Representation Learning of Resting State fMRI with Variational Autoencoder

Jung-Hoon Kim[1,3], Yizhen Zhang[2], Kuan Han[2], Minkyu Choi[2], Zhongming Liu[1,2,3,4*]

[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

[4]School of Electrical and Computer 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 (rs-fMRI) data exhibits complex but structured patterns. However, the underlying origins are unclear and entangled in rs-fMRI data. Here we establish a variational auto-encoder, as a generative model trainable with unsupervised learning, to disentangle the unknown sources of rs-fMRI 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. Of the latent representation, its distribution reveals overlapping functional networks, and its geometry is unique to each individual. Our results support the functional opposition between the default mode network and the task-positive network, while such opposition is asymmetric and non-stationary. Correlations 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 subject.

INTRODUCTION

The brain is active even at rest, showing complex activity patterns measurable with resting state fMRI (rs-fMRI)[1]. It is widely recognized that rs-fMRI activity is shaped by how the brain is wired, or the brain connectome[2]. Inter-regional correlations of rs-fMRI activity are often used to report functional connectivity[3] and map brain networks for individuals[4] or populations in various behavioral[5] or disease states[6]. However, it remains largely unclear where rs-fMRI activity comes from[7, 8], whereas understanding the underlying origins is critical to interpretation of any rs-fMRI pattern or dynamics[9].

Prior findings suggest a multitude of sources (or causes) for rs-fMRI activity[10], including but not limited to fluctuations in neurophysiology[11], arousal[12], unconstrained cognition[13], non-neuronal physiology[14], head motion[15] etc. These sources only partially account for rs-fMRI 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 the task-free resting state[7]. An inclusive study would benefit from using a data-driven approach to uncover and disentangle all plausible but hidden sources from rs-fMRI data itself, without having to presume the sources to whatever are accessible for empirical observations. To be effective, such an approach should be able to infer sources from rs-fMRI data and generate new rs-fMRI 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[16]. In addition to its success in artificial intelligence, deep learning has also been increasingly applied to brain research[17]. Despite its great potential[18-20], deep learning applied to resting state fMRI analysis has arguably limited progress relative to what is attainable with conventional and simpler methods[21]. A challenge is inherent to the absence of any task in the resting state as well as the lack of sufficient knowledge usable for training deep neural networks with supervised learning.

To mitigate this challenge, we chose to use Variational Auto-Encoder (VAE)[22, 23], a type of deep learning model, for unsupervised learning of the ever-increasing "big data" in rs-fMRI. Briefly, we designed and trained a VAE model to represent rs-fMRI data in terms of its hidden (or latent) sources and tested its ability to explain and generate rs-fMRI data. We also explored the functional organization of rs-fMRI data in the latent space to reveal network interactions in the brain. Lastly, we tested the utility of this model for identifying individuals from their rs-fMRI data[4], as a starting example of its applications.

Results

VAE compressed rs-fMRI maps

Inspired by its success in artificial intelligence[22, 23], we designed a VAE model in order to disentangle the generative factors underlying rs-fMRI activity. The model used a pair of convolutional and deconvolutional neural networks in an encoder-decoder architecture (Figure 1.b). The encoder transformed any rs-fMRI pattern, formatted as an image on a regular 2D grid (Figure 1.a), to the probability distributions of 256 independent 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)[24], we first trained the model with rs-fMRI maps from 100 subjects and then tested it with rs-fMRI data from 500 other subjects.

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 1.c). Such compression resulted in spatial blurring comparable to the effect of spatial smoothing with 4mm full width at half maximum or the effect of linear dimension reduction with principal component analysis (Supplementary Figure 1). As such, the latent representation obtained with VAE preserved the spatiotemporal characteristics of rs-fMRI, despite modest but acceptable loss in spatial resolution and specificity.

VAE synthesized correlated fMRI activity

We asked whether the decoder in the VAE, as a generative model, could have learned the putative mechanisms by which rs-fMRI activity patterns arise presumably 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 rs-fMRI maps. We calculated the seed-based correlations[3] by using the VAE-synthesized data and compared the results with those obtained with length-matched rs-fMRI data concatenated across 10 subjects. Figure 2 shows three examples with the seed region in the primary visual cortex (V1), intraparietal sulcus (IPS), or posterior cingulate cortex (PCC). Both the synthesized and measured data gave rise to similar network patterns (mean±std of z-transformed spatial correlation z = 0.81±0.08, 0.97±0.07, or 0.88±0.05), consistent with early visual network, dorsal attention network, and default mode network reported in prior studies (e.g. by Yeo et al.[25]). Thus, the VAE provided a computational account for the generative process of resting state activity and could synthesize realistic rs-fMRI activity patterns and preserve inter-regional correlations as are observable in experiments.

Clusters in latent space

We further explored the utility of VAE for data-driven discovery of brain networks. We used the VAE to encode the rs-fMRI pattern observed at every time point from 500 subjects, clustered the time points by applying k-means clustering (k=21) to the low-dimensional latent representations, and decoded the cluster centroids to corresponding cortical maps. Each of the resulting maps represented a characteristic pattern of network interaction (see all 21 maps in Supplementary Figure 2).

Among the 21 clusters, 5 clusters (Cluster 5, 6, 8, 16, 19) showed activity increase (positive) at one or multiple regions in the default mode network[26-28], alongside activity decrease (negative) at other regions (Figure 3.a). Both the positive and negative regions showed a varying degree of overlapping across the 5 clusters. The overlapping positivity highlighted the default mode network and revealed sub-divisions of its constituent regions[29]. The overlapping negativity showed the networks presumably involved in attention[30], cognitive or executive control[31-33]. Similarly, we found 5 clusters with activity increase in the so-called frontoparietal control network[31] (Cluster 10), cingulo-opercular network[33] (Cluster 4 and 14), cognitive control network[32] (Cluster 17), and dorsal attention networks[34] (Cluster 1) – collectively referred to as “the task positive network”[35] hereafter (Figure 3.b). These 5 clusters were partially overlapping with respect to their positive regions but varied from one another with respect to their negative regions, while some of them showed either no or little activity decrease. The overlapping positivity and negativity showed strong co-activation of the task positive network alongside weak deactivation of the default mode network. These results indicate patterns of opposition between the default mode network and the task positive network, conceptually similar to the notion of “anti-correlation”[35]. Interestingly, the opposition was asymmetric, being more pronounced when activity increases in the default mode network, but much weakened when activity increases in the task positive network.

