Deep learning models reveal the link between dynamic brain connectivity patterns and states of consciousness

Chloé Gomez, Lynn Uhrig, Vincent Frouin, Edouard Duchesnay, Béchir Jarraya, Antoine Grigis

▶ **To cite this version:**

Chloé Gomez, Lynn Uhrig, Vincent Frouin, Edouard Duchesnay, Béchir Jarraya, et al.. Deep learning models reveal the link between dynamic brain connectivity patterns and states of consciousness. 2024. hal-04512801

HAL Id: hal-04512801
https://hal.science/hal-04512801

Preprint submitted on 20 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open licence - etalab

Deep learning models reveal the link between dynamic brain connectivity patterns and states of consciousness

Chloé Gomez[1*], Lynn Uhrig[1,2], Vincent Frouin[3], Edouard Duchesnay[3], Béchir Jarraya[1,4+], and Antoine Grigis[3+]

[1]Cognitive Neuroimaging Unit, NeuroSpin center, CEA, INSERM U992, Université Paris-Saclay, Gif-sur-Yvette, France
[2]Department of Anesthesiology and Critical Care, Necker Hospital, AP-HP, Université de Paris, Paris, France
[3]BAOBAB Unit, NeuroSpin center, CEA, Université Paris-Saclay, Gif-sur-Yvette, France
[4]Neuroscience Pole, Foch Hospital, Université Paris-Saclay, UVSQ, Suresnes, France
[*]For correspondance : chloe.gomez@cea.fr
[+]These authors co-supervised the study

ABSTRACT

Decoding states of consciousness from brain activity is a central challenge in neuroscience. Dynamic functional connectivity (dFC) allows the study of short-term temporal changes in functional connectivity (FC) between distributed brain areas. By clustering dFC matrices from resting-state fMRI, we previously described "brain patterns" that underlie different functional configurations of the brain at rest. The networks associated with these patterns have been extensively analyzed. However, the overall dynamic organization and how it relates to consciousness remains unclear. We hypothesized that deep learning networks would help to model this relationship. Using low-dimensional variational autoencoders (VAE), recent studies have attempted to learn meaningful representations that can help explain consciousness. Here, we investigated the complexity of selecting such a generative model to study brain dynamics, and extended the available methods for latent space characterization and modeling. Therefore, our contributions are threefold. First, in comparison with probabilistic principal component analysis and sparse VAE, we showed that the selected low-dimensional VAE exhibits balanced performance in reconstructing dFCs and classifying brain patterns. The organization of the obtained low-dimensional dFC latent representations was then explored. We showed how these representations stratify the dynamic organization of the brain patterns as well as the experimental conditions. Finally, we proposed to delve into the proposed brain computational model. A receptive field analysis was first applied to identify preferred directions in the latent space to move from one brain pattern to another. Then, an ablation study was achieved where specific brain areas were virtually inactivated. We demonstrated the efficiency of the model in summarizing consciousness-specific information that is encoded in key inter-areal connections, as described in the global neural workspace theory of consciousness. The proposed framework advocates the possibility to develop an interpretable computational brain model of interest for disorders of consciousness, paving the way for a dynamic diagnostic support tool.

Introduction

"The stream of our consciousness, [. . .] like a bird's life, seems to be made of an alternation of flights and perchings", said the philosopher William James[1]. A fundamental observation, that still puzzles many scientists. Like the seasons that transform our landscapes, the spontaneous fluctuations of the brain reveal very different brain configurations. Brain activity at rest is commonly characterized by the spontaneous fluctuations of regional brain fMRI signals. Dynamic functional connectivity (dFC), a recent expansion to traditional functional connectivity (FC) analysis, explores the short-term temporal changes in FC between distributed brain areas[2] and goes beyond the typical assumption that functional networks are temporally static.

DFC matrices can be clustered into several "brain patterns". These brain patterns are typically grouped using unsupervised k-means clustering, where the centroid of each cluster represents a pattern[2–4]. Temporal analysis of the dFCs shows that wakefulness and loss of consciousness exhibit a reorganizing repertoire of brain patterns. This repertoire corresponds to an ensemble of distinct and repeatable patterns of brain activity[3–5]. The conscious brain is the site of rapidly changing dynamics, within a rich repertoire of brain patterns. Conversely, during anesthesia and disorders of consciousness, brain activity is expressed according to a more rigid and poorer repertoire of brain patterns (i.e., transitions between brain patterns are rare, and some brain patterns are almost never visited). In this case, the brain dynamic connectivity is reduced to the underlying anatomical connectivity[3,4,6].

The representation and interpretation of brain patterns is still an area of ongoing research. Previous dynamic studies have examined the frequency of occurrences or stability of each brain pattern[2,3] and have proposed to project fMRI data into two- or three-dimensional space[7]. However, they either do not take into account the spatiotemporal nature of the data or focus on task fMRI rather than rs-fMRI[8,9]. Consequently, it may be interesting to explore such a low-dimensional space to model a fine-grained representation of brain patterns. In addition, there is no evidence yet on the optimal size of this low-dimensional space (i.e., on the learning capacity of the model).

Some works in the literature support the choice of a low-dimensional model to study brain dynamics. For example, dFCs have been shown to reflect the interplay of a small number of latent processes using clustering or PCA-based reduction techniques[2,10] and latent linear models can also be used to estimate these underlying processes[11]. However, linear models may be inadequate if the mapping is nonlinear or, equivalently, if the learned manifold is curved. To address this issue, several studies have implemented variational autoencoders (VAEs)[7,12,13]. The probabilistic nature of such generative models holds great promise for exploring the data structure. Unlike discriminative models, VAEs are unsupervised models that do not require a labeled dataset. In the proposed work, the choice of architecture is supported by the seminal work of Perl and colleagues[7]. They showed that when a VAE (which parameterized both the encoder and decoder using a multi-layer perceptron (MLP)) is trained with simulated whole-brain data from awake and asleep healthy volunteers, the learned representations showed faded states of wakefulness. The choice of the optimal latent space dimension remained an open question in their work. Overall, this choice is a trade-off between compressing only essential information and preserving data reconstruction.

We proposed a new interpretability framework, called VAE for Visualizing and Interpreting the ENcoded Trajectories (VAE-VIENT) between states of consciousness (Figure 1). We took advantage of a previously acquired resting-state fMRI dataset in which non-human primates were scanned under different experimental conditions: awake state and anesthesia-induced loss of consciousness using different anesthetics (propofol, sevoflurane, ketamine)[3,4]. After presenting the considered low-dimensional generative model, we showed that a 2D VAE has a balanced performance in reconstructing dFCs and classifying brain patterns. We then proposed a discrete and continuous characterization of the latent space. Finally, we showed that this model can translate some virtual modifications or inactivations of inter-areal brain connections into a transition of consciousness.

Methods

Dataset

MRI acquisitions This study used a rs-fMRI dataset previously acquired under different experimental conditions: awake state and anesthesia-induced loss of consciousness using different anesthetics (propofol, sevoflurane, ketamine)[3,4]. The data were collected for a previous project to discover a new signature of anesthesia-induced loss of consciousness. Here, we proposed a retrospective analysis of these data without additional experiments. In this way, we will maximize their use and shed new light on them.

Data were collected from five rhesus macaques (Macaca mulatta), one male (monkey J) and four females (monkeys A, K, Ki, and R), 5 to 8 kg, 8 to 12 years old, either in the awake state or under anesthesia (deep ketamine, moderate/deep propofol, or moderate/deep sevoflurane anesthesia), representing six conditions. Three monkeys were scanned for each arousal state (awake: monkeys A, K, and J - propofol anesthesia: monkeys K, R, and J - ketamine anesthesia: monkeys K, R, and Ki - sevoflurane anesthesia: monkeys Ki, R, and J). Levels of anesthesia were defined by a clinical arousal score and continuous EEG monitoring (for details, see *Uhrig et al. (2018)*[4]). 156 rs-fMRI runs (31 for the awake state, 25 for moderate propofol, 30 for deep propofol, 25 for moderate sevoflurane, 20 for deep sevoflurane, and 25 for deep ketamine) of 500 volumes each were acquired on a 3T Siemens MRI with a customized single transmit-receiver surface coil, and a repetition time of 2.4s.

In the original study, all procedures were conducted in accordance with the European Convention for the Protection of Vertebrate Animals used for Experimental and Other Scientific Purposes (Directive 2010/63/EU) and the National Institutes of Health's Guide for the Care and Use of Laboratory Animals. Animal studies were approved by the institutional Ethical Committee (Commissariat à l'Énergie atomique et aux Énergies alternatives; Fontenay aux Roses, France; protocols 10-003 and 12-086).

Pre-processing The NeuroSpin Monkey (NSM) spatial preprocessing[3,4] was applied, which includes the following steps: slice timing correction, B0 inhomogeneities correction, motion correction, reorientation, masking, realignment, and smoothing. Time series denoising operations were then applied[3,4]. Specifically, the voxel time series were detrended, filtered with low-pass (0.05-Hz cutoff), high-pass (0.0025-Hz cutoff), and zero-phase fast-Fourier notch (0.03 Hz, to remove an artifactual pure frequency present in all the data) filters, regressed against motion confounds, and z-score standardized. The denoised voxel time series were further averaged over the 82 cortical Regions Of Interest (ROIs) of the CoCoMac atlas[14] and sliced into sliding time windows of 35 TR with sliding step of a 1 TR[3]. Finally, sparse correlations between ROIs were estimated using L1-norm regularization for each window[2], resulting in 72,384 dFCs. L1-norm regularization provides sparse interpretable solutions, but requires the choice of a regularization parameter λ. As described in *Barttfeld et al. (2015)*[3], this parameter was set to 0.1.

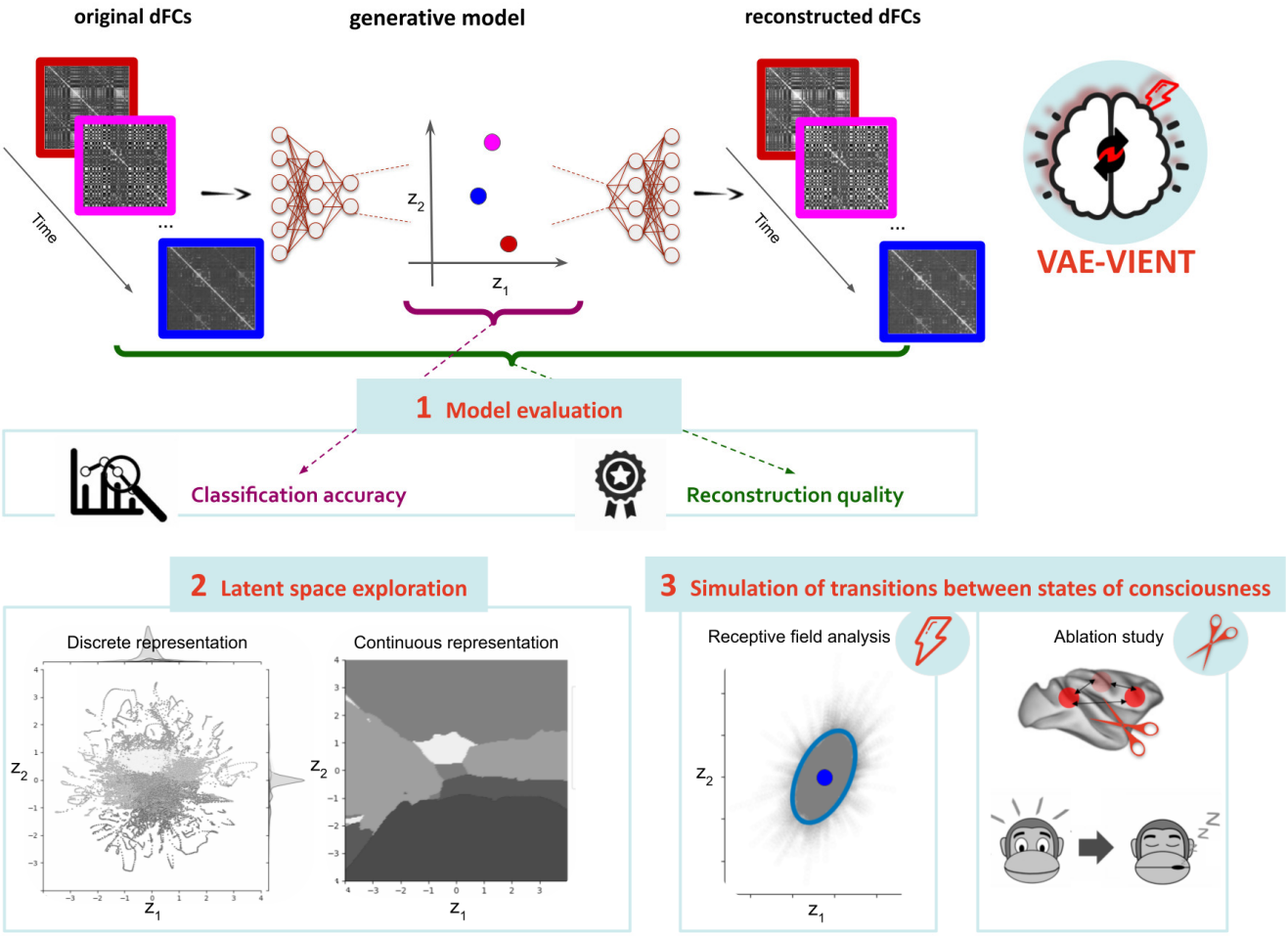


Figure 1. Illustration of the proposed VAE-VIENT framework. A VAE learns 2D latent representations $z = (z1_i, z2_i)$ from dynamic functional connectivity matrices (dFCs), leading to 1) evaluation of the proposed model against other generative models implementing different latent dimensions, 2) exploration of latent space with the ability to view discrete or continuous representations (here we observe how brain patterns are organized in latent space), and 3) two simulation paradigms, including a receptive field analysis that generates tensor representations to study the effect of perturbing input dFCs, and an ablation study of Global Neural Workspace (GNW) connections to study the transition from wakefulness to unconsciousness.

Low-dimensional generative models

The Gaussian VAE The emergence of deep learning-based generative models has spread to many disciplines, including medicine and neurosciences[12, 15–17]. By learning and capturing the underlying probability distribution of the training data, generative models are able to generate novel samples with inherent variability. Three prominent families of generative models can be identified, namely generative adversarial networks, variational autoencoders (VAEs)[18], and diffusion models. We will focus on VAEs in this work. VAE training involves learning both an encoder to transform data as a distribution over the latent space and a decoder to reconstruct the original data (Figure 1). The training minimizes the mean squared error reconstruction term, making the encoding/decoding scheme as effective as possible. Latent space regularity is enforced during the training to avoid overfitting and to ensure continuity (two nearby points in the latent space give similar content once decoded) and completeness (a code sample from the latent space should provide relevant content once decoded). These properties are at the core of the generative process. In practice, a regularization term constrains the encoding distributions to be close to a standard normal distribution using the Kulback-Leibler (KL) divergence.

