HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 November 01.

Published in final edited form as:

Neuroimage. 2021 November 01; 241: 118423. doi:10.1016/j.neuroimage.2021.118423.

Representation learning of resting state fMRI with variational autoencoder

Jung-Hoon Kim[a,c], **Yizhen Zhang**[b], **Kuan Han**[b], **Zheyu Wen**[b], **Minkyu Choi**[b], **Zhongming Liu**[a,b,*]

[a]Department of Biomedical Engineering, University of Michigan, United States

[b]Department of Electrical Engineering and Computer Science, University of Michigan, United States

[c]Weldon School of Biomedical Engineering, Purdue University, United States

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. 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 in each subject. Our results demonstrate that VAE is a valuable addition to existing tools, particularly suited for unsupervised representation learning of resting state fMRI activity.

Keywords

Variational autoencoder; Deep generative model; Unsupervised learning; Latent gradients

*Corresponding author. zmliu@umich.edu (Z. Liu).

Author credits

Zhongming Liu conceived the original idea. Jung-Hoon Kim, Yizhen Zhang, Kuan Han, Minkyu Choi and Zhongming Liu designed and implemented the model and algorithm. Jung-Hoon Kim and Yizhen Zhang performed the analysis. Zheyu Wen helped debugging of the model and the final algorithm. Jung-Hoon Kim and Minkyu Choi documented the model and source code. Jung-Hoon Kim and Zhongming Liu wrote the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.neuroimage.2021.118423.

1. 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 analyzes (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 analyzed the representational gradients 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.

2. Methods

2.1. 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 0.72 s. 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.

2.2. 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 Fig. 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 MAT-LAB 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.

2.3. 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 (Fig. 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 evaluation. 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. Fig. 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.

$$L(\phi, \theta|x) = \| x - x' \|_2^2 + \beta \cdot D_{KL}\left[N(\mu_Z, \sigma_Z) \| N(0, I)\right] \quad (1)$$

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 $\beta = 9$ 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. Note that the training samples included in each batch were randomly selected from different subjects and time points.

We determined the hyperparameters by exploring and testing different parameter settings with the validation dataset. Specifically, we explored seven values (1, 5, 6, 7, 8, 9, 10) for β and chose $\beta = 9$ to balance the reconstruction performance vs. the disentanglement of latent variables (Fig. 2), which corresponded to the two terms in the loss function shown in Eq. (1). We also explored several options for the number of layers (e.g., 6, 8, 12) and the learning rate (e.g., 10^{-3}, 10^{-4}, 10^{-5}), and finalized those parameters based on the loss evaluated with the validation dataset (Supplementary Fig. 2 and 3). Note that with the hyper-parameters described above, only the VAE model with 12 layers were able to reduce both reconstruction loss and D_{KL} when $\beta = 9$.

2.4. Synthesizing rsfMRI functional connectivity

We used the trained VAE to synthesize rsfMRI 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, as previously described (Jarrett, 2009). In addition, we performed a similar analysis without limiting to the seed locations. Instead, we calculated the functional connectivity between each pair of parcels as defined in a 360-parcel atlas of the whole cortex (Glasser et al., 2016).

For comparison, we similarly calculated seed-based or parcel-to-parcel functional connectivity (with the Fisher's z-transform to convert correlation coefficients to z scores) with experimental rsfMRI data concatenated across a varying number (1, 5, 10, 50, and 100) of subjects in HCP. We compared the functional connectivity pattern observed with synthesized and experimental data, and measured the spatial correlation of the vectorized seed-based correlation map or parcel-to-parcel correlation matrix (after z-transform). We repeated the comparison 20 times. At each time, we randomly generated a different set of synthesized data while using experimental data from a different and randomly selected subset of subjects.

2.5. Defining a principal basis set of the latent space

By our design, the VAE model encodes the spatial pattern of fMRI activity and does not represent the temporal dynamics explicitly. The distribution of every latent variable is constrained to be close to a standard normal distribution independent of one another by minimizing 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 and used this time-difference vector as a discrete approximate of the latent gradient. To the latent gradient, we further applied singular vector decomposition and used the singular vectors to identify a unique basis set of the latent space. Each singular vector defined one latent variable, while the corresponding singular value indicated the importance of the latent variable 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 one with a smaller singular value. The concepts of latent trajectory, latent gradient, and principal components of latent gradients are illustrated in Fig. 6.a.

The latent gradients are vectors in the latent space and can be decoded and visualized as spatial patterns on the cortex. We focused interpretation on the top-9 latent variables defined as the singular vectors with the largest 9 singular values. We passed each of these singular vectors as the input to the VAE's decoder and yielded a corresponding cortical pattern for visualization. Since the polarity of each singular vector is arbitrary, the polarity of its cortical visualization should only be interpreted in terms of the opposition between the positivity and the negativity, while reversing positivity and negativity should not affect its interpretation.

We further tested the reproducibility of the principal latent gradients. Specifically, we separated the data from 500 testing subjects into two halves, each including data from 250 subjects. Separately for each half of the dataset, we calculated the top-9 principal latent gradients and decoded them to corresponding cortical patterns. Then we calculated a matrix of pair-wise correlations between the principal gradients from the first half and those from the second half. If the principal gradients were highly reproducible, they should show up with similar patterns and ordering for the first and second halves of the dataset, and the correlation matrix should show high absolute values for diagonal elements but low absolute values for off-diagonal elements.

Note that the latent gradients as well as their principal components approximate the "temporal derivative" of brain-wide dynamics in the latent space. They should not be

interpreted as the “spatial gradients” of either instantaneous activity patterns (Brown et al., 2021) or functional connectivity patterns (Margulies et al., 2016). Instead, the latent gradients described herein share a similar concept as the dynamic functional connectivity previously described as the multivariate statistical dependence of resting state fMRI data between successive time points (Rogers et al., 2010; Liégeois et al., 2017, 2019). We will further elaborate the similarity and distinction in the Discussion section.

2.6. Individual variation

To evaluate the individual variation, we compared the latent representations of the fMRI data from different individuals. In an exploratory analysis, we randomly selected a small (n=20) subset of subjects. We chose 20 subjects to ease visualization and intuitive demonstration, before scaling up the analysis to 500 subjects. For each of the 20 subjects, 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. The t-SNE method attempted to maintain the relative distance between latent representations (regardless of subjects) embedded in the 2-D space to be as close as possible to their distance in the latent space, where the distance was measured as cosine dissimilarity. We calculated the Silhouette index to measure the cosine similarity of latent representation within the same subject relative to the cosine similarity between different subjects.

2.7. Individual identification

After the exploratory analysis above, we evaluated the individual variation across n=500 subjects in the testing set. For the distribution of subject-wise latent representation, the first moment was the mean and the second moment was the covariance. These two statistics report distinct geometrical features of subject-wise latent representations: the mean reports the location, and the covariance reports the geometry.

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 match was evaluated as the cosine similarity or the Pearson correlation coefficient when the query and the key were based on either the mean or covariance of subject-wise latent representation. 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. In addition, we also evaluated the difference in the key-query similarity when the key and the query were from the same subject (within-subject) vs. when they were from different subjects (between-subject). To test the statistical significance of this difference, we ran a permutation test by shuffling the subject identities for all keys and queries for 10,000 times to obtain the null distributions of

both within-subject and between-subject similarity. The randomly shuffled subject identities reduced the matching between the two fMRI sessions of the same subject to a chance level.

