary linear combinations are well-defined, and which is computed using the average static FC across subjects. We then compute, in this approximated space, the linear combination of the projected covariances using w_m. Finally, we project back the resulting linear combination to the original covariance space, ensuring that the geometry of the covariance space is well-preserved. Specific mathematical details can be found in Congedo et al. (2017), where this procedure was applied to EEG data.

To facilitate visualization of the spatial pattern associated with each PC, we used the first eigenvector of the corresponding PC covariance matrix (Q_i), which represented the main pattern of FC for that PC. Note that this is a different eigendecomposition than Eq. (1), which is done on the PC timeseries.

To evaluate the reliability of the results, PCCA was repeated several times and the results across repetitions were compared. We also quantified the relationship between the stability of PCCA and the number of runs, for various choices of R.

2.3. Data availability

The fMRI data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The MEG dataset is part of a larger dataset acquired in Nottingham in the context of the MEG UK Partnership and is not currently available as it contains data from human participants including structural scans. The data are held by the MEG UK Partnership, and access to the MEG UK Database can be requested at http://meguk.ac.uk/contact. Preprocessed (and parcellated) data containing the time series as they were fed to the HMM can be accessed at https://github.com/sonsolesalonsomartinez/reproducibleHMM.

2.4. Software and code availability

All analyses in this paper were carried out using freely available tools in MATLAB (R2021b, Math-Works, Naticks). The software comprises the open-access code repositories https://github.com/OHBA-analysis/HMM-MAR and https://github.com/sonsolesalonsomartinez/reproducibleHMM.

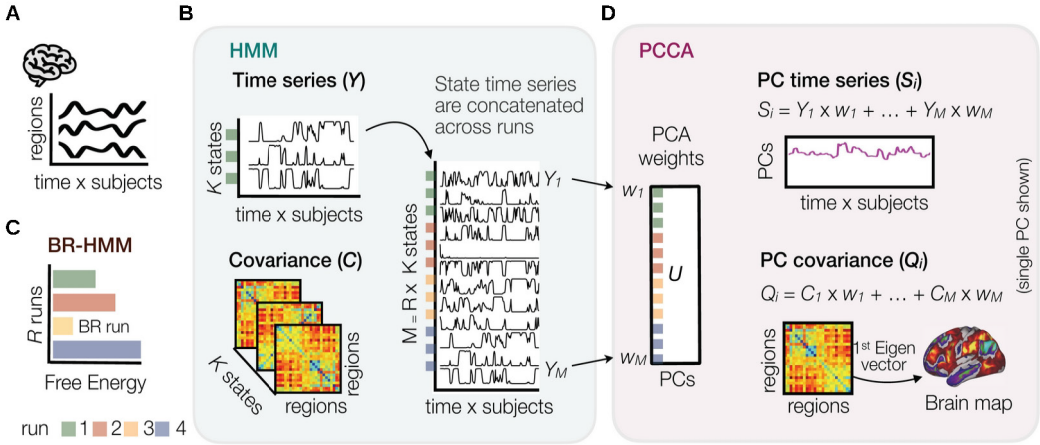

Fig. 1 Graphical summary of the two approaches (BR-HMM and PCCA) to achieve reliable patterns of time-varying FC. A fMRI or MEG timeseries, temporally concatenated across subjects. **B** An HMM with *K* states was inferred from the concatenated brain timeseries to obtain patterns of time-varying FC (i.e., covariance matrices across regions, *C*). For the MEG dataset, the HMM was run on power timeseries. The state timeseries (*Y*) contain the probability of a given state to be active at each time point. The HMM was run several times (*R*), each starting from a different random initialization. **C** BR-HMM approach: the multiple runs are ranked according to the free energy, a quantitative measure that weighs the fitness of the data and simplicity of the model; the best HMM is selected. **D** PCCA approach: PCA was applied to the concatenated state timeseries across the multiple runs ([$R \times K$] by [time × subjects]), projecting the M (=$R \times K$) rows into a reduced set of principal–temporal–components (S) that explain maximum variance. The covariance matrix of each component (Q_i) was estimated as a linear combination of the covariance matrices of the original (M) states, using the PCA coefficients (w_m) as their weights, while preserving the Riemannian geometry of the covariance space (see *Materials and Methods*). Brain connectivity maps were created using the 1st Eigenvector of each PC covariance. BR-HMM: Best-ranked-HMM; HMM: hidden Markov model(ling); PC: principal component; PCA: principal component analysis; PCCA: principal components of connectivity analysis.

