Entropy production of Multivariate Ornstein-Uhlenbeck processes correlates with consciousness levels in the human brain

Matthieu Gilson[1], Enzo Tagliazucchi[2], Rodrigo Cofré[3]
[1] *Institut de Neurosciences des Systèmes INSERM-AMU, Marseille, France*
[2] *Physics Department University of Buenos Aires and Buenos Aires Physics Institute Argentina*
[3] *Institute of Neuroscience (NeuroPSI-CNRS), Paris-Saclay University, Gif sur Yvette 91400, France*

Consciousness is supported by complex patterns of brain activity which are indicative of irreversible non-equilibrium dynamics. While the framework of stochastic thermodynamics has facilitated the understanding of physical systems of this kind, its application to infer the level of consciousness from empirical data remains elusive. We faced this challenge by calculating entropy production in a multivariate Ornstein-Uhlenbeck process fitted to fMRI brain activity recordings. To test this approach, we focused on the transition from wakefulness to deep sleep, revealing a monotonous relationship between entropy production and the level of consciousness. Our results constitute robust signatures of consciousness while also advancing our understanding of the link between consciousness and complexity from the fundamental perspective of statistical physics.

Animal cognition is the most sophisticated example of information processing found in biological and technological systems [1]. Consciousness, understood as the capacity to sustain subjective experience, can be considered a property that emerges when a sufficiently high level of complex cognitive processing is achieved [2]. From the perspective of physics, consciousness and cognition seem unlikely to emerge from regular and predictable systems, such as those which are in thermodynamic equilibrium and obey the detailed balance equations [3]. Instead, recent research draws a close parallel between the level of consciousness and the entropy rate of brain activity time series, highlighting temporal irreversibility as a landmark feature of conscious information processing [4–6]. These results suggest a close link between consciousness and non-equilibrium dynamics, prompting a rigorous evaluation from the perspective of stochastic thermodynamics.

In spite of these exciting results, the direct estimation of entropy production from neural activity recordings is undermined by insufficient spatio-temporal sampling, leading to the adoption of heuristics and approximations which lack rigorous justification [3, 4]. To circumvent these limitations, we adopted a framework based on Multivariate Ornstein-Uhlenbeck (MOU) processes, that are widely used for modeling the multivariate dynamics of time series. The importance of MOU derives from the fact that it is the only continuous stationary stochastic process that is simultaneously Gaussian and Markovian. The MOU process is at the heart of many models used to fit fMRI data and to interpret them in terms of whole-brain communication [7–9], in line with the present methodology. We first characterize the non-equilibrium steady state of a generic MOU process. The irreversibility of the process is encoded in the antisymmetric part of the Onsager matrix, while the linearity of the Langevin equations allows us to derive closed-form expression for the entropy production rate in terms of the matrices that define the MOU. As a result, we obtained a model-based estimation of the entropy production rate for the MOU fitted to fMRI data of subjects transitioning different levels of consciousness during the descent from wakefulness to deep sleep.

Time reversibility and entropy production in the MOU process.— We consider the MOU process closely following the notation in previous work [10]:

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\boldsymbol{B}\,\boldsymbol{x}(t) + \boldsymbol{\eta}(t)\ . \tag{1}$$

Boldfaced symbols denote vectors and matrices. The inputs $\boldsymbol{\eta}(t)$ correspond to Gaussian white noise with covariance

$$\left\langle \boldsymbol{\eta}(t)\,\boldsymbol{\eta}^{\mathrm{T}}\,(t')\right\rangle_t = 2\boldsymbol{D}\,\delta\,(t-t')\ . \tag{2}$$

The angular brackets indicate the mathematical expectation over time and the superscript T the transpose for vectors or matrices. The N-dimensional MOU process is thus defined by two real $N \times N$ matrices, the input covariance matrix $\boldsymbol{D}$, which is symmetric with positive eigenvalues, and the Jacobian $\boldsymbol{B}$, which is not symmetric in general.

Following previous results [11, 12], a sufficient condition for the MOU process in Eq. (1) to be a time reversible stationary process corresponds to a specific relation between the matrices $\boldsymbol{B}$ and $\boldsymbol{D}$:

$$\boldsymbol{B}\,\boldsymbol{D} = \boldsymbol{D}\,\boldsymbol{B}^{\mathrm{T}}. \tag{3}$$

To quantify the time (ir)reversibility of the MOU process, it is advantageous to examine the Onsager matrix $\boldsymbol{L}$ reparameterized using the matrices $\boldsymbol{B}$, $\boldsymbol{D}$, and the pairwise zero-lag covariance $\boldsymbol{S} = \langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle_t$:

$$\begin{aligned} \boldsymbol{L} &= \boldsymbol{B}\,\boldsymbol{S} = \boldsymbol{D} + \boldsymbol{Q}\ , \\ \boldsymbol{L}^{\mathrm{T}} &= \boldsymbol{S}\,\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{D} - \boldsymbol{Q}\ . \end{aligned} \tag{4}$$

Here the antisymmetric part $\boldsymbol{Q}$ of $\boldsymbol{L}$ provides a measure for the irreversibility of the process. When the process

is time reversible $\boldsymbol{Q} = 0$ and $\boldsymbol{L}$ is symmetric. The following expression for the entropy production rate Φ can then be derived (see the supplementary material for a full derivation of this equation)

$$\Phi = \mathrm{tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{Q}\right) = -\,\mathrm{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{Q}\right). \qquad (5)$$

The entropy production rate Φ is non-negative and provides a scalar measure for the (ir)reversibility of the whole network process, vanishing only if the process is reversible.

