Brain functional network connectivity interpolation characterizes the neuropsychiatric continuum and heterogeneity

Xinhui Li,[1,2*] Eloy Geenjaar,[1,2] Zening Fu,[1] Godfrey D. Pearlson,[3] Vince D. Calhoun[1,2]

[1]Tri-institutional Center for Translational Research in Neuroimaging and Data Science,

Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA

[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[3]Department of Psychiatry, School of Medicine, Yale University, New Haven, CT, USA

*Correspondence: xinhuili@gatech.edu

March 2, 2026

Abstract: Psychiatric and neurodevelopmental disorders such as schizophrenia (SZ) and autism spectrum disorder (ASD) are challenging to characterize in part due to their heterogeneous presentation in individuals, with symptoms now believed to exist on a continuum. Conventional diagnostic and neuroimaging analytical approaches rely on subjective assessment or group differences but typically ignore progression between groups or heterogeneity within a group. To estimate the neuropsychiatric continuum and heterogeneity, we propose a functional network connectivity (FNC) interpolation framework based on a variational autoencoder (VAE) using static FNC (sFNC) and dynamic FNC (dFNC) data from controls and patients with SZ or ASD. We demonstrate that VAEs significantly outperform a linear baseline and a semi-supervised counterpart. For both sFNC and dFNC interpolation, the generated results effectively capture representative and generalizable properties in the original data. The interpolated continua from controls to patients in both disorders reveal group-wise gradients characterized by reduced positive correlations within the auditory, sensorimotor, and visual networks, as well as between the subcortical and cerebellar domains. In contrast, anti-correlations weaken between the subcortical domain and sensory domains,

and between the cerebellar domain and sensory regions. Finally, we show examples of how to generate continuous FNC data following pathological or state-based trajectories in the VAE latent space. The proposed framework offers added advantages over traditional methods, including data-driven discovery of hidden relationships, visualization of individual differences, imputation of missing values along a continuous spectrum, and estimation of the stage where an individual falls within the continuum.

Keywords: functional network connectivity interpolation; psychosis continuum; psychosis heterogeneity; variational autoencoder; schizophrenia; autism spectrum disorder

1 Introduction

"Happy individuals are all alike; every unhappy individual is unhappy in their own way."

– adapted from Anna Karenina by Leo Tolstoy

Psychiatric and neurodevelopmental disorders such as schizophrenia (SZ) and autism spectrum disorder (ASD) are recognized as leading causes of the global disease burden, affecting hundreds of millions of individuals worldwide (Collaborators et al., 2022). SZ is a lifelong mental disorder characterized by symptoms such as delusions, hallucinations, disorganized speech, catatonic behavior, and diminished emotional expression (McCutcheon et al., 2020). ASD is a developmental disability defined by social communication difficulties, and repetitive and restrictive behaviors (Lord et al., 2018). SZ and ASD show overlapping features, including genetic risk factors (O'Connell et al., 2018; The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017), symptom profiles (Barneveld et al., 2011; Konstantareas & Hewitt, 2001), and brain abnormalities (H. Chen et al., 2017; Du et al., 2021; Fu et al., 2021; Yoshihara et al., 2020). Moreover, both SZ and ASD are known to be highly heterogeneous across individuals (Masi et al., 2017; McCutcheon et al., 2020). Namely, individuals within the disorder group may differ in various aspects, such as genetic traits (Abrahams & Geschwind, 2008; Bill & Geschwind, 2009; Di Biase et al., 2022; Geschwind et al., 2015; Pulver et al., 2000), clinical symptoms (Ahmed et al., 2018; Georgiades et al., 2013; Kim et al., 2016; Mohr et al., 2004; Morales-Hidalgo et al., 2018), brain structure and function (Alnæs et al., 2019; Amaral et al., 2008; Brugger & Howes, 2017; Hahamy et al., 2015; Hong et al., 2018; Marín, 2012), cognitive function (Bora, 2016; Feczko et al., 2018; Joyce & Roiser, 2007), developmental trajectories (Carpenter Jr & Kirkpatrick, 1988; Fountain et al., 2012; Lord et al., 2015), and other relevant characteristics. In addition, psychiatric symptoms, such as hallucinations and delusions, may exist along a multidimensional

continuum in the general population, varying in severity, frequency, and degree of conviction (DeRosse & Karlsgodt, 2015; Johns & Van Os, 2001; Morales-Hidalgo et al., 2018; Strauss, 1969; Van Os et al., 2000, 2009; Verdoux & van Os, 2002). Estimating such a continuous spectrum and imputing missing values on the continuum remains underexplored in the neuropsychiatric literature. Considering high prevalence and severity of both disorders, as well as their similarity and complexity, it is important to develop reliable diagnostic and analytical methods to characterize the continuum and heterogeneity of SZ and ASD.

However, there are several limitations in conventional evaluation approaches. First, the current diagnostic system based on the Diagnostic and Statistical Manual of Mental Disorders Fifth Edition (DSM-V) (American Psychiatric Association, 2013) relies on expert assessment or patient self-report data. Such subjective measures are insufficient to accurately capture biological deficits and symptom progression of psychiatric disorders, hindering the development of more effective treatments (Craddock & Mynors-Wallis, 2014; Cuthbert & Insel, 2013; Scott & Nelson, 2024; Su et al., 2020). Thus, it is critical to discover objective biomarkers for psychosis characterization. Recent neuroimaging studies have shown that SZ and ASD can be characterized by resting-state static and dynamic brain connectivity, derived from functional magnetic resonance imaging (fMRI) data (H. Chen et al., 2017; Damaraju et al., 2014; Du et al., 2021; Fu et al., 2021; Guo et al., 2019; Hahamy et al., 2015; Hull et al., 2017; Iraji et al., 2019, 2021; Yan et al., 2024; Yoshihara et al., 2020; Yu et al., 2012). Second, group average approaches have been widely used to analyze patient versus control functional connectivity patterns (Damaraju et al., 2014; Du et al., 2020; Y. He et al., 2020; Rabany et al., 2019). Yet, given the heterogeneous nature of SZ and ASD, group averages alone are not sufficient to characterize their multidimensional continua or multifaceted heterogeneity (Segal et al., 2024). Third, supervised learning methods, which utilize diagnostic labels, have been widely developed to predict disorder conditions from functional connectivity data (C. P. Chen et al., 2015; Rabany et al., 2019; Rashid et al., 2016; Zeng et al., 2018). As a result, the success of supervised learning models inherently depends on the validity of diagnostic labels. However, these diagnostic labels derived from current psychiatric nosologies, aside from their subjectivity, indicate only the current status of a disorder and cannot predict future disorder progression for early intervention (Scott & Nelson, 2024). Additionally, binary or multi-class supervised classification approaches alone neither explain individual differences within a group nor estimate psychotic continua between groups. To address these challenges, it is necessary to apply *unsupervised* learning approaches to brain imaging data to help characterize objective biomarkers, individual variability, and psychosis continua.

Variational autoencoders (VAEs) (Kingma & Welling, 2014) are one type of generative model that can learn data distributions in an *unsupervised* manner. A VAE consists of an encoder and a decoder. The encoder projects input data onto a latent space, and the decoder reconstructs the input data from the learned latent distributions. There are two key benefits of using a VAE. First, the encoder can

perform *nonlinear* dimension reduction and approximate low-dimensional latent distributions from high-dimensional data. Second, the decoder can generate continuous synthetic data that closely resemble the observed data by sampling from the learned distributions. Recent studies have demonstrated that latent representations of resting-state fMRI data learned by a VAE can be used to identify SZ (Geenjaar et al., 2021) and ASD (F. Zhang et al., 2022), and characterize spatiotemporal dynamics (X. Zhang et al., 2021). Traditional dimension reduction approaches such as principal component analysis (PCA) (Jolliffe, 1986) and independent component analysis (ICA) (Hyvärinen & Oja, 2000) cannot naturally handle missing values, and thus cannot infer the continuous spectrum from the data. Probabilistic PCA (PPCA) has been developed to address the limitation of PCA by estimating a probabilistic model of observed data (Tipping & Bishop, 1999), but it cannot estimate complex nonlinear relationships due to its linear nature. Deep learning models such as autoencoders nonlinearly compress the data into latent features but do not provide a probabilistic estimate of the data continuum. Other deep generative models also have limitations in this specific task. For example, generative adversarial networks (GANs) (Goodfellow et al., 2020) are known for mode collapse (Che et al., 2016; Salimans et al., 2016) and instability issues (Barnett, 2018). Diffusion models (Ho et al., 2020) learn latent variables of the same dimensionality as the original data, rather than learning a low-dimensional, interpretable latent space. In addition, inference in diffusion models is computationally intensive, which typically requires hundreds of denoising steps. In contrast, the VAE provides a low-dimensional manifold that allows effective *interpretation* and efficient *interpolation* of latent features, making it a promising architecture for estimating psychotic continuum and characterizing individual variability.

Here, we propose an interpretable and generative unsupervised learning framework based on a VAE to interpolate within latent spaces learned from static functional network connectivity (sFNC) and dynamic functional network connectivity (dFNC) derived from resting-state fMRI. Interpolation – mathematically defined as the process of estimating an unknown value based on surrounding known values – can be applied to estimate an FNC matrix from its neighboring FNC matrices, assuming continuous changes between subjects. For sFNC interpolation, individual sFNC matrices are organized on a two-dimensional (2D) grid based on their similarity. A trained VAE is subsequently used to interpolate the sFNC matrices by sampling from the latent distributions along this 2D grid. For dFNC interpolation, we train the VAE using all windowed dFNC matrices from the training subjects, performed k-means clustering (Hartigan & Wong, 1979) on the VAE latent features, and then generated dynamic states based on the k-means clusters. For comparison, dynamic states are also identified by applying the k-means clustering algorithm on the dFNC matrices directly.

Our study highlights the benefits of maximizing data transparency to visualize individual differences within a group and continuous changes between groups. Crucially, sFNC interpolation facilitates examining

individual differences and disorder continua, while dFNC interpolation captures generalizable group-specific dynamic states and intermediate state transition patterns, providing both static and dynamic comprehensive views of the psychiatric illnesses. Moreover, the VAE latent space offers the flexibility to interpolate FNC patterns between samples in the original dataset. We provide examples of how unsupervised generative models can be used to learn new insights by interpolating FNC across subjects, groups, and dynamic states.

2 Methods

2.1 Data

2.1.1 Datasets

We used two resting-state fMRI datasets: the Functional Biomedical Informatics Research Network (FBIRN) dataset (Keator et al., 2016) and the initial release of the Autism Brain Imaging Data Exchange (ABIDE I) dataset (Di Martino et al., 2014).

The FBIRN fMRI scans were acquired using a standard gradient-echo echo-planar imaging (EPI) paradigm with repetition time (TR) $= 2$s, echo time (TE) $= 30$ms, flip angle (FA) $= 77^\circ$, 162 volumes, 32 sequential ascending axial slices of 4mm thickness and 1mm skip. Participants had their eyes closed during the scan. In addition to fMRI data, we utilized Positive and Negative Syndrome Scale (PANSS) scores (Kay et al., 1987), with positive and negative subscale scores summed separately. We also used six cognitive domain scores (speed of processing, attention/vigilance, working memory, verbal learning, visual learning, and reasoning/problem solving), as well as composite scores derived from the Computerized Multiphasic Interactive Neurocognitive System (CMINDS) (van Erp et al., 2015). All CMINDS cognitive scores were z-scored based on the mean and standard deviation of the control group. The original FBIRN dataset includes 311 subjects. 45 subjects were excluded because they had one or more missing or invalid diagnosis and cognitive scores. 266 subjects with valid subject measures were used in the subsequent analysis, including 125 subjects labeled as SZ (age mean $\pm$ s.d.: 38.88 ± 11.19 years; 95 males, 30 females) and 141 controls (CTR) (age mean $\pm$ s.d.: 37.15 ± 11.03 years; 103 males, 38 females).

The original ABIDE I dataset includes 869 subjects. We excluded 107 subjects without valid clinical assessment scores. Then we used 266 subjects with valid clinical assessment scores from three sites, including 133 subjects labeled as ASD (age mean $\pm$ s.d.: 17.01 ± 7.20 years; 119 males, 14 females) and 133 controls (age mean $\pm$ s.d.: 16.05 ± 5.67 years; 104 males, 29 females). The site-specific

demographics and acquisition parameters of the ABIDE I participants are presented in Appendix A.

The statistics of subject measures in each dataset are shown in Appendix B. For each dataset, 85% of subjects were used as the training set ($N_{\text{train}} = 225$) and 15% of subjects with balanced labels were used as the holdout test set ($N_{\text{test}} = 41$). The detailed demographic information of the subjects used in this study is described in Table 1.

2.1.2 Data preprocessing

The fMRI data were preprocessed using the statistical parametric mapping toolbox (SPM12, http://www.fil.ion.ucl.ac.uk/spm/) (Ashburner et al., 2014) in the MATLAB 2016 environment. We removed the first five scans for signal equilibrium and for participants' adaptation to the scanner environment. We then performed rigid body motion correction using SPM, followed by slice timing correction. Subsequently, the fMRI data were registered to the standard Montreal Neurological Institute (MNI) space using an EPI template and resampled to $3 \times 3 \times 3$mm^3 isotropic voxels. The resampled fMRI data were further smoothed using a Gaussian kernel with a full width at half maximum (FWHM) $= 6$mm.

2.1.3 NeuroMark functional network connectivity

We applied the fully automated NeuroMark pipeline (Du et al., 2020) to the preprocessed fMRI data to extract subject-specific functional components and their corresponding time courses (TCs). NeuroMark, implemented in the group ICA of fMRI toolbox (GIFT; http://trendscenter.org/software/gift) leverages network templates (available in GIFT and at http://trendscenter.org/data) originally derived by separately computing group-level independent components (ICs) from two independent datasets of healthy individuals: the Human Connectome Project (HCP) (Van Essen et al., 2013) and the Brain Genomics Superstruct Project (GSP) (Holmes et al., 2015). ICs from the two datasets were matched based on spatial correlations, and 53 intrinsic connectivity networks (ICNs) with correlations greater than 0.4 selected as the network templates. The less noisy ICNs from the GSP dataset were selected for inclusion in the NeuroMark_fMRI_1.0 template (Du et al., 2020). These 53 ICNs were then ordered and assigned to seven functional domains according to their anatomical and functional properties, including the subcortical (SC), auditory (AU), sensorimotor (SM), visual (VI), cognitive control (CC), default mode (DM), and cerebellar (CB) domains (Allen et al., 2014; Du et al., 2020). The components in this template were then used as spatial priors to perform spatially constrained ICA to compute ICNs and TCs for each subject for the FBIRN and ABIDE I datasets. The NeuroMark framework allows us to (1) estimate the ICA separately for each subject, thus preventing data leakage, (2) fully automate the ICA pipeline, and

(3) estimate ICNs for each subject in the FBIRN and ABIDE I datasets which are aligned and comparable. Four postprocessing steps were performed on the extracted TCs, including detrending, removing head motion, despiking, and filtering with a 0.15Hz low-pass filter.

Static functional network connectivity (sFNC) was subsequently computed as the pairwise Pearson correlation between the TCs of 53 ICNs, leading to a 53×53 matrix. Dynamic functional network connectivity (dFNC) was estimated using a sliding window approach and a graphical LASSO method (Friedman et al., 2008). A tapered window to segment the TCs was created by convolving a rectangle (width $= 20$, TR $= 2$s, TRs $= 40$s) with a Gaussian kernel ($\sigma = 3$TRs). The window width covered 20 TRs with a step size of 1 TR, leading to 137 windows in the FBIRN dFNC data and 168 windows in the ABIDE I dFNC data. For further details on data processing, please refer to Fu et al., 2021.

Table 1: **Dataset demographic information.** Demographics in the FBIRN and ABIDE I datasets are shown below. (N: number of total subjects; N_{male}: number of male subjects; N_{female}: number of female subjects; N_{train}: number of subjects in the training set; N_{test}: number of subjects in the test set; s.d.: standard deviation.)

Dataset	Label	N	N_{male}	N_{female}	N_{train}	N_{test}	Age mean $\pm$ s.d. (years)	Age range (years)
FBIRN	SZ	125	95	30	105	20	38.88 ± 11.19	$18 - 62$
FBIRN	CTR	141	103	38	120	21	37.15 ± 11.03	$19 - 59$
ABIDE I	ASD	133	119	14	113	20	17.01 ± 7.20	$7 - 39$
ABIDE I	CTR	133	104	29	112	21	16.05 ± 5.67	$7 - 32$

2.2 Variational autoencoders

We used an unsupervised generative model, a variational autoencoder (VAE) (Kingma & Welling, 2014), as the backbone of the interpolation framework. We trained one VAE for each dataset (FBIRN or ABIDE I) and each data type (sFNC or dFNC). Here, we briefly explain the VAE objective function and provide a step-by-step derivation in Appendix C.

Let $\mathcal{D}_{\text{sFNC}} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}^{(N)}\right\}$ denote a dataset of N sFNC samples $\mathbf{x}^{(i)} \in \mathbb{R}^V$, where N is the number of subjects and V is the number of features in the sample. Each sample $\mathbf{x}^{(i)}$ is the vectorized upper triangle of a symmetric sFNC matrix, thus the dimensionality of each sample can be calculated as $V = \frac{1}{2} \times (53 \times 53 - 53) = 1378$. For dFNC data, we treat each windowed dFNC matrix as an independent sample, and denote the dFNC dataset as $\mathcal{D}_{\text{dFNC}} = \left\{\mathbf{x}^{(1,1)}, \ldots, \mathbf{x}^{(i,t)}, \ldots, \mathbf{x}^{(N,T)}\right\}$. The number of total samples is the product of the number of subjects N and the number of windows T ($N \times T$). Each

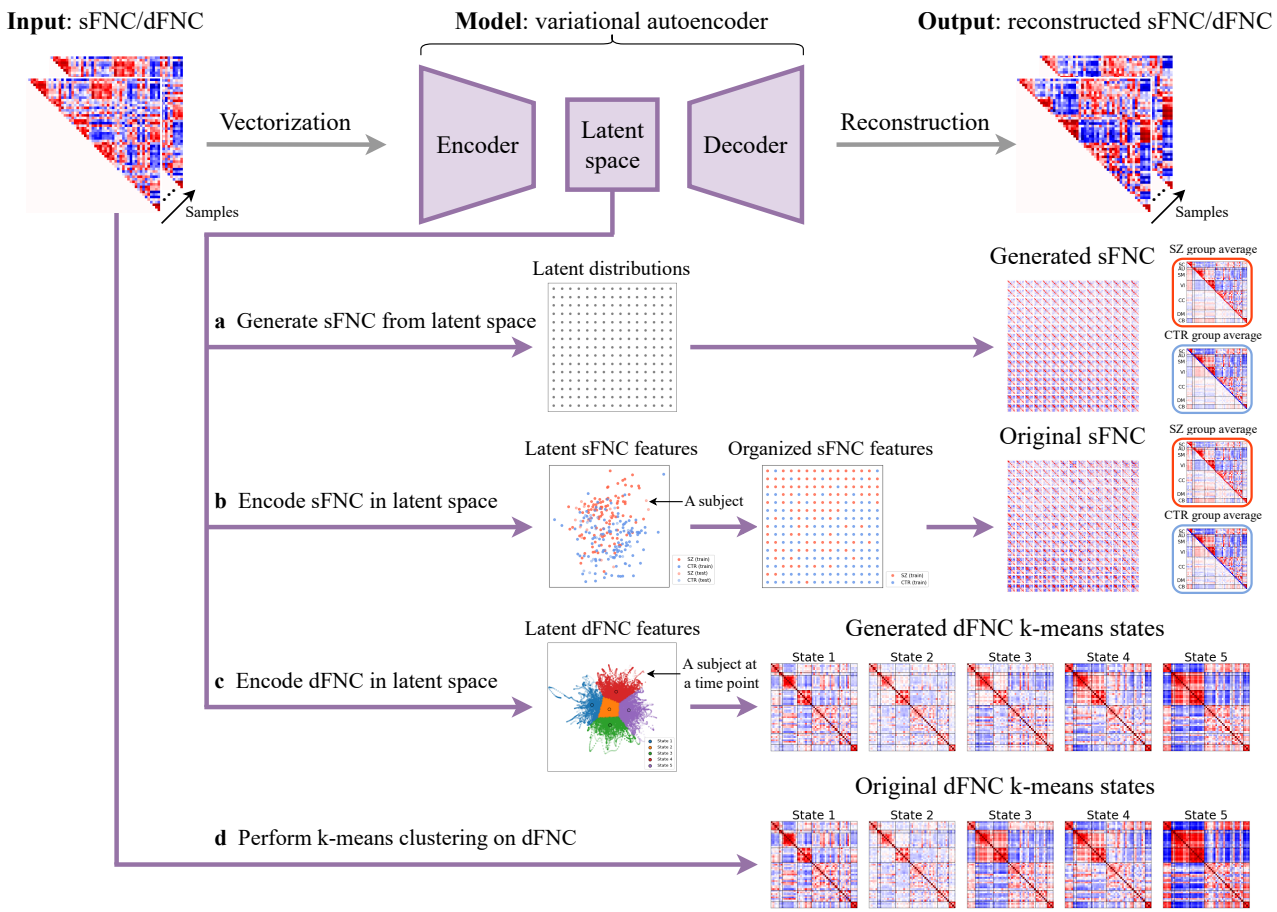


Figure 1: **Overview of functional network connectivity interpolation framework.** The upper triangle of each sFNC or windowed dFNC matrix was vectorized and used as the input to the VAE. During the training stage (gray arrows), a VAE was trained to learn a 2D latent variable $\mathbf{z}$ by minimizing the reconstruction loss between the input and the reconstructed output, while the KL-divergence regularized the difference between the prior distribution $p(\mathbf{z})$ and the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$. During inference time, four pathways (purple arrows) are described as follows: (a) The sFNC matrices were generated by uniformly sampling from the prior distribution $p(\mathbf{z})$ in the latent space and then forward propagating through the trained decoder. (b) The original sFNC matrices were projected onto the latent space by the trained encoder, the corresponding latent features were then mapped to a 2D grid using the JV algorithm, and the corresponding original sFNC matrices were then displayed on the 2D grid. (c) The dFNC matrices were projected onto the latent space by the trained encoder. The k-means clustering algorithm was then performed on the dFNC latent features. The dFNC matrices were generated by sampling from the posterior distributions $q_\phi(\mathbf{z}|\mathbf{x})$. The generated dFNC state was then computed as the element-wise median of the generated dFNC matrices in each cluster. (d) The k-means clustering algorithm was performed on the dFNC matrices directly. The original dFNC state was then computed as the element-wise median of the original dFNC matrices in each cluster.

sample $\mathbf{x}^{(i,t)} \in \mathbb{R}^V$ is the flattened upper triangle of a windowed dFNC matrix from the i-th subject and the t-th window ($V = 1378$). For pedagogical purposes, we will explain the generative model using a single sample $\mathbf{x} \in \mathbb{R}^V$.

Let $\mathbf{z} \in \mathbb{R}^D$ denote a continuous latent variable which represents the underlying factors in the input data. We initially defined a two-dimensional latent space (i.e. $D = 2$) to better visualize the FNC matrices. Note that it is also possible to define a latent space with a different dimension. We assume that the latent variable $\mathbf{z}$ is sampled from a Gaussian prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ and the observed data variable $\mathbf{x}$ is sampled from the conditional likelihood distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ parameterized by $\boldsymbol{\theta}$, with the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. Our goal is to estimate the posterior distribution to perform inference on the latent variable. According to Bayes' theorem, the posterior can be derived as:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}}. \tag{1}$$

However, the computation of the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is analytically intractable because the integral of the marginal likelihood $\int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$ cannot be evaluated in closed form. Hence, we perform *variational inference* to approximate the true posterior distribution with a simpler distribution.

Here, we approximate the posterior distribution using a multivariate Gaussian distribution with diagonal covariance, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \mathrm{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})\right)\right)$. The Kullback-Leibler (KL) divergence is used to quantify the difference between the true posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ parameterized by $\boldsymbol{\theta}$ and the multivariate Gaussian approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ parameterized by $\boldsymbol{\phi}$. The objective function, also known as the *evidence lower bound (ELBO)*, aims to maximize the negative KL divergence:

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}} &= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} -D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right) \\
&\equiv \max_{\boldsymbol{\theta},\boldsymbol{\phi}} -D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\right) \\
&= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right).
\end{aligned} \tag{2}$$

Note that the KL divergence is non-negative, i.e. $D_{\text{KL}}(\cdot) \geq 0$. Thus, the ELBO is the lower bound to the log likelihood of generating the observed data, i.e. $\mathcal{L}_{\text{ELBO}} \leq \log p_{\boldsymbol{\theta}}(\mathbf{x})$. An encoder parameterized by the variational parameters $\boldsymbol{\phi}$ (i.e. $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$) and a decoder parameterized by the generative parameters $\boldsymbol{\theta}$ (i.e. $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$) are trained to maximize the ELBO, thereby simultaneously maximizing the log likelihood of generating the observed data samples and minimizing the KL divergence between the approximate posterior and the true posterior. Maximizing the ELBO is equivalent to minimizing the objective function

as follows (see Appendix C for more details):

$$\mathcal{L}_{\text{VAE}} = \min_{\boldsymbol{\theta},\boldsymbol{\phi}} \ \frac{1}{V}\sum_{i=1}^{V}(x_i - \hat{x}_i)^2 - \frac{1}{2}\sum_{j=1}^{D}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right), \tag{3}$$

where x_i and $\hat{x}_i$ are the i-th element of the observed data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}$, respectively; μ_j and σ_j are the j-th element of $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})$, respectively. To train the model using gradient-based optimization, the *reparameterization trick* (Kingma & Welling, 2014) is used: $\mathbf{z} = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) \odot \boldsymbol{\epsilon}$, where $\odot$ represents the element-wise product, and $\boldsymbol{\epsilon} \in \mathbb{R}^D$ is a noise variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By optimizing the loss, the VAE is capable of learning the underlying factors in the dataset and generating continuous synthetic data by interpolating between latent samples using the learned factors.

A multilayer perceptron (MLP) was implemented as the encoder and decoder in the VAE. We performed a hyperparameter search to select the optimal model architecture that yielded the best training performance from four different architectures for each dataset and each data type separately (see Appendix F.1). The model was trained using the Adam optimizer (Kingma & Ba, 2017) with an initial learning rate of 0.001. We implemented a learning rate scheduler to reduce the learning rate by a factor of 0.1 if the loss value did not improve for 10 epochs. Each model was trained for 1000 epochs with a batch size of 16 and 512 for the sFNC and dFNC data, respectively. We implemented an early stopping mechanism – if the loss value did not decrease after 20 epochs, the training would stop. All training processes stopped within the predefined number of epochs. For each model, we initialized the model weights with 10 different random seeds and recorded the results across 10 runs. The model was implemented in Python using the PyTorch framework (Paszke et al., 2019) and trained with NVIDIA A40 GPUs.

2.3 Model comparison

To comprehensively evaluate model performance, we compared the VAE with a linear baseline method – probabilistic principal component analysis (PPCA) (Tipping & Bishop, 1999) and a semi-supervised alternative – identifiable variational autoencoder (iVAE) (Khemakhem et al., 2020). We review the generative model and optimization approach used in PPCA in Appendix D, and describe the generative model and objective function for iVAE in Appendix E.

2.3.1 Probabilistic principal component analysis

PPCA is a probabilistic extension of PCA. Unlike conventional PCA, PPCA can estimate missing values in the dataset and generate samples from the learned distribution. Recent work has shown that the principal components learned by PPCA can be fully recovered by linear VAEs (Lucas et al., 2019). Thus, PPCA can be viewed as equivalent to linear VAEs. We implemented the expectation-maximization (EM) algorithm to estimate the model parameters, as described in Tipping and Bishop, 1999. We randomly initialized the model parameters with 10 seeds and reported the results across 10 runs.

2.3.2 Identifiable variational autoencoders

Although a VAE is trained to learn latent distributions, there is no guarantee of identifiability[1] of the model parameters in general. The iVAE has been developed to address the identifiability problem by conditioning on an auxiliary variable such as class label or time index. It has been shown that the iVAE recovers conditionally identifiable latent variables up to permutations and pointwise nuonlinear transformations (Khemakhem et al., 2020). Since our main interest lies in psychiatric characteristics, we used the one-hot encoded diagnostic label as the auxiliary variable in the iVAE. We performed a similar hyperparameter search experiment to select the optimal iVAE architecture (Appendix F.2). Each iVAE model was initialized using 10 different random seeds.

2.4 Functional network connectivity interpolation framework

We propose a functional network connectivity interpolation framework based upon the VAE (Figure 1). For each dataset, a VAE was trained on the sFNC data or the dFNC data separately. The vectorized upper triangle of each sFNC or windowed dFNC matrix was used as the input to the VAE. The VAE was trained to learn a latent variable $\mathbf{z}$ by minimizing the reconstruction loss between the input and the reconstructed output, while the KL divergence regularized the difference between the prior distribution $p(\mathbf{z})$ and the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$. After the training stage, we performed sFNC interpolation and dFNC interpolation separately, as described below.

