Supplementary Information for "*Brain network dynamics fingerprints are resilient to data heterogeneity*"

Tommaso Menara
Department of Decoded Neurofeedback,
ATR Computational Neuroscience Laboratories,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
Department of Mechanical Engineering,
University of California Riverside,
900 University Ave, Riverside, CA, 92521, USA

Giuseppe Lisi
Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555, Japan
ATR Brain Information Communication Research Laboratory Group,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan

Aurelio Cortese
Department of Decoded Neurofeedback,
ATR Computational Neuroscience Laboratories,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
cortese.aurelio@gmail.com

January 26, 2020

SI Materials

Traveling-subject Dataset Acquisition and Preprocessing

The Traveling-subject dataset is one of the largest and most heterogeneous multi-site collection of resting-state fMRI data for the same traveling subjects [4]. The following is a list of the acronyms for Table S1, which exhaustively describes the scanning protocols for this dataset. UTO: University of Tokyo; HUH: Hiroshima University Hospital; KUT: Siemens TimTrio scanner at Kyoto University; ATT: Siemens TimTrio scanner at Advanced Telecommunications Research Institute International; ATV: Siemens Verio scanner at Advanced Telecommunications Research Institute International; SWA: Showa University; HKH: Hiroshima Kajikawa Hospital; COI: Center of Innovation in Hiroshima University; KUS: Siemens Skyra scanner at Kyoto University; KPM: Kyoto Prefectural University of Medicine; YC1: Yaesu Clinic 1; YC2: Yaesu Clinic 2; TR: repetition time; TE: echo time.

This dataset consists of 141 scans, collected at 12 different sites for 9 subjects. Each participant underwent three rs-fMRI sessions of 10 min each at nine sites, two sessions of 10 min each at two sites (HKH and HUH), and five cycles (morning, afternoon, next day, next week, and next month) consisting of three 10-min sessions each at a single site (ATT). In the latter situation, one participant underwent four rather than five sessions at the ATT site because of a poor physical condition. Thus, a total of 411 sessions were conducted [8 participants $\times$ ($3 \times 9 + 2 \times 2 + 5 \times 3 \times 1$) + 1 participant $\times$ ($3 \times 9 + 2 \times 2 + 4 \times 3 \times 1$)]. During each session, participants were instructed to maintain a focus on a fixation point at the center of a screen, remain still and awake, and to think about nothing in particular. For sites that could not use a screen in conjunction with fMRI (HKH and KUS), a seal indicating the fixation point was placed on the inside wall of the MRI gantry. Differences between scanning sites include two phase-encoding directions (P$\rightarrow$A and A$\rightarrow$P), three MRI manufacturers (Siemens, GE, and Philips), four different numbers of channels per coil (8, 12, 24, and 32), and seven scanner types (TimTrio, Verio, Skyra, Spectra, MR750W, SignaHDxt, and Achieva). In regards to other scanning parameters, a large effort was made to ensure that imaging was performed using the same variables at all sites.

The fMRI data in the Traveling-subject dataset were preprocessed using *SPM8* implemented in *MATLAB* (R2016b; Mathworks, Natick, MA, USA). The first 10 s of data was discarded to allow for T1 equilibration. Preprocessing steps included slice-timing correction, realignment, coregistration, segmentation of T1-weighted structural images, normalization to Montreal Neurological Institute (MNI) space, and spatial smoothing with an isotropic Gaussian kernel of 6 mm full-width at half-maximum (for additional details, see [4]). If the number of volumes removed after scrubbing exceeded the average

of -3 standard deviations across participants, the sessions were excluded. As a result, 14 sessions were removed from the dataset.