In addition, the other clusters were also informative (Supplementary Figure 2). To name a few examples, Cluster 21 showed activity decrease in the whole brain, thereby a signature of global signal fluctuation. Cluster 13 and 15 showed widespread synchrony across sensory systems. Cluster 7 and 9 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 2 and 20 were unilateral to the right and left prefrontal cortex, respectively. Common to many clusters was the fact 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. These results demonstrate that VAE enables data-driven discovery of overlapping and interacting networks for functional integration, as opposed to networks that limit themselves to anatomical and functional segregation.

Individual identification

We further asked whether functional connectivity (FC) in the latent space could be used as a feature or “fingerprint” for identifying individuals in a population[4, 36]. We calculated the correlation between every pair of latent variables, assembled the pair-wise FC into a FC profile, and evaluated its similarity between two separate sessions within or between subjects. For comparison, we performed similar analyses by evaluating FC between 360 cortical areas in an existing atlas[37]. As shown in Figure 4.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 (z=0.00±0.14) and its reproducibility was high only within the same subject (r=0.32) but not between different subjects (r=0.08). Although less reproducible, the FC profile was more distinctive across subjects when it was evaluated between latent variables rather than cortical areas (Figure 4.b). In the latent space, the FC profile was significantly more consistent within a subject than between subjects (two-sample t-test, t(249,998)=235.81, two-sided p<0.001). The distribution of within-subject correlations was in nearly complete separation from that of between-subject correlations (Figure 4.b, bottom).

Then we compared the performance of individual identification on the basis of the FC profile in the latent vs. cortical space. To identify 1 out of 500 subjects, we compared a target subject's FC profile in the 1st session with every subject's FC profile in the 2nd session and chose the best match in terms of Pearson correlation coefficient. As such, the choice was correct if the correlation with the target subject was higher than the largest correlation with any non-target subject. We found that the FC profile in the cortical space could support 69.3% top-1 accuracy while identification was often done with marginal confidence relative to the decision boundary (Figure 4.c). Using the FC in the latent space allowed us to reach 97.5% top-1 accuracy. The evidence for correct identification was apparent with a large margin from the decision boundary (Figure 4.d). Moreover, the use of FC in the latent space supported reliable and robust performance in top-1 identification given an increasingly larger population (Figure 4.e) or when the data were limited to a short duration (Figure 4.f), being notably superior to the use of FC in the cortical space.

Discussion

Here, we present a method for unsupervised representation learning of cortical rs-fMRI activity. Our results suggest that this method is able to disentangle generative factors underlying spontaneous brain activity, discover overlapping brain networks with opposing or associated functions, and capture individual characteristics or variation. 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[22, 23] (without any label), which is appealing for learning representations of rs-fMRI data. Since rs-fMRI 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 rs-fMRI data[24]. 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[38, 39].

The method herein can be extended in multiple ways. Although it is trained with rs-fMRI data, we hypothesize that the VAE model can encode and decode both rs-fMRI 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[40]. The fact that the VAE can synthesize new data (Figure 2) 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[41, 42] and correction of head motion[15]. It also supports the notion that the learned latent space captures the origins of rs-fMRI and the VAE decoder captures the computational account for how rs-fMRI 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[43]. 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 rs-fMRI 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[19, 44], as an add-on to VAE, for sequence learning based on spatial representations extracted from individual time points.

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[45, 46]. However, existing network analyses still tend to group brain regions into non-overlapping networks[25]. 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[45] but allows for nonlinear relationships between latent variables and the input data they represent[47]. Arguably, finding clusters in the low-dimensional latent space is more desirable than doing so in the higher-dimensional voxel space[48]. Not only is it more computationally efficient, but data representations are also more disentangled in the latent space than in the voxel space to readily reveal the underlying organization, as discussed later.

Clusters in the latent space do not manifest themselves as resting state networks[25] per se but highlight interactions among those networks. Many of the clusters cover more regions and/or reveal finer divisions within regions than are commonly observed in resting state networks (Figure 3). In each cluster, the interactions among its constituent regions should not be interpreted pairwise (e.g. correlation) but as two multivariate modes: co-activation and co-deactivation, which we interpret as the signatures of functional association and opposition, respectively.

Our results suggest the functional opposition between regions in the default mode network and those in cognitive control networks. This finding agrees with the prior finding that attention demanding tasks tend to increase activity in cognitive control networks (also referred to as the task positive network[35]) and decrease activity in the default mode network[26]. It may sound a reminiscence of the anti-correlation between the task positive network and the default mode network[35]. However, the anti-correlation is controversial and confounded by global signal regression[49] – a questionable preprocessing step that causes spuriously negative correlations[50]. Note that global signal regression was not used and thereby not of concern in this study. Our finding provided complementary evidence, supporting a similar but revised view as anti-correlation[35]. We conclude that the functional opposition between the default mode network and the task positive network is indeed real but non-stationary[41, 46]. It occurs at some but not all times. It is also asymmetric in that activity increase in the default mode network tends to co-occur with activity decrease in the task positive network, whereas activity increase in the task positive network unnecessarily or less frequently co-occurs with activity decrease in the default mode network. Interestingly, the global signal fluctuation is also non-stationary and identifiable as a different cluster in the latent space. Together, the functional opposition and the global signal are separable in time; therefore, the latter does not necessarily invalidate or confound the former.

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 describes the correlations among the disentangled sources of resting state activity. This is a new perspective different from the functional connectivity among observable voxels, regions or networks[3, 25]. If the VAE model has fully disentangled the sources in a population level, functional connectivity should be near zero between different latent variables. In other words, the model sets a nearly null baseline such that the latent-space functional connectivity primarily reflects features unique to individuals. Supporting this notion, our results suggest the use of functional connectivity in the latent space leads to a significantly improved accuracy, robustness, and efficiency in individual identification, compared to the use of functional connectivity among cortical parcels[4, 36]. 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.

Methods

Data

We used rs-fMRI data from 602 healthy subjects randomly chosen from the Q2 release by HCP[24]. For each subject, we used two sessions of rs-fMRI data acquired from different days with either right-to-left or left-to-right phase encoding. Each session included 1,200 time points separated by 0.72s. Following minimal preprocessing[51], 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, 2, or 500 subjects for training, validating, or testing the VAE model, respectively.