Let's consider a dataset $D = \{X^{(1)}, ..., X^{(n)}\}$ with $n = 72,384$ dFC samples, where each sample $X^{(i)} = [x_1^{(i)}, ..., x_d^{(i)}]$ is a vector of $d = 3321$ dimensions (the dFC upper triangular elements). An autoencoder learns an identity function in an unsupervised way as follows:

$$\tilde{X^{(i)}} \approx f_\theta(g_\phi(X^{(i)})) \tag{1}$$

where $g_\phi(.)$ denotes the encoder, $f_\theta(.)$ the decoder, and $\tilde{X^{(i)}}$ is the network reconstruction of $X^{(i)}$. The reconstruction loss, expressed as a Mean Squared Error (MSE), can be written as:

$$L_{MSE}(\theta,\phi) = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \tilde{X^{(i)}}) = \frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)} - f_\theta(g_\phi(X^{(i)}))\right) \tag{2}$$

In this work, the VAE relationship between the input dFC data $X^{(i)}$ and the latent encoding vector $z^{(i)}$ was defined with a prior $p_\theta(z^{(i)}) \sim \mathcal{N}(z^{(i)};0,1)$, the likelihood $p_\theta(X^{(i)}|z^{(i)})$, and the posterior $p_\theta(z^{(i)}|X^{(i)})$. Unlike (finite) Gaussian mixture models, the posterior $p_\theta(z^{(i)}|X^{(i)})$ is intractable. Therefore, we used a posterior approximation $q_\phi(z^{(i)}|X^{(i)})$ that outputs what is a likely code given an input $X^{(i)}$. It plays a similar role as $g_\phi(z^{(i)}|X^{(i)})$. In our case of Gaussian VAE, $q_\phi(z^{(i)}|X^{(i)}) = \mathcal{N}(z^{(i)};m_\phi(X^{(i)}),s_\phi(X^{(i)}))$, where m_ϕ and s_ϕ are expressive parameterizations of the conditional mean and variance of $q_\phi(z^{(i)}|X^{(i)})$. The distributions returned by the encoder are further constrained to follow a standard normal distribution as follows:

$$L_{KL}(\theta,\phi) = D_{KL}(q_\phi(z^{(i)}|X^{(i)})||p_\theta(z^{(i)})) \tag{3}$$

where D_{KL} is the KL divergence. To learn disentangled representations and increase interpretability, a regularization parameter β was further introduced[19,20]. The idea is to keep the distance between the real and the estimated posterior distribution small while maximizing the probability of generating real data. A high β value emphasizes statistical independence over reconstruction. The final VAE loss was expressed as follows:

$$L_{VAE}(\theta,\phi) = L_{MSE}(\theta,\phi) - \beta L_{KL}(\theta,\phi) \tag{4}$$

The considered generative models We considered a VAE with a one (VAE_1), two (VAE_2) or three (VAE_3) dimensional latent space, adapting the architecture proposed in *Perl et al. (2020)*[7]. The input was the upper triangular dFCs (as each dFC is symmetric). Then, the encoder part used two hidden fully connected layers (512 and 256 units, respectively) with ReLU activation functions, and the decoder part was implemented with the same structure. The dimension of the latent space corresponds to common neurobiological assumptions made when studying disorders of consciousness[21–23]. Furthermore, we compared our models with the sparse VAE (sVAE)[24], initialized with thirty-two latent dimensions. The sVAE implemented a variational dropout to enforce parsimony and interpretability in the latent representations. A threshold on the dropout rates was used to select the optimal number of latent dimensions. We also applied a baseline machine learning model, the probabilistic PCA (PPCA)[25], and compared the results to those obtained with VAE/sVAE. In fact, PPCA can be considered as a latent variable model. Its assumptions are Gaussian distributions and linear decomposition. The purpose of adding this model was to assess the interest in non-linear models such as VAE or sVAE when working with small datasets.

Model training We trained the VAE and sVAE using an Adam optimizer, with a learning rate starting at 0.001 and a 10% decay every 30 epochs. To limit overfitting during the training, we included early stopping. The model was trained on the training set until its error on the validation set increased, at which point the optimization stopped. As a performance measure to monitor the stopping of training, we considered the sliding median using a 10 epoch interval. In addition, the patience argument allowed training to continue for up to 15 epochs after convergence. This gave the training process a chance to get over flat areas or find additional improvements. Using cross-validation, we studied the effect of the β regularization parameter for the VAE by performing a grid search to determine the better choice for $\beta \in [0.5,20]$ with the following user-defined steps $[0.5,1,4,7,10,20]$. With the intention of building an interpretable model, we kept 8 models: $PPCA_1$, $PPCA_2$, $PPCA_3$, VAE_1, VAE_2, VAE_3 with 1, 2, and 3 latent dimensions respectively, sVAE, and PPCA with the same number of latent variables selected by the sVAE. We performed a leave-one-subject-out to create an independent test set, and a training set with an internal 5-fold cross-validation. In the cross-validation, the stratification of the arousal conditions further strengthens the distribution of the classes in each training split. In the end, only the weights associated with the best validation fold were evaluated on the independent test set.

The labels and pseudo-labels used were the arousal conditions (awake and the different anesthetics) and the brain patterns (BPs) ranked in ascending order of similarity to the structural connectivity (numbered 1 to 7), respectively. Briefly, the use of seven brain patterns has been shown to be effective in representing the different configurations of the brain[3,4]. Choosing the optimal number of brain patterns is challenging. It results from balancing biological assumptions and computational evaluations. These labels are known to be unevenly distributed across experimental conditions. They are also known to be good descriptors of spontaneous fluctuations in brain activity.

Model evaluation

Choosing an appropriate model is a trade-off between compressing only essential information and preserving data reconstruction. Thus, we evaluated the models using two distinct metrics. The first metric was a measure of the reconstruction quality. The second one was the relative entropy of the latent space, measured by considering a classification task. Both used as labels the seven brain patterns previously described[3,4].

Reconstruction quality From the retained trained generative models, we computed the decoded dFC matrices $\tilde{X^{(i)}}$ associated with the test set. Instead of using the MSE training loss, we evaluated the Structural SIMilarity (SSIM) between the averaged decoded dFCs and true decoded dFCs associated with each label. The MSE calculation focuses on pixel values, while the SSIM measurement focuses on and analyzes the structural differences between two dFCs. Unlike the SSIM, the MSE can be very high just because some connection values have changed. Therefore, we prefered the SSIM because we wanted to study global dFC patterns. This metric ranges from 0 to 1, where 1 is a perfect match.

Classification accuracy From the retained trained models, we also computed the latent representations associated with the test set. In addition to the BP labels available from the dataset, we also matched each test dFC latent space location to its nearest location in the train set and retained the corresponding matched label. Balanced accuracy (BAcc) was then used to compare the dataset and matched BP labels.

Consensus metric We proposed a consensus metric $\mathscr{M}$, which is an average between SSIM and BAcc. The goal was to enforce a trade-off that imposed spatial coherence in the latent space without significantly degrading the reconstruction quality.

Latent space exploration

Discrete and continuous descriptors All the information we can transfer in latent space is associated with encoder-generated representations and is discrete by nature. To build a comprehensive whole-brain computational model, semantically continuous representations are required. With the generative capabilities of the VAE (or generative models in general), it is possible to decode the entire latent space. Without losing generality, let us give the formula in 2D. Let's consider a discrete grid $G \in R^2$ with $g \times g$ latent samples and the associated decoded dFCs $\tilde{X_{lm}}$, with $l \in [1,g]$ and $m \in [1,g]$. Using the previously known information on the brain patterns, we could label each $\tilde{X_{lm}}$. To this end, and as suggested by *Perl et al. (2020)*[7], we computed the similarity between each $\tilde{X_{lm}}$ and each brain pattern. To assess the strength of these associations, we used Pearson's correlation. At the end, the label assigned to $\tilde{X_{lm}}$ is the number of the most correlated brain pattern. The obtained continuous labeling reflects the functional reconfiguration of the brain.

Confidence level of continuous descriptors In addition, by quantifying the best association strength, we proposed to compute confidence and reliability maps associated with the continuous descriptor generation process. First, the confidence map $\mathscr{CM}$ was derived at each latent space location by taking the average of the difference between the two largest associations and the correlation between the two closest brain patterns as follows:

$$\mathscr{CM}_{lm} = \frac{1}{2}\left((\mathscr{R}(\tilde{X_{lm}}, \bar{BP_1}) - \mathscr{R}(\tilde{X_{lm}}, \bar{BP_2})) + \mathscr{R}(\bar{BP_1}, \bar{BP_2})\right) \tag{5}$$

where $\mathscr{R}$ is the Pearson correlation, and $\bar{BP_1}$ and $\bar{BP_2}$ are the brain patterns with the first and second highest correlation, respectively. This metric takes into account both the reluctance to label and the objective nature of that reluctance. The model is reasonably confident when $\mathscr{CM}_{lm} \approx 0.5$, and overconfident when $\mathscr{CM}_{lm} \approx 1.0$. Second, the reliability map $\mathscr{RM}$ was expressed at each latent space location by decoding the dFC and targeting the brain pattern with the highest Pearson correlation:

$$\mathscr{RM}_{lm} = \max_{k=1;k\leq 7} \mathscr{R}(\tilde{X_{lm}}, BP_k) = \mathscr{R}(\tilde{X_{lm}}, \bar{BP_1}) \tag{6}$$

The higher $\mathscr{RM}_{lm}$, the more reliable the model is.

Connection-wise simulations

The receptive field analysis If we consider brain dynamics as a physical process characterized by gradual changes in the FC space, then the dFCs used so far are samples of these changes. Physicists tried to model these processes in a principled way by analytically identifying prior knowledge about the underlying processes, e.g. by using differential equations[26]. Instead of incorporating physical knowledge into a deep neural network, we proposed a receptive field (RF) simulation paradigm to generate a tensor model of latent space and thus gain insight into its dynamics. Such a characterization of latent space is essential for building an interpretable model and it can help to understand encoded latent trajectories between states of

consciousness. We proposed to capture the latent space RF at the connection level. In this way, the proposed RF analysis could identify the connections that need to be disrupted in order to move from one state of consciousness to another. In the long run, such an analysis may be a tool to simulate the recovery of consciousness at the individual level.

In more detail, the RF analysis focused on the trained VAE (or another generative model) encoder. A perturbation was simulated at each connection $j \in [1,d]$ of an input dFC matrix $X^{(i)}$ (Figure 2-A). The effect of this perturbation on the encoded latent representations was tracked (Figure 2-B)[15]. In particular, the simulation modified a single connection value $x_j^{(i)}$ p times, by swapping its value with a correlation drawn uniformly in an interval $[-1,1]$, while keeping the other connections fixed. In two dimensions, the latter simulation yielded p latent encoded vectors $z_j^{(i)} = \{z_{j1}^{(i)},\ldots,z_{jk}^{(i)}\} \in \mathbb{R}^{p\times 2}$. The generated latent samples $z_j^{(i)}$ were distributed around a line of varying length. This specific behavior allowed an interesting parameterization of each perturbation, using polar (in 2 dimensions) or spherical (in 3 dimensions) coordinates, through the inclinations $\theta_j^{(i)}$ (Figure 2-B). The perturbation of all connections returned a cloud of points describing the RF. The resulting cloud had an ellipsoidal shape $\mathscr{E}$, estimated with a confidence interval of 0.01. $\mathscr{E}$ can be parameterized by its sorted eigenvalues λ_i and associated eigenvectors $\vec{e}_i$, $i \in [1,N]$, where N is the latent dimension. Finally, since each connection can be related to a direction by the inclination angle $\theta_j^{(i)}$, it was possible to select the connections with high potential for action (i.e., generating the highest brain transitions when perturbed) by identifying the directions aligned with the first eigenvector $\vec{e}_1$ of the ellipsoid $\mathscr{E}$. The procedure described above can be applied to any dFC. That is, any dFC can be projected onto a point in latent space around which an ellipsoid representing the effect of all possible unit perturbations was computed.

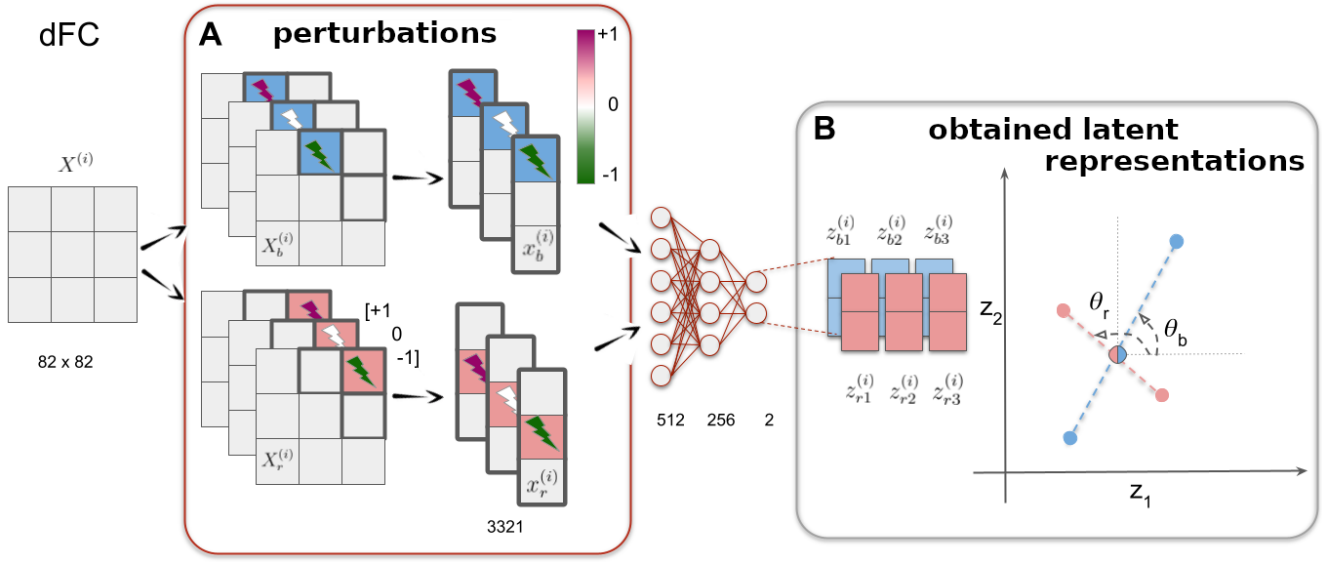


Figure 2. Illustration of the two steps involved in the receptive field analysis. A) From a dFC matrix $X^{(i)}$, or equivalently its upper terms $x^{(i)}$, two perturbations are performed on connections b (blue) and r (red) by swapping one connection with the following three correlations $[-1,0,1]$ ($p=3$). B) Corresponding $z_b^{(i)}$ and $z_r^{(i)}$ latent representations ($N=2$) follow lines and are summarized by their inclination angles with the x-axis θ_b and θ_r, respectively.