MOU-based anatomo-functional model to fit empirical fMRI data.— We fitted a MOU process to the time series of blood oxygen level-dependent (BOLD) activity measured using fMRI for a whole-brain parcellation consisting of $N = 90$ regions of interest (ROIs). The BOLD signals were recorded from 15 healthy participants during wakefulness and three sleep stages of progressively deeper unconsciousness (N1, N2, N3). Further details about the data preprocessing like detrending and filtering can be found in [13]. Example BOLD time series are illustrated in Fig. 1A. Fig. 1B-C show two functional connectivity matrices, here calculated as covariances with zero lag $\hat{\mathbf{S}}(0)$ and lag of 1 timestep $\hat{\mathbf{S}}(1)$. These matrices are the empirical counterparts of the the model pairwise covariance $\mathbf{S}(l) = \langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t+l)\rangle_t$ with lag l, which is symmetric for $l = 0$ and was denoted above by $\mathbf{S} = \mathbf{S}(0)$.

In this application, the activity x_i of the MOU process describes the BOLD activity of node i. Its Jacobian matrix $\mathbf{B}$ quantifies the propagation of BOLD activity between ROIs, ignoring hemodynamics. Specifically, the diagonal elements B_{ii} are related to a time constant τ (identical for all ROIs) and the off-diagonal elements $C_{ij} = -B_{ij}$ correspond to the concept of effective connectivity from ROI j to ROI i (excitatory when $C_{ij} > 0$):

$$-B_{ij} = -\frac{\delta_{ij}}{\tau} + C_{ij}\ , \qquad (6)$$

where δ_{ij} is the Kronecker delta. The input variance D_{ii} reflects the fluctuation amplitude of the spontaneous activity to ROI i.

For each subject and condition, the model was fitted to reproduce the two covariance matrices calculated from the empirical BOLD signals $\hat{\mathbf{S}}(0)$ and $\hat{\mathbf{S}}(1)$ (see Fig. 1B-C). We used a recent estimation method based on gradient descent to iteratively adjust $\boldsymbol{B}$ and $\boldsymbol{D}$ until reaching the best fit [7]. At each optimization step, we calculate the model counterparts of the covariance matrices $\boldsymbol{S}(0)$ and $\boldsymbol{S}(1)$, assuming stationarity over each fMRI session. Importantly, this optimization procedure incorporates topological constraints on $\boldsymbol{B}$, adjusting only existing anatomical connections (see Fig. 1D-E), also keeping the input cross-covariances $D_{ij} = 0$ for $i \neq j$.

Robust decoding of sleep stages from MOU parameters.— Following previous work [15, 16], we used the scikit-learn Python library for the implementations of multinomial logistic regression (MLR) classifier. The input features corresponded to the vectorized

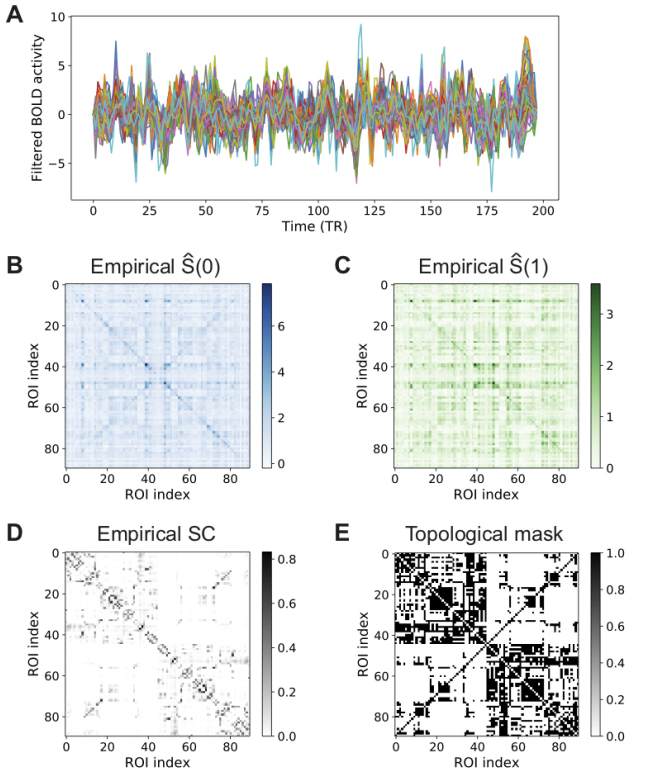

FIG. 1. **Empirical BOLD time series, functional and structural connectivity.** A) Example of the filtered BOLD time series (of duration $T = 198$) corresponding to the 90 ROIs of the AAL parcellation during wakefulness of one participant. B-C) Functional connectivity matrices calculated from the filtered BOLD signals in panel A, $\hat{S}(0)$ with zero lag and $\hat{S}(1)$ with a lag of one timestep (TR=2 s). These matrices are used in the objective functions used to fit the anatomo-functional model. D) Generic structural connectivity (SC) obtained from DTI data as described in [14]. E) Mask for existing directional connections to constrain the topology of the B matrix in the network model (symmetric here).

C/D/S matrices after discarding zero or redundant elements. We implemented a stratified cross-validation scheme with 80% of the samples for the train set and 20% for the test set, where the ratio of classes are the same in both sets. We also use the subject identity as "group information" to avoid mixing subject data between the train and test sets. In practice, we use 100 random splits of the data an report the distribution of accuracies of the 100 splits.

As illustrated in Fig. 2B, both the empirical BOLD variances and the model estimates exhibit global differences across the four sleep stages, although they do not exhibit a clear trend. These differences in global measures, which are averages over all ROIs, may hide more specific changes at the ROI level, as well as interactions between them. Supplementary Figure S1 shows the good fit of the anatomo-functional model to fMRI data that obtained for all sleep stages, with mean correlation be-