3. Results

The following results, for both fMRI and MEG, follow the analysis pipelines detailed in the *Materials and Methods* section. All models were run at the group level, and all results are group-level results.

3.1. fMRI data

The fMRI dataset considered here consisted of 4 sessions of resting-state fMRI data (repetition time, TR=0.72s) from 100 subjects obtained from the HCP (Van Essen et al., 2013). The data were projected onto 25 independent components (ICs), which we used as input time series. Whole-brain patterns of time-varying FC were obtained by applying an HMM to the concatenated (standardized) timeseries for all subjects (25 ICs by [time × subjects]). Here the HMM was modelled with K=12 states, where each state represents a reoccurring pattern of FC.

3.1.1. HMM stability

The stability of the HMM inference was evaluated using the between-run similarities across 200 runs. **Fig. 2A** displays a heatmap of the similarity scores. Each entry of the matrix has a value between 0 and 1, which is proportional to how close the set of states of a run is to the set of states of a different run. The order of the runs was sorted according to a hierarchical clustering on the similarity matrix, represented by a dendrogram on the top and left axis. Two main clusters of runs

were obtained. Most of the runs were included in the first cluster. Red colours highlight clusters of similar runs. Related clusters (i.e., clusters that have similar but not identical runs) can be distinguished as the intensity of red shades into lighter red. Unrelated clusters (i.e., clusters that have different runs) can be appreciated as the intensity of red gradually shades into blue. A histogram of the off-diagonal elements of this matrix revealed a minimum similarity score of 0.48, a maximum of 0.98, and an average score of 0.68. This indicates that the states inferred by the HMM are to some extent different across runs of the inference, as shown previously in Vidaurre et al., 2019.

Fig. 2B displays the free energy values for the 200 runs. The order of the runs was arranged as in Fig. 2A. Within each cluster, some runs ranked better than others with respect to free energy values. However, the BR runs, i.e., the runs with the lowest free energy, all belong to *different* subclusters of the *same* parent cluster (the first cluster, shown in purple). This means that the BR runs across these 200 runs all represent *varying* forms of the *same* underlying HMM solution.

The stability of each specific state was evaluated as the between-state correlation across the K states generated over R runs (200×12=2400 states). **Fig. 2C** shows a matrix of the Pearson correlation coefficients between each pair of state timeseries. Again, the states were ordered according to Ward's hierarchical clustering to help visualize clusters of similar states. By definition, clusters of similar states tend to concentrate around the main diagonal. The stability of each cluster can be evaluated by the size and colour of that cluster. The size indicates how many runs returned the same underlying state, while the colour indicates how correlated they were to each other. If the HMM results were identical (between-run similarity=1), the expected number of clusters would be K=12 states, and the expected size of each cluster would be R=200 runs. Our results show that 8 out of 12 states were relatively stable (see the high correlation coefficients shown in red). **Supplementary Fig. S1** contains details about the underlying FC patterns. Note that a lack of correlation between clusters is expected given that the states derived from the HMM are exclusive i.e., the total probability of activation across states at a given time point is 1.

3.1.2. BR-HMM stability

The BR run of the HMM was selected as the lowest free energy run across 200 runs. This step was repeated 10 times. For each time, 200 HMM runs were estimated and the run with the lowest free energy was chosen. We then computed the between-run similarities across 10 BR runs. **Fig. 2D** shows a significant improvement in the stability of the estimations across BR runs. The between-run similarity now ranges between 0.81 and 0.98, with an average of 0.93 (SD=0.06). As a result, all the K=12 states are now equally reproducible. That is, for each state k, the between-state correlations across the 10 BR runs are very close to 1; see **Fig. 2E** compared to Fig. 2C. These results demonstrate that running the model 200 times and selecting the one with the lowest free energy improve the stability of the results. However, this can be computationally intensive. We estimated the minimum number of runs needed to achieve stable results. A first approximation can be done by plotting the number of runs against the minimum free energy scores from those runs (**Fig. 2F**). We repeated this step 10 times and computed the mean and the standard error of the mean by averaging the lowest free energy values across the 10 repetitions (one value per repetition). As the number of runs increases, the minimum free energy values tend to stabilize quickly. Here, the long tale of stability starts around 225 runs. We then plot the number of runs against the between-run similarity scores from those runs. Again, the mean and the standard error of the mean were obtained by averaging the similarity scores across the 10 repetitions. Our results indicate that reliable estimations (i.e., between-run similarity score above 0.9) can only be guaranteed with more than 225 repetitions for this data set.