[1] A statistical model is identifiable if the mapping from each set of its parameters to the distribution of observed data is injective (Lehmann & Casella, 2006).

2.4.1 Static functional network connectivity interpolation

We first fed the original training or test data to the trained encoder and extracted the corresponding two-dimensional features in the latent space. To generate synthetic data, we defined a 15×15 grid of evenly spaced 2D coordinates in the latent space (as shown in Figure 1a), with the range determined by the 80th percentile of the absolute values of latent features from the training set. To display all training subjects' sFNC matrices on a 2D grid, we select the grid dimension such that the total number of grid points (i.e., the square of the dimension) approximates the number of training samples. These coordinates (values of latent variables) were then forward propagated through the trained decoder to generate continuous synthetic sFNC matrices, which were visualized on the 2D grid. To examine the corresponding original data, the Jonker-Volgenant (JV) algorithm (Jonker & Volgenant, 2005) for the linear sum assignment problem was subsequently used to map these latent features to the nodes on the 2D grid by minimizing the pairwise Euclidean distance between the positions of the latent features and the 2D grid nodes. We then displayed the original sFNC matrices on the 2D grid correspondingly. This allows us to visualize the generated sFNC matrices and the original sFNC matrices on the 2D grid, respectively.

2.4.2 Dynamic functional network connectivity interpolation

We projected the windowed dFNC matrices onto the VAE latent space and extracted their corresponding latent features. Subsequently, the k-means clustering algorithm was applied to these dFNC latent features, using a fixed random seed and varying the number of states k from 2 to 9. The dFNC matrices were generated by sampling from the learned posterior distributions $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The group-specific generated dFNC state was then computed as the element-wise median of the generated dFNC matrices in each cluster for each group. For comparison, the k-means clustering algorithm was performed on the original dFNC matrices directly with the same search window of k, and then the optimal k value was identified using the same elbow criterion. The group-specific original dFNC state was then computed as the element-wise median of the original dFNC matrices in each cluster for each group.

To further quantify the properties of the dFNC states, we calculated the group-specific dwell time (DT) for each state α, defined as the number of FNC matrices per state $N^{\alpha}_{\mathsf{FNC}_g}$ divided by the number of subjects per state N^{α}_{g} for each group g:

$$\mathsf{DT}^{\alpha}_{g} = \frac{N^{\alpha}_{\mathsf{FNC}_g}}{N^{\alpha}_{g}}. \tag{4}$$

We also calculated the group-specific transition probability matrix $\mathbf{P}_g \in \mathbb{R}^{k\times k}$ to evaluate how likely a state changes to another between two consecutive time points. Specifically, we first counted the number of transitions from a state α to a different state β through the entire time course T for each group g:

$$N_g^{\alpha\to\beta} = \frac{1}{N_g}\sum_{i\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\delta\left(S\left(\mathbf{x}^{(i,t)}\right)=\alpha, S\left(\mathbf{x}^{(i,t+1)}\right)=\beta\right), \tag{5}$$

where $\delta(\cdot)=1$ when the transition happens from the state α at time t to the state β at time $t+1$ for the i-th subject, otherwise $\delta(\cdot)=0$. $S(\cdot)\in\{1,\ldots,k\}$ indicates the state of a sample. $\mathcal{G}$ is the set of subject indices for a group and N_g is the corresponding number of subjects for that group.

Then, the off-diagonal entry of the group-specific transition probability matrix $\mathbf{P}_g(\alpha,\beta)$ $(\alpha\neq\beta)$ is defined as the number of transition times from the state α to the state β divided by the total number of transition times from the state α to all the other states:

$$\mathbf{P}_g(\alpha,\beta) = P\left(S\left(\mathbf{x}^{(i,t+1)}\right)=\beta \mid S\left(\mathbf{x}^{(i,t)}\right)=\alpha, i\in\mathcal{G}\right) = \frac{N_g^{\alpha\to\beta}}{\sum_{\gamma\in\{1,\ldots,k\}\setminus\{\alpha\}} N_g^{\alpha\to\gamma}}. \tag{6}$$

3 Results

3.1 sFNC interpolation captures psychosis continuum and heterogeneity

We first compared VAE performance with two alternative methods, probabilistic principal component analysis (PPCA) and identifiable variational autoencoders (iVAEs). We hypothesized that nonlinear VAEs that capture more complex relationships from observed data would achieve better performance than linear PPCA. As expected, the Pearson correlations between the generated and original sFNC matrices from VAEs are significantly higher than those from PPCA (Figure 2, $p<0.0001$ for both FBIRN and ABIDE I, Wilcoxon signed-rank test, Bonferroni correction). More specifically, the average correlations for VAEs are 0.765 and 0.724, while those for PPCA are 0.711 and 0.694 for FBIRN and ABIDE I data, respectively. Surprisingly, VAEs also significantly outperformed iVAEs in this interpolation task (Figure 2, $p<0.05$ for FBIRN, $p<0.01$ for ABIDE I, Wilcoxon signed-rank test, Bonferroni correction). The average correlations for iVAEs are 0.741 for the FBIRN dataset and 0.708 for the ABIDE I dataset, both slightly lower than those for VAEs. By design, we consider VAEs better suited for learning low-dimensional continuous representations in this case due to their *nonlinear* and *unsupervised* nature. Our empirical results support this theoretical advantage, demonstrating that VAEs significantly outperformed both a linear baseline (PPCA) and a semi-supervised alternative (iVAE). Therefore, we

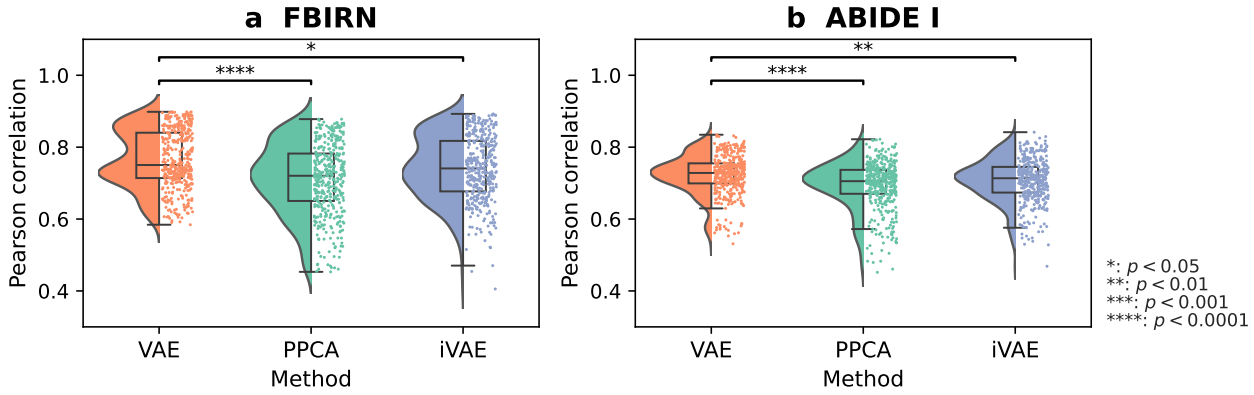


Figure 2: **Pearson correlations between generated and original sFNC matrices.** (a) FBIRN test data. (b) ABIDE I test data. The scatter-box-violin plot shows the data distribution, median and interquartile range of Pearson correlations between the original and generated sFNC matrices in the FBIRN and ABIDE I test sets for VAE, PPCA, and iVAE, across 10 random seeds. In both datasets, VAEs significantly outperformed iVAEs and PPCA in the interpolation task (Wilcoxon signed-rank test, Bonferroni correction).

focused on VAEs for all subsequent experiments in our study.

We confirmed that the VAE representations were robust to potential confounding factors, including age, sex, and collection site (Appendix G), and subsequently performed sFNC interpolation as described in Section 2.4.1. By comparing the generated sFNC matrices and the original sFNC matrices, we observe a high degree of correspondence between the generated and original results for both datasets (Figures 3, 4; zoom-in version in Appendix H). Group-specific patterns, group pattern alterations, and individual differences can be visualized on the 2D grid layout. By examining the group-specific patterns in both the generated and original sFNC matrices, we observe focal modularity in the top half dominated by the patient group, and highly modular and polar patterns in the bottom half dominated by the control group. Specifically, the SZ patient group shows weak functional connectivity among the auditory (AU), sensorimotor (SM), and visual (VI) domains (Figure 3). Similarly, the ASD patient group also shows weak connectivity in these sensory networks (Figure 4). In contrast, the control group shows strong functional connection within the sensory domains (AU, SM, VI), as well as between the subcortical (SC) and cerebellar (CB) domains. The average cell-wise standard deviations of the patient group is slightly higher than that of the control group, suggesting slightly higher variability in the patient group (Appendix I). In addition, the generated sFNC matrices reveal continuous pattern alterations between two groups. The continuous sFNC spectrum from the SZ patient group to the control group demonstrates two key connectivity patterns: (1) stronger positive correlations within the sensory domains (AU, SM, VI) and between the SC and CB domains, and (2) stronger negative correlations between the SC domain and the sensory networks, as well as between the CB domain and these sensory regions. We observe a similar

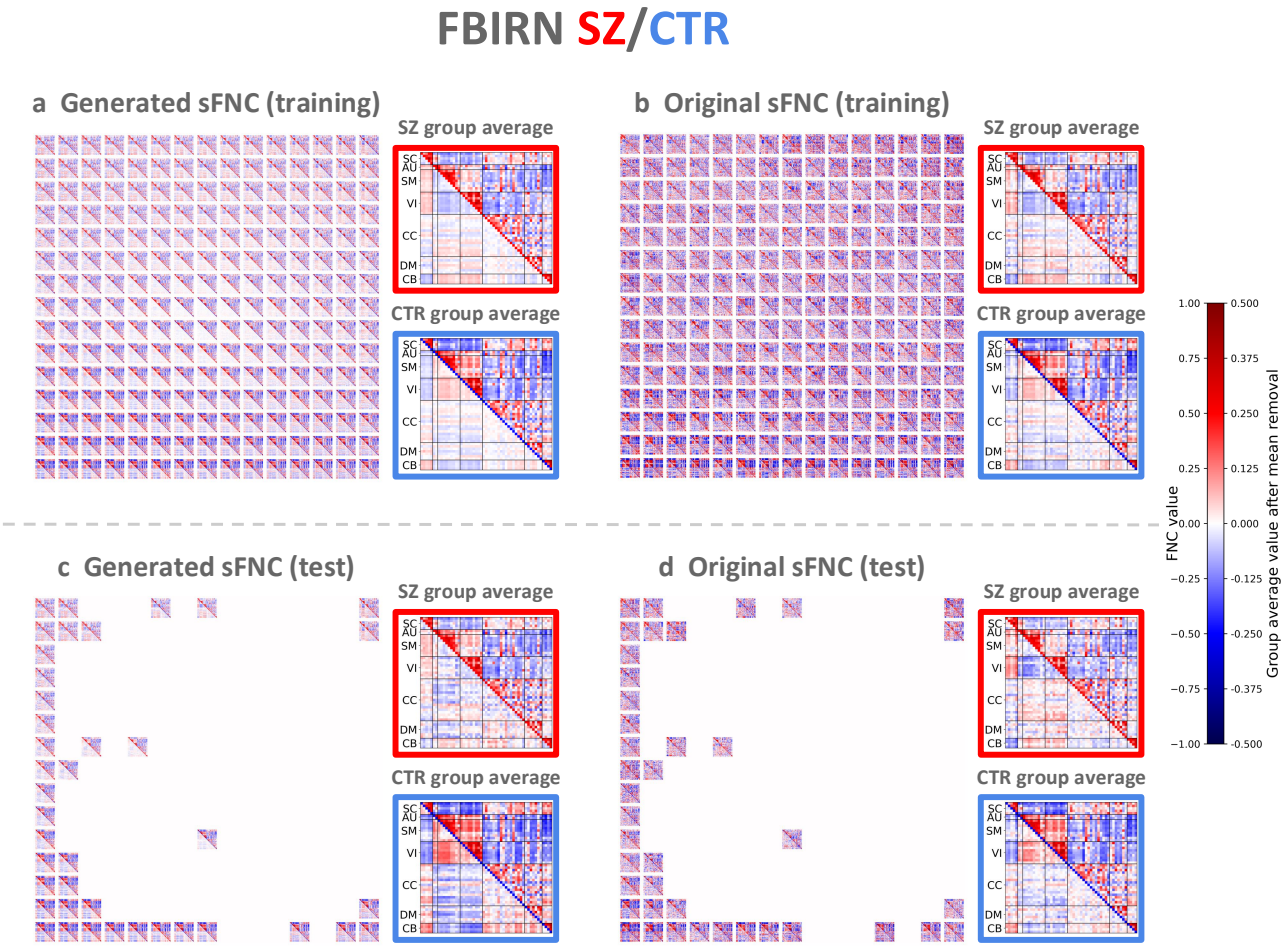


Figure 3: **Generated and original FBIRN sFNC matrices.** (a) Generated sFNC matrices in the training set. (b) Original sFNC matrices in the training set. (c) Generated sFNC matrices in the test set. (d) Original sFNC matrices in the test set. For each sFNC matrix, we retained the upper triangle, removed the element-wise mean across all subjects in the lower triangle to investigate how each subject deviates from the population mean, and color-coded the diagonal based on the diagnosis groups (red: patients; blue: controls). In each red or blue box, the SZ patient or control group average pattern is shown in the upper triangle. The corresponding pattern after removing the element-wise mean across all subjects is shown in the lower triangle (values were multiplied by 2 for better visualization). There is a high degree of correspondence between generated and original sFNC matrices. The generated sFNC matrices capture a continuous spectrum of pattern alterations between groups, while the original sFNC matrices show individual variability across subjects.

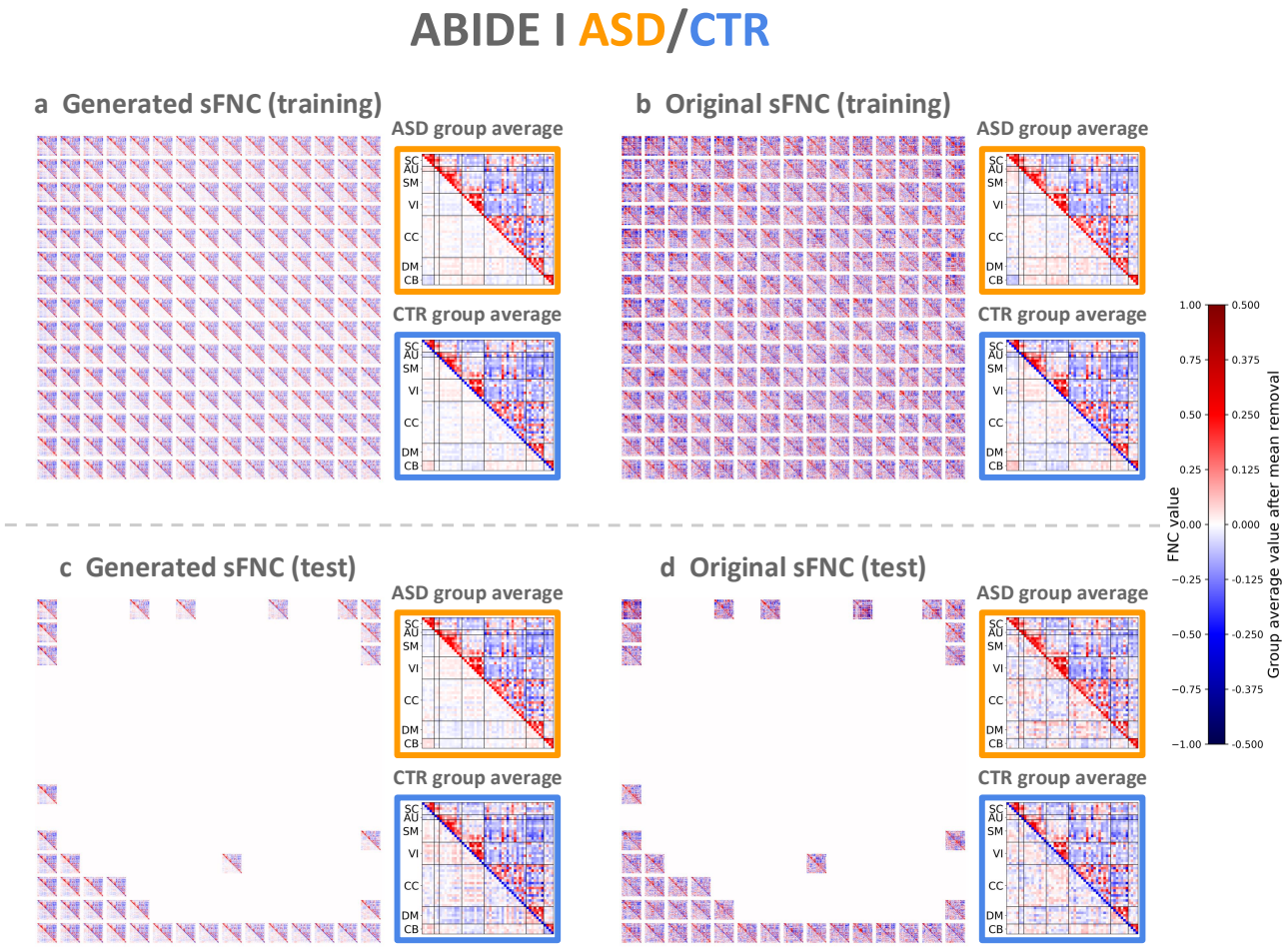


Figure 4: **Generated and original ABIDE I sFNC matrices.** (a) Generated sFNC matrices in the training set. (b) Original sFNC matrices in the training set. (c) Generated sFNC matrices in the test set. (d) Original sFNC matrices in the test set. In each orange or blue box, the ASD or control group average pattern is shown in the upper triangle and the respective group average pattern after removing the element-wise mean across all subjects is shown in the lower triangle. The group average values after mean removal (the lower triangles) in the colored boxes were multiplied by 2 for better visualization.

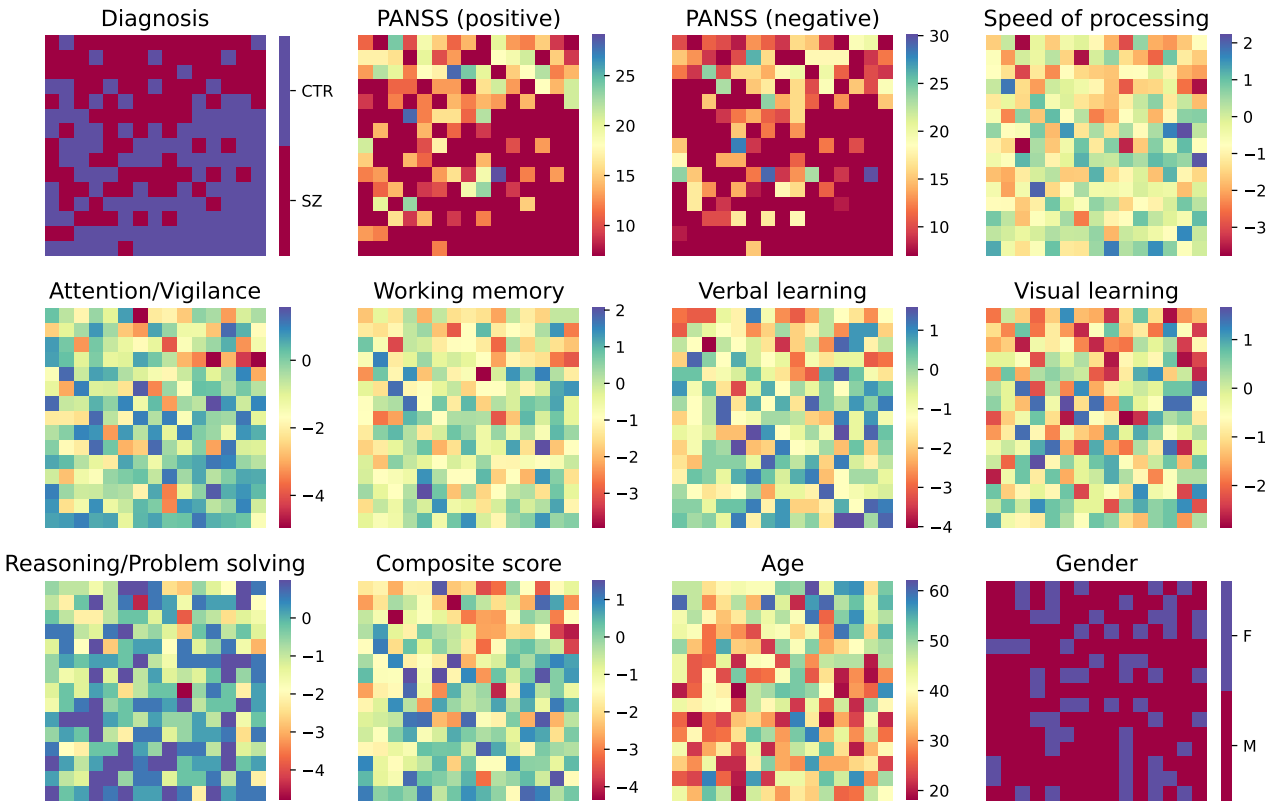


Figure 5: **FBIRN subject measures displayed on the 2D grid.** Based on the individualized mapping between latent features and 2D grid nodes (as in Figure 3), we displayed diagnosis labels, PANSS positive and negative subscale scores, CMINDS cognitive scores (including speed of processing, attention/vigilance, working memory, verbal learning, visual learning, reasoning/problem solving, and composite scores), age and gender information of FBIRN subjects on the 2D grid correspondingly.

continuum of sFNC changes from ASD patients to controls, but the group differences are less pronounced compared to SZ, suggesting that functional connectivity alterations in ASD may not be as substantial as those observed in SZ. Overall, the generated sFNC matrices show representative patterns within a group and continuous alterations between groups, while the original sFNC matrices help to further examine individual variability.

Furthermore, we displayed multiple FBIRN subject measures, including clinical assessment scores, normalized cognitive scores (van Erp et al., 2015), age and gender information, on the 2D grid accordingly (Figure 5), providing interpretable visualization of how each subject measure contributes to the unsupervised learning process. Crucially, we observe that diagnosis is a key factor distinguishing subjects in the latent space, with most patients occupying the upper half of the 2D grid and most controls occupying the lower half. In addition, age appears to be another important factor, as younger subjects tend to cluster in the lower triangle, while older subjects are mostly in the upper triangle. Interestingly, SZ patients located in the lower/left/lower triangular half of the 2D grid have relatively higher (better) cognitive scores than

Table 2: **Cognitive measure statistics of the FBIRN patient cohort.** We partitioned the 2D grid into six regions: the upper 7 rows (upper), the lower 8 rows (lower); the left 7 columns (left), the right 8 columns (right); the upper triangle with the diagonal (upper triangle), and the lower triangle without the diagonal (lower triangle). In the Table, the first row shows the number of SZ patients in each region. The second to eighth rows show the mean (standard deviation) of each cognitive measure for SZ patients across different regions. A higher (better) cognitive score is highlighted in bold for each row. The star ($*$) indicates that cognitive scores in one half are significantly higher than those in the other half ($p < 0.05$, Mann–Whitney U test). Overall, the SZ patients located in the lower/left/lower triangular half of the 2D grid have consistently *higher* cognitive scores compared to those located in the upper/right/upper triangular half, implying that our framework can group SZ patients in the latent space based on their cognitive performance.

Cognitive measure	Upper	Lower	Left	Right	Upper triangle	Lower triangle
Diagnosis	69	36	57	48	67	38
Speed of processing	$-1.35(1.0)$	$\mathbf{-0.92(0.8)}^*$	$\mathbf{-1.13(0.9)}$	$-1.29(1.0)$	$-1.34(1.0)$	$\mathbf{-0.96(0.8)}$
Attention/Vigilance	$-1.41(1.4)$	$\mathbf{-1.05(1.2)}$	$\mathbf{-1.06(1.3)}$	$-1.55(1.4)$	$-1.43(1.4)$	$\mathbf{-1.05(1.2)}$
Working memory	$-1.10(1.0)$	$\mathbf{-0.79(0.9)}$	$\mathbf{-0.91(0.9)}$	$-1.10(1.0)$	$-1.14(1.0)$	$\mathbf{-0.74(0.9)}$
Verbal learning	$-1.38(1.2)$	$\mathbf{-1.11(0.9)}$	$\mathbf{-1.22(1.0)}$	$-1.37(1.2)$	$-1.40(1.2)$	$\mathbf{-1.10(1.0)}$
Visual learning	$-1.09(1.1)$	$\mathbf{-0.81(1.1)}$	$\mathbf{-0.92(1.0)}$	$-1.08(1.2)$	$-1.11(1.1)$	$\mathbf{-0.78(1.0)}$
Reasoning	$-0.87(1.1)$	$\mathbf{-0.40(0.9)}^*$	$\mathbf{-0.66(1.1)}$	$-0.76(1.0)$	$-0.83(1.1)$	$\mathbf{-0.49(0.9)}$
Composite score	$-1.70(1.2)$	$\mathbf{-1.21(1.0)}^*$	$\mathbf{-1.39(1.1)}$	$-1.71(1.3)$	$-1.71(1.3)$	$\mathbf{-1.22(0.9)}^*$

those located in the upper/right/upper triangular half (Table 2). In particular, the speed of processing, reasoning/problem solving, and composite scores in the lower half are significantly higher than those in the upper half, and in addition, composite scores in the lower triangle are significantly higher than those in the upper triangle ($p < 0.05$, Mann–Whitney U test), implying that our framework can group patients in the latent space based on their cognitive ability and further identify subgroups in SZ. To support this application, we applied k-means clustering to the latent representations to identify subgroups, examined their associations with subject measures, and characterized subgroup-specific connectivity patterns (Appendix J). To examine how the dimensionality of the VAE latent space affects downstream subgroup identification, we compared the relationships between k-means subgroups and cognitive scores in both 2D and three-dimensional (3D) latent space, and found that the 3D latent space does not yield additional meaningful subgroups (Appendix J). Moreover, we investigated sFNC patterns across cognitive performance levels and identified both common and distinct connectivity changes across cognitive measures (Appendix K). ABIDE I subject measures (clinical assessment scores, age, and gender) are shown in Figure A.1, Appendix A.

