Supporting Information

Supplementary Methods

Experimental procedure

Hierarchical auditory oddball task (Paris 1 & 2 dataset): Task-related EEG signals were obtained from the 'Local-Global' protocol (Bekinschtein et al., 2009) designed to study unconscious and conscious auditory processing. Accordingly, brain responses to two types of auditory events were recorded: automatically processed short-time-range violations and long-time-range violations whose recognition depends on explicit working memory effort. For optimal cognitive performance, patients underwent recordings at least 24 hours after sedation discontinuation. Medication that could potentially modify the EEG, e.g., myorelaxants and anti-epileptic drugs were not controlled. At the beginning of each recording session, EEG signal quality was assessed. If seizures or any other clearly identifiable abnormal activity were observed, the recording was stopped. EEG recordings were sampled at 250 Hz with a 256-electrode geodesic sponge sensor net (EGI) referenced to the vertex. Recordings were band-pass filtered (from 0.5 to 45Hz using a 6 and 8 order FFT-based Butterworth filter). Data were then epoched from -200*ms* to 1336*ms* relative to the onset of the first sound. Epochs were excluded based on adaptive outlier detection as in Engemann et al. (2015). Subsequently, data were re-referenced using an average reference and baseline correction was applied.

Task-free recordings (Liege dataset): Data were chunked into 1536*ms* pseudo-epochs, matching the length of the task-based epochs, with random intervals between epochs matching the inter-trial intervals of the auditory task. Otherwise, the same acquisition preprocessing scheme was applied.

Statistical Analysis

Statistical Inference: We extended our visualizations into hypothesis tests by employing the percentile bootstrap (Efron & Tibshirani, 1993). Accordingly, we generated 2000 bootstrap samples by drawing with uniform probability and replacement *n* samples from the dataset. The test-statistic of interest was then evaluated on each bootstrap sample. Two-sided 95% confidence intervals were obtained by querying the 2.5 and 97.5 percentiles and the significance-level was then obtained by inversion of the confidence interval that excluded the value under H0. We denoted difference-statistics by ***D***. For correlation analysis, we relied on the non-parametric Spearman's Rank correlation coefficient. In the latter case, we obtained the significance level from analytical p-values.

To assess out-of-sample generalization we used two complementary approaches: A conservative validation on independent data (new cohorts, different protocols and laboratories) and cross-validation. For cross-validation we used a group Monte Carlo sampling scheme with a training set size of 80 percent, a testing set size of 20 percent and 50 iterations. The Monte Carlo procedure is known to minimize estimation variance and has been shown to yield low positive cross-validation bias (Varoquaux et al., 2016). The grouping consisted in exclusively assigning subjects to either the test or the train sets. For assessing the generalization capacity of the DOC-Forest on new datasets we contrasted the performance against empirically estimated chance levels. These were obtained from comparisons against a dummy classifier that generated random predictions based on the observed class-probabilities.

Area under the Curve metric: Univariate and multivariate discrimination performance was summarized with Area Under the Curve (AUC) calculated from the receiver operator characteristic (ROC). For a binary classification system, the ROC pits the detection probability, commonly referred to as *sensitivity* against the probability of false alarm (1 - *sensitivity*). These probabilities are empirically estimated by moving the decision cut-off along the sorted values of a continuous variable, e.g. a score, and evaluating its relation to the true label. In the case of traditional model-free univariate analysis, the score is the EEG-marker itself, in the

case of univariate or multivariate machine learning it is the predicted probability of a given sample to belong to the target class. The AUC can then be conveniently used to summarize the performance, where a score of 0.5 is uninformative and equals to random guessing whereas a score of 1 amounts to perfect classification and 0 to total confusion, indicating negative correlation between the score and the label. We used the AUC in two different contexts, once to summarize the class-probabilities issued by our classification models, once in a direct fashion on single marker values without a classification model. Note that in the latter context, markers often show univariate AUC scores smaller than chance level (0.5) because they are conceptually related to absence of consciousness. To avoid confusion, we rectified the direct AUC in that case by *abs(AUC – 0.5) + 0.5*. Contrastingly, for classification models, AUC values smaller than 0.5 can indicate inconsistencies of patterns between training and testing data or result from low performance of the learned decision rules.

