BRAINS: Building Representations with Autoencoders for Individualized Neuroimaging Spaces

Kajal Singla[1,2], **Dr. Pierre-Louis Bazin**[3], **Dr. Nico Scherf**[1, 2]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig/Dresden, Germany
[3]Full brain picture Analytics, Leiden, Netherlands
singla@cbs.mpg.de, piloubazin@fullbrainpicture.nl, nscherf@cbs.mpg.de

Abstract

Functional Magnetic Resonance Imaging (fMRI) provides rich, high-dimensional data on human brain activity, yet traditional dimensionality-reduction techniques primarily capture group-level structure and overlook individual variability. We introduce BRAINS, a framework based on Convolutional Variational Autoencoders (CVAEs) that learns subject-specific latent spaces directly from BOLD signals. These latent representations effectively denoise voxel-wise time series ($\sim$ 5 % tSNR gain) while preserving functional connectivity and anatomical coherence. Using Procrustes alignment, we show that individual latent spaces can be aligned across participants, revealing both shared and idiosyncratic components of cortical organization. Our approach bridges neuroimaging and deep representation learning, offering a geometry-aware foundation for individualized brain analysis and multimodal integration across subjects, tasks, and models. The code is available at: https://github.com/neural-data-science-lab/NEUROAI_AAAI_BRAINS.git.

Introduction

Functional Magnetic Resonance Imaging (fMRI) is one of the most widely used human neuroimaging techniques, indirectly measuring brain activity over time by detecting changes in blood flow (Glover 2011). It offers high spatial resolution (up to 1–2 mm) but is limited by its lower temporal resolution (1–3 seconds) compared to techniques like electroencephalography (EEG) and magnetoencephalography (MEG). fMRI data can be collected using three primary paradigms: resting-state fMRI (rs-fMRI), which captures spontaneous brain activity without external stimuli; task-based fMRI (tb-fMRI), which measures brain responses to specific cognitive tasks; and naturalistic fMRI, which records brain activity while participants engage with real-world stimuli, such as watching movies or listening to narratives. Naturalistic paradigms strike a balance between the uncontrolled nature of rs-fMRI and the high experimental control of tb-fMRI, offering better ecological validity while still providing time-locked brain activity for analysis (Chang et al. 2020). This time-locked structure enables advanced analysis techniques like inter-subject correlation (Simony et al. 2016).

Traditional fMRI analysis techniques, such as Independent Component Analysis (ICA) and the General Linear Model (GLM), have been widely used to study brain activity. ICA is typically applied to rs-fMRI to extract functional networks and has also been used as a preprocessing tool for noise removal and signal decomposition (Pruim et al. 2015; Glasser et al. 2018). GLM, on the other hand, is primarily used in tb-fMRI for voxel-wise analysis, assessing how well brain activity aligns with predefined task regressors. While both methods have been successful in identifying functional networks and brain states, they have limitations in capturing non-linear relationships and subject-specific variability.

In recent years, deep learning approaches have been increasingly applied to neuroimaging data, offering powerful tools for capturing complex, non-linear patterns in fMRI signals. Models such as Recurrent Neural Networks (RNNs) (Koppe et al. 2019), Transformers (Li, Wang, and Liu 2022), and Variational Autoencoders (VAEs) (Kim et al. 2021; Han et al. 2019) have shown promising results in studying brain dynamics. These deep generative models excel at learning latent representations, detecting outliers, generating synthetic samples, and estimating the probability distribution of input data (Tomczak 2022).

In this study, we leverage a naturalistic fMRI dataset (Finn et al. 2018), sourced from OpenNeuro (ds001338), where 22 participants listened to a 22-minute ambiguous social narrative designed to evoke varying levels of paranoia. Using minimally preprocessed fMRI data, we explore the capacity of deep latent models, specifically Convolutional Variational Autoencoders (CVAEs) (Wang et al. 2024) to reveal individual-specific patterns in the Blood Oxygen Level Dependent (BOLD) signals. CVAEs are effective for fMRI time series analysis because their 1D convolutional architecture captures temporal patterns and dependencies in BOLD signals while reducing noise through dimensionality reduction, whereas the variational component allows us to model the underlying probability distribution of BOLD signals, accounting for the inherent variability in neural responses across time and subjects. Additionally, by processing the time course for each voxel independently, the model can learn consistent temporal features across the brain while preserving spatial variations that reflect individual-specific functional organization. Unlike traditional group-level parcellations, which impose rigid, predefined bound-