The ablation analysis Stemming from our previous work in which we identified key networks underlying consciousness in the macaque brain[27,28], we virtually inactivated specific brain areas and tested the ability of the trained VAE to efficiently predict transitions across consciousness levels. For this purpose, we simulated specific inactivations/ablations of functional connections between brain areas as a virtual experiment. The resulting dFC representations were then used to predict the state of consciousness. Using the framework of the Global Neural Workspace (GNW) theory of consciousness, we previously identified key brain areas (referred to as "macaque GNW nodes") that account for the cortical signature of consciousness realizing a fronto-parieto-cingular network[27,28]. The key brain regions whose associated connections were zeroed in this study were the posterior cingulate cortex (CCp), the anterior cingulate cortex (CCa), the intraparietal cortex (PCip), the frontal eye field (FEF), the dorsolateral prefrontal cortex (PFCdl), the prefrontal polar cortex (PFCpol) and the dorsolateral premotor cortex (PMCdl) of the left and right hemispheres[4,28]. Thus, we proposed a connection-wise ablation study, equivalent to a lesion perturbation, that removed the contribution of connections linked to these regions. It's important to note that removing a region removes all connections associated with that region. By zeroing these connections, we expected to shift dFCs acquired in the awake state to

an anesthetized state.

To predict the awake and anesthetized states, we trained an SVM classifier on the learned latent representations. We then evaluated the performance of the classifier in predicting the awake state using the balance accuracy (BAcc). To focus on the effect of the proposed ablation and to eliminate any unrelated source of variability, the analysis was performed on the training set only. As input to the trained SVM, we took only the raw or perturbed awake dFCs (i.e., awake dFC undergoing the ablation process) encoded with the VAE. We denoted the corresponding prediction scores as $BAcc$ and $B\tilde{A}cc$, respectively. To assess the specificity of zeroing these nodes, we tested the null hypothesis that removing random connections did not result in a significant loss of prediction compared to targeted GNW-associated connections. Let $G = 14$ be the number of GNW-associated connections. Our goal was to modify G connections that were not part of the GNW-related connections. The cardinality of the corresponding universe Ω of all possible combinations is large. Therefore, we drew a subset of $M = 1000$ samples from Ω without replacement. Finally, we evaluated the associated awake prediction performances $B\bar{A}cc^i$, $i \in [1, M]$. Using this null distribution statistic, we computed a one-tailed empirical p-value for $B\tilde{A}cc$ by looking at the proportion of values less than or equal to the observed value when all GNW-related connections were removed[29].

Results

Model evaluation

In our experiments, the final number of epochs varied in the interval $[159, 1359]$ when the early stopping criterion was applied. The variational dropout in the sVAE selected 15 of the 32 latent dimensions (see Appendix S1). To evaluate whether a low-dimensional VAE can achieve reasonable assumptions, we compared several generative models with different parameters ($PPCA_1$, $PPCA_2$, $PPCA_3$, $PPCA_{15}$, VAE_1, VAE_2, VAE_3, and $sVAE_{15/32}$). This allowed quantification of how the latent representations stratify the brain patterns and how a model can reconstruct the input dFCs from the low-dimensional representations.

Balancing reconstruction quality and regularization By looking at the SSIM for all models (Figure 3-A), we observed that i) the chosen β had little effect on the VAE reconstructions, but ii) decreasing the latent space dimension degraded the reconstruction, and iii) nonlinear low-dimensional models had higher reconstruction quality. Higher dimensional models are expected to perform better because they capture more variability, resulting in a better reconstruction. We further quantified the decrease in reconstruction quality by comparing the VAE_2 with the $sVAE_{15/32}$. The cost of a low-dimensional, more interpretable model was approximately a 10% decrease in SSIM. It also appeared that a nonlinear model can reconstruct better with fewer latent dimensions. Monitoring the SSIM brain pattern-wise also showed that not all brain patterns were reconstructed similarly with a SSIM in the $[0.3, 0.9]$ range (Figure 3-D). Interestingly, the reconstruction quality increased with the number associated to each brain pattern. Thus, the models reconstructed more accurately the brain patterns closer to the structural connectivity with a simpler topology.

Balancing classification accuracy and regularization Monitoring the classification accuracy (Figure 3-B), we observed that the BAcc i) increased as β increased, ii) decreased as the latent space dimension decreased, except in high dimensions (i.e., for the $sVAE_{15/32}$), and that iii) the linear PPCA baseline outperformed other models in high dimensions ($PPCA_{15}$). Overall, the classification scores were relatively high for a seven-class classification problem. For all considered models, the BAcc scores ranged from 0.45 to 0.75 (to be compared to the theoretical chance level of 0.14). The classification accuracy metric favored the use of the highest regularization parameter ($\beta = 20$), which promotes coherence in the latent space. Furthermore, better performance (an increase of 6%) and lower interfold variance were observed for the 3D VAE (VAE_3) models. Notably, the $sVAE_{15/32}$ performed poorly, suggesting that a few latent dimensions were preferable to encode the brain pattern information.

Balancing reconstruction and classification As the number of latent dimensions increased, the model captured more variability (possibly noise). Furthermore, limiting the number of latent dimensions improved the brain pattern detection task. This trend confirmed that dFCs reflected the interplay of a small number of latent processes[2,10]. Looking at the consensus metric (Figure 3-C), we specified the following model for the rest of the paper: a 2D VAE (VAE_2) with a $\beta = 20$ regularization parameter. Using these parameters, we enforced a trade-off that imposed spatial coherence in the latent space without significantly degrading the reconstruction quality. Finally, we showed that the reconstructed brain patterns (as the reconstructed dFCs averaged over the different brain patterns) recovered the dominant structures obtained with a k-means clustering of the dFCs (Figure 4). The same model evaluation can be performed using arousal conditions as labels (see Appendix S2). To clarify the notation, the selected β_{20}-VAE_2 will be referred to as VAE in the following.

Latent space exploration

To investigate the potential of latent representations to decode states of consciousness, we considered two types of descriptors: discrete and, by exploiting the generative properties of VAEs, continuous latent representations. Again, we focused on the

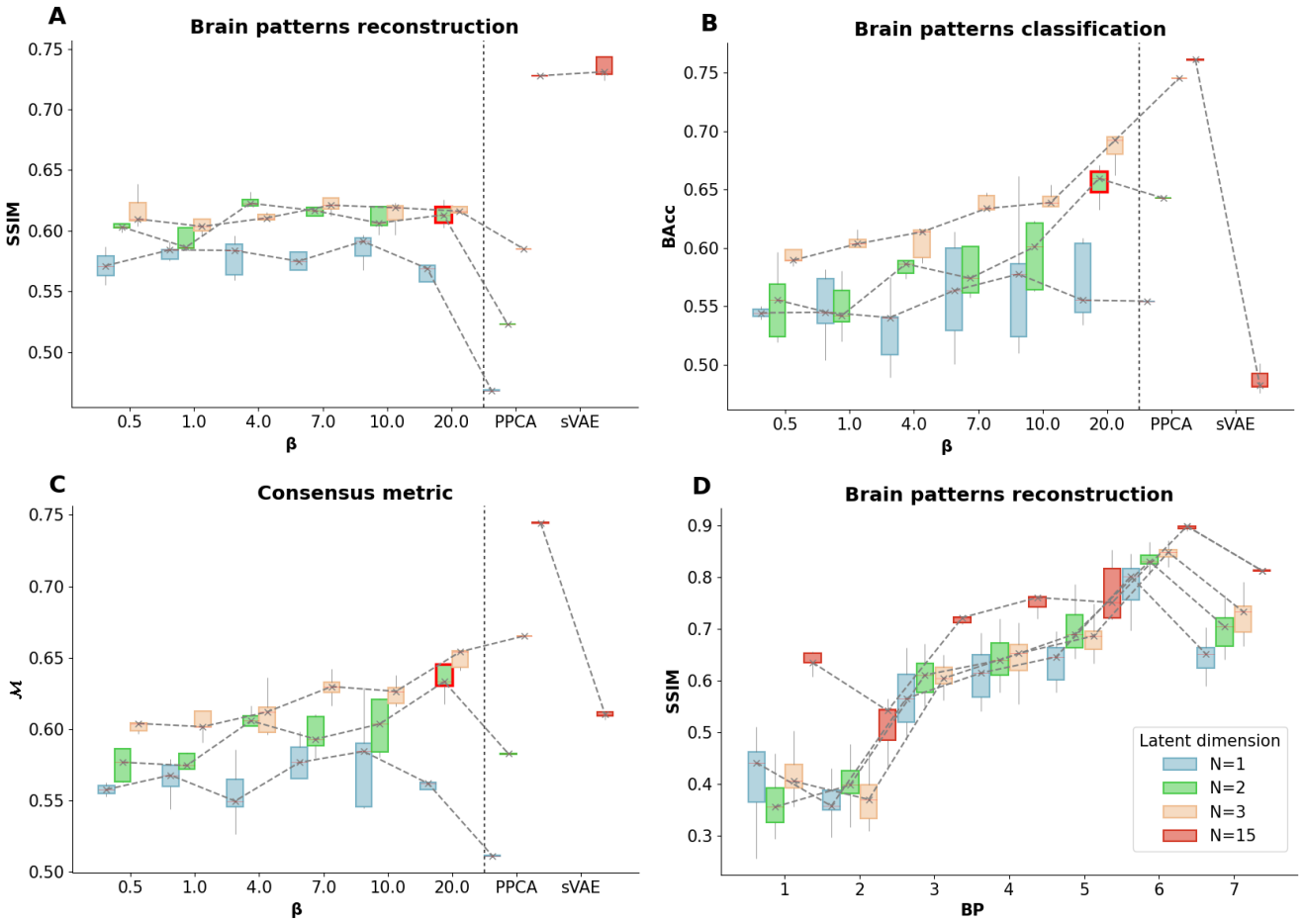


Figure 3. Brain pattern (BP) classification/reconstruction using VAE, PPCA and sVAE models: A) the structural similarity (SSIM) of BP-wise averaged dFCs with respect to the model parameters, B) the balanced accuracy (BAcc) between the ground truth and the matched predicted label, C) the proposed consensus metric $\mathscr{M}$, and D) the SSIM recorded for each BP. In plots A, B and C, the selected VAE_2 is highlighted by a red bounding box. The dashed lines represent the trends obtained for each latent space dimension across the considered models.

stratification of latent representations according to brain patterns. We also considered the reliability of the generated continuous descriptors.

Stratification of brain patterns From the VAE encoder, we obtained discrete latent representations. The ground truth labels were the brain patterns ranked in ascending order of similarity to the structural connectivity (numbered from 1 to 7). We examined the discrete composition of the latent space using the brain pattern labels (Figure 5-A) and the calculated lifetime (Figure 5-B). The lifetime is defined as the time spent continuously in a brain pattern (i.e. when no transition is observed). Therefore, all dFCs on this time axis have the same lifetime. Our focus was on three main properties of latent space. First, the resulting discrete representations formed a cloud of points rather than a set of clearly separable clusters. Second, the generated latent representations were remarkably well stratified when looking at the brain pattern labels (Figure 5-A). Each brain pattern was isolated while no constraint is enforced during training. To quantify the overlap between brain patterns, we chose the Dice similarity coefficient. The Dice metric yields values between 0 (no spatial overlap) and 1 (complete overlap)[30]. Overall, the average Dice metric remained relatively low ($< 0.37_{\pm 0.19}$), confirming that the spatial overlap between brain patterns was small (see Appendix S3 for details). Interestingly, brain pattern 7 (the one closest to the brain structure) occupied a central position in the representation space, and had the highest Dice coefficient. Third, the central locations, aligned with brain pattern 7, had longer lifetimes (Figure 5-B). Note that we verified the absence of subject bias prior to analysis, and also illustrated the stratification of the learned latent space with respect to the acquisition conditions (see Appendix S4). We also verified that the proposed VAE reliably encoded the dFC time courses while no constraint was enforced during training (see Appendix S5).

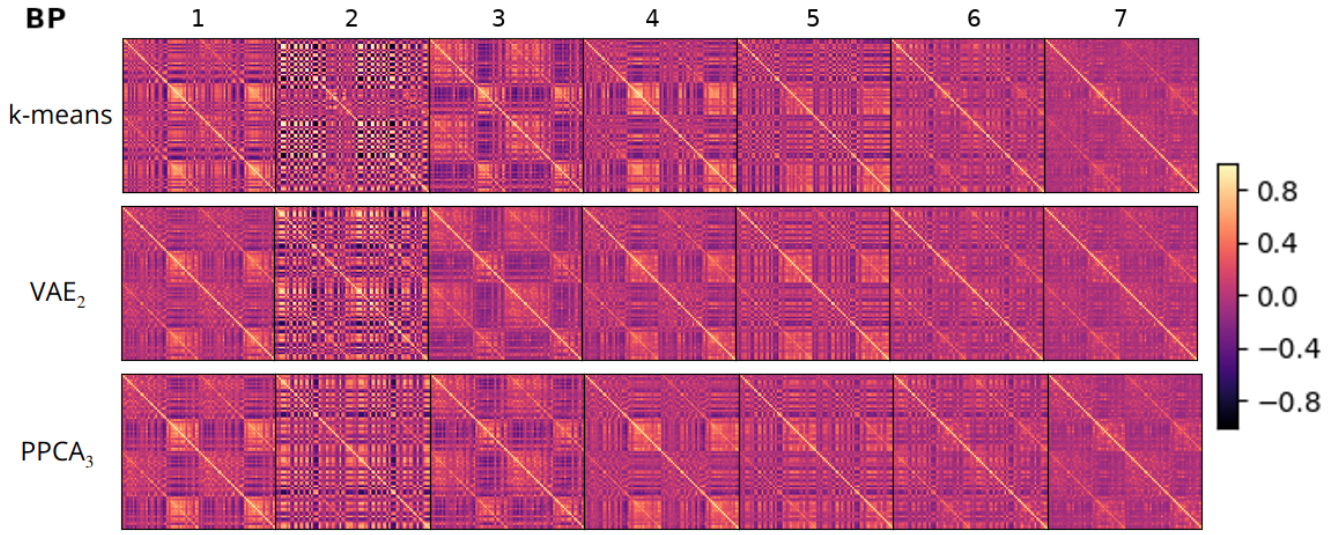


Figure 4. The reconstructed brain patterns (BPs) from the k-means clustering and the low-dimensional (2D or 3D) models maximizing the consensus metric $\mathscr{M}$: VAE_2 and $PPCA_3$.

