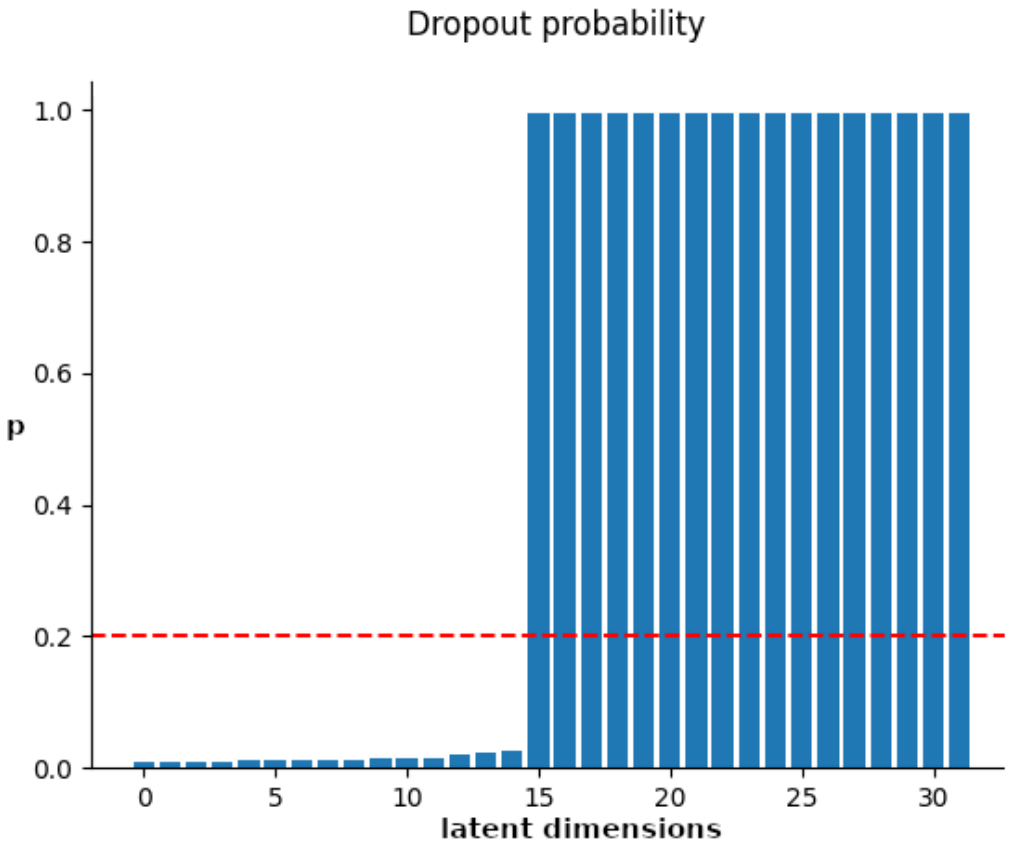
Supplementary information - Deep learning models reveal the link between dynamic brain connectivity patterns and states of consciousness

Appendix S1 Dropout regularization in sVAE training

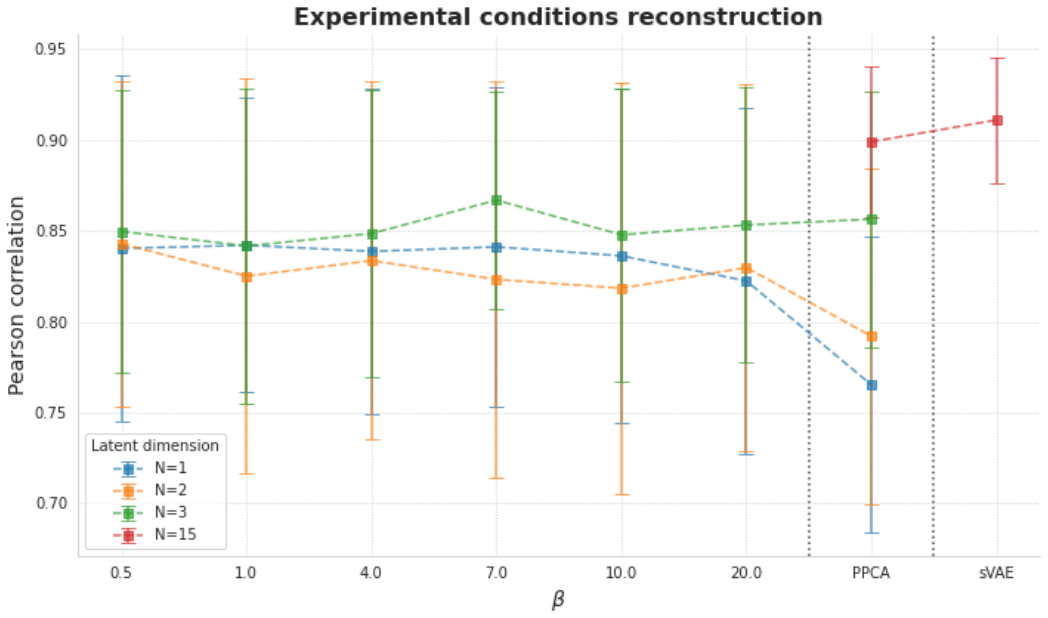
During sVAE training, parsimonious and interpretable representations are enforced by variational dropout. Model selection in latent space can then be achieved using this technique. Here, the dropout rate after convergence is shown when the initial latent dimensions are set to 32 (Appendix S1- Figure 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[25]).