Geometric reformatting

We converted the rs-fMRI data from 3-D cortical surfaces to 2-D grids in order to structure the rs-fMRI 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[52]. 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) reporting the azimuth and elevation angles in a range from $-\pi$ to π and from $-\pi/2$ to $\pi/2$, respectively. We defined a 192×192 grid to resample the spherical surface with respect to azimuth and *sin*(elevation) such that the sampled locations were uniformly distributed at approximation (Supplementary Figure 3). We used the nearest-neighbor interpolation to convert data from the 3-D surface to the 2-D grid, and vice versa.

Variational autoencoder

We designed a β-VAE model[23], a variation of VAE[22], to learn representations of rs-fMRI 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. The decoder sampled the latent distribution to reconstruct the input fMRI map or generate a new map. The encoder stacked five convolutional layers and one fully connected layer. Every convolutional layer applied linear convolution and rectified its output[53]. The 1st layer applied 8×8 convolution separately to the input from each hemisphere and concatenated its output. The 2nd through 5th layers applied 4×4 convolution. 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 to the input space. See Figure 1.b for more details about the 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, $\boldsymbol{\phi}$, and the decoding parameters, $\boldsymbol{\theta}$, to minimize the loss function as below.

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

where $\boldsymbol{x}$ is the input data combined across the left and right hemispheres, $\boldsymbol{x}'$ is the corresponding output from the model, $\mathcal{N}(\boldsymbol{\mu_z},\boldsymbol{\sigma_z})$ is the posterior normal distribution of the latent variables, $\boldsymbol{z}$, with their mean and standard deviation denoted as $\boldsymbol{\mu_z}$ and $\boldsymbol{\sigma_z}$, $\mathcal{N}(\mathbf{0},\mathbf{I})$ is an independent and standard normal distribution as the prior distribution of the latent variables, D_{KL} measures the Kullback-Leibler divergence between the posterior and prior distributions, and β is the hyperparameter balancing the two terms in the loss function. We optimized the model by using stochastic gradient descent (batch size=128, learning rate=10^{-5}, and 500 epochs) and Adam optimizer[54] implemented in PyTorch (v*1.2.0*). We explored four values (1, 2, 5, 10) for β and chose $\beta = 5$ to disentangle the latent variables while minimizing the loss function in training and validation (Supplementary Figure 4).

Synthesizing rs-fMRI functional connectivity

We used the trained VAE to synthesize rs-fMRI data from random samples of latent variables. To synthesize a vector in the latent space, we drew a random sample of every latent variable independently from a standard normal distribution. The synthesized vector passed through the decoder in VAE, generating a cortical pattern. Repeating this process, we synthesized 12,000 cortical patterns as data used for seed-based correlation analysis. As examples, we explored three seed locations within V1, IPS, and PCC and calculated the functional connectivity to each seed based on the Pearson correlation coefficient. The MNI coordinates of the seed in V1, IPS, and PCC were (7, -83, 2), (26, -66, 48), and (0, 57, 27), respectively[55]. For comparison, we evaluated seed-based correlations with length-matched experimental rs-fMRI data concatenated across 10 subjects in HCP. We evaluated the reproducibility of the results by repeating the above analysis 20 times with different synthesized data and the experimental data from different subsets of subjects.

Clustering in the latent space

We encoded the rs-fMRI spatial pattern at every time point for 500 testing subjects, yielding 600,000 vectors in the latent space. We used k-means clustering (with Euclidean distance) to group those vectors to 21 clusters. The choice of k=21 was made empirically in part to be consistent to a prior study with a similar motivation[45] and in part to fall within the range of the number of resting state networks reported in literature. For each of the 21 clusters, the cluster centroid was calculated and converted to a corresponding cortical pattern by using the VAE's decoder; the resulting cortical pattern was scaled such that its maximal absolute value equaled 1.

To evaluate the spatial overlap among clusters, we thresholded the cortical pattern resulting from each cluster by >0.35 (for positivity) or <-0.35 (for negativity). For clusters relevant to the default mode network (5, 19, 8, 6, 16) or the task positive network (17, 1, 14, 4, 10), we calculated the overlapping positivity (or negativity) by counting the number of times that each cortical location was over (or below) 0.35 (or -0.35)

Individual identification

In the testing data set, every individual had rs-fMRI data acquired for two separate sessions. For each session, we encoded the data as (256×1,200) latent representations, calculated the z-transformed correlation between every pair of latent variables, and stored the z-values into a vector, referred to as the FC profile in the latent space.

We tested the utility of this FC profile as the feature for identifying individuals in a population (n=500). For every subject, we used the FC profile collected in one session as the subject-identifying key in a database. Given this database, we tested the accuracy of retrieving any subject's identity by using a query based on the subject's FC profile in the other session. To retrieve the identity, we compared the query to every key to find the best match in terms of the highest correlation. We evaluated the identification accuracy as the percentage by which the correct identity was retrieved. Since we could use either session 1 or session 2 for the key while using the other for the query, we tested both cases and averaged the identification accuracy.

For comparison, we also evaluated the functional connectivity between every pair of 360 cortical parcels defined in an established atlas[37]. Similarly, we used the FC profile in the cortical space as the feature for individual identification and compared the resulting identification accuracy with that based on the FC profile in the latent space. We repeated this comparative evaluation with a varying population size (from n=5 to 500) or a varying length of data (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.

Reference

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

2. Sporns, O., Tononi, G. & Kötter, R. The human connectome: a structural description of the human brain. *PLoS computational biology* **1** (2005).

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

4. Finn, E.S.*, et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience* **18**, 1664 (2015).

5. Smith, S.M.*, et al.* Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences* **106**, 13040-13045 (2009).

6. Fox, M.D.*, et al.* 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 (2014).

7. Leopold, D.A. & Maier, A. Ongoing physiological processes in the cerebral cortex. *Neuroimage* **62**, 2190-2200 (2012).

8. Lu, H., Jaime, S. & Yang, Y. Origins of the resting-state functional MRI signal: potential limitations of the “neurocentric” model. *Frontiers in neuroscience* **13** (2019).

9. Winder, A.T., Echagarruga, C., Zhang, Q. & Drew, P.J. Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. *Nature neuroscience* **20**, 1761-1769 (2017).

10. Bianciardi, M.*, et al.* Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. *Magnetic resonance imaging* **27**, 1019-1029 (2009).

11. Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L. & Corbetta, M. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences* **104**, 13170-13175 (2007).

12. Chang, C.*, et al.* Tracking brain arousal fluctuations with fMRI. *Proceedings of*

the National Academy of Sciences **113**, 4518-4523 (2016).

13. Chou, Y.-h.*, et al.* Maintenance and representation of mind wandering during Resting-State fMRI. *Scientific reports* **7**, 40722 (2017).