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 were extracted with a similar method as previously reported in Finn et al. (2015). Specifically, the functional connectivity (FC) between brain regions (or connectome) was calculated as features for individual identification. Note 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 functional 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.

2.8. 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 as implemented in Group ICA of fMRI Toolbox (GIFT) (Calhoun et al., 2001) (https://trendscenter.org/software/gift) or the software of Multivariate Exploratory Linear Optimized Decomposition into Independent Components (MELODIC) (Beckmann et al., 2004) (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC). For PCA, we applied PCA to the training data (concatenated across 100 subjects) and used the resulting 256 principal components to represent the testing data (concatenated across 500 subjects). Similarly for ICA, we applied the publicly available tool (GIFT or MELODIC) to the training data and used the resulting 256 spatially independent components to represent the testing data. We compared the performance of reconstructing fMRI spatial patterns in the testing dataset based on their representations in the nonlinear (VAE) vs. linear (PCA or ICA) latent space. In addition, we compared PCA, 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.

3. Results

3.1. 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 relative to the error of data reconstruction from the latent variables, was initially explored for different values (1, 5, 6, 7, 8, 9, 10) and tested with the validation dataset. As shown in Fig. 2, the model's ability to represent the data with its posterior distribution of the latent variables was reduced slightly while β increased from 1 to 9. At $\beta = 9$, the model reached a reasonable trade-off between its ability to represent the input data and the independence of latent variables. However, at $\beta = 10$ (or higher), data reconstruction collapsed while the variational posterior distribution was further forced to match the prior – a phenomenon known as the posterior collapse observed in other applications of VAE (Lucas et al., 2019). To avoid the posterior collapse, we set $\beta = 9$ as the final setting for training and testing the VAE with rsfMRI data.

The model used a pair of convolutional and deconvolutional neural networks in an encoder-decoder architecture (Fig. 1.b). The encoder transformed any rsfMRI pattern, formatted as an image on a regular 2D grid (Fig. 1.a), to the posterior 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 different 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 (Fig. 3). The compression resulted in spatial blurring comparable to the effect of spatial smoothing with 4–6 mm full width at half maximum (FWHM) (Fig. 4). Given fMRI data spatially smoothed to a varying extent (FWHM from 1 to 10 mm), VAE showed either comparable or better performance in representing and reconstructing data than its linear counterparts (PCA or ICA obtained with GIFT or MELODIC), when they used the same dimension (256) for their latent spaces (Fig. 4.a). The difference in the reconstruction performance among VAE, PCA or ICA (GIFT or MELODIC) was statistically significant (repeated measures ANOVA followed by post-hoc paired t-test, false discovery rate $q<0.05$), for all levels of spatial smoothing tested in this study (Fig. 4.b). These results suggest that the posterior latent representation obtained with VAE preserved the spatial and temporal characteristics of rsfMRI, despite a modest but acceptable loss in spatial resolution and specificity.

3.2. 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 its prior probability distribution (i.e., the standard normal distribution) and used the decoder to synthesize 12,000 rsfMRI maps (equivalent to time samples from 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 FC maps with the corresponding maps obtained with rsfMRI data concatenated across a different number of subjects. Fig. 5.a shows three examples with the seed region in the primary visual cortex (V1), the intraparietal sulcus (IPS), or the posterior cingulate cortex (PCC). For each of the three seed locations, the synthesized fMRI data showed a similar seeded FC map as that based on length-matched rsfMRI data obtained from 10 subjects (Fig. 5.a). The FC patterns were 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 with respect to a specific seed location (Fig. 5.b) or all cortical parcels (Fig. 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 spatial patterns and preserve inter-regional correlations as are experimentally observable at a group or population level. When this generative process utilizes the latent variables sampled from their prior distributions (i.e., a standard Gaussian distribution), the generated FC patterns reflect the population average, rather than individualized features. 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. In this comparison, temporal ordering is irrelevant to calculation of the temporal correlation coefficient, which ends up with the same measure of temporal dependency after random shuffling in time.

3.3. Latent variables reflected network dynamics

We also examined the time-evolving trajectory of the latent representation and re-defined the latent variables such that they reflected the dynamic changes of fMRI activity. As illustrated in Fig. 6.a, we first evaluated the displacement of the latent representation from every time point to its next and used the resulting time-wise displacement vector as an approximate of the latent gradient at each time point for all 500 subjects in the testing dataset. Then we applied singular value decomposition and used the resulting singular vectors to redefine the latent variables as the unique basis set that spanned the latent space. Such latent variables, ranked in a descending order by their singular values, represented the directions in which the latent representation tended to move along its time-evolving trajectory.

We focused on the top-9 latent variables as the first nine principal latent gradients that explained the latent dynamics of brain network activity in a descending order (Fig. 6.b). Each principal gradient was a vector in the latent space and thus could be visualized by passing itself through the decoder in the VAE, resulting in a corresponding cortical pattern (Fig. 6.c). The 1st latent variable highlighted sensorimotor areas, including primary visual, auditory and motor cortices, in opposite polarity with the lateral intraparietal cortex. The 2nd latent variable was visualized as a pattern of anti-correlation between the dorsal attention network and the default mode network, similar to a finding reported by Fox and

colleagues (Fox et al., 2005) but without using the confounding procedure of global signal regression (Murphy et al., 2009). The 3rd latent variable corresponded to a largely unipolar pattern, likely reflecting the cortical signature of the global signal. The 4th latent variable showed the opposition between the motor cortex and the cognitive control network. The 5th latent variable showed the opposition between a part of the default mode network and the frontoparietal control network. The 6th to 9th latent variables were more complex and less straightforward to interpret in terms of heuristic resting state networks.

Nevertheless, the principal latent gradients were highly reproducible given test and retest. We split the 500 subjects randomly into two groups (250 subjects per group) and obtained the top-9 principal latent gradients and their cortical visualization separately for each group. The matrix of the pair-wise correlations between the principal latent gradients from the first half and those from the second half were very high for the diagonal elements (the sign of correlation was arbitrary) (Fig. 6.d, top), except that the 3rd and 4th gradients switched their order for the first vs. second half of the dataset. The test-retest correlations in the cortical pattern decoded from the principal latent gradients showed generally higher correlations for the diagonal elements than the off-diagonal elements. However, the off-diagonal correlations were not necessarily zeros. This is reasonable, because the VAE is nonlinear and the orthogonality in the latent space does not imply the orthogonality in the cortical space.

3.4. 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 (Fig. 7.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 (Fig. 7.a), but not in the linear latent space obtained with PCA or ICA (Fig. 7.b). Such distinctions were quantitatively confirmed (Fig. 7.c) by using the Silhouette index to measure the degree of clustering by individuals. The Silhouette value for VAE (mean ± std: s=0.057±0.003) was significantly higher (two sample t-test, p<0.001) than that for PCA (s=−0.020±0.015) or ICA (s=−0.009±0.009). Using the center of latent representation as the subject-identifying feature, we found that subject identity could be retrieved with a reasonably high accuracy when the latent representation was extracted by VAE, whereas the linear representation by PCA or ICA failed the same task nearly entirely (Fig. 7.d). These results suggest the feasibility of using VAE to characterize and reveal individual variations of resting state activity in a non-linear latent space.