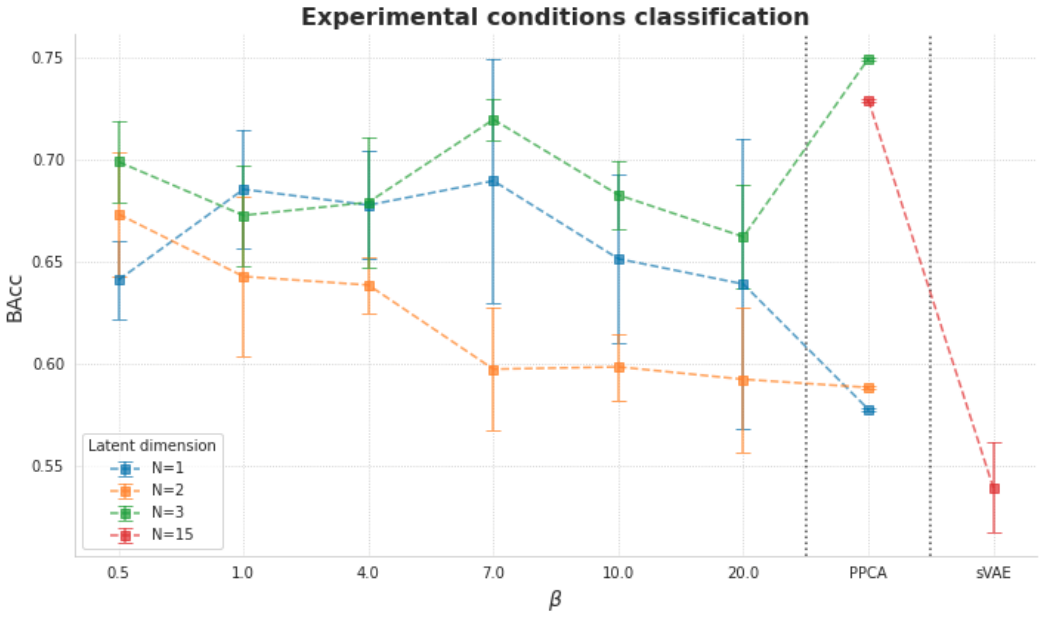
Appendix S1- Figure 1. sVAE estimated dropout rate.

Appendix S2 Model evaluation using experimental conditions

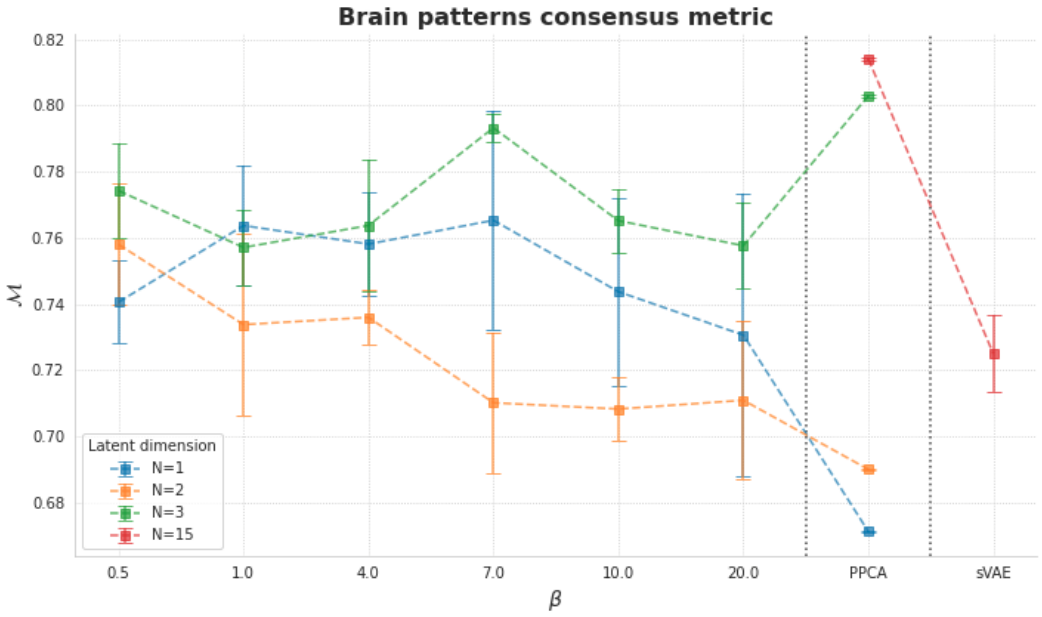
The proposed model evaluation relies on the brain pattern labels. Indeed, it has been shown that these labels can effectively represent the different configurations of the brain[3,4]. Here, we re-evaluate the reconstruction quality, classification accuracy, and consensus metric using a different set of labels composed of the experimental 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 Pearson correlation 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: Pearson correlation 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.