Furthermore, note that the AUC of a clinical score regarding two groups of patients' amounts is closely related to computing the Mann-Whitney-U statistic and can be directly obtained from dividing U by the product of the two groups sample sizes:

$$AUC = \frac{U}{n_1 n_2}$$

Multivariate Pattern Classification: We chose the *Extra-Trees* algorithm (Geurts, Ernst, & Wehenkel, 2006) because this algorithm is well-established and belongs to the most popular machine learning techniques for regression and classification problems, next to Random Forests (Breiman, 2001), Support Vector Machines (SVM) and penalized linear models. We chose this algorithm to improve the robustness of classification. Moreover, we found randomized classification trees to achieve, ad-hoc, without tuning hyper-parameters and without feature-selection a performance equivalent to an SVM with tuned regularization parameter and with explicit feature-selection (Engemann et al., 2015). *Extra-Trees* are non-parametric and robust by design and are not sensitive to the measurement scale of the input data. This algorithm can handle so-called wide

datasets in which more variables than samples are available. Moreover, *Extra-Trees* belong to the family of adaptive algorithms capable of scaling the complexity of the learned model to the amount of data available. The *Extra-Trees* algorithm achieves its efficiency by generalizing the non-linear decision tree approach. Single decision trees are non-parametric rule-based models that automatize variable selection and can be thought of as learning a "regression surface" from the data by recursive orthogonal partitioning (Efron & Hastie, 2016). In other words, decision trees map joint value ranges of the input variables to values of the outcome variable. However, decision trees poorly generalize to new data. The *Extra-Trees* retains all benefits of decision trees while mitigating their excessive variance and poor generalization capability. This is achieved by averaging over many randomly constructed, hence uncorrelated, decision trees, each of which is "grown" on randomly drawn subsets of m input variables when looking for candidate splits at internal nodes (typically, $m = \sqrt{p}$). Unlike Random Forests, their historical predecessor, the Extremely Randomized Trees algorithm does not make use of randomization across samples via bootstrapping. It instead achieves additional randomization at the level of the thresholds used when constructing the trees, which can improve approximation of a fully randomizes tree and can facilitate its interpretation (Louppe, Wehenkel, Sutera, & Geurts, 2013). This principle, effectively, permits a random search through a combinatorial space of variables. Therefore, the *Extra-Trees* algorithm lends itself to exploit interactions between variables if sufficient data is available. Consequently, the ensuing model often consists of thousands of trees which practically renders interpretation difficult.

To improve the interpretability of randomized classification trees, the so-called variable importance metric has been proposed and can be readily computed for a fitted model. Intuitively, the importance score of a variable can be understood as the weighted reduction of entropy of the outcome, over all the internal nodes where that variable has been used for making a split (Louppe et al., 2013). Variable importance can be shown to correspond to a weighted average of the mutual information between that variable and the outcome, conditionally over any possible configuration of any subset of the other variables if entropy is used as impurity criterion. Variable importance is of considerable interest, as its conditional (multivariate) nature complements

marginal (univariate) statistics. In other words, variable importance is multivariate and accordingly a variable can be important either because of a univariate (marginal) correlation with the outcome, or because its interdependencies with other variables are informative about the outcome. This potentially facilitates discovery of interesting variables which would be overlooked otherwise and is a welcomed remedy to the problem that predictive variables are not necessarily significant (Bzdok, Engemann, Grisel, Varoquaux, & Thirion, 2018; Lo, Chernoff, Zheng, & Lo, 2015). Some cautious interpretation is advised as masking effects can occur and not all necessarily influential variables are captured by the variable importance under default settings, especially when using Random Forests (Louppe et al., 2013). However, this effect can be mitigated by using only one variable for splitting (K or max_features = 1), by constraining the maximum tree depth to 3-5 and by approximating full randomization by using Extremely Randomized Trees (Louppe, 2014). This approximation of full randomization has the advantage that the variable importance is only driven by relevant variables, not irrelevant ones and, when combined with entropy as impurity criterion, its interpretation as mutual information holds (Louppe et al., 2013). As performance usually stabilizes at a certain point as trees are added to the model a trade-off between performance and speed has to be made (Geurts, Ernst, & Wehenkel, 2006). Profiling suggested stable and reasonable performance after 1000-2000 trees.

