Validation of non-REM sleep stage decoding from resting state fMRI using linear support vector machines

Altmann A.[1,2,7]*, Schröter M.S. [1,3]*, Spoormaker V.I.[1], Kiem S.A.[1], Jordan D.[4], Ilg R.[5,6], Bullmore E.T.[3], Greicius M.D.[2], Czisch M.[1], Sämann P.G.[1]

*equal contribution

[1] Max Planck Institute of Psychiatry, Department of Translational Research in Psychiatry, Neuroimaging, Munich, Germany

[2] Stanford Center for Memory Disorders, Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

[3] Behavioural and Clinical Neuroscience Institute, Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

[4] Department of Anesthesiology, Klinikum rechts der Isar, Technische Universität München, Munich, Germany

[5] Department of Neurology, Klinikum rechts der Isar, Technische Universität München, Munich, Germany

[6] Asklepios Stadtklinik, Bad Tölz, Germany

[7] Present address: Translational Imaging Group, Centre of Medical Image Computing, University College London, London, UK

Running title: 'Resting state fMRI predicts sleep'

Key Words: resting state fMRI, EEG, EEG-fMRI, sleep, classification

Corresponding author: Andre Altmann

Translational Imaging Group

4 Stephenson Way

NW1 2HE, London, United Kingdom

E-mail: a.altmann@ucl.ac.uk

Phone: +44 7758 542270

This is a pre-copyedited, author-produced PDF of an article accepted for publication in "NeuroImage" following peer review. The version of record "Altmann, A., et al., Validation of non-REM sleep stage decoding from resting state fMRI using linear support vector machines, NeuroImage (2015)" is available online at: http://dx.doi.org/10.1016/j.neuroimage.2015.09.072

Abstract

A growing body of literature suggests that changes in consciousness are reflected in specific connectivity patterns of the brain as obtained from resting state fMRI (rs-fMRI). As simultaneous electroencephalography (EEG) is often unavailable, decoding of potentially confounding sleep patterns from rs-fMRI itself might be useful and improve data interpretation. Linear support vector machine classifiers were trained on combined rs-fMRI/EEG recordings from 25 subjects to separate wakefulness (S0) from non-rapid eye movement (NREM) sleep stages 1 (S1), 2 (S2), slow wave sleep (SW) and all three sleep stages combined (SX). Classifier performance was quantified by a leave-one-subject-out cross-validation (LOSO-CV) and on an independent validation dataset comprising 19 subjects. Results demonstrated excellent performance with areas under the receiver operating characteristics curve (AUCs) close to 1.0 for the discrimination of sleep from wakefulness (S0|SX), S0|S1, S0|S2 and S0|SW, and good to excellent performance for the classification between sleep stages (S1|S2:~0.9; S1|SW:~1.0; S2|SW:~0.8). Application windows of fMRI data from about 70 seconds were found as minimum to provide reliable classifications. Discrimination patterns pointed to subcortical-cortical connectivity and within-occipital lobe reorganization of connectivity as strongest carriers of discriminative information. In conclusion, we report that functional connectivity analysis allows valid classification of NREM sleep stages.

1. Introduction

Studying brain functional connectivity by the use of resting state functional magnetic resonance imaging (rs-fMRI) has become a widely used analytical tool to investigate brain function in vivo (Biswal et al. 2010; van den Heuvel and Hulshoff Pol 2010). Recently, rs-fMRI connectivity has been proposed as an intermediate phenotype for psychiatric and neurological disorders (Meyer-Lindenberg and Weinberger 2006; Seeley et al. 2009; Fornito et al. 2013). However, the quest of translating rs-fMRI to biomarker applications and its use as a phenotyping technique requires standardization of acquisition schemes and postprocessing strategies that control for various signals of non-neuronal origin.