Toward a whole-brain computational model By exploiting the generative capabilities of VAE, we obtained semantically continuous representations in the latent space, which promoted versatility. The generated continuous brain pattern labels covered the entire latent space. They also showed a pooled organization of the brain patterns (i.e., each brain pattern was mostly composed of a single connected component) (Figure 6-A). The accuracy of the brain pattern matching process was measured by the confidence $\mathscr{CM}$ and the reliability $\mathscr{RM}$ maps. Interestingly, the most strinking trend was that brain pattern boundaries were less reliable than central locations (Figure 6-C and D). With these maps, we gained confidence in using continuous descriptors in the latent space. Finally, decoding dFCs on a 19×19 regularly sampled grid in the latent space highlighted the learned manifold structure. It noteworthy exhibited brain patterns gradient toward the origin (Figure 6-B). Overall, the generated low-dimensional representations captured dynamic signatures of fluctuating wakefulness.

Connection-wise simulations

We used external perturbations to further annotate the representation of different states of consciousness. To this end, we first studied the shift in latent space induced by modifying a single connection of a dFC matrix. Using receptive field analysis, we could identify preferred directions for moving from one state to another. Second, we proposed an ablation analysis to ensure that dimension reduction preserves critical information about consciousness. For the latter, specific connections related to the regions highlighted by one of the major theories of consciousness (the GNW) were zeroed, and the induced displacement in latent space was examined.

Perturbation of connections to study transitions Using connection-wise RF analysis, a tensor $\mathscr{E}$ was estimated at each latent space location. We proposed to focus on seven specific latent space locations that were obtained when encoding the seven brain patterns with the VAE (see central plot in Figure 7). From each obtained tensor, we characterized the overall potential for action (i.e., the chance of generating a brain pattern transition) by the mean diffusivity (MD) (obtained by averaging the tensor eigenvalues). We found that this potential for action was always present but was small, lying in the interval [0.0068, 0.023]. Nevertheless, all tensors obtained were anisotropic. Thus, it was possible to select the connections with the highest probability of generating a brain pattern transition. In this study, we kept twenty connections (see circular plots in Figure 7). Interestingly, the MD for BP_7 was minimal, making it a "stable" pattern (i.e., a perturbation of this pattern is unlikely to cause a shift in consciousness).

Ablation of connections for virtual experiments A connection-wise ablation study showed an apparent decrease in BAcc in wakefulness prediction when the GNW-associated connections were removed ($\tilde{BAcc} << BAcc$ in Figure 8). As mentioned above, the zeroed connections involved brain regions that were considered part of a key cortical network for consciousness. We verified the significance of this decrease compared to random connection-wise ablations ($p_{val} = 0.008$). Thus, we showed that a realistic state transition can be obtained by modulating a network involved in consciousness. The connection-wise ablation study highlighted the relevance of the information captured in latent representations and supported the ability of the trained

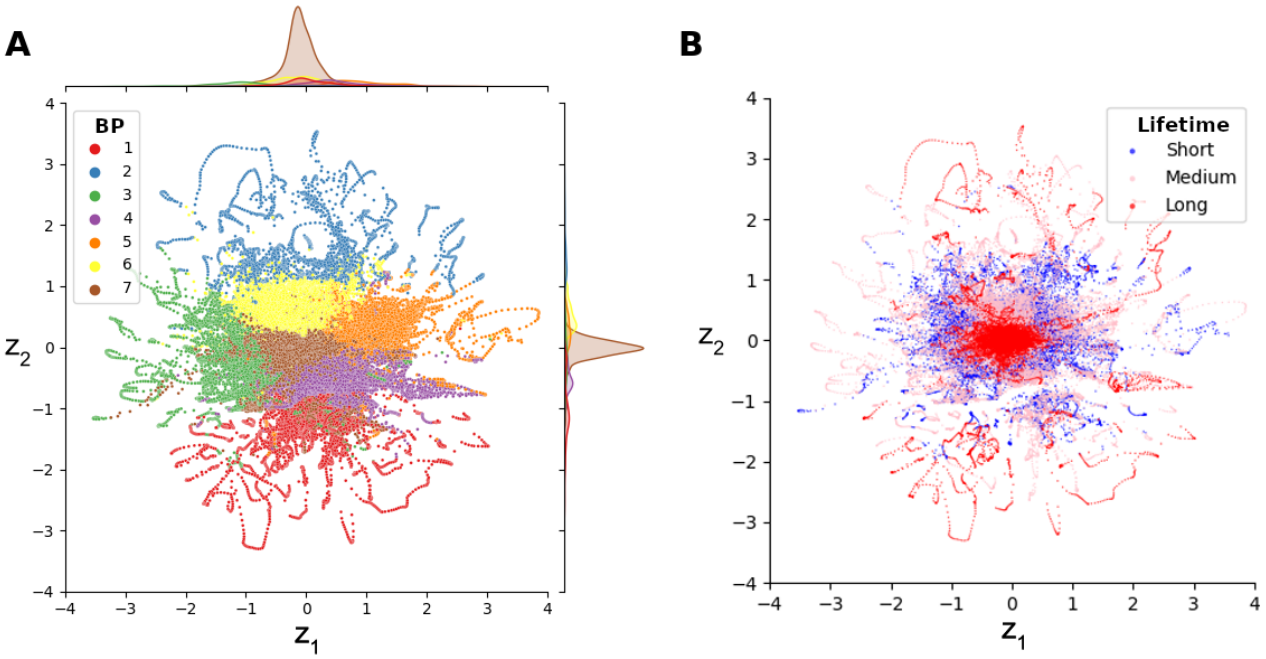


Figure 5. Discrete stratification of the latent space of the selected VAE into a base of A) Brain Patterns (BPs) - the centroids from a seven-class k-means clustering on the dFCs and B) lifetimes - the time spent continuously in the corresponding brain pattern. For the lifetimes, we discretize the values into three categories: the 25% longest (in red), the 25% shortest (in blue), and all others medium (in pink).

VAE to be an attractive computational model to decode and predict states of consciousness. To support this claim, we performed additional virtual ablation experiments (see Appendix S6).

Discussion

We proposed the VAE-VIENT framework as a tool to decode consciousness-related brain patterns from brain activity, to visualize their organization, and the transitions within the patterns that underlie states of consciousness. A VAE generative model has already been used to capture the different states of consciousness in a low-dimensional latent space. Here, we showed that such a model with tailored low-dimensional representations can be used to characterize brain dynamics over the dFCs. With low 2D-dimensional representations, the obtained performances were better than other linear (here, the PPCA) and nonlinear (here, the sVAE) generative models. However, this trend was not confirmed in higher dimensions (especially in 15D). It is generally accepted that simple models of neural mechanisms can be remarkably effective. We showed that a 2D VAE model could i) generate a latent feature space stratified into a base of brain patterns, and ii) reconstruct new brain patterns coherently and stably despite the limited dataset size by exploiting the generative part of the model. We argued that the VAE-VIENT framework provided a simulation-based whole-brain computational model. Indeed, we showed that the tensor fields generated from the RF analysis could model brain pattern transitions and that the proposed ablation analysis provided a unique way to non-invasively select target connections/regions. These findings paved the way for medical applications such as depth of anesthesia monitoring, coma characterization, and accurate diagnosis of disorders of consciousness in patients.

Dataset and preprocessing limitations This study has two major limitations. First, our dataset is relatively small, which increases the risk of overfitting. It will be necessary to perform tests with larger cohorts to validate our observations as studies of sleep and disorders of consciousness in humans[2,6,31] even if these datasets are also limited. However, they can be useful for validating the model with the ultimate goal of clinical translation. Second, we were working with sliding windows-based dFCs and not directly with time series. The former introduces hyperparameters that are not always easy to optimize[32–34]. Nevertheless, from a methodological point of view, we believe that working with sliding windows acts as a natural augmentation scheme that helps during the deep learning training on our limited dataset (5 monkeys - 156 runs - 72384 dFCs). Moreover, from a neuroscientific point of view, we aim to adhere to the dynamic representations of the brain originally described with dFCs[3,4]. Note that identical conclusions have been reached in humans, using a phase-based dynamic functional coordination

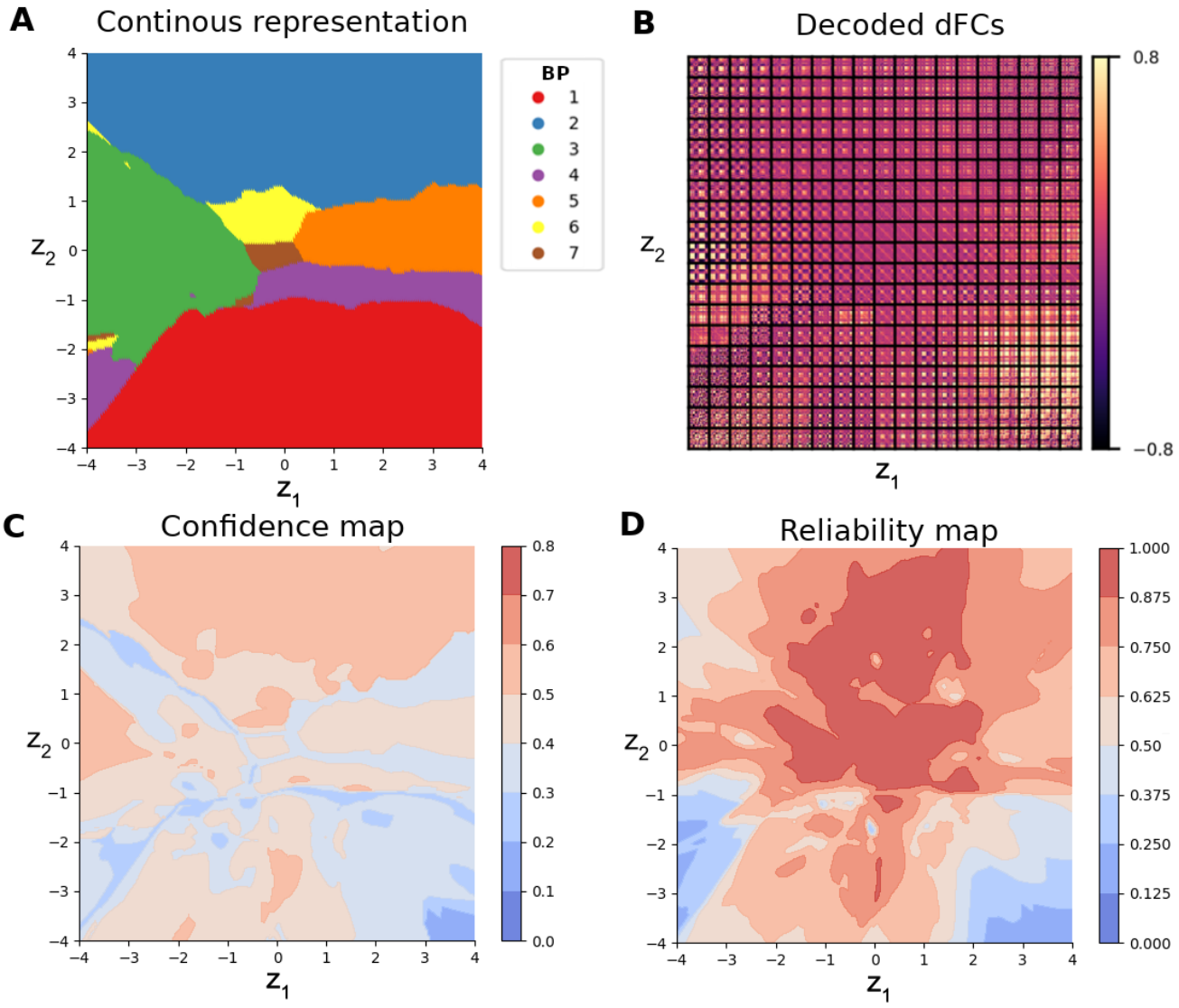


Figure 6. Continuous stratification of the latent space of the selected VAE and corresponding confidence and reliability maps: A) continuous representation of the Brain Patterns (BPs), B) decoded dFCs from a regularly sampled 19×19 grid in the latent space, C) estimated confidence map $\mathscr{CM}$, and D) estimated reliability map $\mathscr{RM}$.

analysis, suggesting only a small bias (if any) induced by sliding windows[6].