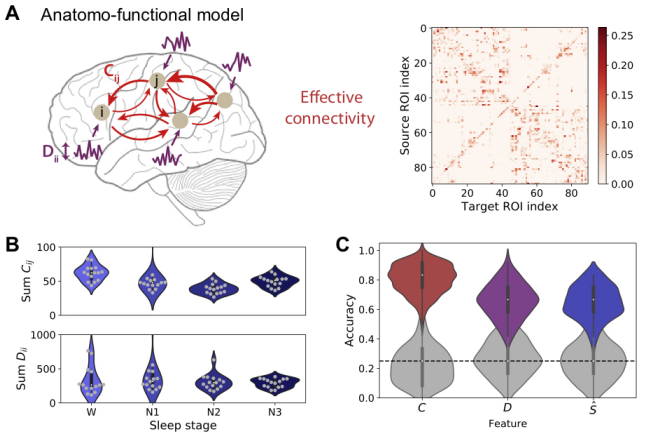

FIG. 2. **Anatomo-functional model fitted to the empirical data.** A) Our dynamic network model has two sets of optimized parameters: the matrix **C**, which describes the causal interaction between brain regions, and the input variance **D**, which represents the spontaneous activity of each brain region. Note that the topology of the matrix. **C** corresponds to the mask inferred from the SC data in Fig. 1D-E, but the weights are estimated from the empirical FC matrices Fig. 1B-C, resulting in an anatomo-functional model. B) Changes in total **C** and **D** weights across sleep stages (x-axis), pooled over the 15 subjects. The sleep stages are represented by the blue contrasts, from light for wake (W) to dark for the deepest sleep (N3). C) Classification accuracy based on the model estimates, **C** and **D**, and the empirical covariance matrices. The classifier is the multinomial logistic regression (MLR), which captures changes in individual features across sleep stages. The gray violin plots correspond to the chance-level accuracy calculated empirically in each case by shuffling the labels of the sleep stages.

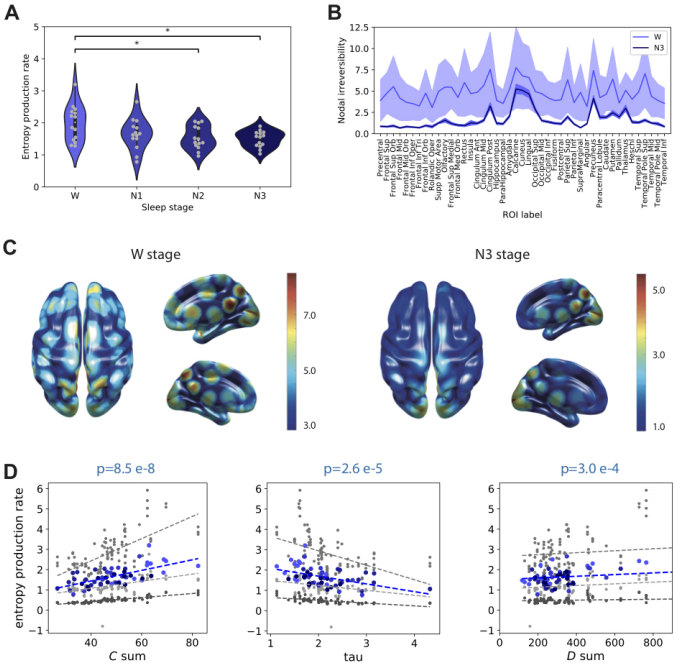

FIG. 3. **Entropy production rate correlates with consciousness levels in the human brain.** A) Violin plots comparing the entropy production rate across sleep stages. Same color coding used in previous plots. The average entropy production values across subjects are for the four sleep stages are: 1.99, 1.65, 1.54 and 1.49, respectively. The stars indicate statistical significance for the Mann-Whitney test with $p < 0.05$. B) Comparison of the nodal irreversibility for each ROI (x-axis) between the W and N3 states (in light and dark blue, respectively). The plotted values correspond to the absolute value of sums over rows of Q, averaged for homotopic regions; error bars indicate the variability across subjects measured as standard error of the mean. C) Heatmap plots of the nodal irreversibility on the cortical surface for the W and N3 sleep stages. Note the different color scales for the two stages, for the purpose of better readability. D) Plots of Φ across sleep stages and subjects as a function of the sum of **C** weights (left panel), the time constant τ on the diagonal of **B** (middle panel) and the sum of **D** variances (right panel). Comparison with surrogate values with randomized (redistributing the total weight/variances keeping the same topology and overall sum) **C** matrices (light gray), randomized **D** matrices (middle gray) and both (dark gray).

tween simulated and empirical FC matrices exceeding 0.6 for all stages. Fig. 2C shows that the model estimates give good classification accuracy, both for C (in red) and D (in purple). This indicates that the model capture the differences in brain dynamics across the sleep stages. Notably, the matrix C gives a better classification accuracy than the empirical functional connectivity $\hat{S}(0)$ (in blue), meaning that the model inversion is robust and captures refined information about the sleep stages. Note that the MLR has a better accuracy than the the 1-nearest-neighbor (1NN) in Suppl Fig S2A, indicating that the changes across sleep stages concern specific features, i.e. connectivity weights (**C**) or nodal spontaneous activity (**D**), rather than their global profile.

Reduced entropy production in the transition from wakefulness to deep sleep.— Using the condition-specific estimated parameters, we calculated the entropy production rate in the MOU model using Eq. (5). These results in Fig 3A show that entropy production decreases as a function of sleep depth, which in turns imply that dynamics become closer to equilibrium.

The model-based approach allows us to dissect this phenomenon. For all ROIs, we observe that the contribution to Φ, as measured via the nodal irreversibility, defined as $\sum_j |Q_{ij}|$ for each ROI i, decreases, as illustrated in Fig 3B. This suggests that the reduction of Φ from W to N3 is a rather global phenomenon, but with a differentiated magnitude across brain regions. Notably, regions in the occipital lobes (cuneus, calcarine, lingual) as well as regions associated to hubs in the default-mode network (precuneus, post cingulate) and the thalamus remain at a high level of nodal irreversibility in the deep sleep N3; these regions have been shown to exhibit sleep-related changes in previous studies [17–19]. See Suppl Fig S3 for a more detailed comparison across sleep stages.