Besides minimizing potentially spurious results through adequate consideration of cardiovascular signals, motion and global signals (Cole et al. 2010; Van Dijk et al. 2012), there remain factors during rs-fMRI acquisition itself that impact the reliability of functional connectivity measures: Besides acquisition length (Van Dijk et al. 2010), one potential obstacle during the acquisition of rs-fMRI data is the type of experimental instruction with regard to eyes closed versus eyes open (with or without fixation) (Patriat et al. 2013; Tagliazucchi and Laufs 2014). Most functional connectivity analyses further assume an analytical signal that is relatively stationary within the recording interval (see Hutchison et al. (2013)), which is in contrast to observations of fluctuating cognitive states and vigilance levels during states referred to as 'wakeful resting' (Christoff et al. 2009; Christoff 2012). By now these temporal fluctuations in functional connectivity have become a subject of investigation themselves (Chang and Glover 2010). Changes of vigilance levels are especially likely if the resting-state scan is either performed after several other MR acquisitions during which no cognitive task is performed, or if it follows a series of cognitively demanding experimental tasks: subjects may then change their focus and drift from wakefulness into periods of light sleep. Additional variability may be expected in patient samples with current medication that promotes or inhibits sleep, symptoms affecting sleep or comorbid sleep disorders, or anxiety during the scan that increases vigilance.

Although the brain's high metabolic activity and most spontaneous activity as measured with fMRI is preserved during periods of light sleep or low-level sedation (Fukunaga et al. 2006; Greicius et al. 2008), there is now an increasing body of evidence for vigilance-stage specific changes of functional connectivity (see Picchioni et al. (2013) for a review). In line with results of earlier studies using simultaneous electroencephalography (EEG) and positron emission tomography (PET) (Braun et al. 1997; Maquet 2000), recent EEG-fMRI experiments reported decreases in frontoparietal (Horovitz et al. 2009; Larson-Prior et al. 2011; Sämann et al. 2011) and thalamocortical connectivity (Spoormaker et al. 2010; Picchioni et al. 2014), but preserved or increased functional connectivity of unimodal sensory cortices (Larson-Prior *et al.* 2011) during non-rapid-eye movement (non-REM) sleep. In REM sleep, functional thalamocortical connectivity was found to differentiate tonic and phasic REM sleep (Wehrle et al. 2007). Furthermore, in non-REM sleep a sleep-stage specific reorganization of nodes of the default mode network (DMN) and its anti-correlated network has been reported (Sämann *et al.* 2011), with a dynamic functional connectivity analysis pointing to significant changes in connectivity already during transition to sleep stage 1. This is closely related to the direct effect of objectively lower vigilance and increased sleep pressure, as induced by sleep deprivation, which modulates functional connectivity of the DMN and other resting state networks (Sämann et al. 2010; De Havas et al. 2012). Finally, higher order measures of network features derived from rs-fMRI time series such as small-worldness (Spoormaker *et al.* 2010) or temporal autocorrelation patterns (Tagliazucchi et al. 2013) were reported to undergo changes in sleep.

To date, most rs-fMRI studies do not control for the state of vigilance during the scan, or document whether subjects have had normal sleep before the scan or control for the subject's circadian rhythm (e.g., morningness-eveningness type). A recent study revealed that in about one third of typical resting state scans loss of wakefulness occurs within 3 minutes (Tagliazucchi and Laufs 2014). Taking into account the aforementioned results of studies on sleep and sleep deprivation, with considerable effect sizes for functional connectivity changes, it is likely that changes in vigilance systematically confound functional connectivity analyses. This risk holds particularly for case-control studies in

which one group is more inclined to have altered vigilance, perhaps due to disease status, medication effects or potential comorbidity with sleep disorders. Similarly, the interpretability of genetic associations gained from imaging genomic samples may be impacted by this confound.

Besides behavioral assessment such as the psychomotor vigilance task, EEG is the most objective method to track levels of vigilance, sleep onset and sleep depth (e.g., according to the criteria by Rechtschaffen and Kales (1968) that recently have been revised by the American Academy of Sleep Medicine (AASM 2007)). However, the special equipment for simultaneous EEG-fMRI acquisition is often unavailable, and even if available, setup of the EEG device is uncomfortable and time-consuming, thus posing an additional burden on fMRI studies. Hence, if behavioral state is controlled at all, most studies rely on post-acquisition self-reported vigilance levels (Fox and Raichle 2007). However, such subjective ratings of sleep features may be unreliable as known from clinical observations (Vanable et al. 2000).