Repertoire of brain patterns and arousal levels Strikingly, the dimensional reduction with a 2D VAE could preserve the information that is related to each brain patterns. Furthermore, the comparison of brain patterns, using Pearson correlation similarity, clearly showed that similar brain patterns in the input space have closer latent representations (see Appendix S3, and Figure 6-A). Looking at the last row of the correlation matrix between brain patterns, we saw that BP_7 is highly correlated with BP_3, BP_4, BP_5, and BP_6. These patterns are also direct neighbors in latent space. Conversely, the less correlated BP_1, and BP_2 are not direct neighbors of BP_7 in the latent space. Thus, the global structure of brain patterns can be revealed by the latent space. Moreover, the performance of brain pattern classification was better than that of arousal level classification (awake vs. anesthetized) (see Appendix S2). We observed a 5% increase in classification performance. Given the difficulty of the task (i.e., a 7-class classification problem vs. a binary classification problem), the model seems to focus on the dynamic information shared between arousal levels. Similar conclusions were reached in our previous work[3,4], where the brain pattern repertoire was described as a set of brain configurations that are unevenly distributed across arousal levels. In other words, compared to arousal levels, brain patterns provide a more detailed description of states of consciousness. On the one hand, this property

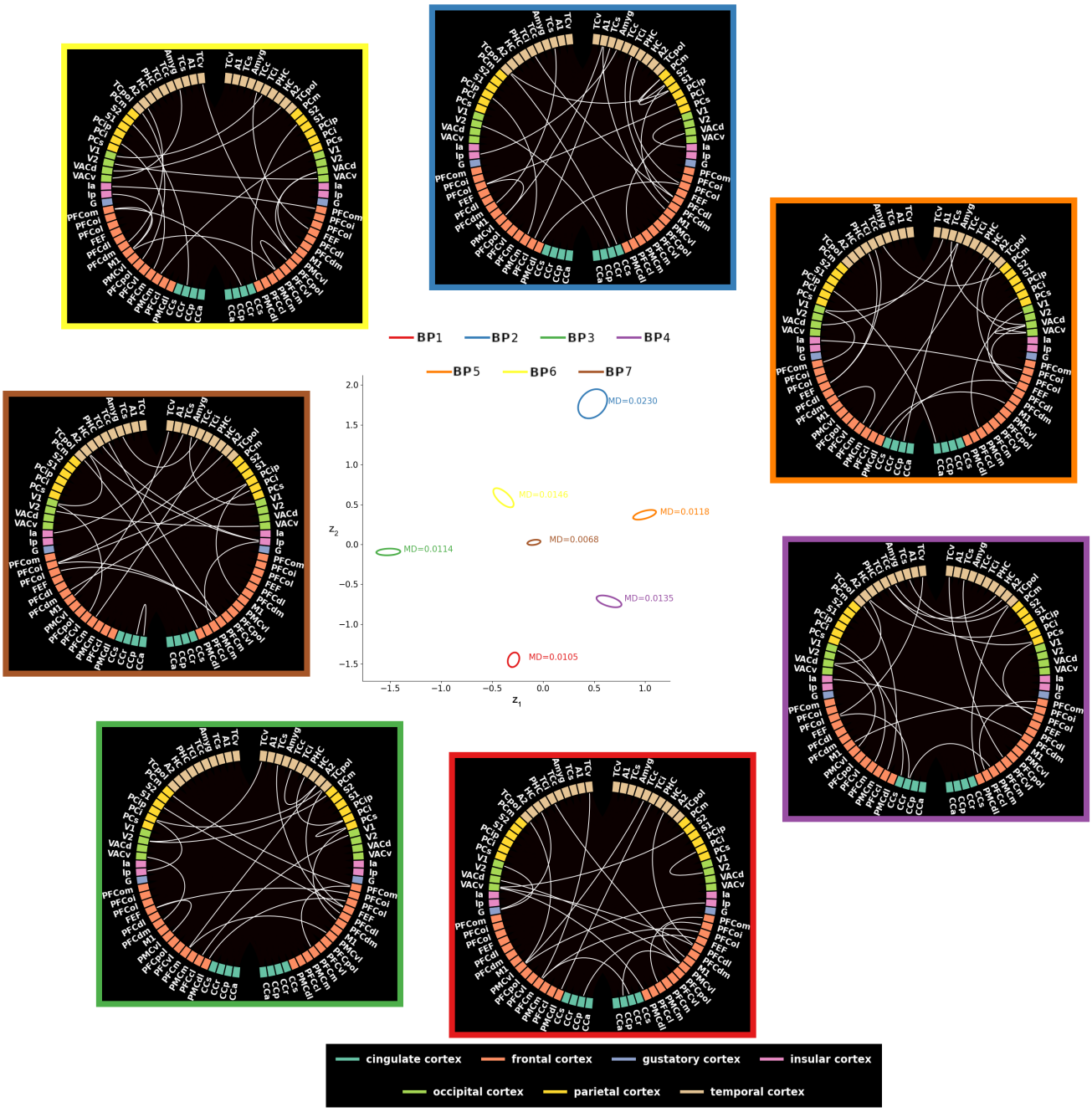


Figure 7. Results of RF analysis of the seven brain patterns and associated connections with a high potential for action. Using the proposed connection-wise RF analysis, a local perturbation model computed as an ellipse is derived at each encoded latent space location. Note that to improve readability, each ellipse is scaled. The associated mean diffusivity (MD) is calculated. For each ellipse, the twenty connections that cause the most displacement in the latent space are displayed using a circular layout.

may be inherited from the nature of the input data. Indeed, dFCs can be directly associated with changes in consciousness over time[2,4]. On the other hand, the difference between levels of sedation (deep and moderate) in the present dataset is small (i.e., only a difference of one level on the monkey behavioral scale)[4]. Such a difference results in changes in reflexes (toe pinch, corneal reflex, shaking) but not in voluntary behavior (response to juice presentation). Therefore, establishing a direct relationship between a subject's level of sedation and his or her level of consciousness may be a more difficult task than characterizing overall brain dynamics. In addition, previous studies on the same dataset have shown that all three anesthetics (propofol, sevoflurane, ketamine), despite different pharmacological molecular mechanisms, implied the same dynamics of

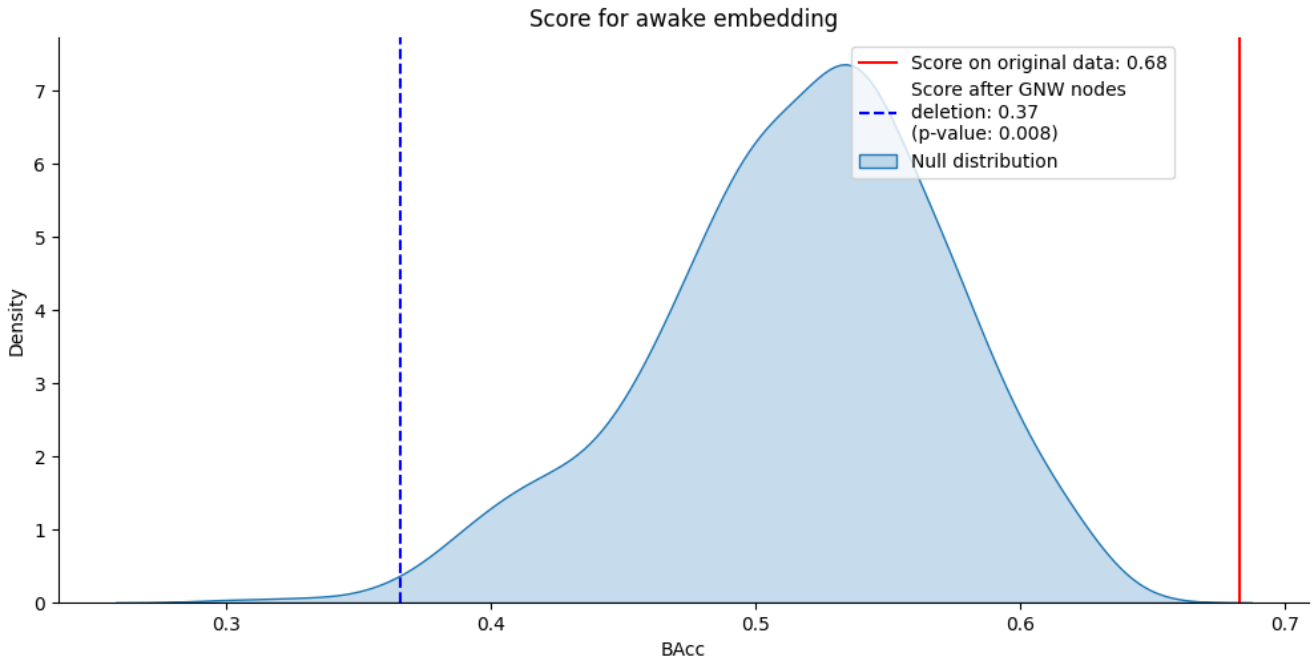


Figure 8. Ablation study performed from the GNW nodes. We evaluate the performance of a trained SVM classifier in predicting the awake state using the balance accuracy (BAcc). As input, we take only the raw or perturbed awake dFCs. We denote the corresponding prediction scores as $BAcc$ (vertical red dot line) and $B\tilde{A}cc$ (vertical blue dot line), respectively. We also display the histogram of $B\bar{A}cc^{i}$ when random connections are removed.

cortical activity measured with dFCs[4]. Thus, it remains to be seen whether we are unable to separate the different levels of consciousness because our data do not contain this information or because our modeling is inadequate.

Temporal modeling The main limitation of the current model is its inability to explicitly model the time course of the cerebral dynamics. We worked with dynamic FC matrices, but did not consider their order in each run. However, inspired by *Tseng et al. (2020)*[35], we investigated how temporal information was encoded by a 2D VAE model (see Appendix S5). Remarkably, the VAE-encoded latent variables had a coherent temporal structure that exhibited transitions characteristic of consciousness, even though no constraint was imposed during training. Other important features are the time spent consecutively in each pattern (previously called lifetime), the frequency of these steady states, and the associated transitions. Interestingly, and as described in the literature[3], the brain pattern closest to the structure (BP_7) was the most stable pattern with the longest lifetime. The latter also occupied a central place in the latent space around which other states were organized. The average lifetime was also significantly higher in the awake state than in all other anesthetized states. Similarly, the number of transitions was higher in the awake state than in all other anesthetized states. Furthermore, there was almost no difference in the number of transitions between different levels of anesthesia or between different anesthetics. In order to interpret such results with more confidence, a time-dependent model seems essential. In the literature, some works have proposed modeling a time series with a VAE, where the encoder and decoder consisted of LSTMs[36]. Other works abandoned the generative property and the decoder. For example, CEBRA is a contrastive learning technique that allows label-informed time series analysis[37]. CEBRA jointly uses auxiliary variables and neural data in a hypothesis-driven manner to generate consistent, time-aware latent representations. In all cases, the goal remains the same: to obtain a consistent picture of the latent space that drives activity and behavior. In future work, we plan to directly consider the time course in the learning phase.

Performing virtual experiments An interesting finding was the ability of the VAE model to simulate shifts in states of consciousness induced by selective virtual ablation of connectivity between pairs of brain areas, or even ablation of connectivity within a larger network. Historically, ablation techniques have been used in animal models to directly test the function of brain areas. For example, ablation techniques have directly linked vision to the occipital lobe and auditory function to the temporal lobe[38,39]. However, physical ablation/deactivation techniques are either irreversible or invasive and lack spatial resolution and specificity, highlighting the need for virtual ablation capabilities through the development of brain simulators[40]. Very few studies have been able to simulate the deactivation of global brain networks with the goal of suppressing consciousness. Here we presented a model capable of simulating a virtual experiment in which deactivation of the "macaque GNW network" leads

to suppression of consciousness. We believe that this simulation strengthens the capabilities of the model and opens up further virtual experiments that can, for example, test the specific effects of brain stimulation on consciousness[41].

Towards new biomarkers of consciousness The 2D VAE model demonstrated its ability to retain information about regions involved in conscious processing, showing that disruption of the "GNW nodes" causes a switch from a conscious to an unconscious state. It should be noted that only virtual inactivation of the entire "GNW network" (and not inactivation of individual node-related connections, see Appendix S6) caused a consciousness transition. We focused on the GNW theory of consciousness in this study because it has been translated from humans to monkeys[27]. In future work, we plan to explore other frameworks of consciousness, such as the Integrated Information Theory (IIT)[42,43]. We can also imagine testing other networks simply by trial-and-error simulations. Setting all links connected to GNW nodes to zero is one of the limitations of the proposed ablation simulation. In fact, it is not realistic to set all connections to zero, and perhaps certain connections should be privileged (using a weighted modulation of true connection values). Conversely, it would be of great interest to show the opposite effect, i.e., to find the regions that should be stimulated to switch from an unconscious to a conscious state. As suggested by[31], this goal is challenging and probably requires simulation. Further analysis of connection-wise RF latent space structure modeling will certainly be valuable in this context. In fact, the RF analysis related the different patterns that reflect the dynamics of the brain (biological markers). The ellipsoids obtained in our work described the most plausible connections to perturb in order to redirect trajectories and potentially restore wakefulness. Studying the sequences of different trajectories in latent space paved the way to a whole-brain computational model of conscious access. In other words, RF analysis provided the unique ability to identify the pairs of nodes involved in consciousness directly from the data. Changing one connection at a time is one of the limitations of the proposed RF simulation. In reality, changing multiple connections at the same time may have a more significant effect. Finally, we believe that the clinical and scientific applications are numerous. First, this approach allows the description of new biomarkers of consciousness. In addition, it is a unique tool to simulate the consequences of targeted modulation of specific brain regions for the loss or recovery of consciousness. In this context, we hypothesize that the latent space structure will be essential for dissecting the mechanisms of Deep Brain Stimulation (DBS) for disorders of consciousness and help to build a general predictive model of the global brain effects of DBS.

References

1. James, W. *The principles of psychology, Vol I.* The principles of psychology, Vol I. (Henry Holt and Co, New York, NY, US, 1890). Pages: xii, 697.
2. Allen, E. A. *et al.* Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex (New York, N.Y.: 1991)* **24**, 663–676, DOI: 10.1093/cercor/bhs352 (2014).
3. Barttfeld, P. *et al.* Signature of consciousness in the dynamics of resting-state brain activity. *Proc. Natl. Acad. Sci.* 19 (2015).
4. Uhrig, L. *et al.* Resting-state Dynamics as a Cortical Signature of Anesthesia in Monkeys. *Anesthesiology* **129**, 942–958, DOI: 10.1097/ALN.0000000000002336 (2018).
5. Preti, M. G., Bolton, T. A. & Van De Ville, D. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage* **160**, 41–54, DOI: 10.1016/j.neuroimage.2016.12.061 (2017).
6. Demertzi, A. *et al.* Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Sci. Adv.* **5**, eaat7603, DOI: 10.1126/sciadv.aat7603 (2019).
7. Perl, Y. S. *et al.* Generative Embeddings of Brain Collective Dynamics Using Variational Autoencoders. *Phys. Rev. Lett.* **125**, 238101, DOI: 10.1103/PhysRevLett.125.238101 (2020).
8. Misra, J. *et al.* Learning brain dynamics for decoding and predicting individual differences. *PLOS Comput. Biol.* **17**, e1008943, DOI: 10.1371/journal.pcbi.1008943 (2021). Publisher: Public Library of Science.
9. Gao, S., Mishne, G. & Scheinost, D. Nonlinear manifold learning in functional magnetic resonance imaging uncovers a low-dimensional space of brain dynamics. *Hum. Brain Mapp.* **42**, 4510–4524, DOI: 10.1002/hbm.25561 (2021).
10. Monti, R. P. *et al.* Decoding Time-Varying Functional Connectivity Networks via Linear Graph Embedding Methods. *Front. Comput. Neurosci.* **11** (2017).
11. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509, DOI: 10.1038/nn.3776 (2014). Number: 11 Publisher: Nature Publishing Group.
12. Kim, J.-H. *et al.* Representation learning of resting state fMRI with variational autoencoder. *NeuroImage* **241**, 118423, DOI: 10.1016/j.neuroimage.2021.118423 (2021).

13. Zhao, Q. *et al.* Variational Autoencoder with Truncated Mixture of Gaussians for Functional Connectivity Analysis. In Chung, A. C. S., Gee, J. C., Yushkevich, P. A. & Bao, S. (eds.) *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, 867–879, DOI: 10.1007/978-3-030-20351-1_68 (Springer International Publishing, Cham, 2019).

14. Bakker, R., Wachtler, T. & Diesmann, M. CoCoMac 2.0 and the future of tract-tracing databases. *Front. Neuroinformatics* **0**, DOI: 10.3389/fninf.2012.00030 (2012). Publisher: Frontiers.

