




Generative Embeddings of Brain Collective Dynamics Using Variational Autoencoders

Yonatan Sanz Perl ^{1,2,3}, Hernán Bocaccio ², Ignacio Pérez-Ipiña,² Federico Zamberlán,² Juan Piccinini,² Helmut Laufs ⁴,
Morten Kringelbach,⁵ Gustavo Deco,³ and Enzo Tagliazucchi²

¹Universidad de San Andrés, Buenos Aires 1644, Argentina

²Physics Department, University of Buenos Aires and Buenos Aires Physics Institute, Buenos Aires 1428, Argentina

³Center for Brain and Cognition, Computational Neuroscience Group, Universitat Pompeu Fabra, Barcelona 08002, Spain

⁴Department of Neurology, Christian-Albrechts-University Kiel, Kiel 24118, Germany

⁵Department of Psychiatry, University of Oxford, Oxford 2JD, United Kingdom



(Received 6 July 2020; revised 29 September 2020; accepted 26 October 2020; published 2 December 2020)

We consider the problem of encoding pairwise correlations between coupled dynamical systems in a low-dimensional latent space based on few distinct observations. We use variational autoencoders (VAEs) to embed temporal correlations between coupled nonlinear oscillators that model brain states in the wake-sleep cycle into a two-dimensional manifold. Training a VAE with samples generated using two different parameter combinations results in an embedding that encodes the repertoire of collective dynamics, as well as the topology of the underlying connectivity network. We first follow this approach to infer the trajectory of brain states measured from wakefulness to deep sleep from the two end points of this trajectory; then, we show that the same architecture was capable of representing the pairwise correlations of generic Landau-Stuart oscillators coupled by complex network topology.

DOI: [10.1103/PhysRevLett.125.238101](https://doi.org/10.1103/PhysRevLett.125.238101)

Many biological systems can be understood in terms of simple dynamical rules coupled by heterogeneous connectivity patterns. Perhaps the most paradigmatic example is the human brain, where complex collective behavior emerges from the nonlinear dynamics of $\approx 10^{10}$ neurons interacting at $\approx 10^{15}$ synaptic connections [1]. In spite of this complexity at the microscopic scale, the brain spontaneously self-organizes into a reduced number of discrete states, such as those in the wake-sleep cycle, which suggests that a low-dimensional manifold could be sufficient to encode its large-scale dynamics [2].

The mechanisms underlying the emergence of different brain states can be probed using whole-brain models based on conceptually simple local dynamical rules coupled according to empirical measurements of anatomical connectivity [3], for instance, by coupling nonlinear oscillators with the long-range white matter tracts inferred from diffusion tensor imaging (DTI) [4,5]. After parameter optimization to reproduce neuroimaging data acquired during different brain states, the models can be used to explore the interplay between local dynamics, long-range structural coupling, and the formation of large-scale activity patterns [6] and as methods for data augmentation to be combined with machine learning techniques for the purpose of brain state classification [7,8].

While whole-brain models can reproduce the functional connectivity of brain states such as those seen in the progression from wakefulness to deep sleep [9,10], it is unclear whether coupled dynamical systems can also capture relationships between these states, encoding them

into a low-dimensional manifold that preserves the ordering within brain state progressions. More generally, we consider a system of coupled units whose dynamics have been optimized to reproduce the second-order statistics (i.e., pairwise correlations) of a real-world system and ask whether different discrete states of such system can be efficiently represented by latent variables that are capable of reproducing the whole repertoire of states from a reduced number of representative examples. In the particular case of collective brain dynamics, this is equivalent to asking whether the end point states of a certain progression, such as the descent from wakefulness into deep sleep, can be used to learn a latent representation that encodes all intermediate stages, and whether this representation can be extrapolated to reproduce states beyond this progression.

We used whole-brain models fitted to empirical data to generate pairwise correlation matrices for the different brain states that comprise the human wake-sleep cycle: wakefulness, $N1$, $N2$, and $N3$ sleep ($N1$ and $N2$ are intermediate stages, while $N3$ is the deepest stage of human sleep) [11,12]. Rapid eye movement sleep data were not included since the required functional magnetic resonance (fMRI) recording time was not possible due to technical constraints. Next, we trained a variational autoencoder (VAE) with matrices corresponding to wakefulness and $N3$ sleep, showing that intermediate ($N1$ and $N2$) sleep stages were embedded continuously in the latent space, and that the resulting two-dimensional manifold could also be extrapolated to capture known results concerning the structure-function relationship during unconsciousness

[13,14]. Finally, we assessed the relationship between latent space variables and the parameters of generic coupled Stuart-Landau oscillators.