In the present study, we hence used the Extremely Randomized Trees with 2000 trees, a tree depth of four, entropy as impurity criterion, and a single feature for splitting. Otherwise we used default parameter values which have been shown to be optimal in most situations and are typically not recommended to be tuned to not jeopardize the computational benefits of Extra Trees over other randomized tree techniques (Geurts et al., 2006). For a detailed description of default values please consider the documentation of the Scikit-Learn software for machine learning (Pedregosa et al., 2011).

Case report of the cognitive motor dissociation patients

Patient P1 This male 40 years old patient was admitted to the hospital with the diagnosis of UWS 11 months after traumatic brain injury resulting from a car accident. The patient behaviorally assessed with the CRS-R five times within a week, during which he showed auditory (1/5) and oral (4/5) reflexive behavior. The patient showed eye opening 3 out of 5 assessments and was diagnosed as comatose the other two assessments. The results of the structural neuroimaging highlighted micro-hemorrhages and diffuse axonal injury. The ventricles were moderately enlarged and atrophy was pronounced in the midbrain and around the cortical sulci. Functional MRI results suggested the presence of brain activity consistent with covert response to command as tested with the tennis paradigm, yet, no activation of resting state networks could be detected.

Patient P2 This female, 36 years old patient was admitted to the hospital with the diagnosis of UWS six years and two months after a post-ischemic vertebrobasilar stroke. The patient was assessed with the CRS-R 5 times within the week of hospitalization, during which she showed motor (4/5) and oral (5/5) reflexes. The patient showed eye opening during all assessments. Structural MRI revealed the presence of severe ischemic lesions. Atrophy was most pronounced at the upper part of the cerebellum, while the frontal, parietal and striatum seemed relatively spared. Ventricles were enlarged. Functional MRI results suggested the presence of brain activity consistent with covert response to command as tested with the tennis paradigm, in addition, activation of the default mode resting state network could be detected.

Supplementary Results

Consistency of classification in diagnostic sub-groups

In order to assess classification accuracy for sub-groups we performed cross-validation on the Paris 1 dataset as described in the main text based on the full EEG-configuration. In each fold we subdivided the test set

according to diagnostic groups and computed separately the average AUC metric across all 50 folds. We first considered the three largest etiological groups (anoxia, stroke and TBI). We obtained an AUC of 0.77 (SEM=0.014) for anoxia patients (n = 34), an AUC of 0.83 (SEM = 0.024) for stroke patients (n = 47) and an AUC of 0.68 (SEM = 0.036) for TBI patients (n = 38). We then subdivided the dataset in acute patients (delay <= 30, n = 71) and chronic patients (delay > 30, n = 71). For acute patients we obtained an AUC of 0.78 (SEM = 0.023) and for chronic patients an AUC of 0.80 (SEM = 0.017).

Detailed comparison between individual markers and DOC-Forest

We assessed how each marker discriminated between UWS and MCS patients as a function of EEG-configurations using the task-EEG dataset recorded in Pitié-Salpêtrière (Paris 1 dataset). Comparing the cross-validated model-based AUC for each marker with the model-free AUC (Figure S1A) revealed a tight positive correlation ($rho_{Spearman}$ = 0.94, 95% CI[0.905, 0.962], p < 0.001). This finding suggests that our univariate forest models performed equivalently to the previous model-free method. Subsequent analyses suggested that, on average, over all EEG, the DOC-Forest performance assumed at least the 98th percentile compared to other markers ($Percentile_{DOC\text{-}Forest}$ = 0.985, 95% CI[0.981, 0.989]) and its fluctuation over configurations was situated around the 83rd percentile ($Percentile_{DOC\text{-}Forest}$ = 0.832, 95% CI[0.796, 0.894]).

