Representation Learning of Resting State fMRI with Variational Autoencoder

Jung-Hoon Kim[1,3], Yizhen Zhang[2], Kuan Han[2],
Zheyu Wen[2], Minkyu Choi[2], Zhongming Liu[1,2*]

[1]Department of Biomedical Engineering, University of Michigan
[2]Department of Electrical Engineering and Computer Science, University of Michigan
[3]Weldon School of Biomedical Engineering, Purdue University

*Correspondence
Zhongming Liu, PhD
Associate Professor
Department of Biomedical Engineering
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
Email: zmliu@umich.edu

Abstract

Resting state functional magnetic resonance imaging (rsfMRI) data exhibits complex but structured patterns. However, the underlying origins are unclear and entangled in rsfMRI data. Here we establish a variational auto-encoder, as a generative model trainable with unsupervised learning, to disentangle the unknown sources of rsfMRI activity. After being trained with large data from the Human Connectome Project, the model has learned to represent and generate patterns of cortical activity and connectivity using latent variables. The latent representation and its trajectory represent the spatiotemporal characteristics of rsfMRI activity. The latent variables reflect the principal gradients of the latent trajectory and drive activity changes in cortical networks. Latent representations are clustered by both individuals and brain states. Representational geometry captured as covariance or correlation between latent variables, rather than cortical connectivity, can be used as a more reliable feature to accurately identify subjects from a large group, even if only a short period of data is available per subjects.

INTRODUCTION

The brain is active even at rest, showing complex activity patterns measurable with resting state fMRI (rsfMRI) (Fox and Raichle, 2007). It is widely recognized that rsfMRI activity is shaped by how the brain is wired, or the brain connectome (Sporns et al., 2005). Inter-regional correlations of rsfMRI activity are often used to report functional connectivity (Biswal et al., 1995) and map brain networks for individuals (Finn et al., 2015) or populations in various behavioral (Smith et al., 2009) or disease states (Fox et al., 2014). However, it remains largely unclear where rsfMRI activity comes from (Leopold and Maier, 2012; Lu et al., 2019), whereas understanding its origins is critical to interpretation of any rsfMRI pattern or dynamics (Winder et al., 2017).

Prior findings suggest a multitude of sources (or causes) for rsfMRI activity (Bianciardi et al., 2009), including but not limited to fluctuations in neurophysiology (Mantini et al., 2007), arousal (Chang et al., 2016), unconstrained cognition (Chou et al., 2017), non-neuronal physiology (Birn et al., 2008), head motion (Power et al., 2014) etc. These sources only partially account for rsfMRI activity and may be entangled not only among themselves but also with other sources that are left out simply because they are hard to specify or probe in a task-free state (Leopold and Maier, 2012). An inclusive study would benefit from using a data-driven approach to uncover and disentangle all plausible but hidden sources from rsfMRI data itself, without having to presume the sources to whatever are experimentally observable. To be effective, such an approach should be able to infer sources from rsfMRI data and generate new rsfMRI data from sources, while being able to account for complex and nonlinear relationships between the sources and the data.

These requirements lead us to deep learning, or representation learning with deep neural networks (LeCun et al., 2015), as a nonlinear method for blind source separation, in contrast to its linear counterparts, e.g. independent component analysis (Beckmann and Smith, 2004; Calhoun et al., 2001; Smith et al., 2012). For brain research, deep learning models has provided testable models of the brain in terms of neural computation for sensory and language processing (Han et al., 2019; Kell et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014; Richards et al., 2019; Wen et al., 2018; Yamins and DiCarlo, 2016; Zhang et al., 2020). Deep learning has also been increasingly used as a generic family of machine learning tools to learn features from fMRI data. See (Khosla et al., 2019b) for a review. Most applications are in the regime of supervised learning. Typically, a neural network takes an fMRI-based input data and is trained to generate an output that optimally matches the ground truth for a task, such as individual identification (Chen and Hu, 2018; Wang et al., 2019), prediction of gender, age, or intelligence (Fan et al., 2020; Gadgil et al., 2020; Plis et al., 2014), disease classification (Seo et al., 2019; Suk et al., 2016; Wang et al., 2020; Yang et al., 2019; Zou et al., 2017). The labels required for supervised learning are often orders of magnitude smaller in size than the fMRI data itself, which has a high dimension in both space and time. As a result, the prior studies often limit the model capacity by using a shallow network and/or limit the input data to activity at the region of interest (ROI) level (Chen and Hu, 2018; Dvornek et al., 2018; Koppe et al., 2019; Matsubara et al., 2019; Suk et al., 2016; Wang et al., 2019; Wang et al., 2020) or reduce it to functional connectivity (D'Souza et al., 2019; Fan et al., 2020; Kawahara et al., 2017; Kim and Lee, 2016; Riaz et al., 2020; Seo et al., 2019; Venkatesh et al., 2019; Yang et al., 2019; Zhao et al., 2018). It is also uncertain to what extent representations

learned for a specific task would be generalizable to other tasks. It is further debatable whether deep neural networks with supervised learning are currently superior to more conventional and simpler methods (He et al., 2020).

For these considerations, unsupervised learning is more preferable for uncovering the underlying causes that drive intrinsic brain activity regardless of any task or disease. We choose to use the Variational Auto-Encoder (VAE) (Higgins et al., 2017; Kingma and Welling, 2013), for unsupervised learning of the increasing “big data” in rsfMRI without requiring any label or narrowly focusing on any downstream task. Unlike auto-encoder, VAE is a generative model capable of synthesizing new data similar to the training data, and it regularizes the latent space with *a priori* spherical Gaussian distributions. These properties allow the representation learned to be expressed in terms of latent variables that encode the disentangled causes of the data. Our emphasis on disentangling latent representations sets this work apart from several prior work based on the auto-encoder implemented in various forms of deep neural networks (Cui et al., 2019; Huang et al., 2017; Liu et al., 2020a; Makkie et al., 2019; Suk et al., 2016; Zhao et al., 2018). Briefly in this study, we designed and trained a VAE model to represent rsfMRI data in terms of its latent sources and tested its ability to explain and generate rsfMRI data. We characterized the time evolving trajectory of latent representation and factorized its gradients by principal components. We also visualized the representational gradients, clusters, and geometries within and across individuals, as a way to characterize brain networks and their dynamic interactions. Lastly, we tested the use of this model for characterizing individual variations and identifying individuals from their rsfMRI data (Finn et al., 2015), as a starting example of its applications.

Methods

Data

We used rsfMRI data from 650 healthy subjects randomly chosen from the Q2 release by HCP (Van Essen et al., 2013). For each subject, we used two sessions of rsfMRI data acquired from different days with either the right-to-left or left-to-right phase encoding. Each session included 1,200 time points separated by 0.72s. Following minimal preprocessing (Glasser et al., 2013) and automatic denoising with ICA (or the ICA-FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014), we applied voxel-wise detrending (regressing out a 3rd-order polynomial function), bandpass filtering (from 0.01 to 0.1 Hz), and normalization (to zero mean and unitary variance). We further separated the data into three sets, including 100, 50, or 500 subjects for training, validating, or testing the VAE model, respectively. The validation dataset was used to determine the hyper-parameters used in the VAE model. The testing data were neither seen nor used by the model during training or validation. This held-out data was used to test the generalizability of the model across different datasets. For an exploratory analysis, we additionally tested the model with rsfMRI data that did not go through denoising with ICA-FIX to evaluate the model performance against presumably noisier rsfMRI data.

Geometric reformatting

We converted the rsfMRI data from 3-D cortical surfaces to 2-D grids in order to structure the rsfMRI pattern as an image to ease the application of convolutional neural networks. As illustrated in Figure 1.a, we inflated each hemisphere to a sphere by using

FreeSurfer (Fischl, 2012). For each location on the spherical surface, we used cart2sph.m in MATLAB to convert its cartesian coordinates (x, y, z) to spherical coordinates (a, e), which reported the azimuth and elevation angles in a range from $-\pi$ to π and from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, respectively. We defined a 192×192 grid to resample the spherical surface with respect to azimuth and *sin*(elevation) such that the resampled locations were uniformly distributed at approximation (Supplementary Figure 1). We used the nearest-neighbor interpolation to convert data from the 3-D surface to the 2-D grid, and vice versa.

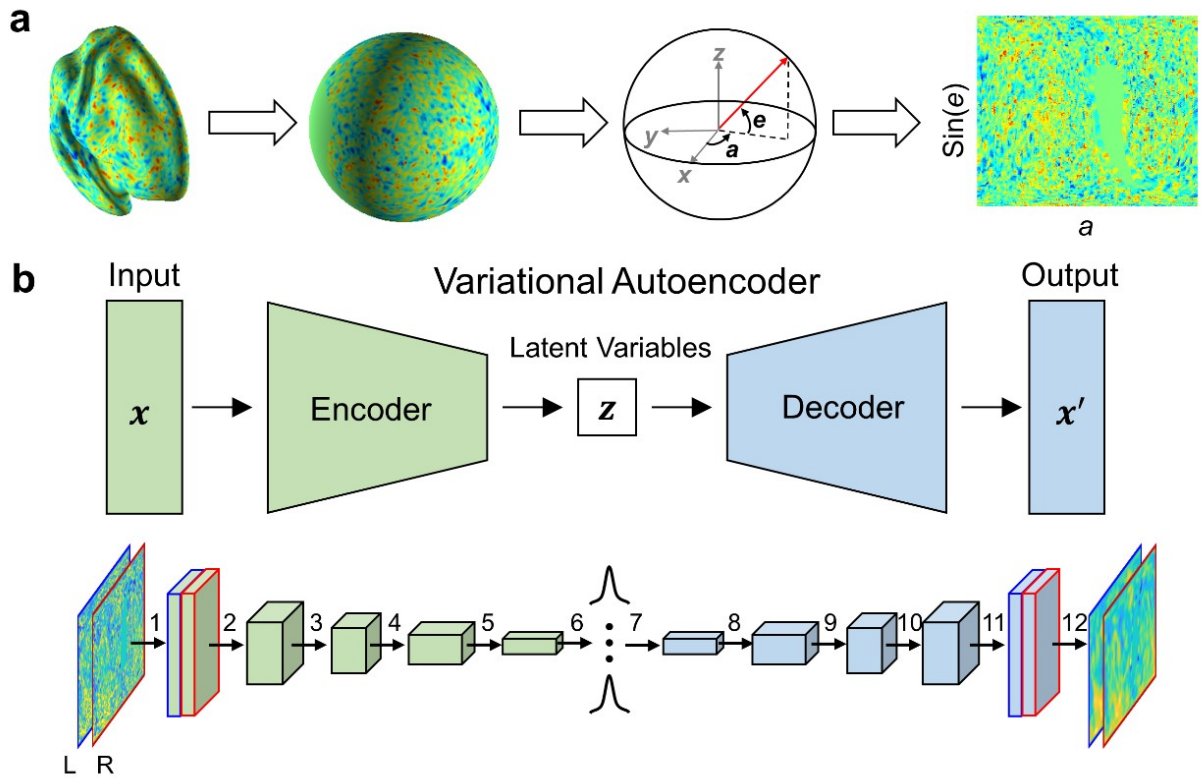


Figure 1. Variational Auto-Encoder (VAE). (a) Geometric reformatting. The cortical distribution of fMRI activity is converted onto a spherical surface and then to an image by evenly resampling the spherical surface with respect to sin(e) and a, where e and a are elevation and azimuth, respectively. **(b) Encoder-decoder architecture.** The encoder and the decoder each contain 5 convolutional layers connected in series. In the encoder, each layer (numbered from 1 to 5) outputs a feature map with the size of 96x96x64, 48x48x128, 24x24x128, 12x12x256, or 6x6x256, respectively. In layer 1, 32 kernels are applied to 192x192 flattened images of each hemisphere separately, and output feature maps are concatenated along the

kernel dimension, resulting in a feature map with 64 channels. In the decoder, each layer (numbered from 8 to 12) outputs a feature map with a size of 6x6x256, 12x12x256, 24x24x128, 48x48x128, or 96x96x64, respectively. The operation at each layer is specified as follows. 1: convolution (kernel size=8, stride=2, padding=3) and rectified nonlinearity; 2-5: convolution (kernel size=4, stride=2, padding=1) and rectified nonlinearity; 6: fully connected layer and re-parameterization; 7: fully connected layer and rectified nonlinearity; 8-11: transposed convolution (kernel size=4, stride=2, padding=1) and rectified nonlinearity; 12: transposed convolution (kernel size=8, stride=2, padding=3). The blue and red boundaries highlight the input and output images for the left and right hemispheres, respectively.

Variational autoencoder

We designed a β-VAE model (Higgins et al., 2017), a variation of VAE (Kingma and Welling, 2013), to learn representations of rsfMRI spatial patterns. This model included an encoder and a decoder (Figure 1.b). The encoder converted an fMRI map to a probabilistic distribution of 256 latent variables. Each latent variable was a Gaussian random variable with a mean and a standard deviation. The decoder sampled the latent distribution to reconstruct the input fMRI map or generate a new map, which appeared similar to what would be observable with fMRI. The encoder stacked five convolutional layers and one fully connected layer. Every convolutional layer applied linear convolution and rectified its output (Nair and Hinton, 2010). The 1st layer applied 8×8 convolution separately to the input from each hemisphere and concatenated its output. To the feature maps concatenated across both hemispheres, the 2nd through 5th layers applied 4×4 convolution. Since a spherical pattern is circularly continuous with respect to the azimuth, we applied circular padding to the boundaries of azimuth for the flattened 2-D map but applied zero padding to the boundaries of elevation. Such padding was intended to avoid artifacts when otherwise applying convolution near those boundaries. The fully connected layer applied linear weighting and yielded the mean and standard deviation that described

the normal distribution of each latent variable. The decoder used nearly the same architecture as the encoder but connected the layers in the reverse order for transformation from the latent space back to the input space. Figure 1.b illustrates the model architecture.

We trained the VAE model to reconstruct input while constraining the distribution of every latent variable to be close to an independent and standard normal distribution. Specifically, using the training data, we optimized the encoding parameters, ϕ, and the decoding parameters, θ, to minimize the loss function as below.

$$L(x) = \|x - x'\|_2^2 + \beta \cdot D_{KL}[N(\mu_z, \sigma_z) \parallel N(0, I)] \qquad (1)$$

where x is the input data combined across the left and right hemispheres, x' is the corresponding output from the model, $N(\mu_z, \sigma_z)$ is the posterior normal distribution of the latent variables, z, with their mean and standard deviation denoted as μ_z and σ_z, $N(0, I)$ is an independent and standard normal distribution as the prior distribution of the latent variables, D_{KL} measures the Kullback-Leibler (K-L) divergence between the posterior and prior distributions, and β is a hyperparameter balancing the two terms in the loss function. Part of the medial cortical surface that corresponds to corpus callosum (i.e. white matter) was excluded from training such that the learned model was intended to merely represent the activity of cortical gray matter. To train the model, we used $\beta = 10$ and stochastic gradient descent (batch size=128, initial learning rate=10^{-4}, and 100 epochs) and Adam optimizer (Kingma and Ba, 2014) implemented in PyTorch (v*1.2.0*). The learning rate decayed by a factor of 10 every 20 epochs.

We determined the hyperparameters by exploring and testing different parameter settings with the validation dataset. Specifically, we explored four values (1, 5, 10, 15) for

β and chose $\beta = 10$ to balance the reconstruction performance vs. the disentanglement (or independence) of latent variables (Figure 2) – the two terms in the loss function shown in Eq. (1). We also explored several options for the number of layers and the learning rate, and finalized those parameters based on the loss evaluated with the validation dataset (Supplementary Figure 2 and 3). Note that with the training strategies described above, only VAE model with 12 layers were able to reduce both reconstruction loss and D_{KL} when $\beta = 10$.

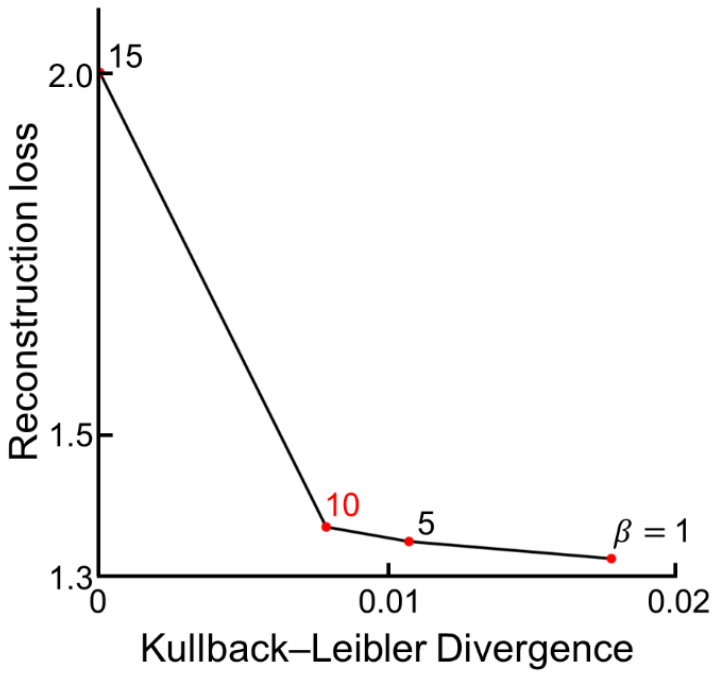


Figure 2. Validation losses for the VAE models with different values for β in Eq. (1). $\beta = 10$ shows a reasonable trade-off between the reconstruction loss and the Kullback–Leibler divergence on the validation dataset.

Synthesizing rsfMRI functional connectivity

We used the trained VAE to synthesize rsfMRI data from random samples of latent

206 variables. To synthesize a vector in the latent space, we drew a random sample of every
207 latent variable independently from a standard normal distribution. The synthesized vector
208 passed through the decoder in VAE, generating a cortical pattern. Repeating this process,
209 we synthesized 12,000 cortical patterns as data used for seed-based correlation analysis.
210 As examples, we explored three seed locations within V1, IPS, and PCC and calculated
211 the functional connectivity to each seed based on the Pearson correlation coefficient. The
212 MNI coordinates of the seed in V1, IPS, and PCC were (7, -83, 2), (26, -66, 48), and (0,
213 57, 27), respectively (Jarrett, 2009). In addition, we performed a similar analysis without
214 limiting to the seed locations. Instead, we calculated the functional connectivity between
215 each pair of parcels as defined in a 360-parcel atlas of the whole cortex (Glasser et al.,
216 2016).

217 For comparison, we similarly calculated seed-based or parcel-to-parcel functional
218 connectivity with experimental rsfMRI data concatenated across a varying number (1, 5,
219 10, 50, and 100) of subjects in HCP. We compared the functional connectivity pattern
220 observed with synthesized and experimental data and repeated the comparison 20 times.
221 At each time, we generated a different set of synthesized data while using experimental
222 data from a different subset of subjects. The comparison was thus randomly repeated.
223

224 **Defining a principal basis set in the latent space**

225 By our design, the VAE model encodes the spatial pattern of fMRI activity and does
226 not represent the temporal dynamics explicitly. The distribution of every latent variable is
227 constrained to be close to a standard normal distribution independent of one another, for
228 the K-L divergence term in the loss function in Eq. (1). This implies that the latent variables

in the VAE model are not unique. An arbitrary rotation of a tentative set of latent variables would arrive at a new set of latent variables that span the same latent space and satisfy the same learning objective.