SI Methods

Hidden Markov Model

Hidden Markov modeling is a powerful technique that enables the description of time series extracted from a system of interest. This can be done utilizing the theory of Markov models to make an educated guess about the structure of the process generating the data. Analyses of hidden Markov models seek to recover the sequence of states from some observed data. The underlying assumption of this class of models is that the observed time series of data can be explained by a discrete sequence of hidden states, which must be finite in number. Additionally, to describe a hidden Markov model, an observation model needs to be chosen. We assume multivariate Gaussian observation model, so that, if $\mathbf{x}_t$ denotes the data at time step t, and s_t represents the state at time step t, we can write, whenever state k is active,

$$\mathbf{x}_t | s_t \sim \text{multivariate Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^c$ is the vector of the mean blood oxygen level-dependent (BOLD) activation for each channel, with c being the number of channels in the data, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{c \times c}$ is the covariance matrix encoding the variances and covariances between channels. The transitions between different brain states depend on which state is active at the previous time step. Specifically, the probability of a state being active at time t depends on which state is active at time step $t-1$. This is encoded in the Transition Probability Matrix $\boldsymbol{\Theta}$, in which the entry $\boldsymbol{\Theta}_{ij}$ – the transition probability – denotes the probability of state i becoming active at the next time step if state j is currently active. Formally, by denoting a probability with Pr, it holds that

$$\Pr(s_t = i) = \sum_j \boldsymbol{\Theta}_{ij} \Pr(s_{t-1} = j)$$

Within the Transition Probability Matrix $\boldsymbol{\Theta}$, the diagonal entries represent the probability of remaining on the actual state and are called persistence probabilities, whereas the off-diagonal entries are called transition probabilities.

For large datasets, it is possible to resort to stochastic Variational Bayes inference to estimate the posterior distribution of each state $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the probability of each state being active at each time step, and the transition probabilities between each pair of states $\boldsymbol{\Theta}_{ij}$ [2]. In conclusion, it is worth noting that, although in this study the model

has been inferred by concatenating all subjects and the brain states are thus an outcome of common brain dynamics, the state time courses are subject-specific. That is, the states are inferred at the group level, but the time instants at which each brain state becomes active is subjective and changes between and across subjects.

Data Preparation for HMM Training

In order to concatenate HCP resting-state fMRI data and the ones from the Traveling-subject dataset, we proceeded as follows. First, we had to match the Traveling-subject voxel coordinates with the spatial maps containing the weights from the ICA decomposition to 50-dimensional space into which the HCP time series have been decomposed. Such decomposition is available as an additional download at the HCP database. The spatial maps were extracted from the group average analysis across all the subjects of the S1200 release and are also provided as an additional download from the HCP website: `https://db.humanconnectome.org`. Because the spatial maps are in a grayordinate CIFTI file [1], it is possible to extract the xyz coordinates in a standard stereotaxic space MNI152 by using a *midthickness* surface file for the surface vertices and the coordinate transformation matrix included in the CIFTI file. Next, we extracted the time series from the Traveling-subject data corresponding to the same xyz coordinates of the aforementioned spatial map in Matlab by using the toolbox DPABI [5]. Finally, we obtained the estimated the 50-dimensional ICs from the extracted time series by means of the HCP group average spatial map. Once the Traveling-subject resting-state fMRI time series have been reduced to 50 ICs, they matched the spatial dimension of the HCP data used to infer the hidden Markov model as in [3].

Control Analysis with Randomized Time Series

To verify the robustness of our analysis in regards to the application of the HMM model to the Travelig-subject dataset, we applied the HCP-trained HMM to randomly permuted time series of the Traveling-subject dataset. Namely, we have applied random permutations to the independent components of the Traveling-subject time series. Next, we proceeded with the HMM decoding on such permuted time series, which yielded state time courses that mostly stay on state 5 for all different scanning sessions. See Fig. S5 for a few examples of the state time courses obtained from HMM decoding of the randomly permute Traveling-subject time series. It is worth noting that state 5 is the state that is mostly uncorrelated from the remaining 11 states and it is the state with the largest variance. In [3], it is shown that state 5 is associated with motion artifacts in the scanner. The outcome of the HMM decoding is in accordance with these observations and supports the robustness of our findings.