Tracking the discrimination performance between UWS and MCS across the EEG configurations revealed fluctuations according to marker subtype, conceptual family (Figure S1, Figure S2) and variance induced by the EEG configuration itself (cf. Figure 2A and 3B in the main text). Indeed, we found a positive correlation between fluctuations of marker estimates and discrimination performance across EEG configurations ($rho_{Spearman}$ = 0.46, 95% CI[0.289, 0.6], p < 0.001, suggesting that performance was more stable when the marker itself was stable too (Figure S1B and S2). Interestingly, evoked response markers appeared more severely affected than other markers.

Supplementary Figures

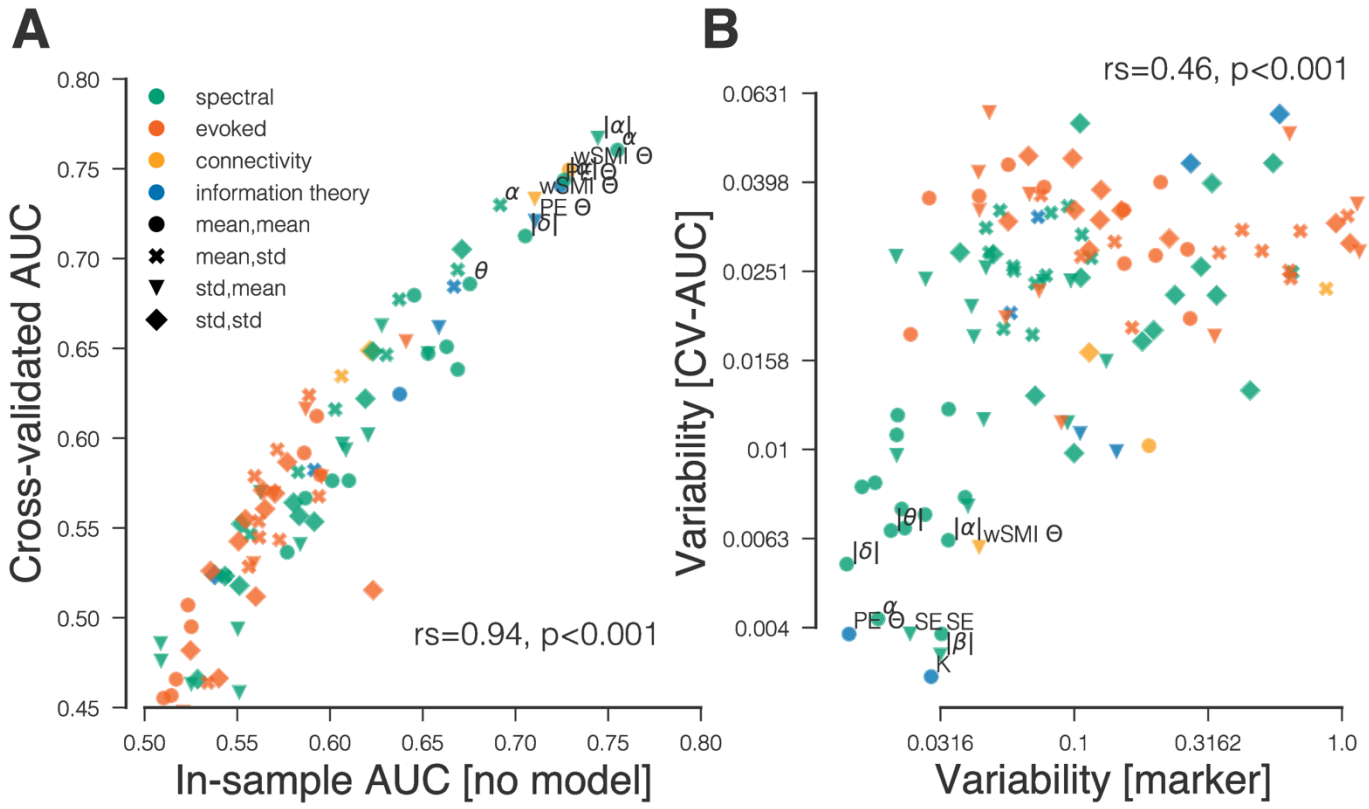

Figure S1: Classification with univariate models. (A) Comparison between traditional model-free classification on the training data using the univariate AUC metric as in Sitt, King et al. 2014 and estimation of out-of-sample performance using univariate classifiers and cross-validation. A positive