3.5. Individual identification

From the t-SNE based visualization (Fig. 7.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 Fig. 8.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 (r=0.41) but not between different subjects (r=0.07). The FC profile was more distinctive across subjects when it was evaluated between latent variables rather than cortical areas (Fig. 8.b). In the latent space, the FC profile was significantly more consistent within a subject than between subjects (permutation test, p<0.001). The distribution of within-subject correlations was in nearly complete separation from that of between-subject correlations (Fig. 8.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 (as a query) against every subject's FC profile in the 2nd session (as a key) and chose the best match between the query and the key in terms of the 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 made with marginal confidence relative to the decision boundary (Fig. 8.c). Using the FC in the latent space allowed us to reach 98.6% top-1 accuracy. The evidence for correct identification was apparent with a large margin from the decision boundary (Fig. 8.d). The use of FC in the latent space supported reliable and robust performance in top-1 identification given an increasingly larger population (Fig. 8.e) or when the data were limited to a short duration (Fig. 8.f), being notably superior to the use of FC in the cortical space.

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 (98.6%). When the key and the query were both based on noisy data (without denoising), the accuracy dropped to 94.1%. When the key and the query were unpaired as denoising applied to one but not the other, the accuracy further dropped to about 91%. 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 no denoising was applied to either the query or the key, the identification accuracy with the cortical-space FC increased to 76.9%, but still significantly lower than the accuracy of 94.1% obtained with the latent-space FC.

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, GIFT, or MELODIC. As shown in Fig. 9, PCA or ICA (either GIFT or MELODIC) was not as effective as VAE. The top-1 accuracy of individual identification was 61.1% for PCA, 50.2% for GIFT, 64.8% for MELODIC in contrast to 98.6% for VAE. The within-subject vs. between-subject similarity in the geometry of linear representation obtained with PCA or ICA (GIFT or MELODIC) exhibited largely overlapping distributions, whereas the corresponding distributions were separated nearly completely for the nonlinear representations obtained with VAE.

4. Discussion

Here, we present a method for unsupervised representation learning of cortical rsfMRI activity. Our results suggest that this method is able to capture and disentangle the generative factors underlying resting state activity, characterize individual variation, and support accurate individual identification. We expect this method to be a valuable addition to the existing tools for characterizing resting state networks and their dynamics. Next, we discuss our findings from the joint perspective of methodology, neuroscience, and applications.

For representation learning of brain activity, we vision that a generalizable system should consist of a base model plus add-on modules. The base model should be trained with self-supervised learning or unsupervised learning and task-free resting state fMRI. Thus, the base model is independent of any specific goals, e.g., behavior or disease prediction, or any specific tasks relevant to perception, action and cognition. After it is trained, the base model is expected to be applicable to fMRI data in different task conditions and to be able to support different goals, not directly by itself, but through add-on extensions. Each add-on should use the representation learned by the based model and be trained to meet a target goal by supervised learning. It is desirable to design and train the base model with a deep architecture to leverage a large amount of unlabeled data, whereas add-ons can be relatively shallow and learnable with fewer labeled data. This strategy would perhaps make the system more scalable, because unlabeled data are much more abundant than labeled data. To support a new goal or condition, the base model should not necessarily have to be retrained or redesigned from scratch but need to pair itself with a new add-on learnable with relatively limited samples.

In the context of this perspective, VAE is a well-suited model to serve as an initial part of the base model described above. It is trainable with unsupervised learning (without any label) (Higgins et al., 2017; Kingma and Welling, 2013). 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 add-on models or 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 design and training of add-on models should be driven by the specific goal of interest and thus be variable across different goals. We intend to confine the scope of this paper to unsupervised learning of the VAE-based base model, while leaving the design and supervised learning of various adds-on models to future studies.

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 (Fig. 5) is also appealing. It can be used as a post-processing strategy for data augmentation and interpolation, when data is short or corrupted, which are 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 plausible origins.

It is worth mentioning two limitations of the VAE model in its current form. First, the model focuses on cortical patterns but excludes subcortical 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, e.g., 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 largely preserves the temporal dynamics (Fig. 3). The trajectory of the latent representation describes the temporal behavior of brain networks, as opposed to the behavior of individual 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 derivative 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. 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 the instantaneous transition in network activity. The direction of this vector encodes a pattern of network interaction that drives the instantaneous change of network

activity. The principal components of the latent gradients uncover the hidden factors that drive the temporal dynamics of brain networks (Fig. 6).

The use of temporal derivative has been exploited in multivariate auto-regressive modeling of voxel-wise fMRI signals. For example, Rogers and colleagues used a first-order auto-regressive (AR-1) model to describe the relationship between signals at successive time points (Rogers et al., 2010). This AR-1 model is a matrix that describes the multivariate dynamics of brain activity – how the signals at present can tell us about the signals upcoming at next. As such, the AR-1 matrix itself provides a dynamic measure of functional connectivity (or namely dynamic functional connectivity), in contrast to static functional connectivity (Biswal et al., 1995; Yeo et al., 2011) or time-varying functional connectivity (Chang and Glover, 2010; Hutchison et al., 2013), as discussed in depth by Liégeois and colleagues (Liégeois et al., 2017). In light of these prior work, the temporal gradients discussed herein can be similarly described by the AR-1 model (or dynamic functional connectivity matrix) minus an identity matrix. It discounts a trivial effect that a voxel-wise signal simply copies itself from one time to its next. In this sense, the temporal gradients describe the temporal dynamics. When evaluated in the latent space, such temporal gradients report dynamics of networks, rather than voxels or regions of interest (Liégeois et al., 2019).

A related analysis or notion has been explored in two independent studies reported in two recent preprints, which we became aware of during the peer review of our paper (Brown et al., 2021; Liu et al., 2020b). Unlike these related studies, it is worth noting that the latent gradients and their principal components discussed in this paper are temporal gradients of latent trajectory, rather than spatial gradients. It reflects changes of network activity patterns, as opposed to the spatial pattern of network activity itself. Further, the temporal gradient of network activity should also be set apart from the spatial gradient of functional connectivity (Margulies et al., 2016). Nevertheless, the conceptually distinctive gradient measures appear to share partly similar patterns – an intriguing observation that remains to be interpreted either mathematically or in terms of brain structure.

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 that compress the spatial patterns of brain activity and explain temporal gradients in brain dynamics. 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 discovered in a data-driven manner, it is currently unclear how to interpret them as specific physiological processes, many of which are not observable. Visualizing each latent variable as a cortical pattern through the VAE's decoder is helpful for heuristic interpretation of the latent variable in terms of resting state networks (Yeo et al., 2011). For example, it is perhaps intuitive to interpret the 3rd latent variable as a major contributor to the global signal (Fox et al., 2005; Murphy et al., 2009), potentially reflecting the arousal fluctuation (Chang et al., 2016). Several latent variables correspond to various patterns of opposition between networks, e.g., cortical areas involved in sensorimotor vs. cognitive functions, attention vs. default-mode. However, such heuristic and post-hoc interpretation should be taken with caution. More mechanistic interpretation awaits future studies to test the causal relationships between the latent variables learned from data and their cortical or behavioral correlates.