To identify a unique set of latent variables, we exploited the trajectory of the latent representation. Specifically, for the fMRI data in the testing set (concatenated across 500 subjects), we encoded the fMRI pattern observed at every time into a point (or vector) embedded in the latent space. As time progressed, this point moved in the latent space along a trajectory that represented the temporal dynamics of fMRI activity.

In a first-order differential analysis, we evaluated the displacement (or difference) of the latent representation from every time point to its next. To this time-difference vector (or the latent gradient), we further applied singular vector decomposition and used the singular vectors to define a unique basis set of the latent space. As such, each singular vector was a re-defined latent variable, while the corresponding singular value indicated its importance in explaining the latent gradient of cortical activity. In other words, the trajectory was more likely to move along the direction represented by a singular vector with a larger singular value than that with a smaller singular value.

We further interpreted and visualized the top-10 latent variables defined as the singular vectors with the largest 10 singular values. For this purpose, we decoded each of these latent variables onto the cortical surface by using the decoder in the VAE model. Note that the latent variables were related to cortical patterns through nonlinear functions. We evenly sampled each latent variable of interest from -5 to 5, while keeping other latent variables to zero. We mapped the decoded cortical pattern and characterized its variation due to the variation of a single latent variable. We quantified the variation separately for

each cortical location in terms of the standard deviation multiplied by a sign. The sign of standard deviation map was determined by measuring the sign of Pearson correlation coefficient between the decoded values of each cortical location and the samples of the given latent variable.

Clustering in the latent space

We encoded the rsfMRI spatial pattern at every time point for 500 testing subjects, yielding 600,000 vectors in the latent space. We used k-means clustering with 1-cosine distance (based on “kmeans” in Matlab) to group those vectors to 21 clusters. The choice of k=21 was empirical but made intentionally to be consistent to a prior study with a similar motivation (Smith et al., 2012). This choice was within a reasonable range for the number of resting state networks as reported in literature (Smith et al., 2009; Yeo et al., 2011). Beyond this single choice, we explored other numbers of clusters to ensure that k=21 was a reasonable choice for the distribution of latent representation. Specifically, we varied k from 1 to 100. Given each choice, we ran the k-means clustering for the testing data (n=500 subjects), identified the clusters, calculated the centroid of each cluster, and summed the distance to the centroid within every cluster. We further plotted the sum of distance as a function of k and ensured that k=21 was around the “elbow” of the plot, as a useful, but not strict, rule of thumb.

Given k-means clustering with k=21, we re-ordered individual time points by their cluster membership and compared the distance between time points within and between clusters. We also evaluated the relationships among different clusters by calculating the distance between the centroids of individual clusters and we further grouped clusters into

275 super-clusters organized in a multi-level hierarchy visualized as a dendrogram (by using
276 “linkage” in Matlab).

277 To visualize and interpret each cluster, we further converted the cluster centroid to
278 a corresponding cortical pattern by using the VAE’s decoder. The resulting cortical pattern
279 was scaled such that its maximal absolute value equaled 1. This pattern was considered
280 as a functional cortical network. To further evaluated how each cortical network changed
281 its activity in time, we defined and evaluated the cluster-wise activity as the cosine affinity
282 between the centroid of each cluster and the latent representation of fMRI activity at every
283 time. As such, a cluster increased its activity when the latent representation moved toward
284 the centroid of that cluster or decreased its activity when the representation moved away
285 from that centroid but towards the centroid of another cluster. After this analysis was done
286 separately for each session and subject, we averaged the cluster-wise activity across
287 subjects. Then we compared the group-level activity between sessions and across
288 clusters, and tested the statistical significance with a non-parametric permutation test
289 (false discovery rate $q<0.01$), for which the time points were randomly shuffled for 10,000
290 trials to yield a null distribution.

291

292 **Individual variation**

293 To evaluate the individual variation, we compared the latent representations of the
294 fMRI data from different individuals. In an exploratory analysis, we randomly selected a
295 small (n=20) subset of subjects. We chose 20 subjects to ease visualization and intuitive
296 demonstration, before scaling up the analysis to 500 subjects. For each of the 20 subjects,
297 we converted the fMRI activities, instance by instance, to the representations in the latent

space. To visualize and compare subject-wise representations, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) method to visualize the 256-dimensional latent representations (color-coded by subjects) in a two-dimensional space. We calculated the Silhouette index to measure how similar the latent representation was within the same subject vs. between different subjects.

Individual identification

After the exploratory analysis above, we evaluated the individual variation across n=500 subjects. For the distribution of subject-wise latent representation, the first moment was the mean and the second moment was the covariance, which indicated the location and geometry of the subject-wise latent representation, respectively. We tested the use of the first moment (mean) or the second moment (covariance) as the subject-identifying feature.

In the testing data set, every individual had rsfMRI data acquired for two separate sessions. From the first session, we extracted the feature from every subject and stored it as the subject-identifying key in a database that included a population of 500 subjects. Given this database, we tested the accuracy of retrieving any subject's identity by using the feature extracted from the second session as a query to match against all keys in the database. The goodness of match was evaluated as the cosine similarity or the Pearson correlation coefficient when the query and the key were based on the first moment (mean) or the second moment (covariance) of the subject-wise representation, respectively. The accuracy of individual identification was evaluated as the percentage by which the correct identity was retrieved as one of the best 1, 5, or 10 matches, yielding the namely top-1,

5, or 10 accuracy.

For comparison, we compared the performance of individual identification based on the above latent-space feature vs. the similar feature evaluated in the cortical space. The cortical-space features extracted with a similar method as previously reported in (Finn et al., 2015). Specifically, the FC between brain regions (or connectome) was calculated as features for individual identification. It is worth noting that the cortical connectome and covariance of latent representation, although they are nominally different terms, can both be viewed as the representational geometry of brain activity in the cortical space (for the connectome) or the latent space (for the covariance of latent representation). In addition, we may also cast both notions as the functional connectivity profile in the cortical space or the latent space. Given such conceptual connections, we evaluated the FC between every pair of 360 cortical parcels defined in an established atlas (Glasser et al., 2016) and used the FC-based connectome as the feature for individual identification (Finn et al., 2015). We compared the connectome-based identification accuracy with that based on the FC profile (or representational geometry) in the latent space for a varying population size (from n=5 to 500 subjects) or a varying length of data per subject (from 9 to 180 s). We repeated the above analysis 100 times, each time with a different subset of the testing data and averaged the identification accuracy across the repeated tests.

Comparison with linear latent space

The VAE model described herein provided nonlinear mapping from the cortical space to the latent space (through the encoder) and in reverse (through the decoder).

Such reversible mapping could be conventionally done through linear matrix operations, such as the principal component analysis (PCA) and independent component analysis (ICA). Hence, we compared the distribution and geometry of the rsfMRI representation in the nonlinear latent space obtained with VAE vs. the linear latent space obtained with PCA or ICA. For such comparison, we used PCA or ICA trained with the training data to represent rsfMRI data in the testing dataset, while keeping the linear latent space of the same (256) dimension as its nonlinear counterpart. We compared the performance of reconstructing fMRI patterns from their latent representations (see results in Fig. 4). In addition, we also compared PCA or ICA vs. VAE for characterizing individual variation or performing individual identification by using the representation in the PCA or ICA-derived linear latent space for the same analyses as used for the representation in the VAE-based nonlinear latent space (see results in Figs. 8 and 10).

Results

VAE compressed rsfMRI maps

Inspired by its success in artificial intelligence (Higgins et al., 2017; Kingma and Welling, 2013), we designed a VAE model in order to disentangle the generative factors underlying rsfMRI activity. The model was trained to represent and reconstruct rsfMRI data with a set of latent variables that were constrained to be as independent as possible. The hyper-parameter, β, which expressed the weighting of independence among latent variables in the overall learning objective, was initially explored for different values (1, 5, 10, 15) before being finalized to β=10 – a setting that led to a reasonable trade-off of the model performance vs. constraint as demonstrated with the validation dataset (Fig. 2).

The model used a pair of convolutional and deconvolutional neural networks in an encoder-decoder architecture (Figure 1.b). The encoder transformed any rsfMRI pattern, formatted as an image on a regular 2D grid (Figure 1.a), to the probability distributions of 256 latent variables. The decoder used samples of the latent variables to reconstruct or generate an fMRI map. Using data from HCP (WU-Minn HCP Quarter 2) (Van Essen et al., 2013), we first trained the model with rsfMRI maps from 100 subjects and then tested it with rsfMRI data from 500 other subjects.

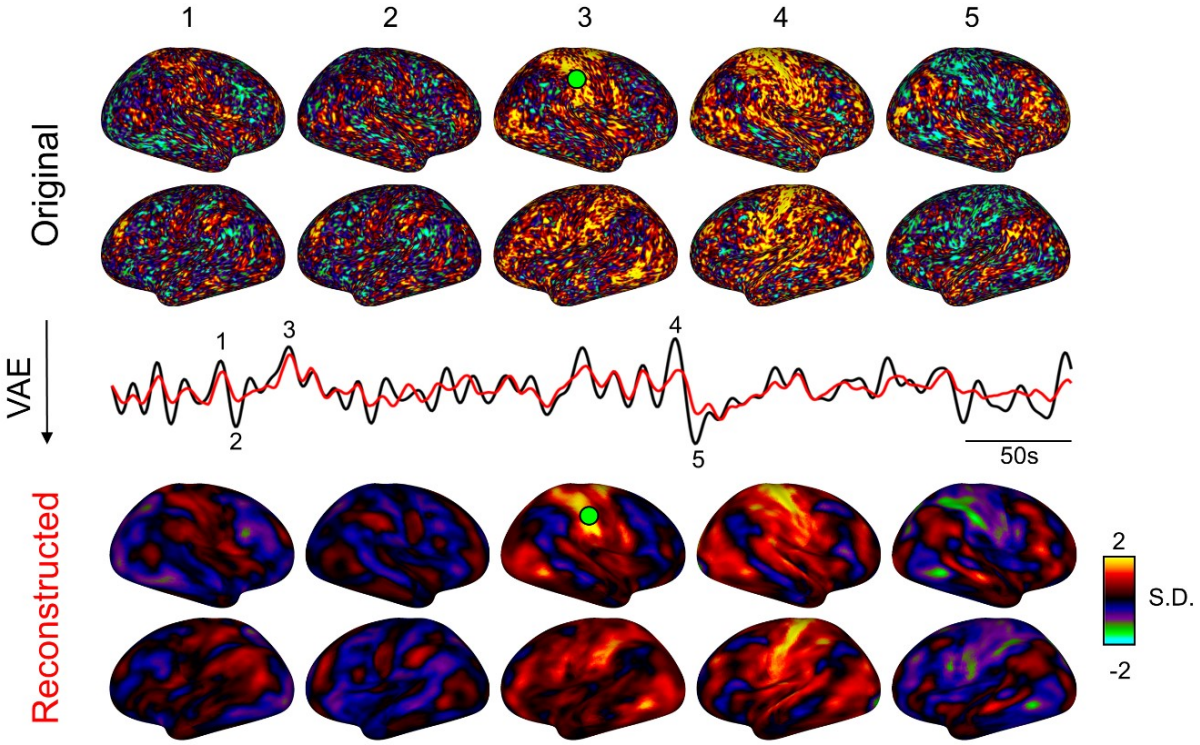


Figure 3. Image reconstruction using VAE. A series of cortical patterns are reconstructed through the VAE model. Among them, five original cortical patterns (upper panel) and their corresponding reconstruction through VAE (bottom panel) are visualized for comparison. For an example region (green circle), the time series of the original activity (black line) and the reconstructed activity (red line) are plotted for comparison.

After being trained, the model could compress any fMRI map to a low-dimensional latent space and restore the map from the latent representation separately for every time point

(Figure 3). The compression resulted in spatial blurring comparable to the effect of spatial smoothing with 4-6 mm full width at half maximum (FWHM) (Figure 4). Given fMRI data spatially smoothed to a varying extent (FWHM from 1 to 10 mm), VAE showed either comparable or better performance of reconstruction than its linear counterparts (PCA and ICA), when VAE, PCA, and ICA all used the same dimension (256) for their latent spaces (Figure 4.a). The difference in reconstruction performance between VAE and PCA or ICA was marginal but statistically significant (repeated measures ANOVA followed by post-hoc paired t-test, false discovery rate $q<0.05$), for all smoothing levels except FWHM=1 mm (Figure 4.b). These results suggest that the latent representation obtained with VAE preserved the spatial and temporal characteristics of rsfMRI, despite a modest but acceptable loss in spatial resolution and specificity.

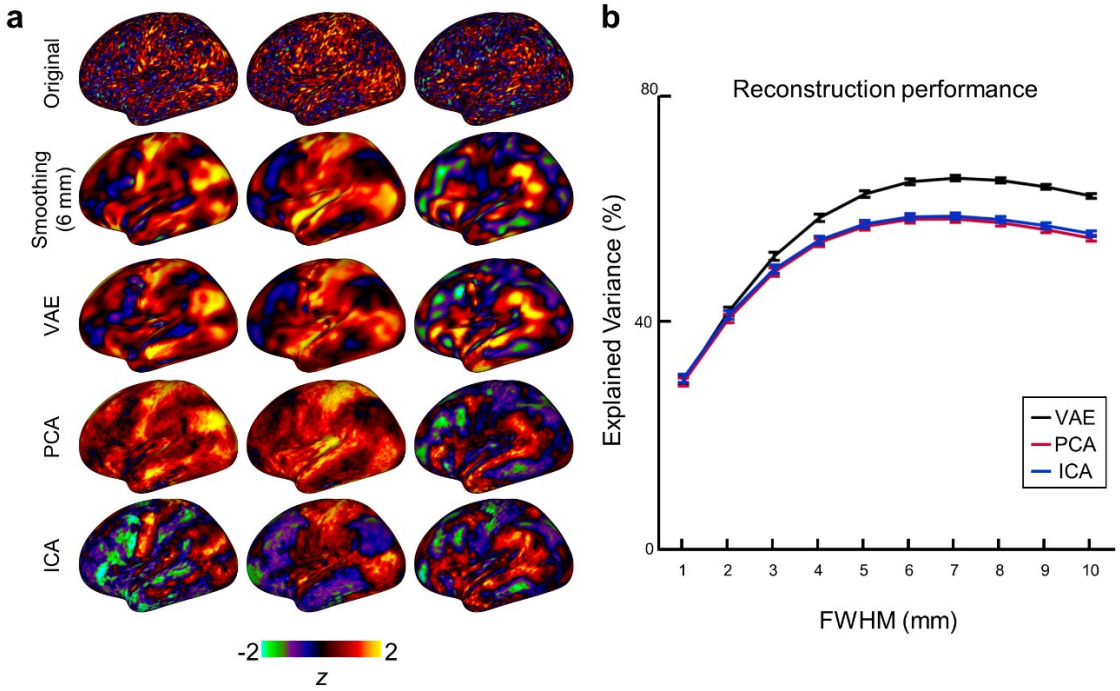


Figure 4. **RsfMRI data compression and reconstruction with VAE vs. PCA and ICA**. (**a**) For illustration, three example maps of fMRI activity, before (1st row) and after (2nd row) being smoothed (FWHM=6mm), are shown in comparison with the corresponding maps reconstructed with VAE (3rd row), PCA (4th row), and

ICA (5th row) trained to compress and reconstruct the training data from 100 subjects with 256 variables or components. (**b**) For quantitative comparison, the reconstruction performance, in terms of the percentage of variance in the fMRI images as explained by the model reconstruction, is shown for VAE, PCA, and ICA as a function of FWHM (from 1 to 10 mm) applied to the spatial smoothing of the fMRI images. The error bar stands for the standard error of mean.

VAE synthesized correlated fMRI activity

We asked whether the decoder in the VAE, as a generative model of fMRI activity, had learned the putative mechanisms by which rsfMRI activity patterns arise from brain networks. To address this question, we randomly sampled every latent variable from a standard normal distribution and used the decoder to synthesize 12,000 rsfMRI maps (equivalent to 10 subjects at 1,200 time points per subject).

We calculated the seed-based correlations by using the VAE-synthesized data and compared the resulting maps of correlations with those obtained with rsfMRI data concatenated across a different number of subjects. Figure 5.a shows three examples with the seed region in the primary visual cortex (V1), intraparietal sulcus (IPS), or posterior cingulate cortex (PCC). For each of the three seed locations, the synthesized fMRI data showed a similar correlational map as that based on length-matched rsfMRI data obtained from 10 subjects (Figure 5.a), and the correlational map was consistent with the literature (Yeo et al., 2011). The measured FC patterns were more similar to the synthesized FC patterns, when the measured FC was based on data from increasingly more subjects, regardless of whether the FC was evaluated and compared with respect to a specific seed location (Figure 5.b) or across all cortical parcels (Figure 5.c). These results suggest that the VAE provided a computational account for the generative process of resting state activity and could synthesize realistic rsfMRI activity patterns and preserve

inter-regional correlations as are experimentally observable at a group or population level. However, it is worth mentioning that the temporal ordering of the synthesized data is not meaningful, since the VAE model does not explicitly model the temporal dynamics.