relationship between model-free in-sample performance and univariate out-of-sample performance was observed. Markers that performed better as measured by traditional model-free AUC also showed higher cross-validated performance. This relationship was less tight at lower performance. Note that markers performing close to chance level tended to score below a cross-validated AUC of 0.5. The 10 best markers are indicated with labels for convenience. **(B)** Relationship between relative marker variance and performance over 36 EEG configurations displayed on double logarithmic scale. Markers at each configuration were standardized to the reference configuration of 100% epochs and 256 sensors. A non-linear positive relationship emerged, suggesting that markers whose values were more strongly changed by the EEG configuration also showed stronger fluctuations in performance. The 10 least variable markers are labeled for convenience.

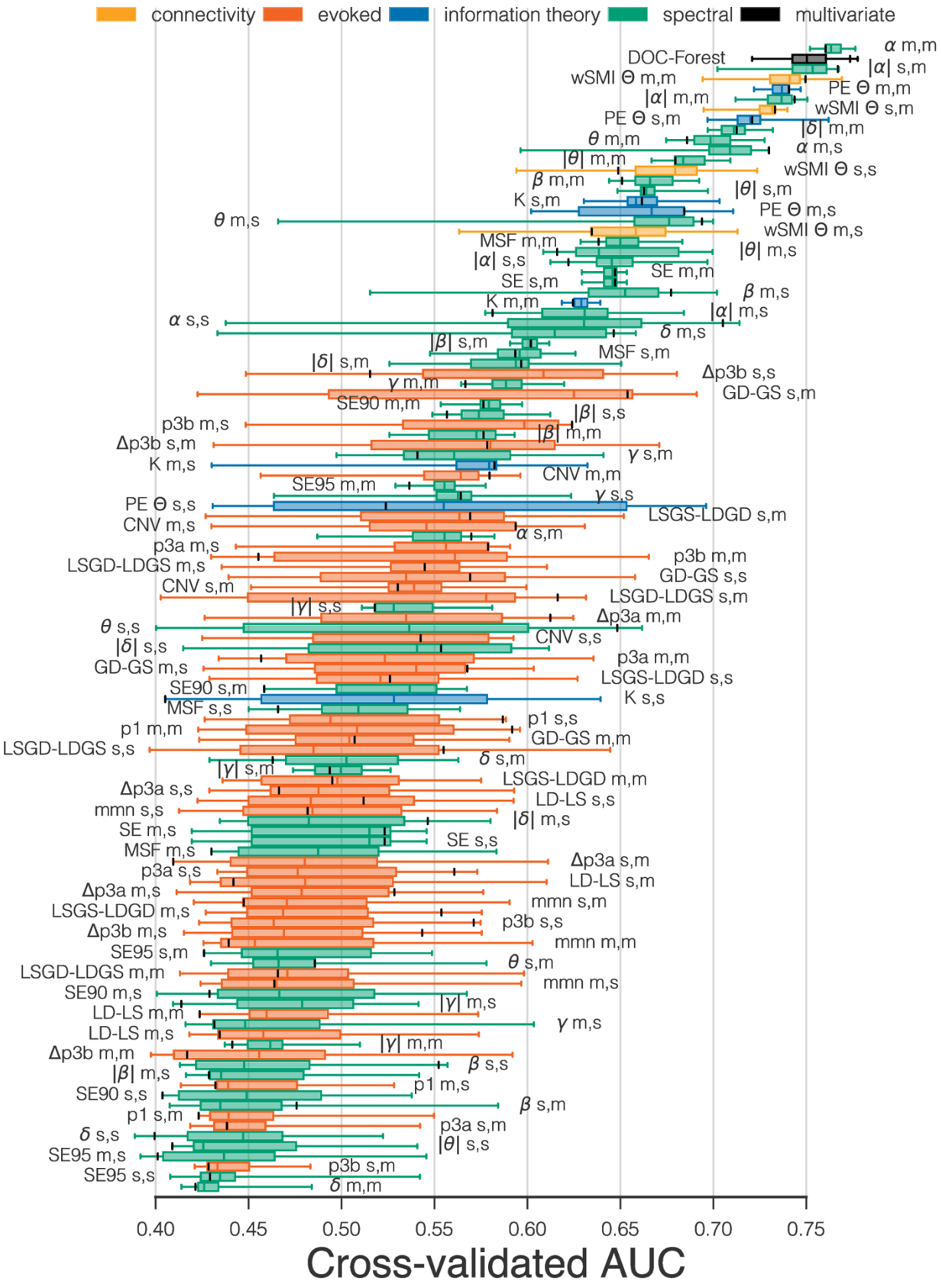
connectivity
evoked
information theory
spectral