In addition, we measured Pearson correlations between the generated and original sFNC matrices to quantify the correspondence (Figure 6). For FBIRN, all 225 matrices (100.0%) used for training have

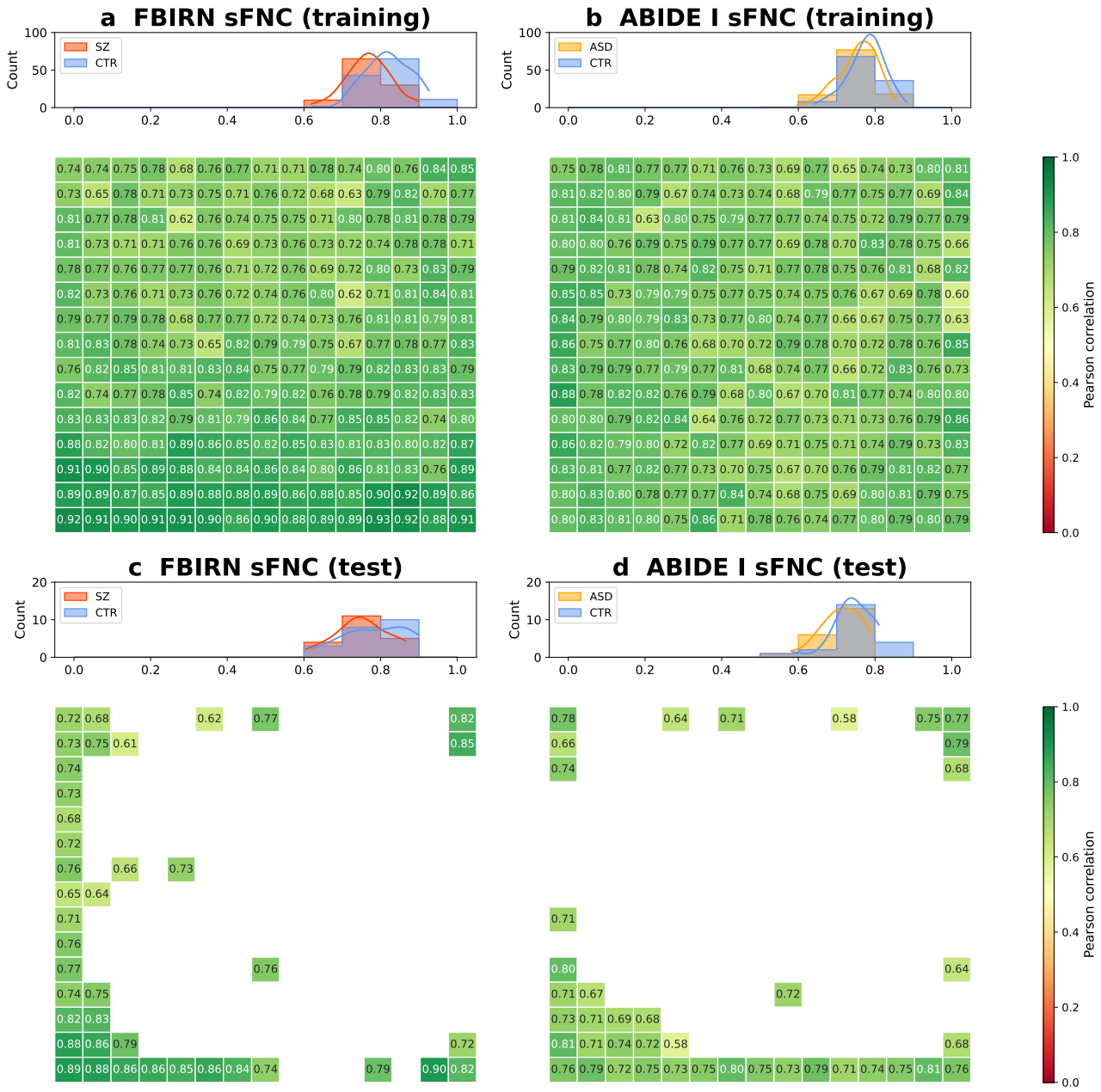


Figure 6: **Pearson correlations between generated and original sFNC matrices.** (a) FBIRN training set. (b) ABIDE I training set. (c) FBIRN test set. (d) ABIDE I test set. Each panel shows Pearson correlations between individual generated and original sFNC matrices on the 2D grid and the corresponding group-specific histogram distributions. High correlations across all panels suggest that the generated sFNC matrices well aligned with the original sFNC matrices ordered by the JV algorithm, especially for controls who show less variability in functional connectivity.

correlations greater than 0.6 and 106 out of 225 matrices (47.1%) have correlations greater than 0.8, with the median correlations of 0.819 and 0.768 for controls and SZ patients, respectively (Figure 6a). For ABIDE I, 224 out of 225 matrices (99.6%) have correlations greater than 0.6 and 54 out of 225 matrices (24.0%) have correlations greater than 0.8, with the median correlations of 0.784 and 0.762 for the control and ASD patient groups, respectively (Figure 6b). We also observe high correlations from the unseen test sets (Figure 6c,d). In the FBIRN test set, the median correlations for the control group and the patient group are 0.792 and 0.747, respectively, while in the ABIDE I test set, they are 0.740 and 0.711, respectively. Mean squared errors (MSE) between the generated and original sFNC matrices also demonstrate their similarity, with the median MSEs of 0.031, 0.034, 0.029, and 0.034 for the FBIRN training set, FBIRN test set, ABIDE I training set, and ABIDE I test set, respectively (Appendix L). These comparisons indicate that the generated sFNC matrices well aligned with the original ones, especially for controls, who show more homogeneity in their functional connectivity. Thus, the VAE latent space can be subsequently used to generate continuous samples by interpolating along pathological trajectories.

3.2 dFNC interpolation reveals group-specific dynamic states

To further evaluate how these group-specific patterns change over time, we performed a dynamic state analysis using the k-means clustering algorithm on the dFNC data (Section 2.4.2). We searched k from 2 to 9, and considered that five clusters were optimal for both FBIRN and ABIDE I datasets based upon the elbow criterion on the L_2 distance (Appendix M). The group-specific dFNC states and group differences for FBIRN and ABIDE I datasets are presented in Figures 7 and 8, respectively.

Similar to the sFNC results, we demonstrate high correlations between the generated dFNC states and the original ones. Specifically, for the FBIRN training set, we first applied k-means clustering to the dFNC latent features from all subjects (including SZ patients and controls) to derive five shared states. We then computed the Pearson correlations between the generated and original dFNC states (1 to 5) separately for each group. The resulting correlations are $0.990, 0.990, 0.882, 0.983, 0.967$ for the SZ patient group, and $0.958, 0.982, 0.872, 0.994, 0.974$ for the control group, respectively (Figure 7a,b). For the FBIRN test set, the correlations from states 1 to 5 are $0.902, 0.968, 0.891, 0.935, 0.908$ for the patient group and $0.820, 0.955, 0.882, 0.967, 0.955$ for the control group, respectively (Figure 7c,d). In state 1, SZ patients, compared with controls, exhibit stronger negative correlations between the subcortical (SC) and sensorimotor (SM) domains and between the SM and visual (VI) domains. Conversely, they show increased positive correlations between the SC and VI domains and between the SM and cerebellar (CB) domains. In state 5, SZ patients show stronger negative correlations between the SM and VI domains and between the SC and CB domains, and stronger positive correlations between the SC domain and

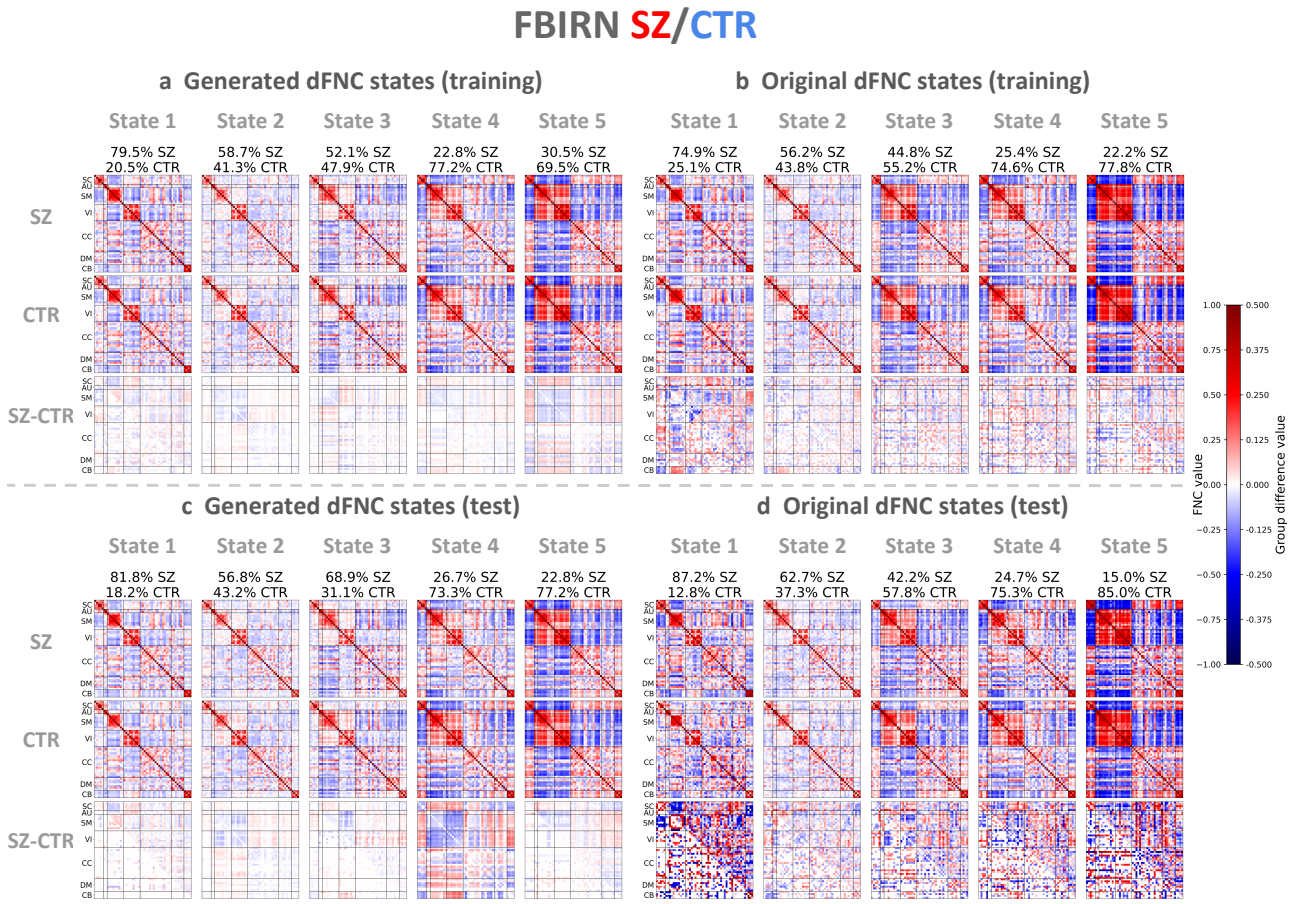


Figure 7: **Generated and original FBIRN dFNC states.** (a) Five generated dFNC states of SZ patients, controls and group differences in the training set. (b) Five original dFNC states in the training set. (c) Five generated dFNC states in the test set. (d) Five original dFNC states in the test set. The generated and original dFNC states represent the element-wise medians of dFNC patterns, obtained by applying k-means clustering to the VAE latent features and the dFNC matrices, respectively. The original dFNC states were sorted by the increasing percentage of control windows relative to total windows per state, and the generated states were ordered based on their similarity to the original states. Group differences between patient and control states are shown in the upper triangle, and significant differences are shown in the lower triangle ($p < 0.05$, two-sample t-test, Bonferroni correction), with all values multiplied by 2 for better visualization. The generated dFNC states revealed state-based group differences, similar to the original ones. Moreover, the VAE effectively captured group-specific representations that are generalizable to the unseen test set.

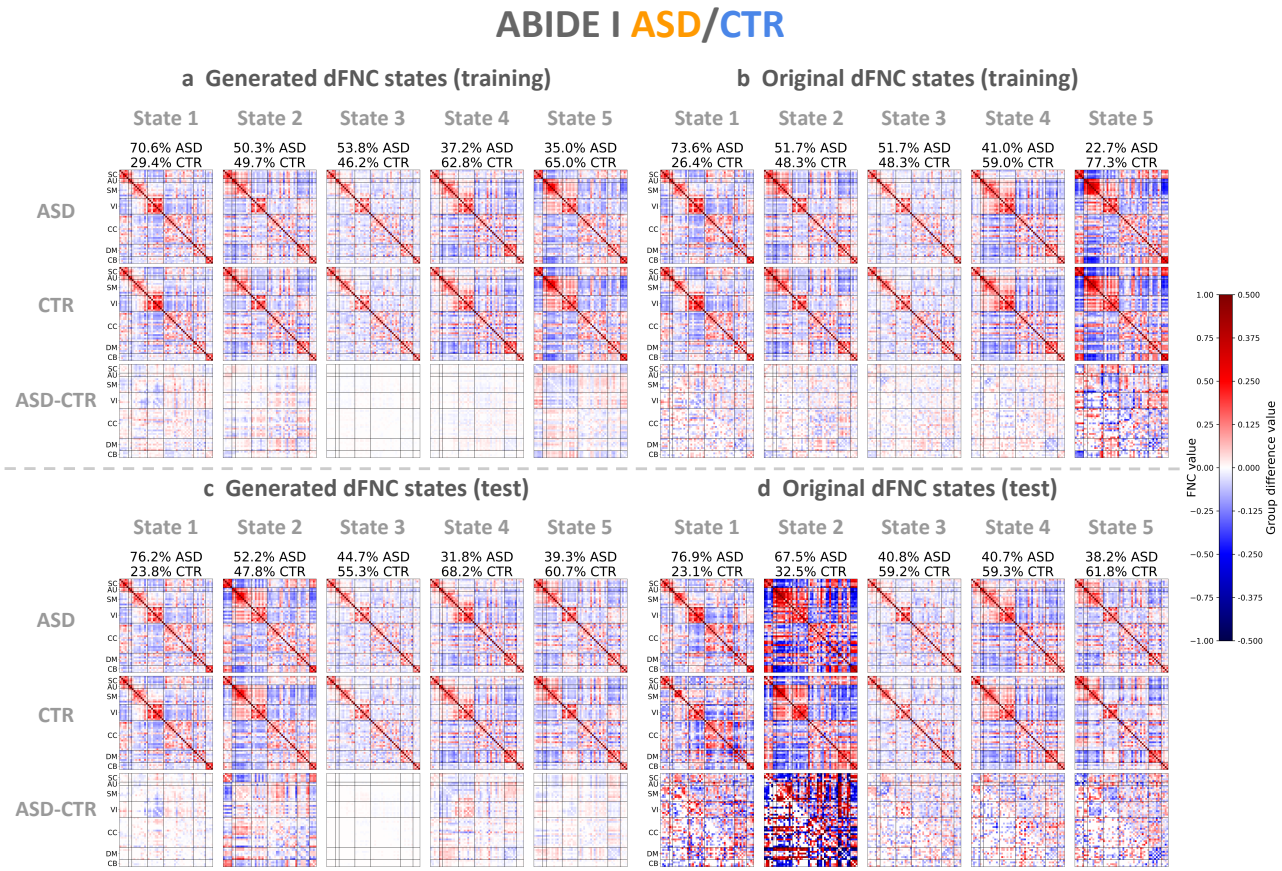


Figure 8: **Generated and original ABIDE I dFNC states.** (a) Five generated dFNC states of ASD patients, controls and group differences in the training set. (b) Five original dFNC states in the training set. (c) Five generated dFNC states in the test set. (d) Five original dFNC states in the test set. The generated and original dFNC states represent the element-wise medians of dFNC patterns, obtained by applying k-means clustering to the VAE latent features and the dFNC matrices, respectively. The original dFNC states were sorted by the increasing percentage of control windows relative to total windows per state, and the generated states were ordered based on their similarity to the original states. Group differences between patient and control states are shown in the upper triangle, and significant differences are shown in the lower triangle ($p < 0.05$, two-sample t-test, Bonferroni correction), with all values multiplied by 2 for better visualization.

both the SM and VI domains, and between the CB domain and these two sensory areas. When evaluated on the unseen test set, the original group difference patterns appear noisy due to the small sample size (Figure 7d). Nonetheless, the generated group difference patterns (Figure 7c) closely resemble those in the training set, particularly between test state 2 and training state 3 (Pearson correlation coefficient $r = 0.811$), and test state 4 and training state 4 ($r = 0.793$). This finding suggests that the VAE can effectively capture *generalizable* group-specific representations.

For the ABIDE I training set, the correlations between the generated and original dFNC states 1 to 5 are $0.985, 0.994, 0.995, 0.982, 0.924$ for the ASD patient group, and $0.971, 0.990, 0.993, 0.984, 0.947$ for the control group, respectively (Figure 8a,b). For the ABIDE I test set, the correlations from states 1 to 5 are $0.957, 0.820, 0.974, 0.947, 0.961$ for the patient group, and $0.885, 0.867, 0.980, 0.969, 0.960$ for the control group, respectively (Figure 8c,d). In state 1 which is mainly occupied by ASD patients, we observe significant group differences, including negative correlations between the SM and VI domains and between the SC and SM domains, as well as positive correlations between the VI domain and the SC, CB and DM domains. These patterns partially overlap with those in the FBIRN state 1, indicating that ASD and SZ may share similar neurobiological deficits.

Subsequently, we evaluated two group-specific state-based dynamic measures – dwell time and transition probability matrix – to further quantitatively compare the generated and original dFNC states (Figure 9). The group-specific dwell time measures the average amount of time that a group lingers in a state. The dwell time statistics of the generated states are largely consistent with those of the original states in both datasets. For the FBIRN training set (Figure 9a,b), states 1, 2, 4 and 5 show significant group differences in both the generated and original states ($p < 0.01$, two-sample t-test, FDR correction). In particular, SZ patients stay in weakly connected states 1 and 2 significantly longer, while controls stay in densely connected states 4 and 5 significantly longer. For the FBIRN test set (Figure 9c,d), we note that the generated results show significant group differences in state 5 ($p < 0.05$, two-sample t-test, FDR correction), whereas the original results do not show any significant differences across five states.

For the ABIDE I training set (Figure 9e,f), the dwell time for the ASD group is significantly longer than that for the control group in both the generated and original state 1 ($p < 0.0001$, two-sample t-test, FDR correction). In contrast, the dwell time for the control group is significantly longer than that for the ASD group in both the generated and original states 4 and 5 ($p < 0.05$, two-sample t-test, FDR correction). No significant group differences are observed in states 2 and 3. For the ABIDE I test set (Figure 9g,h), the dwell time statistics in the generated results are also consistent with the original ones. Only the patient-dominant state 1 shows significant group differences ($p < 0.05$, two-sample t-test, FDR correction).

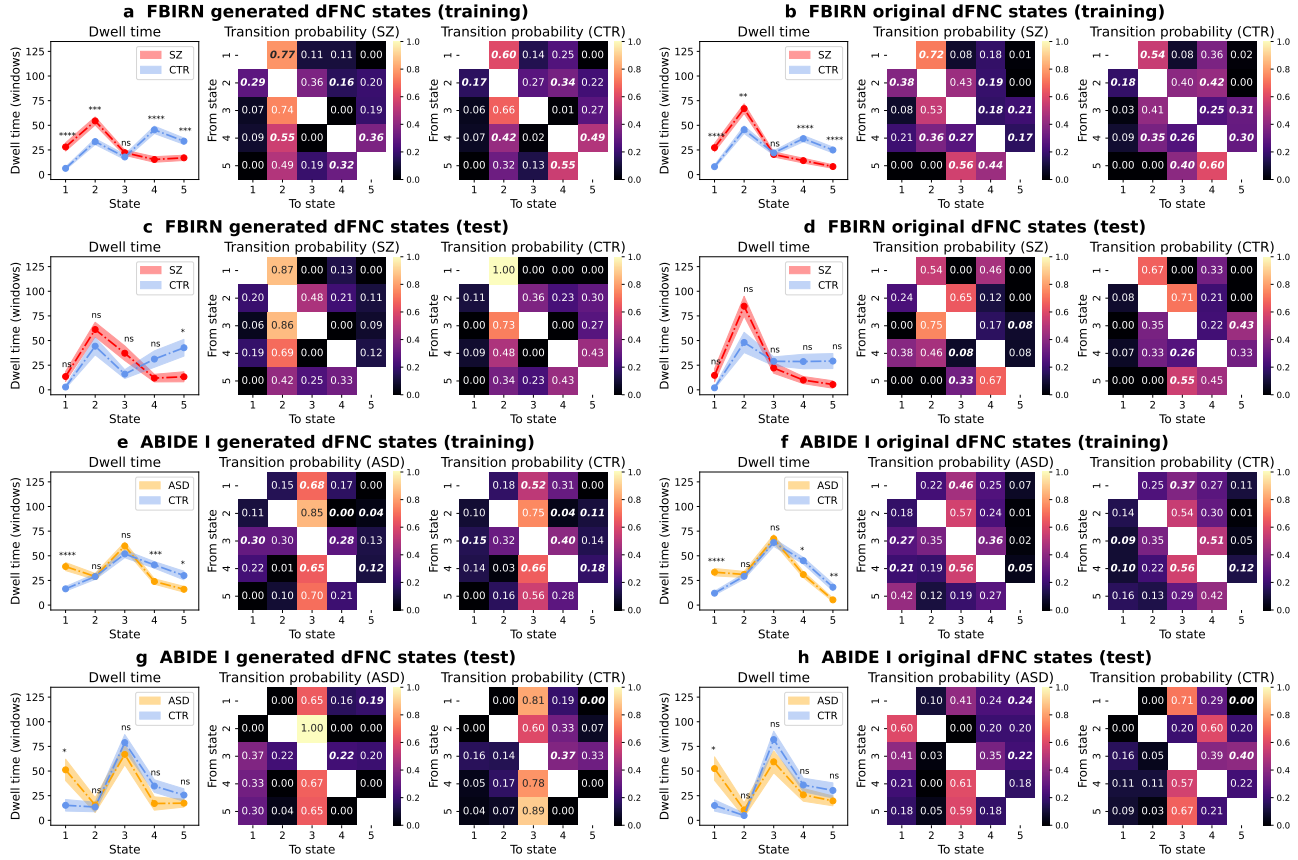


Figure 9: **Group-specific dwell time and transition probability matrices of generated and original dFNC states.** Parts a,b,c,d show the results from the FBIRN training and test sets, while parts e,f,g,h show the results from the ABIDE I training and test sets. Parts a,c,e,g show the metrics derived from the generated data and parts b,d,f,h present the results from the original data. In each part, the left panel shows the mean $\pm$ standard error of dwell time measured by windows for different groups (red: SZ patients; orange: ASD patients; blue: controls). Two-sample t-tests with the false discovery rate (FDR) correction for five states were used to assess group differences (state labels are consistent with Figures 7 and 8). The number of asterisks (1, 2, 3, 4) indicates the level of significance ($p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 0.0001$, FDR correction). The abbreviation "ns" stands for "not significant". The middle and right panels show the transition probabilities between every two states for patients and controls, respectively. The transition probability is calculated by counting the number of transition times from one state to another state divided by the number of subjects from each group, and then normalizing the sum of each row to 1. Significant transition count differences between patients and controls are highlighted in bold italics ($p < 0.05$, two-sample t-test). The generated states exhibited group-specific properties, such as dwell time and transition probabilities, that closely matched the original states. These metrics indicate that VAEs effectively captured state-based dynamic properties.

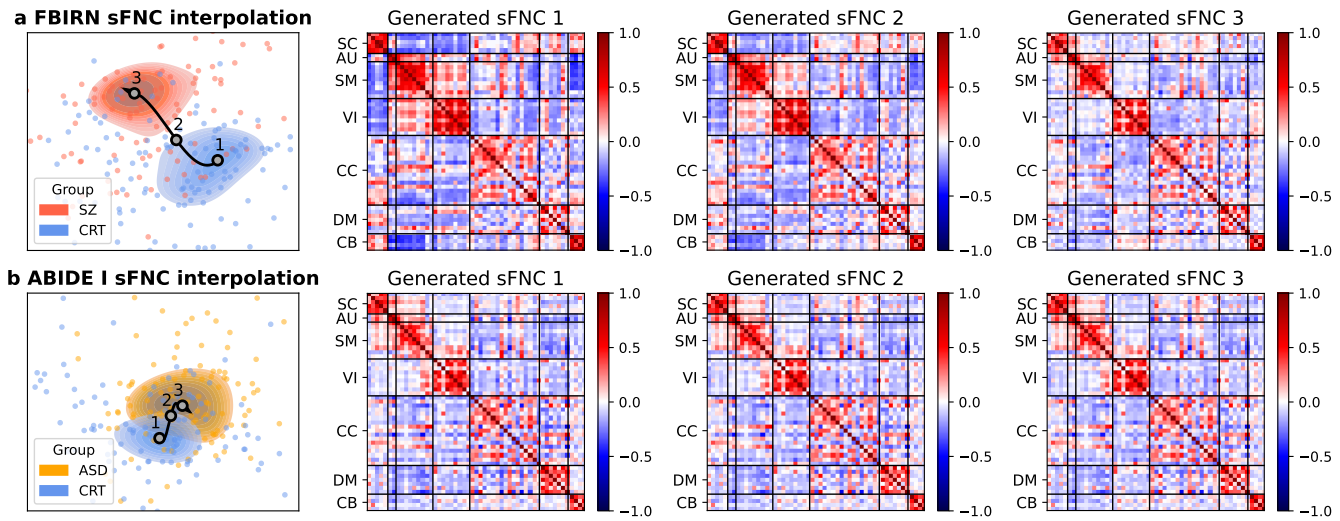


Figure 10: **sFNC interpolation between diagnostic groups in the VAE latent space.** (a) FBIRN training data. (b) ABIDE I training data. In each row, the scatter plot shows latent features colored by the diagnostic label (red: SZ patients; orange: ASD patients; blue: controls), overlaid with contour plots of two clusters connected by a trajectory. For each group, we generated three representative sFNC patterns by uniformly sampling from the derived trajectory.

The group-specific transition probability matrix shows the probability that a group switches from one state to another. For the FBIRN training set (Figure 9a,b), we note that SZ patients are more likely to transition between loosely connected states 1 and 2 (state 1 to state 2: $P_{\text{generated}} = 0.77$, $P_{\text{original}} = 0.72$; state 2 to state 1: $P_{\text{generated}} = 0.29$, $P_{\text{original}} = 0.38$). In contrast, controls show significantly lower transition probabilities when moving between state 1 and state 2 (state 1 to state 2: $P_{\text{generated}} = 0.60$, $P_{\text{original}} = 0.54$; state 2 to state 1: $P_{\text{generated}} = 0.17$, $P_{\text{original}} = 0.18$). By comparison, controls are more likely to switch between densely connected states 4 and 5 (state 4 to state 5: $P_{\text{generated}} = 0.49$, $P_{\text{original}} = 0.30$; state 5 to state 4: $P_{\text{generated}} = 0.55$, $P_{\text{original}} = 0.60$), whereas patients with SZ have relatively lower probabilities of switching between states 4 and 5 (state 4 to state 5: $P_{\text{generated}} = 0.36$, $P_{\text{original}} = 0.17$; state 5 to state 4: $P_{\text{generated}} = 0.32$, $P_{\text{original}} = 0.44$).

Interestingly, for the ABIDE I training and test data (Figure 9e-h), both the ASD patient and control groups tend to switch from other states to state 3, suggesting that it may act as a hub for dynamic state transitions. In general, controls exhibit significantly higher transition probabilities from other states to state 4, while autistic patients are more likely to transition to state 1.

Taken together, the generated states show highly similar group-specific properties, such as dwell time and transition probabilities, as the original ones. These results indicate that VAEs effectively learn state-based dynamic properties without explicit incorporation of time information during the training stage.

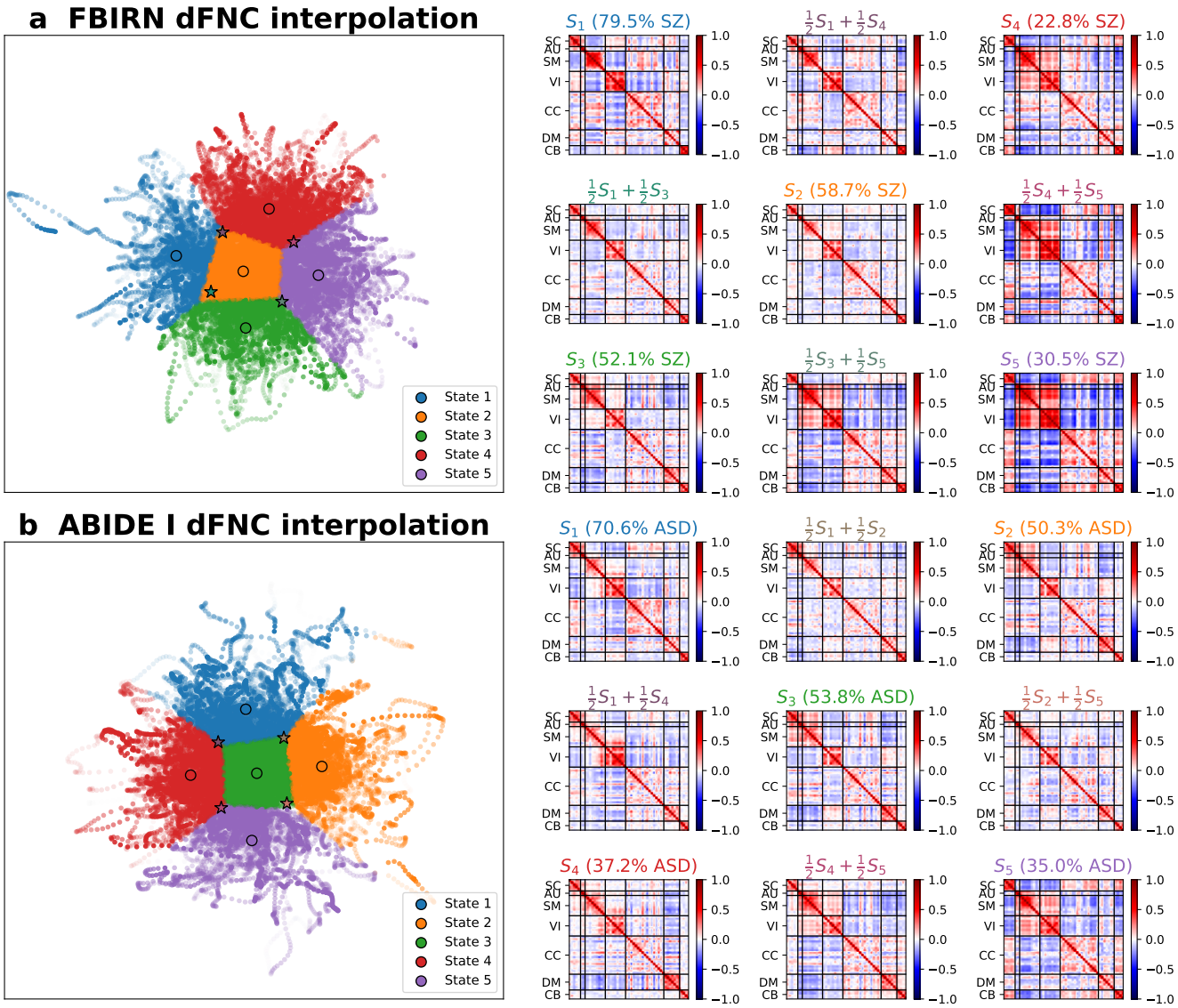


Figure 11: **dFNC interpolation between two dynamic states in the VAE latent space.** (a) FBIRN training data. (b) ABIDE I training data. Left, we show the dFNC latent features color-coded by k-means states. We performed interpolation at the midpoint between two state centroids to examine the transition patterns. Stars ($\star$) represent the relative positions in the latent space where the samples were generated. Right, we show the generated samples from each of five state centroids and each of four midpoints between two state centroids.