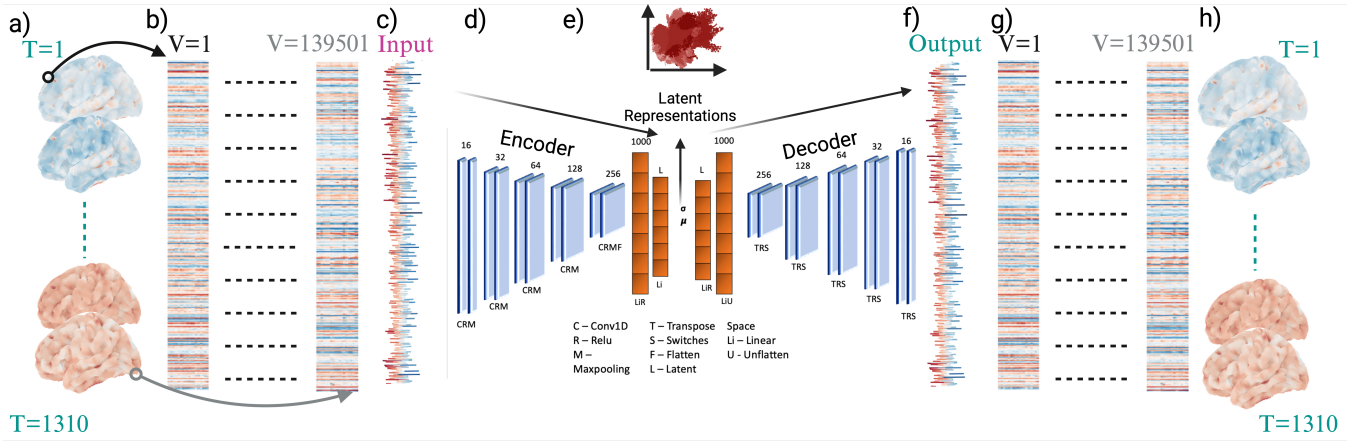

Figure 1: **Overview of the BRAINS pipeline.** Each subject's fMRI time series is encoded into latent coordinates using a convolutional variational autoencoder (CVAE), which learns to reconstruct and denoise BOLD signals while preserving spatial structure. a,b) 4D fMRI data of 22 subjects were converted into a 139501×1310 (voxels $\times$ time points) data matrix. Time series, positions and brain region labels were extracted based on Schaefer atlas 2018 with 200 parcellations of the Yeo 17 networks and MNI152NLin2009cAsym. c) Each time series was normalized (minmax scaling) across time series. Data was divided into batches 280×1310 (voxels $\times$ time points). d) A separate CVAE network was trained for each of the 22 subjects. Both encoder and decoder consist of 8 layers. e) The encoder output (μ) defines the latent coordinates for a given input BOLD signal. f) The decoder (re-)generate the input data. We train the network using an L2 reconstruction loss (mean squared error). g) The trained model learns to embed and reconstruct input BOLD signals that h) can then be mapped back to the brain coordinates.

aries, CVAE may provide a simplified latent space for functional alignment, either in comparison or in conjunction with hyper-alignment (Nastase et al. 2020) and shared response modeling (Chen et al. 2015).

Our main contributions are:

- A deep latent-space framework based on Convolutional Variational Autoencoders (CVAEs) for **individualized fMRI representations**.
- Demonstration that CVAEs **denoise BOLD signals without distorting functional connectivity**, yielding an average $\sim$5 % tSNR improvement.
- A procedure to **align subject-specific latent spaces** using Procrustes analysis, revealing shared and individual cortical components.
- A unified perspective connecting **representation learning and functional neuroimaging**, enabling geometry-aware comparison across subjects.

Together, these contributions advance individualized and multimodal analysis of brain activity, bridging deep generative modeling and cognitive neuroscience.

Methods

Dataset

We re-analyzed a naturalistic fMRI dataset from OpenNeuro (https://openneuro.org/datasets/ds001338/versions/00002) featuring 22 healthy participants (11 females; age range: 19–35 years) (Finn et al. 2018). During the experiment, participants listened to an original 22-minute audio narrative depicting an ambiguous social scenario designed to elicit varying interpretations and emotional responses. The narrative was divided into three continuous segments, lasting 8:46, 7:32, and 5:32 minutes, respectively, and was presented as a single uninterrupted functional run.

This dataset is particularly suitable for deep learning analyses due to its naturalistic design, which captures ecologically valid neural responses while preserving high temporal synchronization across participants. Leveraging this data, we investigated the ability of deep latent models to uncover subject-specific patterns in BOLD signals and analyze functional connectivity at the individual level.

Preprocessing

We utilized minimally preprocessed fMRI data of the narrative dataset available on OpenNeuro (Finn et al. 2018), as provided by the Naturalistic Data Analysis repository. The preprocessing pipeline primarily involved the use of `fmriprep` for standard fMRI corrections, followed by smoothing and `GLM` denoising.

The `fmriprep` pipeline performed the following steps: 1) Realignment, 2) Spatial normalization to the MNI152NLin2009cAsym template, 3) Motion correction to account for participant movement during scanning, 4) Time slice correction to address acquisition timing differences between slices. After these corrections, spatial smoothing was applied with a full-width at half-maximum (FWHM) of 6 mm. To further denoise the data, `GLM` is used to remove voxel-wise motion parameters and global artifacts.

The preprocessed fMRI data consists of 4D whole-brain volume activations over time (Figure 1a). We flattened this 4D data into a 2D matrix (voxels $\times$ time points) for each subject. Specifically, we used the Schaefer atlas (2018) with 200 parcellations mapped to Yeo 17 networks (Schaefer et al.

2018) and applied the MNI152NLin2009cAsym brain mask to extract relevant voxels and their associated time series. This resulted in a 139,501 (voxels) × 1,310 (time points) matrix for each participant.

To standardize the input for the deep learning model, we normalized the BOLD time series to a range of [0,1] using the MinMaxScaler from `sklearn`. Additionally, we extracted the spatial coordinates of each voxel to maintain a mapping between voxel IDs and their brain locations (Figure 1b), which facilitated visualization and analysis in later stages. Notably, the deep learning architecture was designed to learn functional connectivity patterns from the data, without relying on predefined spatial or temporal priors. This design choice ensured that connectivity emerges organically through the model's latent representations, rather than being imposed through explicit structural assumptions.

Model Architecture and Training

To model the latent structure of fMRI data, we employed a Variational Autoencoder (VAE) framework (Kingma, Welling et al. 2019). We used a 1D Convolutional Variational Autoencoder (CVAE) to process the temporal BOLD signals (Figure 1d). Both the encoder and decoder networks consist of five convolutional layers, 1 flatten and two fully connected (FC) layers. The data for each subject was randomly split into training (70%), validation (10%), and test (20%) sets. We trained the models using mini-batches, where each batch comprised 280 voxels × 1,310 time points (Figure 1c).

Encoder Architecture The encoder processes the 1D input time series through five successive Conv1d layers with increasing channel sizes: 16, 32, 64, 128, and 256. The output from the final convolutional layer is flattened and passed through two fully connected layers to produce the parameters (mean and variance) of a Gaussian variational approximation to the latent space (Figure 1e). Using the reparameterization trick, we sample from this latent distribution and feed the sample into the decoder.

Decoder Architecture The decoder mirrors the encoder structure, starting with two fully connected layers followed by five ConvTranspose1d layers with decreasing channels (256, 128, 64, 32, 16) to reconstruct the input BOLD signals (Figure 1f).

All layers used small kernels (3×1), ReLU activations, and pooling/upsampling for temporal compression.