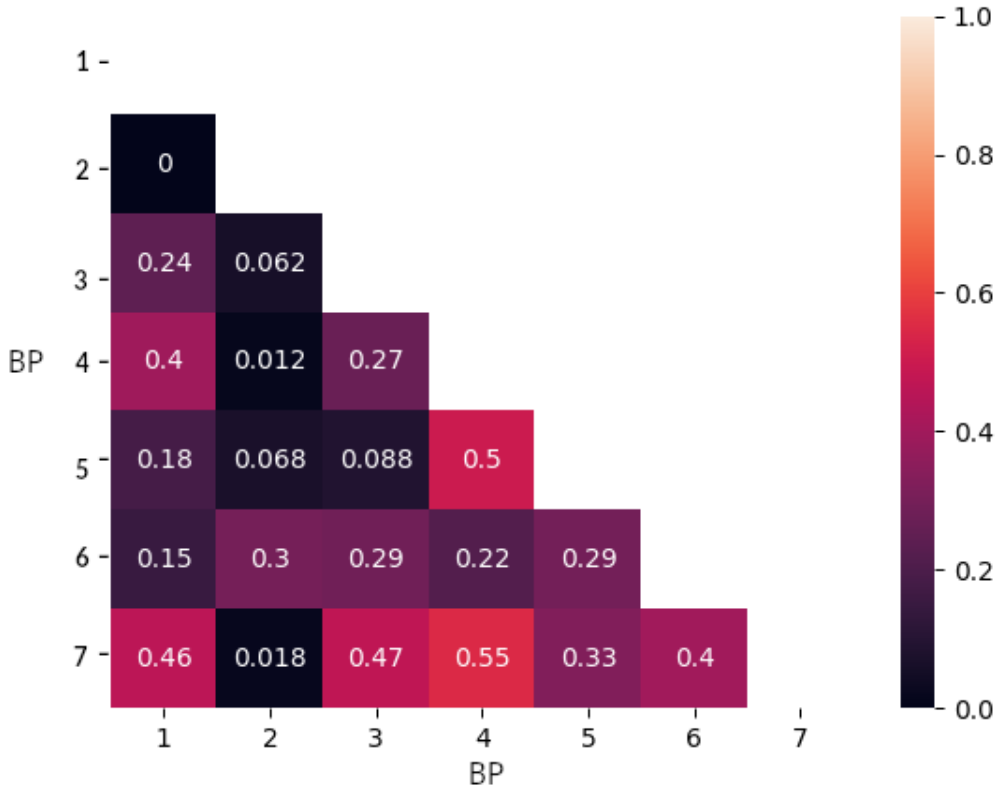
Appendix S2- Figure 3. The proposed consensus metric $\mathscr{M}$.