multivariate
DOC-Forest
wSMI Θ m,m
|α| m,m
PE Θ s,m
θ m,m
|θ| m,m
β m,m
K s,m
θ m,s
MSF m,m
|α| s,s
SE s,m
K m,m
α s,s
|β| s,m
|δ| s,m
γ m,m
SE90 m,m
p3b m,s
Δp3b s,m
K m,s
SE95 m,m
PE Θ s,s
CNV m,s
p3a m,s
LSGD-LDGS m,s
CNV s,m
|γ| s,s
θ s,s
|δ| s,s
GD-GS m,s
SE90 s,m
MSF s,s
p1 m,m
LSGD-LDGS s,s
|γ| s,m
Δp3a s,s
mmn s,s
SE m,s
MSF m,s
p3a s,s
Δp3a m,s
LSGS-LDGD m,s
Δp3b m,s
SE95 s,m
LSGD-LDGS m,m
SE90 m,s
LD-LS m,m
LD-LS m,s
Δp3b m,m
|β| m,s
SE90 s,s
p1 s,m
δ s,s
SE95 m,s
SE95 s,s
α m,m
|α| s,m
PE Θ m,m
wSMI Θ s,m
|δ| m,m
α m,s
wSMI Θ s,s
|θ| s,m
PE Θ m,s
wSMI Θ m,s
|θ| m,s
SE m,m
β m,s
|α| m,s
δ m,s
MSF s,m
Δp3b s,s
GD-GS s,m
|β| s,s
|β| m,m
γ s,m
CNV m,m
γ s,s
LSGS-LDGD s,m
α s,m
p3b m,m
GD-GS s,s
LSGD-LDGS s,m
Δp3a m,m
CNV s,s
p3a m,m
LSGS-LDGD s,s
K s,s
p1 s,s
GD-GS m,m
δ s,m
LSGS-LDGD m,m
LD-LS s,s
|δ| m,s
SE s,s
Δp3a s,m
LD-LS s,m
mmn s,m
p3b s,s
mmn m,m
θ s,m
mmn m,s
|γ| m,s
γ m,s
|γ| m,m
β s,s
p1 m,s
β s,m
p3a s,m
|θ| s,s
p3b s,m
δ m,m
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Cross-validated AUC