15. Liu, R. *et al.* A generative modeling approach for interpreting population-level variability in brain structure. *bioRxiv* 2020.06.04.134635, DOI: 10.1101/2020.06.04.134635 (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results.

16. Qiang, N., Dong, Q., Sun, Y., Ge, B. & Liu, T. Deep Variational Autoencoder for Modeling Functional Brain Networks and ADHD Identification. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 554–557, DOI: 10.1109/ISBI45749.2020.9098480 (2020). ISSN: 1945-8452.

17. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat. Commun.* **12**, 5684, DOI: 10.1038/s41467-021-26017-0 (2021). Number: 1 Publisher: Nature Publishing Group.

18. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]* (2014). ArXiv: 1312.6114.

19. Higgins, I. *et al.* β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework. *ICLR* 22 (2017).

20. Burgess, C. P. *et al.* Understanding disentangling in β-VAE. *arXiv:1804.03599 [cs, stat]* (2018). ArXiv: 1804.03599.

21. Plum, F. & Posner, J. B. *The Diagnosis of Stupor and Coma* (Oxford University Press, 1982). Google-Books-ID: Pbl4CH4NlQsC.

22. Laureys, S. The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn. Sci.* DOI: 10.1016/j.tics.2005.10.010 (2005).

23. Demertzi, A., Laureys, S. & Boly, M. Coma, Persistent Vegetative States, and Diminished Consciousness. *Encycl. Conscious.* 147 (2009).

24. Antelmi, L., Ayache, N., Robert, P. & Lorenzi, M. Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data. In *International Conference on Machine Learning*, 302–311 (PMLR, 2019).

25. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *J. Royal Stat. Soc. Ser. B* **61**, 611–622 (1999).

26. Deco, G., Kringelbach, M. L., Jirsa, V. K. & Ritter, P. The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core. *Scientific Reports* **7**, DOI: 10.1038/s41598-017-03073-5 (2017).

27. Uhrig, L., Dehaene, S. & Jarraya, B. A Hierarchy of Responses to Auditory Regularities in the Macaque Brain. *J. Neurosci.* **34**, 1127–1132, DOI: 10.1523/JNEUROSCI.3165-13.2014 (2014).

28. Uhrig, L., Janssen, D., Dehaene, S. & Jarraya, B. Cerebral responses to local and global auditory novelty under general anesthesia. *NeuroImage* **141**, 326–340, DOI: 10.1016/j.neuroimage.2016.08.004 (2016).

29. Ojala, M. & Garriga, G. C. Permutation Tests for Studying Classifier Performance. In *2009 Ninth IEEE International Conference on Data Mining*, 908–913, DOI: 10.1109/ICDM.2009.108 (IEEE, Miami Beach, FL, USA, 2009).

30. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302, DOI: 10.2307/1932409 (1945). Publisher: Ecological Society of America.

31. Perl, Y. S. *et al.* Low-dimensional organization of global brain states of reduced consciousness. *Cell Reports* **42**, 112491, DOI: 10.1016/j.celrep.2023.112491 (2023).

32. Shakil, S., Lee, C.-H. & Keilholz, S. D. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *NeuroImage* **133**, 111–128, DOI: 10.1016/j.neuroimage.2016.02.074 (2016).

33. Savva, A. D., Kassinopoulos, M., Smyrnis, N., Matsopoulos, G. K. & Mitsis, G. D. Effects of motion related outliers in dynamic functional connectivity using the sliding window method. *J. Neurosci. Methods* **330**, 108519, DOI: 10.1016/j.jneumeth.2019.108519 (2020).

34. Mokhtari, F., Akhlaghi, M. I., Simpson, S. L., Wu, G. & Laurienti, P. J. Sliding window correlation analysis: Modulating window shape for dynamic brain connectivity in resting state. *NeuroImage* **189**, 655–666, DOI: 10.1016/j.neuroimage.2019.02.001 (2019).

35. Tseng, J. & Poppenk, J. Brain meta-state transitions demarcate thoughts across task contexts exposing the mental noise of trait neuroticism. *Nat. Commun.* **11**, 3480, DOI: 10.1038/s41467-020-17255-9 (2020). Number: 1 Publisher: Nature Publishing Group.

36. Lin, S. *et al.* Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4322–4326, DOI: 10.1109/ICASSP40776.2020.9053558 (IEEE, Barcelona, Spain, 2020).

37. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368, DOI: 10.1038/s41586-023-06031-6 (2023). Number: 7960 Publisher: Nature Publishing Group.

38. Panizza, B. Osservazioni sul nervo ottico. *Gior. I. R. Ist Lomb. Sci. Lett. Arti.* *7* 237–252. (1855).

39. Munk, H. OF THE VISUAL AREA OF THE CEREBRAL CORTEX, AND ITS RELATION TO EYE MOVEMENTS *. *Brain* **13**, 45–70, DOI: 10.1093/brain/13.1.45 (1890).

40. Fan, X. & Markram, H. A Brief History of Simulation Neuroscience. *Front. Neuroinformatics* **13** (2019).

41. Deco, G. *et al.* Awakening: Predicting external stimulation to force transitions between different brain states. *Proc. Natl. Acad. Sci.* **116**, 18088–18097, DOI: 10.1073/pnas.1905534116 (2019).

42. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **5**, 42, DOI: 10.1186/1471-2202-5-42 (2004).

43. Balduzzi, D. & Tononi, G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLOS Comput. Biol.* **4**, e1000091, DOI: 10.1371/journal.pcbi.1000091 (2008). Publisher: Public Library of Science.

Acknowledgements

We thank Morgan Dupont for help with animal experiments, Alexis Amadon, Hauke Kolster, Laurent Laribière, and the NeuroSpin MRI and informatics teams for help with imaging tools, Christophe Joubert and Jean-Marie Helies for animal facilities, and Jean-Robert Deverre for support.

Author contributions statement

B.J. and L.U. conceived and designed the experiments, B.J. and L.U. collected the data, C.G. and A.G. analyzed the data, A.G. and C.G. wrote the paper, and V.F. and E.D. contributed to the critical appraisal of the paper. All authors reviewed the manuscript.

Additional information

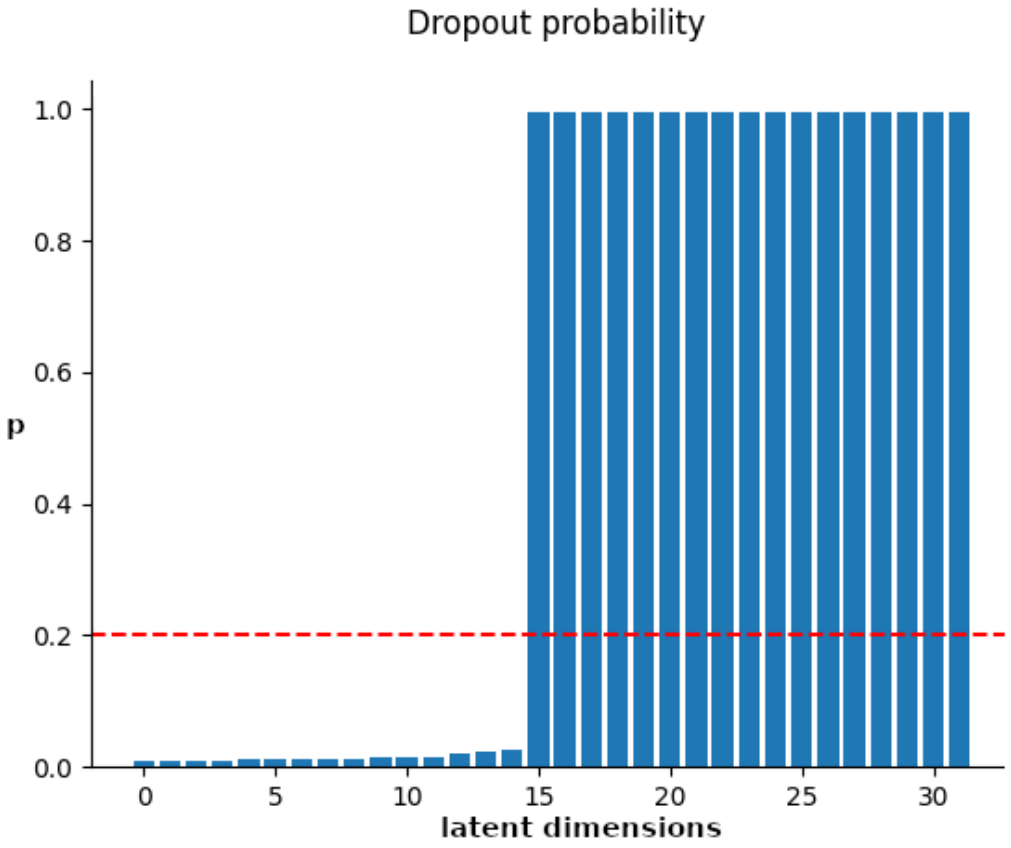
Accession codes All data and codes needed to evaluate the conclusions of the paper are publicly available at `https://zenodo.org/records/10423725`.

Competing interests The authors declare no competing interests.

Funding statement: This work was supported by Institut National de la Santé et de la Recherche Médicale, the Inserm Avenirprogram (B.J.), Commissariat à l'Energie Atomique, Collège de France, ERC Grant NeuroConsc (to S.D.), Foundation Bettencourt-Schueller, French National Research Agency for the project Big2Small (Chair in AI, ANR-19-CHIA-0010-01), the project RHU-PsyCARE (French government's "Investissements d'Avenir" program, ANR-18-RHUS-0014), and European Union's Horizon 2020 for the project R-LiNK (H2020-SC1-2017, 754907).

Appendix S1 Dropout regularization in sVAE training

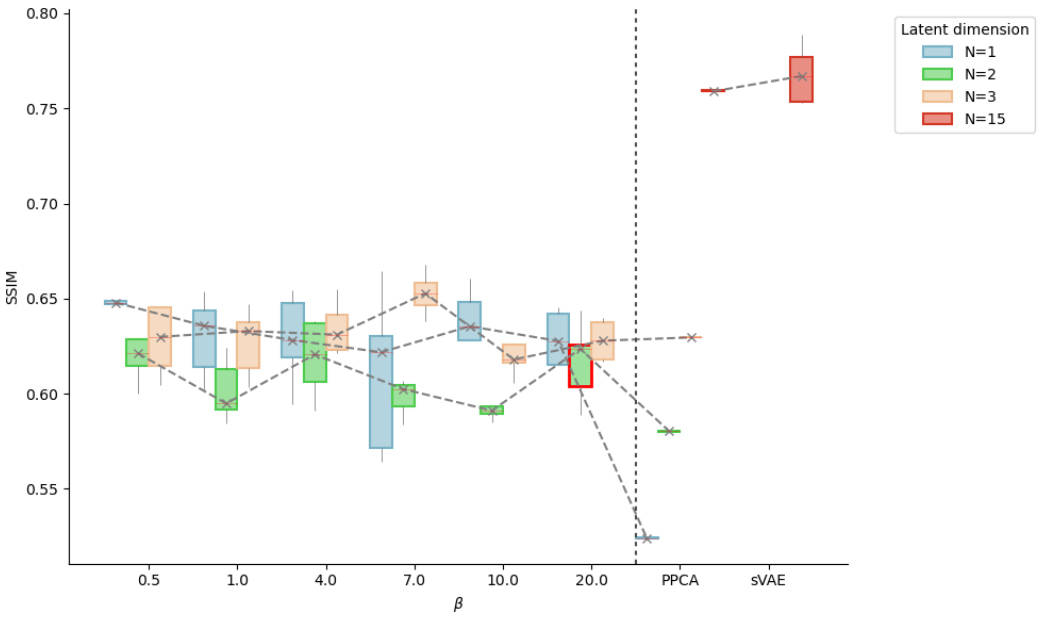
During sVAE training, parsimonious and interpretable representations are enforced by variational dropout. Model selection in latent space can then be achieved using this technique. Here, the dropout rate after convergence is shown when the initial latent dimensions are set to 32 (Appendix S1- Figure 1). Note that the learned dropout rate is highly contrasted. For this reason, model selection can be done by keeping the latent dimensions that meet a suitable dropout rate threshold. We can see that it is possible to safely select the best model with a threshold $p < 0.2$ (as proposed in the original paper[24]).



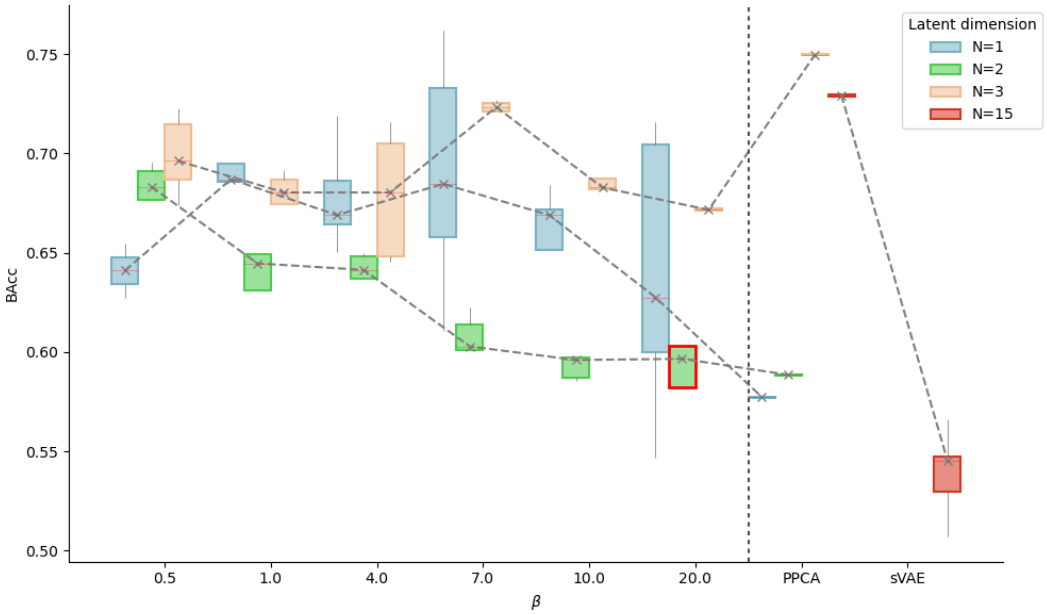
Appendix S1- Figure 1. sVAE estimated dropout rate.