Appendix S3 Additional brain pattern analyses

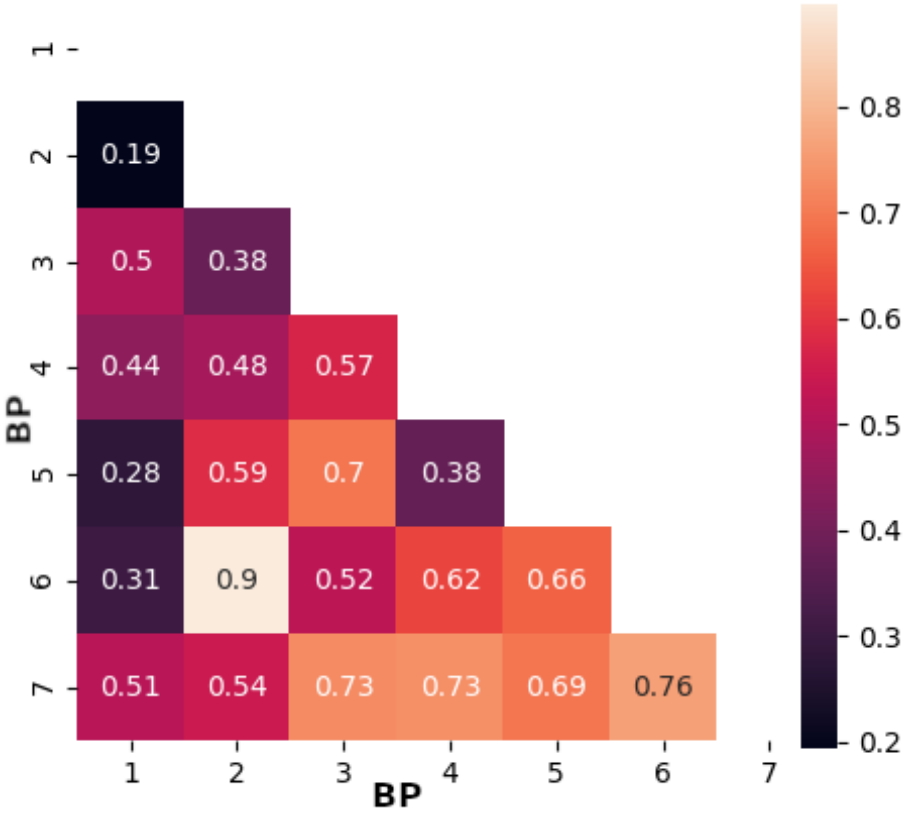
The 2D latent representations obtained with a VAE_2 can be stratified using the brain pattern (BP) labels. To quantify the overlap between brain pattern locations in latent space, we use the Dice similarity coefficient (Appendix S3- Table 1). The Dice metric yields values between 0 (no spatial overlap) and 1 (complete overlap)[37]. 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 resulting Dice coefficients for each brain pattern are averaged and the associated standard deviations are calculated. The Pearson correlation matrix between the brain patterns further describes the studied brain repertoire (Appendix S3- Figure 2). Overall, these two experiments allow us to better characterize the learned latent space in terms of brain patterns.

Subjects	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
Averaged Dice coefficients across all other BPs and associated standard deviations for each BP embedding.



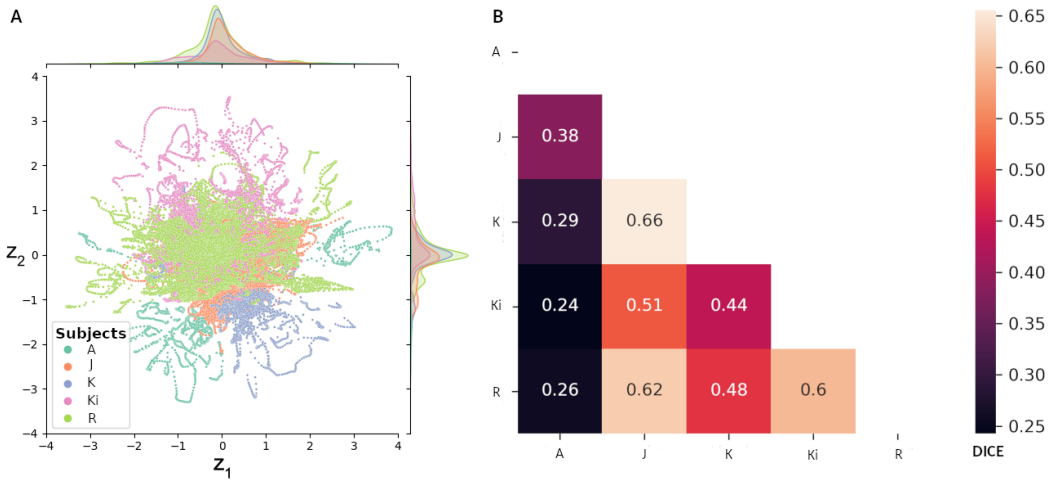
Appendix S3- Figure 1. Dice coefficients for each BP pair



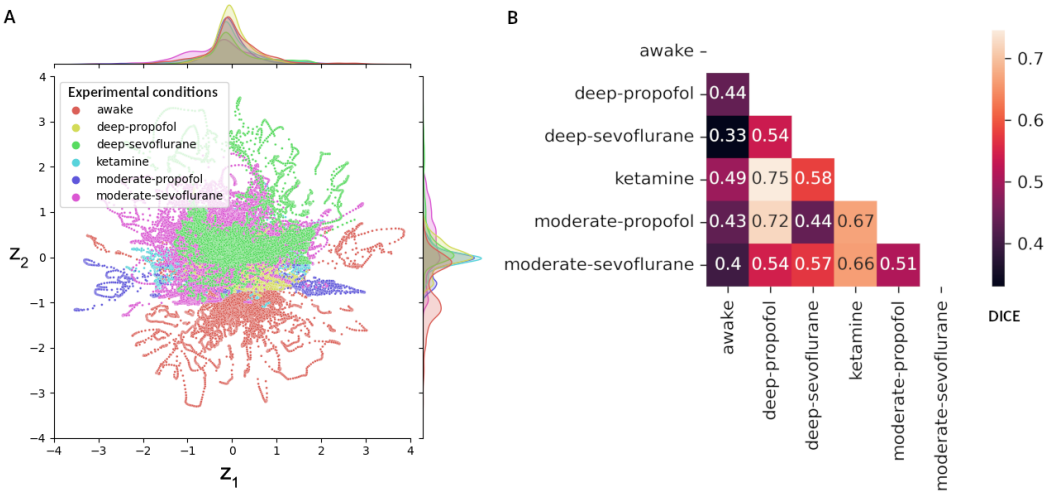
Appendix S3- Figure 2. Correlation matrix between brain patterns.