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

15. Power, J.D.*, et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320-341 (2014).

16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436-444 (2015).

17. Richards, B.A.*, et al.* A deep learning framework for neuroscience. *Nature neuroscience* **22**, 1761-1770 (2019).

18. Suk, H.-I., Wee, C.-Y., Lee, S.-W. & Shen, D. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage* **129**, 292-307 (2016).

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

20. Plis, S.M.*, et al.* Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience* **8**, 229 (2014).

21. He, T.*, et al.* Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* **206**, 116276 (2020).

22. Kingma, D.P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

23. Higgins, I.*, et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr* **2**, 6 (2017).

24. Van Essen, D.C.*, et al.* The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62-79 (2013).

25. Thomas Yeo, B.*, et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* **106**, 1125-1165 (2011).

26. Raichle, M.E.*, et al.* A default mode of brain function. *Proceedings of the National Academy of Sciences* **98**, 676-682 (2001).

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

28. Greicius, M.D., Krasnow, B., Reiss, A.L. & Menon, V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences* **100**, 253-258 (2003).

29. Andrews-Hanna, J.R., Reidler, J.S., Sepulcre, J., Poulin, R. & Buckner, R.L. Functional-anatomic fractionation of the brain's default network. *Neuron* **65**, 550-562 (2010).

30. Corbetta, M. & Shulman, G.L. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience* **3**, 201-215 (2002).

31. Dixon, M.L.*, et al.* 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 (2018).

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

33. Dosenbach, N.U.*, et al.* Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences* **104**, 11073-11078 (2007).

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

35. Fox, M.D.*, et al.* The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences* **102**, 9673-9678 (2005).

36. Venkatesh, M., Jaja, J. & Pessoa, L. Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *NeuroImage* **207**, 116398 (2020).

37. Glasser, M.F.*, et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171-178 (2016).

38. Garrity, A.G.*, et al.* Aberrant “default mode” functional connectivity in schizophrenia. *American journal of psychiatry* **164**, 450-457 (2007).

39. Zhang, D., *et al.* Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**, 856-867 (2011).

40. Gonzalez-Castillo, J., *et al.* Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Sciences* **112**, 8762-8767 (2015).

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

42. Allen, E.A., *et al.* Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex* **24**, 663-676 (2014).

43. Rakic, P. Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews Neuroscience* **10**, 724-735 (2009).

44. Sutskever, I., Vinyals, O. & Le, Q.V. Sequence to sequence learning with neural networks. in *Advances in neural information processing systems* 3104-3112 (2014).

45. Smith, S.M., *et al.* Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences* **109**, 3131-3136 (2012).

46. Liu, X. & Duyn, J.H. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences* **110**, 4392-4397 (2013).

47. Khemakhem, I., Kingma, D.P. & Hyvärinen, A. Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809* (2019).

48. Liu, X., Chang, C. & Duyn, J.H. Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns. *Frontiers in systems neuroscience* **7**, 101 (2013).

49. Liu, T.T., Nalci, A. & Falahpour, M. The global signal in fMRI: Nuisance or Information? *Neuroimage* **150**, 213-229 (2017).

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

51. Glasser, M.F., *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105-124 (2013).

52. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774-781 (2012).

53. Nair, V. & Hinton, G.E. Rectified linear units improve restricted boltzmann

machines. in *Proceedings of the 27th international conference on machine learning (ICML-10)* 807-814 (2010).

54. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

55. Jarrett, C. The restless brain. *The Psychologist* (2009).

Figures

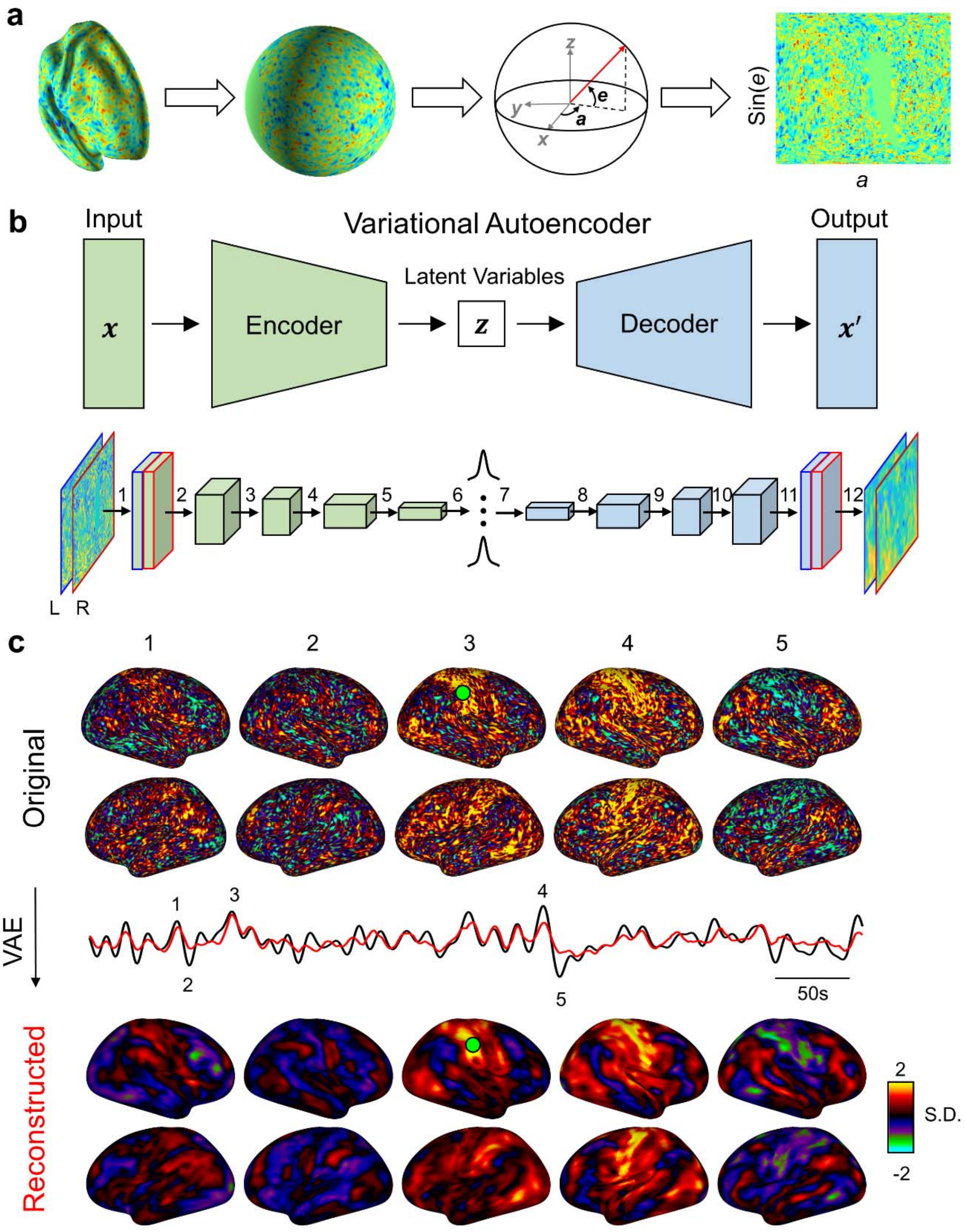

a
z
y
x
e
a
Sin(e)
a
b
Input
Variational Autoencoder
Output
x
Encoder
Latent Variables
z
Decoder
x′
1
2
3
4
5
6
7
8
9
10
11
12
L
R
c
1
2
3
4
5
Original
VAE
1
2
3
4
5
50s
Reconstructed
2
S.D.
-2

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 convolutional layer (numbered from 1 to 5) outputs a feature map with the size of 96x96x64, 48x48x128, 24x24x128, 12x12x256, or 6x6x256, respectively. In the decoder, each convolutional 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). Blue and red boundaries highlight the input/out images for the left and right hemispheres, respectively. **(c) Reconstruction of rs-fMRI.** For a typical rs-fMRI dataset, the activity patterns observed are shown in the top and their reconstructions through VAE are shown in the bottom. The observed and reconstructed patterns correspond to 5 time points as shown in the voxel time series from the intra-parietal sulcus. The time series of the observed and reconstructed activity are shown in black and red, respectively.

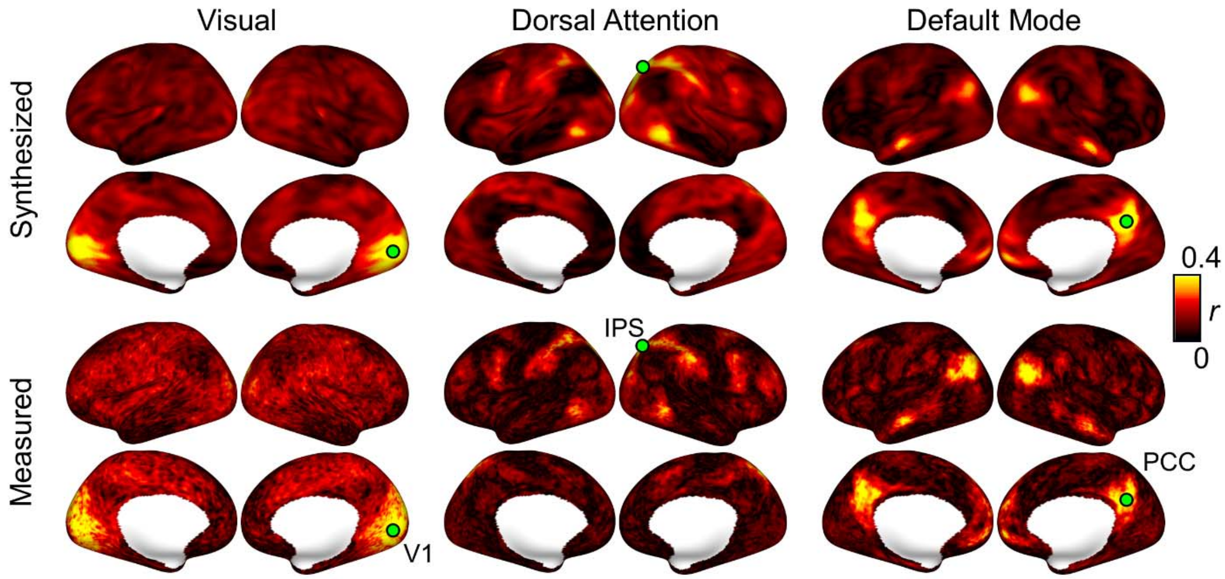


Figure 2. Synthesis of correlated rs-fMRI activity. Seed-based correlations based on VAE-synthesized (upper panel) and experimentally measured (lower panel) rs-fMRI data given three seed locations in the primary visual cortex, intra-parietal sulcus and posterior cingulate cortex, as example locations in the visual network, dorsal attention network, and default-mode network, respectively. The color indicates the correlation coefficient.