Loss Function The standard VAE loss function combines a Reconstruction Loss (Mean Squared Error (MSE) between the input and reconstructed data) and a KL Divergence (Measures the difference between the learned latent distribution and a standard multivariate normal distribution) (Kingma, Welling et al. 2019). However, for this study, we set the KL divergence term to zero for two main reasons: (i) Training Stability: KL divergence often requires annealing strategies to ensure stable training (Joas et al. 2024). (ii) Focus on Latent Structure: Our primary goal was not to train a generative model but to analyze the latent embeddings produced by the encoder.

Training Procedure We trained a separate CVAE model for each of 22 subjects to capture subject-specific latent representations. The models were implemented using Lightning PyTorch and trained on a A100 GPU using Adam optimizer with a learning rate of 0.001. Early stopping was applied with patience of 100 epochs to prevent overfitting. The code for the CVAE architecture, training pipeline, and latent representation analyses are available on https://github.com/neural-data-science-lab/NEUROAI_AAAI_BRAINS.git

Latent Dimensionality

To explore the impact of latent space size on reconstruction quality, we trained CVAE for each subject using latent dimensions ranging from 2 to 16. We evaluated the reconstruction performance on the test set for each configuration, measuring the MSE between the input and reconstructed BOLD signals (Figure 1g,h).

We observed that reconstruction accuracy improved with increasing latent dimensionality, but gains plateaued beyond a certain point (as shown in Supplementary Figure A.1). Notably, we observed several outliers during training, where certain subjects showed elevated test loss at specific latent dimensionalities. This variability is likely due to random weight initializations at the start of training. Based on the trade-off between reconstruction fidelity and model complexity, we selected a latent dimension of 9 for subsequent analyses. This choice provided a balance between capturing sufficient variance in the data while avoiding overfitting or introducing unnecessary complexity.

Results

Deep autoencoders for BOLD signal denoising

To further assess the denoising performance of CVAE, we analyzed reconstructed BOLD signals from Subject 1. A subset of 2,000 randomly selected voxels was used for visualization and error analysis. Figure 2a presents heatmaps comparing the input (left) and reconstructed (right) time series, along with a close-up of two randomly chosen voxels (Figure 2a, bottom). The average MSE across the sample was 0.0011, indicating that the CVAE effectively captured temporal dynamics of BOLD signals while reducing noise. We conducted a cross-subject analysis for validation, where second subject's data was passed through the pre-trained model of the first subject. The MSE error between the input and the decoded data was 0.019, about an order of magnitude higher than the reference MSE error of 0.0011 when using subject 1 data for training and reconstruction. We also performed a control test where we generated a surrogate, randomized signal by taking the mean voxel-wise time series for a subject and added Gaussian noise (std = 2), followed by smoothing with a Gaussian filter (std = 1). This noisy signal was then passed through the pre-trained model of subject 1 for its reconstruction, resulting in MSE error of 1.20, higher than both within and cross-subject reconstruction errors.

We also examined whole-brain reconstructions at a specific time point ($t = 900$) (Figure 2b), which yielded an

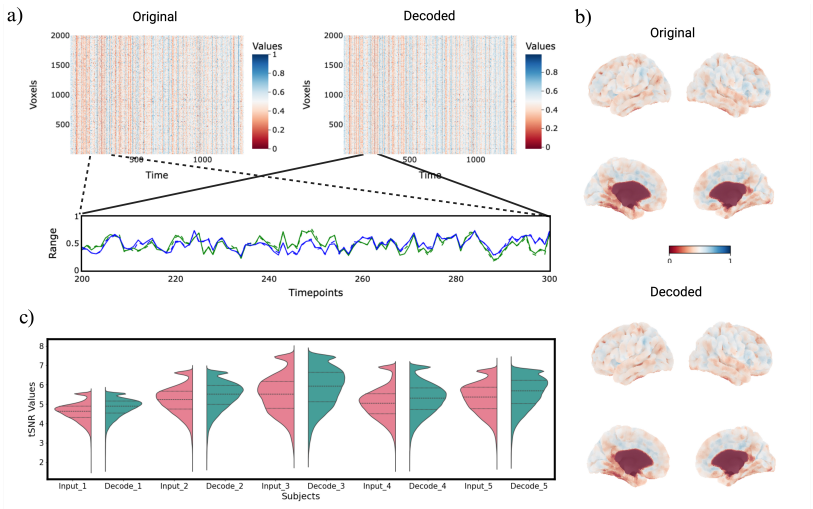

Figure 2: **CVAE denoises voxel-wise BOLD signals, improving tSNR while preserving structure.** a) Heatmaps showing time series for a random subset of voxels for input (left) and decoded data (right). For more detail, input and reconstructed BOLD data are shown for two random voxels (green and blue) as line plots (zoomed in to show only 100 time points for better visibility). b) Input data and reconstruction of signal across whole brain (bottom) for a random time point (t=900, EMSE=0.15%). c) Violin plot of voxel-wise tSNR for input (red) and decoded (green) data for subjects 1-5.

Empirical MSE (EMSE) of 0.15% between the input and reconstructed volumes. This suggests that the reconstruction quality is consistent across both voxels and time points.

To quantitatively assess the denoising capability of the CVAE, we computed the temporal Signal-to-Noise Ratio (tSNR) for each voxel, defined as the mean intensity of the time series divided by its standard deviation. Figure 2c shows the tSNR distributions for Subjects 1–5, comparing the original and reconstructed data. In all cases, the reconstructed BOLD signals demonstrated higher tSNR values, confirming that the CVAE effectively reduces noise while preserving relevant neural signals. We evaluated the significance of tSNR improvements using the Wilcoxon signed rank test, which yielded significant results ($p < 1e{-}308$) in all cases. The average tSNR improvements were 4.98%, 4.83%, 6.46%, 4.61%, and 5.36%, respectively.

These results demonstrate that CVAE can successfully embed, reconstruct, and denoise BOLD time series at voxel level. The learned latent representations not only capture essential temporal dynamics, but also enhance signal quality, providing a robust basis for further analysis of individual brain activity.