Appendix S2 Model evaluation using arousal conditions

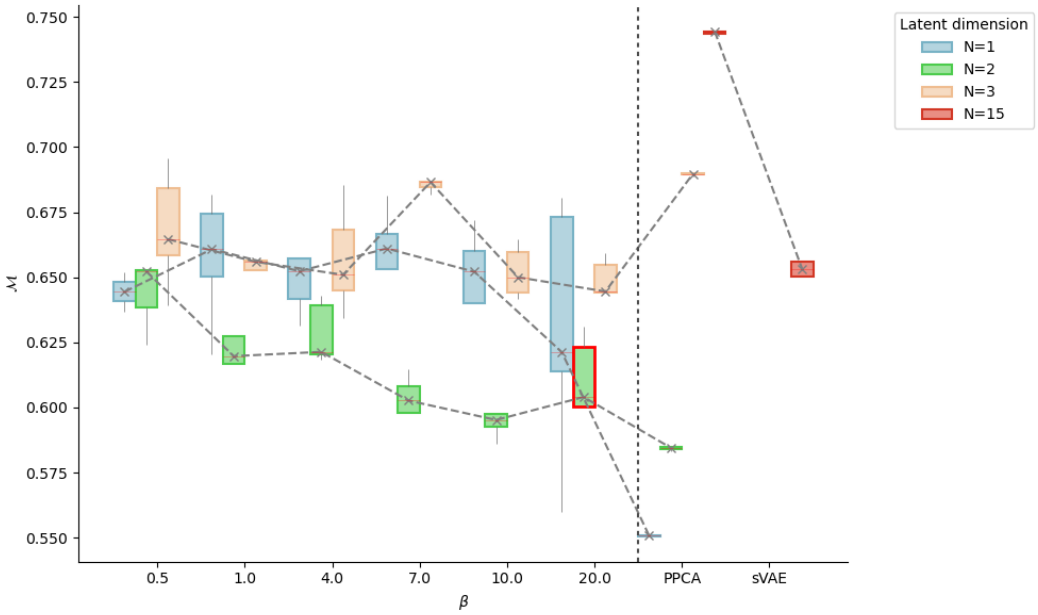
The proposed model evaluation relies on the brain pattern labels. Indeed, it has been shown that these labels can effectively represent the different configurations of the brain[3,4]. Here, we re-evaluate the reconstruction quality, classification accuracy, and consensus metric using a different set of labels composed of the arousal conditions. In particular, we address a binary classification problem between the awake and the anesthetized data (where all the associated acquisition conditions are grouped together). First, the dFCs are well reconstructed for all models when looking at the reconstruction quality (using the structural similarity (SSIM) metric) (Appendix S2- Figure 1). We note that i) for the VAEs, the chosen β has little effect on the reconstruction, ii) for the PPCA baseline, increasing the latent space improves the reconstruction, but this is not the case for the nonlinear models, and iii) the sVAE with the chosen fifteen dimensions performs best. Second, it does not seem trivial to classify a dFC matrix as belonging to the conscious or unconscious category. This is consistent with previous findings showing that dFC matrices of one condition can be associated with different brain patterns[4]. Finally, looking at the consensus metric, the least constrained VAE models ($\beta = 0.5$ and $\beta = 1$) perform best in low dimensions (Appendix S2- Figure 3). Beyond three dimensions, the PPCA stands out. This may be due to the "relative" simplicity of our task.



Appendix S2- Figure 1. VAE, PPCA, sVAE reconstruction quality: SSIM of label-wise averaged dFCs with respect to the β regularization parameters.



Appendix S2- Figure 2. VAE, PPCA, sVAE classification accuracy: BAcc between the ground truth and the matched predicted labels.



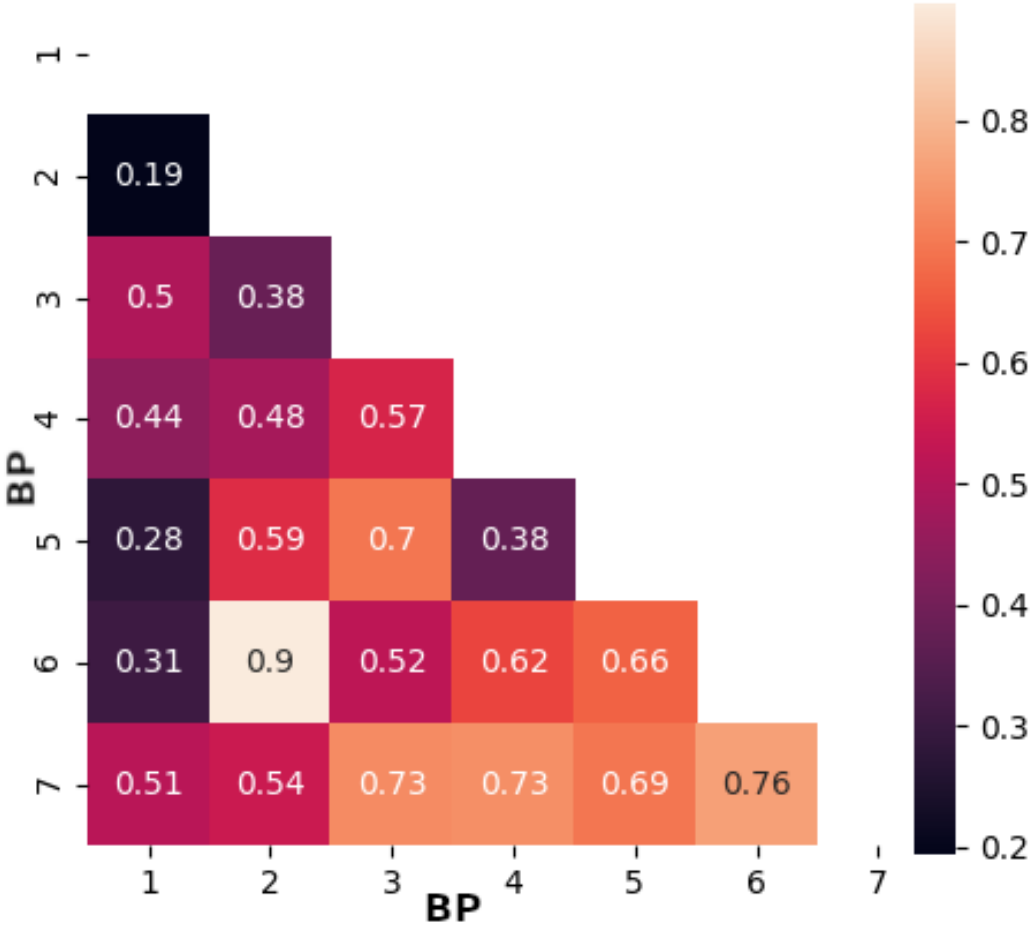
Appendix S2- Figure 3. The proposed consensus metric $\mathscr{M}$.

Appendix S3 Additional brain pattern analyses

The 2D latent representations obtained with a VAE_2 can be stratified using the brain pattern (BP) labels. To quantify the overlap between brain pattern locations in latent space, we use the Dice similarity coefficient (Appendix S3- Table 1). The Dice metric yields values between 0 (no spatial overlap) and 1 (complete overlap)[30]. Unfortunately, the Dice metric can only be applied to array-like data. Therefore, we choose to perform a brain pattern-wise regridding (using a 60 x 60 grid) of the obtained latent representations, which produces a binary array-like support per brain pattern (numbered 1 to 7). The Pearson correlation matrix between the brain patterns further describes the studied brain repertoire (Appendix S3- Figure 1). Overall, these two experiments allow us to better characterize the learned latent space in terms of brain patterns.

BP	1	2	3	4	5	6	7
Dice	$0.24_{\pm 0.17}$	$0.07_{\pm 0.11}$	$0.24_{\pm 0.15}$	$0.32_{\pm 0.20}$	$0.24_{\pm 0.16}$	$0.28_{\pm 0.08}$	$0.37_{\pm 0.19}$

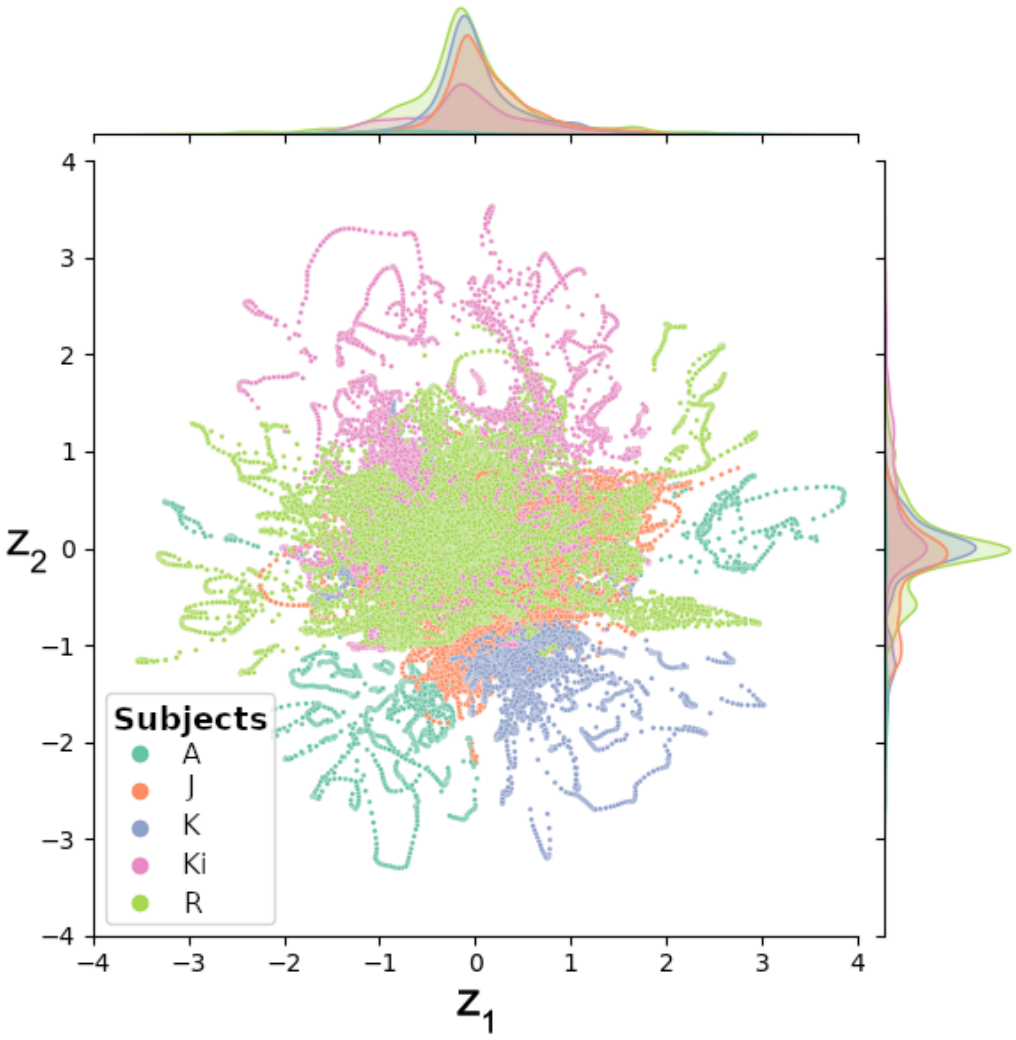
Appendix S3- Table 1. table
Averaged across folds Dice coefficients and associated standard deviations for each BP embedding.



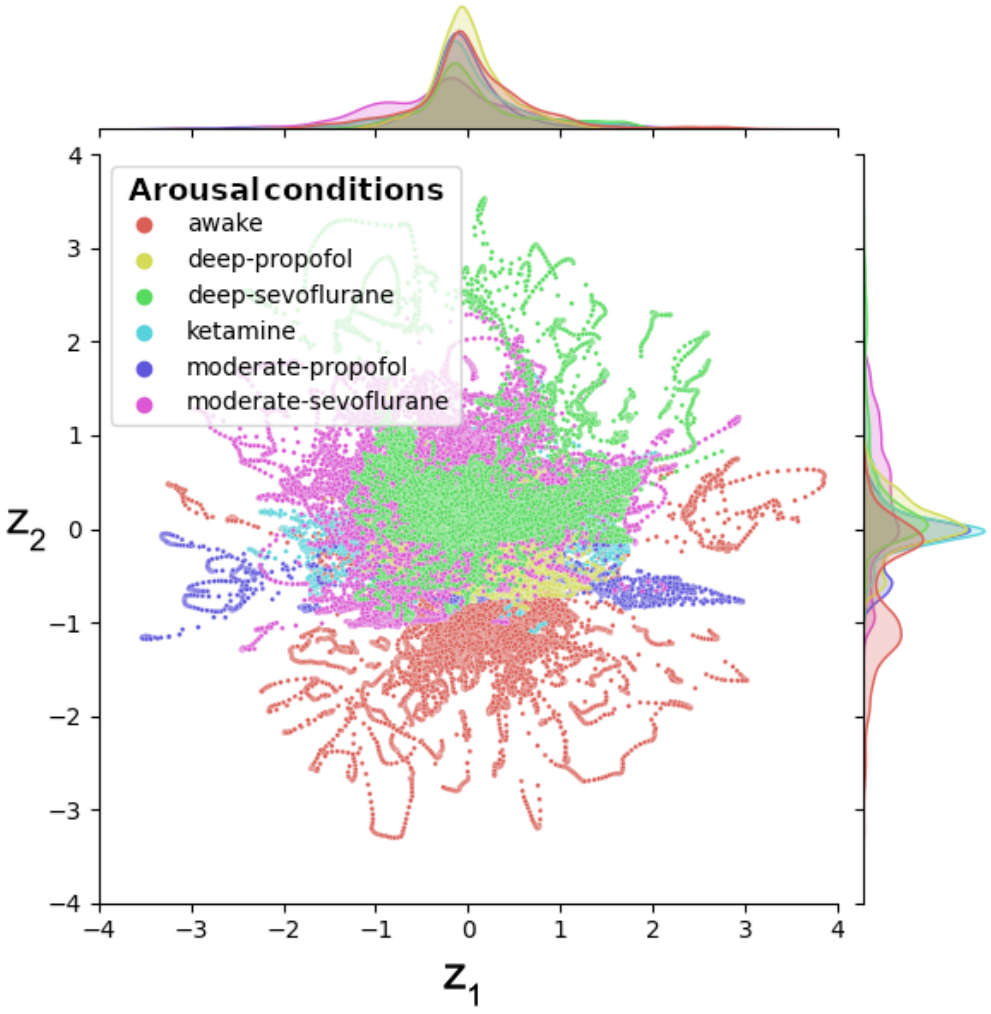
Appendix S3- Figure 1. Correlation matrix between brain patterns.

Appendix S4 Explore learned latent representations with different labels

The present study focuses on brain pattern labels. Indeed, it has been shown that these labels can effectively represent the different configurations of the brain[3,4]. Two different sets of labels are considered below. First, the subjects to investigate whether the VAE_2 has incorrectly learned subject-specific information (i.e., there is some overfitting during training in our small dataset) (Appendix S4- Figure 1). Second, the acquisition conditions to verify that no anesthetic effect (known to have different vascular effects) can be observed in the latent representations (Appendix S4- Figure 2).