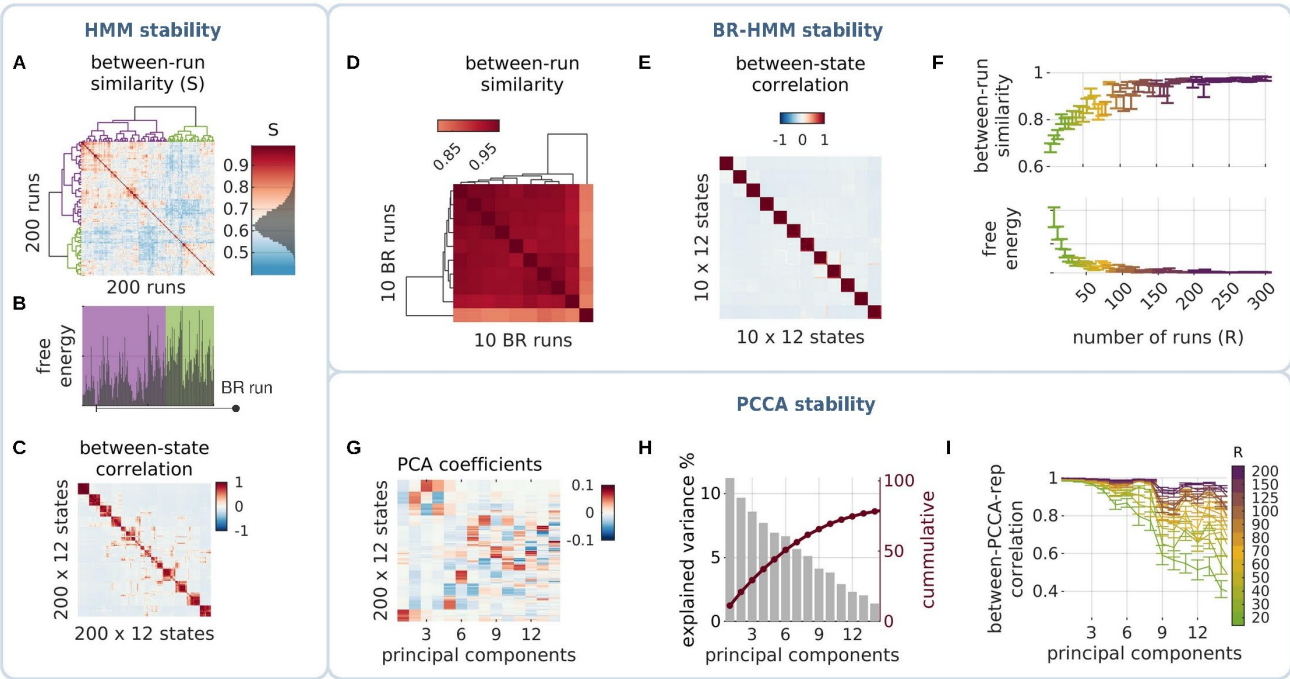

Fig. 2 Capturing reliable time-varying FC estimations from resting-state fMRI data. A Matrix of the between-run similarities across pair of runs; R=200 HMM runs. States are ordered according to hierarchical clustering to help visualize clusters of similar runs. The similarity was computed as the sum of the joint probabilities across two sets of K=12 state time series (see *Materials and Methods*). At the right, a histogram of the upper off-diagonal elements from the similarity matrix. **B** Free energy levels for each HMM run as sorted in A. **C** Matrix of the Pearson correlation coefficients between pairs of state timeseries (total number of states $M=R\times K$ states); states were ordered according to hierarchical clustering on this matrix to help visualize clusters of similar states. **D** Similarity across 10 BR runs. **E** Between-state correlation across states from BR runs. **F** Number of runs against the minimum free-energy scores across runs and corresponding between-run similarity. **G** State contribution (PCA coefficients) to the final PCCA model for the top 14 PCs; states are ordered according to the hierarchical clustering of the between-state correlation matrix as shown in panel C. **H** Percentage of the individual (bars) and cumulative (line) explained variance by the top 14 PCs that resulted from applying PCA to the M states. **I** Pearson correlation (y-axis) of the 14 PCs (x-axis) between every pair of (J=) 30 PCCA repetitions, for several choices of R (see legend); error bars are the standard error of the mean correlation across all pairs of PCCA repetitions.