The structure of encoded latent representations

We next examined the structure of the subject-specific CVAE embeddings to understand how the latent space captures the underlying fMRI data patterns.

Variance Analysis Across Latent Dimensions We first analyzed the variance explained by each of the nine latent dimensions across all subjects (Figure 3a). The percentage of variance decreases gradually with higher dimensions, suggesting that the latent space forms a homogeneous subspace rather than being dominated by specific directions.

To explore this structure in more depth, we focused on Subject 1 (red line in Figure 3a). Figure 3b presents the mean and standard deviation of the latent coordinates for each of the nine dimensions. Despite training with KL divergence set to zero, the latent dimensions remain compact and centered around zero, indicating inherent regularization through the reconstruction objective alone.

Correlation between dimensions and Spatial Mapping We further assessed the relationships between latent dimensions by calculating pairwise correlations (Figure 3c). The correlations were relatively low (mostly below 0.3), suggesting that the latent dimensions capture generally distinct features, despite the absence of orthogonality constraints. To visualize the spatial patterns encoded in the latent space, we mapped the latent coordinates of each voxel back onto the brain surface for each dimension separately (Figure 3d). These maps revealed smooth or clustered patterns across the cortex, indicating that the CVAE captures anatomically coherent features. Compared to previous studies using diffusion map embedding for cortical gradient analysis (Margulies et al. 2016), our subject-specific latent coordinates suggest plausible anatomic patterns. Compared to diffusion embedding, which typically emphasizes large-scale cortical gradients, our CVAE-based representations appear to reflect more individual-level variations and finer-scale structures.

Dimensionality Reduction and Visualization To investigate whether the latent space reflects known brain organization (e.g. functional clusters or networks), we projected the 9D latent representations and the original high-dimensional inputs into 2D spaces using principal component analysis (PCA), ICA, and t-distributed stochastic neighbor embedding (t-SNE) (Figure 4).

The PCA and ICA visualizations of the latent space show

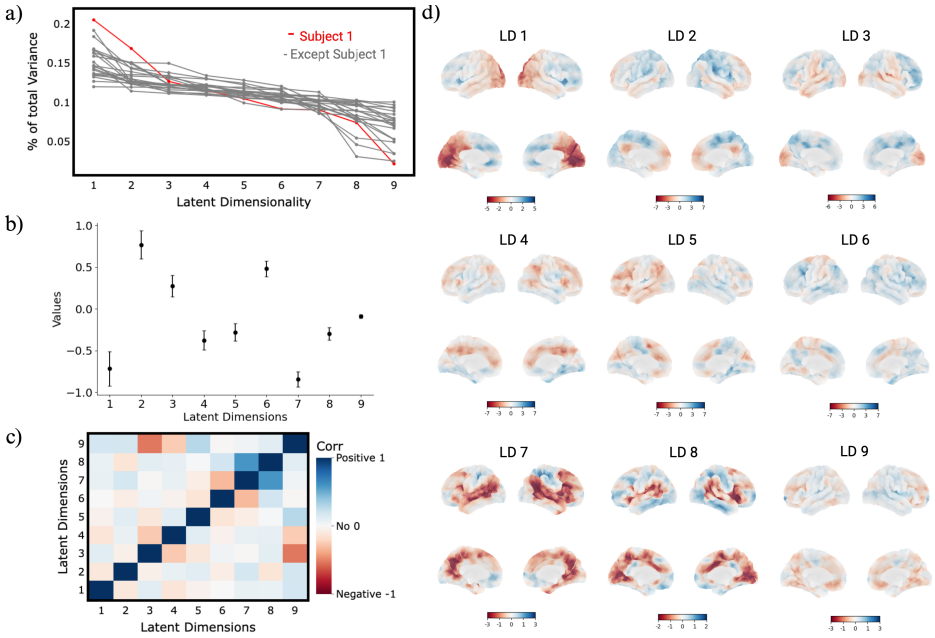

Figure 3: **Latent dimensions capture distinct, spatially coherent features.** a) Percentage of total variance for each of the nine latent dimensions in descending order for all subjects (gray lines) and subject 1 (red). b) Mean (dot) and standard deviation (line) of latent coordinates for subject 1. c) Pairwise correlation between latent coordinates of different latent dimensions of subject 1. d) Latent coordinate of each voxel along a specific latent dimension shown on the brain map. Each plot separately shows one of the nine latent dimensions.

a more uniform distribution of voxels with respect to the Yeo 17 networks, suggesting that the CVAE reduces noise and captures generalized patterns across networks. However, we did not observe clear clustering based on network labels, indicating that the latent space preserves functional structure without explicitly aligning to predefined network boundaries. t-SNE projections did not reveal clear clusters or discernible patterns in either the input or latent spaces, which is unsurprising given the non-linear and often unpredictable behavior of t-SNE in high-dimensional data visualization. To evaluate clustering performance for Yeo 17 networks and Schaefer 200 parcellation (Schaefer et al. 2018), we calculated Adjusted Purity (ARI) (Hubert and Arabie 1985) and Silhouette scores (Rousseeuw 1987). ARI for the VAE latent was nearly double the input, while Silhouette increased only slightly (Table B.1 & B.2).

Overall, CVAE encodes BOLD signals into low-dimensional space that captures smooth, anatomically coherent patterns while maintaining minimal correlations between dimensions. Although latent dimensions do not directly map onto known functional networks, they appear to retain aspects of brain organization at subject level, potentially providing more individualized view of brain activity. This suggests that CVAE captures latent structure that reflects individual neurofunctional architecture beyond standard parcellation schemes, with potential utility for individualized neuroimaging analysis.

Aligning latent representations across subjects

Our CVAE representations are inherently subject-specific, since each model is trained independently. Previous work by (Moschella et al. 2022) demonstrated that even retraining simple autoencoders on the same data with different random initializations leads to misaligned latent spaces. To examine the alignment of latent spaces across subjects, we applied Orthogonal Procrustes (Sasse et al. 2024) analysis to evaluate the consistency of learned embeddings (Figure 5).