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).



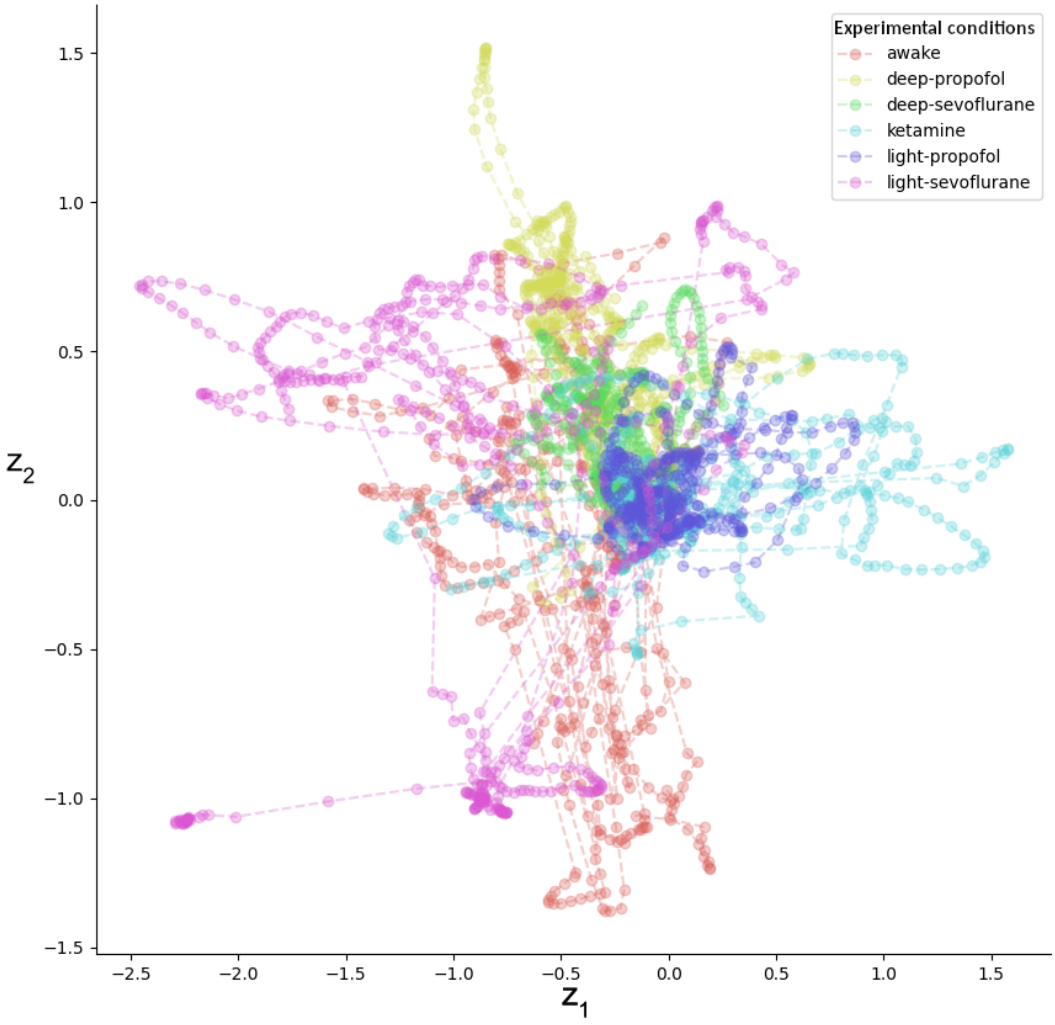
Appendix S4- Figure 1. A) Stratification of VAE_2 latent representations by subject. Density curves show that the different subjects are almost superimposed in latent space, indicating the absence of subject bias. B) Dice coefficient for each subject pair. The dice coefficient quantifies the overlay, which is large except for subject A, for whom we only have awake runs.