Whole-brain model.—We start from a model constructed from 90 Stuart-Landau nonlinear oscillators, each representing the dynamics within a macroscopic brain region of interest [4]. The coupled dynamics are given by

$$\begin{aligned} \frac{dx_j}{dt} &= (a - x_j^2 - y_j^2)x_j - \omega_j y_j \\ &\quad + G \sum_i J_{ij}(x_i - x_j) + \beta \eta_j, \\ \frac{dy_j}{dt} &= (a - x_j^2 - y_j^2)y_j + \omega_j x_j \\ &\quad + G \sum_i J_{ij}(y_i - y_j) + \beta \eta_j, \end{aligned} \quad (1)$$

where x_j is the dynamical variable that simulates the fMRI signal of region j , and J_{ij} represents the symmetrical coupling matrix that weights the connectivity between regions i and j . This matrix is inferred from the diffusion of water molecules in white matter from DTI recordings and represents the empirical distribution of long-range axon bundles in the brain (further information on the experimental procedures to obtain and analyze fMRI and DTI data can be found in the Supplemental Material [15]). The bifurcation parameter (a) controls the proximity to oscillatory dynamics and G scales the global coupling between oscillators. The ω_j values were estimated from the empirical fMRI data. These frequencies ranged from 0.04 to 0.07 Hz and were determined by the averaged peak frequency of the bandpass-filtered fMRI signals of each individual brain region [10]. Finally, η_j is an additive Gaussian noise term that is scaled by $\beta = 0.04$. An exploration of another model presenting an order-disorder transition can be found in the Supplemental Material [15].

Equation (1) can be optimized to reproduce the second-order statistics of fMRI data acquired during different brain states. In previous work, we proposed to reduce the complexity of the model by grouping brain regions into well-studied functional networks, known as resting state networks (RSNs) [10]. We encoded the 90 bifurcation parameters (a_j) into six parameters representing the contribution of each RSN to the local dynamics as $a_j = \sum_k \Delta_k \mathbb{1}_{jk}$, where $\mathbb{1}_{jk}$ equals 1 if the node j belongs to the k th RSN and zero otherwise. We then applied a stochastic optimization algorithm to determine the Δ_k and G that best reproduce the correlation matrix C_{ij} of each state in the progression from wakefulness to deep sleep. The C_{ij} contains in its i, j entry the linear correlation between the empirical and simulated fMRI time series corresponding to nodes i and j [10]. The Supplemental Material contains a comparison between empirical and simulated C_{ij} [15].