The correlations on the anti-diagonal in Figure 5b (see also Figure C.1) show that dimensions 5-8 vary more between subjects, while 1-4 are more consistently shared. The latent dimensions that did not align well between the individuals, could be reflecting differences in brain function rather than noise, suggesting that the embeddings capture meaningful, subject-specific information. The latent dimensions that are aligned between subject-specific latent spaces demonstrate that CVAE embeddings also capture shared underlying structure in fMRI data, despite being trained in individual subjects.

For validation, we examined the correlation between the CVAE-derived latent vectors and the first principal gradient from Margulies et al. (2016) (Margulies et al. 2016), a well-established measure of large-scale cortical organization. Across all subjects, the second latent dimension showed a strong negative correlation (mean $r = -0.49 \pm 0.03$) with the first cortical gradient, indicating that the CVAE captures a fundamental axis of brain organization, a transition from

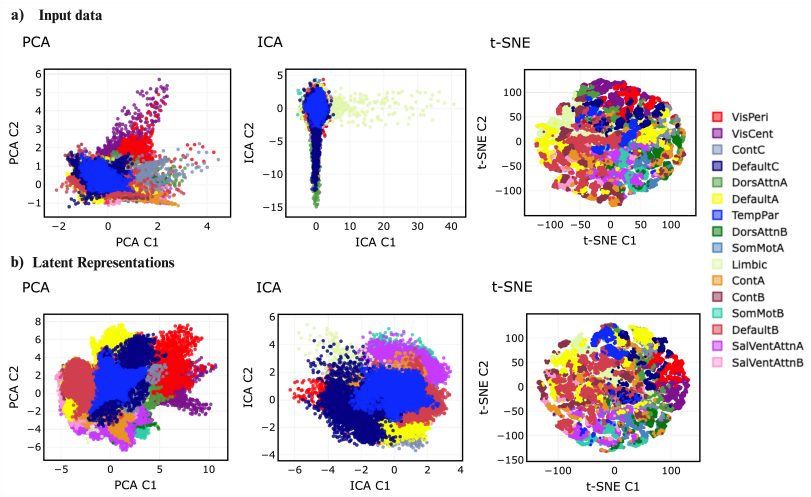

Figure 4: **Distribution of functional brain networks over latent representations.** 2D visualization of a) input data and b) latent coordinates using PCA, ICA and t-SNE, color code indicates Yeo 17 networks. Latent representations show smoother, less noisy network organization compared to input space.

sensory to associative regions consistent with prior findings. This alignment suggests that the learned latent space not only denoises BOLD signals but also encodes biologically meaningful, spatially coherent patterns that reflect known principles of cortical hierarchy.

Together, these findings demonstrate that CVAE latent space simultaneously preserves shared, population-level structure and encodes individual-specific functional variation, providing a powerful, interpretable framework for both group-level analysis and personalized modeling of brain activity.

Deep latent functional connectivity

Functional connectivity refers to the temporal coherence in activation patterns between anatomically distinct brain regions, reflecting their interactions over time (Friston, Frith, and Frackowiak 1993). To compute functional connectivity, voxel-wise BOLD signals are typically averaged within predefined brain regions, and pairwise Pearson correlations are calculated between these region-averaged time series.

We first evaluated whether the CVAE-based denoising process altered the fundamental structure of functional connectivity. Figure 6a displays pairwise Pearson correlations between brain regions, computed from average input data (in 22 subjects, voxels × time points) in the lower triangular matrix, and from CVAE-reconstructed data (averaged in subjects) in the upper triangular matrix. The correlation patterns between the original input and reconstructed data are highly consistent, indicating that the CVAE preserves functional connectivity while effectively denoising the BOLD signals. Quantitatively, the 90th percentile mean correlation for the input data was 0.928, compared to 0.931 for the reconstructed data, reflecting a minor improvement of 0.3%. A paired t-test comparing the 90th percentile correlations between original input and reconstructed data yielded a highly significant result ($p = 6.99 \times 10^{-8}$), confirming that the denoising process leads to a small but meaningful enhancement in signal consistency. This slight increase suggests that the denoising process enhances signal quality without distorting functional relationships.

Given that the CVAE is trained to map similar BOLD signals into nearby regions of the latent space, we hypothesized that the learned latent embeddings could serve as a novel and informative basis for functional connectivity analysis. To test this, we first aligned the latent representations of all 22 subjects on a common reference (Subject 7) using Procrustes alignment and ensures spatial consistency across individuals. After alignment, we computed the average latent representation across subjects to obtain a population-level latent space. Within this averaged latent space, we quantified functional connectivity using two similarity measures: Pearson correlation to assess linear relationships between region-specific latent representations and Euclidean distance to quantify the spatial proximity of latent vectors in the embedding space.

Figure 6b presents these analyses: The lower triangular matrix shows Pearson correlations between brain regions in the averaged latent space. The upper triangular matrix displays Euclidean distances (normalized using MinMaxScaler with a reversed color map) between averaged latent representations. Both similarity measures revealed connectivity patterns consistent with those observed in the input and reconstructed BOLD data (Figure 6a). Notably, the latent space appears to capture more refined and subtle interregional relationships, suggesting that the CVAE enhances the resolution of functional connectivity by preserving fine-grained structural and dynamic relationships that may be obscured by noise or variability in the raw data.