Subject Classification Using Brain Dynamics Fingerprints

To support our findings, and the robustness of the subject-specific fingerprints to data heterogeneity, we used machine learning on such fingerprints to perform subject classification. Specifically, we performed a classification task on the subjects in the Traveling-subject dataset based on the two fingerprints used in this study (Metastate Profiles and Fractional Occupancies). Our simple classification experiment revealed that subject classification by means of fingerprints that are based on the subjects' brain dynamics is possible, and that the accuracy of our classifier is way above the baseline chance level.

For each scanning factor, we trained a logistic regression classifier – which minimizes the cross-entropy loss – with the `scikit-learn` maching learning package in Python 3 with the following parameters: default L2 penalty, default L-BFGS-B algorithm [6], and 'multi_class' option set to 'multinomial'. The classification task is repeated multiple times by splitting the data into different training and validation datasets as follows. We repeat the training and validation of the linear regression classifier for each factor attribute (e.g., for the scanner parameter, we repeat the procedure for each scanner model) by performing a leave-one-attribute-out cross-validation: we choose as validation set all the samples (i.e. fingerprints) belonging to one factor attribute, and we used as training set all the remaining samples. A summary of the classification results based on the two fingerprints is in Fig. S6, whereas the classification results for each scanning factor are reported in Table S3. It is worth noting that (1) even with different scanning protocols and heterogeneous data, the classification based on brain dynamics fingerprints performs way above the baseline chance level, and (2) the scanning factors site and day are the ones with the lowest accuracy among all the classification results, in accordance with our claim that some factors tend to make data collections more spurious than other factors.

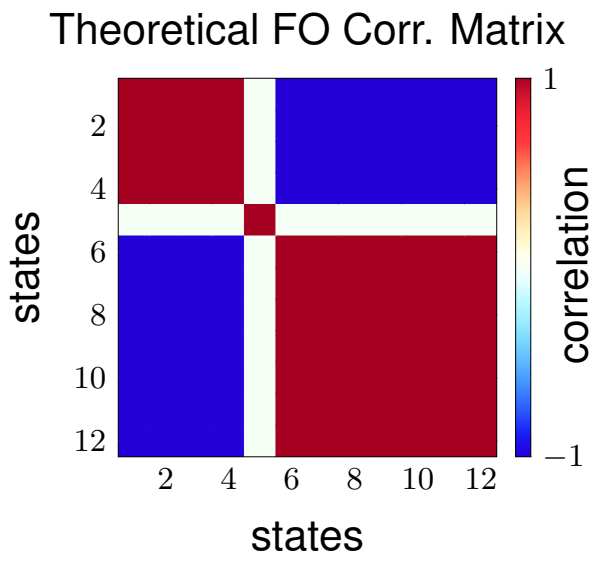

Figure S1: Ideal FO Correlation Matrix. For each model inferred from our datasets, we compute the Euclidean distance between the model's FO Correlation Matrix and the ideal one. We rank our models based on this distance with the aim of selecting the model that has the most clear emergence of the 2-metastate structure.