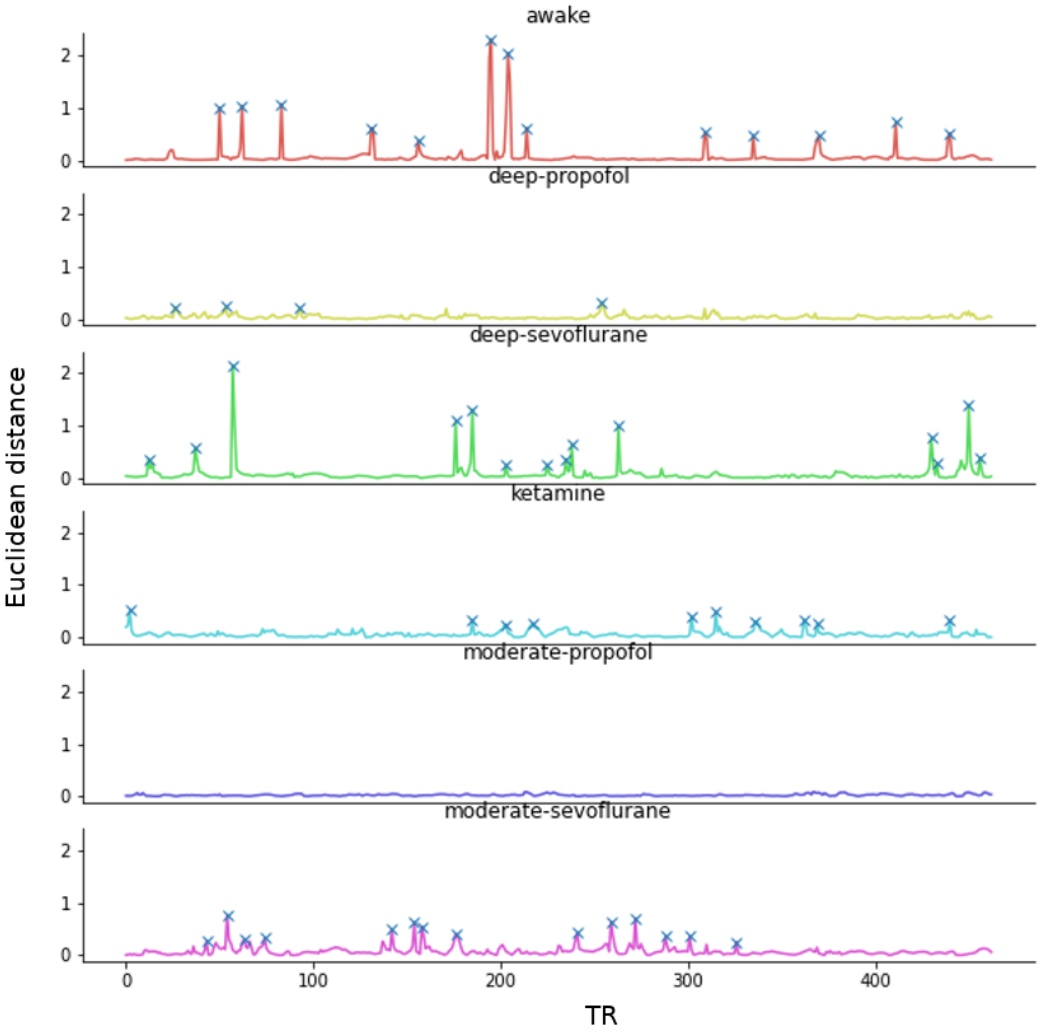
Appendix S4- Figure 2. A) Stratification of VAE_2 latent representations by acquisition conditions. Density curves indicate that the different conditions of anesthesia are almost superimposed in the latent space, showing the absence of anesthetic bias. B) Dice coefficient for each acquisition condition pair. The Dice coefficient quantifies the overlay between the different anesthetics.

Appendix S5 Temporal structure encoded in latent space

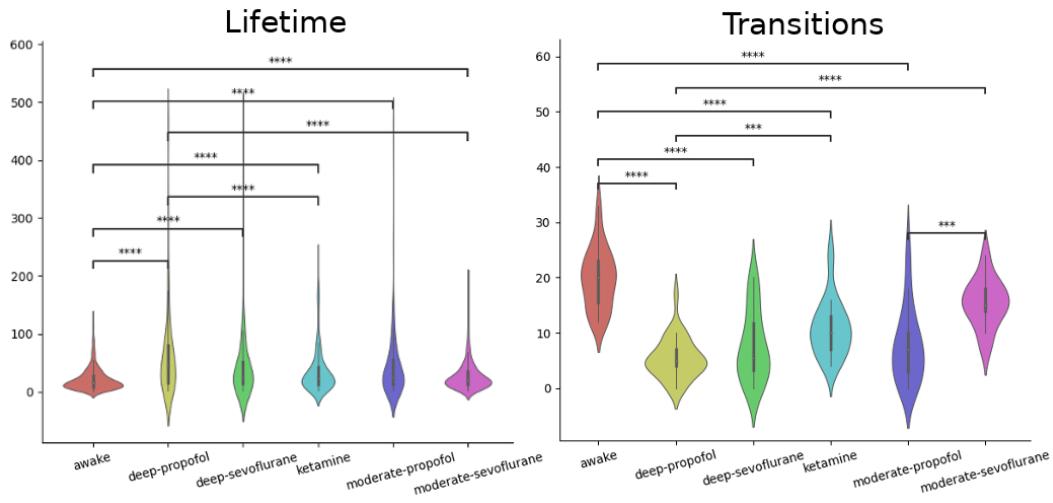
Following the idea of *Tseng et al. (2020)*[49] to study meta-transitions between brain configurations, we examine the encoding of temporal information using the selected VAE_2. While no modeling of the acquisition time course is enforced during training, the encoded latent variables have a coherent 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 an Euclidean distance close to zero, and a transition occurs when a jump in the metric is observed (indicated by blue crosses). We used a peak-finding algorithm to locate transitions. This algorithm finds local maxima by simply applying a threshold α defined as the 95th percentile of the Euclidean distance distribution. 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 lower in the awake state than in all other anesthetized states (Appendix S5- Figure 3). It is also interesting to note that the distribution tails are significantly larger in the anesthetized states than in the awake state. 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 experimental condition.