To quantitatively evaluate the correspondence between functional connectivity patterns in the original input and latent spaces, we applied Fisher's r-to-z transformation to Pearson correlation matrices derived from the averaged

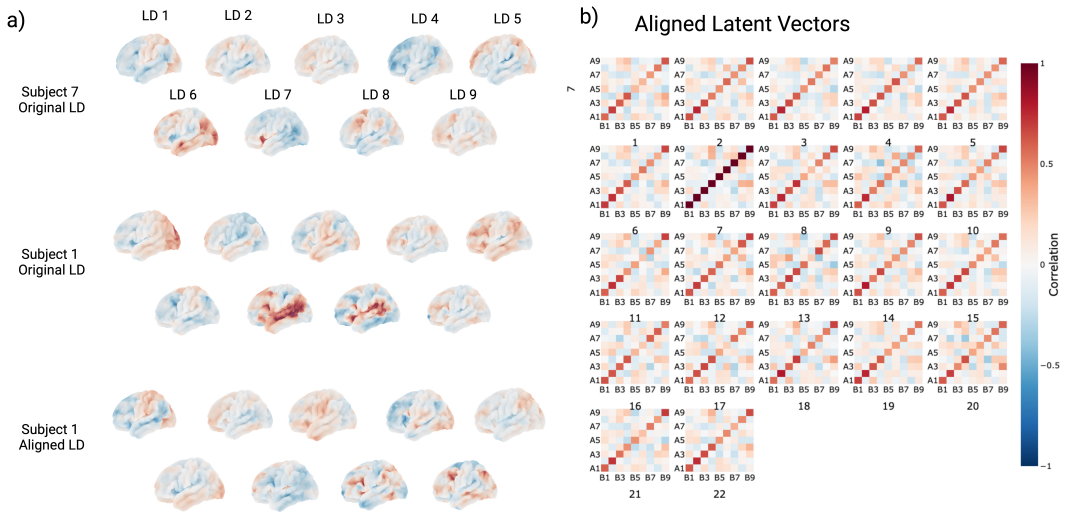

Figure 5: **Aligned latent spaces reveal shared and individual-specific structure across subjects.** a) Latent representations of subject 1 were aligned to the latent representations of subject 7 using Procrustes alignment over corresponding time points. b) Pairwise correlation of latent representations of all subjects after alignment with subject 7. Subject 7 was chosen as an illustration, results are similar across target subjects.

BOLD signals (lower triangle, Figure 6a) and the averaged latent space representations (lower triangle, Figure 6b), both computed across subjects. This transformation stabilized the variance on correlation coefficients, enabling reliable statistical comparison between pairs. A paired Z-test was conducted for each of the 19,100 unique, non-redundant brain region pairs (excluding self-connections and duplicate pairs), yielding a p-value for every pairwise relationship.

After controlling for the false discovery rate (FDR) at 0.05, we identified approximately 19,725 connections that exhibited statistically significant differences between the input and latent spaces, indicating that these relationships were either suppressed or reweighted in the latent representation. This suggests that CVAE effectively filters out noise-driven or spurious correlations, enhancing the specificity of functional connectivity estimates. In contrast, the remaining 155 connections showed no significant change between these two, reflecting strong preservation of core, robust functional relationships. These results demonstrate the CVAE's dual capacity: it retains the most salient, biologically meaningful connectivity patterns while attenuating weaker or unreliable associations. As a result, the latent space provides a cleaner, more interpretable representation of brain network organization, offering a powerful framework for individualized and noise-resilient functional connectivity analysis. Figure 6 (c, d) shows a scatterplot of the different connectivity measures for all pairs of brain regions. The measures tend to converge for high correlation values, meaning that what is highly correlated and likely important in input space is similar in latent space while the latent space correlations and distances are more spread out, giving a more nuanced view of the connectivity. Full x-axis range for Figure 6c,d is shown in Appendix Figure D.1 and D.2.

These results suggest several key insights: (i) The CVAE maintains core functional connectivity patterns post-denoising, with minor improvements in signal quality. (ii) The latent embeddings offer an alternative yet consistent representation of functional connectivity, capable of revealing nuanced inter-regional relationships. (iii) Using both Pearson correlation and Euclidean distance in the latent space provides complementary perspectives on regional interactions.

Taken together with findings from Figure 4b), these results indicate that the subject-specific latent spaces learned by the CVAE serve as a robust coordinate system for analyzing voxel- and region-level BOLD signal similarities.

Discussion

In this study, we demonstrated that Convolutional Variational Autoencoders (CVAEs) provide an effective framework for individualized fMRI representation learning. Our models denoise voxel-wise BOLD signals, reconstruct subject-specific latent spaces, and enable cross-subject alignment via Procrustes analysis. These embeddings improve signal quality while preserving the biological integrity of functional connectivity.

Beyond denoising, CVAEs holds potential for detecting outliers and generating new BOLD samples, areas that warrant future exploration. Since our CVAE approach learns subject-specific representations that can later be aligned. It could potentially preserve individual anatomical variations that are typically lost during spatial normalization and enables cross-subject analysis. This framework provides a more compact representation that could improve computational efficiency for subsequent analyses. It opens new avenues for studying individual differences in brain connec-