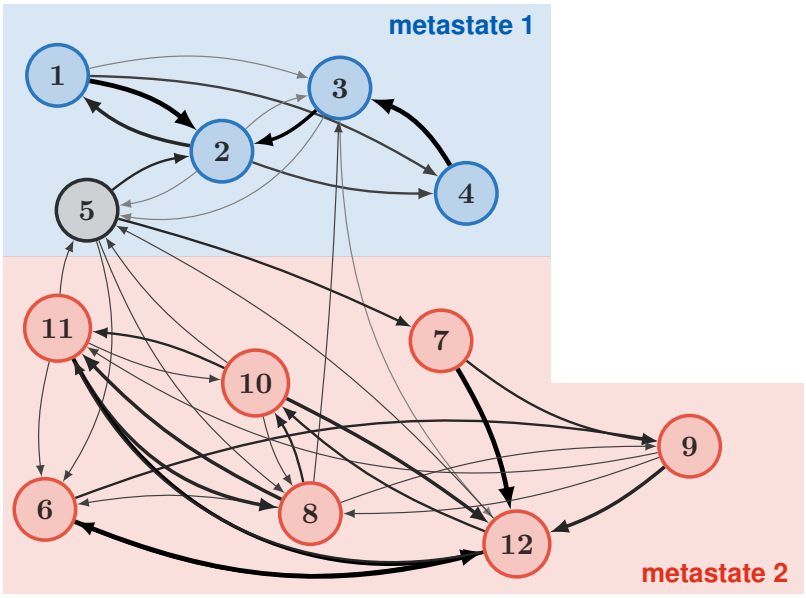

Figure S2: Network associated with the Transition Probability Matrix of the HMM used in this study (FIg. S1A). The first 4 states comprise metastate 1, whereas states 6 to 11 comprise metastate 2. State 5 is mostly uncorrelated to the other states, is associated with head motion, and has the highest variance [3]. The interconnections depicted in this graph represent probabilities higher than 10% (i.e. > 0.1) in the HMM's Transition Probability Matrix, and their color and thickness are proportional to their magnitude.