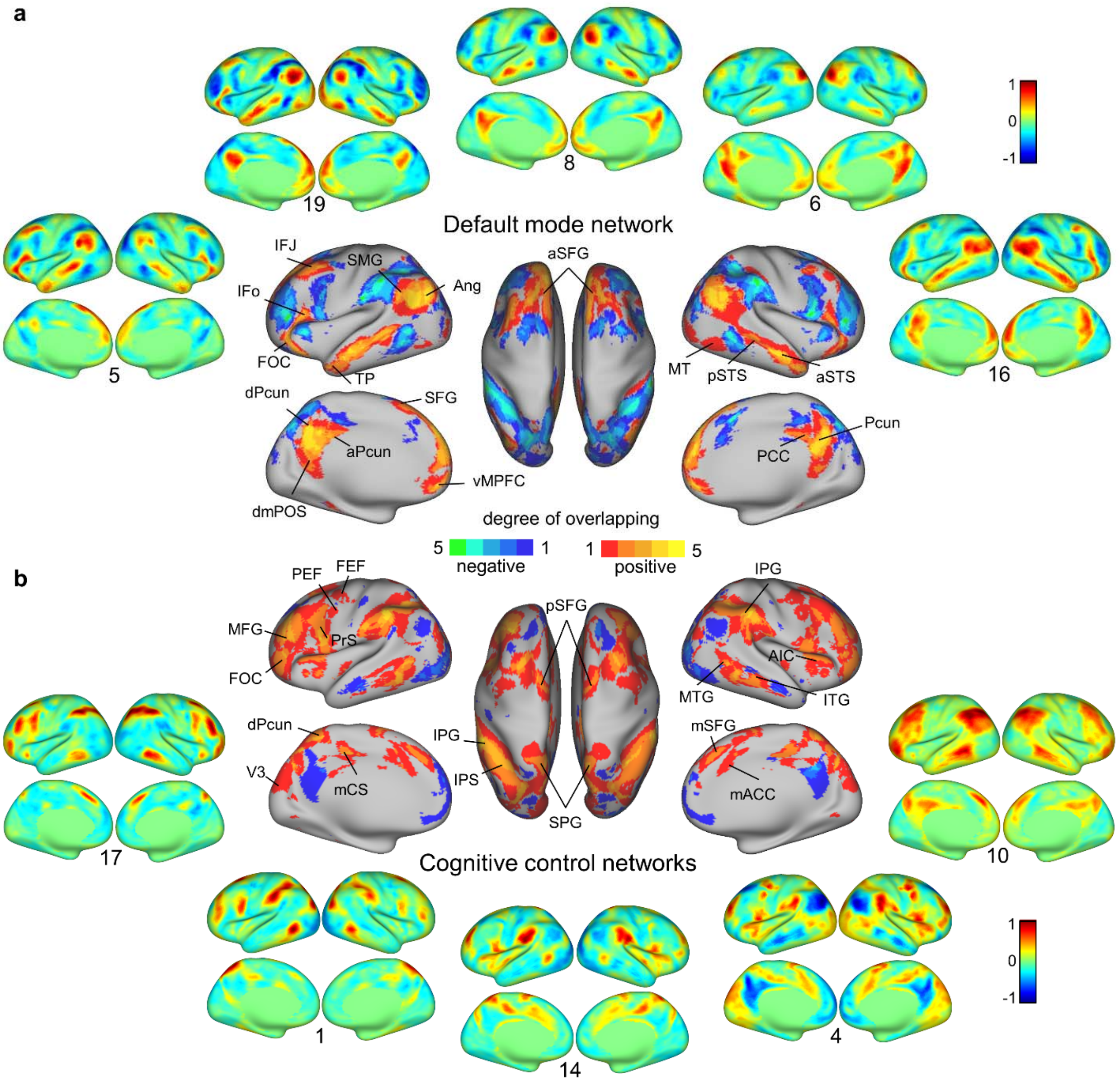
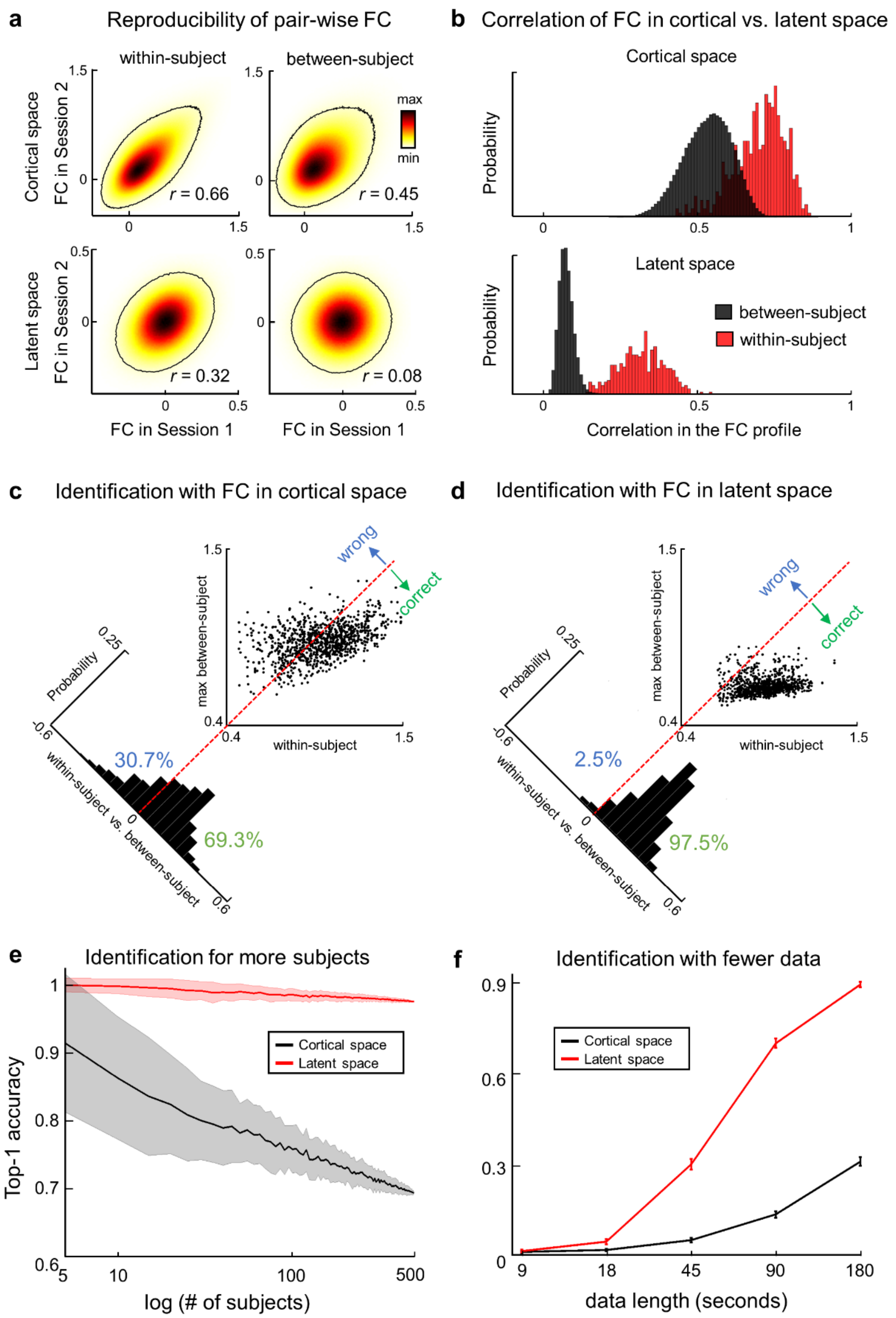