A recent development in the analysis of rs-fMRI data is the application of multivariate models such as support vector machines (SVM) (Pereira et al. 2009; Richiardi et al. 2011). Examples for the application of multivariate models in the analysis of rs-fMRI include maturation of the brain (Dosenbach et al. 2010), classifying healthy controls from patients with various psychiatric and neurodegenerative disorders (Supekar et al. 2008; Cecchi et al. 2009; Richiardi et al. 2012) or to decode task specific brain activities in healthy subjects (Richiardi *et al.* 2011; Shirer et al. 2012). Recently, Tagliazucchi et al. (2012) presented a report on predicting sleep stages from connectivity patterns in rs-fMRI using a pre-defined set of functional regions-of-interest (ROIs) and SVM with non-linear kernels. Later, Tagliazucchi and Laufs (2014) entered a broader set of connections from a whole brain parcellation to train and later apply SVM-based classifiers on independently acquired rs-fMRI samples.

In line with this work, the present study further elaborates the possibilities of classifying sleep-wake states within rs-fMRI data. To this end, we trained a statistical model for inferring the respective sleep stage on the basis of a previously published EEG-validated rs-fMRI dataset and a second

independent sample acquired later. We systematically assessed the classification performance by using separate training and validation data, and employing a leave-one-subject-out cross-validation (LOSO-CV) procedure to avoid overestimating the classifier's performance. Moreover, we investigated the minimum acquisition length required for reliable predictions and unlike previous studies we did not use the same length for training and applying the classifier but investigated effects of different training length and application length separately.

For training the sleep-stage classifier, we made use of linear support vector machines (SVMs), which have the inherent advantage that resulting models are interpretable, i.e., one can derive a weight value per single rs-fMRI connection that can be interpreted in terms of importance of that connection for the classification. While previous work has successfully demonstrated the feasibility of predicting sleep stages from rs-fMRI connectivity patterns using support vector classification (Tagliazucchi *et al.* 2012), the black-box nature of the employed non-linear SVM kernels hampers the elucidation of the respective underlying predictive connectivity networks. Although there are methods for deriving importance values of single features even from such black-box classifiers (Guyon et al. 2002), these are often computationally intensive and typically do not take advantage of the non-linear nature of the classifier. Even though interpreting linear SVM is feasible in theory, it is often challenging with a large number of features in play. This holds in particular for settings in which the features themselves are highly correlated among each other, such as functional connections in the brain. Thus, in most empirical settings the actual importance value of correlated features is distributed across the entire correlated group (Altmann et al. 2010; Tolosi and Lengauer 2011). As a result in our application, a single connection may seem unimportant, while in fact it merely belongs to a larger correlated group of important functional connections. Thus, in addition to extracting importance weights for single connections, we employed the concept of reduced and increased connectivity indices (Richiardi *et al.* 2012) to provide a two dimensional view of resulting discriminative connectivity patterns.

2. Material and Methods

2.1. EEG-fMRI Acquisition, EEG Sleep Stage Rating and FMRI Preprocessing

EEG-fMRI acquisition: We used two datasets comprising simultaneous EEG-fMRI recordings across wakefulness and NREM sleep stages. The training set was based on previously published recordings (Spoormaker *et al.* 2010) (25 subjects, 24.7±2.8 years, instructed to try to fall asleep). A second, independent validation set was derived from an EEG-fMRI study unrelated to sleep that included longer resting state recordings (Kiem et al. 2013) (19 males, mean 27.0±2.7 years, instructed to try to stay awake). In both cases, subjects were instructed to rest in the scanner with their eyes closed. Recording times were 9 pm in the training sample (Spoormaker *et al.* 2010) and 4 pm in the validation sample (Kiem *et al.* 2013). EEG-fMRI acquisition techniques of the two datasets were highly similar and postprocessing steps as detailed below were identical. Polysomnographic recordings were performed using an MR-compatible EEG system placed according to the international 10/20-electrode system with 19 channels and additional electrooculography, submental electromyography and electrocardiography (sampling rate 5 kHz; bandpass filter of 0.5–30 Hz; EasyCAP modified for sleep, Hersching, Germany; VisionRecorder Version 1.03, BrainProducts, Gilching, Germany). After offline artefact removal (VisionAnalyzer 1.05, Brain Products, Gilching, Germany), EEG data was scored according to the criteria of Rechtschaffen and Kales (1968) in 20-s epochs (Vision RecView, BrainProducts, Gilching, Germany). FMRI of both datasets were recorded on a 1.5 Tesla clinical MR scanner (General Electric, Milwaukee, USA; training dataset: model Signa; validation dataset: recorded after upgrade to Signa Excite) using echo planar imaging through an 8-channel head coil with parameters as follows: Training dataset: 25 AC-PC-oriented oblique slices, 3 mm thickness, 1 mm gap, TR 2000 ms, TE 40 ms, flip angle 90°, FOV 20 × 20 cm^2, 64 × 64 matrix, in-plane resolution 3.44 × 3.44 mm^2) obtained as time series of 800 images (26.7 min). Validation dataset: 28 AC-PC-oriented oblique slices, 4 mm thickness, 0.5 mm gap, TR 2000 ms, TE 30 ms, flip