Last, we examine how the model parameters **C** and **D** contribute to Φ and its reduction across sleep stages. Fig 3D shows a positive relationship between Φ and the sum of weights in **C**, as well as the sum of variances in **D**; conversely, a larger τ (directly calculated from the empirical BOLD signals) corresponds to a lower Φ. Then we assess the importance of the detailed structures in the **C** and **D** estimates by randomizing them spatially, namely redistributing the total weight/variances across non-zero elements while keeping the same topology and overall sum. We observe the same trends with respect to the **C** and **D** sums, but shifted up or down depending on the surrogates in Fig 3C: randomizing **C** (light gray) decreases slightly Φ, whereas randomizing **D** (middle gray) increases Φ; randomizing both (dark gray) decreases Φ. This indicates that Φ strongly depends on the detailed structures of the **C** and **D** estimates, being larger in the data than in the randomized surrogates. The opposing effects in randomizing **C** and **D** also suggest a balance implemented by the detailed brain dynamics, which results in a controlled level of Φ. Together, our results hint at a positive relationship between the measured Φ and the different levels of consciousness.

Discussion.— We measured the entropy production using our anatomo-functional MOU process associated to resting-state fMRI activity recorded from human subjects in different sleep stages. The advantage of our model-based approach is that the entropy production has a closed-form expression from first principles of stochastic thermodynamics for the MOU process, which is numerically fitted to the fMRI data. Our results show high entropy production rate in conscious wakefulness, i.e. correlating positively with the presumed level of cognitive processing. This is consistent with converging theoretical accounts that identify consciousness with an emergent property of a highly complex physical system [2]. These results are also consistent with previous findings relating entropy production with states of consciousness [4–6], with the advantage that do not depend on heuristic approximations. Importantly, our approach allows to identify the brain regions that contribute most to entropy production. The fulfilment of detailed balance in the brain is scale-dependent [3]. At the large scale, its violation might relate to the large-scale circuit operations critical for healthy cognition and for the global broadcasting of information which is identified with the computational aspect of consciousness [20]. Because of this, metrics related to the departure from detailed balance (such as entropy production rate) might offer valuable tools to determine levels of consciousness in brain injured patients and other neurological populations.

In summary, assessing temporal irreversibility through entropy production of MOU processes derived from fMRI signals has the potential to highlight different states of consciousness and cognition. More generally, can bridge brain dynamics and thermodynamics, and ultimately help to understand fundamental questions about the brain and consciousness.

[1] G. Piccinini and A. Scarantino, Journal of biological physics **37**, 1 (2011).
[2] A. K. Seth and T. Bayne, Nature Reviews Neuroscience , 1 (2022).
[3] C. W. Lynn, E. J. Cornblath, L. Papadopoulos, M. A. Bertolero, and D. S. Bassett, PNAS (2021).
[4] Y. Sanz Perl, H. Bocaccio, C. Pallavicini, I. Pérez-Ipiña, S. Laureys, H. Laufs, M. Kringelbach, G. Deco, and E. Tagliazucchi, Physical Review E (2021).
[5] L. de la Fuente, F. Zamberlan, H. Bocaccio, M. Kringelbach, G. Deco, Y. S. Perl, and E. Tagliazucchi, Cerebral Cortex (2022).
[6] R. N. Muñoz, A. Leung, A. Zecevik, F. A. Pollock, D. Cohen, B. van Swinderen, N. Tsuchiya, and K. Modi, Physical Review Research **2**, 023219 (2020).
[7] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco, PLoS Computational Biology **12** (2016).
[8] K. J. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, and P. Zeidman, NeuroImage **199** (2019).
[9] S. Frässle, Z. M. Manjaly, C. T. Do, L. Kasper, K. P. Pruessmann, and K. E. Stephan, NeuroImage **225** (2021).
[10] C. Godrèche and J. M. Luck, Journal of Physics A: Mathematical and Theoretical (2019).
[11] H. Risken and T. Frank, Springer Series in Synergetics (1996).
[12] H. Qian, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 10.1098/rspa.2001.0811 (2001).
[13] E. Tagliazucchi and H. Laufs, Neuron (2014).
[14] I. P. Ipiña, P. D. Kehoe, M. Kringelbach, H. Laufs, A. Ibañez, G. Deco, Y. S. Perl, and E. Tagliazucchi, NeuroImage (2020).
[15] V. Pallarés, A. Insabato, A. Sanjuán, S. Kühn, D. Mantini, G. Deco, and M. Gilson, NeuroImage **178** (2018).
[16] M. Gilson, G. Zamora-López, V. Pallarés, M. H. Adhikari, M. Senden, A. T. Campo, D. Mantini, M. Corbetta, G. Deco, and A. Insabato, Network Neuroscience **4** (2020).
[17] S. G. Horovitz, A. R. Braun, W. S. Carr, D. Picchioni, T. J. Balkin, M. Fukunaga, and J. H. Duyn, Proceedings of the National Academy of Sciences of the United States of America **106**, 10.1073/pnas.0901435106 (2009).
[18] T. T. Dang-Vu, M. Schabus, M. Desseilles, V. Sterpenich, M. Bonjean, and P. Maquet, Functional neuroimaging insights into the physiology of human sleep (2010).
[19] L. Mirandola, G. Cantalupo, A. E. Vaudano, P. Avanzini, A. Ruggieri, F. Pisani, G. Cossu, C. A. Tassinari, P. F. Nichelli, F. Benuzzi, and S. Meletti, Epilepsy and Behavior Case Reports **1**, 10.1016/j.ebcr.2013.06.005 (2013).
[20] S. Dehaene, M. Kerszberg, and J. P. Changeux, Proceedings of the National Academy of Sciences of the United States of America **95**, 10.1073/pnas.95.24.14529 (1998).

Supplemental Material for: Entropy production of Multivariate Ornstein-Uhlenbeck processes correlates with consciousness levels in the human brain

Matthieu Gilson[1], Enzo Tagliazucchi[3], Rodrigo Cofré[4]
[1] *Institut de Neurosciences des Systèmes INSERM-AMU, Marseille, France*
[2] *Physics Department University of Buenos Aires and Buenos Aires Physics Institute Argentina*
[3] *Institute of Neuroscience (NeuroPSI-CNRS), Paris-Saclay University, Gif sur Yvette 91400, France*

In this supplemental material, we present a detailed derivations of the equations used in this article, methods to compute the parameteres of the model, and detailed description of the fMRI data.