Figure 3. **Latent-space clusters related to the default-mode network (DMN) and the task positive network (TPN).** (**a**) Five clusters (#5, 19, 8, 6, 16) project onto

cortical patterns with positivity in one or multiple regions of DMN. Each pattern is shown as a map normalized to [-1, 1] (or divided by the maximum of the absolute voxel value). The cortical locations with values >0.35 or <-0.35 are labeled as “positive” or “negative”, respectively. For each location, the number of times it appears “positive” (or “negative”) is displayed as red to yellow (or blue to green) to show the degree of overlapping positivity (or negativity) across the five clusters. (**b**) Similarly, five clusters project onto positive patterns in TPN, including the cognitive control network (#17), attention network (#1), cingulo-opercular network (#14, 4), frontoparietal control network (#10). The degree of overlapping positivity (or negativity) is evaluated and displayed in the same way as (a). IFJ: inferior frontal junction, SMG: supramarginal gyrus, IFo: inferior frontal gyrus (pars opercularis), Pcun: precuneus, pSTS: posterior superior temporal sulcus, TP: temporal pole, SFG: superior temporal gyrus, FOC: frontal orbital cortex, dmPOS: dorsomeidal parietooccipital sulcus, IPG: inferior parietal gyrus, MTG: middle temporal gyrus, MFG: middle frontal gyrus, Ang: Angular gyrus, PrS: precentral sulcus, IPS: intraparietal sulcus, ITG : inferior temporal gyrus, IFt: inferior frontal gyrus (pars triangularis), AIC: anterior insular cortex, IFS: inferior frontal sulcus, PHT: Area PHT, SPG: superior parietal gyrus, mCS: margin of the cingulate sulcus, FEF: frontal eye field, PEF: parietal eye field.