Nevertheless, we expect the latent variables extracted by VAE to provide the computational basis for understanding the origins of resting state activity. We do not suggest that the latent variables should be interpreted with any dichotomy such as signals vs. noise, neuronal vs. non-neuronal, and brain vs. body. Increasingly evidence suggests that such cases of dichotomy are either overly simplified or questionable (Bright and Murphy, 2015; Azzalini et al., 2019). We also do not necessarily expect one-to-one correspondence between latent variables and observable physiological sources, because those sources are likely entangled. For example, systemic fluctuations, such as changes in cardiac, respiratory, or gastric rhythms, may arise from sympathetic or parasympathetic neuromodulation mediated by neural pathways ranging from the brainstem to the cortex (Özbay et al., 2019; Rebollo et al., 2018). Instead, it is more reasonable to expect many-to-one correspondence such that physiological sources, e.g., arousal or respiration, may be predicted from the latent variables up to linear and sparse projection.

It is worth noting that the VAE is trained with data from a population of subjects, instead of a single subject. Every training example is a spatial pattern of cortical activity. Different training examples may reflect different brain states or different subjects. Hence, the learned latent variables may reflect some sources that explain individual variation and those that explain the characteristics of the brain common across individuals. It is likely and desirable that the latent space might be further separated into two sub-spaces: one characterizes individual variation, whereas the other reports population characteristics of the brain. Although we are unable to separate the individual vs. population characteristics, the VAE is not biased by individual variation in terms of its architecture, learning objective, or training strategy. During training, each batch uses training examples randomly drawn from different subjects. Thus, embeddings in the latent space should reflect the intrinsic structure of resting state fMRI data, as opposed to any analysis artifact.

The VAE model used in this study learns to compress high-dimensional data into much lower-dimensional space spanned by orthogonal basis. In this sense, VAE is similar to temporal ICA (Smith et al., 2012) but allows for nonlinear relationships between the latent variables and the input data they represent (Khemakhem et al., 2019). Similar to the notion of “learn to compress and compress to learn” (Yu et al., 2020), the VAE pushes for disentanglement to maximize the degree of compression. If the latent variables were not independent of one another, the low-dimensional representations expressed in terms of the latent variables could be further compressed. In practice, this objective is measured by using the K-L divergence between the latent variables’ prior and posterior distributions. The latter is dependent on the input data, whereas the former does not. When the K-L divergence is given a high weight (β) relative to the reconstruction loss, the model forces the posterior distribution to match the prior and thus tends to ignore the input data. We have observed this phenomenon when β was set to 10 or higher with our current model. This phenomenon, also known as the posterior collapse (Lucas et al., 2019), is a dilemma for training the VAE. Since disentanglement is not perfect (to avoid the posterior collapse), the posterior is not exactly a spherical Gaussian distribution. When VAE represents data from a large population of subjects, the representations follow a distribution close to, but not exactly, a spherical Gaussian distribution. However, given input data from a single subject (or a few subjects) or a short period of time, the VAE-obtained representations do not necessarily

follow a spherical Gaussian distribution, as demonstrated in Fig. 7. In this study, we used a fixed β during training. It might be of interest to vary β dynamically during learning to mitigate the dilemma of reconstruction vs. disentanglement (Shao et al., 2020). In addition, the architecture of the VAE model also has room for improvement, including its depth, the number of channels or the kernel size in each layer.

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 still 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 such applications would be fruitful and potentially impactful to cognitive sciences and clinical applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The research is supported by National Institute of Mental Health R01MH104402 and the University of Michigan.

Data and code availability statement

The data used in this paper are publicly accessible from the website from the Human Connectome Project. The source code, model and documentation for the method described in this paper are publicly available https://github.com/libilab/rsfMRI-VAE.

Reference

Allen EA, Damaraju E, Plis SM, Erhardt EB, Eichele T, Calhoun VD, 2014. Tracking whole-brain connectivity dynamics in the resting state. Cerebral Cortex24, 663–676. [PubMed: 23146964]

Amico E, Goñi J, 2018. The quest for identifiability in human functional connectomes. Scienti. Rep8, 1–14.

Azzalini D, Rebollo I, Tallon-Baudry C, 2019. Visceral signals shape brain dynamics and cognition. Trends Cognit. Sci. 23, 488–509. [PubMed: 31047813]

Beckmann CF, Smith SM, 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging23, 137–152. [PubMed: 14964560]

Bianciardi M, Fukunaga M, van Gelderen P, Horovitz SG, de Zwart JA, Shmueli K, Duyn JH, 2009. Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magnet. Resonance Imaging27, 1019–1029.

Birn RM, Smith MA, Jones TB, Bandettini PA, 2008. The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage40, 644–654. [PubMed: 18234517]

Biswal B, Zerrin Yetkin F, Haughton VM, Hyde JS, 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnet. Resonance Med. 34, 537–541.

Bright MG, Murphy K, 2015. Is fMRI “noise” really noise? Resting state nuisance regressors remove variance with network structure. NeuroImage114, 158–169. [PubMed: 25862264]

Brown JA, Lee AJ, Pasquini L, Seeley WW, 2021. A dynamic gradient architecture generates brain activity states. bioRxiv2020.08.12.248112.

Byrge L, Kennedy DP, 2019. High-accuracy individual identification using a “thin slice” of the functional connectome. Netw. Neurosci. 3, 363–383. [PubMed: 30793087]

Calhoun VD, Adali T, Pearlson G, Pekar JJ, 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Human Brain Map. 13, 43–53.

Chang C, Glover GH, 2010. Time–frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage50, 81–98. [PubMed: 20006716]

Chang C, Leopold DA, Schölvinck ML, Mandelkow H, Picchioni D, Liu X, Frank QY, Turchi JN, Duyn JH, 2016. Tracking brain arousal fluctuations with fMRI. Proc. Natl. Acad. Sci. 113, 4518–4523. [PubMed: 27051064]

Chen S, Hu X, 2018. Individual identification using the functional brain fingerprint detected by the recurrent neural network. Brain Connect. 8, 197–204. [PubMed: 29634323]

Chou Y. h., Sundman M, Whitson HE, Gaur P, Chu M-L, Weingarten CP, Madden DJ, Wang L, Kirste I, Joliot M, 2017. Maintenance and representation of mind wandering during Resting-State fMRI. Scient. Rep. 7, 40722.

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 Trans. Cogn. Devel. Syst.

D’Souza NS, Nebel MB, Wymbs N, Mostofsky S, Venkataraman A, 2019. Integrating neural networks and dictionary learning for multidimensional clinical characterizations from functional connectomics data. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 709–717.

Dvornek NC, Yang D, Ventola P, Duncan JS, 2018. Learning generalizable recurrent neural networks from small task-fmri datasets. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 329–337.

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. Front. Neurosci14, 881. [PubMed: 33013292]

Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT, 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neurosci. 18, 1664. [PubMed: 26457551]

Fischl B, 2012. FreeSurfer. Neuroimage62, 774–781. [PubMed: 22248573]

Fox MD, Buckner RL, Liu H, Chakravarty MM, Lozano AM, Pascual-Leone A, 2014. Resting-state networks link invasive and noninvasive brain stimulation across diverse psychiatric and neurological diseases. Proc. Natl. Acad. Sci111, E4367–E4375. [PubMed: 25267639]

Fox MD, Raichle ME, 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature Rev. Neurosci8, 700–711. [PubMed: 17704812]

Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME, 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci102, 9673–9678. [PubMed: 15976020]

Gadgil S, Zhao Q, Adeli E, Pfefferbaum A, Sullivan EV, Pohl KM, 2020. Spatio-temporal graph convolution for functional MRI analysis. arXiv preprint arXiv:2003.10613.