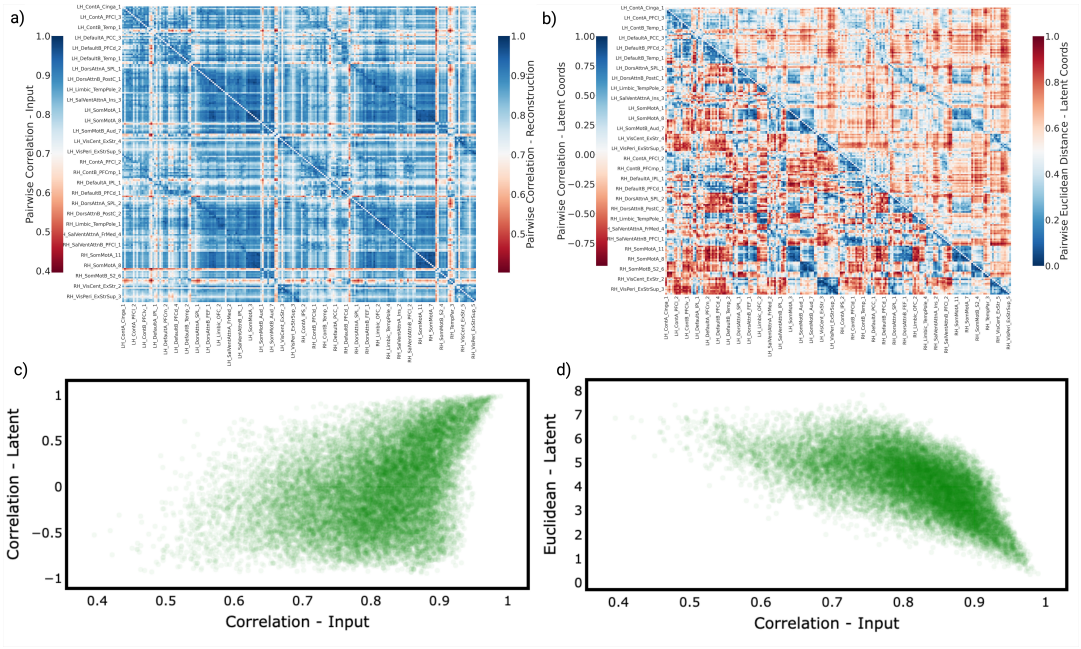

Figure 6: **Functional connectivity in latent space matches input while reducing noise-driven edges.** a) Functional connectivity based on pairwise correlation of BOLD signals using the input data (lower triangular part) and CVAE reconstructions (upper triangular). b) Functional connectivity based latent coordinates using correlation (lower triangular) and Euclidean distance (upper triangular, normalized values with reverse colorbar). (c,d) show tight correspondence between input and latent connectivity patterns. c) Scatter plot between correlation values of input (a, lower triangular part) and latent encodings (b, lower triangular part). d) Scatter plot between correlation values of input (a, lower triangular part) and Euclidean distance of latent encodings (b, upper triangular part).

tivity, offering the detection of clinically relevant deviations in network organization linked to psychiatric and neurological disorders. However, our methodology is conceptually applicable to un-normalized data as well, which represents a significant potential advantage.

However, VAEs come with known challenges, such as posterior collapse and the "hole problem" (Tomczak 2022), which could affect latent representations. Addressing these issues remains crucial for improving the reliability of VAE-based models. Additionally, since we trained the CVAE with the KL-divergence term set to zero, the model behaved as a regularized autoencoder. In future work, we aim to examine the influence of the KL divergence term on reconstruction performance and explore the extent to which it affects the separability of latent representations (Higgins et al. 2017).

In summary, our work demonstrates that deep latent models like the CVAE are not merely dimensionality reduction tools but represent a paradigm shift in functional connectivity analysis. By integrating denoising, dimensionality reduction, and individual-level modeling within a unified framework, they offer a more accurate, biologically plausible, and interpretable view of brain network dynamics. Future work will explore the utility of these latent representations in decoding cognitive states, predicting individual behavior, and identifying biomarkers of brain disorders, positioning deep latent modeling as a cornerstone of next-generation neuroimaging analysis.

Conclusion

This study demonstrates that deep latent variable models, specifically CVAEs, offer a principled and flexible approach to reducing the complexity of fMRI analysis by generating meaningful subject-specific latent space representations, capture smooth, anatomically coherent neural dynamics with minimal inter-dimensional correlation. Our results highlight the potential of these models to denoise BOLD signals while preserving and enhancing the integrity of functional connectivity patterns.

Crucially, the latent space maintains strong correspondence with established functional architecture, preserving the most robust, biologically relevant connections while simultaneously suppressing noise driven or spurious correlations. Quantitative comparison revealed that the vast majority of connectivity relationships were either reweighted or suppressed in the latent space, with only 155 connections showing no significant change. This indicates that CVAE does not merely compress data but actively disentangles meaningful neural signals from noise, yielding a more specific and interpretable representation of brain network organization. Together, these findings establish CVAE as a powerful tool for individualized functional connectivity analysis, offering a data-driven, low-dimensional framework that enhances signal quality without distorting underlying neural architecture.

References

Chang, L.; Fleetwood, G.; Manning, J.; Parkinson, C.; Finn, E.; Wager, T.; Haxby, J.; Geerligs, L.; Vega, A. d. l.; Lahnakoski, J.; et al. 2020. Neuroimaging analysis methods for naturalistic data. In *Annual meeting of the Organization for Human Brain Mapping 2020*, FZJ-2020-04493. Gehirn & Verhalten.

Chen, P.-H. C.; Chen, J.; Yeshurun, Y.; Hasson, U.; Haxby, J.; and Ramadge, P. J. 2015. A reduced-dimension fMRI shared response model. *Advances in neural information processing systems*, 28.

Finn, E. S.; Corlett, P. R.; Chen, G.; Bandettini, P. A.; and Constable, R. T. 2018. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature communications*, 9(1): 2043.

Friston, K.; Frith, C.; and Frackowiak, R. 1993. Time-dependent changes in effective connectivity measured with PET. *Human Brain Mapping*, 1(1): 69–79.

Glasser, M. F.; Coalson, T. S.; Bijsterbosch, J. D.; Harrison, S. J.; Harms, M. P.; Anticevic, A.; Van Essen, D. C.; and Smith, S. M. 2018. Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. *Neuroimage*, 181: 692–717.

Glover, G. H. 2011. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2): 133–139.

Han, K.; Wen, H.; Shi, J.; Lu, K.-H.; Zhang, Y.; Fu, D.; and Liu, Z. 2019. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, 198: 125–136.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of classification*, 2(1): 193–218.

Joas, M.; Jurenaite, N.; Praščević, D.; Scherf, N.; and Ewald, J. 2024. A generalized and versatile framework to train and evaluate autoencoders for biological representation learning and beyond: AUTOENCODIX. *bioRxiv*, 2024–12.

Kim, J.-H.; Zhang, Y.; Han, K.; Wen, Z.; Choi, M.; and Liu, Z. 2021. Representation learning of resting state fMRI with variational autoencoder. *NeuroImage*, 241: 118423.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Koppe, G.; Toutounji, H.; Kirsch, P.; Lis, S.; and Durstewitz, D. 2019. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLoS computational biology*, 15(8): e1007263.

Li, W.; Wang, S.; and Liu, G. 2022. Transformer-based model for fMRI data: ABIDE results. In *2022 7th International Conference on Computer and Communication Systems (ICCCS)*, 162–167. IEEE.

Margulies, D. S.; Ghosh, S. S.; Goulas, A.; Falkiewicz, M.; Huntenburg, J. M.; Langs, G.; Bezgin, G.; Eickhoff, S. B.; Castellanos, F. X.; Petrides, M.; et al. 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44): 12574–12579.