3.3 A low-dimensional manifold to interpret and interpolate FNC

In our proposed interpolation framework, the VAE latent space serves as a low-dimensional manifold to interpret and interpolate sFNC or dFNC matrices, providing novel insights into how FNC patterns change along a trajectory of interest. Here, we show two potentially useful applications of the interpolation framework: (1) sFNC interpolation between patient and control groups to estimate a psychosis continuum (Figure 10) and (2) dFNC interpolation between two states to investigate state transition (Figure 11).

To visualize how sFNC continuously changes between the control group and the patient group, we interpolated along a trajectory from the control center to the patient center (Figure 10). When interpolating along the trajectory from the control center to the SZ patient center, we observe significant connectivity changes across multiple domains ($p < 0.0001$, Wilcoxon signed-rank test, Bonferroni correction). Specifically, anti-correlations become significantly weaker in the SC–SM, SC–VI, and SM–CB domains, whereas positive correlations become significantly weaker within the SM, VI, and CC domains. In addition, positive correlations gradually shift to negative correlations in the SC–CB and SM–VI domains, while negative correlations shift to positive correlations in the VI–CB domain (Figure 10a). When interpolating along the trajectory from the control center to the ASD patient center, we observe partially overlapping but distinct connectivity patterns. Similar to SZ, negative correlations are significantly stronger in the SC–CB domain but weaker in the VI-CB domain. In contrast, ASD shows unique alterations, including significantly weaker negative correlations in the SM–DM and VI–DM domains, stronger negative correlations in the SC-VI and VI-CC domains, decreased positive correlations within the DM domain, and increased positive correlations within the CC domain (Figure 10b).

To evaluate how a dynamic state switches to another, we generated an FNC matrix by sampling from the distribution at the midpoint between two state centroids in the latent space (Figure 11). When an SZ-dominant state (e.g., state 1) changes to a control-dominant state (e.g., state 5), we observe weaker negative correlations between the SM and VI domains and between the VI and CC domains, and stronger negative correlations between the CB domain and the AU and SM domains (Figure 11a). When an ASD-dominant state (e.g., state 1) switches to a state occupied by more controls (e.g., state 2), we notice a connectivity shift from positive correlations between the SM and VI domains to positive correlations among the SC, AU, and SM domains (Figure 11b).

4 Discussion

This work presents a functional network connectivity interpolation framework using a VAE to investigate neuropsychiatric continuum and heterogeneity from static and dynamic FNC data. Our framework provides additional advantages that complement existing diagnostic and analytical methods. To identify neurobiological markers of SZ and ASD, we leveraged functional network connectivity derived from fMRI – an objective measure of brain abnormalities under pathological conditions. To address the limitations of diagnostic labels and supervised prediction, we implemented an *unsupervised generative* model to learn hidden relationships in a fully data-driven manner and generate continuous FNC patterns along the latent disorder spectrum. The correlations between the generated and original sFNC matrices demonstrate that the VAE captured representative and generalizable latent distributions in both datasets, outperforming a linear baseline (PPCA) and a semi-supervised alternative (iVAE) (Figures 2, 6). Notably, the unsupervised VAE slightly outperformed the semi-supervised iVAE that conditioned on the diagnostic label. One reason could be that the iVAE is trained to reduce the total correlation of the latent variable given the auxiliary variable except for optimizing reconstruction (see Khemakhem et al., 2020 Appendix F), and thus its reconstruction performance might not be optimal. In addition, we chose the diagnostic label as the auxiliary variable, yet the most appropriate auxiliary information in the neuroimaging data remains unknown. Regardless, we consider VAEs to be the most suitable for our interpolation task because of their ability to learn nonlinear latent relationships in an unsupervised manner, as supported by empirical evidence.

Unlike prior studies that focus on group-level comparisons, our framework provides a generative approach to model a functional connectivity continuum between groups. Specifically, the interpolated continua from controls to patients in both disorders show two common functional network connectivity gradients: (1) weaker positive correlations within sensory networks (auditory, sensorimotor, visual) and between the subcortical and cerebellar domains; (2) weaker negative correlations between the subcortical domain and the sensory domains, as well as between the cerebellar domain and these sensory regions (Figures 3, 4). Consistent with these findings, previous studies using NeuroMark sFNC in SZ and ASD have reported overlapping group-level alterations in both disorders, particularly between the subcortical domain and the cerebellar, auditory, and sensorimotor domains (Du et al., 2020; Jensen et al., 2024; Yan et al., 2024). We also observe disorder-specific patterns. For example, SZ is characterized by reduced functional connectivity in the visual domain, whereas ASD shows decreased functional connectivity within the default mode network, aligning with prior work (Du et al., 2020).

Another key benefit of our interpolation framework is to visualize and characterize individual differences. Instead of solely relying on group averages, we designed a 2D grid display to organize sFNC matrices,

supporting examination of subject-level variability and inter-subject relationships (Figures 3, 4). Apart from sFNC matrices, we projected multiple subject measures, including clinical and cognitive scores, on the same 2D grid (Figure 5). Notably, SZ patients grouped in the lower/left/lower triangular half of the 2D grid consistently show higher cognitive scores, indicating that these patients have better cognitive functions (Table 2).

Moreover, we identified subgroups among SZ patients by applying the k-means clustering algorithm on the learned latent features and then examined subgroup differences in subject measures and functional network connectivity patterns (Appendix J). Changes related to working memory were observed across multiple network connections, including sensorimotor–default mode, sensorimotor–cerebellar, and cognitive control–cerebellar networks. Previous studies have similarly linked working memory deficits in SZ to weaker connectivity between the fronto-parietal network and cerebellar network (Repovs et al., 2011), as well as stronger within-network connectivity in the fronto-parietal and default mode networks (Unschuld et al., 2014). The default mode network was also associated with speed of processing, consistent with previous reports (Mwansisya et al., 2013; Wang et al., 2014). In addition, attention performance was associated with connectivity changes between the default mode and cognitive control networks and each of the sensorimotor, cerebellar, and auditory domains. This finding is in line with prior work proposing the insula as a key hub that modulates activity between the default mode network and task-positive networks (Fox et al., 2005; Z. He et al., 2013; Moran et al., 2013).

We further examined how sFNC patterns vary across cognitive performance levels by interpolating along a trajectory from the higher-score group to the lower-score group (Appendix K). Existing literature suggests that cognitive deficits in SZ are associated with reduced resting-state functional connectivity between cortical and subcortical regions (Sheffield & Barch, 2016). Consistent with this evidence, we found that, compared with the higher-score group, the lower-score group exhibited significantly attenuated negative correlations in the subcortical–sensorimotor, subcortical–visual, and sensorimotor–cerebellar domains, as well as reduced positive correlations within the sensorimotor, visual, and cognitive control domains, supporting the role of disrupted cortical–subcortical integration in cognitive impairment.

In addition to static brain connectivity, we evaluated whether the VAE can capture dynamic properties from windowed dFNC data. We observe highly similar k-means states between the generated and original results (Figures 7, 8). These generated states exhibit consistent group-specific properties, measured by dwell time and transition probability, as the original states (Figure 9). In particular, the dynamic state analysis in both the generated and original states suggests that SZ and ASD patients tend to stay longer in sparsely connected states while controls tend to stay longer in densely connected states, consistent with prior literature (Damaraju et al., 2014; Fu et al., 2019, 2021).

A VAE is designed to learn latent distributions and capture hidden relationships from observed data in an unsupervised manner. By leveraging the VAE latent space, we performed interpolation along a pathological trajectory to visualize continuous sFNC alterations from the control center to the patient center (Figure 10). Similarly, we generated samples between two dFNC state centroids to examine how one state gradually switches to another (Figure 11). Overall, we demonstrate the ability of the unsupervised generative model to capture representative group-specific functional network connectivity profiles and generate continuous pattern alterations between different subjects, groups, dynamic states, and time points.

Despite the promising potential of our interpolation framework, we acknowledge limitations in this work. First, we implemented a vanilla VAE with a multivariate Gaussian prior assumption, but it might not be fully capable of capturing the heterogeneity or potential hierarchical structures in the SZ and ASD datasets. Second, in order to generate continuous sFNC data, we sampled from linearly spaced coordinates in the latent space, rather than the posterior distributions, so the generated sFNC matrices and the original sFNC matrices ordered by the JV algorithm were not one-to-one matched due to uniform sampling. Lastly, we made an assumption that the differences in FNC between groups, which reflect the underlying neurobiological changes in psychiatric disorders, are continuous and can be interpolated. However, there has been debate about whether psychosis can be viewed as continua or categories (David, 2010; DeRosse & Karlsgodt, 2015; Lawrie et al., 2010). More recently, there is growing evidence that suggests the shift from the concept of schizophrenia to a broader psychotic spectrum (Guloksuz & van Os, 2018) that captures a continuum of psychosis proneness from normal to abnormal (Van Os et al., 2009). Hence, it is worth further quantifying the associations between the interpolated FNC patterns and the subject measures, including clinical symptoms and cognitive functions.

In the future, we aim to extend the framework to further characterize psychotic syndromes using hierarchical clustering or mixture model priors in VAEs, and to automatically capture dynamic states using dynamical VAEs (Girin et al., 2021). We will also more systematically quantify the relationships between functional network connectivity patterns and cognitive functions across patient subgroups. To improve generalizability and establish more reliable continuum prediction, we will analyze other large-scale longitudinal datasets with multiple mental disorders, such as the Adolescent Brain and Cognitive Development (ABCD) dataset (Barch et al., 2018; Casey et al., 2018; Garavan et al., 2018).

5 Conclusions

We present a functional network connectivity interpolation framework using an unsupervised generative model to interpolate sFNC and dFNC data from controls and patients with SZ or ASD. We demonstrated a high degree of correspondence between the generated and original individual sFNC patterns, as well as the group-specific dFNC states. We showed that our interpolation framework captured individual variability within a group, sFNC pattern alterations between groups, and dFNC dynamic states between groups. In addition, we provided examples of interpolating along trajectories of interest in the latent manifold to visualize the corresponding continuous FNC alterations in the original space. Our approach offers new insights into the neuropsychiatric continuum and heterogeneity, marking a promising step toward personalized characterization of mental disorders.

Ethics statement

The authors have no conflict of interest to declare. Large language models such as Claude and ChatGPT were used to correct grammar mistakes at the sentence level.

Data and code availability

The FBIRN dataset can be accessed at https://www.nitrc.org/projects/fbirn/. The ABIDE I dataset can be accessed at https://fcon_1000.projects.nitrc.org/indi/abide/. The NeuroMark network templates are available at http://trendscenter.org/software. The analysis and visualization code is publicly available at https://github.com/XinhuiLi/interpolation.git.

Author contributions

Xinhui Li: Conceptualization; formal analysis; investigation; methodology; software; visualization; writing - original draft; writing - review and editing. **Eloy Geenjaar:** Investigation; methodology; writing - review and editing. **Zening Fu:** Data curation; investigation; writing - review and editing. **Godfrey D. Pearlson:** Investigation; writing - review and editing. **Vince D. Calhoun:** Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; writing - review and editing.

Funding

This work was supported by the National Institutes of Health (NIH) grants (R01MH118695 and R01EB006841) and the National Science Foundation (NSF) grant (NSF2112455). Additionally, X.L. and E.G. were supported by the Georgia Tech/Emory NIH/NIBIB Training Program in Computational Neural-engineering (T32EB025816).

Declaration of competing interests

The authors have no competing interest to declare.

Acknowledgments

We thank the FBIRN team for collecting and sharing the dataset, including Adrian Preda, Aysenil Belger, Bryon A. Mueller, Daniel H. Mathalon, Daniel S. O'Leary, Jessica A. Turner, Juan R. Bustillo, Judith M. Ford, Kelvin O. Lim, Steven G. Potkin, and Theo G.M. van Erp. We also thank the ABIDE team for collecting, organizing and sharing the dataset. We acknowledge the NeuroImaging Tools and Resources Collaboratory (NITRC) team for providing the data sharing platform.

References

Abrahams, B. S., & Geschwind, D. H. (2008). Advances in autism genetics: On the threshold of a new neurobiology. *Nature reviews genetics*, *9*(5), 341–355.

Ahmed, A. O., Strauss, G. P., Buchanan, R. W., Kirkpatrick, B., & Carpenter, W. T. (2018). Schizophrenia heterogeneity revisited: Clinical, cognitive, and psychosocial correlates of statistically-derived negative symptoms subgroups. *Journal of psychiatric research*, *97*, 8–15.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, *24*(3), 663–676.

Alnæs, D., Kaufmann, T., Van Der Meer, D., Córdova-Palomera, A., Rokicki, J., Moberget, T., Bettella, F., Agartz, I., Barch, D. M., Bertolino, A., et al. (2019). Brain heterogeneity in schizophrenia and its association with polygenic risk. *JAMA psychiatry*, *76*(7), 739–748.

Amaral, D. G., Schumann, C. M., & Nordahl, C. W. (2008). Neuroanatomy of autism. *Trends in neurosciences*, *31*(3), 137–145.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: Dsm-5* (Vol. 5). American psychiatric association Washington, DC.

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., Kiebel, S., Kilner, J., Litvak, V., Moran, R., et al. (2014). Spm12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, *2464*(4).

Barch, D. M., Albaugh, M. D., Avenevoli, S., Chang, L., Clark, D. B., Glantz, M. D., Hudziak, J. J., Jernigan, T. L., Tapert, S. F., Yurgelun-Todd, D., et al. (2018). Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description. *Developmental cognitive neuroscience*, *32*, 55–66.

Barnett, S. A. (2018). Convergence problems with generative adversarial networks (gans). *arXiv preprint arXiv:1806.11382*.

Barneveld, P. S., Pieterse, J., de Sonneville, L., van Rijn, S., Lahuis, B., van Engeland, H., & Swaab, H. (2011). Overlap of autistic and schizotypal traits in adolescents with autism spectrum disorders. *Schizophrenia research*, *126*(1-3), 231–236.

Bill, B. R., & Geschwind, D. H. (2009). Genetic advances in autism: Heterogeneity and convergence on shared pathways. *Current opinion in genetics & development*, *19*(3), 271–278.

Bishop, C. M., & Bishop, H. (2023). *Deep learning: Foundations and concepts*. Springer Nature.

Bora, E. (2016). Differences in cognitive impairment between schizophrenia and bipolar disorder: Considering the role of heterogeneity. *Psychiatry and clinical neurosciences*, *70*(10), 424–433.

Brugger, S. P., & Howes, O. D. (2017). Heterogeneity and homogeneity of regional brain structure in schizophrenia: A meta-analysis. *JAMA psychiatry*, *74*(11), 1104–1111.

Carpenter Jr, W. T., & Kirkpatrick, B. (1988). The heterogeneity of the long-term course of schizophrenia. *Schizophrenia bulletin*, *14*(4), 645–652.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. (2018). The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, *32*, 43–54.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., & Li, W. (2016). Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*.

Chen, C. P., Keown, C. L., Jahedi, A., Nair, A., Pflieger, M. E., Bailey, B. A., & Müller, R.-A. (2015). Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism. *NeuroImage: Clinical*, *8*, 238–245.

Chen, H., Uddin, L. Q., Duan, X., Zheng, J., Long, Z., Zhang, Y., Guo, X., Zhang, Y., Zhao, J., & Chen, H. (2017). Shared atypical default mode and salience network functional connectivity between autism and schizophrenia. *Autism Research*, *10*(11), 1776–1786.

Collaborators, G. 2. M. D., et al. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, *9*(2), 137–150.

Craddock, N., & Mynors-Wallis, L. (2014). Psychiatric diagnosis: Impersonal, imperfect and important. *The British Journal of Psychiatry*, *204*(2), 93–95.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of rdoc. *BMC medicine*, *11*, 1–8.

Damaraju, E., Allen, E., Belger, A., Ford, J., McEwen, S., Mathalon, D., Mueller, B., Pearlson, G., Potkin, S., Preda, A., Turner, J., Vaidya, J., van Erp, T., & Calhoun, V. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical*, *5*, 298–308. https://doi.org/https://doi.org/10.1016/j.nicl.2014.07.003

David, A. (2010). Why we need more debate on whether psychotic symptoms lie on a continuum with normality. *Psychological medicine*, *40*(12), 1935–1942.

DeRosse, P., & Karlsgodt, K. H. (2015). Examining the psychosis continuum. *Current behavioral neuroscience reports*, *2*, 80–89.

Di Biase, M. A., Geaghan, M. P., Reay, W. R., Seidlitz, J., Weickert, C. S., Pébay, A., Green, M. J., Quidé, Y., Atkins, J. R., Coleman, M. J., et al. (2022). Cell type-specific manifestations of cortical thickness heterogeneity in schizophrenia. *Molecular psychiatry*, *27*(4), 2052–2060.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, *19*(6), 659–667.

Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., Salman, M., Abrol, A., Rahaman, M. A., Chen, J., Hong, L. E., Kochunov, P., Osuch, E. A., & Calhoun, V. D. (2020). Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders. *NeuroImage: Clinical*, *28*, 102375. https://doi.org/https://doi.org/10.1016/j.nicl.2020.102375

Du, Y., Fu, Z., Xing, Y., Lin, D., Pearlson, G., Kochunov, P., Hong, L. E., Qi, S., Salman, M., Abrol, A., et al. (2021). Evidence of shared and distinct functional and structural brain signatures in schizophrenia and autism spectrum disorder. *Communications biology*, *4*(1), 1073.

Feczko, E., Balba, N., Miranda-Dominguez, O., Cordova, M., Karalunas, S., Irwin, L., Demeter, D., Hill, A., Langhorst, B., Painter, J. G., et al. (2018). Subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm. *Neuroimage*, *172*, 674–688.

Fountain, C., Winter, A. S., & Bearman, P. S. (2012). Six developmental trajectories characterize children with autism. *Pediatrics*, *129*(5), e1112–e1120.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Fu, Z., Sui, J., Turner, J. A., Du, Y., Assaf, M., Pearlson, G. D., & Calhoun, V. D. (2021). Dynamic functional network reconfiguration underlying the pathophysiology of schizophrenia and autism spectrum disorder. *Human brain mapping*, *42*(1), 80–94.

Fu, Z., Tu, Y., Di, X., Du, Y., Sui, J., Biswal, B. B., Zhang, Z., de Lacy, N., & Calhoun, V. D. (2019). Transient increased thalamic-sensory connectivity and decreased whole-brain dynamism in autism. *Neuroimage*, *190*, 191–204.

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R., Heeringa, S., Jernigan, T., Potter, A., Thompson, W., & Zahs, D. (2018). Recruiting the abcd sample: Design considerations and procedures. *Developmental cognitive neuroscience*, *32*, 16–22.

Geenjaar, E., Lewis, N., Fu, Z., Venkatdas, R., Plis, S., & Calhoun, V. (2021). Fusing multimodal neuroimaging data with a variational autoencoder. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3630–3633.

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P., et al. (2013). Investigating phenotypic heterogeneity in children with autism spectrum disorder: A factor mixture modeling approach. *Journal of Child Psychology and Psychiatry*, *54*(2), 206–215.

Geschwind, D. H., et al. (2015). Gene hunting in autism spectrum disorder: On the path to precision medicine. *The Lancet Neurology*, *14*(11), 1109–1120.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends® in Machine Learning*, *15*(1-2), 1–175. https://doi.org/10.1561/2200000089

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Guloksuz, S., & van Os, J. (2018). The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. *Psychological medicine*, *48*(2), 229–244.

Guo, X., Duan, X., Suckling, J., Chen, H., Liao, W., Cui, Q., & Chen, H. (2019). Partially impaired functional connectivity states between right anterior insula and default mode network in autism spectrum disorder. *Human brain mapping*, *40*(4), 1264–1275.

Hahamy, A., Behrmann, M., & Malach, R. (2015). The idiosyncratic brain: Distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nature neuroscience*, *18*(2), 302–309.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

He, Y., Byrge, L., & Kennedy, D. P. (2020). Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. *Human brain mapping*, *41*(5), 1334–1350.

He, Z., Deng, W., Li, M., Chen, Z., Jiang, L., Wang, Q., Huang, C., Collier, D., Gong, Q., Ma, X., et al. (2013). Aberrant intrinsic brain activity and cognitive deficit in first-episode treatment-naive patients with schizophrenia. *Psychological medicine*, *43*(4), 769–780.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Holmes, A. J., Hollinshead, M. O., O'keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., Fischl, B., Rosen, B. R., Mair, R. W., Roffman, J. L., et al. (2015). Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific data*, *2*(1), 1–16.

Hong, S.-J., Valk, S. L., Di Martino, A., Milham, M. P., & Bernhardt, B. C. (2018). Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cerebral Cortex*, *28*(10), 3578–3588.

Hull, J. V., Dokovna, L. B., Jacokes, Z. J., Torgerson, C. M., Irimia, A., & Van Horn, J. D. (2017). Resting-state functional connectivity in autism spectrum disorders: A review. *Frontiers in psychiatry*, *7*, 205.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural networks*, *13*(4-5), 411–430.

Iraji, A., Deramus, T. P., Lewis, N., Yaesoubi, M., Stephen, J. M., Erhardt, E., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., et al. (2019). The spatial chronnectome reveals a dynamic interplay between functional segregation and integration. *Human brain mapping*, *40*(10), 3058–3077.

Iraji, A., Faghiri, A., Lewis, N., Fu, Z., Rachakonda, S., & Calhoun, V. D. (2021). Tools of the trade: Estimating time-varying connectivity patterns from fmri data. *Social cognitive and affective neuroscience*, *16*(8), 849–874.

Jensen, K. M., Calhoun, V. D., Fu, Z., Yang, K., Faria, A. V., Ishizuka, K., Sawa, A., Andrés-Camazón, P., Coffman, B. A., Seebold, D., et al. (2024). A whole-brain neuromark resting-state fmri analysis of first-episode and early psychosis: Evidence of aberrant cortical-subcortical-cerebellar functional circuitry. *NeuroImage: Clinical*, *41*, 103584.

Johns, L. C., & Van Os, J. (2001). The continuity of psychotic experiences in the general population. *Clinical psychology review*, *21*(8), 1125–1141.

Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis* (pp. 115–128). Springer New York.

Jonker, R., & Volgenant, A. (2005). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, *38*, 325–340.

Joyce, E. M., & Roiser, J. P. (2007). Cognitive heterogeneity in schizophrenia. *Current opinion in psychiatry*, *20*(3), 268.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, *13*(2), 261–276. https://doi.org/https://doi.org/10.1093/schbul/13.2.261

Keator, D. B., van Erp, T., Turner, J. A., Glover, G. H., Mueller, B. A., Liu, T. T., Voyvodic, J. T., Rasmussen, J., Calhoun, V. D., Lee, H. J., Toga, A. W., McEwen, S., Ford, J. M., Mathalon, D. H., Diaz, M., O'Leary, D. S., Jeremy Bockholt, H., Gadde, S., Preda, A., . . . FBIRN. (2016). The function biomedical informatics research network data repository. *NeuroImage*, *124*, 1074–1079.

Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. *International Conference on Artificial Intelligence and Statistics*, 2207–2217.

Kim, S. H., Macari, S., Koller, J., & Chawarska, K. (2016). Examining the phenotypic heterogeneity of early autism spectrum disorder: Subtypes and short-term outcomes. *Journal of Child Psychology and Psychiatry*, *57*(1), 93–102.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes.

Konstantareas, M. M., & Hewitt, T. (2001). Autistic disorder and schizophrenia: Diagnostic overlaps. *Journal of autism and developmental disorders*, *31*, 19–28.

Lawrie, S. M., Hall, J., McIntosh, A. M., Owens, D. G., & Johnstone, E. C. (2010). The 'continuum of psychosis': Scientifically unproven and clinically impractical. *The British Journal of Psychiatry*, *197*(6), 423–425.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lopez-Paz, D., Hennig, P., & Schölkopf, B. (2013). The randomized dependence coefficient. *Advances in neural information processing systems*, *26*.

Lord, C., Bishop, S., & Anderson, D. (2015). Developmental trajectories as autism phenotypes. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *169*(2), 198–208.

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *The lancet*, *392*(10146), 508–520.

Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, *32*.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, *30*, 3.

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nature Reviews Neuroscience*, *13*(2), 107–120.

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience bulletin*, *33*, 183–193.

McCutcheon, R. A., Marques, T. R., & Howes, O. D. (2020). Schizophrenia—an overview. *JAMA psychiatry*, *77*(2), 201–210.

Mohr, P. E., Cheng, C. M., Claxton, K., Conley, R. R., Feldman, J. J., Hargreaves, W. A., Lehman, A. F., Lenert, L. A., Mahmoud, R., Marder, S. R., et al. (2004). The heterogeneity of schizophrenia in disease states. *Schizophrenia research*, *71*(1), 83–95.

Morales-Hidalgo, P., Ferrando, P. J., & Canals, J. (2018). Assessing the heterogeneity of autism spectrum symptoms in a school population. *Autism Research*, *11*(7), 979–988.

Moran, L. V., Tagamets, M. A., Sampath, H., O'Donnell, A., Stein, E. A., Kochunov, P., & Hong, L. E. (2013). Disruption of anterior insula modulation of large-scale brain networks in schizophrenia. *Biological psychiatry*, *74*(6), 467–474.

Mwansisya, T. E., Wang, Z., Tao, H., Zhang, H., Hu, A., Guo, S., & Liu, Z. (2013). The diminished interhemispheric connectivity correlates with negative symptoms and cognitive impairment in first-episode schizophrenia. *Schizophrenia research*, *150*(1), 144–150.

O'Connell, K. S., McGregor, N. W., Lochner, C., Emsley, R., & Warnich, L. (2018). The genetic architecture of schizophrenia, bipolar disorder, obsessive-compulsive disorder and autism spectrum disorder. *Molecular and Cellular Neuroscience*, *88*, 300–307.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pulver, A. E., Mulle, J., Nestadt, G., Swartz, K., Blouin, J., Dombroski, B., Liang, K.-Y., Housman, D., Kazazian, H., Antonarakis, S., et al. (2000). Genetic heterogeneity in schizophrenia: Stratification of genome scan data using co-segregating related phenotypes. *Molecular psychiatry*, *5*(6), 650–653.

Rabany, L., Brocke, S., Calhoun, V. D., Pittman, B., Corbera, S., Wexler, B. E., Bell, M. D., Pelphrey, K., Pearlson, G. D., & Assaf, M. (2019). Dynamic functional connectivity in schizophrenia and autism spectrum disorder: Convergence, divergence and classification. *NeuroImage: Clinical*, *24*, 101966.

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., & Calhoun, V. D. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fmri brain connectivity. *NeuroImage*, *134*, 645–657. https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.04.051

Repovs, G., Csernansky, J. G., & Barch, D. M. (2011). Brain network connectivity in individuals with schizophrenia and their siblings. *Biological psychiatry*, *69*(10), 967–973.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29*.

Scott, I., & Nelson, B. (2024). Are psychiatric nosologies limiting the success of clinical prediction models? *JAMA psychiatry*.

Segal, A., Tiego, J., Parkes, L., Holmes, A. J., Marquand, A. F., & Fornito, A. (2024). Embracing variability in the search for biological mechanisms of psychiatric illness. *Trends in Cognitive Sciences*.

Sheffield, J. M., & Barch, D. M. (2016). Cognition and resting-state functional connectivity in schizophrenia. *Neuroscience & Biobehavioral Reviews*, *61*, 108–120.

Strauss, J. S. (1969). Hallucinations and delusions as points on continua function: Rating scale evidence. *Archives of general psychiatry*, *21*(5), 581–586.

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1). https://doi.org/https://doi.org/10.1038/s41398-020-0780-3

The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. (2017). Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular autism*, *8*, 1–17.

Tiao, L. C. (2017). A Tutorial on Variational Autoencoders with a Concise Keras Implementation. *tiao.io*. https://tiao.io/post/tutorial-on-variational-autoencoders-with-a-concise-keras-implementation/

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(3), 611–622.

Unschuld, P. G., Buchholz, A. S., Varvaris, M., Van Zijl, P. C., Ross, C. A., Pekar, J. J., Hock, C., Sweeney, J. A., Tamminga, C. A., Keshavan, M. S., et al. (2014). Prefrontal brain network connectivity indicates degree of both schizophrenia risk and cognitive dysfunction. *Schizophrenia bulletin*, *40*(3), 653–664.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: An overview. *Neuroimage*, *80*, 62–79.