I. MULTIVARIATE ORNSTEIN-UHLENBECK PROCESS

A. Description of the state evolution

Knowing the initial condition $\boldsymbol{x}(0)$ and the realization of the stochastic input $\boldsymbol{\eta}$ over time, the trajectory of the solution of the Eq.1 in the main text is given by:

$$\boldsymbol{x}(t) = \boldsymbol{G}(t)\,\boldsymbol{x}(0) + \int_0^t \boldsymbol{G}(t-s)\,\boldsymbol{\eta}(s)\,\mathrm{d}s \ , \qquad (1)$$

where $\boldsymbol{G}(t) = \mathrm{e}^{-\boldsymbol{B}t}$ is the Green's function, also known as propagator. In addition to its mean value $\langle \boldsymbol{x}(t) \rangle = \boldsymbol{G}(t)\,\boldsymbol{x}(0)$, the process is also characterized by its covariance matrix $\boldsymbol{S}(t,t') = \langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t') \rangle$. The zero-lag covariance, denoted by $\boldsymbol{S}(t,t)$, obeys the following deterministic differential equation:

$$\frac{\mathrm{d}\boldsymbol{S}(t,t)}{\mathrm{d}t} = -\boldsymbol{B}\,\boldsymbol{S}(t,t) - \boldsymbol{S}(t,t)\,\boldsymbol{B}^{\mathrm{T}} + 2\boldsymbol{D} \ . \qquad (2)$$

Meanwhile, the lagged covariance with $t' > t$ exhibits an exponential decay as a function of the lag $t' - t$:

$$\boldsymbol{S}(t,t') = \boldsymbol{S}(t,t)\,e^{-\boldsymbol{B}^{\mathrm{T}}(t'-t)} \ . \qquad (3)$$

A standard method for analysing Eq.1 in the main text, consists in describing the evolution of the probability distribution $P(\boldsymbol{x},t)$ via the Fokker–Planck equation:

$$\frac{\partial P(\boldsymbol{x},t)}{\partial t} = \nabla \cdot \left[\boldsymbol{B}\,\boldsymbol{x}(t)\,P(\boldsymbol{x},t) + \boldsymbol{D}\,\nabla P(\boldsymbol{x},t)\right] \ , \qquad (4)$$

where ∇ denotes the spatial derivative with respect to $\boldsymbol{x}$. Eq. (4) can be rewritten as a continuity equation of the form

$$\frac{\partial P(\boldsymbol{x},t)}{\partial t} + \nabla \cdot \boldsymbol{J}(\boldsymbol{x},t) = 0 \ . \qquad (5)$$

with the following expression for the probability current (or flux)

$$\boldsymbol{J}(\boldsymbol{x},t) = -\boldsymbol{D}\nabla P(\boldsymbol{x},t) - \boldsymbol{B}\boldsymbol{x}(t)P(\boldsymbol{x},t) \qquad (6)$$

B. Stationary state and probability current

The Gauss-Markov property of the Ornstein-Uhlenbeck process ensures that the mean and covariances converge exponentially fast toward their respective fixed points, provided the eigenvalues of $\boldsymbol{B}$ (which may be complex) have positive real part. The stationary state of the MOU process exhibits Gaussian fluctuations around a mean equal to zero. This corresponds to the time-independent multivariate probability density

$$P(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}(\det \boldsymbol{S})^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}\right) \ , \qquad (7)$$

where $\boldsymbol{S}$ denotes the fixed point of the zero-lag covariance matrix $\boldsymbol{S}(t,t)$. From Eq. (7), the gradient of $P(\boldsymbol{x})$ simply reads

$$\nabla P(\boldsymbol{x}) = \frac{\partial P(\boldsymbol{x})}{\partial \boldsymbol{x}} = -P(\boldsymbol{x})\,\boldsymbol{S}^{-1}\,\boldsymbol{x} \qquad (8)$$

From Eq. (6), the stationary probability current $\boldsymbol{J}(\boldsymbol{x})$ can thus be rewritten in a compact form

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{x}) &= \boldsymbol{D}\,P(\boldsymbol{x})\,\boldsymbol{S}^{-1}\,\boldsymbol{x} - \boldsymbol{B}\,\boldsymbol{x}\,P(\boldsymbol{x}) \qquad (9) \\ &= \boldsymbol{\mu}\,\boldsymbol{x}\,P(\boldsymbol{x}) \ , \end{aligned}$$

with

$$\boldsymbol{\mu} = \boldsymbol{D}\,\boldsymbol{S}^{-1} - \boldsymbol{B} \qquad (10)$$

C. Entropy production rate

Going a step further, the (ir)reversibility can be described using thermodynamic variables evaluated for the dynamic process. Using the well-known definition for entropy for the probability distribution $P(\boldsymbol{x},t)$, now considering its time dependent version, we have

$$e[P] = -\int_{\mathbb{R}^n} P(\boldsymbol{x},t)\,\log P(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} \ . \qquad (11)$$

It can be shown that the rate of the increase of entropy over time can be decomposed into two factors, namely $\dot{e}[P] = EPR - HDR$, where EPR is the entropy production rate and HDR the heat-dissipation rate [1–3]. The

EPR is the main quantity of interest here, which we denote by Φ. Now calculating Φ for the time-independent distribution $P(\boldsymbol{x})$, we have

$$\Phi = \int \frac{\boldsymbol{J}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{D}^{-1}\boldsymbol{J}(\boldsymbol{x})}{P(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \left\langle \boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\Pi}\right\rangle \tag{12}$$

where $\boldsymbol{\Pi}$ is called the the thermodynamic force and is related to $\boldsymbol{J}$ by the Onsager's reciprocal relations [1]:

$$\boldsymbol{\Pi} = \frac{\boldsymbol{D}^{-1}\boldsymbol{J}}{P} \tag{13}$$

The heat-dissipation rate can be computed as follows:

$$HDR = \int_{\mathbb{R}^n} \boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{x}\cdot\boldsymbol{J}\mathrm{d}x \tag{14}$$