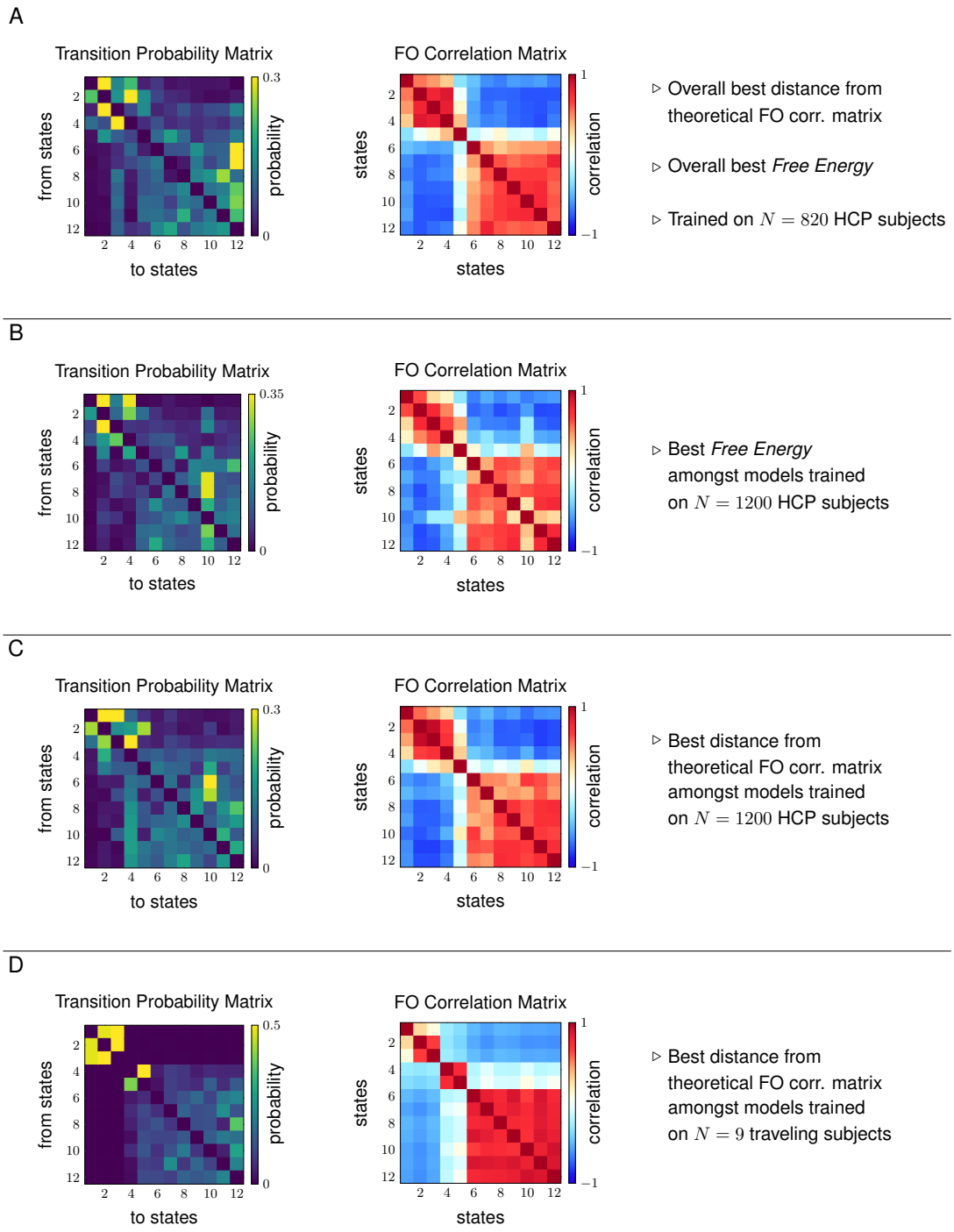

Figure S3: Transition Probability Matrix and Fractional Occupancy Correlation Matrix for different HMM models. (A) The best model among all the models trained in terms of free energy and Euclidean distance from the ideal FO Correlation Matrix. (B) The best model in terms of free energy among all models trained on the subjects of the HCP 1200-subject distribution. (C) The best model in terms of Euclidean distance from the ideal FO Correlation Matrix among all models trained on the subjects of the HCP 1200-subject distribution. (D) The best model in terms of Euclidean distance from the ideal FO Correlation Matrix among all models trained on the Traveling-subject dataset by using the model (A) as a prior. Notice that, while the model in this last panel displays very distinct metastate separation in the TPM matrix, such a matrix is not irreducible (i.e. there does not exist a path connecting the two groups of states), making it not suitable to represent any biological system.