Van Os, J., Hanssen, M., Bijl, R. V., & Ravelli, A. (2000). Strauss (1969) revisited: A psychosis continuum in the general population? *Schizophrenia research*, *45*(1-2), 11–20.

Van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., & Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: Evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. *Psychological medicine*, *39*(2), 179–195.

van Erp, T. G., Preda, A., Turner, J. A., Callahan, S., Calhoun, V. D., Bustillo, J. R., Lim, K. O., Mueller, B., Brown, G. G., Vaidya, J. G., et al. (2015). Neuropsychological profile in adult schizophrenia measured with the cminds. *Psychiatry research*, *230*(3), 826–834.

Verdoux, H., & van Os, J. (2002). Psychotic symptoms in non-clinical populations and the continuum of psychosis. *Schizophrenia research*, *54*(1-2), 59–65.

Wang, X., Xia, M., Lai, Y., Dai, Z., Cao, Q., Cheng, Z., Han, X., Yang, L., Yuan, Y., Zhang, Y., et al. (2014). Disrupted resting-state functional connectivity in minimally treated chronic schizophrenia. *Schizophrenia research*, *156*(2-3), 150–156.

Yan, W., Pearlson, G. D., Fu, Z., Li, X., Iraji, A., Chen, J., Sui, J., Volkow, N. D., & Calhoun, V. D. (2024). A brainwide risk score for psychiatric disorder evaluated in a large adolescent population reveals increased divergence among higher-risk groups relative to control participants. *Biological Psychiatry*, *95*(7), 699–708.

Yoshihara, Y., Lisi, G., Yahata, N., Fujino, J., Matsumoto, Y., Miyata, J., Sugihara, G.-i., Urayama, S.-i., Kubota, M., Yamashita, M., et al. (2020). Overlapping but asymmetrical relationships between schizophrenia and autism revealed by brain connectivity. *Schizophrenia Bulletin*, *46*(5), 1210–1218.

Yu, Q., Allen, E. A., Sui, J., Arbabshirani, M. R., Pearlson, G., & Calhoun, V. D. (2012). Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current topics in medicinal chemistry*, *12*(21), 2415–2425. https://doi.org/https://doi.org/10.2174/156802612805289890

Zeng, L.-L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., Wang, K., & Hu, D. (2018). Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri. *EBioMedicine*, *30*, 74–85. https://doi.org/https://doi.org/10.1016/j.ebiom.2018.03.017

Zhang, F., Wei, Y., Liu, J., Wang, Y., Xi, W., & Pan, Y. (2022). Identification of autism spectrum disorder based on a novel feature selection method and variational autoencoder. *Computers in Biology and Medicine*, *148*, 105854.

Zhang, X., Maltbie, E. A., & Keilholz, S. D. (2021). Spatiotemporal trajectories in resting-state fmri revealed by convolutional variational autoencoder. *NeuroImage*, *244*, 118588.

Appendix A ABIDE I data information

In the ABIDE I dataset, we used 266 subjects from the following three sites: NYU Langone Medical Center (NYU), University of California, Los Angeles: Sample 1 (UCLA), and University of Utah School of Medicine (USM). Site-specific demographic information is shown in Table A.1. fMRI acquisition parameters for each site are shown in Table A.2.

Based on the mapping of latent features to 2D grid nodes (as in Figure 4), we show diagnosis labels, autism diagnostic observation schedule (ADOS) scores, age and gender information of ABIDE I subjects on the 2D grid (Figure A.1). We note that the diagnosis label plays a key role in grouping subjects, with most ASD patients clustered in the upper triangle and most controls in the lower triangle.

Table A.1: **ABIDE I site-specific demographic information.** N: number of total subjects; N_{male}: number of male subjects; N_{female}: number of female subjects; s.d.: standard deviation.

Site	N	N_{male}	N_{female}	Age mean $\pm$ s.d. (years)	Age range (years)
			ASD		
NYU	64	55	9	15.72 ± 7.11	$7.15 - 39.10$
UCLA	41	36	5	13.54 ± 2.25	$9.26 - 17.94$
USM	28	28	0	25.05 ± 6.11	$15.93 - 37.78$
			CTR		
NYU	95	70	25	16.45 ± 6.12	$7.29 - 31.78$
UCLA	29	25	4	13.35 ± 1.96	$9.50 - 17.79$
USM	9	9	0	20.57 ± 4.65	$13.81 - 27.60$

Table A.2: **ABIDE I site-specific fMRI acquisition parameters.** Parameters (units) are shown as follows: TR (ms), TE (ms), FA (degree), voxel size (mm).

Site	Scanner	TR	TE	FA	Voxel size	Slices
NYU	Siemens Magnetom Allegra syngo MR 2004A	2000	15	90	$3.0 \times 3.0 \times 4.0$	33
UCLA	Siemens Magnetom TrioTim syngo MR B15	3000	28	90	$3.0 \times 3.0 \times 4.0$	34
USM	Siemens Magnetom TrioTim syngo MR B17	2000	28	90	$3.4 \times 3.4 \times 3.0$	40

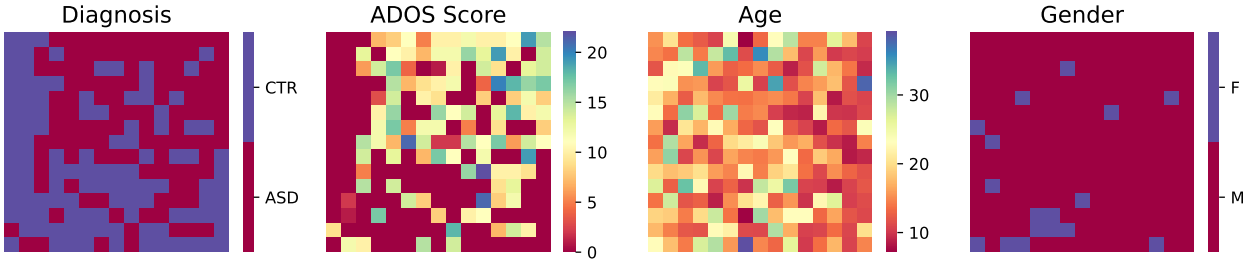


Figure A.1: **ABIDE I subject measures displayed on the 2D grid.** Diagnosis labels, autism diagnostic observation schedule (ADOS) scores, age and gender information of ABIDE I subjects were displayed on the 2D grid correspondingly.

Appendix B Statistics of subject measures

Table B.1 summarizes the group-level mean, standard deviation, and range for each subject measure (diagnostic or cognitive score) in each dataset.

Table B.1: **Statistics of subject measures.** PT: patient; CTR: control.

Subject measure	PT mean ± s.d.	PT range	CTR mean ± s.d.	CTR range
		FBIRN		
PANSS (positive)	14.976 ± 4.75	$7 - 29$	7.000 ± 0.00	$7 - 7$
PANSS (negative)	13.928 ± 5.00	$7 - 30$	7.000 ± 0.00	$7 - 7$
Speed of processing	-1.217 ± 0.98	$-3.682 - 1.165$	0.038 ± 1.00	$-3.764 - 2.195$
Attention/Vigilance	-1.277 ± 1.36	$-4.965 - 1.390$	0.026 ± 0.98	$-4.965 - 1.542$
Working memory	-1.018 ± 0.97	$-3.947 - 2.053$	0.020 ± 0.96	$-3.019 - 2.053$
Verbal learning	-1.313 ± 1.17	$-4.034 - 1.561$	0.042 ± 0.99	$-3.101 - 1.561$
Visual learning	-1.002 ± 1.13	$-2.873 - 1.152$	0.026 ± 0.99	$-2.873 - 1.646$
Reasoning/Problem solving	-0.756 ± 1.11	$-4.443 - 0.981$	0.022 ± 0.97	$-4.805 - 0.981$
Composite score	-1.562 ± 1.22	$-4.502 - 1.496$	0.044 ± 0.95	$-4.023 - 1.550$
		ABIDE I		
ADOS score	11.594 ± 4.00	$2 - 22$	0.090 ± 0.47	$0 - 4$

Appendix C Variational autoencoders

Here, we provide the step-by-step derivation of the VAE objective function. We refer readers to Kingma and Welling, 2014 for the original theory and to Tiao, 2017 for a detailed tutorial.

Let $\mathbf{x} \in \mathbb{R}^V$ be an observed data variable and $\mathbf{z} \in \mathbb{R}^D$ be its corresponding latent variable. We assume that the latent variable $\mathbf{z}$ is sampled from a prior distribution $p(\mathbf{z})$ and the observed variable $\mathbf{x}$ is sampled from the conditional likelihood distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ parameterized by $\boldsymbol{\theta}$. We consider a latent variable model with the following joint distribution:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}). \tag{7}$$

We would like to estimate the posterior distribution in order to perform inference on the latent variable. According to Bayes' theorem, the posterior can be computed as:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}}. \tag{8}$$

The denominator $p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$ is the marginal distribution of the observed data, also known as the *evidence*. It is analytically intractable because there is no closed form for the integral term.

To overcome this intractability, we leverage *variational inference*, which assumes a simpler distribution to approximate the true posterior. We want to estimate the optimal variational parameters $\boldsymbol{\phi}^*$ by minimizing the Kullback-Leibler (KL) divergence[2] between the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and the true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} D_{\mathsf{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right), \tag{9}$$

where $D_{\mathsf{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right) = 0$ if and only if $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. This is again intractable due to $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. However, note that the true posterior is proportional to the joint distribution, $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) \propto p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. Thus, we can minimize the following KL divergence instead:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} D_{\mathsf{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\right). \tag{10}$$

The *evidence lower bound* (ELBO) is defined to maximize the negative KL divergence in Equation 10 by

[2]The KL divergence measures how a probability distribution differs from another probability distribution. For a discrete random variable, $D_{\mathsf{KL}}(p(x)\|q(x)) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$; for a continuous random variable, $D_{\mathsf{KL}}(p(x)\|q(x)) = \int_{-\infty}^{+\infty} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$.

optimizing the generative parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$:

$$\begin{aligned}
\max \mathcal{L}_{\mathsf{VAE}} &= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} -D_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})) \\
&= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z})\right] \\
&= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\
&= \max_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})).
\end{aligned} \tag{11}$$

Crucially, maximizing the ELBO with respect to the generative parameters $\boldsymbol{\theta}$ approximately maximizes the log marginal likelihood of the observed data, while maximizing it with respect to the variational parameters $\boldsymbol{\phi}$ equivalently minimizes the KL divergence. Also, note that the KL divergence is non-negative, i.e. $D_{\mathsf{KL}}(\cdot) \geq 0$. Thus, the ELBO is the lower bound to the log likelihood of generating the observed data, i.e. $\mathcal{L}_{\mathsf{ELBO}} \leq \log p_{\boldsymbol{\theta}}(\mathbf{x})$.

In this work, we assume that both the variational posterior and the prior follow multivariate Gaussian distributions:

$$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \mathsf{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})\right)\right), \quad p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}|\mathbf{0}, \mathbf{I}\right). \tag{12}$$

Then the KL term in Equation 11 can be rewritten as:

$$\begin{aligned}
D_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z})\right] && (13) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{1}{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mathbf{z}-\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})}{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})}\right)^2}\right) - \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mathbf{z}^2}\right)\right] && (14) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{1}{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\mathbf{z}-\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})}{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})}\right)^2 - \log\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{1}{2}\mathbf{z}^2\right] && (15) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{\sqrt{2\pi}}{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})\sqrt{2\pi}}\right)\right] - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\frac{(\mathbf{z}-\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}))^2}{\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})}\right] + \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\mathbf{z}^2\right] && (16) \\
&= -\log\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) - \frac{1}{2} + \frac{1}{2}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\mu}^2_{\boldsymbol{\phi}}(\mathbf{x})\right) && (17) \\
&= -\frac{1}{2}\cdot\left(1 + \log\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) - \boldsymbol{\mu}^2_{\boldsymbol{\phi}}(\mathbf{x}) - \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})\right). && (18)
\end{aligned}$$

To simplify Equation 16 to Equation 17, we leverage the definition of variance,

$$\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{z}-\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}))^2\right] = \mathbb{E}\left[\mathbf{z}^2\right] - \mathbb{E}\left[\mathbf{z}\right]^2. \tag{19}$$

Thus, $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[(\mathbf{z}-\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}))^2\right] = \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})$ and $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\mathbf{z}^2\right] = \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\mu}^2_{\boldsymbol{\phi}}(\mathbf{x})$.

In order to use gradient descent, the VAE objective function is defined to minimize the reconstruction error measured by the mean squared error (MSE) and the KL divergence in Equation 18, which equivalently minimizes the negative ELBO in Equation 11:

$$\mathcal{L}_{\text{VAE}} = \min_{\boldsymbol{\theta},\boldsymbol{\phi}} \frac{1}{V}\sum_{i=1}^{V}(x_i - \hat{x}_i)^2 - \frac{1}{2}\sum_{j=1}^{D}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right), \tag{20}$$

where x_i and $\hat{x}_i$ are the i-th element of the observed data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}$, respectively; μ_j and σ_j are the j-th element of $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})$, respectively.

To train the ELBO with gradient-based optimization, we use the *reparameterization trick* which expresses the random variable $\mathbf{z}$ as a deterministic transformation of the input $\mathbf{x}$ and a Gaussian random variable with zero mean and unit variance $\boldsymbol{\epsilon} \in \mathbb{R}^D$:

$$\mathbf{z} = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{21}$$

where $\odot$ represents the element-wise product.

Appendix D Probabilistic principal component analysis

In PPCA, we assume the prior distribution over the latent variable $\mathbf{z} \in \mathbb{R}^D$ and the conditional distribution of the observed variable $\mathbf{x} \in \mathbb{R}^V$ given $\mathbf{z}$ follow Gaussian distributions:

$$p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}|\mathbf{0}, \mathbf{I}\right), \quad p(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right), \tag{22}$$

where the matrix $\mathbf{W} \in \mathbb{R}^{V\times D}$ and the vector $\boldsymbol{\mu} \in \mathbb{R}^V$ define the mean of $\mathbf{x}$, and the scalar $\sigma^2 \in \mathbb{R}$ determines the variance of the conditional distribution.

The marginal distribution of the observed data is also a Gaussian:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}(\mathbf{z}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}\right), \tag{23}$$

where the covariance matrix $\mathbf{C} \in \mathbb{R}^{V\times V}$ is defined by $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$. Note that $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Thus, $\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$ and $\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbb{E}[\mathbf{W}\mathbf{z}\mathbf{z}^\top\mathbf{W}^\top] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

Using Bayes' rule, the posterior distribution can be derived as:

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right), \tag{24}$$

where the matrix $\mathbf{M} \in \mathbb{R}^{D\times D}$ is given by $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Given a dataset of observed samples $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}^{(N)}\right\}$, the corresponding log likelihood function can be written as:

$$\ln p\left(\mathcal{D}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = \sum_{i=1}^{N} \ln p\left(\mathbf{x}^{(i)}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) \tag{25}$$

$$= -\frac{NV}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}\right)^\top \mathbf{C}^{-1}\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}\right). \tag{26}$$

Using maximum likelihood estimation, we obtain the model parameter $\boldsymbol{\mu} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}$ by setting the derivative of the log likelihood with respect to $\boldsymbol{\mu}$ equal to zero.

We implement the expectation-maximization (EM) algorithm to estimate the model parameters $\mathbf{W}$ and σ^2. In the E step, we use the old parameter values to evaluate the expectation of the complete-data log

likelihood with respect to the posterior distribution:

$$\mathbb{E}\left[\ln p\left(\mathcal{D}, \mathcal{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)\right] = -\sum_{i=1}^{N}\left(\frac{V}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2}\operatorname{tr}\left(\mathbb{E}\left[\mathbf{z}^{(i)}\mathbf{z}^{(i)\top}\right]\right) + \frac{1}{2\sigma^2}\left\|\mathbf{x}^{(i)} - \boldsymbol{\mu}\right\|^2 \right.$$
$$\left. - \frac{1}{\sigma^2}\mathbb{E}\left(\mathbf{z}^{(i)}\right)^\top \mathbf{W}^\top\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}\right) + \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbb{E}\left[\mathbf{z}^{(i)}\mathbf{z}^{(i)\top}\right]\mathbf{W}^\top\mathbf{W}\right) + \frac{D}{2}\ln(2\pi)\right), \tag{27}$$

$$\mathbb{E}\left[\mathbf{z}^{(i)}\right] = \mathbf{M}^{-1}\mathbf{W}^\top\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right), \tag{28}$$

$$\mathbb{E}\left[\mathbf{z}^{(i)}\mathbf{z}^{(i)\top}\right] = \sigma^2\mathbf{M}^{-1} + \mathbb{E}\left[\mathbf{z}^{(i)}\right]\mathbb{E}\left[\mathbf{z}^{(i)}\right]^\top, \tag{29}$$

where $\mathbf{z}^{(i)}$ is the i-th element in the set of latent variables $\mathcal{Z} = \left\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(i)}, \ldots, \mathbf{z}^{(N)}\right\}$.

In the M step, we maximize the the expected complete-data log likelihood with respect to $\mathbf{W}$ and σ^2, yielding the new parameter values:

$$\mathbf{W}_{\text{new}} = \left[\sum_{i=1}^{N}\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right)\mathbb{E}\left[\mathbf{z}^{(i)}\right]^\top\right]\left[\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{z}^{(i)}\mathbf{z}^{(i)\top}\right]\right]^{-1}, \tag{30}$$

$$\sigma^2_{\text{new}} = \frac{1}{VN}\sum_{i=1}^{N}\left(\left\|\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right\|^2 - 2\mathbb{E}\left[\mathbf{z}^{(i)}\right]^\top\mathbf{W}_{\text{new}}^\top\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right) + \operatorname{tr}\left(\mathbb{E}\left[\mathbf{z}^{(i)}\mathbf{z}^{(i)\top}\right]\mathbf{W}_{\text{new}}^\top\mathbf{W}_{\text{new}}\right)\right). \tag{31}$$

We iteratively alternate between the E step and the M step, until the parameter values converge. For more details, we refer readers to Tipping and Bishop, 1999 or Chapter 16, Bishop and Bishop, 2023.

Appendix E Identifiable variational autoencoders

In iVAE, we consider the following conditional generative model:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\epsilon}, \tag{32}$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}), \tag{33}$$

$$p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\boldsymbol{\epsilon}}\left(\mathbf{x} - \mathbf{f}(\mathbf{z})\right), \tag{34}$$

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{D} \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp\left[\sum_{j=1}^{s} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\right], \tag{35}$$

where $\mathbf{x} \in \mathbb{R}^V$ is the observed data variable, $\mathbf{u} \in \mathbb{R}^U$ is the observed auxiliary variable, $\mathbf{z} \in \mathbb{R}^D$ is the latent variable, $\boldsymbol{\epsilon} \in \mathbb{R}^V$ is a noise variable with probability density function $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$, $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ is a set of parameters of the generative model, and $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^V$ is a nonlinear mixing function. Since our main interest lies in psychiatric characteristics, we use the one-hot encoded diagnostic label as the auxiliary variable $\mathbf{u}$ in the iVAE ($U = 2$). We assume that the prior on the latent variable $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$ is conditionally independent, and each z_i follows a univariate exponential family distribution given the auxiliary variable $\mathbf{u}$, where Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant, $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,s})$ are the sufficient statistics, $\boldsymbol{\lambda}_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}), \ldots, \lambda_{i,s}(\mathbf{u}))$ are the parameters depending on $\mathbf{u}$, and s is the dimension of each sufficient statistic.

Similar to VAE, the iVAE objective function aims to learn the parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ that maximize the data generation likelihood by maximizing the ELBO:

$$\mathcal{L}_{\text{iVAE}} = \max_{\boldsymbol{\theta},\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) - D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})\right), \tag{36}$$

where $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u})\, \mathrm{d}\mathbf{z}$ is the conditional marginal distribution of the observed data, and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ is the variational posterior of its true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$.

For more information, please refer to Khemakhem et al., 2020.

Appendix F Model architecture search

F.1 VAE architecture search

We evaluated four VAE architectures based on multilayer perceptrons (MLPs). Each architecture has a different set of output units in the VAE encoder and the corresponding symmetric output units in the VAE decoder. The number of output units in each encoder layer is listed in Table F.1. In each layer, we used a Leaky Rectified Linear Unit (Leaky ReLU) activation function (Maas et al., 2013) with a negative slope of 0.5.

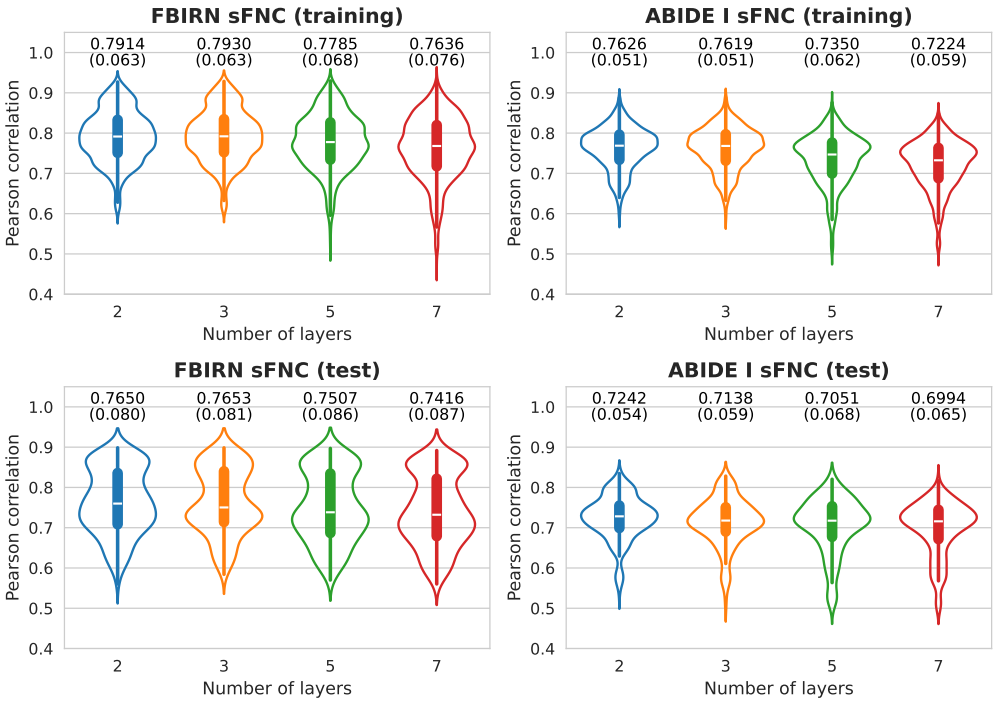
For each dataset (FBIRN or ABIDE I) and each data type (sFNC or dFNC), we evaluated each of four potential model architectures separately. For each architecture, we used 10 different random seeds to initialize model weights. As shown in Figure F.1, we find that smaller models with 2 or 3 layers achieved optimal performance for sFNC data, while larger models with 5 or 7 layers worked better for dFNC data. Based on the training performance across 10 runs, we selected the following VAE architectures as shown in Table F.2.

Table F.1: **Evaluated VAE architectures.**

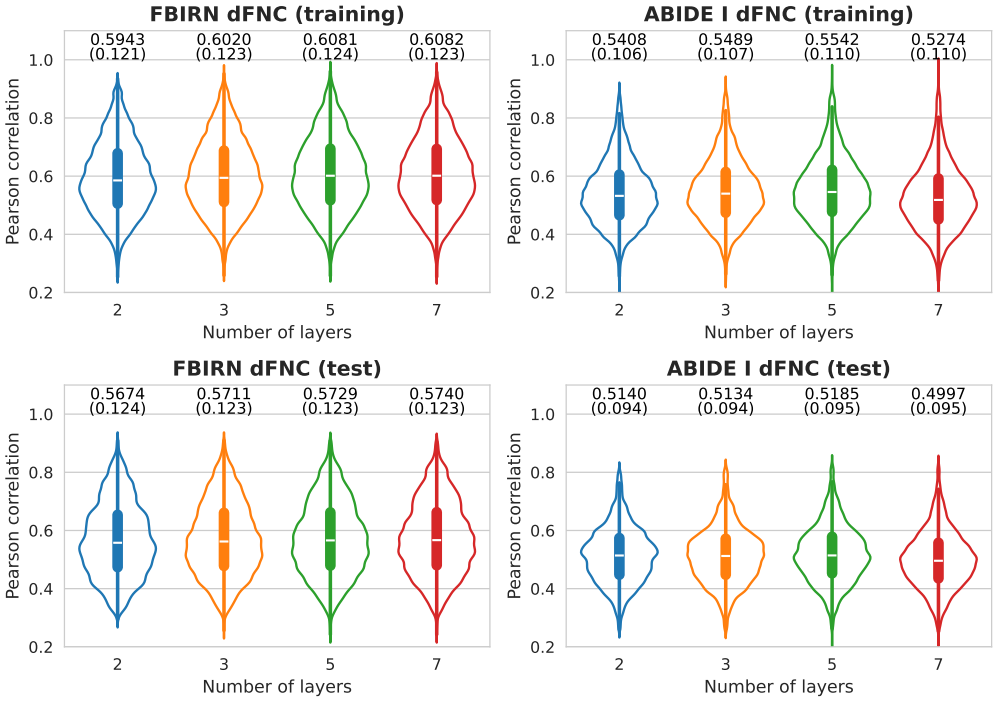
Number of layers	Output units in each encoder layer
2	64, 16
3	128, 64, 32
5	256, 128, 64, 32, 16
7	512, 256, 128, 64, 32, 16, 8

Table F.2: **Selected VAE architectures.**

Dataset	Data type	Number of layers	Output units in each encoder layer
FBIRN	sFNC	3	128, 64, 32
FBIRN	dFNC	7	512, 256, 128, 64, 32, 16, 8
ABIDE I	sFNC	2	64, 16
ABIDE I	dFNC	5	256, 128, 64, 32, 16



(a) sFNC data.



(b) dFNC data.

Figure F.1: **VAE architecture search.** Each violin plot shows a symmetric kernel density estimate of Pearson correlations between original and generated sFNC matrices across 10 random seeds, along with the quartiles represented by a box plot.

F.2 IVAE architecture search

We evaluated 12 MLP-based encoder-decoder architectures, varying the number of layers (2, 3, 4) and hidden dimensions (16, 32, 64, 96). According to the training performance (Figure F.2), we selected the iVAE model with 3 layers and 16 hidden units for FBIRN, and the model with 2 layers and 32 hidden units for ABIDE I.

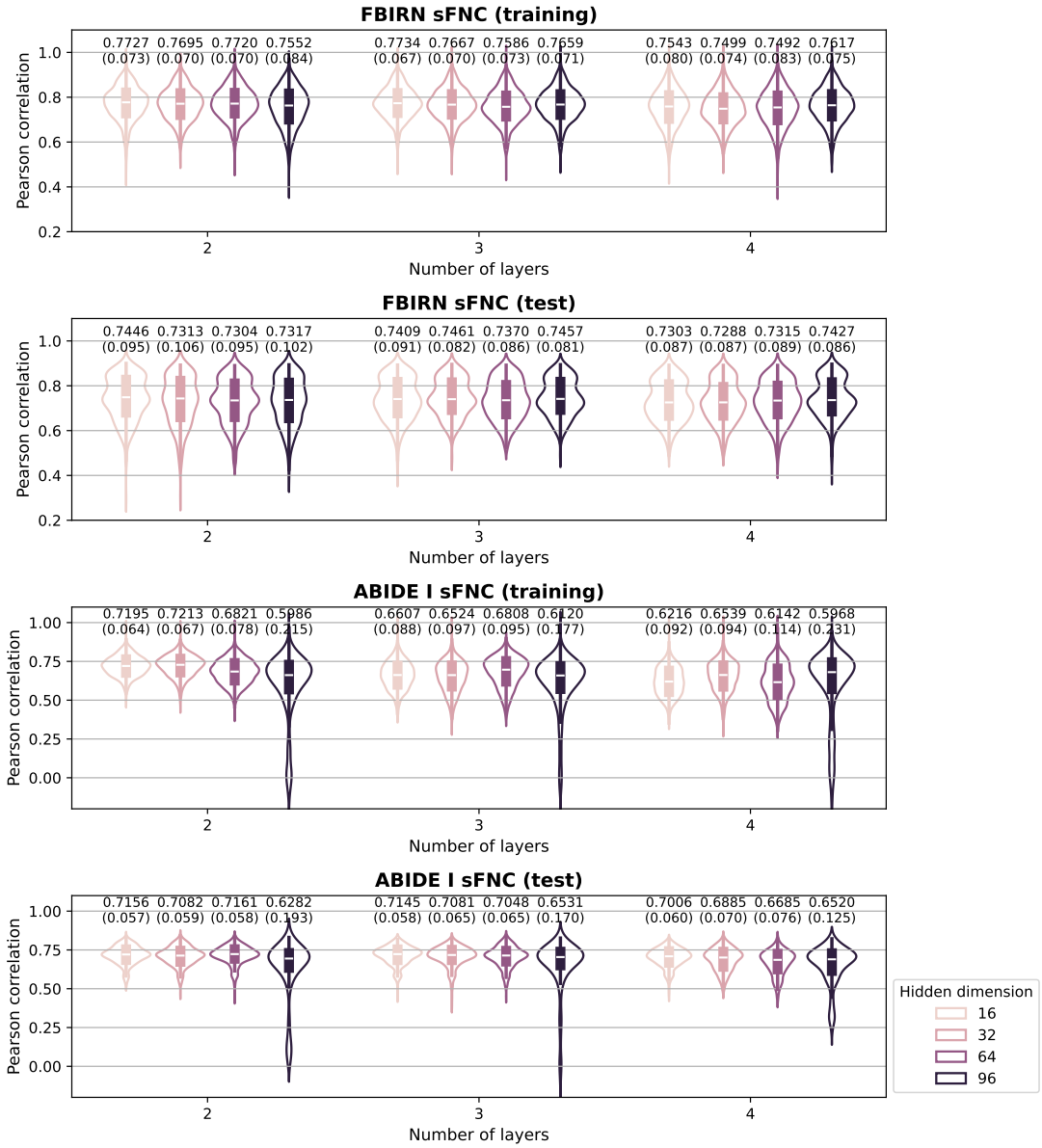


Figure F.2: **IVAE architecture search for sFNC data.** Each violin plot shows a symmetric kernel density estimate of Pearson correlations between original and generated sFNC matrices across 10 random seeds, along with the quartiles represented by a box plot.