Moschella, L.; Maiorca, V.; Fumero, M.; Norelli, A.; Locatello, F.; and Rodolà, E. 2022. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*.

Nastase, S. A.; Liu, Y.-F.; Hillman, H.; Norman, K. A.; and Hasson, U. 2020. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, 217: 116865.

Pruim, R. H.; Mennes, M.; van Rooij, D.; Llera, A.; Buitelaar, J. K.; and Beckmann, C. F. 2015. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage*, 112: 267–277.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Sasse, L.; Paquola, C.; Dukart, J.; Hoffstaedter, F.; Eickhoff, S. B.; and Patil, K. R. 2024. Procrustes Alignment in Individual-level Analyses of Functional Gradients. *bioRxiv*, 2024–11.

Schaefer, A.; Kong, R.; Gordon, E. M.; Laumann, T. O.; Zuo, X.-N.; Holmes, A. J.; Eickhoff, S. B.; and Yeo, B. T. 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex*, 28(9): 3095–3114.

Simony, E.; Honey, C. J.; Chen, J.; Lositsky, O.; Yeshurun, Y.; Wiesel, A.; and Hasson, U. 2016. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, 7(1): 12141.

Tomczak, J. 2022. *Deep Generative Modeling*. Germany: Springer. ISBN 978-3-030-93157-5.

Wang, Y.; Li, D.; Li, L.; Sun, R.; and Wang, S. 2024. A novel deep learning framework for rolling bearing fault diagnosis enhancement using VAE-augmented CNN model. *Heliyon*, 10(15).

A Latent Dimensionality

All subjects were trained with the CVAE model, and test loss was evaluated across latent dimensions ranging from 2 to 16. The mean test loss was computed by averaging across all subjects. The results indicate that the mean test loss remains nearly constant beyond 8 latent dimensions. A few subjects exhibited slightly higher losses due to variations in parameter initialization. Since the mean test loss stabilized after 9 latent dimensions, we opted for a smaller latent dimension rather than larger ones (Figure A.1).

B Adjusted purity and silhouette score for Input data and Latent Representations

To assess clustering and separation quality between the input data and latent representations, we computed Adjusted Purity (ARI) and the Silhouette score. Both datasets were reduced to low dimensions using PCA (2 and 9 components),

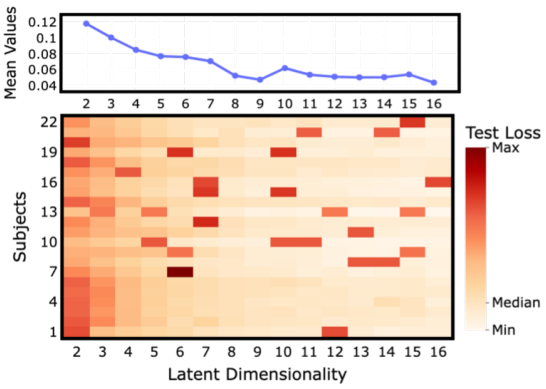

Figure A.1: The test loss of the CVAE model was evaluated for each subject across latent dimensions ranging from 2 to 16. The results show that beyond 8 latent dimensions, the mean test loss remains nearly constant (indicated by the blue line). The dark red markers represent cases with higher test loss, which are attributed to poor parameter initialization.

and the evaluation was performed with respect to the Yeo 17 networks and the Schaefer 200 parcellation. Notably, the ARI values for the VAE latent representations were approximately twice those of the input data, indicating that the latent space preserves more clustering-relevant information (Table B.1). This means clustering in reduced space is only slightly better than random. In contrast, the Silhouette score for the VAE latent increased only marginally (by 0.02) compared to the input (Table B.2).

Table B.1: Adjusted purity for input data and latent representations

Dim.	Type	PCA of Input	VAE Latent
2	Yeo 17 Networks	0.08	**0.11**
2	Schaefer 200	0.02	**0.05**
9	Yeo 17 Networks	0.11	**0.20**
9	Schaefer 200	0.05	**0.19**

Table B.2: Silhouette score for input data and latent representations

Type	PCA of Input (2)	VAE Latent
Yeo 17 Networks	0.33	**0.35**
Schaefer 200 parcellation	0.32	**0.34**

C Distribution of Diagonal Correlations over 9 aligned Latent dimensions across 22 subjects

We extracted the diagonal correlation values from Figure 5b and plotted them to identify which latent dimensions are more consistently correlated across subjects. Subject 7 (pink line), which shows a correlation of 1 across all 9 latent representations, serves as the reference. Latent dimensions 5–8 exhibit lower correlations across subjects, reflecting inter-subject variability. In contrast, latent dimensions 1–4 display higher correlations (0.5–0.8), indicating that they are largely shared across subjects and may encode information relevant to the naturalistic dataset (Figure C.1).

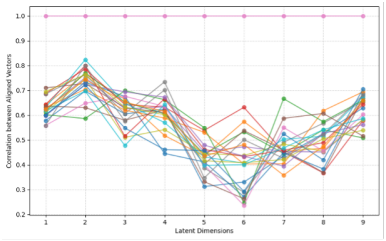

Figure C.1: Diagonal correlations (Subject 7 as reference) show variability in dimensions 5–8 and shared naturalistic information in dimensions 1–4.

D X-axis range adjusted to 1 to 1 for Figure 6c,d

In the main figure, panels 6c and 6d do not display the full x-axis range; therefore, we provide the complete visualization here (Figure D.1 and D.2).

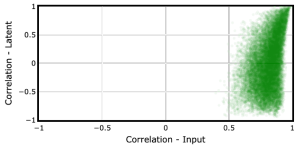

Figure D.1: Scatter plot between correlation values of input (Figure 6a, lower triangular part) and latent encodings (Figure 6b, lower triangular part).

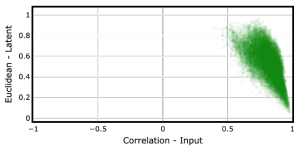

Figure D.2: Scatter plot between correlation values of input (Figure 6a, lower triangular part) and Euclidean distance of latent encodings (Figure 6b, upper triangular part).