Garrity AG, Pearlson GD, McKiernan K, Lloyd D, Kiehl KA, Calhoun VD, 2007. Aberrant "default mode" functional connectivity in schizophrenia. Am. J. Psych. 164, 450–457.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, 2016. A multi-modal parcellation of human cerebral cortex. Nature536, 171–178. [PubMed: 27437579]

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage80, 105–124. [PubMed: 23668970]

Gonzalez-Castillo J, Hoy CW, Handwerker DA, Robinson ME, Buchanan LC, Saad ZS, Bandettini PA, 2015. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. Proc. Natl. Acad. Sci112, 8762–8767. [PubMed: 26124112]

Griffanti L, Salimi-Khorshidi G, Beckmann CF, Auerbach EJ, Douaud G, Sexton CE, Zsoldos E, Ebmeier KP, Filippini N, Mackay CE, 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage95, 232–247. [PubMed: 24657355]

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. Neuroimage198, 125–136. [PubMed: 31103784]

He T, Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, Bzdok D, Feng J, Yeo BT, 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage206, 116276. [PubMed: 31610298]

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. ICLR2, 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 Trans. Med. Imaging37, 1551–1561. [PubMed: 28641247]

Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, Della Penna S, Duyn JH, Glover GH, Gonzalez-Castillo J, 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage80, 360–378. [PubMed: 23707587]

Jarrett C, 2009. The restless brain. 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 Neurosci. 4, 448–466.

Kawahara J, Brown CJ, Miller SP, Booth BG, Chau V, Grunau RE, Zwicker JG, Hamarneh G, 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage146, 1038–1049. [PubMed: 27693612]

Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH, 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron98, 630–644 e616. [PubMed: 29681533]

Khaligh-Razavi S-M, Kriegeskorte N, 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol10, e1003915. [PubMed: 25375136]

Khemakhem I, Kingma DP, Hyvärinen A, 2019. Variational autoencoders and nonlinear ICA: a unifying framework. arXiv preprint arXiv:1907.04809.

Khosla M, Jamison K, Kuceyeski A, Sabuncu MR, 2019a. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. Neuroimage199, 651–662. [PubMed: 31220576]

Khosla M, Jamison K, Ngo GH, Kuceyeski A, Sabuncu MR, 2019b. Machine learning in resting-state fMRI analysis. Magnet. 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. In: Proceedings of the International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, pp. 1–4.

Kingma DP, Ba J, 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma DP, 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 Comput. Biol15, e1007263. [PubMed: 31433810]

Kriegeskorte N, Kievit RA, 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cognit. Sci17, 401–412. [PubMed: 23876494]

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. Nature521, 436–444. [PubMed: 26017442]

Leopold DA, Maier A, 2012. Ongoing physiological processes in the cerebral cortex. Neuroimage62, 2190–2200. [PubMed: 22040739]

Liégeois R, Laumann TO, Snyder AZ, Zhou J, Yeo BT, 2017. Interpreting temporal fluctuations in resting-state functional connectivity MRI. NeuroImage163, 437–435. [PubMed: 28916180]

Liégeois R, Li J, Kong R, Orban C, Van De Ville D, Ge T, Sabuncu MR, Yeo BT, 2019. Resting brain dynamics at different timescales capture distinct aspects of human behavior. Nature Commun. 10, 2317. [PubMed: 31127095]

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. Neuroimage208, 116459. [PubMed: 31837471]

Liu S, Zhao L, Wang X, Xin Q, Zhao J, Guttery DS, Zhang Y-D, 2020b. Deep spatio-temporal representation and ensemble classification for attention deficit/Hyperactivity disorder. IEEE Trans. Neural Syst. Rehabil. Eng.

Lu H, Jaime S, Yang Y, 2019. Origins of the resting-state functional MRI signal: potential limitations of the "neurocentric" model. Front. Neurosci13.

Lucas J, Tucker G, Grosse R, Norouzi M, 2019. Don't blame the ELBO! A linear VAE perspective on posterior collapse. Adv. Neural Inf. Process. Syst32.

Makkie M, Huang H, Zhao Y, Vasilakos AV, Liu T, 2019. Fast and scalable distributed deep convolutional autoencoder for fMRI big data analytics. Neurocomputing325, 20–30. [PubMed: 31354187]

Mantini D, Perrucci MG, Del Gratta C, Romani GL, Corbetta M, 2007. Electrophysiological signatures of resting state networks in the human brain. Proc. Natl. Acad. Sci104, 13170–13175. [PubMed: 17670949]

Margulies DS, Ghost SS, Goulas A, Falkiewicz M, Huntenburg JM, Langs G, Bezgin G, Eickhoff SB, Castellanos XF, Petrides M, Jefferies E, Smallwood J, 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proc. Natl. Acad. Sci113, 12574–12579. [PubMed: 27791099]

Matsubara T, Tashiro T, Uehara K, 2019. Deep neural generative model of functional MRI images for psychiatric disorder diagnosis. IEEE Trans. Biomed. Eng66, 2768–2779. [PubMed: 30703004]

Mejia AF, Nebel MB, Barber AD, Choe AS, Pekar JJ, Caffo BS, Lindquist MA, 2018. Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. Neuroimage172, 478–491. [PubMed: 29391241]

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. Neuroimage104, 398–412. [PubMed: 25312773]

Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA, 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced?Neuroimage44, 893–905. [PubMed: 18976716]

Nair V, Hinton GE, 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.

Özbay PS, Chang C, Picchioni D, Mandelkow H, Chappel-Farley MG, van Gelderen P, de Zwart JA, Duyn JH, 2019. Sympathetic activity contributes to the fMRI signal. Commun. Biol2, 421. [PubMed: 31754651]

Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, Johnson HJ, Paulsen JS, Turner JA, Calhoun VD, 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci8, 229. [PubMed: 25191215]

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE, 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage84, 320–341. [PubMed: 23994314]

Rakic P, 2009. Evolution of the neocortex: a perspective from developmental biology. Nature Rev. Neurosci10, 724–735. [PubMed: 19763105]

Rebollo I, Devauchelle AD, Beranger B, Tallon-Baudry C, 2018. Stomach-brain synchrony reveals a novel, delayed connectivity resting-state network in humans. eLife7, e33321. [PubMed: 29561263]

Riaz A, Asad M, Alonso E, Slabaugh G, 2020. DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. J. Neurosci. Methods335, 108506. [PubMed: 32001294]

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S, 2019. A deep learning framework for neuroscience. Nature Neurosci. 22, 1761–1770. [PubMed: 31659335]

Rogers BP, Katwal SB, Morgan VL, Asplund CL, Gore JC, 2010. Functional MRI and multivariate autoregressive models. Magnet. Reson. Imaging28, 1058–1065.

Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM, 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage90, 449–468. [PubMed: 24389422]

Seo Y, Morante M, Kopsinis Y, Theodoridis S, 2019. Unsupervised pre-training of the brain connectivity dynamic using residual D-Net In: Proceedings of the International Conference on Neural Information Processing. Springer, pp. 608–620.