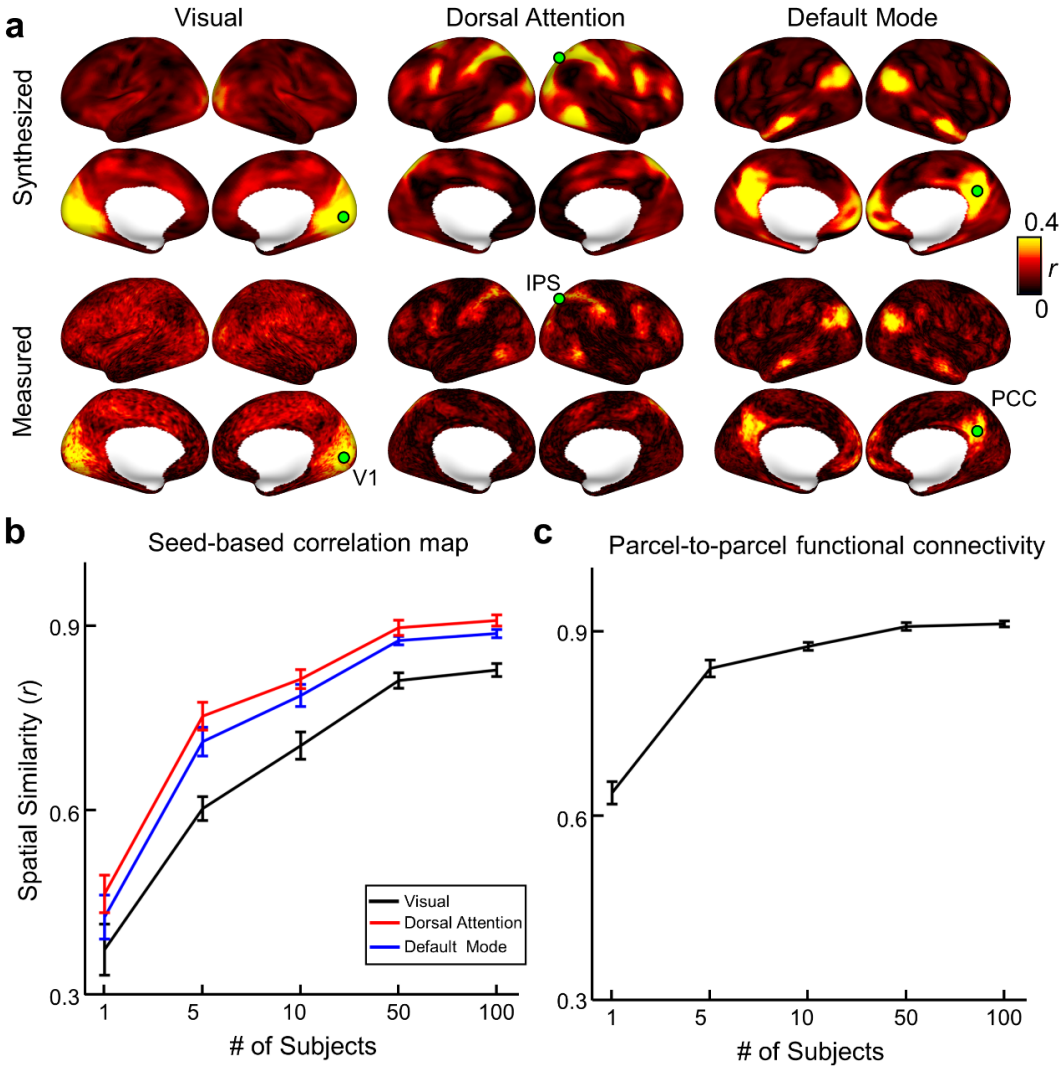


Figure 5. VAE synthesizes correlated fMRI activity. (**a**) Seed-based correlations of VAE-synthesized fMRI data (top row) vs. experimental fMRI data (bottom row) with the seed location (green circle) at V1 (left), IPS (middle), or PCC (right). (**b**) Spatial correlations between the seed-based functional connectivity based on VAE-synthesized data and those based on measured fMRI data concatenated across 1, 5, 10, 50, or 100 subjects. The colors indicate different seed locations (V1: black; IPS: red ; PCC: blue). Similarly, (**c**) shows the spatial correlation between the synthesized vs. measured functional connectivity among 360 cortical parcels. The error bar indicates the standard error of the mean averaged across 20 repeated trials.

434 **Latent variables reflected network dynamics**

435 We also examined the time-evolving trajectory of the latent representation and re-
436 defined the latent variables such that they reflected the dynamic changes of fMRI activity.
437 We first evaluated the displacement of the latent representation from every time point to
438 its next. Then we applied singular value decomposition and used the resulting singular
439 vectors to redefine the latent variables as a new basis set that spanned the latent space.
440 These redefined latent variables, ranked in a descending order by their singular values,
441 represented the principal directions in which the instantaneous latent representation
442 tended to move along its time-evolving trajectory.

443 We chose the top-10 latent variables for further visualization and interpretation.
444 For each latent variable, we uniformly sampled its value in a range from -5 to 5 and
445 visualized each sample by decoding it to a cortical pattern. We found that as the latent
446 variable increased its value linearly, the decoded cortical pattern changed in a non-linear
447 way that differed across cortical locations (see illustrative examples in Figure 6.b). To
448 visualize how each latent variable controlled the activity at each cortical location, we
449 calculated the standard deviation of the voxel-wise activity change given an increasing
450 value for the given latent variable and multiplied by a sign (+1 or -1) depending on whether
451 the activity tended to increase or decrease as the latent variable increased. For example,
452 the 1st latent variable was visualized as a cortical pattern that resembled the default mode
453 network (Figure 6.a). Using the same visualization method, we found that the 2nd through
454 10th latent variables all corresponded to distinct but partially overlapping cortical patterns
455 (Figure 6.d). However, the top-10 latent variables were found to be inadequate to explain
456 the dynamics of latent representations. The percentage of the variance explained by each

latent variable was around 1% or less, and the total variance collectively explained by top 10, 20, 50, and 100 latent variables were 9.6, 17.5, 37.6, and 62.3% (Figure 6.c). These results suggest that the dynamics of rsfMRI is complex and high-dimensional in nature. Nevertheless, the latent variables derived from the above analysis represent distinctive factors that drive the dynamic change in resting state activity.

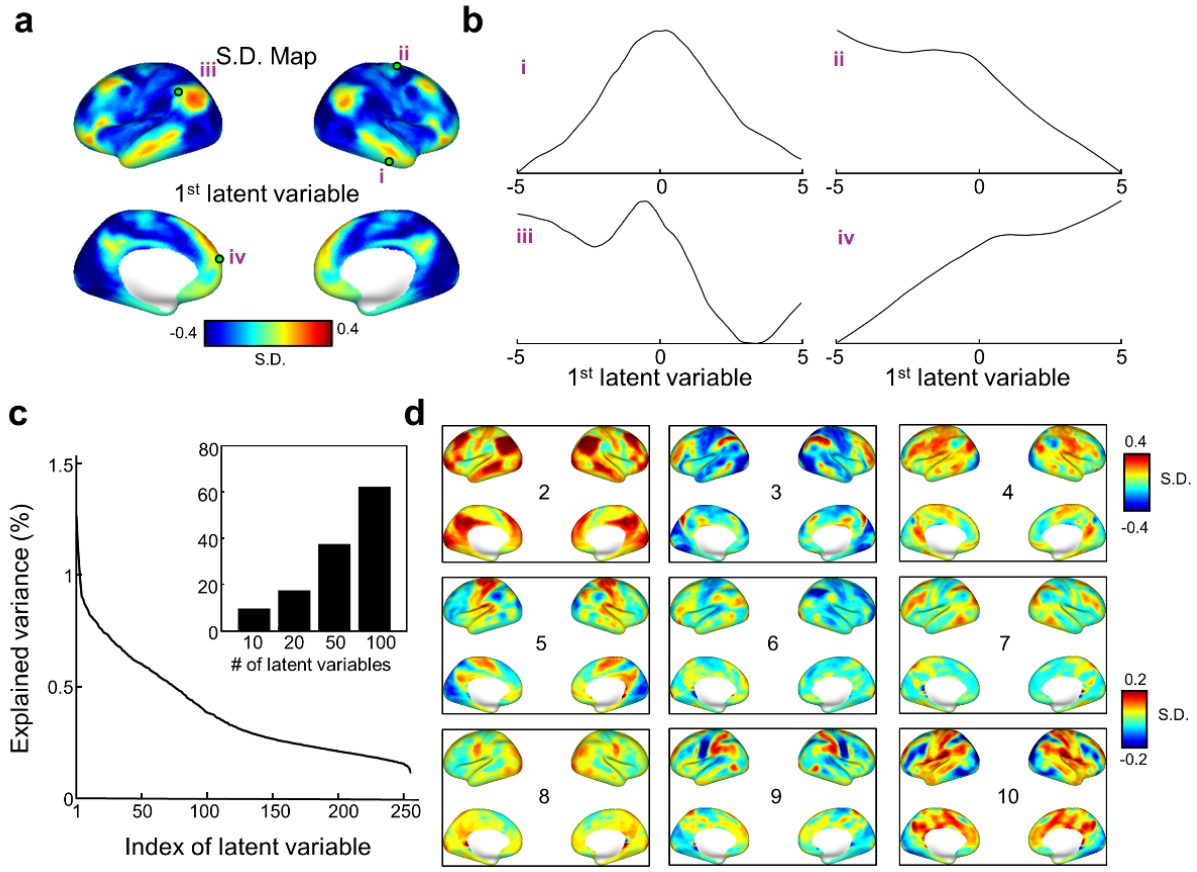


Figure 6. **Latent variables drive the dynamics of latent representation.** The latent variables are defined as the principal directions for the dynamic change of latent representation across adjacent time points. **(a)** Visualization of the 1st latent variable as a cortical pattern of the signed standard deviation. (**b**) For each of the four cortical locations, denoted as i through iv and shown as green circles in (**a**), the activity change is shown as a function of the 1st latent variable, indicating a varying nonlinear relationship. (**c**) The percentage of the variance that each latent variable explains the first order dynamics of latent representation. The inset shows the percentage of the total variance explained by top 10, 20, 50, or 100 latent variables. (**d**) The visualization of the 2nd through 10th latent variables as cortical patterns.

472 **Clusters in latent space**

473 We further characterized the distribution of latent representation and attempted to
474 identify clusters in the latent space. We used the VAE to encode the rsfMRI pattern
475 observed at every time point from 500 subjects, clustered the time points by applying k-
476 means clustering (k=21) to the latent representations, and decoded the cluster centroids
477 to corresponding cortical maps. The number of clusters (k=21) was close to the "elbow"
478 indicative of a reasonable balance between reducing variation within clusters and
479 avoiding too many clusters (Figure 7.a). The 1-cosine dissimilarity between latent
480 representations reordered by their cluster membership shows not only close affinity within
481 every cluster but also a varying level of affinity between different clusters (Figure 7.b).
482 This motivated us to hierarchically merge clusters into super-clusters (or "clusters of
483 clusters") based on the cosine affinity between cluster centroids (Figure 7.c).

484 For each of the 21 clusters, we decoded and visualized the cluster centroid as a
485 cortical pattern as shown in Figure 7.d. Among the 21 clusters, 5 clusters (Cluster 4, 6,
486 11, 12, 18) showed activity increase (positive) at one or multiple regions in the default
487 mode network (Buckner et al., 2008; Greicius et al., 2003; Raichle et al., 2001), alongside
488 activity decrease (negative) at other regions. Similarly, we found 5 clusters with activity
489 increase in the so-called frontoparietal control network (Cluster 8) (Dixon et al., 2018),
490 cingulo-opercular network (Cluster 7 and 9) (Dosenbach et al., 2007), cognitive control
491 network (Cluster 1) (Cole and Schneider, 2007), and dorsal attention networks (Cluster
492 10) (Fox et al., 2006) – collectively referred to as "the task positive network" (Fox et al.,
493 2005). In addition, cluster 13 and 16 showed activity decrease in the whole brain, thereby
494 a signature of global signal fluctuation (Murphy et al., 2009; Schölvinck et al., 2010; Wen

and Liu, 2016). Cluster 5 and 17 showed widespread synchrony across sensory systems. Cluster 2 and 21 showed the networks for sensorimotor control of the limbs and of the mouth, pharynx, and visceral organs, respectively. Whereas most clusters were bilaterally symmetric, Cluster 15 and 3 were unilateral to the right and left prefrontal cortex, respectively. A common observation for many clusters was that a cluster could highlight the positive interactions among a set of well-defined cortical regions alongside their negative interactions with a different set of regions. Given the above interpretation of individual clusters, we further interpreted the three super-clusters as sensorimotor, default mode, and task positive networks (Figure 7.C).

In addition, we evaluated the temporal dynamics of latent representation in terms of the dynamics of individual clusters or their corresponding cortical networks. Intuitively, we considered the time-evolving trajectory of the latent representation as the movement towards or away from each cluster. In this regard, we expressed the cluster-wise activity as the time series of cosine affinity between the instantaneous latent representation and the cluster centroid. During a rsfMRI session, different clusters expressed similar activity levels (Figure 7.e), except in the initial period of the session. In that period of 20 seconds, clusters presumably related to task positive networks showed a transition from a high activity level to a lower steady state; the clusters related to sensorimotor networks showed a transition from a low activity level to a higher steady state; in contrary, clusters related to the default mode network remained roughly unchanged. These (somewhat incidental) observations were consistent and reproducible across individuals and sessions. On one hand, this result suggests that the first 20 seconds in a rsfMRI session are not necessarily the steady state under a resting condition. On the other hand, this exploratory analysis

shows the feasibility of using the VAE-extracted latent representations to identify brain networks and reveal their individual dynamics.

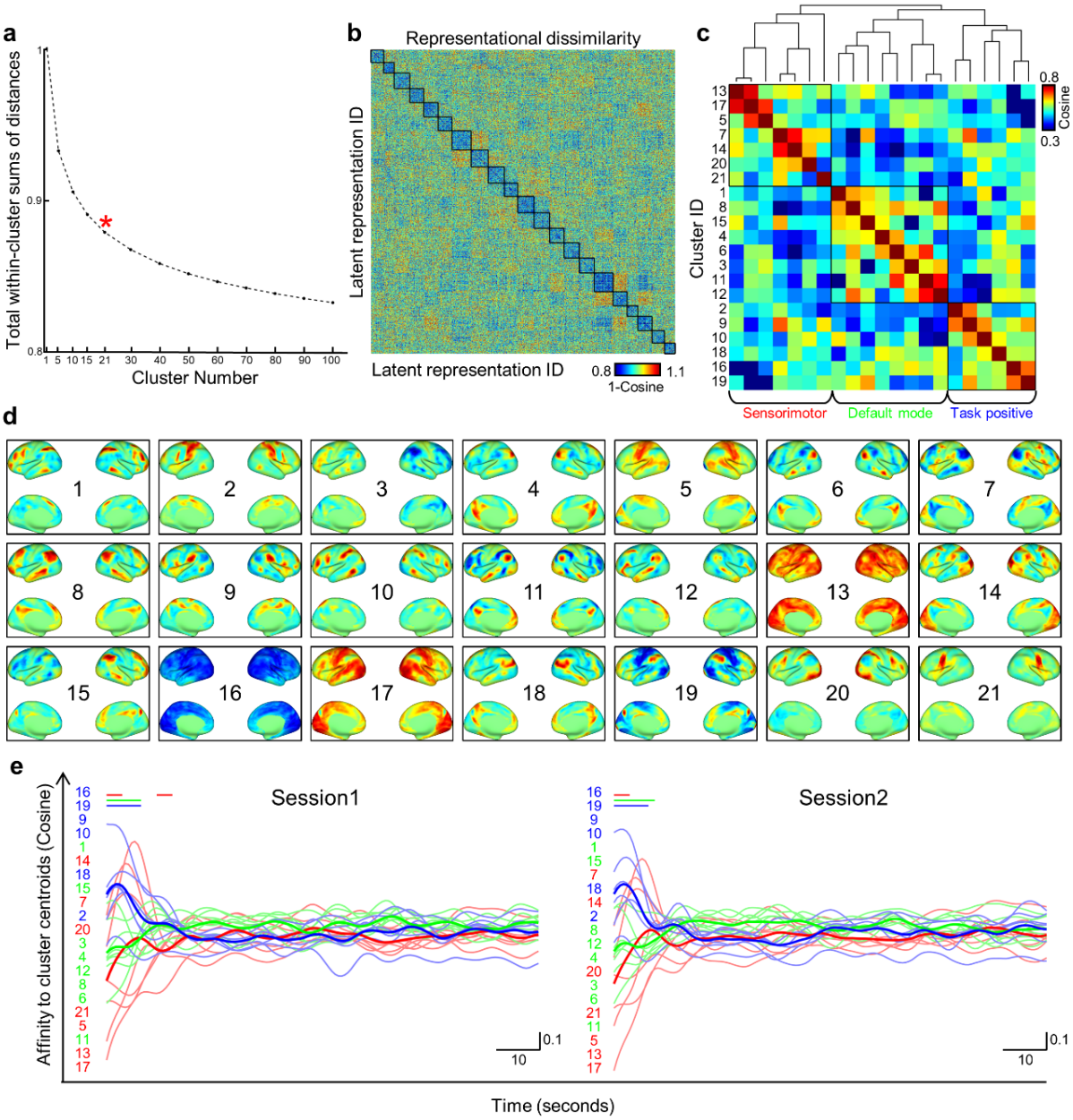


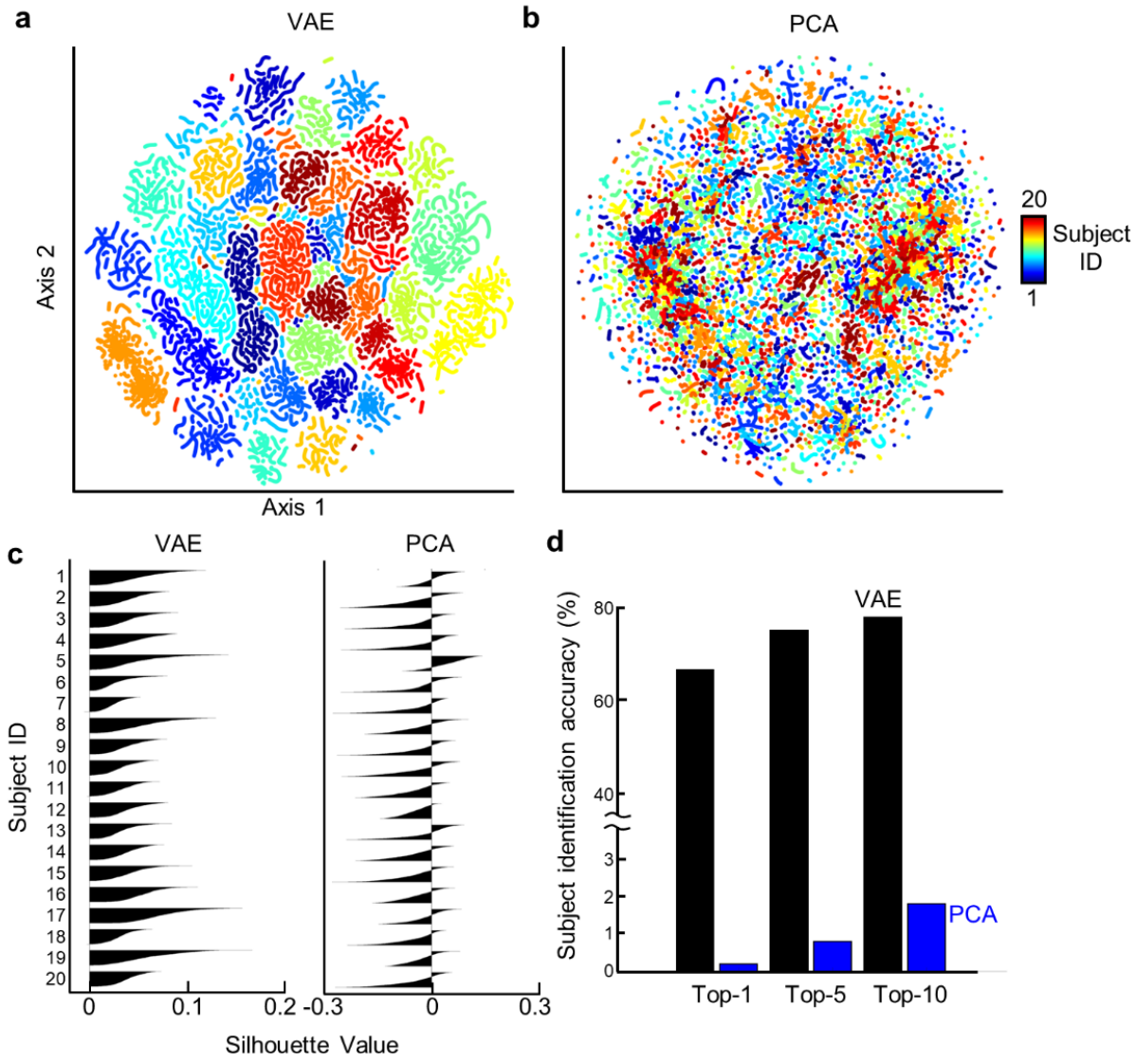
Figure 7. Clusters of latent representations. (**a**) The sum of within-cluster distance is shown as a function of the number of clusters (or K). The choice of k=21 is about where an elbow is observable in the plot. (**b**) Representational dissimilarity matrix of the 1-cosine distance between instantaneous latent representations (from 100 subjects with 1,200 time points per subject) reordered by cluster membership. (**c**) Hierarchical

clustering of 21 cluster centroids. The clusters are grouped into three super-clusters, provisionally labeled as sensorimotor (red), default mode (green), and task positive (blue) networks, based on the cortical visualization of individual clusters in (**d**). (**d**) Cortical patterns decoded from every cluster centroid. The number shows the cluster index. The scale of each pattern is normalized by its maximal absolute value. (**e**) The group-averaged cluster-wise activity, described as the cosine affinity of instantaneous representation to the centroid of each cluster. Each thin line corresponds to one cluster; the color indicating the super-cluster (sensorimotor: red; default mode: green; task positive: blue) that each cluster belongs to. The thick lines show the average within super-clusters. The left and right panels show the activity patterns averaged across all subjects for session 1 and session 2, respectively. Colored lines on the top of panel **(e)** highlight the periods in which the sensorimotor (red), default-mode (green) or task positive (blue) super-cluster was statistically significant in the group level (permutation test, false discovery rate $q<0.01$).