Appendix G Confound effects on latent representations

To assess the robustness of the learned representations to potential confounding variables (e.g., age, sex, and collection site), we regressed these variables out of the latent representation ($\mathbf{z}$) using linear regression. For the FBIRN dataset, we constructed a confound matrix $\mathbf{C} = [\mathbf{1}, \mathbf{c}_{\text{age}}, \mathbf{c}_{\text{sex}}] \in \mathbb{R}^{N\times 3}$, where $\mathbf{1} \in \mathbb{R}^N$ is a column vector of ones, $\mathbf{c}_{\text{age}} \in \mathbb{R}^N$ is the demeaned linear age vector, and $\mathbf{c}_{\text{sex}} \in \mathbb{R}^N$ is the binary sex vector (0: female; 1: male). For the multi-site ABIDE I dataset, we also included the site information in the confound matrix, $\mathbf{C} = [\mathbf{1}, \mathbf{c}_{\text{age}}, \mathbf{c}_{\text{sex}}, \mathbf{c}_{\text{site}}] \in \mathbb{R}^{N\times 5}$, where $\mathbf{c}_{\text{site}} \in \mathbb{R}^{N\times 2}$ encodes the site information ($[0, 0]$: NYU; $[1, 0]$: UCLA; $[0, 1]$: USM). The latent representation after regressing out confounding variables ($\mathbf{z}'$) is calculated as follows:

$$\mathbf{z}' = \mathbf{z} - \mathbf{C}\left(\mathbf{C}^\top\mathbf{C}\right)^{-1}\mathbf{C}^\top\mathbf{z}. \tag{37}$$

Next, we calculated the randomized dependence coefficient (RDC) (Lopez-Paz et al., 2013), a measure of nonlinear dependence, between the representations before and after regression. An RDC value is bounded between 0 and 1, with higher values indicating greater similarity.

As shown in Figure G.1, the RDC values along the diagonal were very high (FBIRN: $0.989, 0.970$; ABIDE I: $0.976, 0.955$), indicating high similarity between the representations before and after regression. These results suggest that the learned representations appear to be largely insensitive to the examined confounding factors.

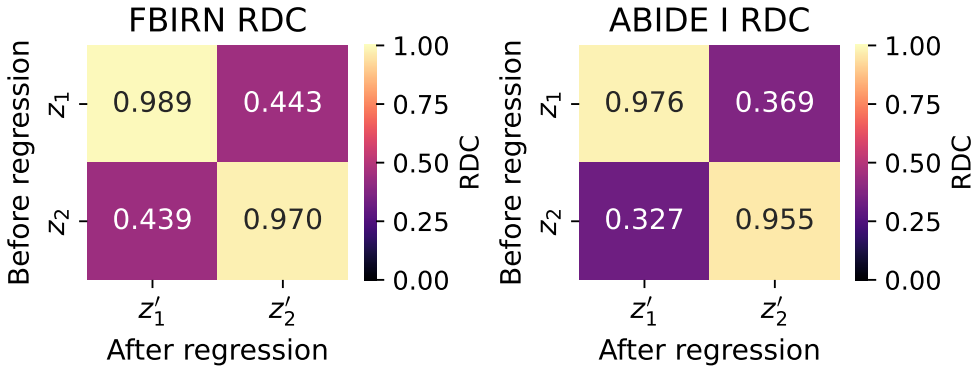


Figure G.1: **Randomized dependence coefficient (RDC) between the latent representations before and after regressing out confounding variables.** High diagonal RDC values indicate strong similarity between representations before and after regression, suggesting that the learned representations are largely unaffected by the examined confounding factors.