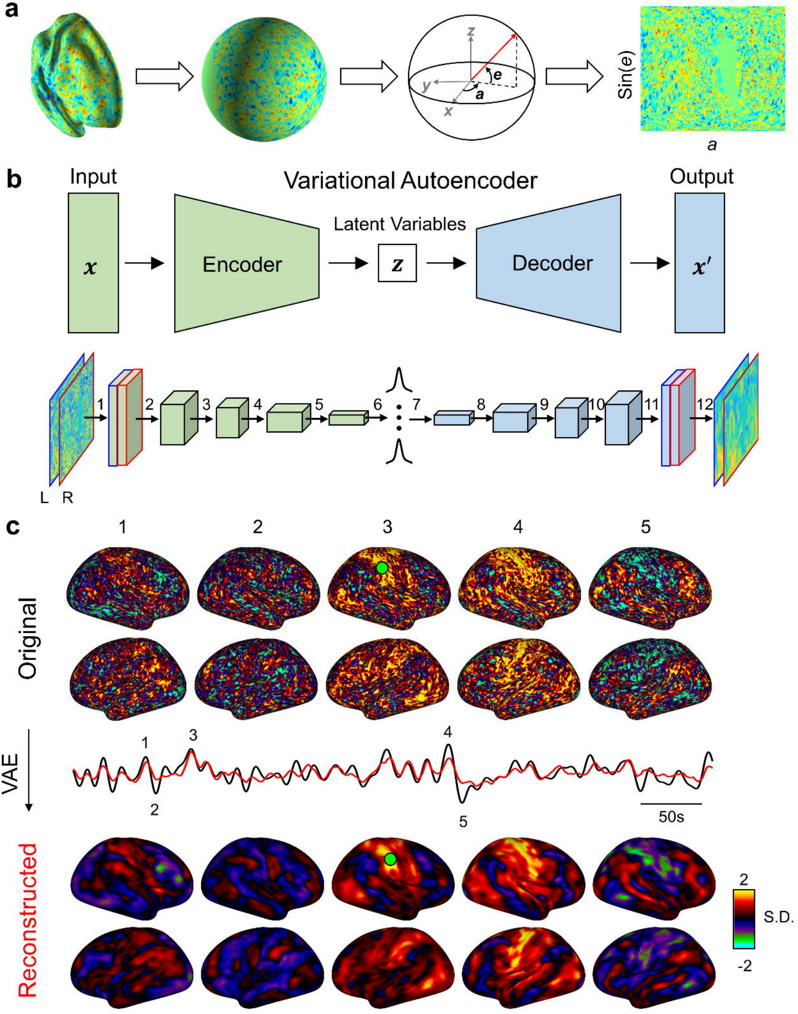


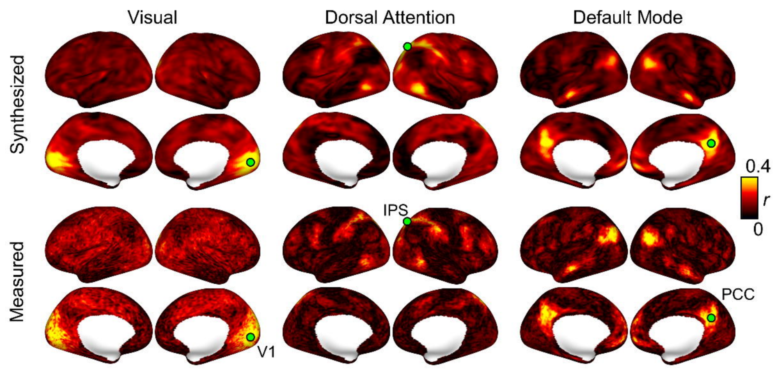
647

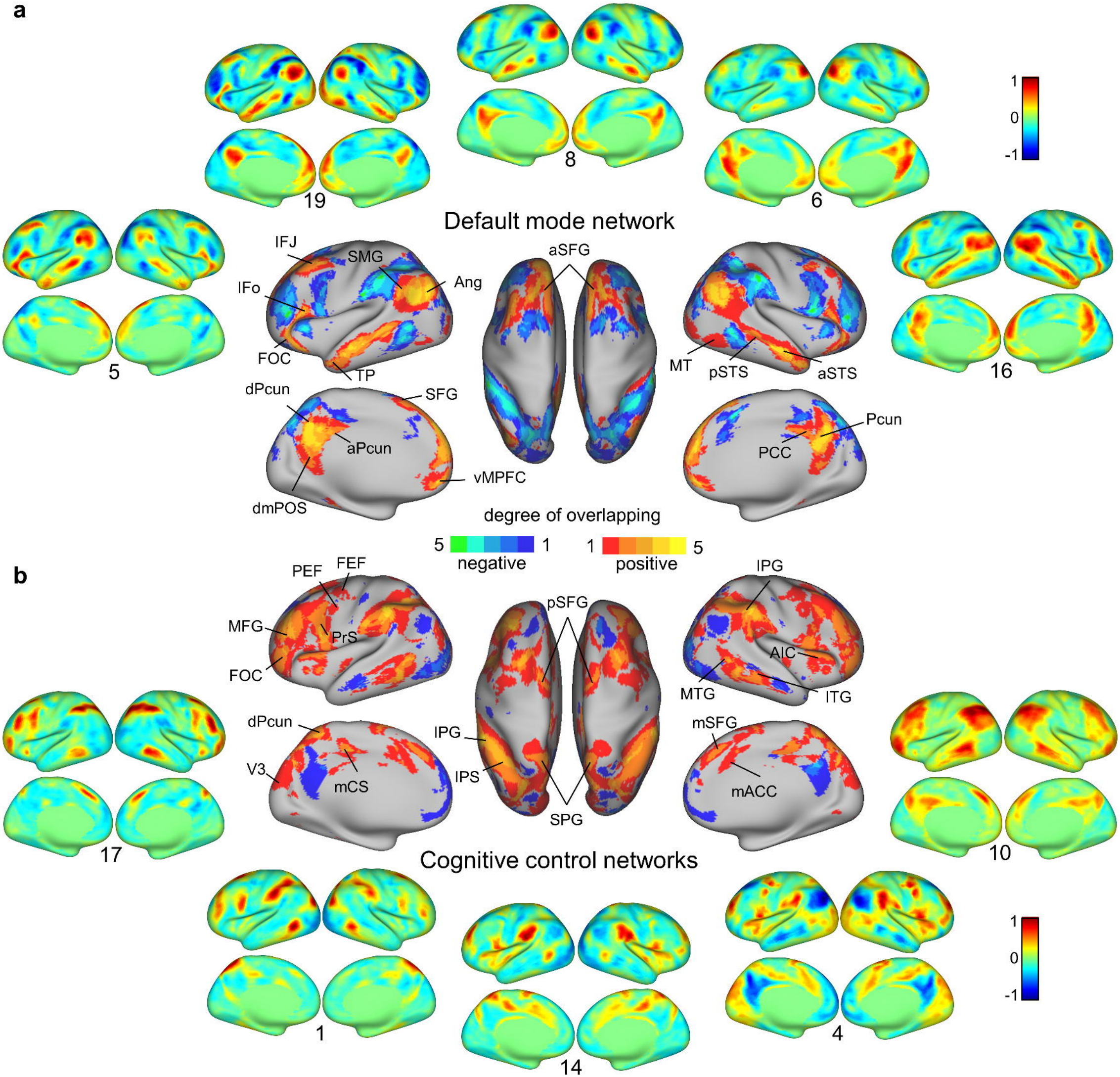
648 **Figure 4. Individual identification based on correlations between latent variables**

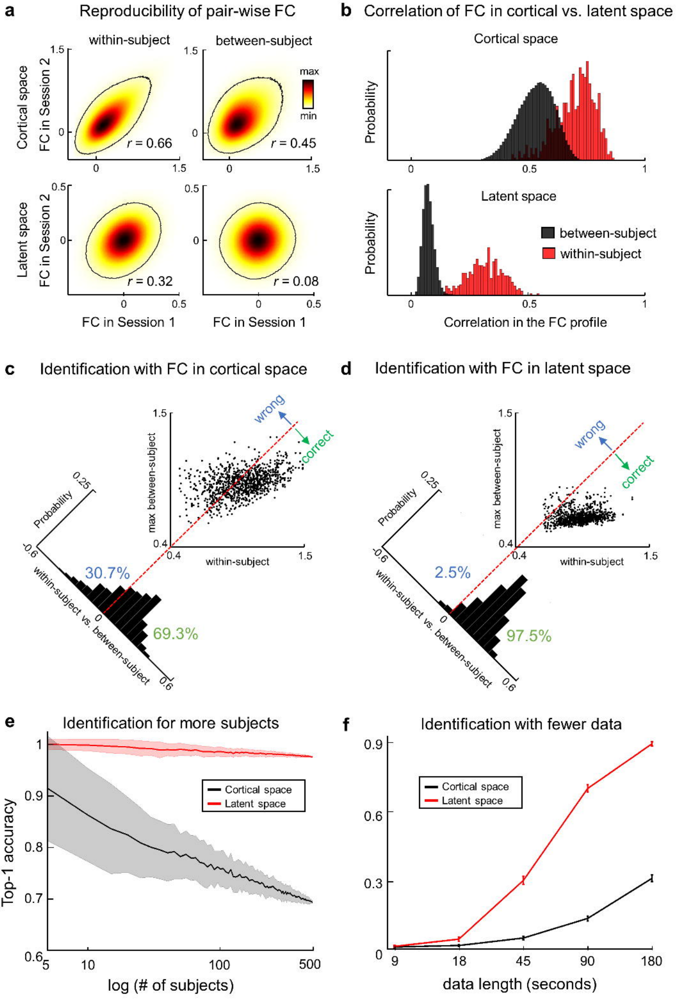
649 **or cortical parcels. (a)** Density distributions of z-transformed correlations between

every pair of cortical parcels (top) or latent variables (bottom). For each pair, the correlation 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. **(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 rs-fMRI 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.


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


Visual
Dorsal Attention
Default Mode
Synthesized
Measured
IPS
PCC
V1
0.4
r
0


a
19
8
6
1
0
-1
Default mode network
5
16
IFJ
SMG
Ang
IFo
FOC
TP
aSFG
MT
pSTS
aSTS
dPcun
SFG
aPcun
vMPFC
dmPOS
Pcun
PCC
degree of overlapping
5
1
negative
1
5
positive
b
PEF
FEF
MFG
PrS
FOC
pSFG
IPG
AIC
MTG
ITG
dPcun
V3
mCS
IPG
IPS
SPG
mSFG
mACC
17
10
Cognitive control networks
1
14
4
1
0
-1


a
Reproducibility of pair-wise FC
within-subject
between-subject
Cortical space
FC in Session 2
1.5
0
r = 0.66
r = 0.45
max
min
Latent space
FC in Session 2
0.5
r = 0.32
r = 0.08
FC in Session 1
FC in Session 1
b
Correlation of FC in cortical vs. latent space
Cortical space
Probability
0
0.5
1
Latent space
Probability
between-subject
within-subject
Correlation in the FC profile
c
Identification with FC in cortical space
1.5
max between-subject
0.4
within-subject
wrong
correct
0.25
Probability
-0.6
within-subject vs. between-subject
0
0.6
30.7%
69.3%
d
Identification with FC in latent space
1.5
max between-subject
0.4
within-subject
wrong
correct
0.25
Probability
-0.6
within-subject vs. between-subject
0
0.6
2.5%
97.5%
e
Identification for more subjects
Top-1 accuracy
1
0.9
0.8
0.7
0.6
5
10
100
500
log (# of subjects)
Cortical space
Latent space
f
Identification with fewer data
0.9
0.6
0.3
0
9
18
45
90
180
data length (seconds)
Cortical space
Latent space