Shao H, Lin H, Yang Q, Yao S, Zhao H, Abdelzaher T, 2020. Dynamic-VAE: Decoupling reconstruction error and disentangled representation learning. arXiv:2009.06795

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. Neuroimage49, 3110–3121. [PubMed: 19931396]

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 Map. 39, 2269–2282.

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR, 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci106, 13040–13045. [PubMed: 19620724]

Smith SM, Miller KL, Moeller S, Xu J, Auerbach EJ, Woolrich MW, Beckmann CF, Jenkinson M, Andersson J, Glasser MF, 2012. Temporally-independent functional modes of spontaneous brain activity. Proc. Natl. Acad. Sci109, 3131–3136. [PubMed: 22323591]

Sporns O, Tononi G, Kötter R, 2005. The human connectome: a structural description of the human brain. PLoS Comput. Biol1.

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. Neuroimage129, 292–307. [PubMed: 26774612]

Sutskever I, Vinyals O, Le QV, 2014. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp. 3104–3112.

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium W-MH, 2013. The WU-Minn human connectome project: an overview. Neuroimage80, 62–79. [PubMed: 23684880]

Venkatesh M, Jaja J, Pessoa L, 2019. Brain dynamics and temporal trajectories during task and naturalistic processing. Neuroimage186, 410–423. [PubMed: 30453032]

Venkatesh M, Jaja J, Pessoa L, 2020. Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. Neuroimage207, 116398. [PubMed: 31783117]

Wang L, Li K, Chen X, Hu XP, 2019. Application of convolutional recurrent neural network for individual recognition based on resting state fMRI data. Front. Neurosci13, 434. [PubMed: 31118882]

Wang X, Liang X, Jiang Z, Nguchu BA, 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 Map. 41, 1505–1519.

Wen H, Shi J, Chen W, Liu Z, 2018. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Scient. Rep. 8, 1–17.

Winder AT, Echagarruga C, Zhang Q, Drew PJ, 2017. Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. Nature Neurosci. 20, 1761–1769. [PubMed: 29184204]

Yamins DL, DiCarlo JJ, 2016. Using goal-driven deep learning models to understand sensory cortex. Nature Neurosci. 19, 356–365. [PubMed: 26906502]

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 Trans. Cybern..

Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol..

Yu Y, Chan KHR, You C, Song C, Ma Y, 2020. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. arXiv: 2006.08558.

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. Neuroimage55, 856–867. [PubMed: 21236349]

Zhang Y, Han K, Worth R, Liu Z, 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. Nature Commun. 11, 1–13. [PubMed: 31911652]

Zhao Q, Honnorat N, Adeli E, Pfefferbaum A, Sullivan EV, Pohl KM, 2019. Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. In: Proceedings of the 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). In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 181–189.

Zou L, Zheng J, Miao C, Mckeown MJ, Wang ZJ, 2017. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. IEEE Access5, 23626–23636.

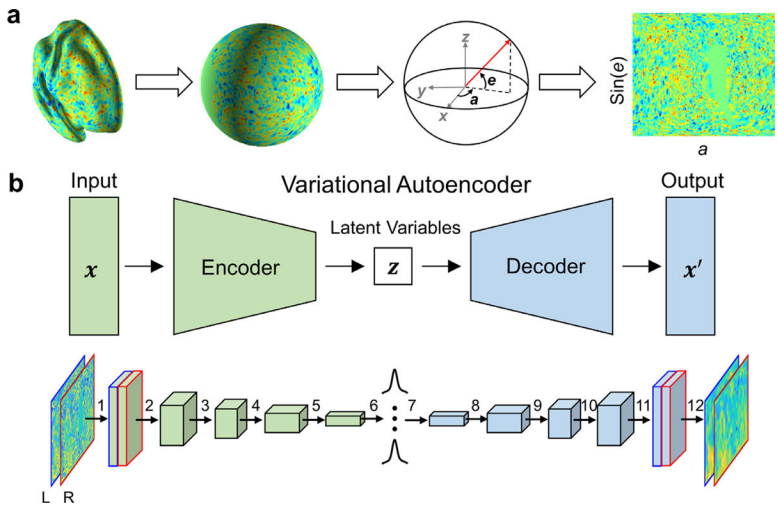

Fig. 1. Variational Auto-Encoder (VAE).

(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 contains 5 convolutional layers connected in series. In the encoder, each layer (numbered from 1 to 5) outputs a feature map with the size of 96 × 96 × 64, 48 × 48 × 128, 24 ×24 ×128, 12 × 12 ×256, or 6 ×6 ×256, respectively. In layer 1, 32 kernels are applied to 192 × 192 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 12 × 12 × 256, 24 × 24 × 128, 48 × 48 × 128, 96 × 96 × 64, or 192 × 192 × 2, 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 (yielding two 256-vectors as the mean and the standard deviation of 256 latent variables) and re-parameterization; 7: fully connected layer and rectified nonlinearity (yielding a 6 × 6 × 256 feature map); 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.

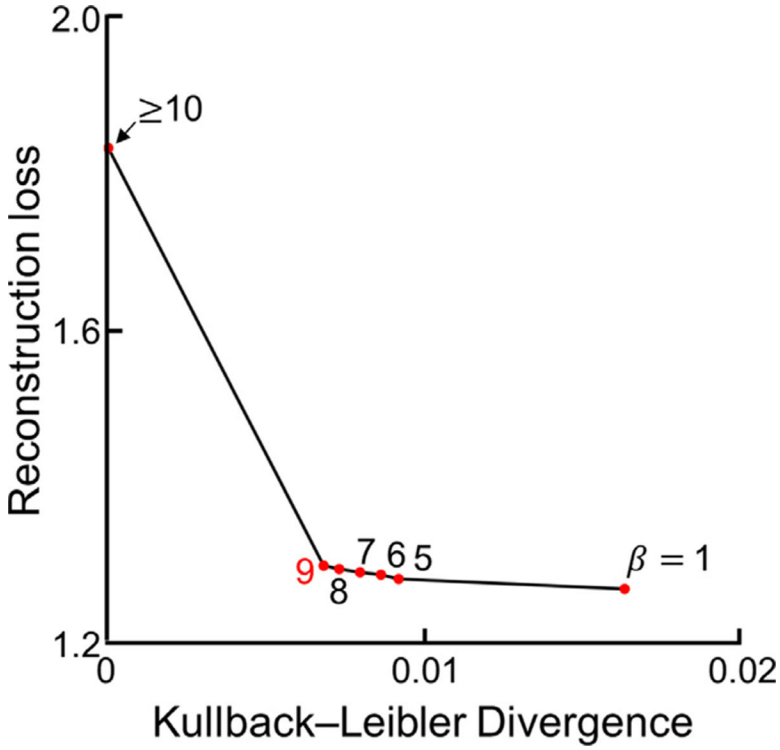

Fig. 2. Reconstruction vs. disentanglement for the VAE models trained with different values for β in the loss function.

$\beta = 9$ shows a reasonable trade-off between the reconstruction loss and the Kullback–Leibler divergence on the validation dataset. Also see Supplementary Fig. 3 for related results with VAE that include different numbers of layers.