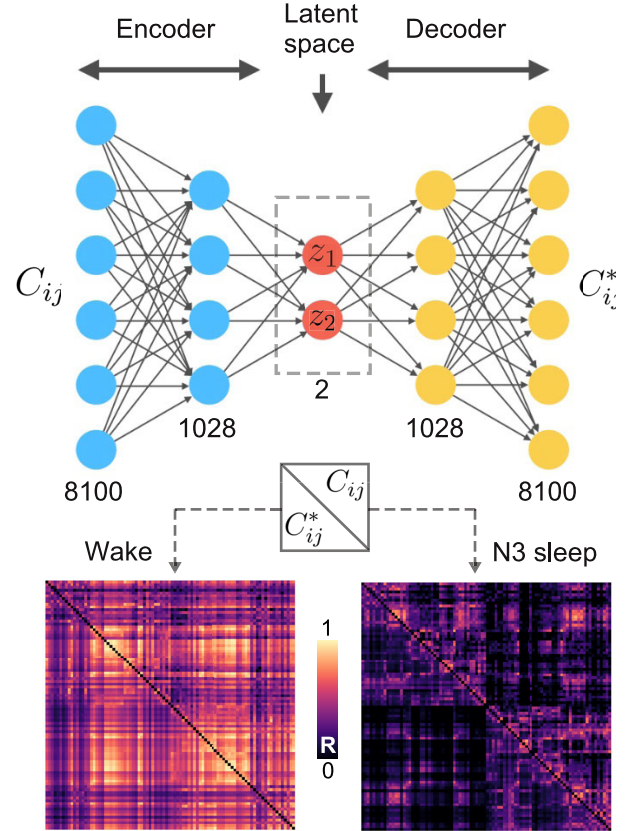


FIG. 1. VAE architecture. The inputs are correlation matrices C_{ij} obtained from the model [Eq. (1)] fitted to wakefulness and $N3$ sleep. The input layer has 8100 units, followed by an intermediate layer with 1028 neurons and a two-dimensional latent space. The next two layers reverse the encoding process, yielding a matrix C_{ij} for each z_1, z_2 pair in the latent space. The bottom panel presents input matrices C_{ij} (above diagonal) and their reconstructed versions C_{ij}^* (below diagonal) for the model fitted to wakefulness and $N3$ sleep.

Encoding the C_{ij} with a VAE.—We implemented a VAE to find a low-dimensional representation encoding the progression of brain states. VAEs are autoencoders trained to map inputs to probability distributions in latent space, which can be regularized during the training process to produce meaningful outputs after the decoding step [34]. The architecture of the implemented VAE (shown in Fig. 1) can be subdivided into three parts: the encoder network, middle variational layer, and decoder network. The encoder consists of a deep neural network with rectified linear units as activation functions and two dense layers, which bottlenecks into the two-dimensional variational layer, where units z_1 and z_2 span the latent space. The encoder network applies a nonlinear transformation to map the C_{ij} into Gaussian probability distributions in latent space, and the decoder network mirrors the encoder architecture to produce reconstructed matrices C_{ij}^* from samples of these distributions.

To train the network, the errors were backpropagated via gradient descent with the purpose of minimizing a loss function composed of two terms: a standard reconstruction error term (computed from the units in the output layer of the decoder) and a regularization term computed as the Kullback-Leibler divergence between the distribution in latent space and a standard Gaussian distribution. The regularization term ensures continuity and completeness in the latent space, i.e., that similar values are decoded into similar outputs and that those outputs represent meaningful combinations of the encoded inputs. [34].

We generated 5000 correlation matrices C_{ij} corresponding to wakefulness and $N3$ sleep using the model described in Eq. (1), each one using a different random seed. We then created 80%/20% random splits to obtain training and test sets and used the training set to optimize the VAE parameters. The training procedure consisted of batches with 128 samples and 50 training epochs using the loss function described in the previous paragraph and an Adam optimizer (gradient descent with parameter-specific learning rate and a running average of gradients and their second moments to attenuate the effects of noise) [35]. A comparison with principal component analysis (PCA) is shown in the Supplemental Material [15].

The latent space encodes the progression of brain states during sleep.—The encoding process applied to the test set with wakefulness and $N3$ sleep data generated two distinct clusters in the latent space [Fig. 2(a)]. We then applied the encoding to simulated correlation matrices obtained by fitting empirical data corresponding to intermediate sleep stages not used to train the VAE ($N1$ and $N2$ sleep). This procedure resulted in separate clusters organized according to sleep depth [Fig. 2(a)]. The emergence of a manifold in latent space where the sequence of correlation matrices was mapped preserving its continuity suggests that a low-dimensional representation can capture the signatures of progressively fading wakefulness.

We applied the decoder exhaustively throughout the latent space, obtaining a pairwise correlation matrix for each z_1, z_2 pair [Fig. 2(b)]. Next, we computed the structural similarity index (SSIM) to compare each matrix obtained from the latent space to the matrices corresponding to wakefulness, $N1$, $N2$, and $N3$ sleep. SSIM is defined as $[(2\mu_x\mu_y+0.01)/(\mu_x^2+\mu_y^2+0.01)][(2\sigma_x\sigma_y+0.03)/(\sigma_x^2+\sigma_y^2+0.03)]/[(\sigma_{xy}+0.015)/(\sigma_x\sigma_y+0.015)]$ [36], where x stands for each C_{ij} matrix shown in Fig. 2(b) and y is the average C_{ij} computed for each brain state. The variables μ_x , μ_y , σ_x , σ_y , and σ_{xy} correspond to the local means, standard deviations, and covariances of matrices x and y , respectively. SSIM has the advantage of simultaneously weighting the Euclidean and correlation distances between matrices [10]. For each z_1, z_2 pair, we determined the brain state with the highest SSIM value and constructed the latent space parcellation shown in Fig. 2(c). Again, we observe that the latent space could be divided into regions