3.1.3. PCCA stability

PCCA is designed to capture the dominant, or principal, patterns of time-varying FC by integrating information across multiple HMM runs. Specifically, PCA was applied to the concatenated sets of K=12 state timeseries across R=200 runs. The total number of states M was 2400 ($=R\times K$). This produced a set of orthogonal components that optimally explained maximum time-varying FC variability across all M states. **Fig. 2G** displays the contributions (PCA coefficients) of each of these states to each of the PCs. The states were ordered according to the hierarchical clustering on the between-state-correlations matrix shown in Fig. 2C. This panel shows that the group of states belonging to the eight most stable groups of states highly contribute to the top eight PCs shown in panel 2G. The spatial patterns of time-varying FC associated with each of the top PCs can be seen in **supplementary Fig. S2**. As revealed in **Fig. 2H**, the top eight PCs alone retained 70% of the total temporal variance present across all the states. We then assessed the reliability of these PCs across 30 different PCCA repetitions. Each repetition was applied on a different set of 200×12 state timeseries. The reliability of each PC was assessed as the Pearson correlation between the PC timeseries across repetitions. The average correlation was r≈1.0 for PC1 and PC2; r≈0.99 for PC3 to PC5; and r≈0.95 for PC6 to PC8; see the purple line in **Fig. 2I**. We further quantified the relationship between PCCA stability and the number of runs (R referenced on the colour map). We

observed that 40 runs are sufficient to make the top eight PCs highly stable (r>0.9). While this stability decreased for higher-order PCs, adding more HMM runs did increase the number of stable PCs.

3.1.4. Interim conclusion

Taken together, despite the stochastic nature of the HMM inference, momentary changes in FC can also be reliably captured in fMRI using the two proposed approaches. On the one hand, the accuracy of BR-HMM was bound to multiple inference realizations in order to adequately describe the landscape of possible HMM solutions. PCCA, on the other hand, proved to be exceptionally stable against estimation noise, which other techniques suffer systematically; see Hindriks et al. (2016) and Vidaurre et al. (2019).

3.2. MEG data

The MEG dataset considered here consists of 5-minute resting state recordings from 10 subjects. The data were mapped to a 42-region parcellation. HMMs with K=6 states were inferred from the temporally concatenated data across subjects to obtain whole-brain patterns of time-varying FC. For easier comparison with the fMRI results, the HMM was applied on the power (band-limited to 8–12 Hz) instead of the raw timeseries; see *Materials and Methods* for further details. A higher number of fixed states in the model (K=12) led to a much higher inference variability; see results in **supplementary Fig. S3**, where 1000 runs of this model were not sufficient to obtain reliable estimations. The average between-run similarity score across 1000 runs was 0.36 (SD=0.03); see **supplementary Fig. S4**.

3.2.1. HMM stability

To evaluate the stability of the HMM inference for a model with K=6 states, the model was run 200 times and the similarity across runs was computed. **Fig. 3A** displays a heatmap of the similarity scores. The order of the runs was sorted according to a hierarchical clustering on the similarity matrix, represented by a dendrogram on the top and left axis. Two main clusters of runs were obtained. Most of the runs were included in the first cluster. A histogram of the off-diagonal elements of this matrix shows a minimum similarity score of 0.39, a maximum of 0.94, and an average score of 0.65 (SD=0.09). These results indicate a slightly higher inference variability compared to the first dataset. The free energy values for each of the 200 runs are displayed in **Fig. 3B**. The order of runs is arranged according to Fig. 3A. Whereas low free energy values were found across the two clusters, the minimum free energy values were found in the first cluster— shown in purple. The stability of each specific state was evaluated as the between-state correlation across all K=6 states generated over 200 runs. **Fig. 3C** displays a heatmap of the correlation coefficients across pairs of states. The states were rearranged according to the hierarchical structure of the correlations. Groups of correlated states can be seen along the main diagonal. Three of these groups or clusters were characterized by high levels of within-cluster correlation and lower levels of between-cluster correlation, whereas the correlations of other clusters were more heterogeneous. This indicates that there are some states more stable than others, i.e., less affected by the random initialization of the model. The pattern of FC representing these states can be seen in **supplementary Fig. S5**.