Individual variation of latent representation

Whereas the aforementioned analyses focused on the group-level characteristics of the latent representations, we further asked how the distribution and geometry of latent representation varied across individuals. Only for the sake of demonstration, we randomly selected 20 subjects in the testing dataset and visualized their individual representations in the latent space after reducing its dimension from 256 to 2 by using t-SNE (Figure 8.a). Strikingly, the latent representations were grouped by and separable across individuals. The clustering by individuals was noticeable in the nonlinear latent space obtained with VAE (Figure 8.a), but not in the linear latent space obtained with PCA (Figure 8.b). Such distinctions were quantitatively confirmed (Figure 8.c) by using the Silhouette value to measure the degree of clustering by individuals. The Silhouette value for VAE (mean ± std: s = 0.044±0.002, 50 bootstrapping trials) was significantly higher ($p<0.001$, two-sample t-test) than that for PCA (s = -0.020±0.015). Using the center of latent

550 representation as the subject-identifying feature, we found that subject identity could be
551 retrieved with a reasonably high accuracy when the latent representation was extracted
552 by VAE, whereas the linear representation by PCA failed the same task nearly entirely
553 (Figure 8.d). These results suggest the feasibility of using VAE to characterize and reveal
554 individual variations of resting state activity in a non-linear latent space.



555

556 **Figure 8**. **Individual variation of latent representation obtained with VAE vs. PCA. (a-b)** Subject-wise
557 latent representations visualized in a 2-D space obtained with t-SNE, when **(a)** VAE or **(b)** PCA is used to
558 extract representations of rsfMRI activity from 20 subjects. **(c)** The Silhouette value shows how similar a

representation is similar to each other within the same subject as opposed to between different subjects for VAE (left) or PCA (right). **(d)** The top-1, 5, and 10 accuracy of using the time-averaged representation as the feature to identify individuals in a large group of (n=500) subjects, for the representations obtained with VAE (black) or PCA (blue).

Individual identification

From the t-SNE based visualization (Fig. 8.a), it was noticeable that subject-wise representations exhibited different geometries. Some were more elongated or scattered than others. This observation motivated us to ask whether the representational geometry (Kriegeskorte and Kievit, 2013) could be an individual-specific feature (or "fingerprint") to allow for more accurate individual identification. Specifically, we calculated the covariance between every pair of latent variables and assembled the pair-wise covariance into a vector as the feature of the representational geometry and evaluated the similarity in this feature between two sessions within or between subjects. The representational geometry evaluated in this way could be interpreted as the functional connectivity (FC) between latent variables. This interpretation related this approach to a conceptually similar approach: the "connectome-based fingerprinting" (Finn et al., 2015; Venkatesh et al., 2020), in which the functional connectivity was evaluated between cortical parcels. So, we evaluated the use of either the latent-space or cortical-space FC for individual identification in comparison.

As shown in Figure 9.a, FC between any pair of cortical areas was mostly positive (mean ± std of z-transformed correlation: z=0.26±0.3) and highly reproducible not only within the same subject (r=0.66) but also between different subjects (r=0.45). On the other hand, FC between latent variables had both positive and negative values (mean ± std of covariance: σ^2=0.00±0.13) and its reproducibility was high only within the same subject

584 (r=0.33) but not between different subjects (r=0.07). The FC profile was more distinctive
585 across subjects when it was evaluated between latent variables rather than cortical areas
586 (Figure 9.b). In the latent space, the FC profile was significantly more consistent within a
587 subject than between subjects (two-sample t-test, $t(249,998)=254.05$, two-sided $p<0.001$).
588 The distribution of within-subject correlations was in nearly complete separation from that
589 of between-subject correlations (Figure 9.b, bottom).

590 Then we compared the performance of individual identification on the basis of the
591 FC profile in the latent vs. cortical space. To identify 1 out of 500 subjects, we compared
592 a target subject's FC profile in the 1st session (as a query) against every subject's FC
593 profile in the 2nd session (as a key) and chose the best match between the query and the
594 key in terms of the Pearson correlation coefficient. As such, the choice was correct if the
595 correlation with the target subject was higher than the largest correlation with any non-
596 target subject. We found that the FC profile in the cortical space could support 69.3% top-
597 1 accuracy while identification was often made with marginal confidence relative to the
598 decision boundary (Figure 9.c). Using the FC in the latent space allowed us to reach 97.8%
599 top-1 accuracy. The evidence for correct identification was apparent with a large margin
600 from the decision boundary (Figure 9.d). The use of FC in the latent space supported
601 reliable and robust performance in top-1 identification given an increasingly larger
602 population (Figure 9.e) or when the data were limited to a short duration (Figure 9.f), being
603 notably superior to the use of FC in the cortical space.

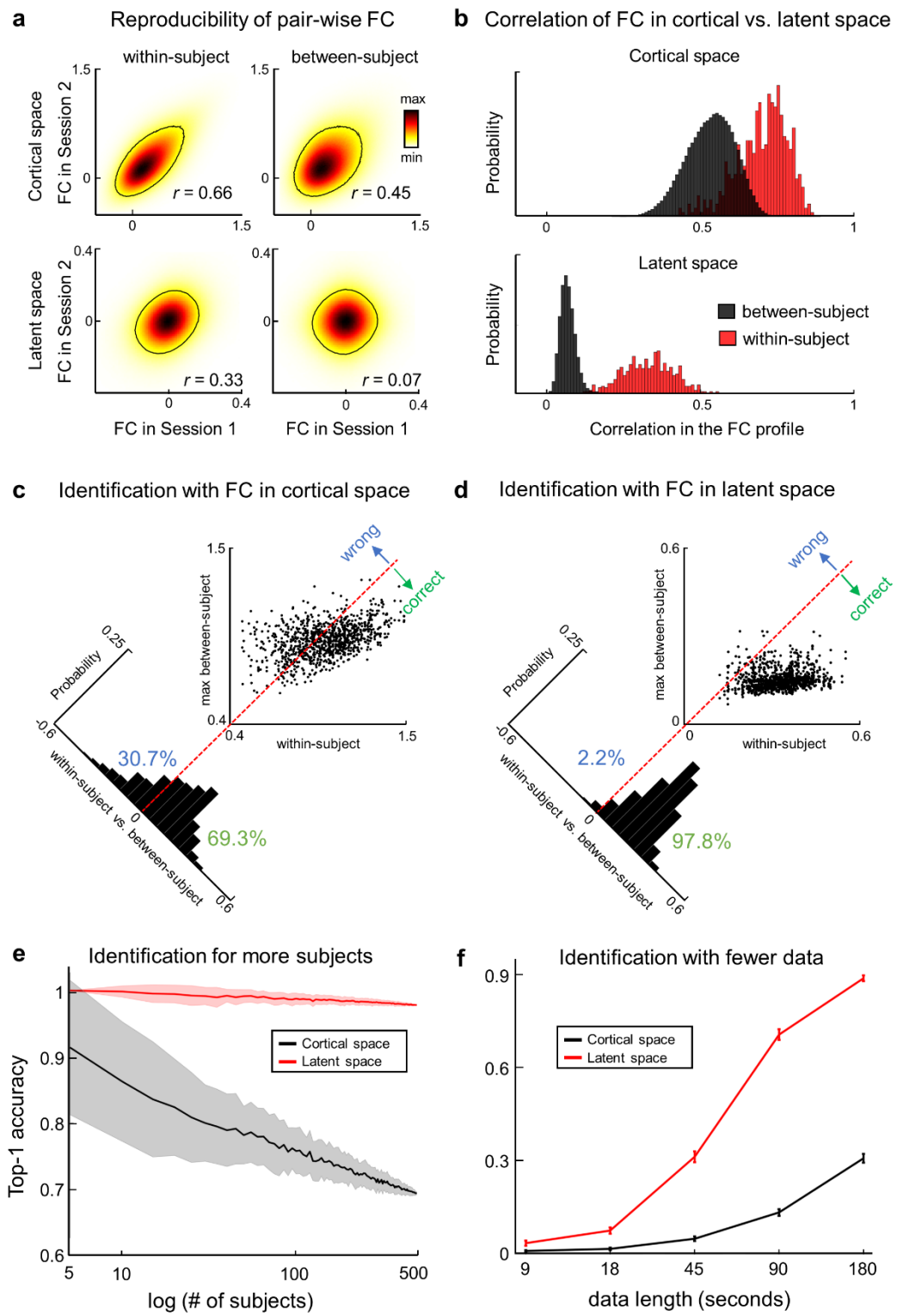


Figure 9. Individual identification based on correlations between latent variables or cortical parcels. **(a)** Density distributions of z-transformed correlations between every pair of cortical parcels (top) or

covariance between every pair of latent variables (bottom). For each pair, the correlation and covariance in one session is plotted against the corresponding correlation in the other session for the same subject (within-subject, left) or different subjects (between-subject, right) given the testing dataset with n=500 subjects. Contour line stands for 20% of the maximal density. **(b)** Within-subject (red) and between-subject (black) correlations in the FC among cortical parcels (top) or latent variables (bottom) are shown as histograms with the width of each bin at 0.01. **(c)** In the scatter plot, each dot indicates one subject, plotting the maximal correlation in the cortical FC profile between that subject and a different subject against the corresponding correlation within that subject. The red-dashed line indicates y=x, serving as a decision boundary, across which identification is correct (x>y) or wrong (y>x). The histogram shows the distribution of y-x (0.05 bin width) with the decision boundary corresponding to 0. Similarly, (**d**) presents the results obtained with latent-space FC in the same format as (**c**). (**e**) Top-1 identification accuracy evaluated with an increasing number of subjects (n=5 to 500) given the latent-space (red) or cortical-space (black) FC profile. The solid line and the shade indicate the mean and the standard deviation of the results with different testing data. (**f**) Top-1 identification accuracy given rsfMRI data of different lengths (from 9s to 180s). The line and the error bar indicate the mean and the standard deviation with different testing data.

We further tested to what extent the performance of individual identification relied on the use of ICA-FIX to preprocess and denoise the rsfMRI data. For this purpose, we applied ICA-FIX to one or both of the two sessions in every subject and then tested the individual identification with n=500 subjects. As shown in Table 1, when the FC profile in the latent space was derived from the (ICA-FIX denoised) clean data for both the keys and queries, the identification has the highest accuracy (97.5%). When the key and the query were both based on noisy data (without denoising), the accuracy dropped to 91.3%. When the key and the query were unpaired as denoising applied to one but not the other, the accuracy further dropped to about 88%. Nevertheless, this performance obtained with the latent-space FC was still notably higher than the performance based on the cortical-space FC. For the latter, the use of unpaired preprocessing for the query and the key significantly dropped the identification performance from 69.3% to 47.5%. Counter-intuitively, when the denoising was applied to neither the query nor the key, the

identification accuracy with the cortical-space FC increased to 76.9%, but still significantly lower than the accuracy of 91.3% obtained with the latent-space FC.

Table 1. Subject identification accuracy across different conditions

			Session 1	
			Clean	Noisy
Session 2	Latent Space	Clean	97.8%	87.9%
		Noisy	88.4%	91.3%
	Cortical Space	Clean	69.3%	47.2%
		Noisy	47.5%	76.9%

Lastly, we explored whether the representational geometry (based on the profile of the covariance between latent variables) would yield a similar level of distinction across individuals for a linear latent space obtained with PCA or ICA. As shown in Figure. 10, PCA or ICA was not as effective as VAE. The top-1 accuracy of individual identification was 61.1% for PCA, 63.6% for ICA, in contrast to 97.8% for VAE. The within-subject vs. between-subject similarity in the geometry of linear representation obtained with PCA or ICA exhibited largely overlapping distributions, whereas the corresponding distributions were separated nearly completely for the nonlinear representations obtained with VAE.

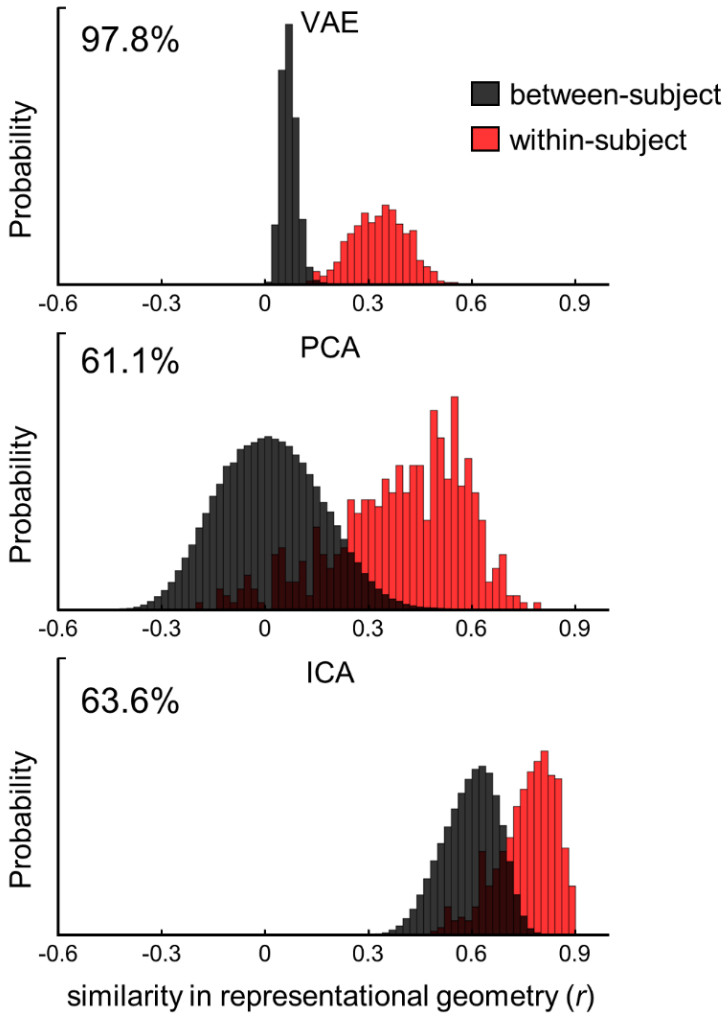


Figure 10. **Individual identification with nonlinear vs. linear representations.** Each plot shows the histogram of the similarity in the representational geometry between sessions within the same subject (red) vs. across different subjects (black), for representations in the nonlinear latent space obtained by VAE (top) or in the linear latent space obtained by PCA (middle) or ICA (bottom). The similarity reported is based on the inter-session correlation coefficient (or *r*). The histogram is discretized by bins with a width of 0.02.

Discussion

Here, we present a method for unsupervised representation learning of cortical rsfMRI activity. Our results suggest that this method is able to disentangle generative

factors underlying spontaneous brain activity, discover overlapping brain networks, capture individual characteristics or variation, and support accurate individual identification. We expect this method to be a valuable addition to the existing tools for investigating the origins of resting state activity, mapping functional brain networks, and potentially supporting individualized prediction of disease phenotypes and progression. Next, we discuss our findings from the joint perspective of methodology, neuroscience, and applications.

VAE is trainable with unsupervised learning (without any label) (Higgins et al., 2017; Kingma and Welling, 2013), which is appealing for learning representations of rsfMRI data. Since rsfMRI measures spontaneous brain activity unconstrained by any task, labels as required for supervised learning are either unavailable or far fewer than the data itself. Unsupervised learning with VAE can leverage the ever-increasing amount of rsfMRI data (Van Essen et al., 2013). The latent representations extracted from VAE can serve as the input to other algorithms to further support more specific goals such as classification of brain disorders and prediction of their phenotypes (Garrity et al., 2007; Moradi et al., 2015; Shen et al., 2010; Zhang et al., 2011).

The method herein can be extended in multiple ways. Although it is trained with rsfMRI data, we hypothesize that the VAE model can encode and decode both rsfMRI and task-fMRI data but with different latent distributions. If this is true, one may use this model to classify different perceptual, behavioral, or cognitive states and to reveal the distinctive network interactions underlying various states (Gonzalez-Castillo et al., 2015). The fact that the VAE can synthesize new data (Figure 5) is also appealing. It can be used as a post-processing strategy for data augmentation and interpolation, when data is short

or corrupted, of interest for evaluation of dynamic functional connectivity (Allen et al., 2014; Chang and Glover, 2010) and correction for head motion (Power et al., 2014). It also supports the notion that the learned latent space captures the origins of rsfMRI and the VAE decoder captures the computational account for how rsfMRI arises from its origins.

It is worth mentioning two limitations of the VAE model in its current form. First, the model focuses on cortical patterns but excludes sub-cortical and white-matter voxels. This design is not only for the ease of model implementation but also for the predominant role of the neocortex in brain functions (Rakic, 2009). However, this precludes the model from accounting for subcortical networks or their interactions with the cortex. Addressing this limitation awaits future studies to redesign the model as a 3-D neural network that takes volumetric fMRI data as the input. Second, the VAE model only represents spatial patterns but ignores temporal dynamics inherent to rsfMRI data. Modeling the temporal dynamics is desirable but non-trivial, since it is highly irregular, complex and variable. To fill this gap, we direct future studies to designing a recurrent neural network (Chen and Hu, 2018; Cui et al., 2019; Shi et al., 2018; Sutskever et al., 2014; Zhao et al., 2019), as an add-on to VAE, to further learn sequence representation, for example, with a self-supervised predictive learning strategy (Kashyap and Keilholz, 2020; Khosla et al., 2019a).

Although VAE does not explicitly model the temporal dynamics, the representation obtained with VAE preserves the temporal dynamics (Figure 3). The trajectory of the latent representation describes the temporal behavior of brain networks, as opposed to voxels or regions. This trajectory is amenable to the use of many methods previously described for voxel-wise or region-wise analysis. To note a few examples explored in this study, the first-order temporal difference in the latent representation captures the gradient of latent

trajectory that drives the brain to change its activity pattern from one time point to the next. As the latent gradient is also represented as a vector in the latent space, the length of this vector measures the displacement in the latent space and presumably the magnitude of network activity, and the direction of this vector encodes a pattern of network interaction that drives the instantaneous change of brain activity. The principal components of the displacement in representation uncover the important hidden factors that drive the temporal dynamics of brain networks (Figure 6). Similar analysis or notion has also been explored in two independent studies discussed in two very recent papers published or in preprint during the peer review of our paper (Brown et al., 2020; Liu et al., 2020b). These initial analyses are expected to merit and direct future studies upon predictive modeling of the trajectory of the VAE-derived latent representation, for example, by using Multivariate Auto-Regressive models (Liégeois et al., 2019; Rogers et al., 2010), Hidden Markov Models (Eavani et al., 2013; Suk et al., 2016).