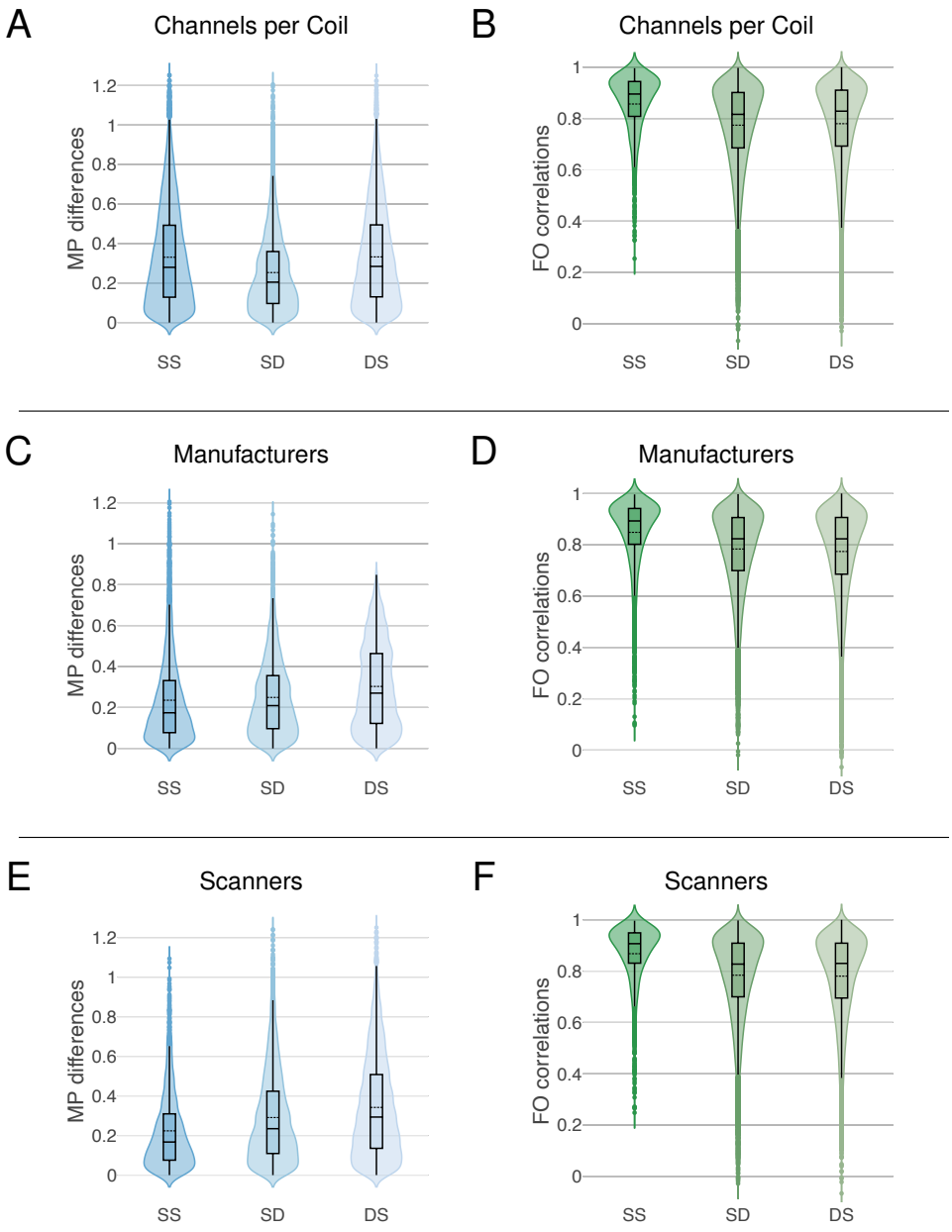

Figure S4: Panels (A) to (F) illustrate the distributions of values for MP Differences and FO Correlations, for the factors: numbers of channels per coil, manufacturers, and scanner model. The set SS comprises the MP Differences (resp., FO Correlations) computed for each subject within the same factor attribute, and the SS distribution displays these values for all subjects; the set SD consists of the MP Differences (resp., FO Correlations) computed for each subject across different attributes of the same factor, and the SD distribution displays these values for all subjects; finally, the set DS consists of the MP Differences (resp., FO Correlations) computed across all subjects within the same factor attribute, and the DS distribution displays these values for all attributes of the same factor. Further, for all the distributions, the black dashed lines illustrate the mean. The difference between SS and DS distributions in panel (A), and the difference between SD and DS distributions in panel (F), are not statistically significant (see also Table 1 in the main text).