3.2.2. BR-HMM stability

For a set of 200 runs, the BR run was selected as the lowest free-energy run. We repeated this process 10 times, obtaining 10 different sets of 200 runs and selecting the BR run for each set. The

stability of BR runs was evaluated by computing the between-run similarity score for each pair of BR runs. The average similarity score across BR runs was 0.89 (SD=0.03), the minimum 0.83 and the maximum 0.94, which was significantly higher than that (0.65) across individual runs (**Fig. 3D**). Consequently, the K=6 states are now equally reproducible; see **Fig. 3E** compared to Fig. 2C. Our results demonstrate that running the model 200 times and selecting the one with the lowest free energy significantly improves the stability of the estimations. However, not all the comparisons across BR runs reached a 0.9 threshold of reliability (between-run similarity). The minimum number of runs needed to obtain reliable results can be estimated by the probability of finding the minimum energy value across those numbers of runs. **Fig. 3F** shows that 300 runs were not enough to exceed this level.

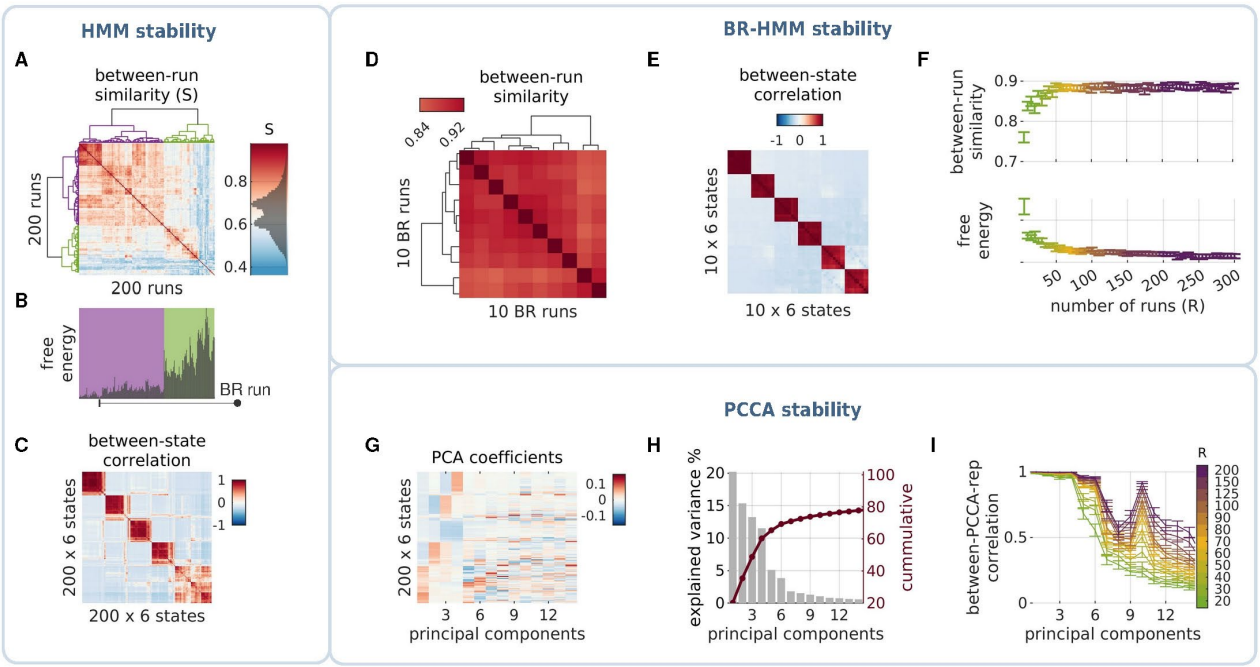

Fig. 3 Capturing reliable time-varying FC estimations from resting-state (alpha-band) MEG data. A Matrix of between-run similarities across pair of runs; R=200 HMM runs. States are ordered according to hierarchical clustering to help visualize clusters of similar runs. Similarity was computed as the sum of the joint probabilities across two sets of K=6 state time series. Histogram of the upper off-diagonal elements in the similarity matrix. **B** Free energy levels for each HMM run as sorted in A. **C** Matrix of the Pearson correlation coefficients between pairs of state timeseries (total number of states $M=R\times K$ states). States were ordered according to hierarchical clustering to help visualize clusters of similar states. **D** Between-run similarity across 10 BR runs. **E** Between-state correlation across states from BR runs. **F** Number of runs against the minimum free-energy scores across runs and corresponding between-run similarity. **G** State contribution (PCA coefficients) to the final PCCA model for the top 14 PCs; states are ordered according to the hierarchical clustering of the between-state correlation matrix as shown in C. **H** Percentage of the individual (bars) and cumulative (line) explained variance by the top 14 PCs that resulted from applying PCA to the M states. **I** Pearson correlation (y-axis) of the 14 PCs (x-axis) between each pair of (J=) 30 PCCA repetitions, for several choices of R (see colorbar); error bars are the standard error of the mean correlation across all pairs of PCCA repetitions.

3.2.3. PCCA stability

PCA was applied to the concatenated K=6 state timeseries across R=200 runs. This produced a set of orthogonal components that optimally explained maximum time-varying FC variability across all states. **Fig. 3G** shows the PCA coefficients of each state to the top 14 components. The states were