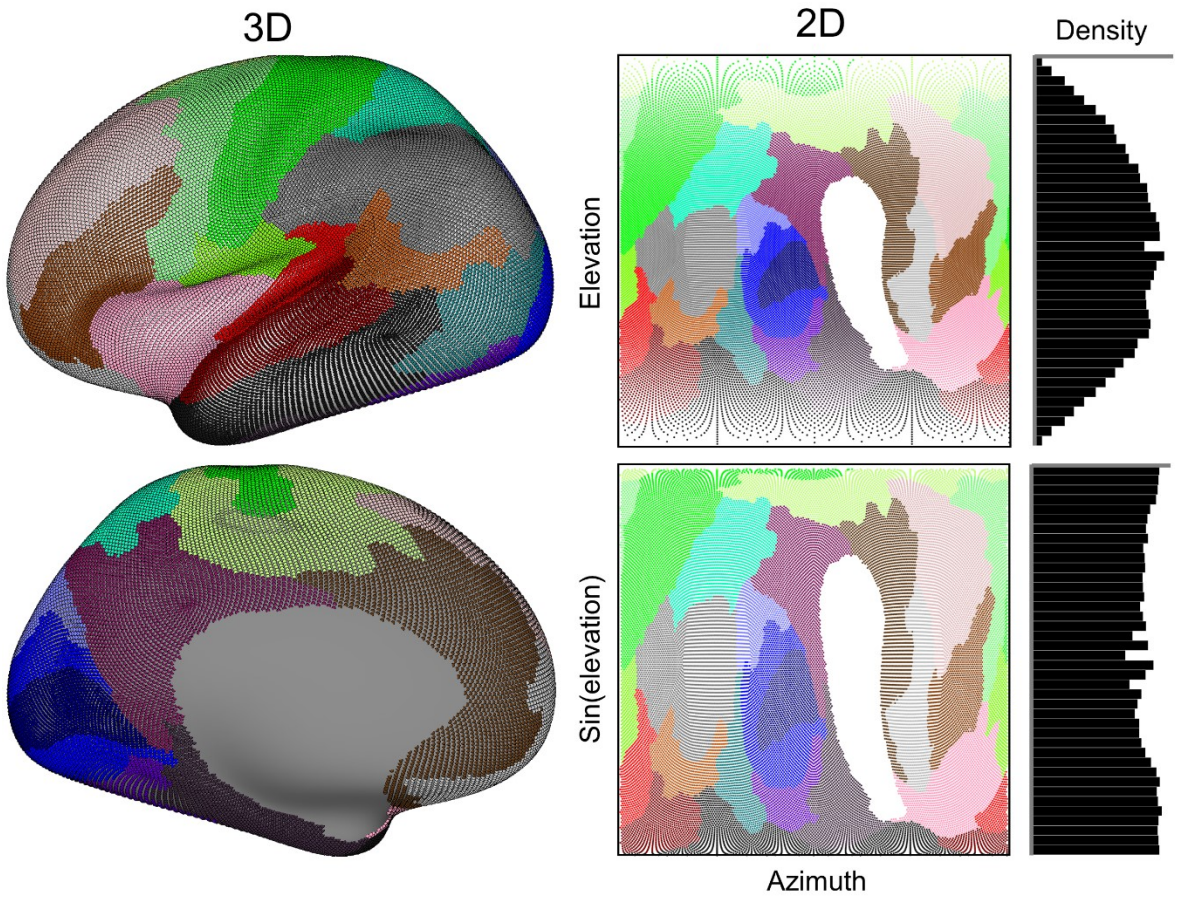
VAE provides a new tool for mapping overlapping functional networks in the brain. A brain region may be involved in multiple networks each supporting a distinctive function (Liu and Duyn, 2013; Smith et al., 2012). However, existing network analyses still tend to group brain regions into non-overlapping networks (Yeo et al., 2011). VAE allows us to discover overlapping networks as clusters in the latent space spanned by independent latent variables. As such, VAE is conceptually similar to temporal ICA (Smith et al., 2012) but allows for nonlinear relationships between latent variables and the input data they represent (Khemakhem et al., 2019). Arguably, finding clusters in the low-dimensional latent space is more desirable than doing so in the higher-dimensional voxel space (Liu et al., 2013). Not only is it more computationally efficient, but representations are also

more disentangled in the latent space than in the voxel space to readily reveal the underlying organization. However, it is not readily straightforward to attribute a cluster in the latent space to a distinct brain state (Hutchison et al., 2013) or an individual (Xie et al., 2018). Both are plausible. Our results show that individual variation manifests itself as the latent representation is in part clustered by subject (Figure 8), suggesting individual variation is a contributing factor to the clustering of latent representation. Our results also suggest that cluster-wise activity shows a consistent pattern across all subjects, in particular for the first 20 seconds of each session (Figure 7). Moreover, the clusters seem to group themselves hierarchically into presumably functional domains: sensorimotor, default-mode and task-positive networks (Figure 7). Together These results lead us to speculate that variation in brain states and individuals both contribute to the clustering of brain activity in the latent space. It is challenging to fully separate them and awaits future studies.

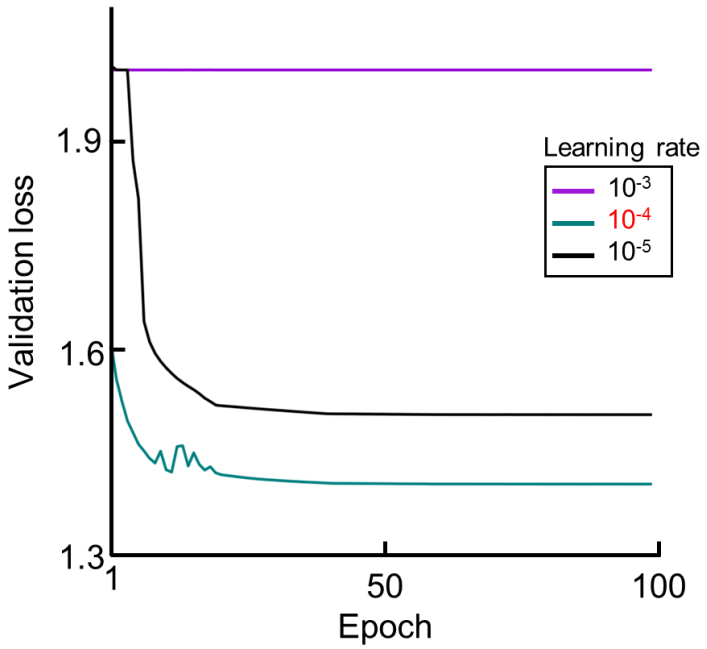
Central to this study is the efficacy of using VAE to disentangle what causes resting state activity. In the VAE model, the sources are the latent variables; the decoder describes how the sources generate the observed activity; the encoder models the inverse inference of the sources from the activity. Since the latent variables are data-driven, it is currently unclear how to interpret them as specific physiological processes, many of which are not observable. Nevertheless, we expect the latent variables extracted by VAE to provide the computational basis for further understanding the origins of resting state activity. We hypothesize that the truly disentangled physiological origins, whether observable or not, are individually describable as the latent variables up to linear and sparse projection. This hypothesis awaits confirmation by future studies.

In the latent space, functional connectivity between latent variables describes the geometry of the latent representation of rsfMRI activity. This is a new perspective different from the functional connectivity among observable voxels, regions or networks (Biswal et al., 1995; Yeo et al., 2011). If the VAE model has fully disentangled the sources in a population level, functional connectivity should be near zero between different latent variables and thus reflect a spherical geometry. In other words, the model sets a nearly null population-level baseline, against which individual variation stands out. The latent-space functional connectivity given data from a single subject becomes a unique feature of that subject. Supporting this notion, the use of functional connectivity in the latent space allows for a significantly improved accuracy, robustness, and efficiency in individual identification, compared to the use of functional connectivity among cortical parcels (Amico and Goñi, 2018; Byrge and Kennedy, 2019; Finn et al., 2015; Mejia et al., 2018; Venkatesh et al., 2020).

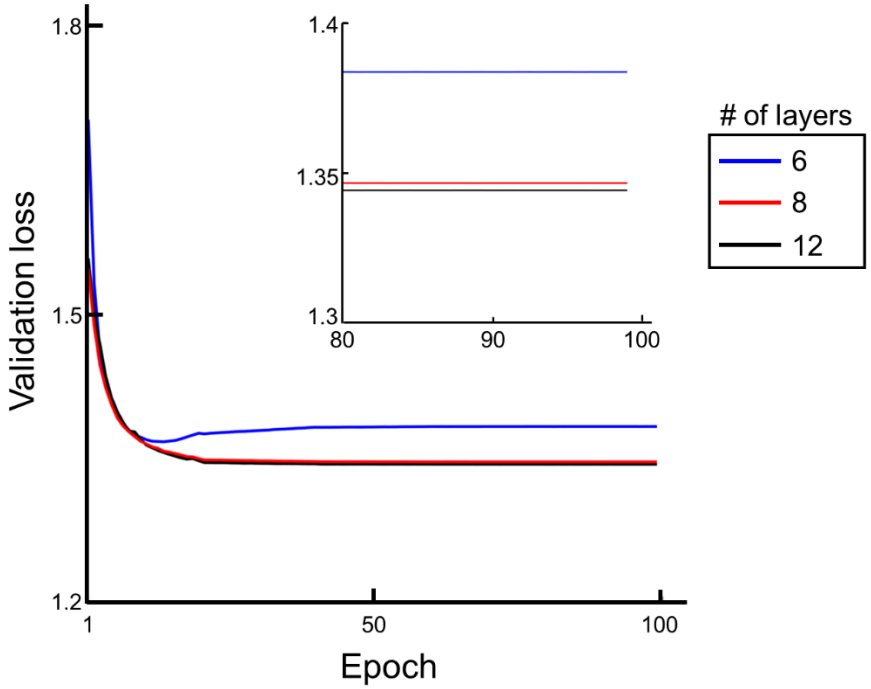
Note that our main purpose is not to push for a higher identification accuracy but to understand the distribution and geometry of data representations in the feature space. Therefore, we opt for minimal preprocessing and the simplest strategy for individual identification. There is room for methodological development to further improve the identification accuracy or to extend it for many other tasks, including classification of the gender or disease states, prediction of behavioral and cognitive performances, to name a few examples. We expect that such applications would be fruitful and potentially impactful to cognitive sciences and clinical applications.



Supplementary Figure 1. Grayordinates (# = 29,696) in left hemisphere (left panel) is projected into 2D space (middle panel) based on their azimuth- and elevation levels. Each color stands for different brain atlas based on the literature (Glasser et al., 2016). The imbalance in data density along the varying elevation level (upper in right panel) is alleviated by further applying elevation to Sine function (bottom in right panel).



Supplementary Figure 2. Validation curve of VAE at varying learning rate.



Supplementary Figure 3. Validation curve of VAE at different layers

Reference

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. Cerebral cortex 24, 663-676.

Amico, E., Goñi, J., 2018. The quest for identifiability in human functional connectomes. Scientific reports 8, 1-14.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE transactions on medical imaging 23, 137-152.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S.G., de Zwart, J.A., Shmueli, K., Duyn, J.H., 2009. Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magnetic resonance imaging 27, 1019-1029.

Birn, R.M., Smith, M.A., Jones, T.B., Bandettini, P.A., 2008. The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage 40, 644-654.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic resonance in medicine 34, 537-541.

Brown, J.A., Lee, A.J., Pasquini, L., Seeley, W.W., 2020. Intrinsic brain activity gradients dynamically coordinate functional connectivity states. bioRxiv.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease.

Byrge, L., Kennedy, D.P., 2019. High-accuracy individual identification using a “thin slice” of the functional connectome. Network Neuroscience 3, 363-383.

Calhoun, V.D., Adali, T., Pearlson, G., Pekar, J.J., 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Human brain mapping 13, 43-53.

Chang, C., Glover, G.H., 2010. Time–frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage 50, 81-98.

Chang, C., Leopold, D.A., Schölvinck, M.L., Mandelkow, H., Picchioni, D., Liu, X., Frank, Q.Y., Turchi, J.N., Duyn, J.H., 2016. Tracking brain arousal fluctuations with fMRI. Proceedings of the National Academy of Sciences 113, 4518-4523.

Chen, S., Hu, X., 2018. Individual identification using the functional brain fingerprint detected by the recurrent neural network. Brain connectivity 8, 197-204.

Chou, Y.-h., Sundman, M., Whitson, H.E., Gaur, P., Chu, M.-L., Weingarten, C.P., Madden, D.J., Wang, L., Kirste, I., Joliot, M., 2017. Maintenance and representation of mind wandering during Resting-State fMRI. Scientific reports 7, 40722.

Cole, M.W., Schneider, W., 2007. The cognitive control network: integrated cortical regions with dissociable functions. Neuroimage 37, 343-360.

Cui, Y., Zhao, S., Chen, Y., Han, J., Guo, L., Xie, L., Liu, T., 2019. Modeling brain diverse and complex hemodynamic response patterns via deep recurrent autoencoder. IEEE Transactions on Cognitive and Developmental Systems.

D’Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., Venkataraman, A., 2019. Integrating Neural Networks and Dictionary Learning for Multidimensional Clinical Characterizations from Functional Connectomics Data. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 709-717.

Dixon, M.L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, R.N., Cole, M.W., Christoff,

K., 2018. Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. Proceedings of the National Academy of Sciences 115, E1598-E1607.

Dosenbach, N.U., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A., Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2007. Distinct brain networks for adaptive and stable task control in humans. Proceedings of the National Academy of Sciences 104, 11073-11078.

Dvornek, N.C., Yang, D., Ventola, P., Duncan, J.S., 2018. Learning generalizable recurrent neural networks from small task-fmri datasets. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 329-337.

Eavani, H., Satterthwaite, T.D., Gur, R.E., Gur, R.C., Davatzikos, C., 2013. Unsupervised learning of functional network dynamics in resting state fMRI. International conference on information processing in medical imaging. Springer, pp. 426-437.

Fan, L., Su, J., Qin, J., Hu, D., Shen, H., 2020. A Deep Network Model on Dynamic Functional Connectivity With Applications to Gender Classification and Intelligence Prediction. Frontiers in neuroscience 14, 881.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature neuroscience 18, 1664.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774-781.

Fox, M.D., Buckner, R.L., Liu, H., Chakravarty, M.M., Lozano, A.M., Pascual-Leone, A., 2014. Resting-state networks link invasive and noninvasive brain stimulation across diverse psychiatric and neurological diseases. Proceedings of the National Academy of Sciences 111, E4367-E4375.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2006. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proceedings of the National Academy of Sciences 103, 10046-10051.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature reviews neuroscience 8, 700-711.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences 102, 9673-9678.

Gadgil, S., Zhao, Q., Adeli, E., Pfefferbaum, A., Sullivan, E.V., Pohl, K.M., 2020. Spatio-Temporal Graph Convolution for Functional MRI Analysis. arXiv preprint arXiv:2003.10613.

Garrity, A.G., Pearlson, G.D., McKiernan, K., Lloyd, D., Kiehl, K.A., Calhoun, V.D., 2007. Aberrant "default mode" functional connectivity in schizophrenia. American journal of psychiatry 164, 450-457.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. nature 536, 171-178.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105-124.

Gonzalez-Castillo, J., Hoy, C.W., Handwerker, D.A., Robinson, M.E., Buchanan, L.C., Saad, Z.S., Bandettini, P.A., 2015. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. Proceedings of the National Academy of Sciences 112, 8762-8767.

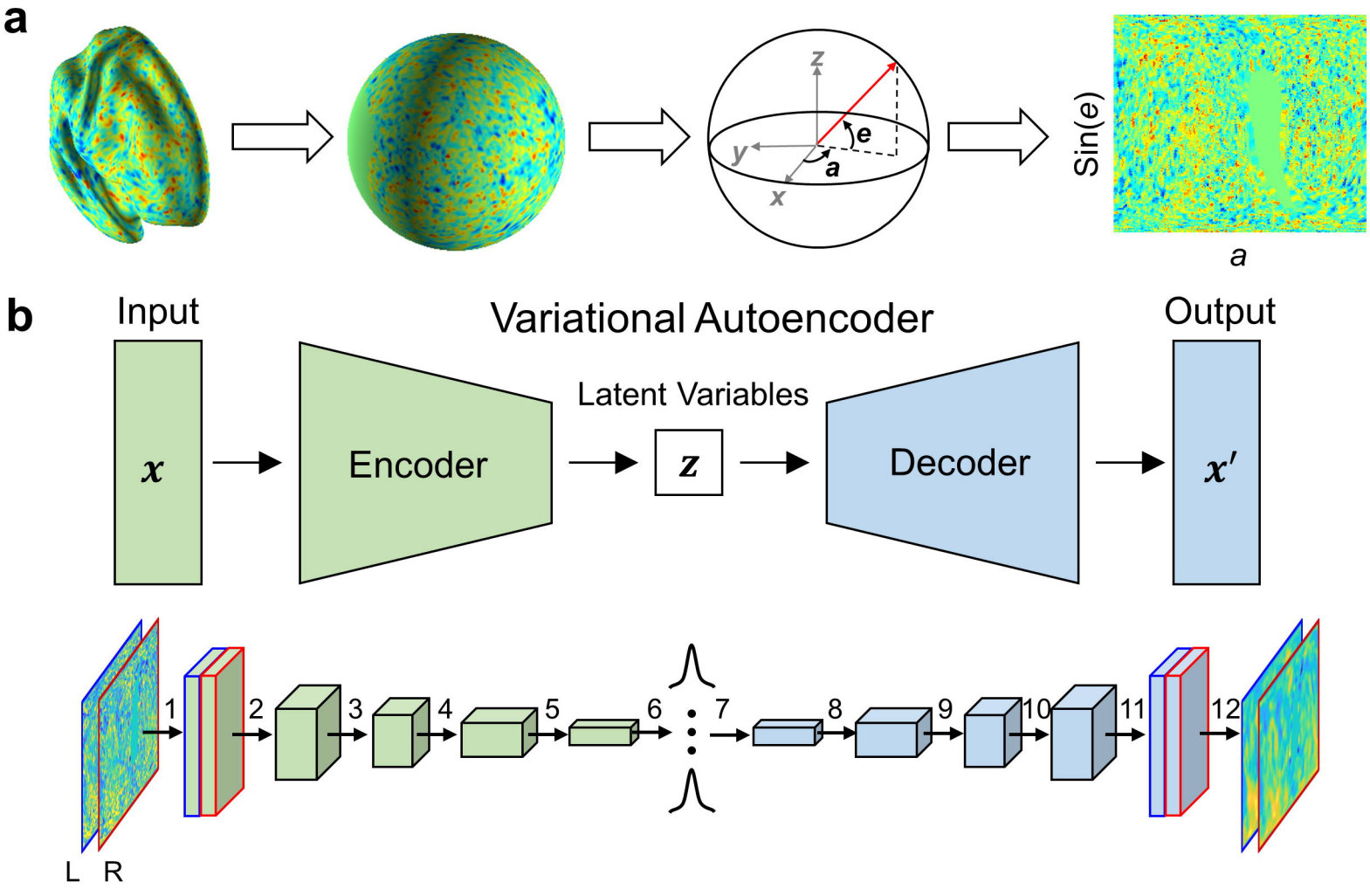
Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting

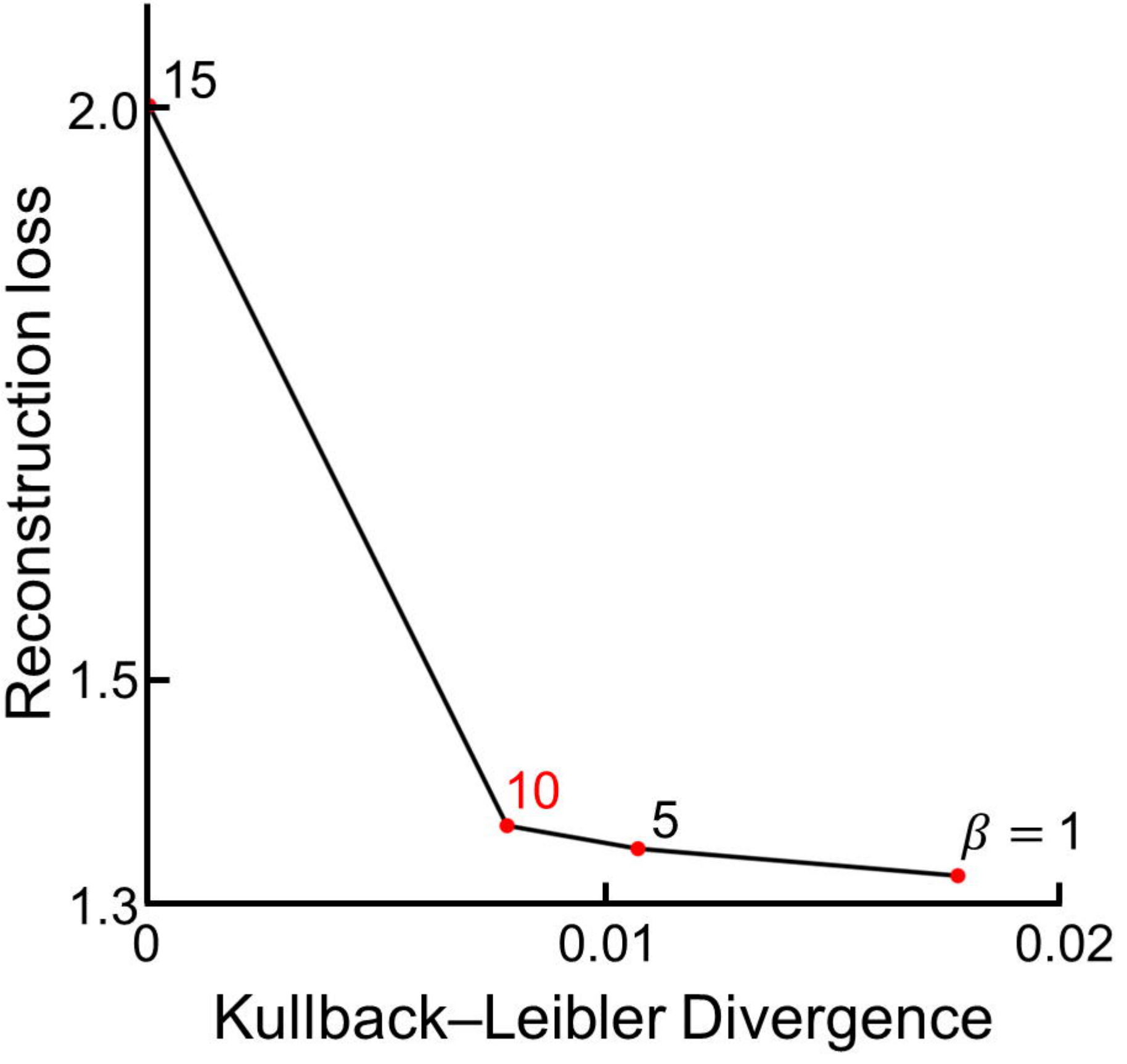
brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100, 253-258.
Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232-247.
Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., Liu, Z., 2019. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. Neuroimage 198, 125-136.
He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206, 116276.
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Iclr 2, 6.
Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., Liu, T., 2017. Modeling task fMRI data via deep convolutional autoencoder. IEEE transactions on medical imaging 37, 1551-1561.
Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360-378.
Jarrett, C., 2009. The restless brain. The Psychologist.
Kashyap, A., Keilholz, S., 2020. Brain network constraints and recurrent neural networks reproduce unique trajectories and state transitions seen over the span of minutes in resting-state fMRI. Network Neuroscience 4, 448-466.
Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146, 1038-1049.
Kell, A.J., Yamins, D.L., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630-644. e616.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915.
Khemakhem, I., Kingma, D.P., Hyvärinen, A., 2019. Variational autoencoders and nonlinear ica: A unifying framework. arXiv preprint arXiv:1907.04809.
Khosla, M., Jamison, K., Kuceyeski, A., Sabuncu, M.R., 2019a. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. Neuroimage 199, 651-662.
Khosla, M., Jamison, K., Ngo, G.H., Kuceyeski, A., Sabuncu, M.R., 2019b. Machine learning in resting-state fMRI analysis. Magnetic resonance imaging.
Kim, H.-c., Lee, J.-h., 2016. Evaluation of weight sparsity control during autoencoder training of resting-state fMRI using non-zero ratio and Hoyer's sparseness. 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, pp. 1-4.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint

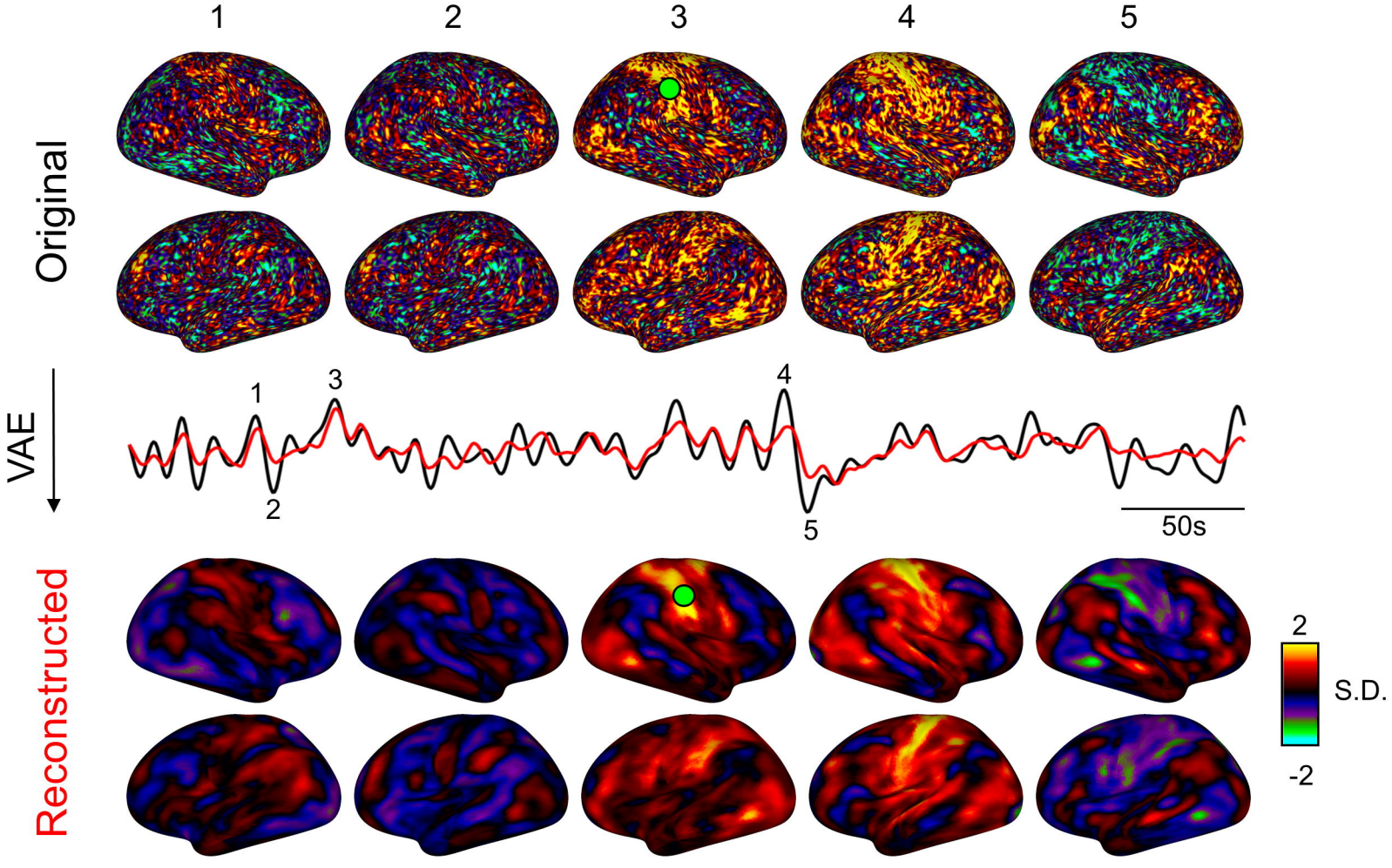
arXiv:1312.6114.
Koppe, G., Toutounji, H., Kirsch, P., Lis, S., Durstewitz, D., 2019. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. PLoS computational biology 15, e1007263.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends in cognitive sciences 17, 401-412.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436-444.
Leopold, D.A., Maier, A., 2012. Ongoing physiological processes in the cerebral cortex. Neuroimage 62, 2190-2200.
Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M.R., Yeo, B.T., 2019. Resting brain dynamics at different timescales capture distinct aspects of human behavior. Nature communications 10, 1-9.
Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., Xu, M., Initiative, A.s.D.N., 2020a. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. Neuroimage 208, 116459.
Liu, S., Zhao, L., Wang, X., Xin, Q., Zhao, J., Guttery, D.S., Zhang, Y.-D., 2020b. Deep Spatio-Temporal Representation and Ensemble Classification for Attention deficit/Hyperactivity disorder. IEEE Transactions on Neural Systems and Rehabilitation Engineering.
Liu, X., Chang, C., Duyn, J.H., 2013. Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns. Frontiers in systems neuroscience 7, 101.
Liu, X., Duyn, J.H., 2013. Time-varying functional network information extracted from brief instances of spontaneous brain activity. Proceedings of the National Academy of Sciences 110, 4392-4397.
Lu, H., Jaime, S., Yang, Y., 2019. Origins of the resting-state functional MRI signal: potential limitations of the "neurocentric" model. Frontiers in neuroscience 13.
Makkie, M., Huang, H., Zhao, Y., Vasilakos, A.V., Liu, T., 2019. Fast and scalable distributed deep convolutional autoencoder for fMRI big data analytics. Neurocomputing 325, 20-30.
Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., Corbetta, M., 2007. Electrophysiological signatures of resting state networks in the human brain. Proceedings of the National Academy of Sciences 104, 13170-13175.
Matsubara, T., Tashiro, T., Uehara, K., 2019. Deep neural generative model of functional MRI images for psychiatric disorder diagnosis. IEEE Transactions on Biomedical Engineering 66, 2768-2779.
Mejia, A.F., Nebel, M.B., Barber, A.D., Choe, A.S., Pekar, J.J., Caffo, B.S., Lindquist, M.A., 2018. Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. Neuroimage 172, 478-491.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.s.D.N., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398-412.
Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? Neuroimage 44, 893-905.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation

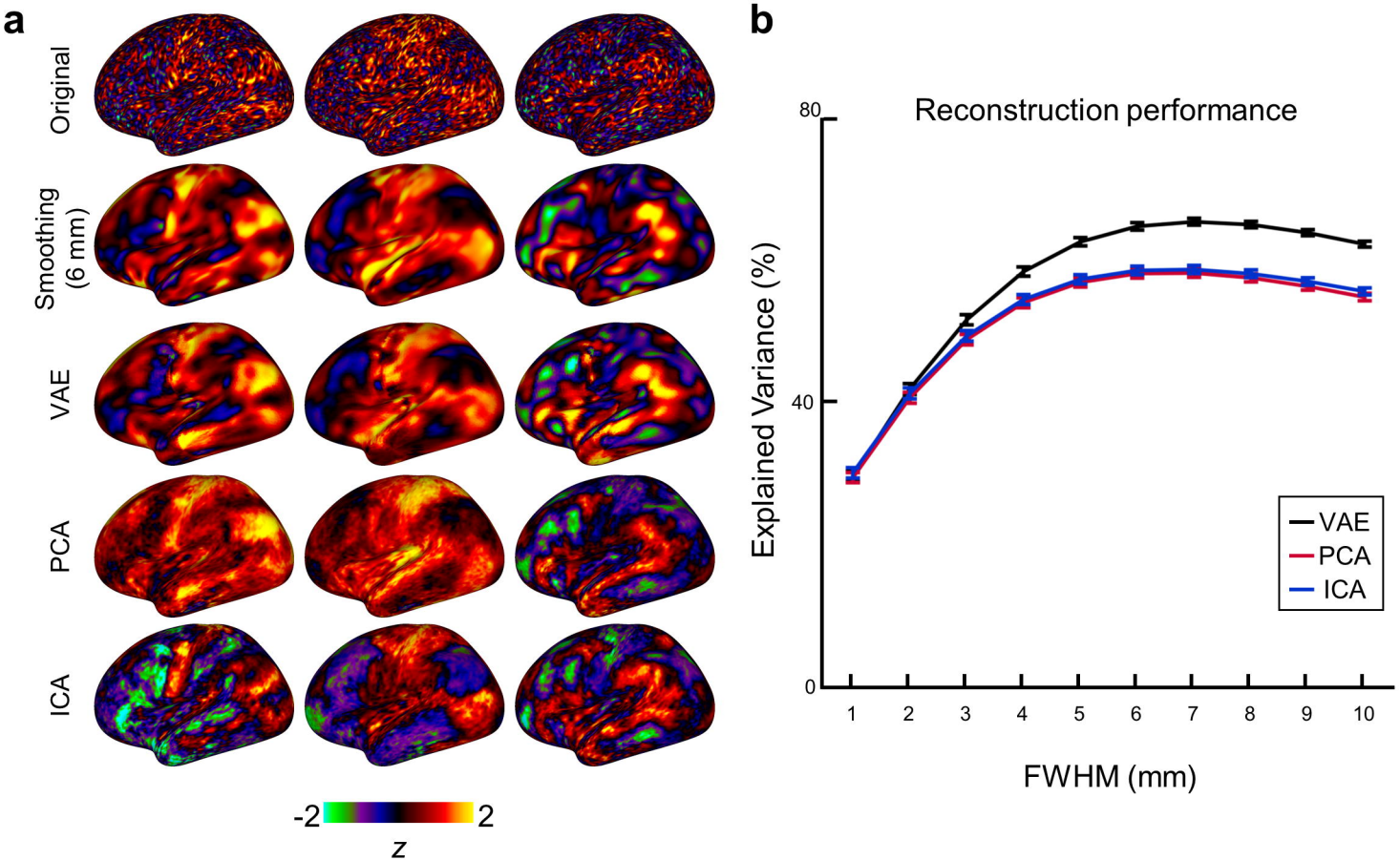
study. Frontiers in neuroscience 8, 229.
Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320-341.
Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proceedings of the National Academy of Sciences 98, 676-682.
Rakic, P., 2009. Evolution of the neocortex: a perspective from developmental biology. Nature reviews neuroscience 10, 724-735.
Riaz, A., Asad, M., Alonso, E., Slabaugh, G., 2020. DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. Journal of Neuroscience Methods 335, 108506.
Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., 2019. A deep learning framework for neuroscience. Nature neuroscience 22, 1761-1770.
Rogers, B.P., Katwal, S.B., Morgan, V.L., Asplund, C.L., Gore, J.C., 2010. Functional MRI and multivariate autoregressive models. Magnetic resonance imaging 28, 1058-1065.
Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449-468.
Schölvinck, M.L., Maier, A., Frank, Q.Y., Duyn, J.H., Leopold, D.A., 2010. Neural basis of global resting-state fMRI activity. Proceedings of the National Academy of Sciences 107, 10238-10243.
Seo, Y., Morante, M., Kopsinis, Y., Theodoridis, S., 2019. Unsupervised Pre-training of the Brain Connectivity Dynamic Using Residual D-Net. International Conference on Neural Information Processing. Springer, pp. 608-620.
Shen, H., Wang, L., Liu, Y., Hu, D., 2010. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. Neuroimage 49, 3110-3121.
Shi, J., Wen, H., Zhang, Y., Han, K., Liu, Z., 2018. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. Human brain mapping 39, 2269-2282.
Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., 2009. Correspondence of the brain's functional architecture during activation and rest. Proceedings of the National Academy of Sciences 106, 13040-13045.
Smith, S.M., Miller, K.L., Moeller, S., Xu, J., Auerbach, E.J., Woolrich, M.W., Beckmann, C.F., Jenkinson, M., Andersson, J., Glasser, M.F., 2012. Temporally-independent functional modes of spontaneous brain activity. Proceedings of the National Academy of Sciences 109, 3131-3136.
Sporns, O., Tononi, G., Kötter, R., 2005. The human connectome: a structural description of the human brain. PLoS computational biology 1.
Suk, H.-I., Wee, C.-Y., Lee, S.-W., Shen, D., 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. Neuroimage 129, 292-307.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems, pp. 3104-3112.
Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62-79.
Venkatesh, M., Jaja, J., Pessoa, L., 2019. Brain dynamics and temporal trajectories during task and naturalistic processing. Neuroimage 186, 410-423.
Venkatesh, M., Jaja, J., Pessoa, L., 2020. Comparing functional connectivity matrices: A geometry-