Appendix H sFNC visualization

H.1 FBIRN

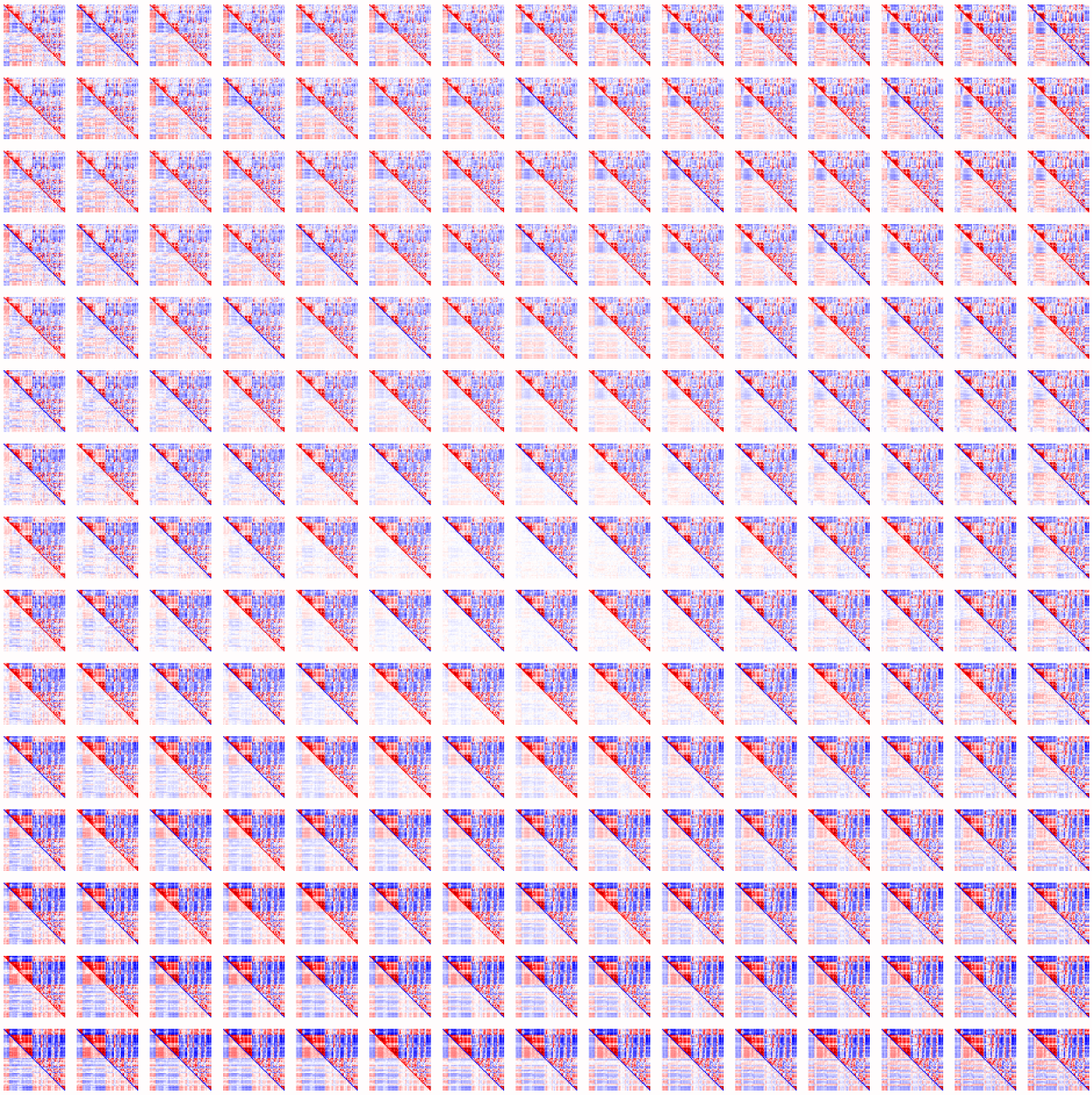


Figure H.1: **FBIRN generated sFNC (training).**

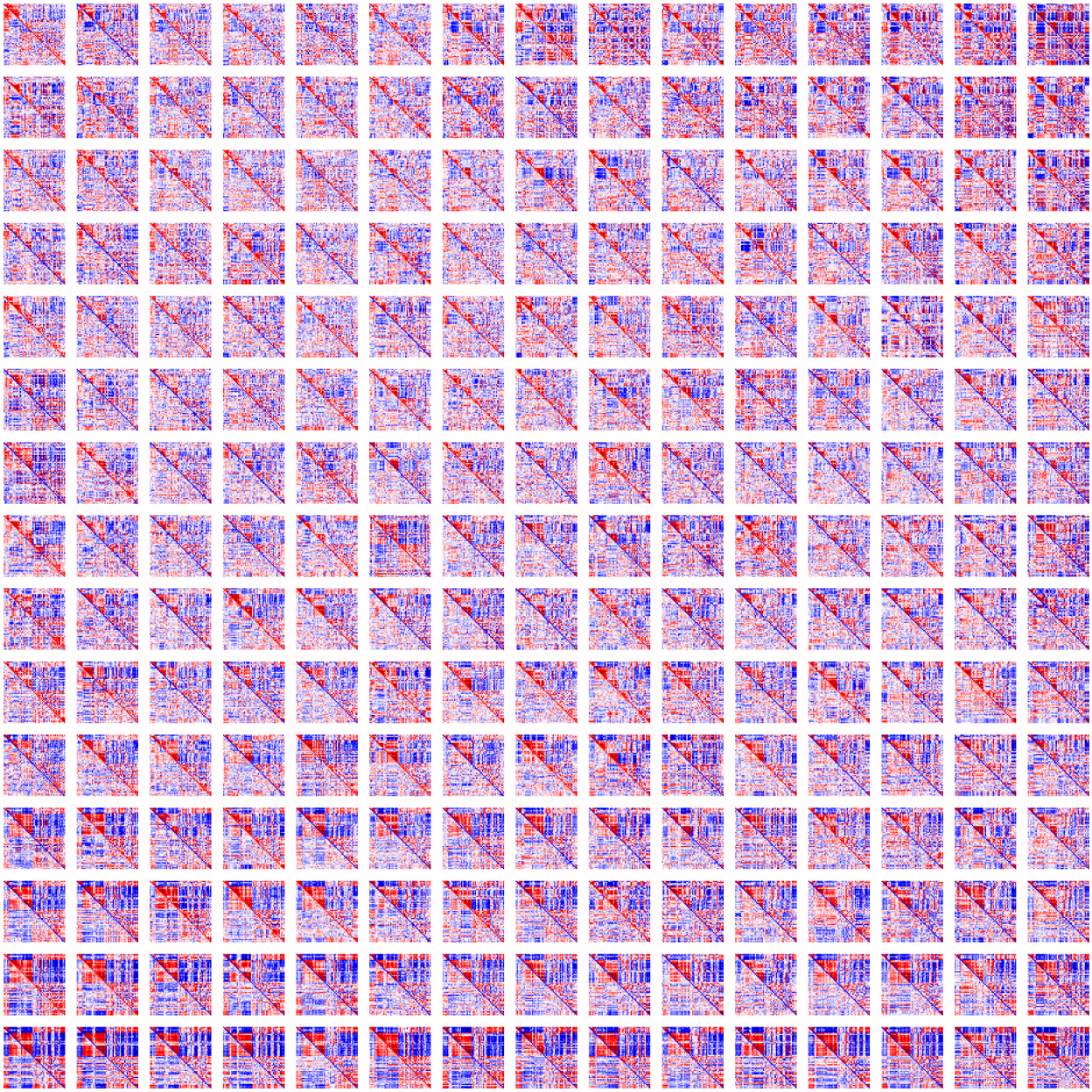


Figure H.2: **FBIRN original sFNC (training).**

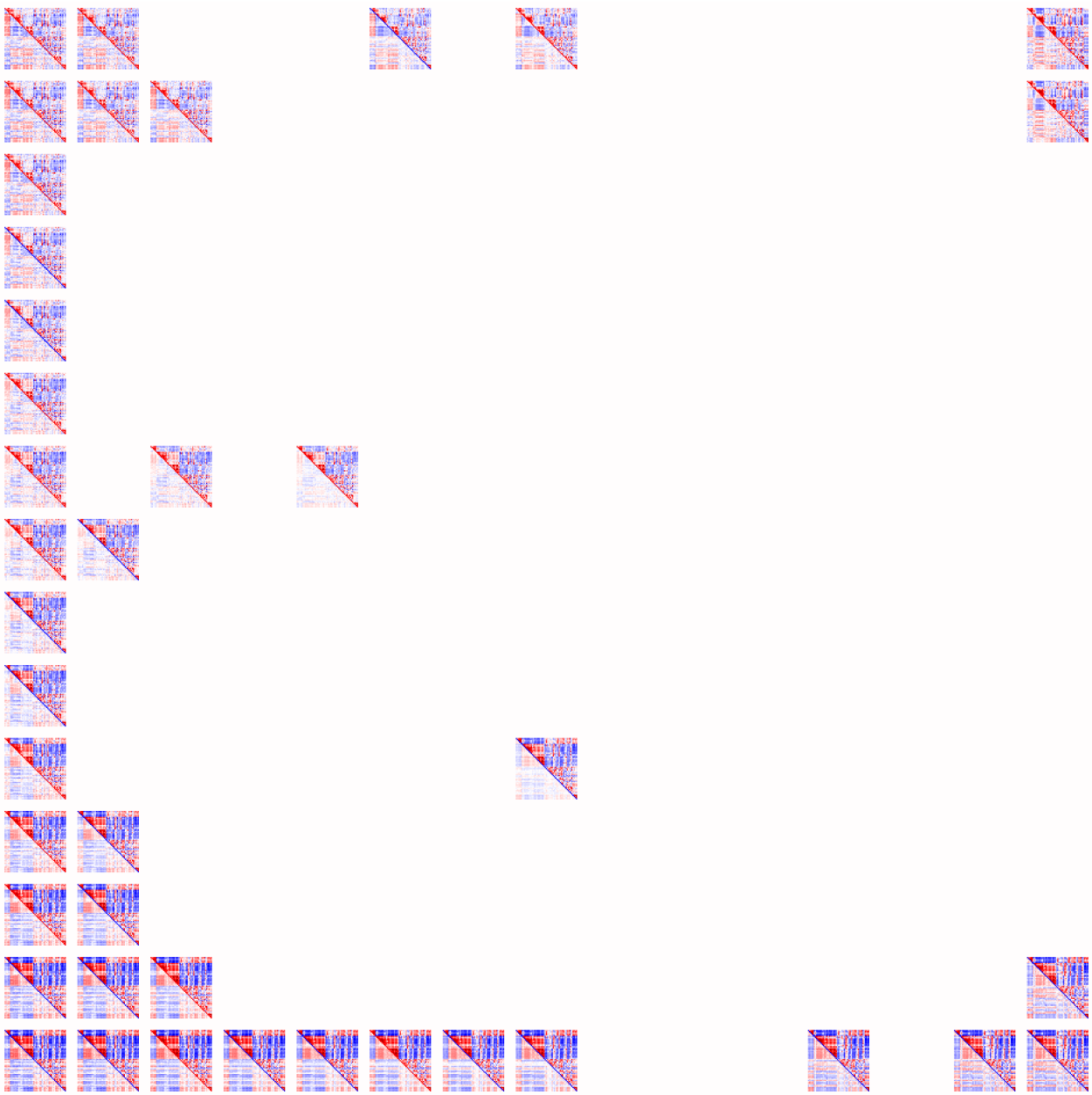


Figure H.3: **FBIRN generated sFNC (test).**

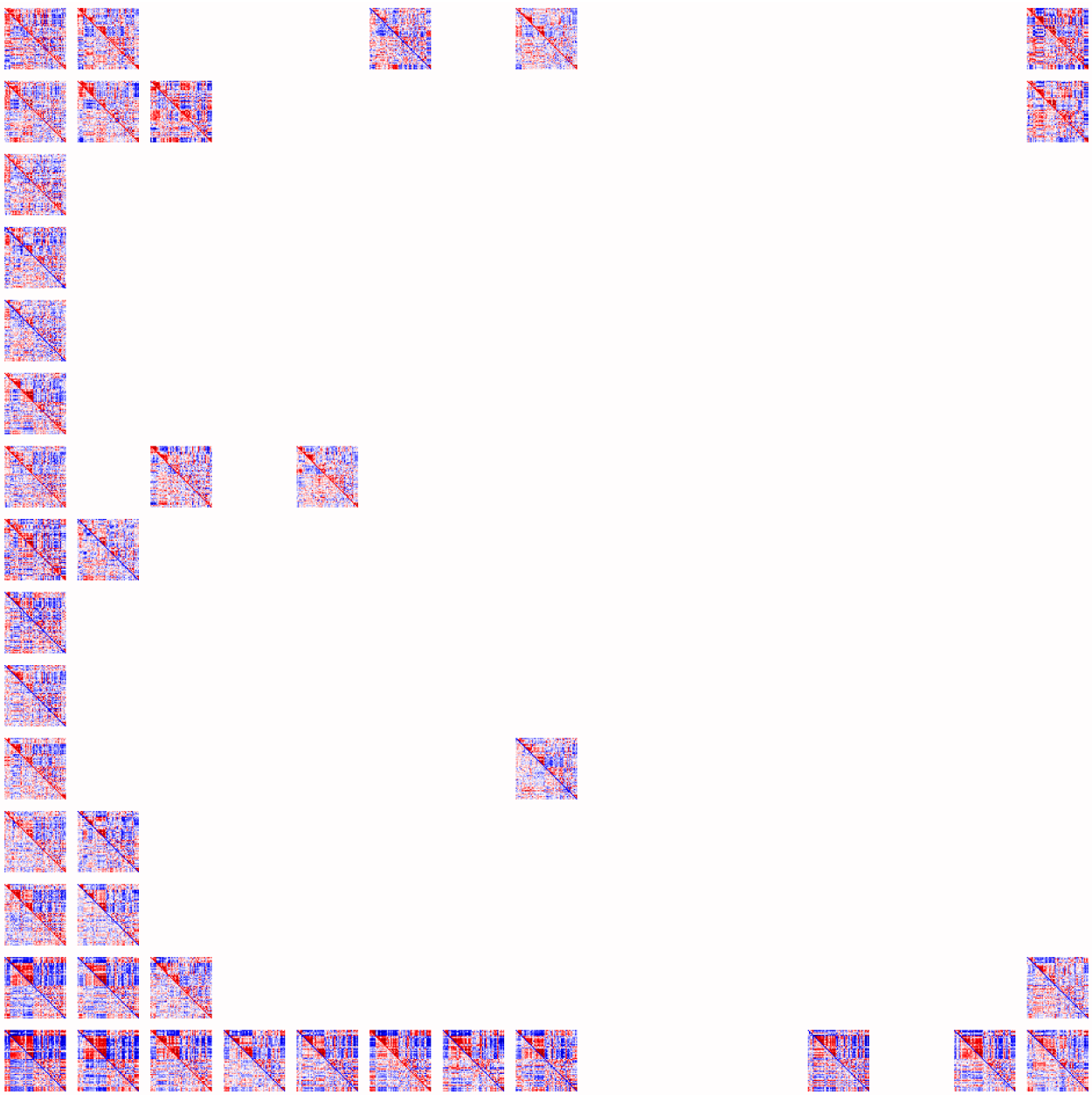


Figure H.4: **FBIRN original sFNC (test).**

H.2 ABIDE I

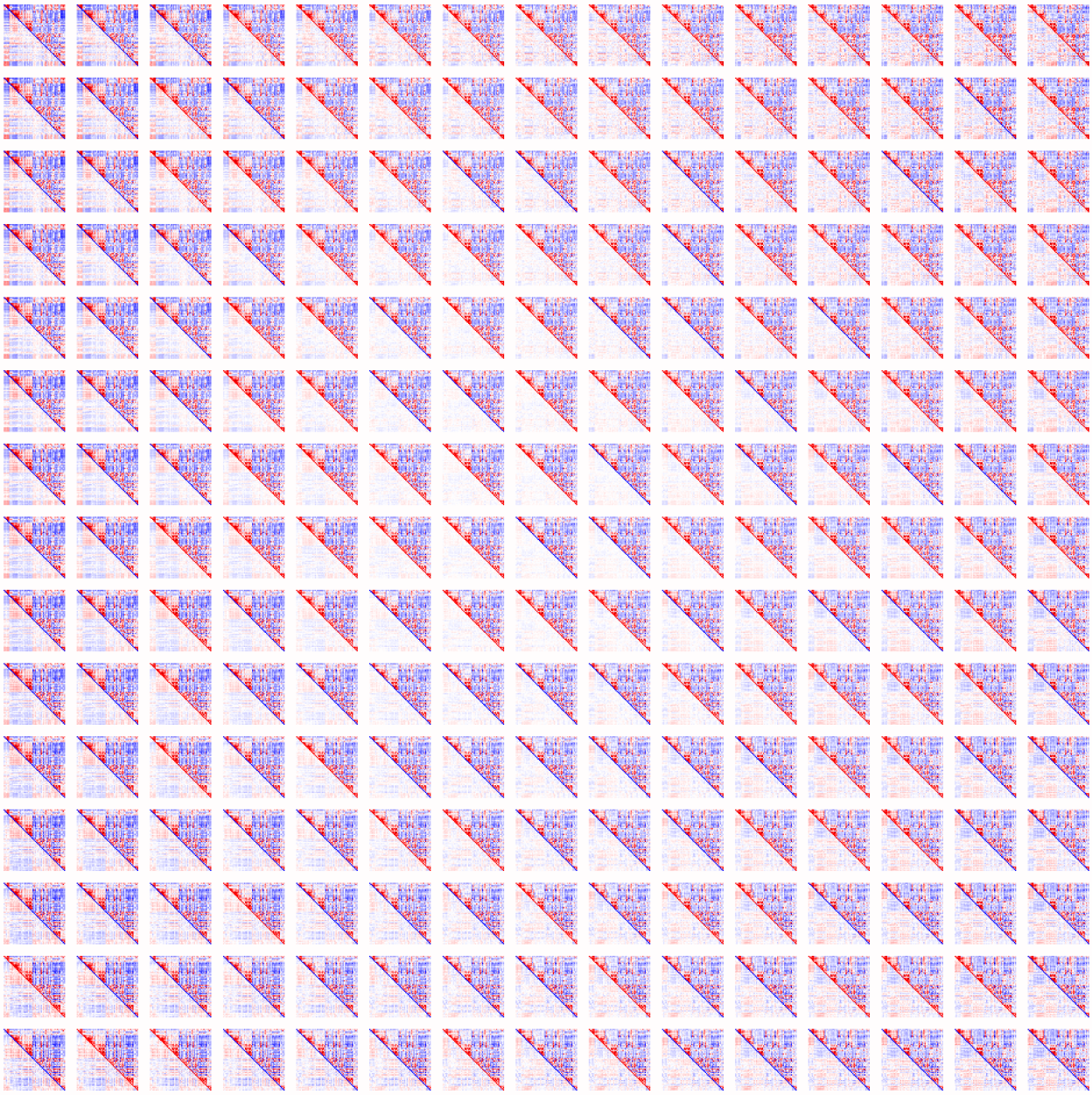


Figure H.5: **ABIDE I generated sFNC (training).**

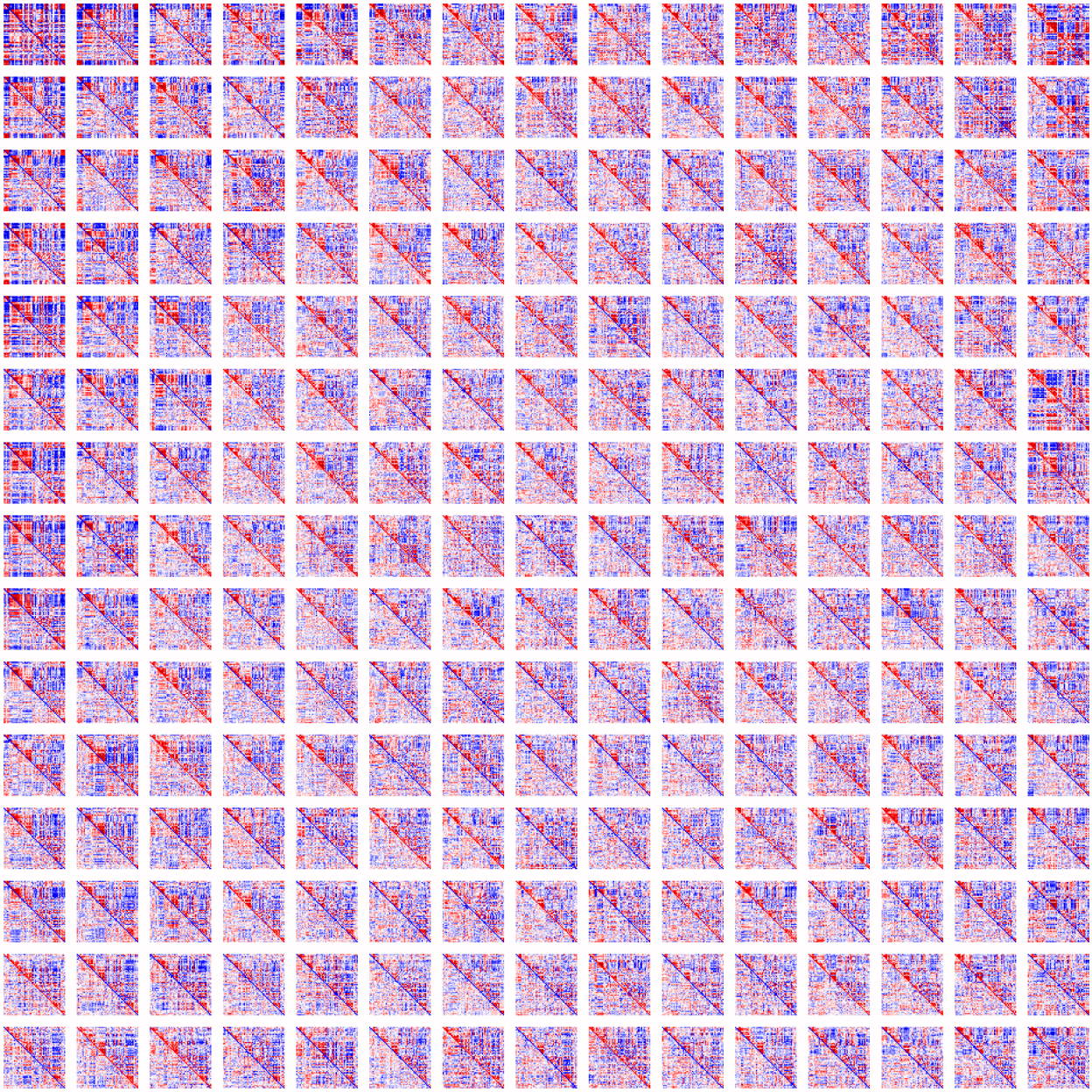


Figure H.6: **ABIDE I original sFNC (training).**

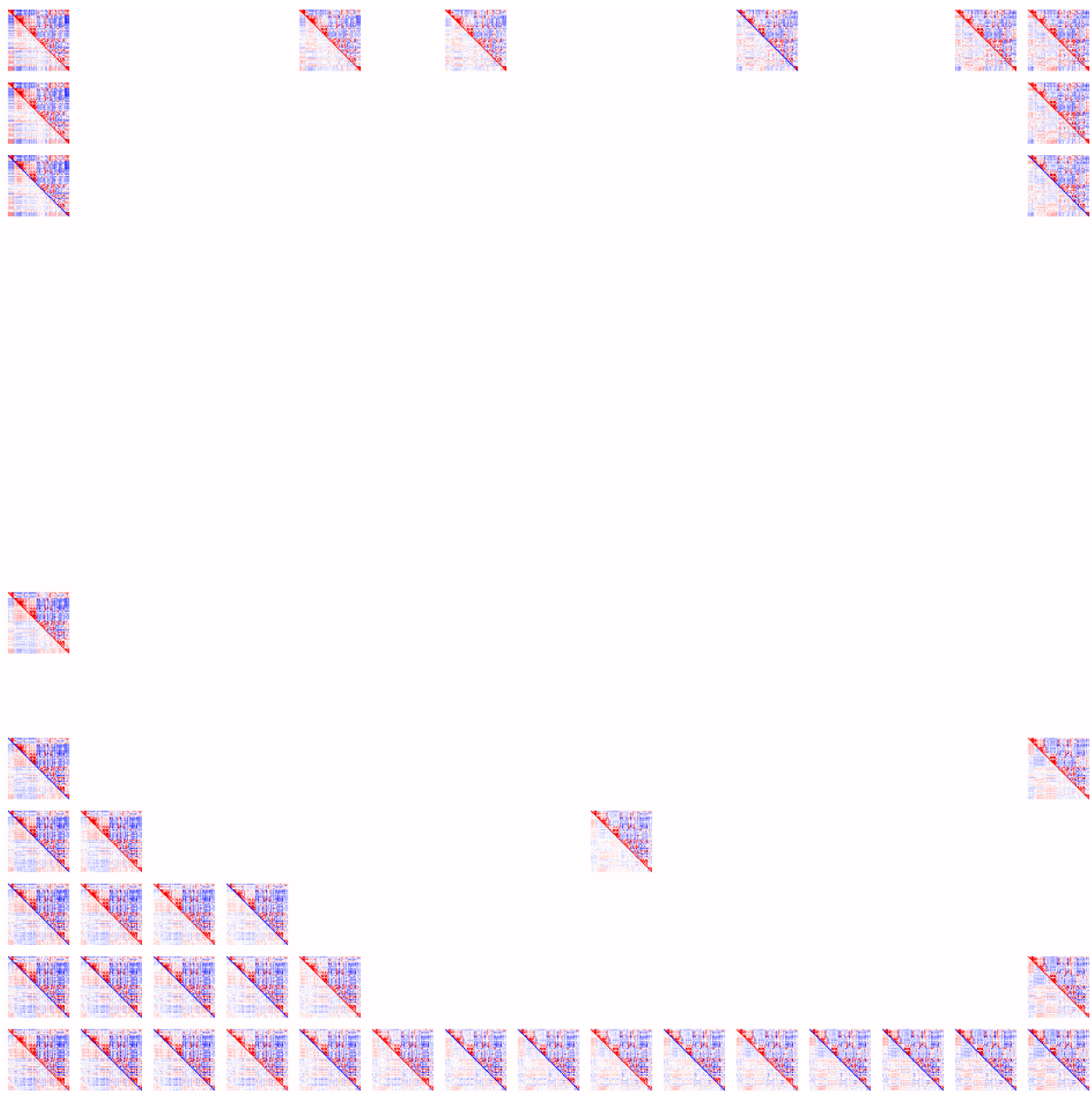


Figure H.7: **ABIDE I generated sFNC (test).**

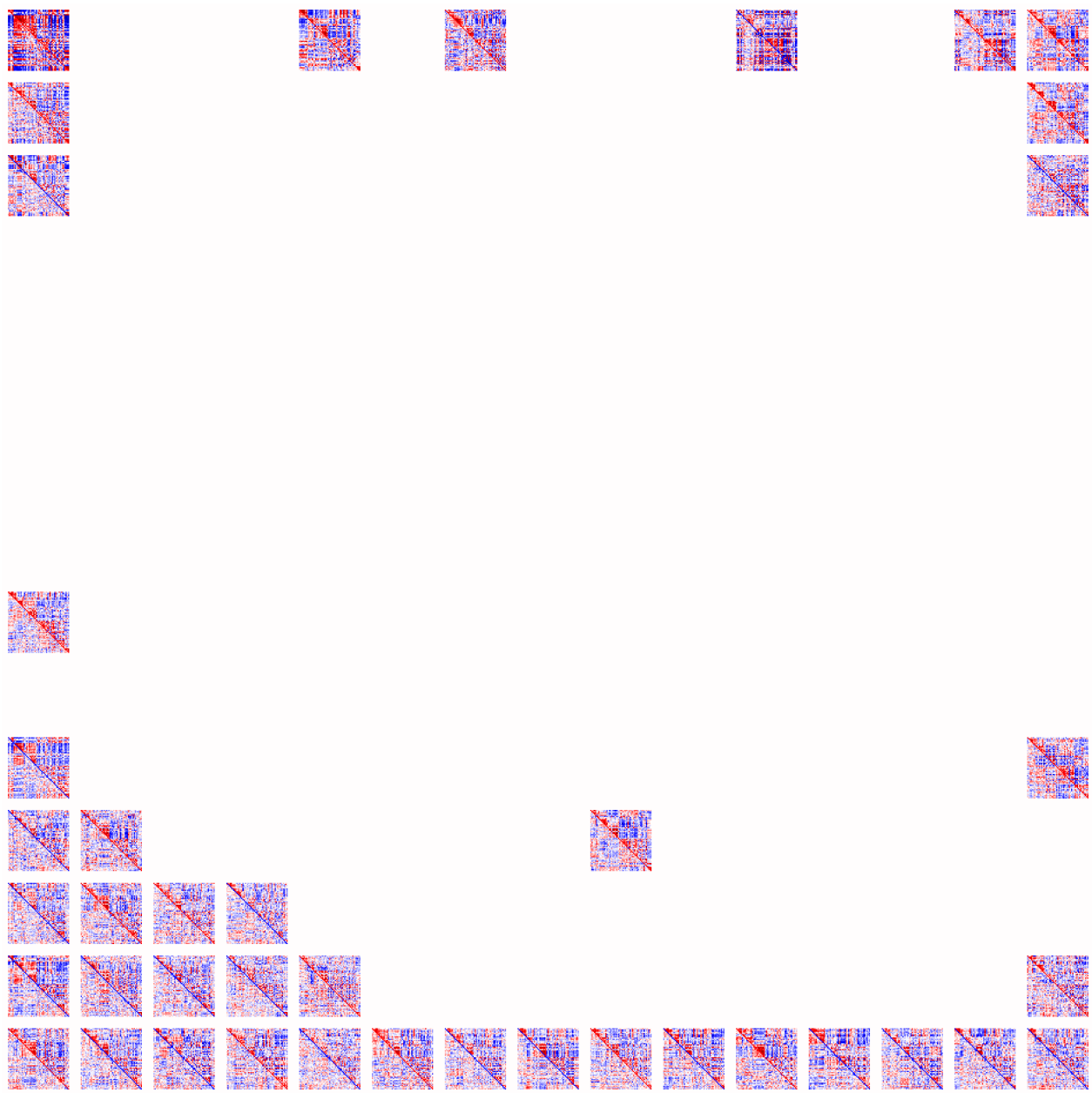


Figure H.8: **ABIDE I original sFNC (test).**

Appendix I Group differences in sFNC variability

To quantify heterogeneity in sFNC, we computed the cell-wise standard deviation of sFNC for each group and each dataset (Figure I.1). In the FBIRN dataset, the average standard deviation was slightly higher in the patient group (0.201) than in the control group (0.199). A similar pattern was observed in ABIDE I, where the patient group also showed a marginally higher average standard deviation (0.185) compared with controls (0.183). These results indicate slightly greater sFNC variability in the patient group across both datasets.

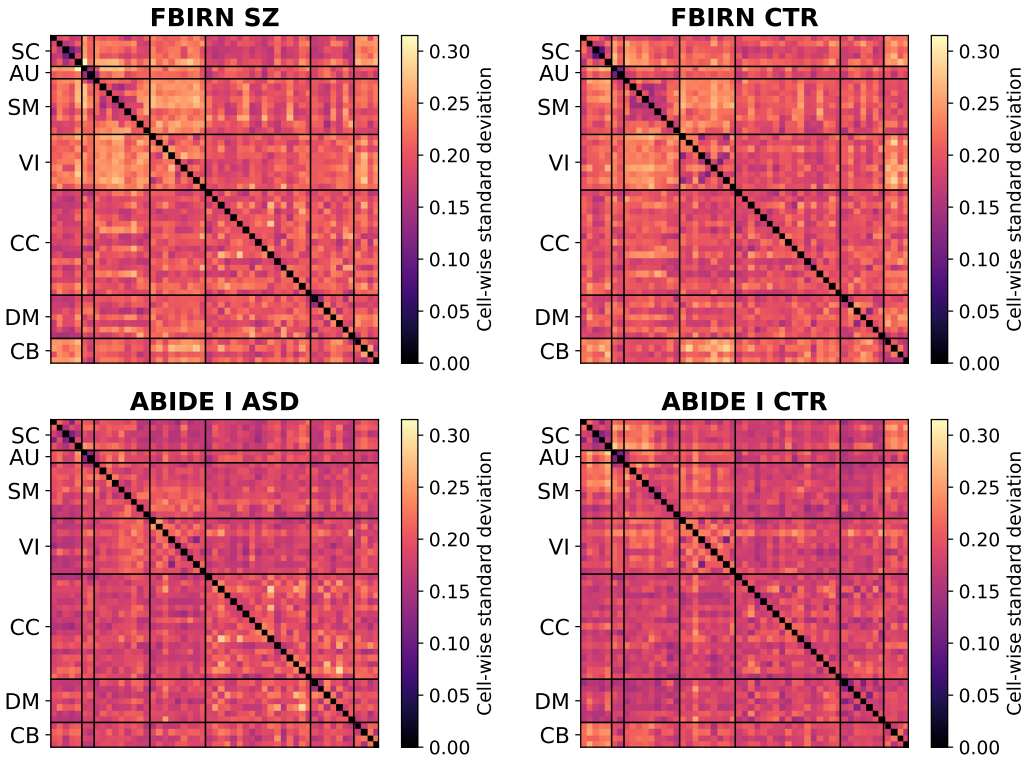


Figure I.1: **sFNC cell-wise standard deviation.** In both FBIRN and ABIDE I datasets, the average cell-wise standard deviation in the patient group is slightly higher than that in the control group, indicating slightly greater sFNC variability in the patient group.

Appendix J Subgroup analysis using two- and three-dimensional latent representations

We examined how the dimensionality of the VAE latent space affects downstream subgroup identification within the SZ patient cohort. First, we trained two VAEs to learn latent representations, one with a two-dimensional (2D) latent space and one with a three-dimensional (3D) latent space. Second, for each set of latent representations, we applied k-means clustering, varying the number of clusters k from 2 to 9. Third, we selected the optimal number of clusters using the elbow criterion based on the L_2 loss. Finally, we assessed whether pairs of clusters differed significantly in subject measures, including PANSS positive and negative scores, CMINDS cognitive scores, age, and gender.

Based on the elbow criterion, four clusters were identified for the 2D representations and six clusters for the 3D representations (Figure J.1). Figures J.2 and J.5 show the resulting subgroup (cluster) assignments and the corresponding average sFNC pattern for each subgroup within each diagnostic group. Figures J.3 and J.6 illustrate the corresponding p-values for subject measures between subgroup pairs. Figures J.4 and J.7 then display distributions of subject measures across subgroups. Statistical relationships among subgroup pairs derived from the 2D latent space are largely consistent with those reported in Table 2. For both 2D and 3D representations, fifteen subgroup pairs exhibited statistically significant differences ($p < 0.05$, Mann–Whitney U test), suggesting that the higher-dimensional (3D) latent space does not yield additional meaningful subgroups compared with the 2D latent space.

Furthermore, we examined FNC patterns across four subgroups derived from 2D latent representations (Figure J.2). Table J.1 summarizes pairwise subgroup differences in subject measures and functional network connectivity. Subgroups 1 and 2 differed significantly in age and attention scores ($p < 0.05$, Mann–Whitney U test), with associated FNC changes primarily involving sensorimotor and cognitive control networks, including SC-SM, AU-CC, SM-VI, SM-DM, and CC-CB domains ($p < 0.0001$, Wilcoxon signed-rank test with Bonferroni correction). Subgroups 1 and 3 showed significant differences in age, visual learning scores, and composite scores. Statistically significant differences were observed in both within-network alterations (SM, VI) and between-network interactions. Subgroups 1 and 4 differed significantly in visual learning scores, with FNC differences dominated by visual network connectivity (VI-VI, VI-CB, VI-SM, VI-SC) and subcortical, sensorimotor, and cerebellar interactions (SC-SM, SC-CB, SM-CB). Subgroups 2 and 3 were significantly different in visual learning and reasoning scores, with FNC differences spanning within-network connectivity (SM, VI, and CC) and between-network interactions involving subcortical, sensorimotor, visual, and cerebellar-related connections. Subgroups 2 and 4 differed in working memory and visual learning scores, particularly among subcortical, sensorimotor, visual, and

cerebellar domains. Subgroups 3 and 4 exhibited multiple cognitive score differences, including speed of processing, working memory, verbal learning, visual learning, and composite scores, with FNC differences mainly involving sensorimotor and visual networks. Table J.2 summarizes pairwise subgroup differences in subject measures and functional network connectivity patterns using 3D latent representations.

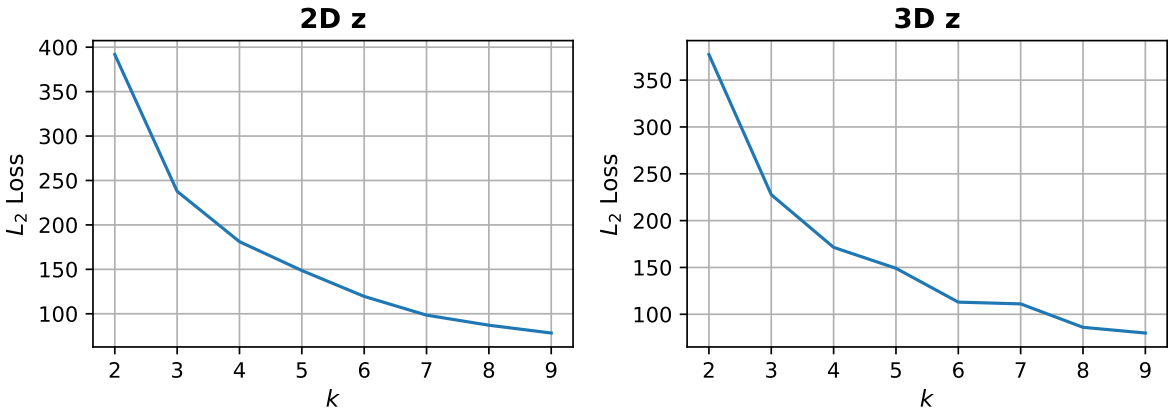


Figure J.1: **k-means clustering L_2 loss with respect to different k values (2D vs 3D).** According to the elbow criterion, four clusters were identified for the 2D representations and six clusters for the 3D representations.

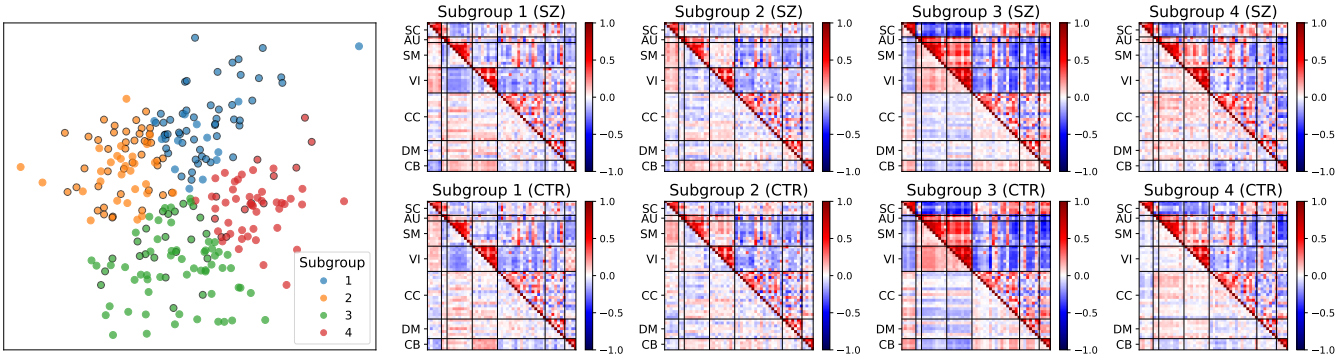


Figure J.2: **Subgroups based on 2D latent representations.** In the scatter plot, SZ patients are highlighted in black circles, while controls are shown in their original colors. The average sFNC pattern is shown for each of the four k-means subgroups (clusters) and for each of the two diagnostic groups (SZ patients and controls). Subgroups were sorted by the increasing proportion of controls among all subjects.

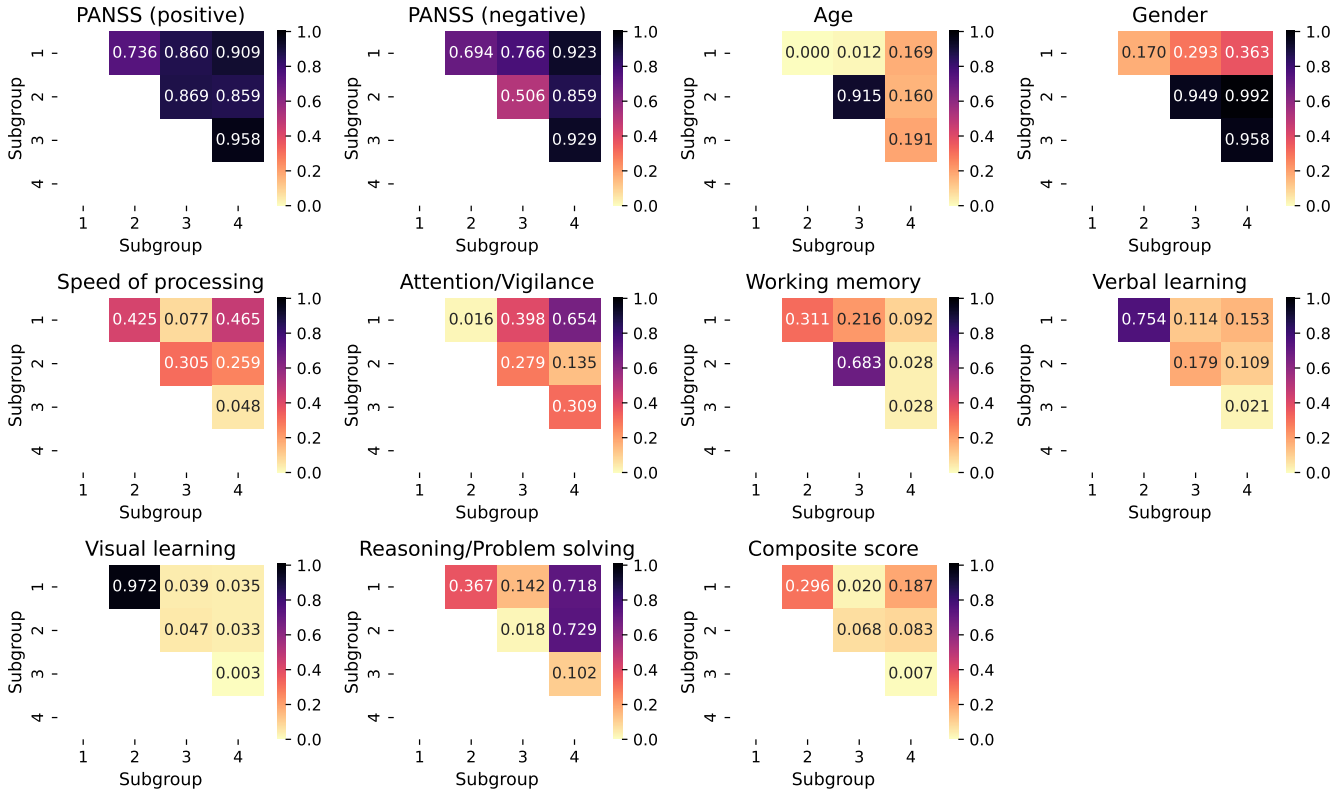


Figure J.3: **Statistical significance of each score type between two SZ subgroups (2D latent representations).** The heatmaps show p-values for each score type between all pairs of subgroups (Mann–Whitney U test). Fifteen cluster pairs exhibit statistically significant differences ($p < 0.05$).

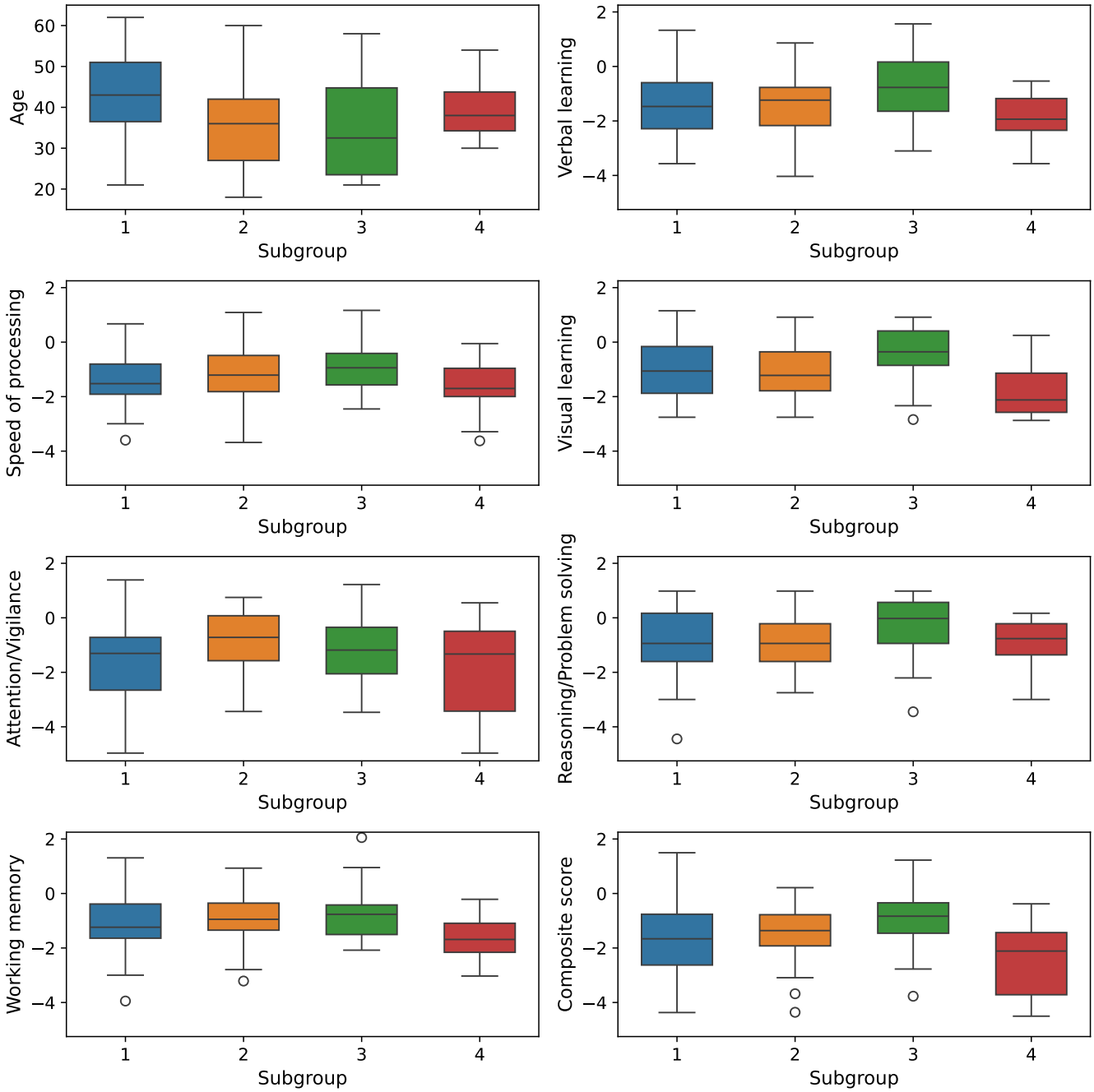


Figure J.4: **Subject measure distributions for SZ subgroups (2D latent representations).** Box plots display subject measure (age, cognitive score) distributions (quartiles) for SZ subgroups.

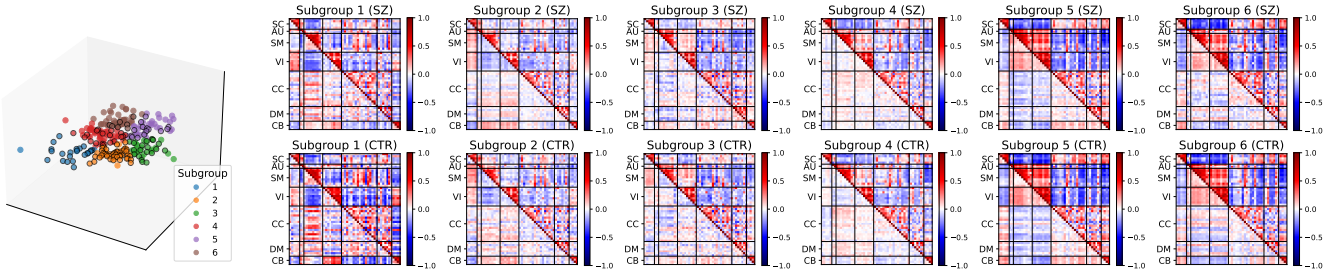


Figure J.5: **Subgroups based on 3D latent representations.** In the scatter plot, SZ patients are highlighted in black circles, while controls are shown in their original colors. The average sFNC pattern is shown for each of the four k-means subgroups (clusters) and for each of the two diagnostic groups (SZ patients and controls). Subgroups were sorted by the increasing proportion of controls among all subjects.

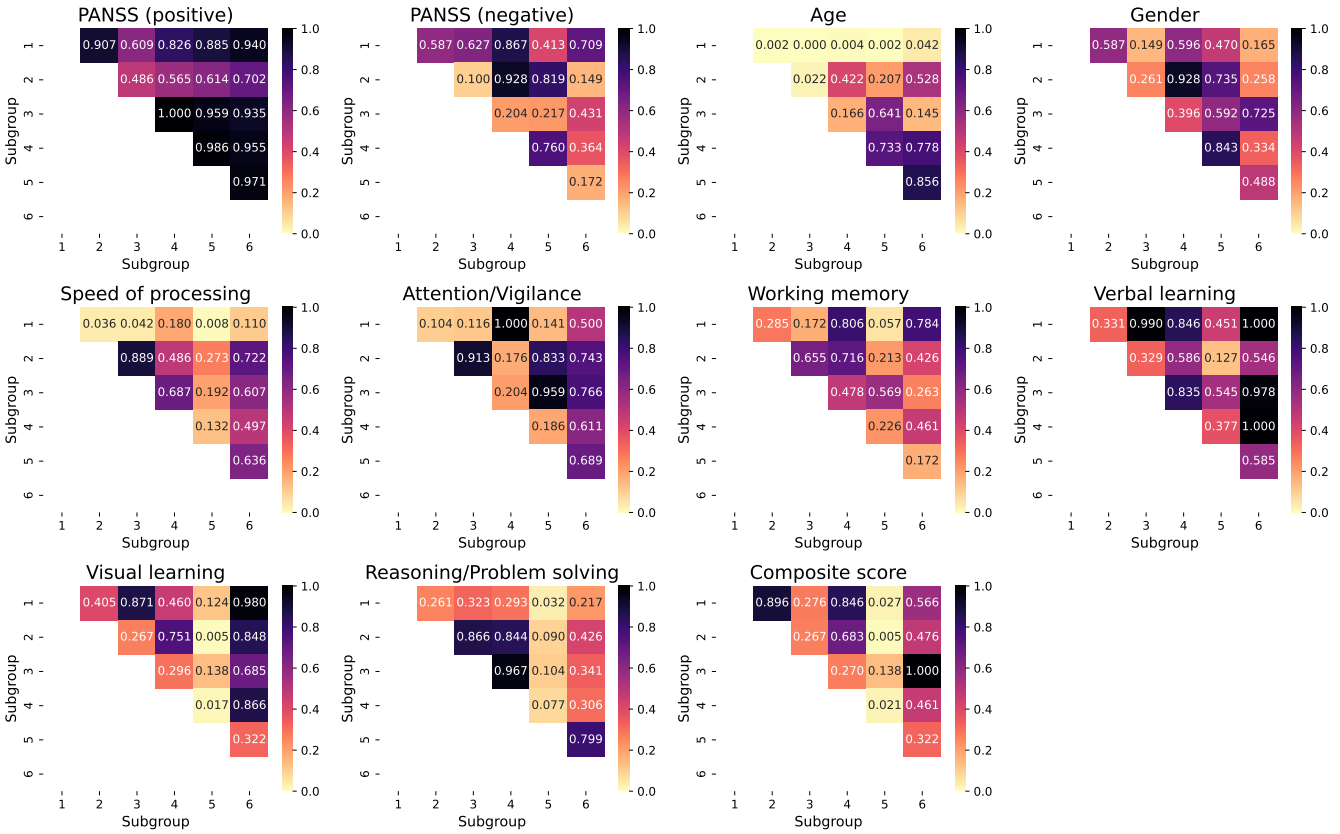


Figure J.6: **Statistical significance of each score type between two subgroups of 3D latent representations in the patient group.** The heatmaps show p-values for each score type between all pairs of subgroups (Mann–Whitney U test). Fifteen cluster pairs exhibit statistically significant differences ($p < 0.05$).