ordered according to the hierarchical clustering on the between-state correlation matrix displayed in Fig. 3C. The states that most contribute to the top three components represent the three groups or clusters of states found to be most stable. These three components jointly explained almost half of the total temporal variance present across all (200×6) states. When we consider the top 6 components, we can explain almost 70% of the total variance (**Fig. 3H**). The spatial patterns of time-varying FC associated with each of these PCs can be seen in **supplementary Fig. S6**. We assessed the reliability of the top 14 components across 30 different PCCA repetitions. Each PCCA repetition was applied on a different set of 200×6 state timeseries. We computed the Pearson correlation between the PC timeseries across PCCA repetitions. The average correlation was $r \approx 1.0$ for PC1 to PC4 and $r \approx 0.96$ for PC5 and PC6; see the purple line in **Fig. 3I**. We observed that 20 runs are sufficient to make the top four components reliable ($r > 0.95$). While this stability decreased for higher-order components, adding more HMM runs did increase the number of reliable components.

3.2.4. Interim conclusion

In summary, our results demonstrate that both approaches can produce reliable patterns of time-varying FC even when the estimations across runs differ substantially. However, when the estimations across runs are that different, obtaining a reliable BR-HMM can become computationally intensive, making PCCA a more suitable approach for this data set.

4. Discussion

Much previous work applies sliding windows to compute time-varying FC. However, the detection of genuine dynamics using this approach is limited because of the great statistical variability of these estimations. An alternative that does not suffer from such limitations and allows modelling time point by time point variability of FC, is the hidden Markov model (HMM), which uses a finite number of states to characterize time-varying FC (Vidaurre et al., 2017). However, a critical limitation of the HMM, as well as many other models, is the fact that different runs of the inference can potentially produce different results even for the same data set. We here proposed two approaches, BR-HMM and PCCA, to tackle this limitation.

Results from two fMRI and MEG datasets demonstrate that the two proposed approaches can significantly improve the stability of time-varying FC estimations obtained from an HMM. The main advantage of using BR-HMM is that the result is a regular HMM, which we can interpret and use as in previous work (Vidaurre et al., 2021). This is helpful, for example, if we wish to characterize the individual profiles of FC states in terms of their probability of activation or its duration, or the specific switches between states, for example in response to a task. The main disadvantage of this approach, however, is that numerous runs of the time-varying model might be required to reach a level of stability that is considered sufficient. We observed that the higher the inference variability, the more runs are needed to successfully identify runs with minimum free energy. On the other hand, PCCA can return reliable components of time-varying FC with just a fraction of the runs, at the expense of producing a different model from the HMM. For example, the PCCA time courses are not any more probabilities, and the model does not include a proper transition probability matrix; however, the resulting components do have the structure of an HMM state.