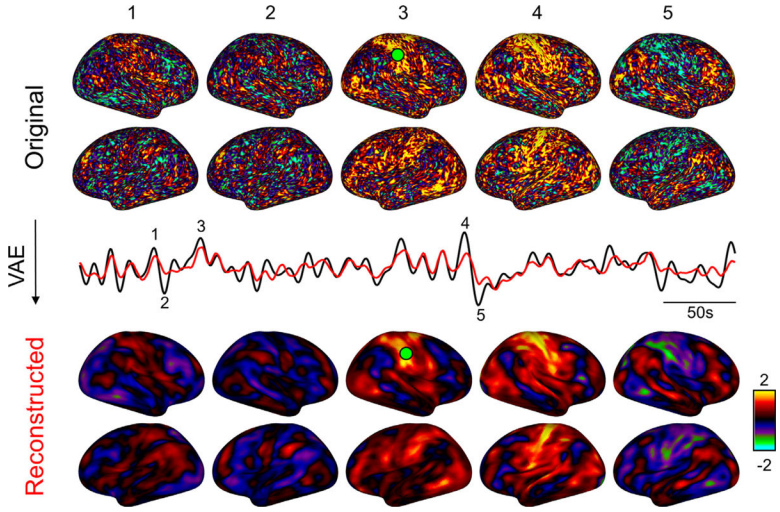

Fig. 3. Image reconstruction using VAE.

A series of cortical patterns are reconstructed through the VAE model given the posterior latent distributions learned from the original data from rsfMRI experiments. 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 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

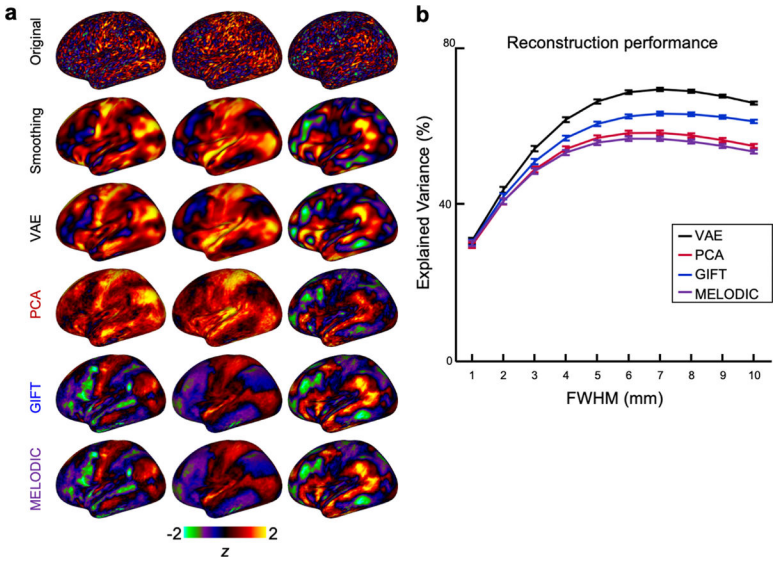

Fig. 4. RsfMRI data compression and reconstruction with VAE vs. PCA and ICA.

(**a**) For illustration, three example maps of fMRI activity, before (1st row) and after (2nd row) spatial smoothing (FWHM=6mm), are shown in comparison with the corresponding maps reconstructed with VAE (3rd row), PCA (4th row), ICA by GIFT (5th row) and MELODIC (6th row) with 256 nonlinear latent variables or linear components. (**b**) For quantitative comparison, the reconstruction performance, in terms of the percentage of variance in the fMRI images explained by the model reconstruction, is shown for VAE, PCA, GIFT, and MELODIC a function of the FWHM (from 1 to 10 mm) used for spatial smoothing of fMRI images. The error bar stands for the standard error of mean.

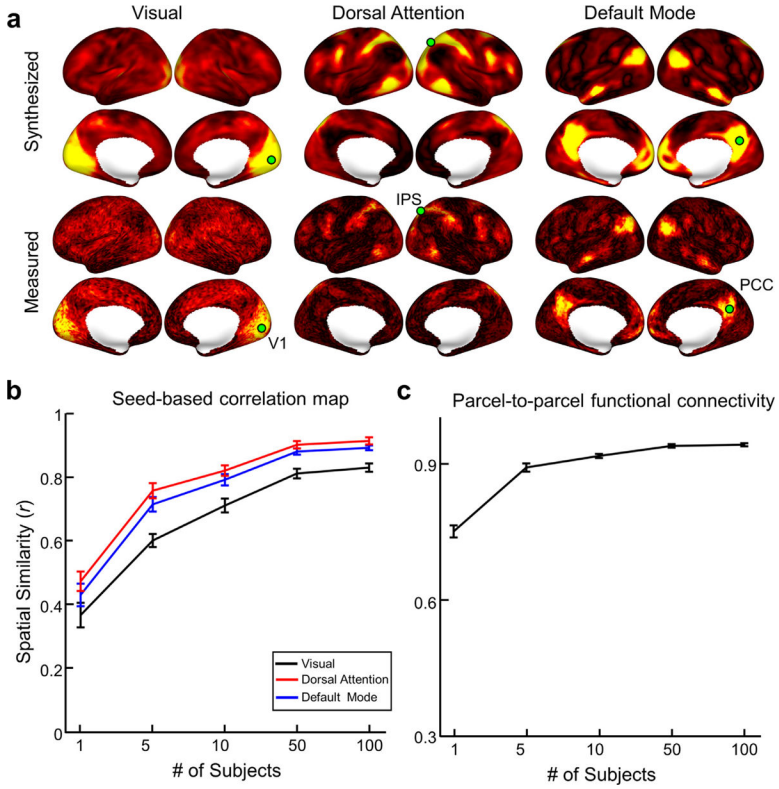

Fig. 5. VAE synthesizes correlated fMRI activity.

(**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 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

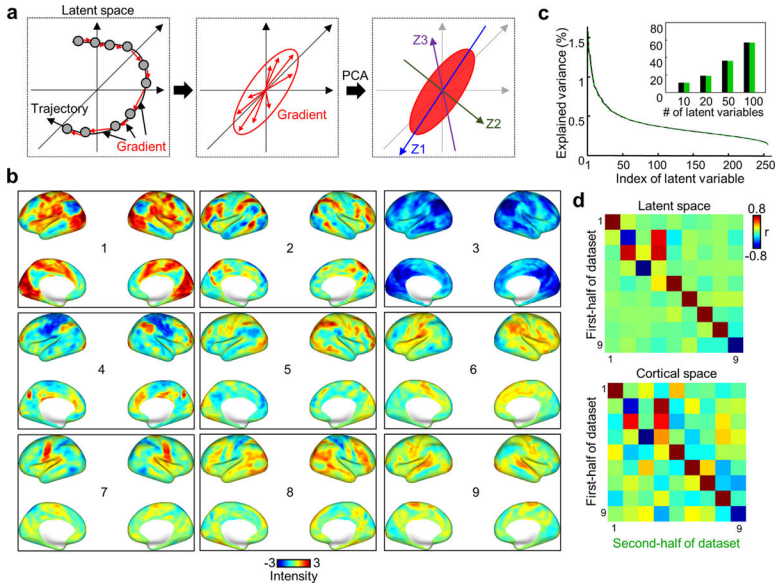

Fig. 6. Latent gradients drive the dynamics of latent representation.