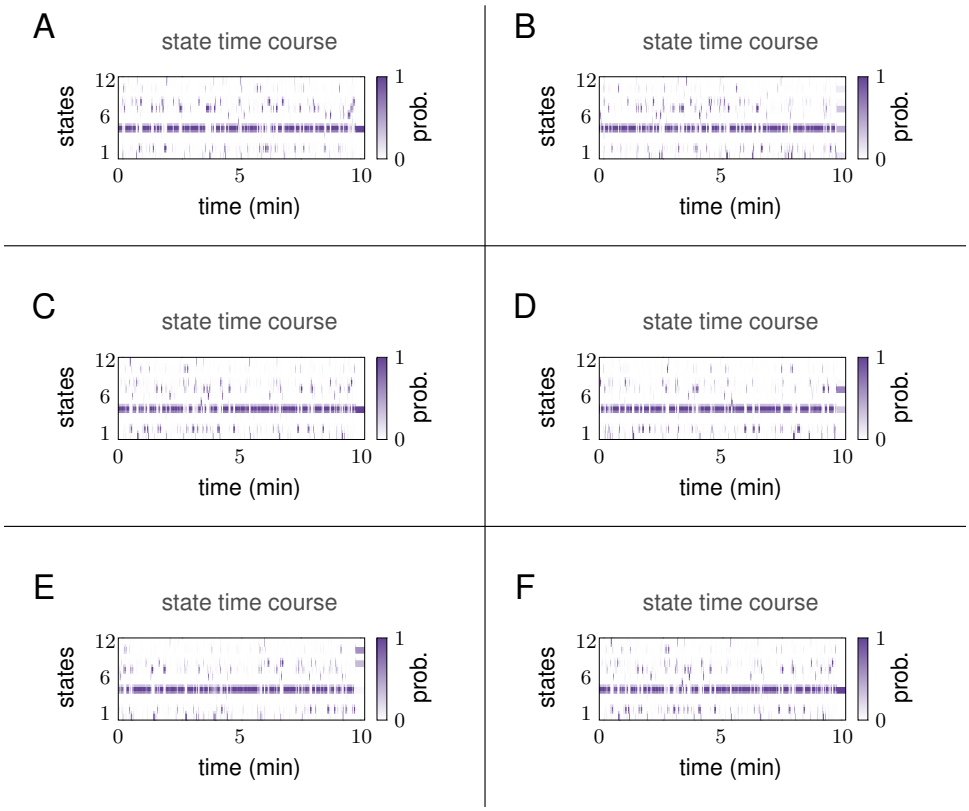

Figure S5: Examples of state time courses after HMM decoding on randomized time series. To provide a baseline for our study, we applied the HMM model used in this work to random permutations of the 50 independent components for each run in the Traveling-subject dataset (see SI Methods). The HMM decoding yields state time courses that stay most of the time in state 5, which is the state that is highly uncorrelated from the other 11 states and the one with the highest variance. This fact supports the goodness of fit of the inferred model, as randomized time series do not provide meaningful state time courses. In this figure, we show 6 casually chosen state time courses after random permutation of the time series in the Traveling-subject dataset. (A) subject 8 run 4. (B) subject 4 run 2. (C) subject 3 run 35. (D) subject 2 run 1. (E) subject 9 run 15. (F) subject 5 run 48.

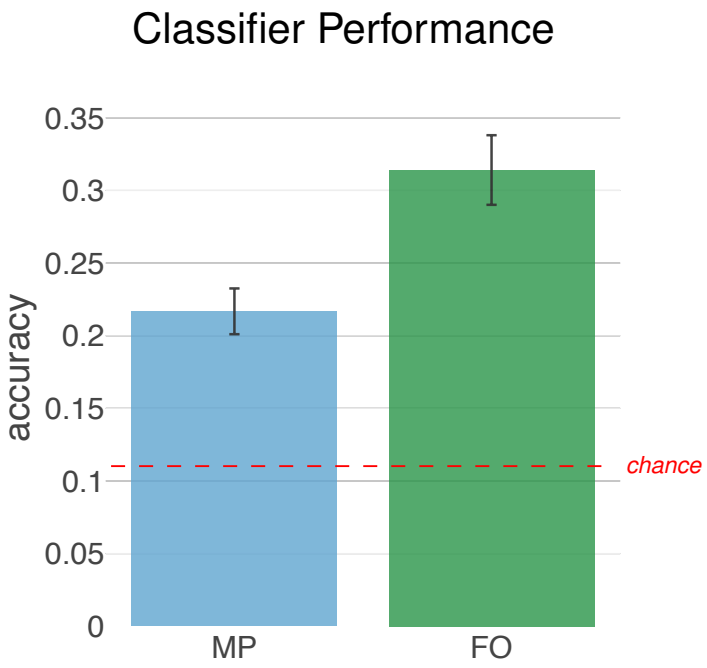

Figure S6: Summary of the leave-one-attribute-out cross-validation for all scanning factors. The two bars represent the average classification accuracy amongst all different scanning factors, along with the standard deviation, for the two fingerprints. The red dashed line indicates the baseline chance level. These results show that the personal signature of brain dynamics fingerprints emerges even with a simple logistic regression classifier.

Table S1: Imaging protocols for resting-state fMRI in the Traveling-subject dataset.