We note that the inference variability of the HMM can be due to two causes: the complexity of the landscape of FC dynamics, which cannot be accurately described with the parameters of the model nor sufficiently explained by a single realization of such model; and pure estimation noise, due to

limited data being available or high signal-to-noise ratio. While this is not a problem regarding the interpretation of our results, it can be if we wish to compare the complexity of the landscape of FC dynamics between two conditions with different levels of noise and amount of data. Although not explored in this paper, the two sources of variability might potentially be disentangled by looking at the PCA eigenspectra patterns. Another avenue for improvement is to explore alternatives for PCA that might be better suited for time series of probabilities (Lee et al., 2010). Also, note that both the HMM and PCA are run at the group level, meaning that while the HMM and PCCA time series are estimated at the subject level, all the spatial estimates are at the group level. Future approaches could aim at obtaining a hierarchical estimation to boost between-subject or between-condition differences; similar to Harrison et al. (2015).

Overall, we observed that there is much information that can be obtained from the various distinct HMM solutions, and this variability can be leveraged for prediction or statistical testing (Vidaurre et al., 2019). Depending on the research goals, it might be useful to exploit all this inference variability to extract additional information rather than just discard it. Here, PCCA has proven handy in making use of the inference variability to return a small set of orthogonal components that reliably capture the dominant fluctuations of FC. Since PCCA components can concentrate more information than FC patterns from a single HMM solution, PCCA may be a powerful approach to characterize cognition and behaviour more exhaustively.

5. Conclusion

Reproducibility is a growing concern in neuroimaging, and for time-varying FC methods this problem can be exacerbated by the stochastic nature of the estimations. Focusing on the hidden Markov model, we present two methods to achieve stable estimations that do not vary from run to run: BR-HMM and PCCA. While BR-HMM needs many inference runs to adequately describe the landscape of possible HMM solutions, PCCA can be more stable using fewer inference runs. BR-HMM, however, is the only method that returns a proper HMM solution.

References

Beckmann CF, Smith SM. 2004. Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Trans Med Imaging* **23**:137–152. doi:10.1109/TMI.2003.822821

Bellaïche A. 1996. The tangent space in sub-Riemannian geometry. *Sub-Riemannian Geom* 1–78. doi:10.1007/978-3-0348-9210-0_1

Congedo M, Barachant A, Bhatia R. 2017. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *http://dx.doi.org/101080/2326263X20171297192* **4**:155–174. doi:10.1080/2326263X.2017.1297192

Dafflon J, F. Da Costa P, Váša F, Monti RP, Bzdok D, Hellyer PJ, Turkheimer F, Smallwood J, Jones E, Leech R. 2022. A guided multiverse study of neuroimaging analyses. *Nat Commun 2022 131* **13**:1–13. doi:10.1038/s41467-022-31347-8

Fornito A, Bullmore ET. 2010. What can spontaneous fluctuations of the blood oxygenation-level-dependent signal tell us about psychiatric disorders? *Curr Opin Psychiatry* **23**:239–49. doi:10.1097/YCO.0b013e328337d78d

Friston KJ. 1994. Functional and effective connectivity in neuroimaging: A synthesis. *Hum Brain Mapp* **2**:56–78. doi:10.1002/hbm.460020107

Griffanti L, Salimi-Khorshidi G, Beckmann CF, Auerbach EJ, Douaud G, Sexton CE, Zsoldos E, Ebmeier KP, Filippini N, Mackay CE, Moeller S, Xu J, Yacoub E, Baselli G, Ugurbil K, Miller KL, Smith SM. 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* **95**:232–247. doi:10.1016/j.neuroimage.2014.03.034

Harrison SJ, Woolrich MW, Robinson EC, Glasser MF, Beckmann CF, Jenkinson M, Smith SM. 2015. Large-scale probabilistic functional modes from resting state fMRI. *Neuroimage* **109**:217–231. doi:10.1016/j.neuroimage.2015.01.013

Hindriks R, Adhikari MHH, Murayama Y, Ganzetti M, Mantini D, Logothetis NKK, Deco G. 2016. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage* **127**:242–256. doi:10.1016/j.neuroimage.2015.11.055