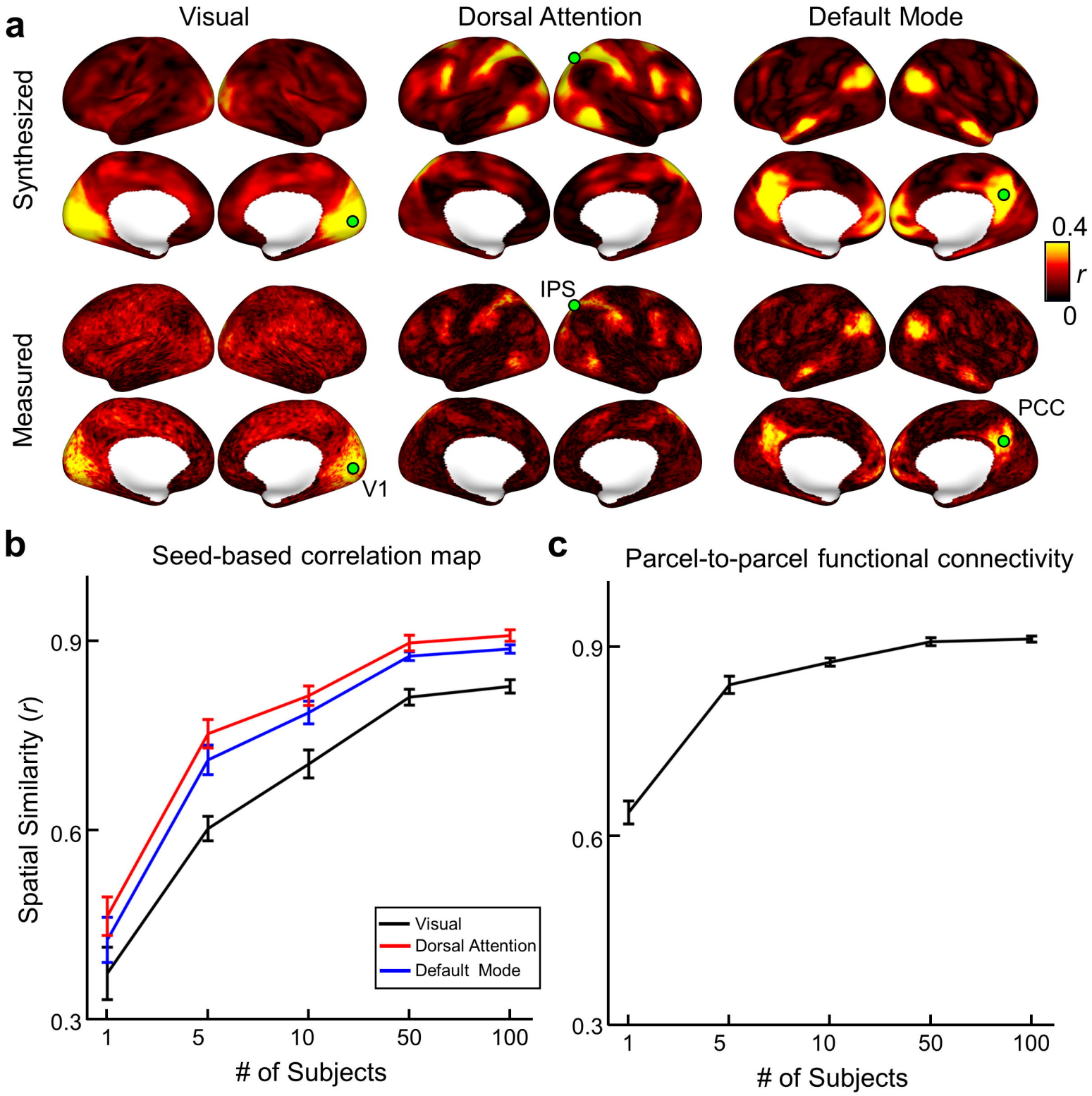
aware approach applied to participant identification. Neuroimage 207, 116398.
Wang, L., Li, K., Chen, X., Hu, X.P., 2019. Application of convolutional recurrent neural network for individual recognition based on resting state fmri data. Frontiers in neuroscience 13, 434.
Wang, X., Liang, X., Jiang, Z., Nguchu, B.A., Zhou, Y., Wang, Y., Wang, H., Li, Y., Zhu, Y., Wu, F., 2020. Decoding and mapping task states of the human brain via deep learning. Human brain mapping 41, 1505-1519.
Wen, H., Liu, Z., 2016. Broadband electrophysiological dynamics contribute to global resting-state fMRI signal. Journal of Neuroscience 36, 6030-6040.
Wen, H., Shi, J., Chen, W., Liu, Z., 2018. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Scientific reports 8, 1-17.
Winder, A.T., Echagarruga, C., Zhang, Q., Drew, P.J., 2017. Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. Nature neuroscience 20, 1761-1769.
Xie, H., Calhoun, V.D., Gonzalez-Castillo, J., Damaraju, E., Miller, R., Bandettini, P.A., Mitra, S., 2018. Whole-brain connectivity dynamics reflect both task-specific and individual-specific modulation: A multitask study. Neuroimage 180, 495-504.
Yamins, D.L., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience 19, 356-365.
Yang, P., Zhou, F., Ni, D., Xu, Y., Chen, S., Wang, T., Lei, B., 2019. Fused sparse network learning for longitudinal analysis of mild cognitive impairment. IEEE transactions on cybernetics.
Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of neurophysiology.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.s.D.N., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856-867.
Zhang, Y., Han, K., Worth, R., Liu, Z., 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. Nature communications 11, 1-13.
Zhao, Q., Honnorat, N., Adeli, E., Pfefferbaum, A., Sullivan, E.V., Pohl, K.M., 2019. Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. International Conference on Information Processing in Medical Imaging. Springer, pp. 867-879.
Zhao, Y., Li, X., Zhang, W., Zhao, S., Makkie, M., Zhang, M., Li, Q., Liu, T., 2018. Modeling 4d fmri data via spatio-temporal convolutional neural networks (st-cnn). International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 181-189.
Zou, L., Zheng, J., Miao, C., Mckeown, M.J., Wang, Z.J., 2017. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. IEEE Access 5, 23626-23636.

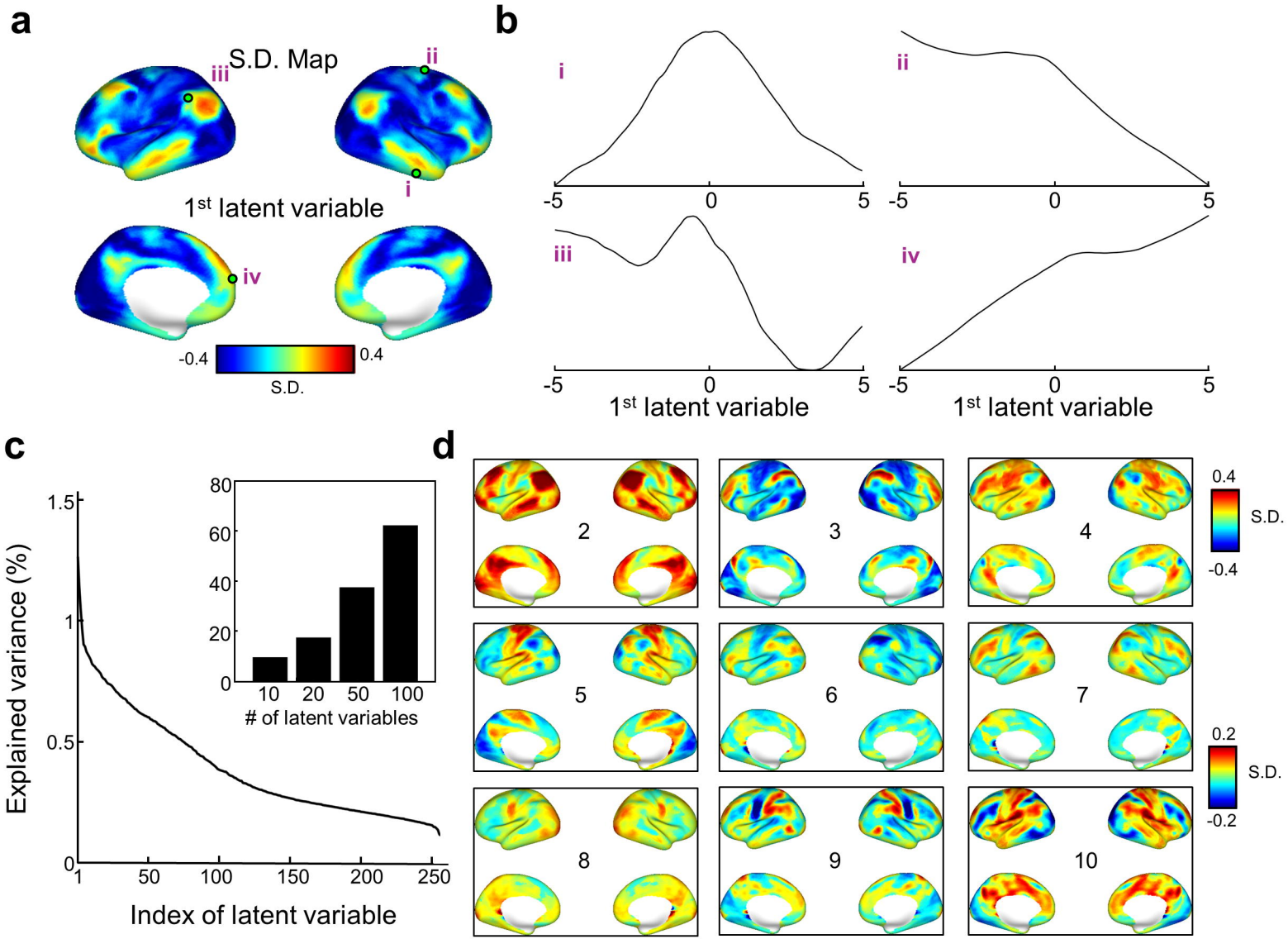

a
z
y
x
e
a
Sin(e)
a
b
Input
Variational Autoencoder
Output
Latent Variables
x
Encoder
z
Decoder
x′
1
2
3
4
5
6
7
8
9
10
11
12
L
R

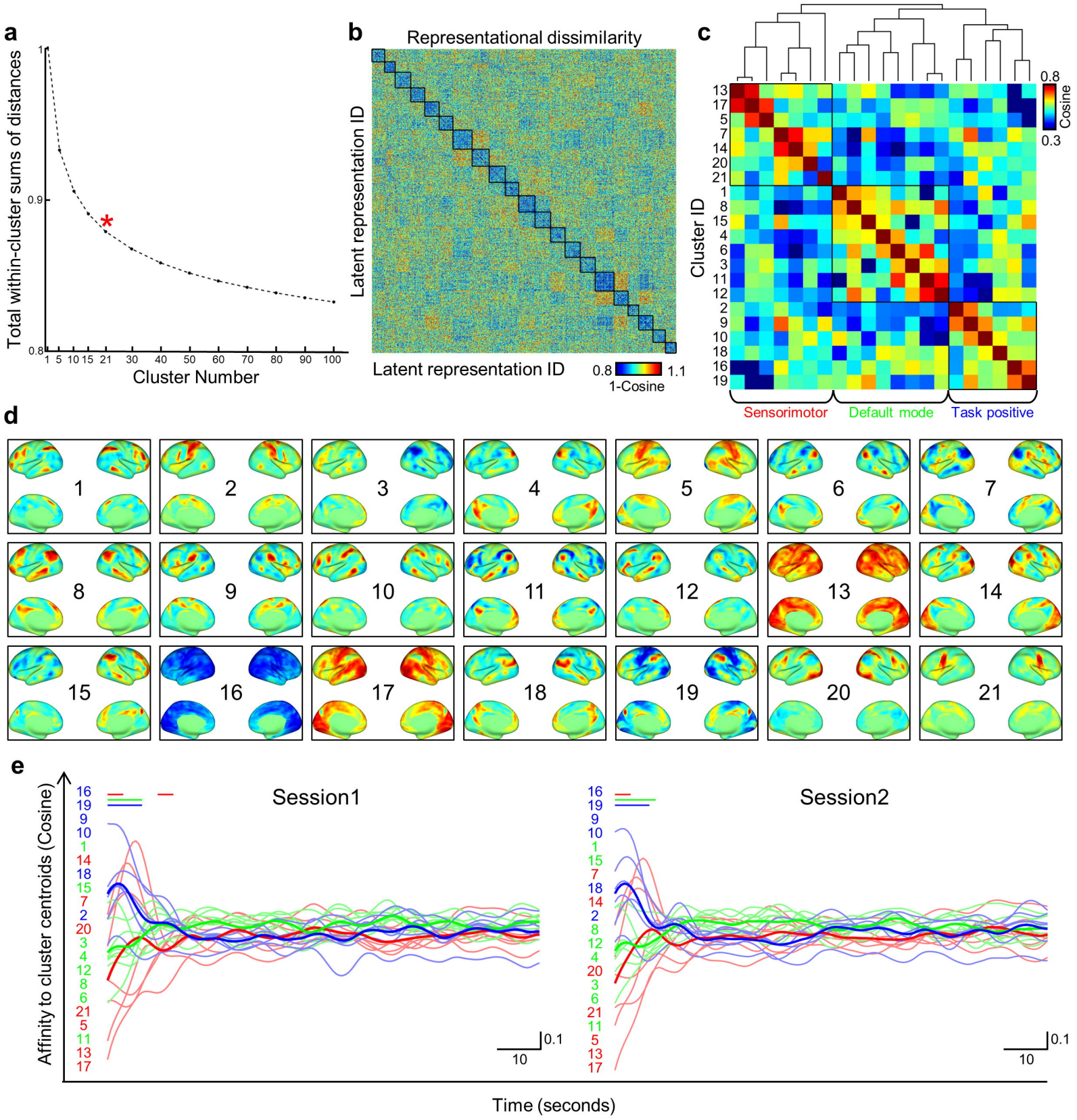

15
10
5
$\beta = 1$
2.0
1.5
1.3
0
0.01
0.02
Reconstruction loss
Kullback–Leibler Divergence

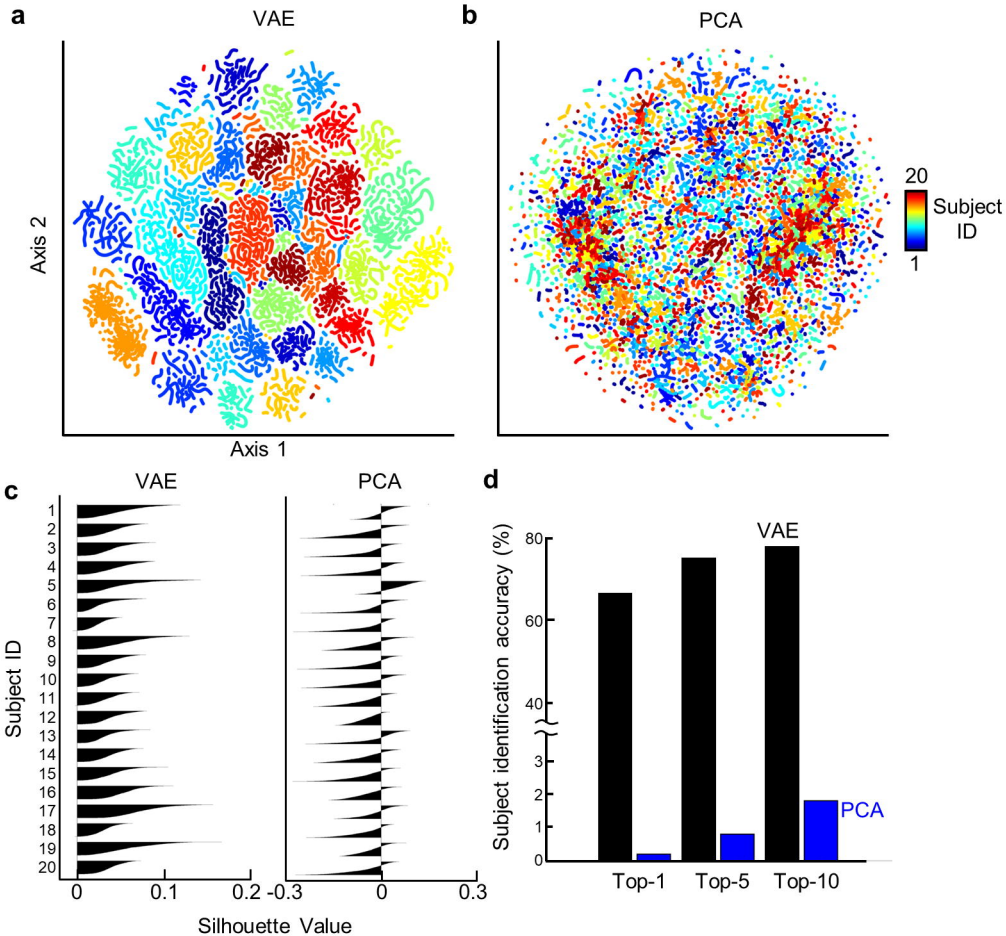

1
2
3
4
5
Original
VAE
1
2
3
4
5
50s
Reconstructed
2
S.D.
-2


a
Original
Smoothing (6 mm)
VAE
PCA
ICA
-2
2
z
b
Reconstruction performance
Explained Variance (%)
80
40
0
1
2
3
4
5
6
7
8
9
10
FWHM (mm)
VAE
PCA
ICA


a
Visual
Dorsal Attention
Default Mode
Synthesized
Measured
IPS
V1
PCC
0.4
r
0
b
Seed-based correlation map
Spatial Similarity (r)
0.9
0.6
0.3
1
5
10
50
100
# of Subjects
Visual
Dorsal Attention
Default Mode
c
Parcel-to-parcel functional connectivity
0.9
0.6
0.3
1
5
10
50
100
# of Subjects


a
S.D. Map
iii
ii
i
1st latent variable
iv
-0.4
0.4
S.D.
b
i
ii
iii
iv
-5
0
5
1st latent variable
c
Explained variance (%)
1.5
1
0.5
0
1
50
100
150
200
250
Index of latent variable
80
60
40
20
0
10
20
50
100
# of latent variables
d
2
3
4
5
6
7
8
9
10
0.4
S.D.
-0.4
0.2
S.D.
-0.2


a
Total within-cluster sums of distances
Cluster Number
b
Representational dissimilarity
Latent representation ID
Latent representation ID
1-Cosine
c
Cluster ID
Cosine
Sensorimotor
Default mode
Task positive
d
e
Session1
Session2
Affinity to cluster centroids (Cosine)
Time (seconds)


a
VAE
Axis 2
Axis 1
b
PCA
20
Subject
ID
1
c
VAE
PCA
Subject ID
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
0.1
0.2
-0.3
0
0.3
Silhouette Value
d
VAE
PCA
Subject identification accuracy (%)
80
60
40
3
2
1
0
Top-1
Top-5
Top-10