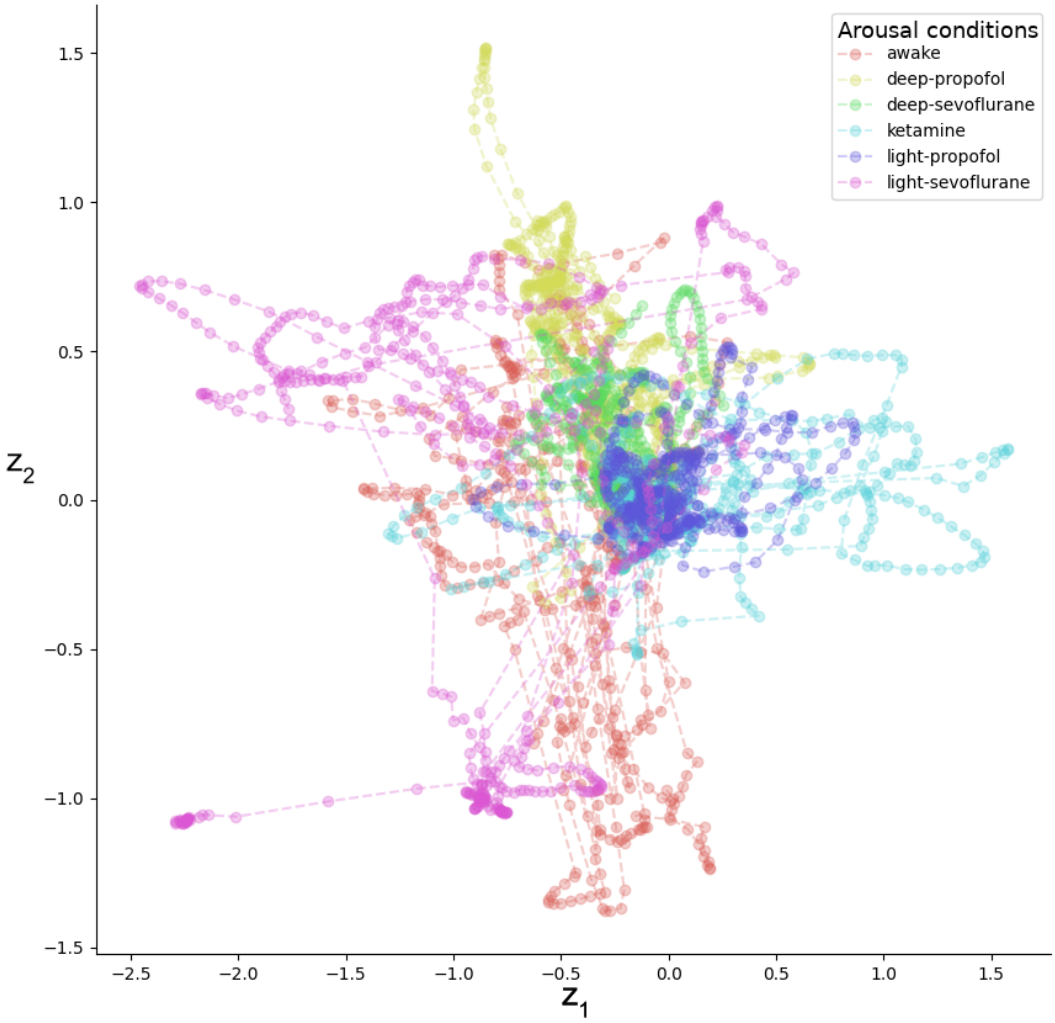
Appendix S4- Figure 1. Stratification of VAE_2 latent representations by subject.



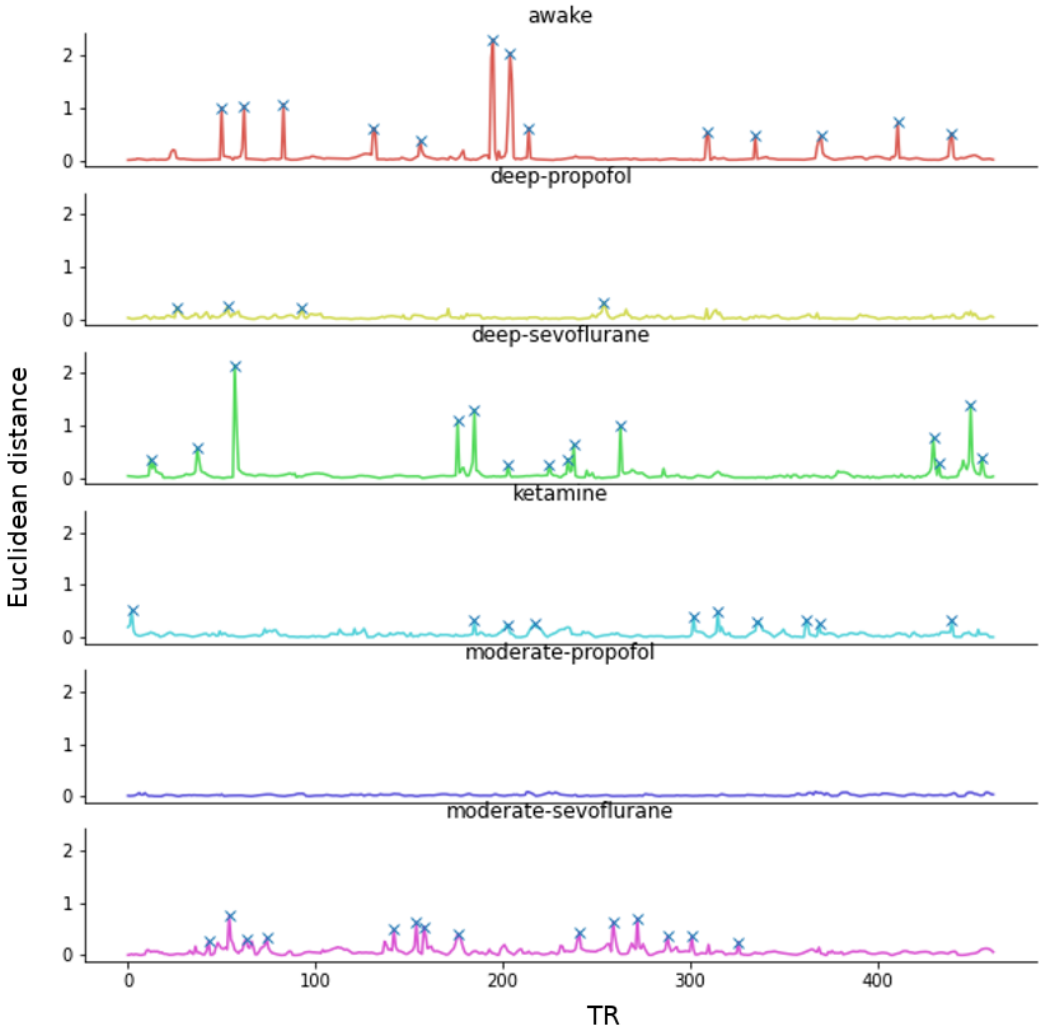
Appendix S4- Figure 2. Stratification of VAE_2 latent representations by acquisition conditions.

Appendix S5 Temporal structure encoded in latent space

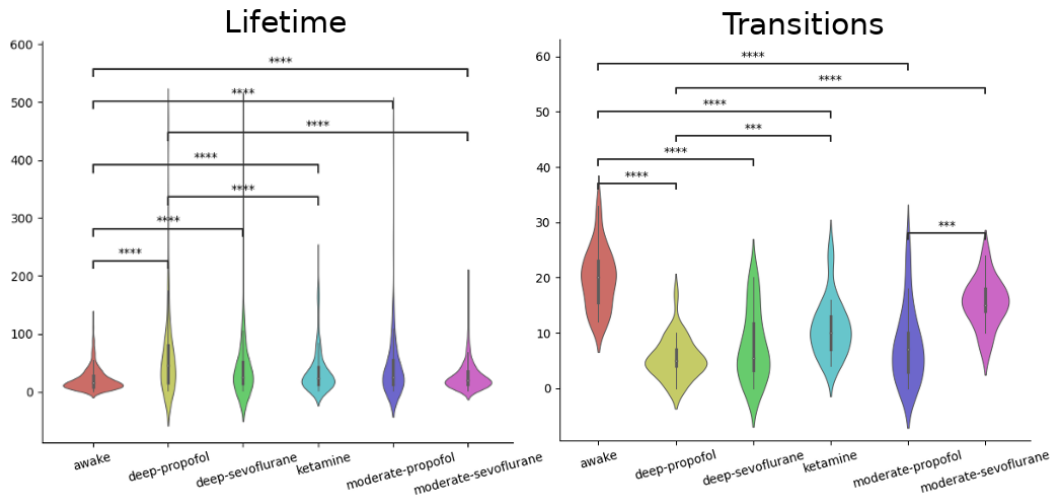
Following the idea of *Tseng et al. (2020)*[35] to study meta-transitions between brain configurations, we examine the encoding of temporal information using the selected VAE_2. While no modeling of the acquisition time course is enforced during training, the encoded latent variables have a coherent temporal structure (Appendix S5- Figure 1). This is a sign of successful modeling. It is then possible to examine the temporal transitions within the same run by calculating the Euclidean distance between each successive encoded time point (i.e., replaying the dFC movie) (Appendix S5- Figure 2). Stable periods have a Euclidean distance close to zero, and a transition occurs when a jump in the metric is observed (indicated by blue crosses). Transitions between brain configurations have stable periods, which we call meta-stable states. The average lifetime of a meta-stable state, i.e. the time spent continuously in this state, is significantly higher in the awake state than in all other anesthetized states (Appendix S5- Figure 3). Similarly, the number of transitions is higher in the awake state than in all other anesthetized states (Appendix S5- Figure 3). Interestingly, there is almost no difference between different levels of anesthesia or between different anesthetics.



Appendix S5- Figure 1. Latent representations of six runs, one per arousal condition.



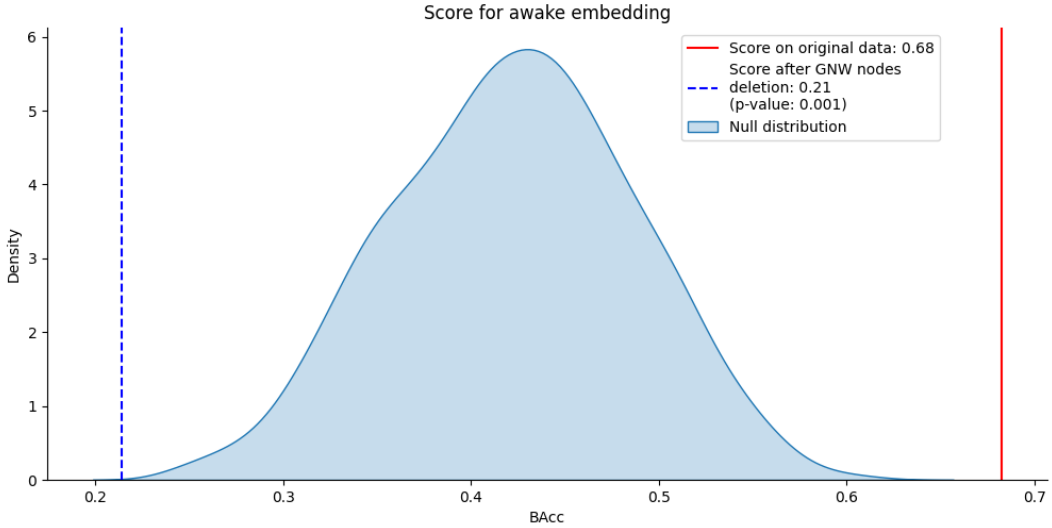
Appendix S5- Figure 2. The Euclidean distance between two consecutive time points for each run previously selected. Blue crosses mark transitions.



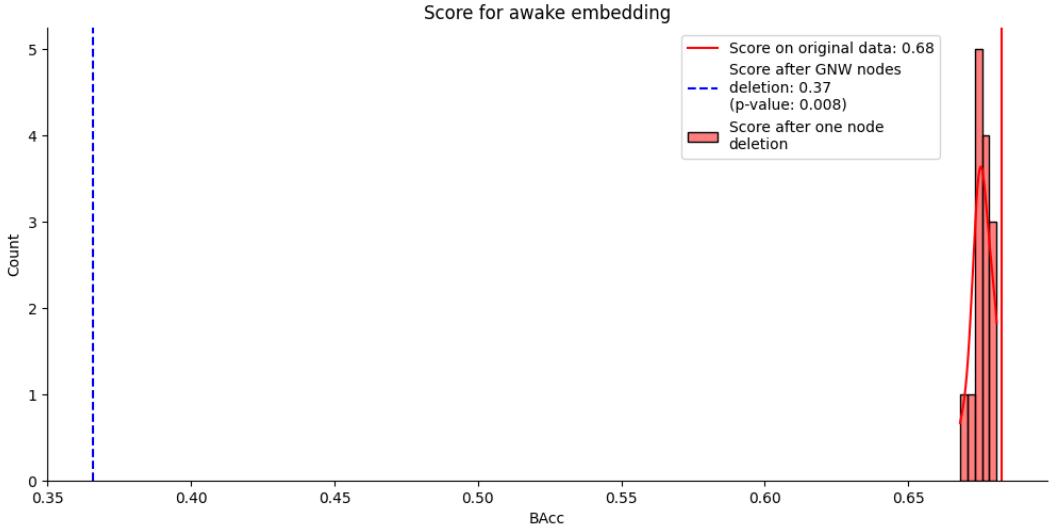
Appendix S5- Figure 3. Meta-stable state lifetime and transition occurrence distributions across acquisition conditions.

Appendix S6 Additional virtual ablation experiments

The proposed ablation analysis provides a unique way to non-invasively select target connections/regions. First, the GNW key regions are expanded to include primary sensory regions, primary somatosensory cortex S1, primary auditory cortex A1, and visual area V1. When the awake state is altered, the addition of these target regions in the ablation study further increases the statistical significance of the results (Appendix S6- Figure 1). However, perhaps only a few important regions drive the prediction performance. Thus, in the proposed analysis, instead of removing all GNW-related connections, only one GNW-related connection is removed. By using all combinations, the histogram of prediction performance is computed (Appendix S6- Figure 2). In conclusion, removing one GNW-related connection is insufficient to produce a significant shift in awareness.



Appendix S6- Figure 1. Ablation study considering GNW and sensory related connections.



Appendix S6- Figure 2. Ablation study considering GNW-related connections one by one. The resulting histogram of prediction performance is shown in red.