(**a**) From the latent trajectory (black curve), the time-difference vectors (red arrows) are the difference between the representations (gray circles) from successive time points. These time-difference vectors approximate the temporal gradients as shown in the middle panel. Applying SVD to the temporal gradients extracts the principal components of the temporal gradients (also referred to as the principal gradients). These principal gradients are orthogonal to each another and thus form a basis set of the latent space and redefine the latent variables. (**b**) The percentage of variance in the latent gradient that each redefined latent variable explains. The inset shows the percentage of the total variance explained by top 10, 20, 50, or 100 latent variables. (**c**) Cortical pattern decoded from each of the top-9 latent variables. (**d**) Test-retest reproducibility of latent variables. Each element in the matrix shows the correlation between a latent variable (or principal latent gradient) derived from a set of 250 subjects (first half) and a latent variable derived from a different set of 250 subjects (second half). The ordering of latent variable is based on the corresponding singular value (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

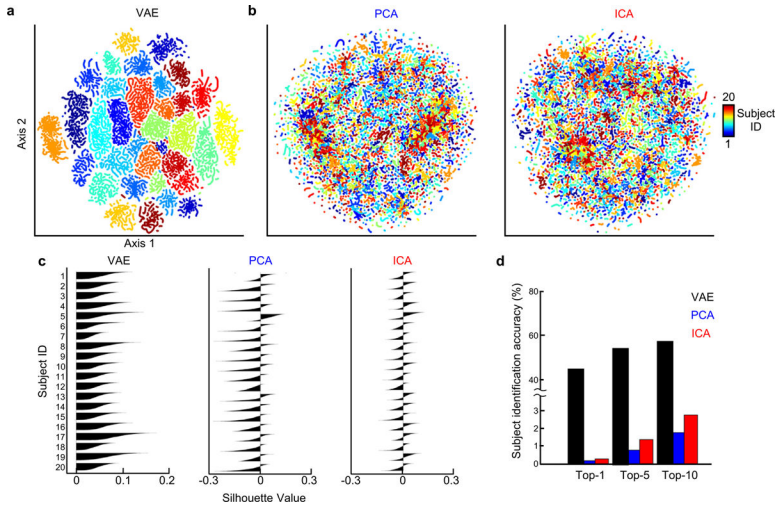

Fig. 7. Individual variation of latent representation obtained with VAE, PCA and ICA.

(a-b) Subject-wise latent representations visualized in a 2-D space obtained with t-SNE, when **(a)** VAE, or **(b)** PCA or ICA is used to extract representations of rsfMRI activity from 20 subjects. **(c)** The Silhouette value indicates to what extent a representation is similar to each other within the same subject as opposed to between different subjects for VAE (left), PCA (middle) or ICA (right). **(d)** Top-1, 5, and 10 accuracy of using the time-averaged representation as the feature to identify individuals in a large group (n=500) of subjects, given VAE (black), PCA (blue), or ICA. ICA is based on GIFT. Similar results are obtained with ICA implemented by MELODIC (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

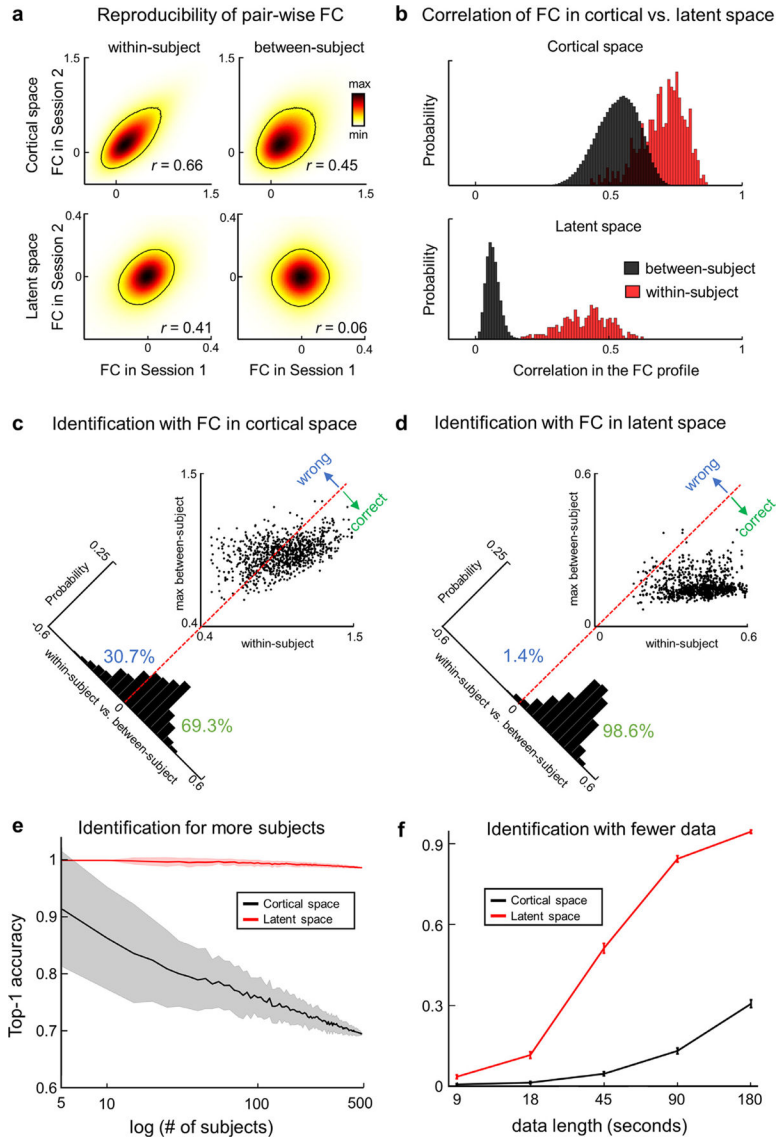

Fig. 8. Individual identification based on correlations between latent variables or cortical parcels.

(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 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

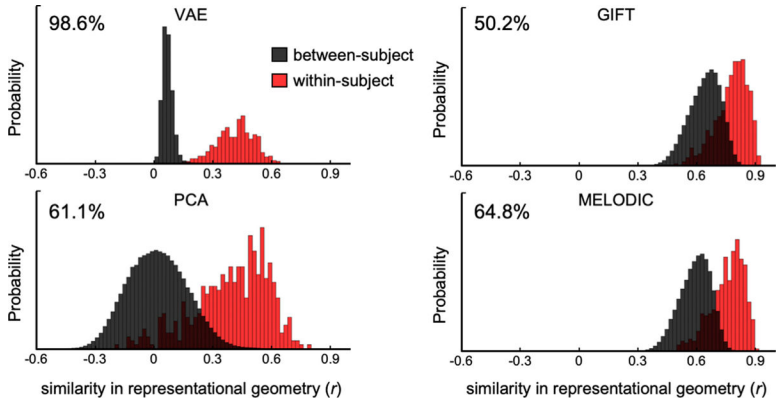

Fig. 9. Individual variation shown in nonlinear vs. linear representations.

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 left) or in the linear latent space obtained by PCA (bottom left), ICA by GIFT (top right) or MELODIC (bottom right). 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 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 1

The accuracies of subject identification when ICA-FIX based denoising was applied to the rsfMRI data for both sessions, only one session, and neither sessions for each subject.

			Session 1	
			Clean(%)	**Noisy(%)**
Session 2	Latent Space	Clean	98.6	90.7
		Noisy	91.5	94.1
	Cortical Space	Clean	69.3	47.2
		Noisy	47.5	76.9