In the context of the stationary MOU diffusion processes, a general expression for the entropy production rate per unit time in the stationary state is the following [1, 2, 4]

$$\Phi = \int \left(\nabla \log P(\boldsymbol{x}) - \boldsymbol{D}\boldsymbol{B}\boldsymbol{x}\right)^{T}\boldsymbol{D}\left(\nabla \log P(\boldsymbol{x}) - \boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{x}\right)P(\boldsymbol{x})d\boldsymbol{x} \tag{15}$$

which can be obtained from (10), (12) and (13) as follows:

$$\begin{aligned}
\mu &= \boldsymbol{D}\boldsymbol{S}^{-1} - \boldsymbol{B} & \\
\boldsymbol{D}^{-1}\mu &= \boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B} & \boldsymbol{D}^{-1}\cdot \\
\boldsymbol{D}^{-1}\mu\boldsymbol{x} &= (\boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B})\boldsymbol{x} & \cdot\,\boldsymbol{x} \\
\boldsymbol{D}^{-1}\mu\boldsymbol{x}P &= (\boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B})\boldsymbol{x}P & \cdot\,P \\
\boldsymbol{D}^{-1}\boldsymbol{J} &= (\boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B})\boldsymbol{x}P & \text{from (9)} \\
\boldsymbol{\Pi} &= (\boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B})\boldsymbol{x} & \text{from (13)} \\
\boldsymbol{\Pi} &= \boldsymbol{S}^{-1}\boldsymbol{x} - \boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{x} &
\end{aligned} \tag{16}$$

Now, as $\nabla \log P(\boldsymbol{x}) = \boldsymbol{S}^{-1}\boldsymbol{x}$, we obtain (15). From (12)

$$\left\langle \boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\Pi}\right\rangle = \left\langle \boldsymbol{x}^{\mathrm{T}}\left(\boldsymbol{D}^{-1}\boldsymbol{B} - \boldsymbol{S}^{-1}\right)^{\mathrm{T}}\boldsymbol{D}\left(\boldsymbol{D}^{-1}\boldsymbol{B} - \boldsymbol{S}^{-1}\right)\boldsymbol{x}\right\rangle,$$

we obtain that

$$\Phi = \left\langle \boldsymbol{x}^{\mathrm{T}}\left(\boldsymbol{D}^{-1}\boldsymbol{B} - \boldsymbol{S}^{-1}\right)^{\mathrm{T}}\boldsymbol{D}\left(\boldsymbol{D}^{-1}\boldsymbol{B} - \boldsymbol{S}^{-1}\right)\boldsymbol{x}\right\rangle, \tag{17}$$

where the average is taken over the stationary state of the process. From this equation we can verify that when $\boldsymbol{S} = \boldsymbol{B}^{-1}\boldsymbol{D}$, then $\Phi = 0$. From Eq.4 in the main text, and (10), we have $\boldsymbol{D}^{-1}\boldsymbol{B} - \boldsymbol{S}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{Q}\boldsymbol{S}^{-1} = -\boldsymbol{D}^{-1}\boldsymbol{\mu}$. Thus, from Eq. (17) considering that $\boldsymbol{S}$ and $\boldsymbol{D}$ are symmetric and $\boldsymbol{Q}$ is anti-symmetric we obtain:

$$\Phi = -\left\langle \boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}\boldsymbol{S}^{-1}\boldsymbol{x}\right\rangle = \left\langle \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{\mu}\boldsymbol{x}\right\rangle \tag{18}$$

The entropy production rate Φ is non-negative. It is strictly positive if the process is irreversible, and it vanishes only if the process is reversible. Since the stationary state of the MOU is Gaussian with covariance matrix $\boldsymbol{S}$, we have the following property: $\left\langle \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right\rangle = \mathrm{tr}(\boldsymbol{S}\boldsymbol{A})$, and so

$$\Phi = -\operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}\right) = \operatorname{tr}\left(\boldsymbol{S}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{\mu}\right), \tag{19}$$

which can be written into the following equivalent expressions, that does not involve the covariance matrix $\boldsymbol{S}$ nor its inverse explicitly:

$$\Phi = \operatorname{tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{Q}\right) = -\operatorname{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{Q}\right), \tag{20}$$

II. METHODS

A. Empirical covariance from fMRI data

The model is fitted to reproduce the two covariance matrices calculated from the empirical BOLD signals, with zero lag and a lag equal to 1 TR:

$$\widehat{S}_{ij}(0) = \frac{1}{T-2}\sum_{1\le t\le T-1}\left[x_i(t) - \bar{x}_i\right]\left[x_j(t) - \bar{x}_j\right] \ , \tag{21}$$

$$\widehat{S}_{ij}(1) = \frac{1}{T-2}\sum_{1\le t\le T-1}\left[x_i(t) - \bar{x}_i\right]\left[x_j(t+1) - \bar{x}_j\right] \tag{22}$$

Here $\bar{x}_i$ denotes the mean empirical signal: $\bar{x}_i = \frac{1}{T}\sum_t x_i(t)$ for all i, which is used to center the data as all variables x_i have mean zero in the model. These are the empirical counterparts of the model covariances $S_{ij}(t,t)$ and $S_{ij}(t,t+1)$ averaged over time t.

B. Parameter estimation of the MOU process

We fit the MOU process from the fMRI time series data for each subject in each sleep condition. We rely on a recent estimation method that tunes the MOU model such that its covariance structure reproduces the matrices in Eq. (21), optimizing its parameters the Jacobian matrix $-\boldsymbol{B}$ as well as the input covariance matrix $2\boldsymbol{D}$ [5]. Importantly, this optimization procedure incorporates topological constraints on $\boldsymbol{B}$, adjusting only existing anatomical connections, also keeping the input cross-covariances $D_{ij} = 0$ for $i \neq j$. Note that our current notation corresponds to a previous publication [5], using the following $-\boldsymbol{B} \leftrightarrow \boldsymbol{J}$ and $2\mathbf{D} \leftrightarrow \Sigma$; note that $-\boldsymbol{B} \leftrightarrow \boldsymbol{J}^{\mathrm{T}}$ in the subsequent paper [6].