Hunt BAE, Tewarie PK, Mougin OE, Geades N, Jones DK, Singh KD, Morris PG, Gowland PA, Brookes MJ. 2016. Relationships between cortical myeloarchitecture and electrophysiological networks **113**:13510–13515.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. 2012. FSL. *Neuroimage* **62**:782–790. doi:10.1016/j.neuroimage.2011.09.015

Karapanagiotidis T, Vidaurre D, Quinn AJ, Vatansever D, Poerio GL, Turnbull A, Ho NSP, Leech R, Bernhardt BC, Jefferies E, Margulies DS, Nichols TE, Woolrich MW, Smallwood J. 2020. The psychological correlates of distinct neural states occurring during wakeful rest. *Sci Rep* **10**:21121. doi:10.1038/s41598-020-77336-z

Laughlin SB, Sejnowski TJ. 2003. Communication in Neuronal Networks. *Science (80-)* **301**:1870–1874. doi:10.1126/science.1089662

Lee S, Huang JZ, Hu J. 2010. Sparse logistic principal components analysis for binary data. *Ann Appl Stat* **4**:1579–1601. doi:10.1214/10-AOAS327

Liégeois R, Li J, Kong R, Orban C, Van De Ville D, Ge T, Sabuncu MR, Yeo BTT. 2019. Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nat Commun* **10**:2317. doi:10.1038/s41467-019-10317-7

Lurie DJ, Kessler D, Bassett DS, Betzel RF, Breakspear M, Keilholz S, Kucyi A, Liégeois R, Lindquist MA, McIntosh AR, Poldrack RA, Shine JM, Thompson WH, Bielczyk NZ, Douw L, Kraft D, Miller RL, Muthuraman M, Pasquini L, Razi A, Vidaurre D, Xie H, Calhoun VD. 2020. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Netw Neurosci* **4**:30–69. doi:10.1162/NETN_A_00116

Munkres J. 1957. Algorithms for the Assignment and Transportation Problems. *J Soc Ind Appl Math* **5**:32–38. doi:10.1137/0105003

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. 2013. The WU-Minn Human Connectome Project: An overview. *Neuroimage* **80**:62–79. doi:10.1016/j.neuroimage.2013.05.041

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, Della Penna S, Feinberg D, Glasser MF, Harel N, Heath AC, Larson-Prior L, Marcus D, Michalareas G, Moeller S, Oostenveld R, Petersen SE, Prior F, Schlaggar BL, Smith SM, Snyder AZ, Xu J, Yacoub E. 2012. The Human Connectome Project: A data acquisition perspective. *Neuroimage* **62**:2222–2231. doi:10.1016/j.neuroimage.2012.02.018

Vidaurre D, Hunt LT, Quinn AJ, Hunt BAEE, Brookes MJ, Nobre AC, Woolrich MW. 2018. Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. *Nat Commun* **9**:2987. doi:10.1038/s41467-018-05316-z

Vidaurre D, Llera A, Smith SMM, Woolrich MWW. 2021. Behavioural relevance of spontaneous, transient brain network interactions in fMRI. *Neuroimage* **229**:117713. doi:10.1016/j.neuroimage.2020.117713

Vidaurre D, Quinn AJ, Baker AP, Dupret D, Tejero-Cantero A, Woolrich MW. 2016. Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage* **126**:81–95. doi:10.1016/j.neuroimage.2015.11.047

Vidaurre D, Smith SM, Woolrich MW. 2017. Brain network dynamics are hierarchically organized in time. *Proc Natl Acad Sci* **114**:12827–12832. doi:10.1073/pnas.1705120114

Vidaurre D, Woolrich MW, Winkler AM, Karapanagiotidis T, Smallwood J, Nichols TE. 2019. Stable between-subject statistical inference from unstable within-subject functional connectivity estimates. *Hum Brain Mapp* **40**:1234–1243. doi:10.1002/hbm.24442

Xie H, Gonzalez-Castillo J, Handwerker DA, Bandettini PA, Calhoun VD, Chen G, Damaraju E, Liu X, Mitra S. 2018. Time-varying whole-brain functional network connectivity coupled to task engagement. *Netw Neurosci* **3**:49–66. doi:10.1162/netn_a_0005