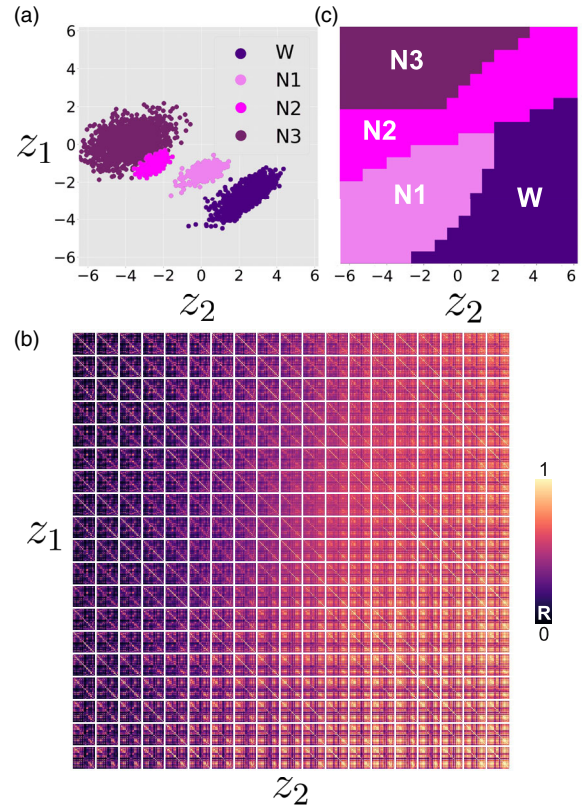


FIG. 2. The latent space obtained from wakefulness and $N3$ sleep contains the orderly progression of intermediate brain states. (a) Latent space representation obtained by encoding the test set (wakefulness and $N3$ sleep) and the encoding obtained for the two intermediate states that were not used to train the VAE ($N1$ and $N2$ sleep). (b) Correlation matrices obtained by decoding an exhaustive exploration of the latent space variables z_1 and z_2 . (c) Latent space divided into regions with maximal similarity to wakefulness, $N1$, $N2$, and $N3$ correlation matrices.

corresponding to wakefulness and all sleep stages, while also respecting the ordering of brain states in the descent to deep sleep only from the VAE fitted to wakefulness and $N3$ sleep [Fig. 2(c)].

Extreme latent space values predict collapse into structural connectivity.—After mapping the progression of brain states during sleep into the latent space, we investigated whether the variables z_1, z_2 could be extrapolated to reproduce signatures of other unconscious states. We hypothesized that moving past $N3$ sleep in the latent space manifold where the progression of brain states is represented would increase the similarity between C_{ij}^* (decoded correlation matrices computed from the dynamics) and J_{ij} (structural coupling matrix). As previously shown both in humans and nonhuman primates [13,14], states of deep unconsciousness are characterized by the collapse of functional coupling to the underlying anatomical connectivity structure.

We decoded a wider range of latent space variables and computed the SSIM between the output correlations and