The model is first calibrated by calculating the time constant τ from the empirical signals.

$$\tau = -\frac{N}{\sum_{1\le i\le N} a\left(v_i \mid u\right)} \ , \tag{23}$$

where $a\left(v_i \mid u\right)$ is the slope of the linear regression of $v_i = \left[\log\left(\widehat{S}^0_{ii}\right), \log\left(\widehat{S}^1_{ii}\right)\right]$ by $u = [0, 1]$.

We rely on a gradient descent to iteratively adjust $\boldsymbol{B}$ and $\boldsymbol{D}$ until reaching the best fit [5]. At each optimization step, we calculate the model counterparts of the covariance matrices in Eq. (21) $\boldsymbol{S}(0)$ and $\boldsymbol{S}(1)$, assuming stationarity over each fMRI session. They can be calculated by solving the Lyapunov equation using e.g. the Bartels-Stewart algorithm, which yields here

$$\boldsymbol{B}\,\boldsymbol{S}(0) + \boldsymbol{S}(0)\,\boldsymbol{B}^{\mathrm{T}} = 2\boldsymbol{D}\ , \tag{24}$$

once again equating the derivative with zero in Eq. (2), and the equation involving the propagator. We calculate the lagged covariance rewriting Eq. (3) for the time-lag equation here as

$$\boldsymbol{S}(1) = \boldsymbol{S}(0)\,e^{\boldsymbol{B}^{\mathrm{T}}}\ . \tag{25}$$

We then calculate the difference between the model and empirical covariances, $\Delta\boldsymbol{S}(t) = \hat{\boldsymbol{S}}(t) - \boldsymbol{S}(t)$ with $t \in \{0, 1\}$. The parameter update is given by differentiating Eqs. (25) and (24):

$$\begin{aligned}\Delta\boldsymbol{B} &= \epsilon_B\,[\boldsymbol{S}(0)]^{-1}\left[\Delta\boldsymbol{S}(0) - \Delta\boldsymbol{S}(1)\,e^{\boldsymbol{B}^{\mathrm{T}}}\right]\ , \\ \Delta\boldsymbol{D} &= \epsilon_D\boldsymbol{B}\,\Delta\boldsymbol{S}(0) + \epsilon_D\Delta\boldsymbol{S}(0)\,\boldsymbol{B}^{\mathrm{T}}\ ,\end{aligned} \tag{26}$$

with ϵ_B and ϵ_D small learning rates. The best fit corresponds to minimising the squared norm of both $\Delta\boldsymbol{S}(0)$ and $\Delta\boldsymbol{S}(1)$.

The model fitting is quantified by two measures: the model error defined using the matrix distance and the Pearson correlation between the vectorized FC matrices (model versus data). As shown in Fig S1, all sleep states have goodness of fit, with Pearson correlation above 0.6, corresponding to R2 of 0.36.

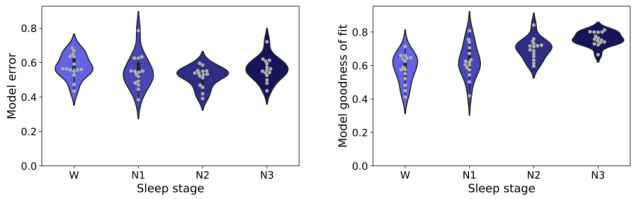

FIG. S1. **Model error and goodness of fit across sleep stages.** Left panel: Model error for each sleep stage (x-axis) across the 15 subjects. The sleep stages are represented by the blue contrasts, from light for wake (W) to dark for the deepest sleep (N3). Right panel: Goodness of fit measured by the Pearson correlation between the vectorized model and empirical FC matrices: $S(0)$ with $\hat{S}(0)$, and $S(1)$ with $\hat{S}(1)$. The average of the two Pearson correlation values is reported.

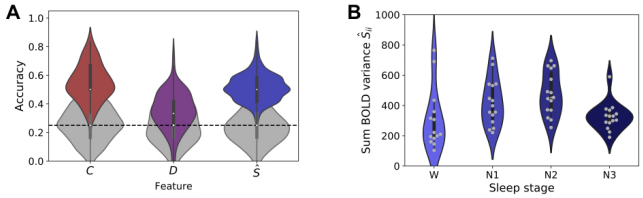

FIG. S2. **Model-based decoding performance outperforms phenomenological classification of sleep stages.** A) Similar plot to Fig 2C in the main text with the classification accuracy for the 1-nearest-neighbor (1NN) classifier, which relies on a similarity measure (here the Pearson correlation) between the input features to predict the class of the test sample. The x-axis indicates the features: the model estimates **C** and **D**, as well as the empirical FC denoted by $\hat{\mathbf{S}}$. B) Similar plot to Fig 2B in the main text but for the model input variance summed over all ROIs.

C. Decoding of sleep stages

We use the same approach as in previous work [6, 7]. We use the scikit-learn Python library for the implementations of classifiers [8].

We rely on two usual classifiers: multinomial logistic regression (MLR) and the 1-nearest-neighbor (1NN). The features correspond to vectorized **C**/**D**/**S** matrices after discarding zero or redundant elements. The MLR is a canonical tool for high-dimensional linear classification, which tunes a weight for each feature, thus selecting the important ones to discriminate the classes. In addition, we use L2-regularization for the MLR ($C = 1.0$ in the scikit-learn implementation). In contrast, the 1NN assigns to a new sample the class to which belongs its closest neighbors with respect to a similarity metric, here chosen as the Pearson correlation coefficient between the feature vectors. It thus relies on the global profile of features to cluster samples into classes.

Following standards, we use a stratified cross-validation scheme with 80% of the samples for the train set and 20% for the test set, where the ratio of classes are the same in both sets. We also use the subject identity as "group information" to avoid mixing subject data between the train and test sets. In practice, we use 50 random splits of the data an report the distribution of accuracies of the 50 splits (see the violin plots).