Site	ATT	ATV	COI	HUH	HKH	KPM	SWA	KUT	KUS	UTO	YC1	YC2
1. Scanner Manufacturer	Siemens	Siemens	Siemens	GE	Siemens	Philips	Siemens	Siemens	Siemens	GE	Philips	Philips
2. Scanner Model	TimTrio	Verio	Verio	Signa HDxt	Spectra	Achieva	Verio	TimTrio	Skyra	MR750W	Achieva	Achieva
3. Magnetic Field Strength	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T	3.0 T
4. No. Channels per Coil	12	12	12	8	12	8	12	32	32	24	8	8
5. Field-of-view (mm)	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212	212 × 212
6. Matrix	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64	64 × 64
7. No. of Slices	40	39	40	35	35	40	40	40	40	40	40	40
8. No. of Volumes	240	240	240	240	240	240	240	240	240	240	240	240
9. In-plane Resolution (mm)	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125	3.3125× 3.3125
10. Slice Thickness (mm)	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
11. Slice Gap (mm)	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
12. TR (ms)	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
13. TE (mm)	30	30	30	30	30	30	30	30	30	30	30	30
14. Total Scan Time (min:s)	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00	10 : 00
15. Flip Angle (deg)	80	80	80	80	80	80	80	80	80	80	80	80
16. Slice Acquisition Order	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending
17. Phase Encoding	PA	PA	AP	PA	PA	AP	PA	PA	AP	PA	AP	AP
18. Eyes Closed/Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate	Fixate
19. Field Map	yes	yes	yes	-	-	yes	yes	yes	yes	yes	yes	-

Table S2: Kolmogorov-Smirnov test statistics for MP Differences and FO Correlations

Parameter	MP Differences			FO Correlations		
	SS vs SD	SS vs DS	SD vs DS	SS vs SD	SS vs DS	SD vs DS
1. Site	0.2808	0.1756	0.1552	0.5535	0.4549	0.1416
2. Day	0.4438	0.2183	0.2468	0.4708	0.2424	0.2745
3. Phase	0.0428	0.1904	0.154	0.044	0.2074	0.1687
4. Channels/Coil	0.1422	0.0078	0.1472	0.2513	0.2168	0.0361
5. Manufacturer	0.0735	0.1836	0.1419	0.2184	0.2156	0.0257
6. Scanner	0.1454	0.231	0.0929	0.2616	0.2686	0.0117

Table S3: Logistic regression accuracy results

Parameter	MP Differences	FO Correlations
1. Site	0.2096	0.3056
2. Day	0.1889	0.2917
3. Phase	0.2222	0.3119
4. Channels/Coil	0.2341	0.3289
5. Manufacturer	0.2217	0.3430
6. Scanner	0.2243	0.3220
Mean	0.2168	0.3140
Standard Deviation	0.0158	0.0239

References

[1] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[2] D. Vidaurre, R. Abeysuriya, R. Becker, A. J. Quinn, F. Alfaro-Almagro, S. M. Smith, and M. W. Woolrich. Discovering dynamic brain networks from big data in rest and task. *Neuroimage*, 180:646–656, 2018.

[3] D. Vidaurre, S. M. Smith, and M. W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017.

[4] A. Yamashita, N. Yahata, T. Itahashi, G. Lisi, T. Yamada, N. Ichikawa, M. Takamura, Y. Yoshihara, A. Kunimatsu, N. Okada, H. Yamagata, K. Matsuo, R. Hashimoto, G. Okada, Y. Sakai, J. Morimoto, J. Narumoto, Y. Shimada, K. Kasai, N. Kato, H. Takahashi, Y. Okamoto, S. C. Tanaka, M. Kawato, O. Yamashita, and H. Imamizu.

Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology*, 17(4):1–34, 04 2019.

[5] C.-G. Yan, X.-D. Wang, X.-N. Zuo, and Y.-F. Zang. Dpabi: Data processing & analysis for (resting-state) brain imaging. *Neuroinformatics*, 14(3):339–351, Jul 2016.

[6] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.