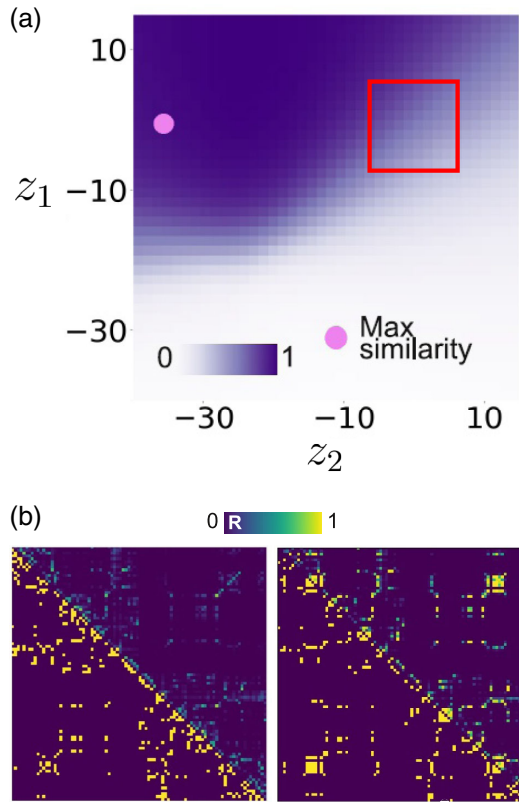


FIG. 3. Latent space variables can be extrapolated to reproduce increased structure-function coupling as a signature of unconsciousness. (a) An exhaustive exploration of the SSIM between the decoded correlation matrices and the empirical structural connectivity matrix. High z_1 and low z_2 maximize this similarity. The red rectangle indicates the range of z_1 and z_2 reproduced in Fig. 2(a). (b) The empirical structural connectivity (left) and the best connectivity matrix reconstructed from the latent space (right) with the lower triangular part representing the matrices thresholded at 0.2.

the structural connectivity. As shown in Fig. 3(a), moving beyond the $N3$ region (high z_1 , low z_2) increased the similarity of the generated correlations with the structural connectivity. Exploring a wider region of the latent space, we found the highest similarity between empirical (J_{ij}) and reconstructed (J_{ij}^*) structural connectivity given by SSIM (J_{ij}, J_{ij}^*) = 0.81. Figure 3(b) (left) shows the empirical J_{ij} and Fig. 3(b) (right) shows the best connectivity matrix reconstructed from the latent space variables; in both cases, the part below the diagonal corresponds to the matrices thresholded at 0.2. As hypothesized, moving past the $N3$ region in latent space reproduced a well-known signature of deep unconsciousness. This suggests that the latent space constructed from wakefulness and $N3$ sleep not only represented intermediate stages, but also captured a manifold where an ampler range of levels of consciousness can be represented.

Mapping the homogeneous model into the latent space.—To gain further understanding concerning how

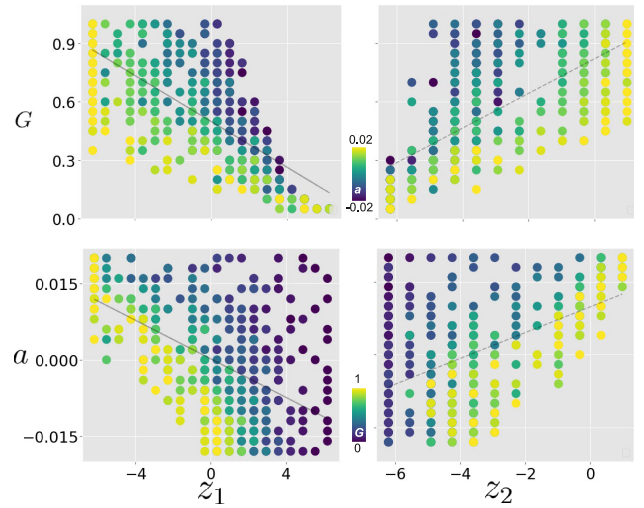


FIG. 4. Relationship between latent space variables (z_1, z_2) and the parameters of the homogeneous model (a, G).

the VAE successfully captured the progression of brain states from few parameter combinations, we trained a VAE using a homogeneous version of the nonlinear coupled oscillators in Eq. (1) (i.e., same a for all oscillators) and compared the latent space encoding in variables z_1, z_2 with the parameters a and G [9]. While the resulting correlation matrices do not reflect those obtained from the empirical data, the homogeneous model can be used to gain insight on the mapping performed by the VAE.

We trained a VAE with 8000 correlation matrices randomly extracted from a set of 10000 matrices generated with the homogeneous model. Half of these matrices were generated using a high coupling factor ($G = 0.8$) and a bifurcation parameter in the oscillatory regime ($a = 0.015$), while the other half were generated using low coupling ($G = 0.2$) and a bifurcation parameter corresponding to fixed-point dynamics ($a = -0.015$).

We decoded the latent space in 20 steps from -6.2 to 6.2 for each variable, obtaining a correlation matrix for each parameter combination. We also constructed several correlation matrices from the model with a between -0.02 and 0.02 and G between 0 and 1 . For each parameter combination, we found the combination of latent space variables that maximized the SSIM between both matrices. In this way, we related each pair (a, G) in the parameter space with each pair (z_1, z_2) in the latent space. We found that both sets of variables were related by approximately linear relationships (G vs z_1 , $r = -0.70$, $p < 0.001$; G vs z_2 , $r = 0.69$, $p < 0.001$; a vs z_1 , $r = -0.56$, $p < 0.001$; a vs z_2 , $r = 0.52$, $p < 0.001$) (Fig. 4). This shows that, for the simplified case of homogeneous a , the latent space approximates a linear transformation of the parameters governing the dynamics of the coupled oscillators.