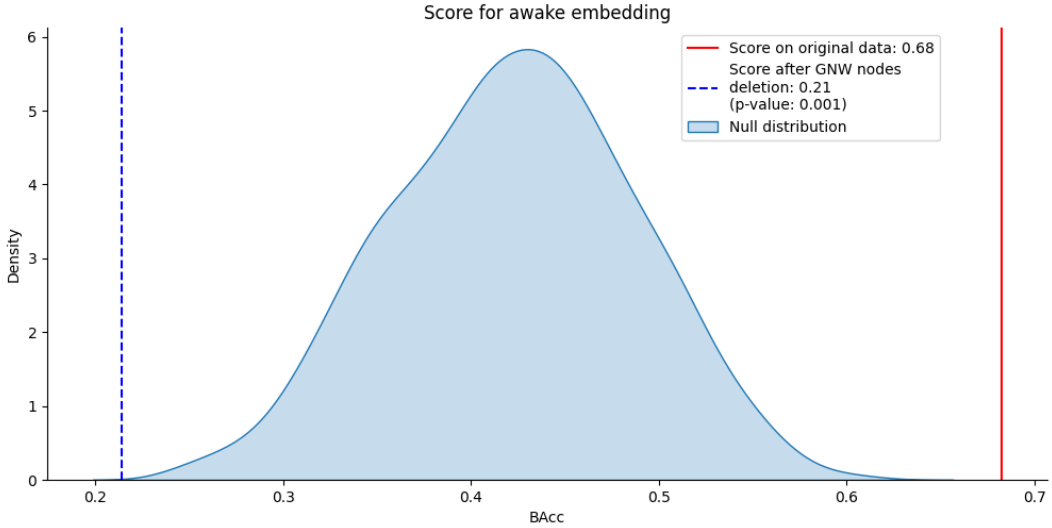
Appendix S5- Figure 2. The Euclidean distance between two consecutive time points for the six runs previously selected. Blue crosses mark transitions.



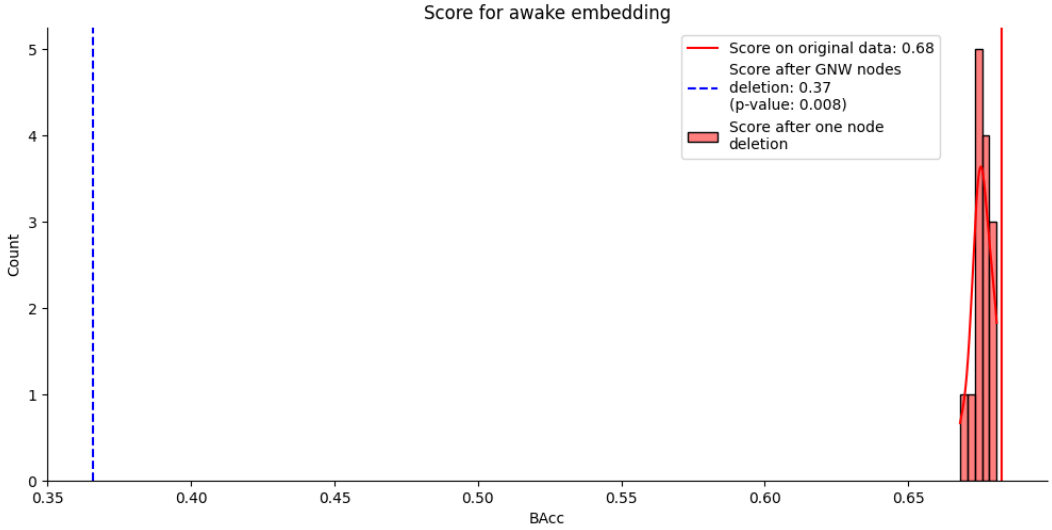
Appendix S5- Figure 3. Meta-stable state lifetime and transition occurrence distributions across acquisition conditions.