Fig S2 shows that the decoding of the sleep states by the 1NN classifier, which assigns to a new sample the class to which belongs its closest neighbors with respect to a similarity metric, here chosen as the Pearson correlation coefficient between the feature vectors. It thus relies on the global profile of features to cluster samples into classes.

D. Complementary analysis

The comparison of the nodal irreversibility across sleep states shows an overall decrease for all ROIs from W to N3 (Fig S3A), with a pronounced decreases from W to N1 (Fig S3B) and to a lesser extent from N2 to N3 (Fig S3D), although N1 and N2 are rather similar (Fig S3C). Importantly, we can see heterogeneity in the reduction of irreversibility across the ROIs in the transition to deep

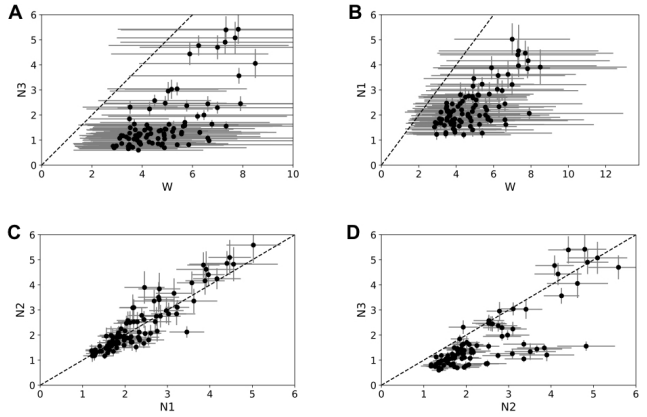

FIG. S3. **Comparison of ROI-specific nodal irreversibility across sleep states.** A) Plot of the sum of absolute values in Q over each row across the W and N1 states. The error bars indicate the s.e..m. across subjects. This plot is another view of the same data in Fig 3B (main text). B-C) Similar plots to panel A for N1 versus N2, N2 versus N3 and W versus N3.

sleep.

E. Resting-State fMRI signals

Participants

A total of 63 healthy subjects (36 females, mean SD, 2343.3 years) were selected from a data set previously described in a sleep-related study by Tagliazucchi and Laufs [9]. Participants entered the scanner at 7 PM and were asked to relax, close their eyes, and not fight the sleep onset. A total of 52 minutes of resting state activity were measured with a simultaneous combination of EEG and fMRI. According to the rules of the American Academy of Sleep Medicine [10], the polysomnography signals (including the scalp potentials measured with EEG) determine the classification of data into four stages (wakefulness, N1, N2 and N3 sleep). We selected 15 subjects with contiguous resting state time series of at least 200 volumes to perform our analysis. The local ethics committee approves the experimental protocol (Goethe-Universität Frankfurt, Germany, protocol number: 305/07), and written informed consent was asked to all participants before the experiment. The study was conducted according to the Helsinki Declaration on ethical research.

MNI data acquisition

MRI images were acquired on a 3-T Siemens Trio scanner (Erlangen, Germany) and fMRI acquisition parameters were 1505 volumes of T2-weighted echo planar images, $TR/TE = 2080\ ms/30\ ms$, matrix 6464, voxel size 3×3×3 mm^3, distance factor 50%; FOV 192 mm^2. An optimized polysomnographic setting was employed (chin and tibial EMG, ECG, EOG recorded bipolarly [sampling rate 5 kHz, low pass filter 1 kHz] with 30 EEG channels recorded with FCz as the reference [sampling rate 5 kHz, low pass filter 250 Hz]. Pulse oxymetry and respiration were recorded via sensors from the Trio [sampling rate 50 Hz]) and MR scanner compatible devices (BrainAmp MR+, BrainAmpExG; Brain Products, Gilching, Germany), facilitating sleep scoring during fMRI acquisition.

Brain parcellation AAL 90 to extract BOLD time series and filtering

To extract the time series of BOLD signals from each participant in a coarse parcellation, we used the AAL90 parcellation with 90 brain areas anatomically defined in [11]. BOLD signals (empirical or simulated) were filtered with a Butterworth (order 2) band-pass filter in the 0.01-0.1 frequency range.

[1] H. Qian, Mathematical formalism for isothermal linear irreversibility, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 10.1098/rspa.2001.0811 (2001).

[2] C. Godrèche and J. M. Luck, Characterising the nonequilibrium stationary states of Ornstein-Uhlenbeck processes, Journal of Physics A: Mathematical and Theoretical (2019).

[3] R. Cofré and C. Maldonado, Information entropy production of maximum entropy markov chains from spike trains, Entropy **20**, 10.3390/e20010034 (2018).

[4] H. Qian, M. Qian, and X. Tang, Thermodynamics of the general diffusion process: Time-reversibility and entropy production, Journal of Statistical Physics (2002).

[5] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco, Estimation of Directed Effective Connectivity from fMRI Functional Connectivity Hints at Asymmetries of Cortical Connectome, PLoS Computational Biology **12** (2016).

[6] M. Gilson, G. Zamora-López, V. Pallarés, M. H. Adhikari, M. Senden, A. T. Campo, D. Mantini, M. Corbetta, G. Deco, and A. Insabato, Model-based whole-brain effective connectivity to study distributed cognition in health and disease, Network Neuroscience **4** (2020).

[7] V. Pallarés, A. Insabato, A. Sanjuán, S. Kühn, D. Mantini, G. Deco, and M. Gilson, Extracting orthogonal subject- and condition-specific signatures from fMRI data using whole-brain effective connectivity, NeuroImage **178** (2018).

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour-

napeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12** (2011).

[9] E. Tagliazucchi and H. Laufs, Decoding Wakefulness Levels from Typical fMRI Resting-State Data Reveals Reliable Drifts between Wakefulness and Sleep, Neuron (2014).

[10] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. Lloyd, C. L. Marcus, and B. V. Vaughn, Aasm — scoring manual version 2.2 the aasm manual for the scoring of sleep and associated events. rules, terminology and technical specifications., American Academy of Sleep Medicine **176** (2015).

[11] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, NeuroImage **15**, 10.1006/nimg.2001.0978 (2002).