Discussion.—Several recent studies demonstrated that low-dimensional models suffice to capture the large-scale correlation structure of neural activity seen during different

brain states [9,10]. We went a step further, showing that these models implicitly represent different brain states as points in a low-dimensional manifold. This was established following a constructive process that consisted of training a VAE with correlation matrices belonging to a reduced set of brain states and showing that the latent space represented intermediate states and could be extrapolated to produce hypothesized signatures of deeper unconsciousness. More generally, we showed that complex nonlinear dynamics depending on two parameters could be represented by a latent space that approximated a linear transformation of these parameters. Our results suggest that other (e.g., pathological [37]) brain states could be captured and understood in terms of trajectories within a low-dimensional latent space, with potential applications in diagnosis, prognosis, and data augmentation for automated classification. Generally, we propose that whenever complex collective dynamics are suspected to emerge from few independent parameters, VAEs can be applied to reconstruct these parameters as a trajectory embedded in a low-dimensional latent space.

Authors acknowledge funding from Agencia Nacional De Promocion Cientifica Y Tecnologica (Argentina), Grant No. PICT-2018-03103.

-
- [1] D. R. Chialvo, *Nat. Phys.* **6**, 744 (2010).
- [2] J. M. Shine, L. J. Hearne, M. Breakspear, K. Hwang, E. J. Müller, O. Sporns, R. A. Poldrack, J. B. Mattingley, and L. Cocchi, *Neuron* **104**, 849 (2019).
- [3] M. Breakspear, *Nat. Neurosci.* **20**, 340 (2017).
- [4] G. Deco, M. L. Kringelbach, V. K. Jirsa, and P. Ritter, *Sci. Rep.* **7**, 3095 (2017).
- [5] P. J. Hellyer, M. Shanahan, G. Scott, R. J. Wise, D. J. Sharp, and R. Leech, *J. Neurosci.* **34**, 451 (2014).
- [6] R. Cofré, R. Herzog, P. A. Mediano, J. Piccinini, F. E. Rosas, Y. Sanz Perl, and E. Tagliazucchi, *Brain Sci.* **10**, 626 (2020).
- [7] Y. S. Perl, C. Pallacivini, I. P. Ipiña, M. L. Kringelbach, G. Deco, H. Laufs, and E. Tagliazucchi, *Chaos Solitons Fractals* **139**, 110069 (2020).
- [8] D. M. Arbabiyazd, K. Shen, Z. Wang, M. Hofman-Apitius, A. R. McIntosh, D. Battaglia, V. Jirsa, A. D. N. Initiative *et al.* (2020), <https://doi.org/10.1101/2020.01.18.911248>.
- [9] B. M. Jobst, R. Hindriks, H. Laufs, E. Tagliazucchi, G. Hahn, A. Ponce-Alvarez, A. B. Stevner, M. L. Kringelbach, and G. Deco, *Sci. Rep.* **7**, 4634 (2017).
- [10] I. P. Ipiña, P. D. Kehoe, M. Kringelbach, H. Laufs, A. Ibañez, G. Deco, Y. S. Perl, and E. Tagliazucchi, *NeuroImage* **215**, 116833 (2020).
- [11] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan *et al.*, *J. Clin. Sleep Med.* **8**, 597 (2012).
- [12] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. V. Vaughn *et al.*, The AASM manual for the scoring of sleep and associated events (2012), <https://aasm.org/resources/pdf/scoring-manual-preface.pdf>.
- [13] J. L. Vincent, G. H. Patel, M. D. Fox, A. Z. Snyder, J. T. Baker, D. C. Van Essen, J. M. Zempel, L. H. Snyder, M. Corbetta, and M. E. Raichle, *Nature (London)* **447**, 83 (2007).
- [14] P. Barttfeld, L. Uhrig, J. D. Sitt, M. Sigman, B. Jarraya, and S. Dehaene, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 887 (2015).