Figure S2: Discrimination performance of EEG-markers across experiments. The cross-validated AUC for all 112 EEG-markers of consciousness and the DOC-Forest across all 36 EEG configurations on the Paris 1 dataset ordered by average performance. The marker family is indicated by the colors. The marker subtype by the letter suffix, e. g., mean over epochs and standard deviation over channels reads “m,s”. The boxplot whiskers indicate value ranges. The black notches indicate performance at the reference configuration of 100% epochs and 256 sensors. It can be seen that some markers rather improved their performance as sensors and epochs were changed (e.g. θ m,m) while others rather decreased their performance (e.g., α m,s). Note that performance fluctuated less when markers performed higher on average.

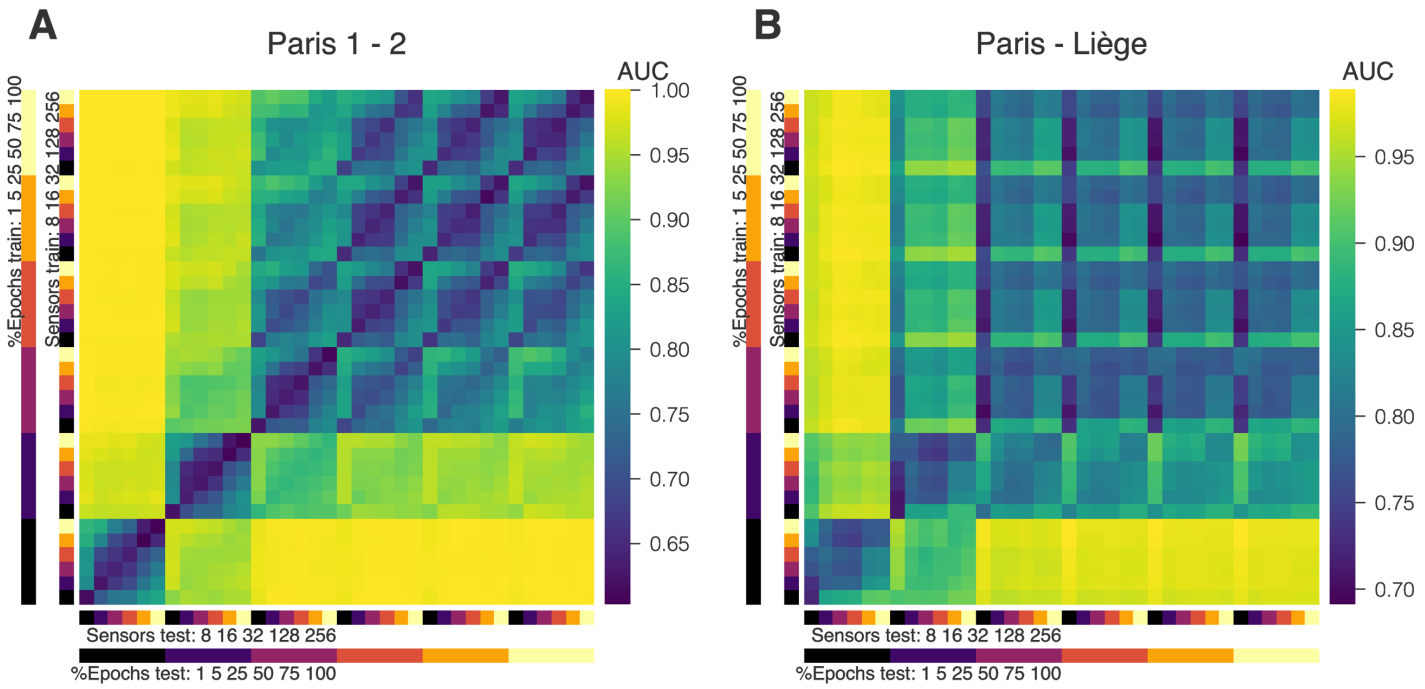

Figure S3: Distribution shifts from mismatching EEG-configuration. To ensure that using different combinations of sensors and epochs for training and testing induces nontrivial differences between the datasets, we applied the DOC-Forest to the origin of an EEG-recording instead of the diagnosis of the patient using 5-fold cross-validation over all combinations of EEG-configurations (**A** Paris 1 versus 2, **B** Paris versus Liège). It can be seen that the origin is almost perfectly decodable when EEG-configurations maximally diverge, suggesting that changing the number of sensors and epochs impacts the distribution of the datasets.