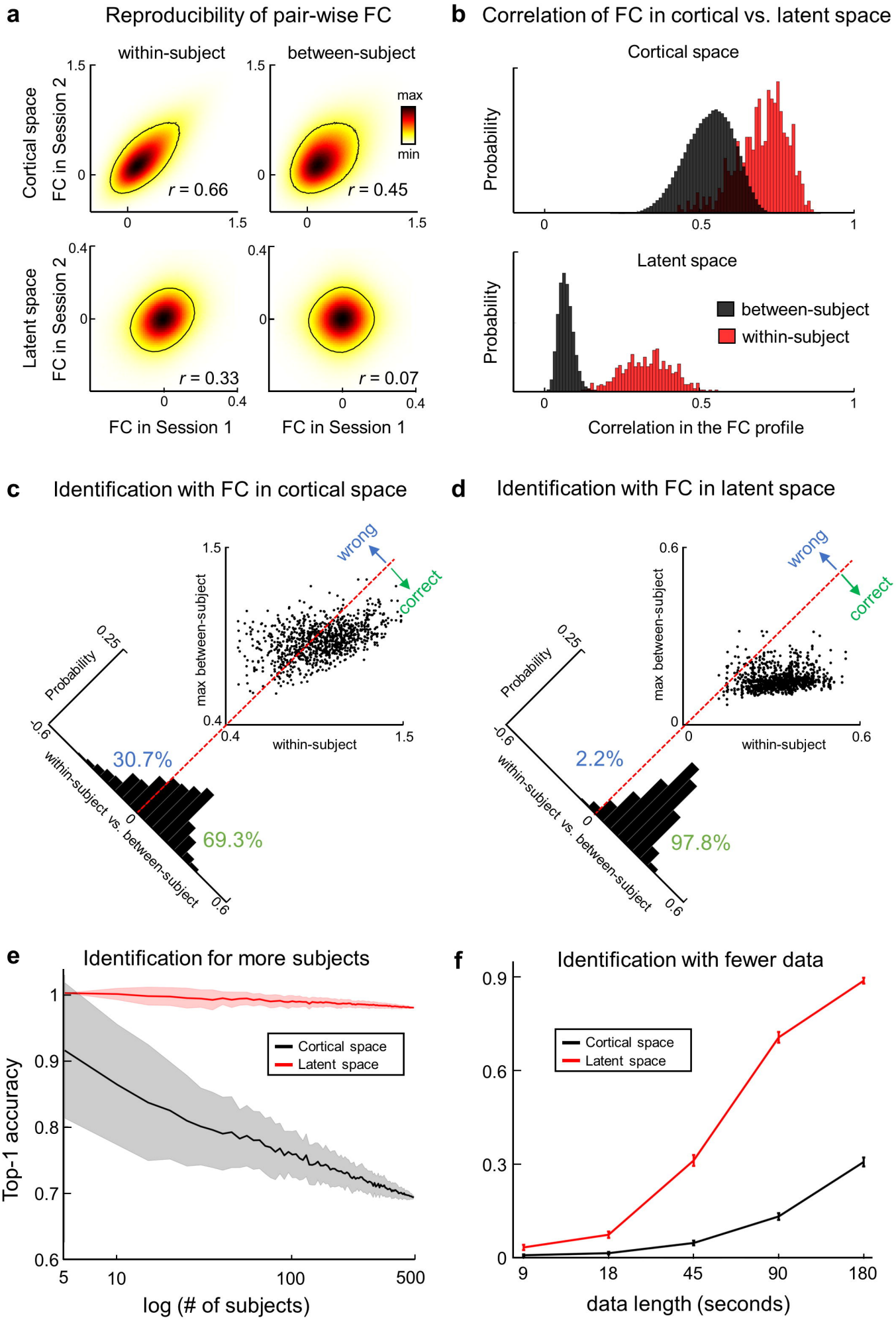
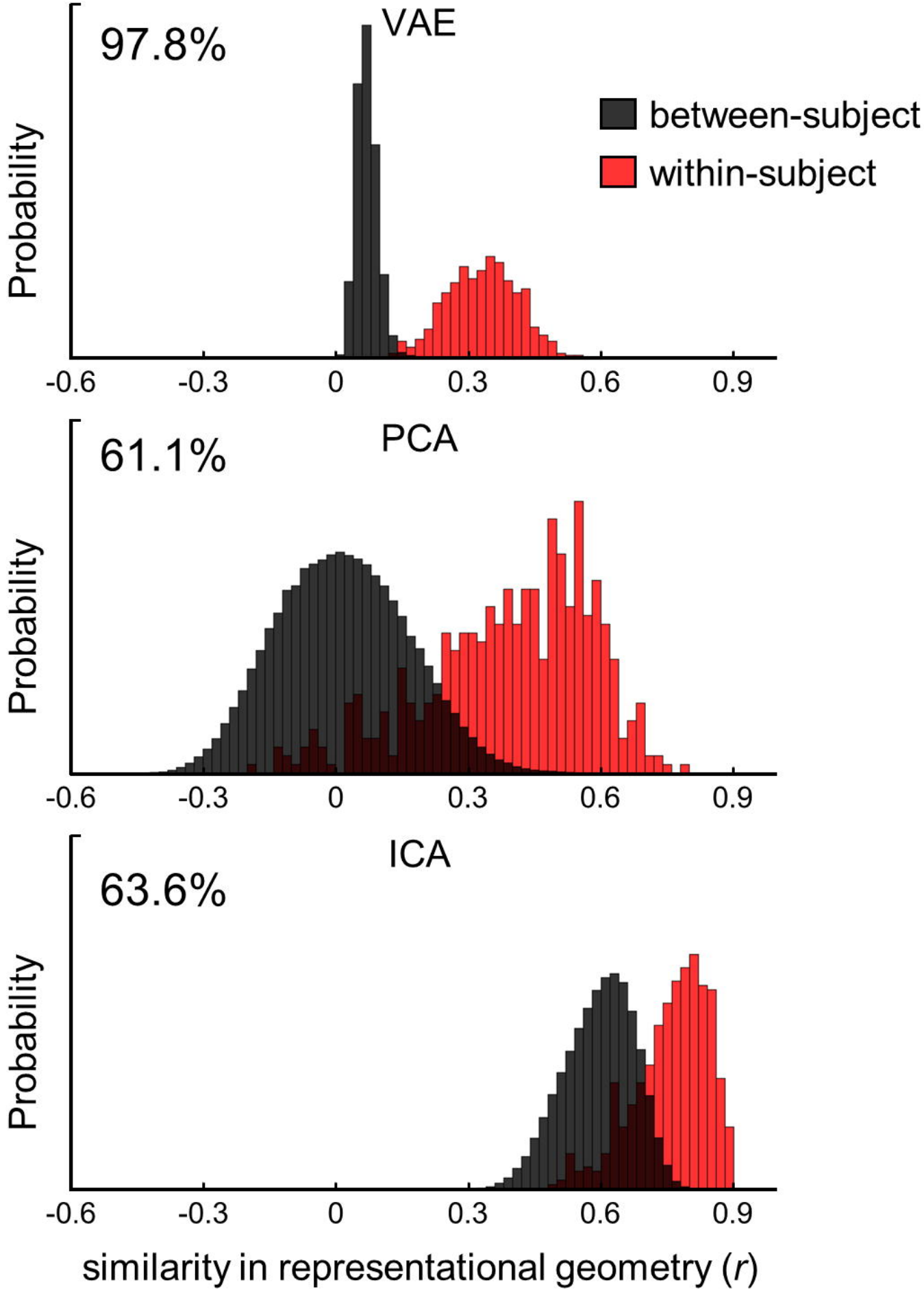
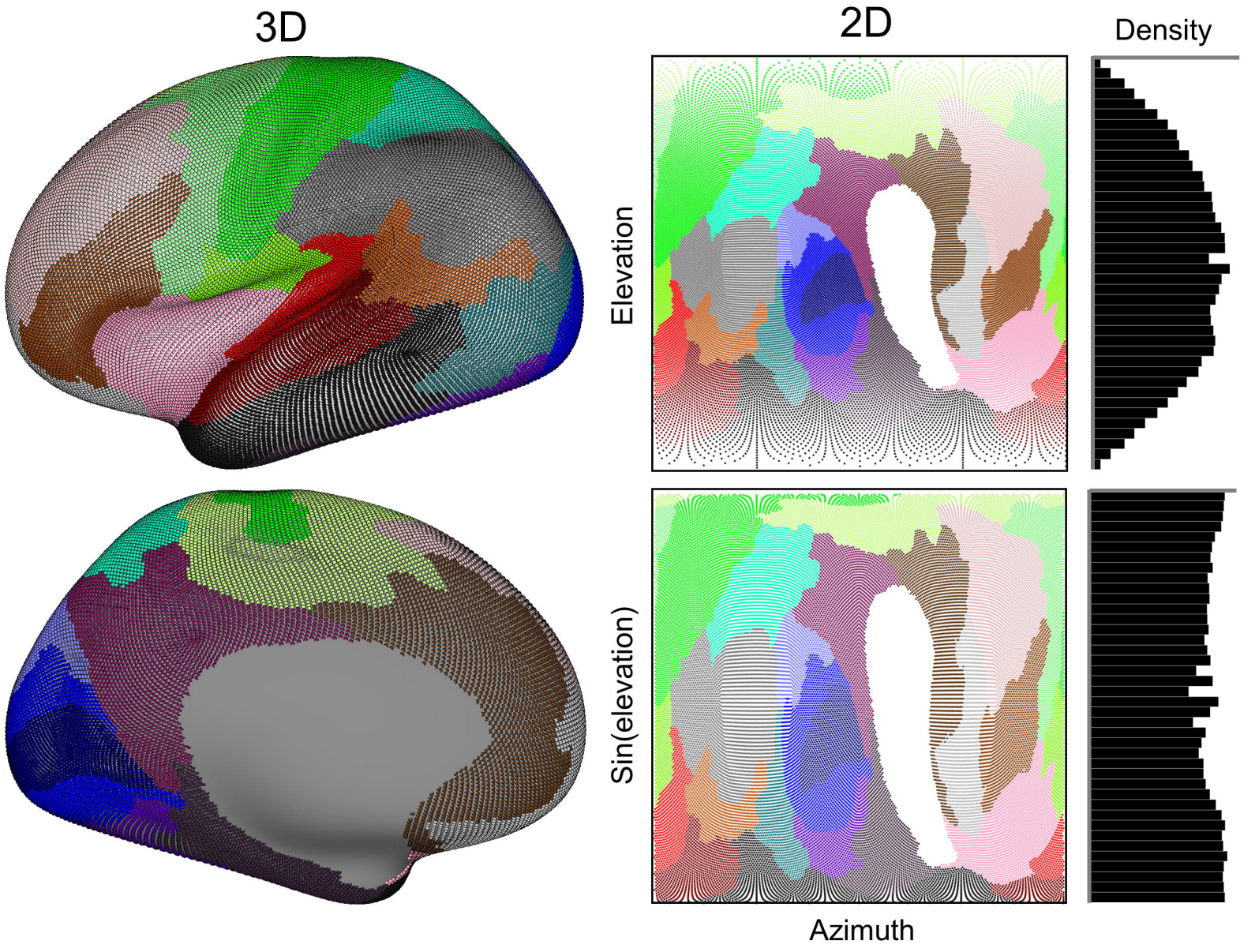
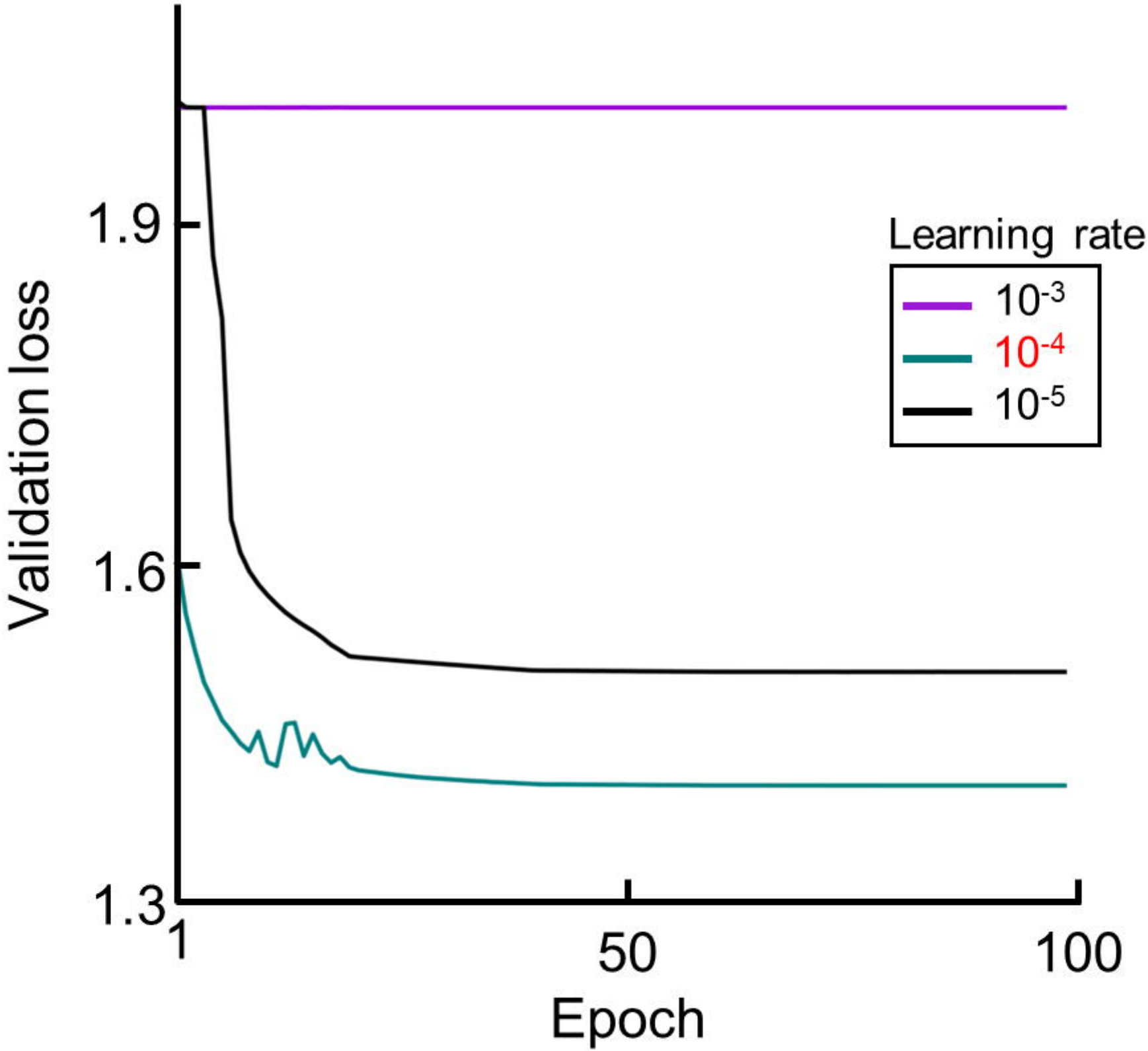

a
Reproducibility of pair-wise FC
within-subject
between-subject
Cortical space
FC in Session 2
Latent space
FC in Session 2
max
min
r = 0.66
r = 0.45
r = 0.33
r = 0.07
FC in Session 1
FC in Session 1
b
Correlation of FC in cortical vs. latent space
Cortical space
Latent space
Probability
between-subject
within-subject
Correlation in the FC profile
c
Identification with FC in cortical space
wrong
correct
max between-subject
within-subject
Probability
within-subject vs. between-subject
30.7%
69.3%
d
Identification with FC in latent space
wrong
correct
max between-subject
within-subject
Probability
within-subject vs. between-subject
2.2%
97.8%
e
Identification for more subjects
Cortical space
Latent space
Top-1 accuracy
log (# of subjects)
f
Identification with fewer data
Cortical space
Latent space
data length (seconds)

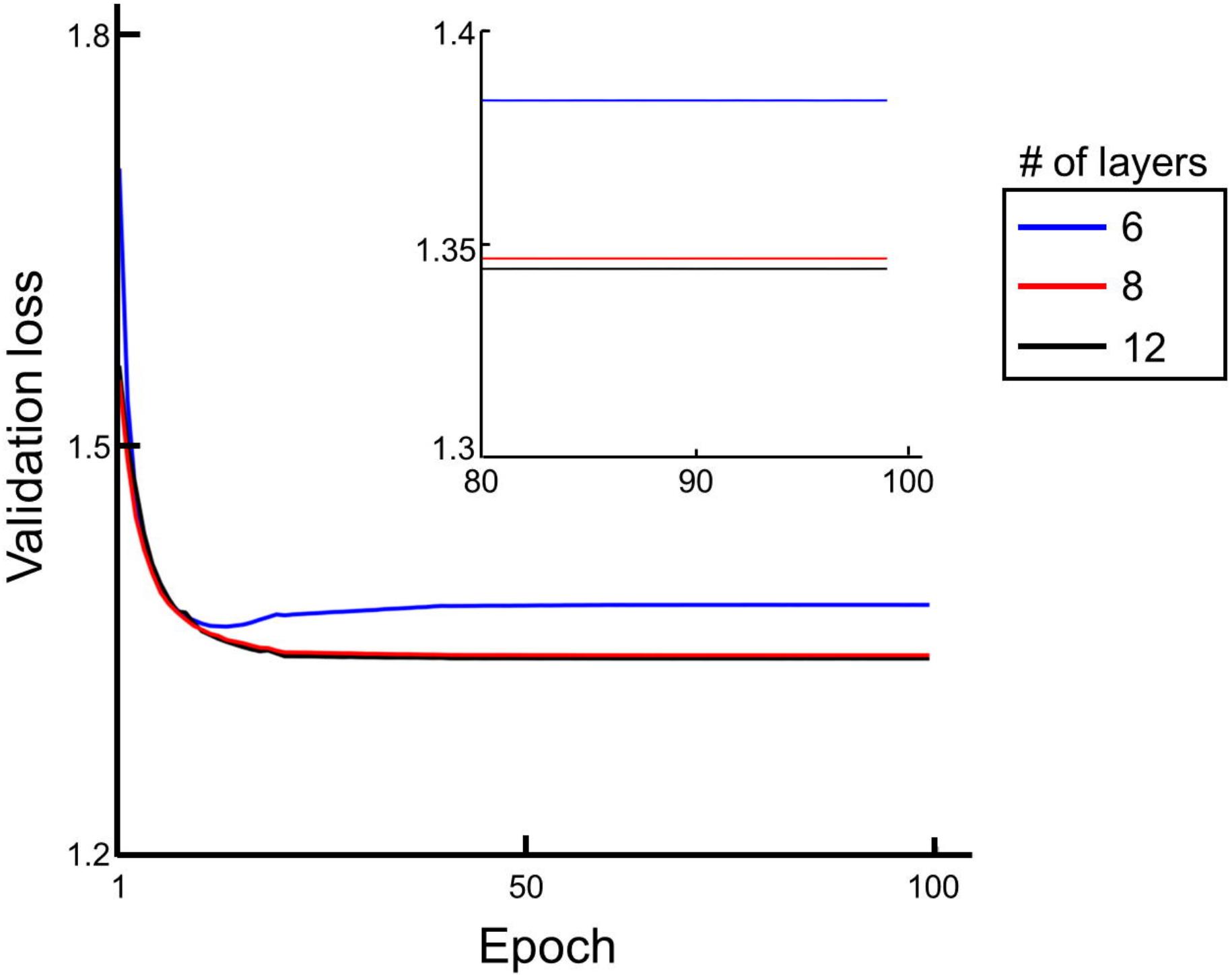
Table 1. Subject Identification Accuracy across different conditions

			Session 1	
			Clean	Noisy
Session 2	Latent Space	Clean	97.8%	87.9%
		Noisy	88.4%	91.3%
	Cortical Space	Clean	69.3%	47.2%
		Noisy	47.5%	76.9%


97.8%
VAE
between-subject
within-subject
Probability
-0.6
-0.3
0
0.3
0.6
0.9
61.1%
PCA
Probability
-0.6
-0.3
0
0.3
0.6
0.9
63.6%
ICA
Probability
-0.6
-0.3
0
0.3
0.6
0.9
similarity in representational geometry (r)


3D
2D
Density
Elevation
Sin(elevation)
Azimuth


Validation loss
1.9
1.6
1.3
1
50
100
Epoch
Learning rate
10-3
10-4
10-5


Validation loss
1.8
1.5
1.2
1
50
100
Epoch
1.4
1.35
1.3
80
90
100
# of layers
6
8
12