Appendix S6 Additional virtual ablation experiments

The proposed ablation analysis provides a unique way to select target connections/regions non-invasively. 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.

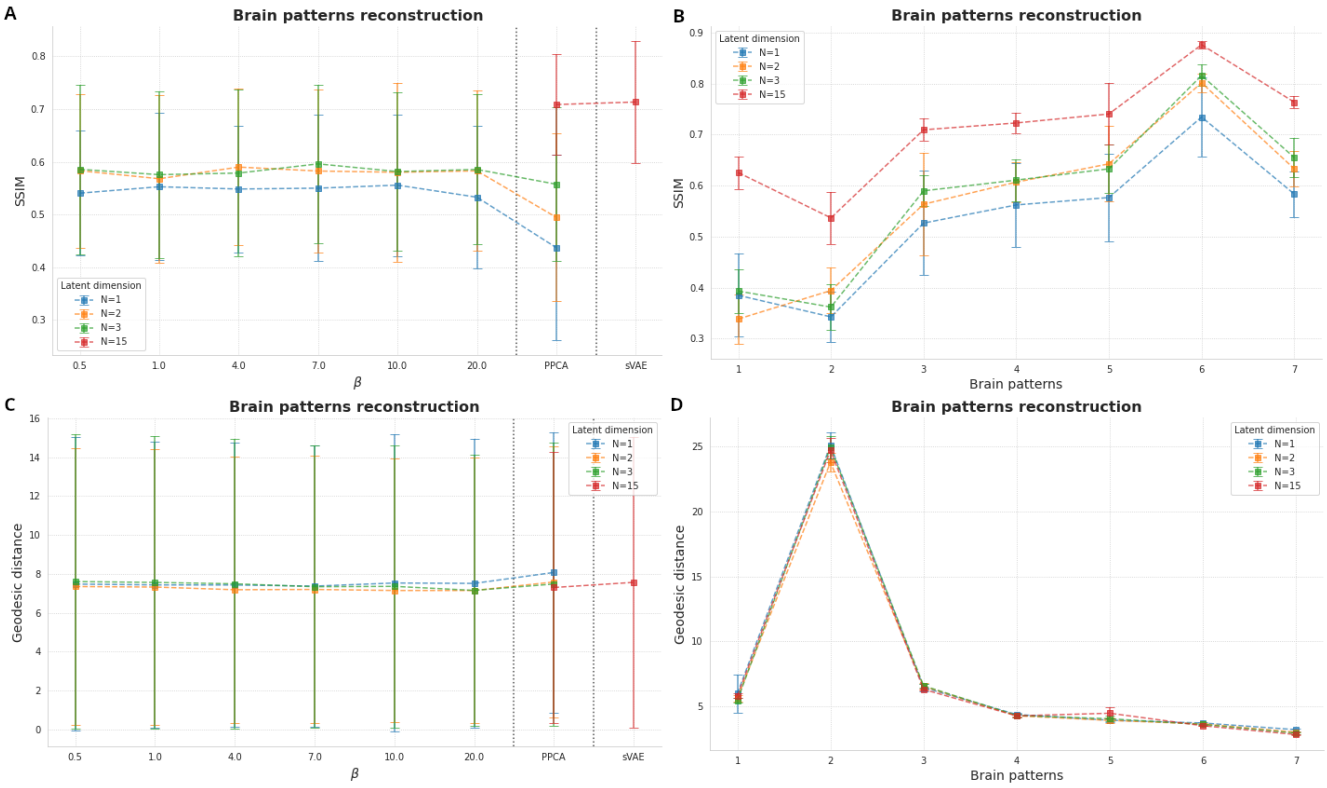


Appendix S6- Figure 1. Ablation study considering GNW and sensory-related connections.



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

Appendix S7 Reconstruction similarity with SSIM and geodesic distance



Appendix S7- Figure 1. Brain pattern (BP) classification/reconstruction using VAE, PPCA, and sVAE models: A) the SSIM of BP-wise averaged dFCs with respect to the model parameters, B) the SSIM recorded for each BP, C) the geodesic distance of BP-wise averaged dFCs with respect to the model parameters and D) the geodesic distance recorded for each BP. The dashed lines represent the trends obtained for each latent space dimension across the considered models.

Appendix S8 Check for out-of-distribution data after ablation

BP	1	3	4	5	6	7
Pearson correlation difference	0.30	0.23	0.31	0.30	0.14	0.26

Appendix S8- Table 1. Difference between reconstruction metric score (with Pearson correlation) of the brain pattern-wise average dFCs after and before ablation study. The brain pattern 2 is not represented in awake data.