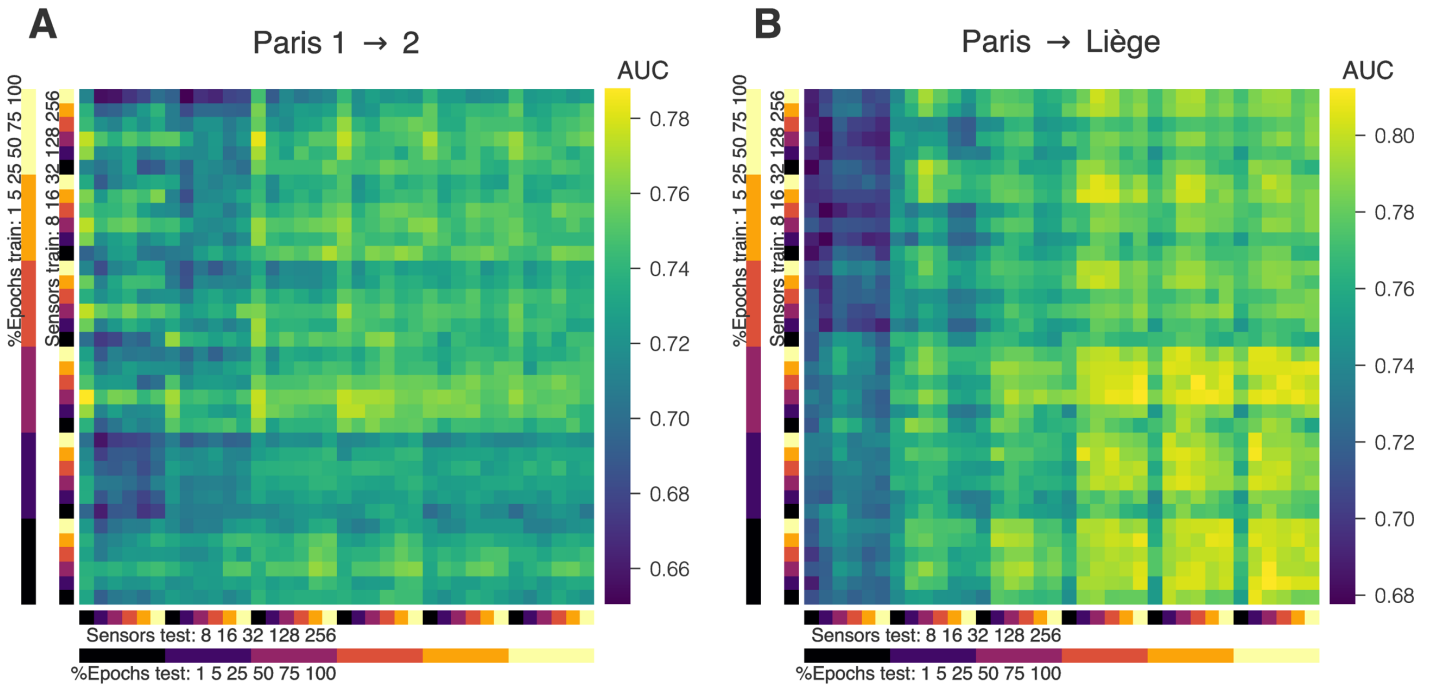

Figure S4: Cross-configuration generalization. (A) DOC-Forest generalization performance when using different EEG configurations for training on Paris 1 (rows) and testing on Paris 2 (columns). The 36 combinations of sensors and epochs are ordered such that, from the left corner, epochs grow slowly (1% to 100%) while sensors repeat (1 to 256). We observed reasonable generalization performance for majority of combinations of EEG-configuration. Note the horizontal and vertical stripes in the upper right part of the matrix, which point out peak performance with fewer sensors (8-32) and at least 25% of the epochs. **(B)** The same analysis for generalization from the combined Paris 1 & 2 dataset to the Liège dataset. Again, reasonable generalization performance was observed for the majority of combinations. However, the overall generalization pattern was markedly different, suggesting peak performance when less epochs were used on for training but more epochs and at least 16 sensors for testing.

References

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0809667106

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bzdok, D., Engemann, D.-A., Grisel, O., Varoquaux, G., & Thirion, B. (2018). Prediction and inference diverge in biomedicine: Simulations and real-world data. *BioRxiv*. https://doi.org/10.1101/327437

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference* (Vol. 5). Cambridge University Press.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Engemann, D., Raimondo, F., King, J.-R., Jas, M., Gramfort, A., Dehaene, S., … Sitt, J. (2015). Automated Measurement and Prediction of Consciousness in Vegetative and Minimally Conscious Patients. In *Automated Measurement and Prediction of Consciousness in Vegetative and Minimally Conscious Patients*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees.

Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, *112*(45), 13892–13897.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Understanding variable importances in forests of randomized trees* (pp. 431–439).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*.

Wannez, S., Heine, L., Thonnard, M., Gosseries, O., Laureys, S., & Coma, S. G. collaborators. (2017). The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Ann Neurol*, *81*(6), 883–889.