- [15] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.125.238101> for the experimental procedures followed to obtain and analyze the neuroimaging data, replication using a different model for data augmentation, comparison between empirical and simulated correlation matrices, comparison of linear (PCA) vs nonlinear (VAE) dimensionality reduction and justification for the use of VAEs and latent space dimensionality justification, which includes Refs. [6,10,12–14,16–33].
- [16] E. Tagliazucchi and H. Laufs, *Neuron* **82**, 695 (2014).
- [17] R. MacLaren, J. M. Plamondon, K. B. Ramsay, G. M. Rocker, W. D. Patrick, and R. I. Hall, *Pharmacotherapy: J. Human Pharmacology Drug Therapy* **20**, 662 (2000).
- [18] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, *NeuroImage* **17**, 825 (2002).
- [19] T. E. Behrens, H. J. Berg, S. Jbabdi, M. F. Rushworth, and M. W. Woolrich, *NeuroImage* **34**, 144 (2007).
- [20] M. E. Tipping and C. M. Bishop, *J. R. Stat. Soc. Ser. B* **61**, 611 (1999).
- [21] B. Schölkopf, A. Smola, and K.-R. Müller, in *International Conference on Artificial Neural Networks* (Springer, New York, 1997), pp. 583–588.
- [22] S. Laureys, M. Boly, G. Moonen, and P. Maquet, *Encycl. Neurosci.* **2**, 1133 (2009), <https://www.semanticscholar.org/paper/Dimensions-of-Consciousness-%3A-Nosology-of-Disorders-Coma/2491e79435793bae3853b481eff92ad8f11ceab4?p2df>.
- [23] M. Magnin, M. Rey, H. Bastuji, P. Guillemant, F. Mauguière, and L. Garcia-Larrea, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3829 (2010).
- [24] C. Reveley, A. K. Seth, C. Pierpaoli, A. C. Silva, D. Yu, R. C. Saunders, D. A. Leopold, and Q. Y. Frank, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2820 (2015).
- [25] A. Messé, D. Rudrauf, H. Benali, and G. Marrelec, *PLoS Comput. Biol.* **10**, e1003530 (2014).
- [26] M. A. Carskadon and W. C. Dement, in *Principles and Practice of Sleep Medicine*, edited by M. H. Kryger, T. Roth, and W. C. Dement (Elsevier Saunders, St Louis, 2011), 5th ed., pp. 16–26.
- [27] D. A. Steyn-Ross, M. L. Steyn-Ross, J. W. Sleigh, M. T. Wilson, I.-P. Gillies, and J. Wright, *J. Biol. Phys.* **31**, 547 (2005).
- [28] A. Haimovici, E. Tagliazucchi, P. Balenzuela, and D. R. Chialvo, *Phys. Rev. Lett.* **110**, 178101 (2013).
- [29] E. Tagliazucchi, N. Crossley, E. T. Bullmore, and H. Laufs, *Brain Struct. Funct.* **221**, 4221 (2016).
- [30] M. Murphy, M.-A. Bruno, B. A. Riedner, P. Boveroux, Q. Noirhomme, E. C. Landsness, J.-F. Bricchant, C. Phillips, M. Massimini, S. Laureys *et al.*, *Sleep* **34**, 283 (2011).
- [31] L. Nelson, T. Guo, J. Lu, C. Saper, N. Franks, and M. Maze, *Nat. Neurosci.* **5**, 979 (2002).
- [32] J. Schrouff, V. Perlberg, M. Boly, G. Marrelec, P. Boveroux, A. Vanhaudenhuyse, M.-A. Bruno, S. Laureys, C. Phillips, M. Péligrini-Issac *et al.*, *NeuroImage* **57**, 198 (2011).

- [33] E. Tagliazucchi, D. R. Chialvo, M. Siniatchkin, E. Amico, J.-F. Brichant, V. Bonhomme, Q. Noirhomme, H. Laufs, and S. Laureys, *J. R. Soc. Interface* **13**, 20151027 (2016).
- [34] D. P. Kingma and M. Welling, [arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
- [35] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [36] Z. Wang, E. P. Simoncelli, and A. C. Bovik, in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (IEEE, New York, 2003), Vol. 2, pp. 1398–1402.
- [37] H. Shen, L. Wang, Y. Liu, and D. Hu, *NeuroImage* **49**, 3110 (2010).