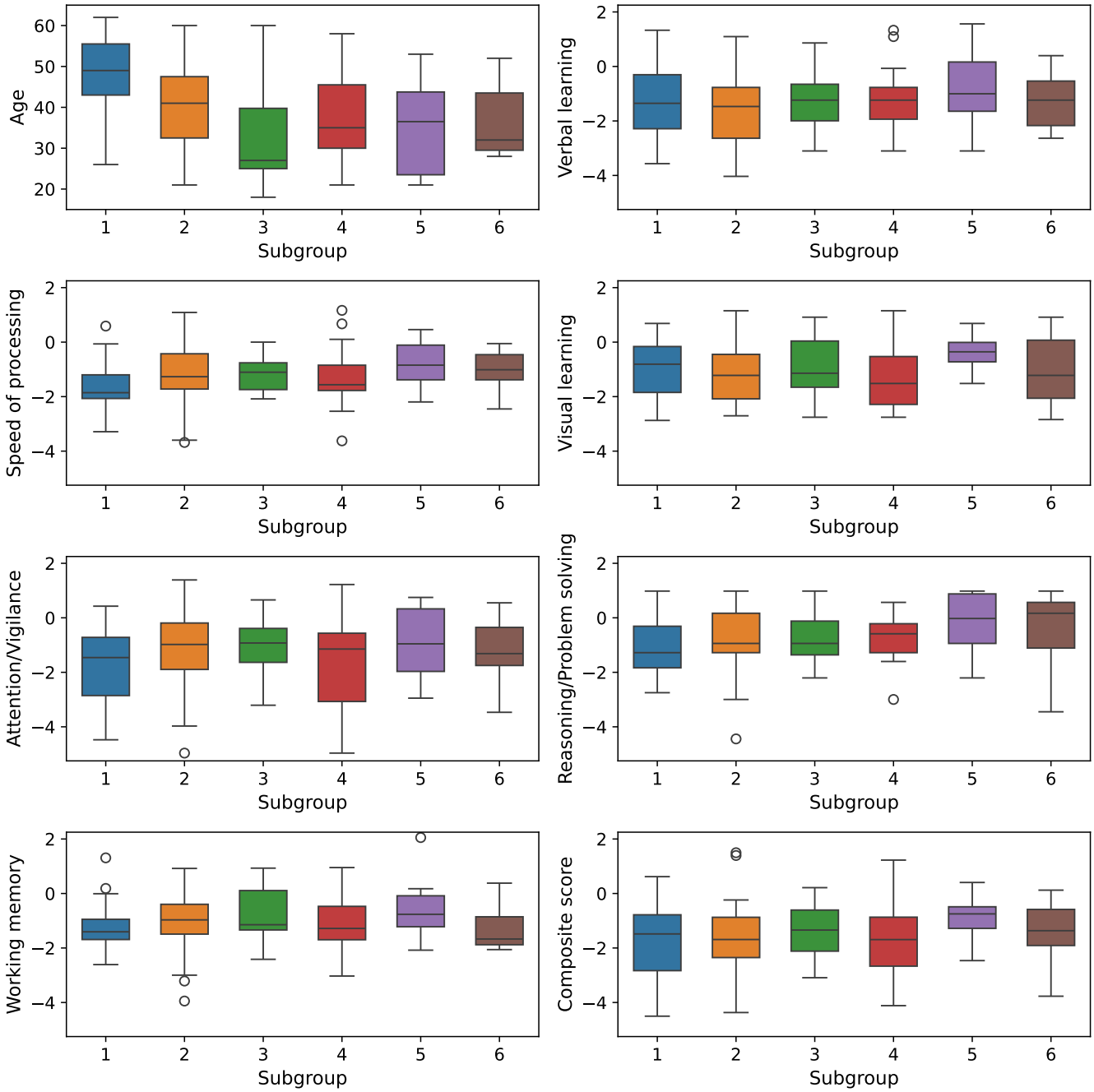


Figure J.7: **Subject measure distributions for SZ subgroups (3D latent representations).** Box plots display subject measure (age, cognitive score) distributions (quartiles) for SZ subgroups.

Table J.1: **Pairwise subgroup differences in subject measures and functional network connectivity (2D latent representations).** Each row lists a subgroup pair, significantly different subject measures ($p < 0.05$, Mann–Whitney U test), and corresponding functional network connectivity patterns ($p < 0.05$, Wilcoxon signed-rank test with Bonferroni correction).

Subgroup pairs	Subject measures	Functional network connectivity
1, 2	Age Attention	SC-SM, AU-CC, SM-VI, SM-DM, CC-CB ($p < 0.0001$); AU-SM, SM-CC ($p < 0.001$); SC-CB, VI-CB, DM-CB ($p < 0.01$); SC-VI, AU-DM, CB-CB ($p < 0.05$)
1, 3	Age Visual learning Composite score	SC-SM, SC-VI, SC-CB, SM-SM, SM-VI, SM-CC, SM-DM, SM-CB, VI-VI, VI-CC, VI-CB, CC-CB, DM-CB ($p < 0.0001$); AU-SM, AU-VI, AU-CC ($p < 0.001$); SC-SC ($p < 0.01$); SC-AU, SC-CC, CC-CC ($p < 0.05$)
1, 4	Visual learning	SC-SM, SC-VI, SC-CB, SM-VI, SM-CB, VI-VI, VI-CB ($p < 0.0001$); CC-CC, CB-CB ($p < 0.05$)
2, 3	Visual learning Reasoning	SC-SM, SC-VI, SC-CB, SM-SM, SM-VI, SM-CB, VI-VI, VI-CC, VI-CB, CC-CC ($p < 0.0001$); SC-CC, AU-VI ($p < 0.001$); SC-SC, AU-SM, SM-CC, CC-CB ($p < 0.01$); SC-AU, CB-CB ($p < 0.05$)
2, 4	Working memory Visual learning	SC-SM, SC-VI, SM-VI, SM-DM, SM-CB, VI-VI, VI-CB, CC-CB ($p < 0.0001$); AU-CC, SM-CC ($p < 0.001$); SC-CB, SM-SM, CC-CC, DM-DM ($p < 0.01$); SC-SC, AU-DM, CB-CB ($p < 0.05$)
3, 4	Speed of processing Working memory Verbal learning Visual learning Composite score	SC-SM, SC-VI, SC-CB, AU-CC, SM-SM, SM-VI, SM-CC, SM-DM, SM-CB, VI-VI, VI-CC, VI-CB, CC-CB, DM-CB ($p < 0.0001$); SC-CC, AU-SM, AU-VI, DM-DM ($p < 0.001$); AU-DM ($p < 0.01$); SC-SC, SC-AU, VI-DM, CB-CB ($p < 0.05$)

Table J.2: **Pairwise subgroup differences in subject measures and functional network connectivity (3D latent representations).** Each row lists a subgroup pair, significantly different subject measures ($p < 0.05$, Mann–Whitney U test), and corresponding functional network connectivity patterns ($p < 0.05$, Wilcoxon signed-rank test with Bonferroni correction).

Subgroup pairs	Subject measures	Functional network connectivity
1, 2	Age Speed of processing	SC-SM, SC-VI, AU-CC, SM-VI, SM-CB, CC-CB ($p < 0.0001$); AU-SM, VI-VI ($p < 0.001$); SM-CC, SM-DM ($p < 0.01$); SM-SM, CB-CB ($p < 0.05$)
1, 3	Age Speed of processing	SC-SM, SC-VI, AU-CC, SM-VI, SM-DM, VI-CB, CC-CB ($p < 0.0001$); AU-SM, SM-CC ($p < 0.001$); SC-CB, DM-DM, DM-CB ($p < 0.01$); AU-DM, CB-CB ($p < 0.05$)
1, 4	Age	SC-VI, SM-VI, SM-CB, VI-VI, VI-CB, CC-CB ($p < 0.0001$); AU-SM, DM-CB ($p < 0.001$); SC-SC, SC-CB ($p < 0.01$); AU-VI, SM-DM, CB-CB ($p < 0.05$)
1, 5	Age Speed of processing Reasoning Composite score	SC-VI, SC-CB, AU-CC, SM-VI, SM-CC, SM-DM, SM-CB, VI-VI, VI-CC, VI-CB, CC-CB, DM-CB ($p < 0.0001$); AU-SM, AU-VI ($p < 0.001$); SC-SC, SC-SM ($p < 0.01$); SC-AU, CB-CB ($p < 0.05$)
1, 6	Age	SC-SM, SC-VI, SC-CB, SM-VI, SM-CC, SM-CB, VI-VI, VI-CB, CC-CB, DM-CB ($p < 0.0001$); AU-SM, AU-VI ($p < 0.001$); SC-SC, VI-CC ($p < 0.01$)
2, 3	Age	SC-VI, SC-CB, SM-VI, SM-DM, SM-CB, VI-VI, VI-DM, VI-CB, CC-CB, DM-DM, DM-CB ($p < 0.0001$); AU-SM, AU-VI, SM-CC ($p < 0.001$); SC-SC, SC-DM, AU-DM ($p < 0.01$); SC-AU, SC-SM, CB-CB ($p < 0.05$)
2, 5	Visual learning Composite score	SC-SM, SC-VI, SC-CB, SM-SM, SM-VI, SM-CC, SM-CB, VI-VI, VI-CC, VI-CB, CC-CC, CC-CB ($p < 0.0001$); SC-CC, AU-SM, AU-VI, AU-CC, DM-CB ($p < 0.001$); SC-SC, SM-DM ($p < 0.01$); SC-AU, SC-DM, CB-CB ($p < 0.05$)
4, 5	Visual learning Composite score	SC-SM, SC-VI, SC-CC, SC-CB, AU-CC, SM-SM, SM-VI, SM-CC, SM-DM, SM-CB, VI-VI, VI-CC, VI-CB, CC-CB ($p < 0.0001$); AU-SM, AU-VI, DM-CB ($p < 0.001$); AU-DM ($p < 0.01$); SC-AU, SC-DM, VI-DM, DM-DM ($p < 0.05$)

Appendix K sFNC interpolation along cognitive score trajectories

To examine how sFNC patterns vary across cognitive performance levels, we divided subjects in the FBIRN dataset into two groups based on the median cognitive score: a lower-score group (scores $\leq$ median) and a higher-score group (scores $>$ median). Within each group, we derived a cluster in the latent space according to the distribution of cognitive scores. For each pair of clusters (higher- and lower-score groups), we constructed a trajectory connecting their centroids in the latent space. Finally, we generated one representative sFNC matrix from each cluster centroid to characterize connectivity patterns associated with different cognitive levels. Most cognitive scores show moderate to strong correlations (Figure K.1), therefore, the resulting clusters and trajectories are largely similar across different cognitive measures (Figure K.2).

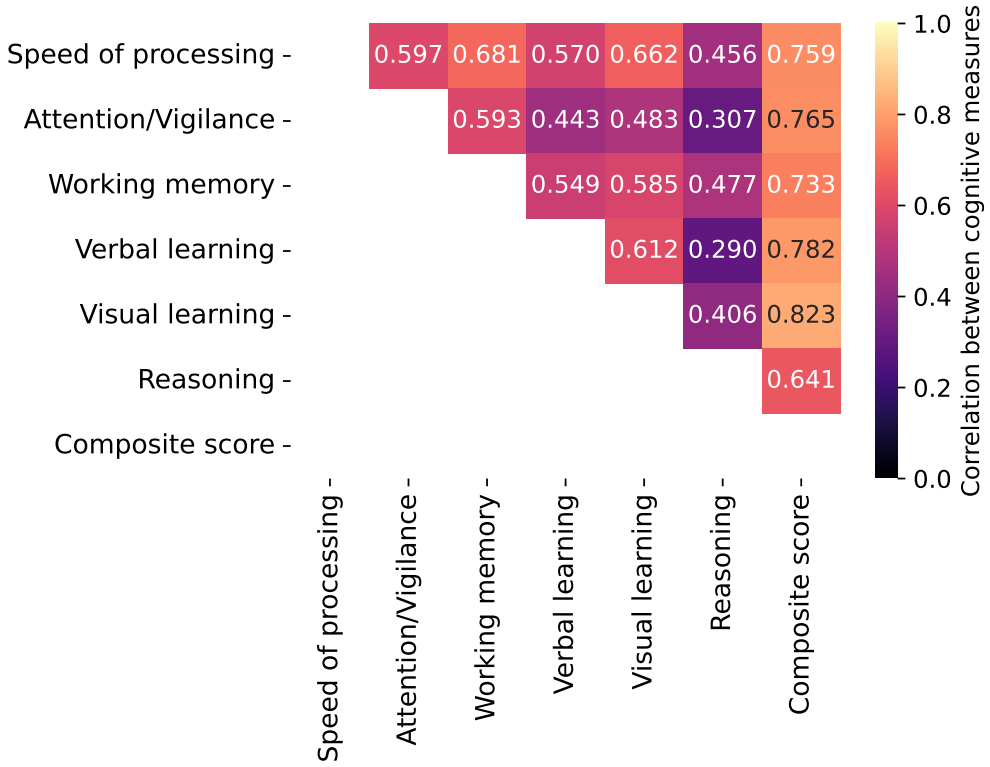


Figure K.1: **Correlations between cognitive measures.** Cognitive scores exhibit varying degrees of similarity, with correlation coefficients ranging from 0.290 to 0.823. The composite score shows strong correlations with the other scores.

As shown in Figure K.2, clusters corresponding to different cognitive score groups occupy highly consistent spatial locations in the latent space, with the lower-score cluster located in the upper-left region and the higher-score cluster in the lower-right region. For each cluster, we generated one representative sFNC

matrix by sampling from the cluster centroid. For each group, we also computed the average sFNC matrix across the original sFNC matrices. We demonstrate that, for each group, the generated sFNC matrices are highly similar to the corresponding original sFNC matrices (Figure K.3).

Across the two sampled states, we observe systematic and consistent changes in sFNC patterns associated with cognitive performance. From group 1 (with higher scores) to group 2 (with lower scores), negative correlations are significantly attenuated in the SC–SM, SC–VI, and SM–CB domains ($p < 0.0001$, Wilcoxon signed-rank test with Bonferroni correction). We also observe a transition from positive to negative correlations in the SC–CB and SM–VI domains ($p < 0.0001$), as well as a transition from negative to positive correlations in the VI–CB domain ($p < 0.0001$) and the SC–AU domain ($p < 0.05$). In addition, positive correlations within the SM, VI, and CC domains are significantly reduced ($p < 0.0001$), as are correlations within the SC domain and the AU–SM domain ($p < 0.01$), and within the CB domain ($p < 0.05$).

Several connectivity changes are specific to individual cognitive domains. For attention ($p < 0.05$), working memory ($p < 0.05$), and reasoning ($p < 0.01$), we observe a shift from positive to negative correlations in the AU–VI domain. Moreover, working memory ($p < 0.001$) and reasoning ($p < 0.0001$) scores are associated with significantly weakened negative correlations in the VI–CC domain.

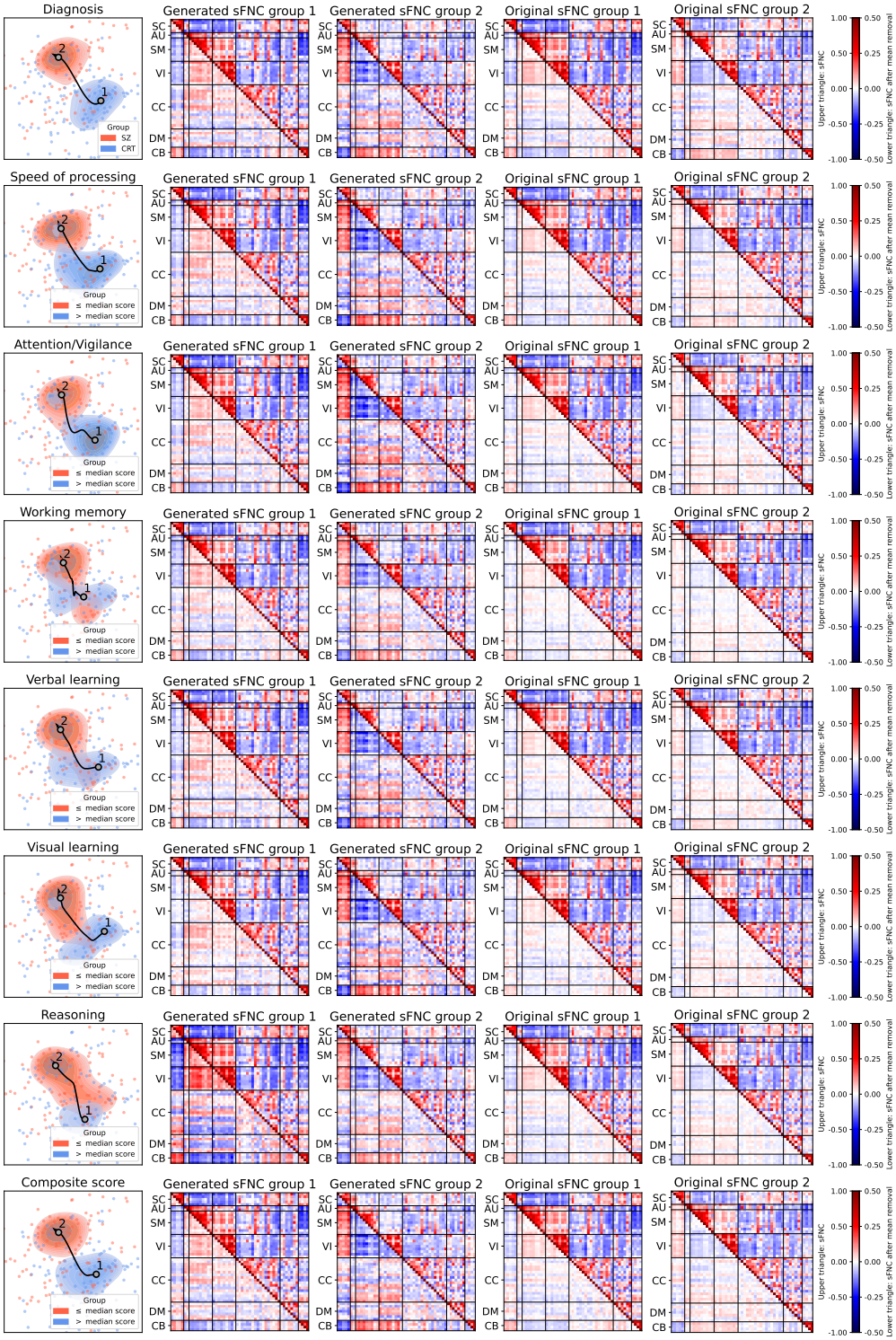


Figure K.2: **Interpolation along a trajectory derived from two clusters defined by cognitive scores.** In each row, the scatter plot shows latent features colored by cognitive scores (red: $\leq$ median score; blue: $>$ median score), overlaid with contour plots of two clusters connected by a trajectory. For each group, we generated one representative sFNC matrix by sampling from the cluster centroid. In addition, we computed the average sFNC matrix across the original sFNC matrices.

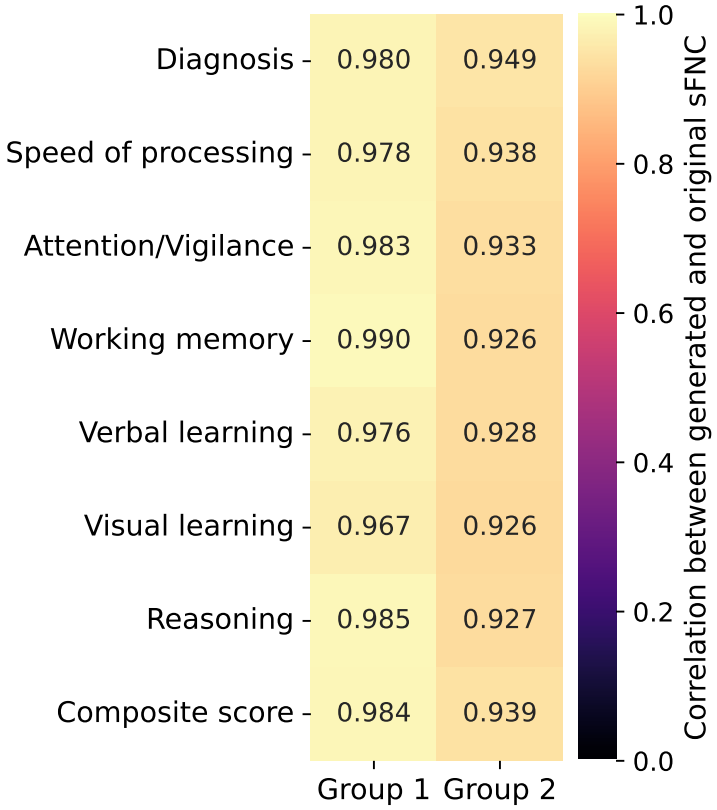


Figure K.3: **Correlations between generated and original sFNC.** For each group, the generated sFNC matrices are highly similar to the corresponding original sFNC matrices.

Appendix L Similarity measured by mean squared error

Mean squared error (MSE) was used to measure the similarity between the generated and original sFNC matrices. The median MSEs for the FBIRN training set, FBIRN test set, ABIDE I training set, and ABIDE I test set are 0.031, 0.034, 0.029, and 0.034, respectively. These results highlight that the generated sFNC matrices are highly similar to the original ones, as measured by MSE.

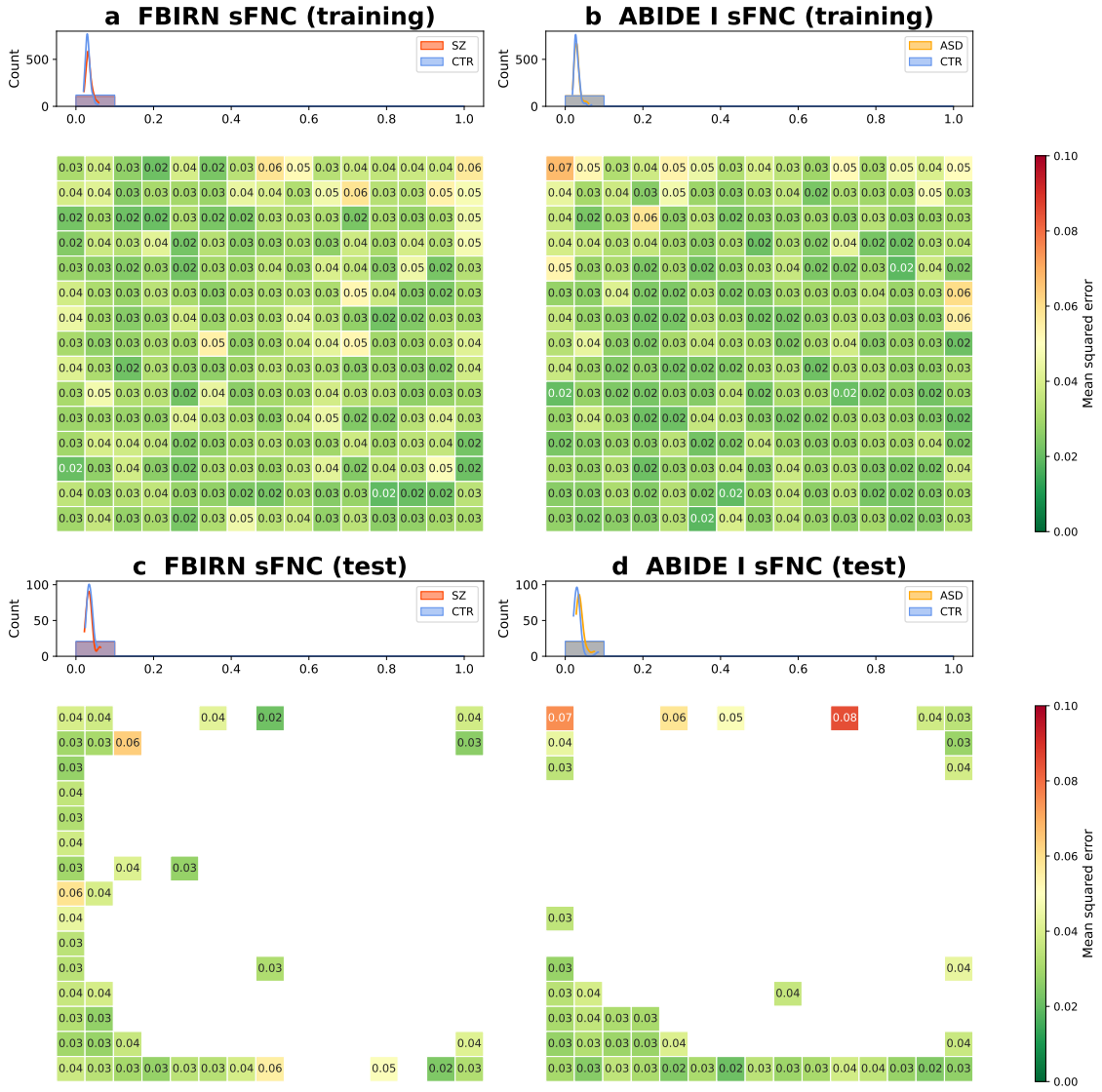


Figure L.1: **Mean squared errors between generated and original sFNC matrices.** (a) FBIRN training set. (b) ABIDE I training set. (c) FBIRN test set. (d) ABIDE I test set. Each panel shows mean squared errors (MSEs) between individual generated and original sFNC matrices on the 2D grid and the corresponding group-specific histogram distributions. Low MSEs across all panels suggest a high degree of correspondence between generated and original sFNC matrices.

Appendix M k-means clustering

The k-means clustering algorithm was separately applied to generated or original dFNC data in each dataset to identify dynamic states by searching the number of states k from 2 to 9. The optimal number of states was then identified using the elbow criterion based on the L_2 loss, which measures the Euclidean distance between each sample and its assigned centroid. As shown in Figure M.1, the elbow point occurred at $k = 5$ in most cases (Figure M.1a,b,c). Thus, an optimal value $k = 5$ was used for both FBIRN and ABIDE I.

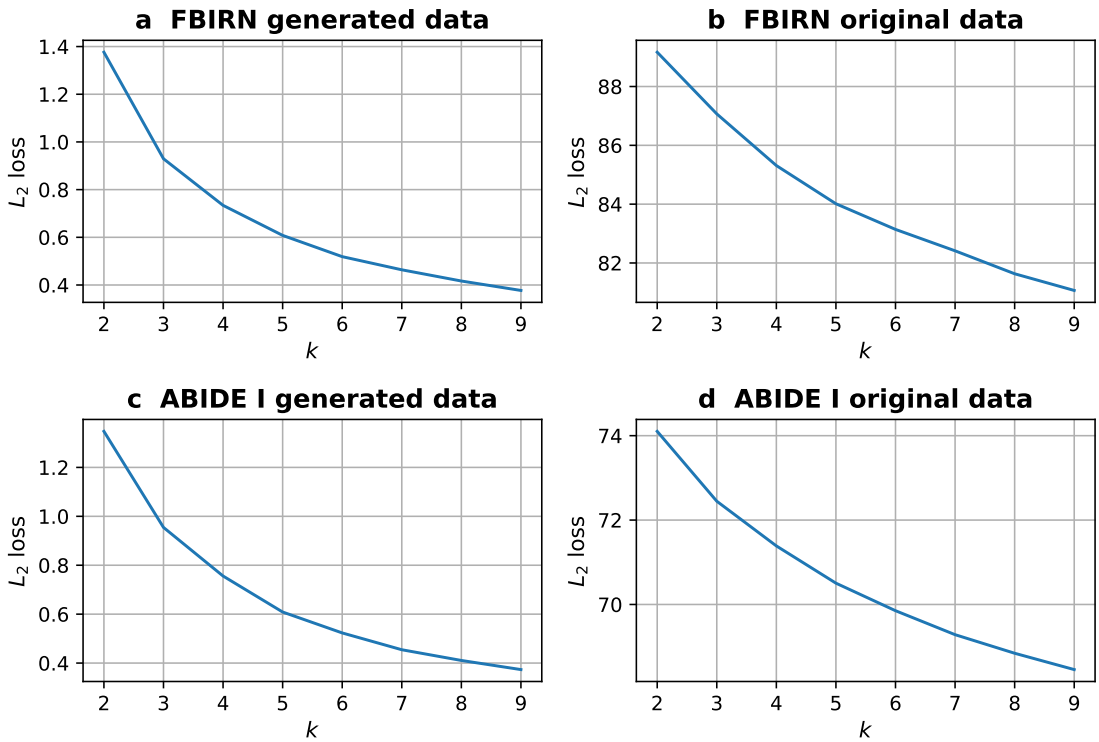


Figure M.1: **k-means clustering L_2 loss with respect to different k values.** (a) FBIRN generated data. (b) FBIRN original data. (c) ABIDE I generated data. (d) ABIDE I original data.