angle 90°, FOV 20 × 20 cm^2, 64 × 64 matrix, in-plane resolution 3.44 × 3.44 mm^2) obtained as time series of 735 images (24.5 min).

Sleep stage rating: Generally, EEG data were screened by independent raters (not involved in the later fMRI analysis) for 5-minute-epochs that comprised at least 85% of one behavioural state (wakefulness, sleep stage 1, sleep stage 2 and slow wave sleep). Eventually, for the training dataset, 92 usable epochs were identified (27 wakefulness (S0), 23 sleep stage 1 (S1), 24 sleep stage 2 (S2), and 18 slow-wave sleep (SW)); for the validation dataset, 42 epochs were identified (17 S0, 13 S1, 12 S2). Supplementary Table S1 provides a detailed account of the amount and stage of epochs contributed by each subject in the training and the validation dataset.

FMRI preprocessing: FMRI data were processed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm), FSL v4.1 (http://www.fmrib.ox.ac.uk), and in-house software written in Matlab 2008b (http://www.mathworks.de/products/matlab) and IDL v6.4 (http://www.creaso.com). The time series comprised 144 images (150 images with first 6 images generally excluded due to potential incomplete T1-equilibrium in wakeful epochs collected from the very beginning of an fMRI recording) and underwent the following processing: (1) Correction of slice time differences, (2) realignment of time series to the corresponding first image using rigid body transformation, (3) spatial normalization to a standard EPI template in MNI space with interpolation to 3 × 3 × 3 mm^3 using standard linear and nonlinear deformations (all in SPM5) and (4) brain extraction to remove soft tissues (Brain Extraction Tool, FSL4.1). The time series was then residualized against the following regressors to remove signal fluctuations of non-neural origin: (1–4) average time courses of the white matter and CSF compartment and their 1^{st} order derivatives, (5-16) movement regressors (six parameters derived from the realignment procedure) and their 1^{st} order derivatives. Eventually, time series of 90 automated anatomical labeling (AAL) atlas regions (Tzourio-Mazoyer et al. 2002) were extracted (Supplementary Table S2). The cerebellum was excluded due to heterogeneous coverage across subjects in the training dataset. Next, the time series were band-pass filtered in the range of 0.01–0.1

Hz, using a 3rd order Buttherworth filter and the Pearson's product moment correlation coefficient (*r*) was computed between each pair of ROIs. Finally, the feature vector was derived from the upper triangle of the 90 × 90 correlation matrix; *r* values were converted to z-scores by Fisher's transform and stored in a vector of 4,005 elements per epoch.

2.2. Classifier and Classification Performance Measure

Linear support vector machine (SVM) classification with L2-regularization as implemented in libsvm (Chang and Lin 2011) was selected as the classifier for this study (libsvm was accessed via the R-package e1071). Prior to training the classifiers, each feature in the training dataset was scaled to mean=0 and variance=1. The same transformation was then applied to the unseen instances in the validation dataset. The cost parameter for punishing misclassifications on the training epochs (*C*) was set to 1.0 prior to the respective experiment. Thus, no parameter optimization was performed. Platt's algorithm (Platt 1999) was used to extract class probabilities from the SVM classifiers. All computations were carried out in R (Team 2013).

Performance of trained classifiers was measured using the area under the receiver operating characteristics (ROC) curve (referred to as AUC). Classifiers provide decision scores and, as a user, one is free to choose a threshold for the switch between the two classes to occur. The thresholds result in different sensitivity and specificity of the classifier for the whole dataset. ROC curves simply depict the sensitivity and specificity of a classifier at all possible thresholds. The area under this ROC curve (AUC) is a convenient summary measure for the entire curve. The AUC ranges between 0.0 and 1.0. A value of 1.0 indicates the existence of a threshold that leads to perfect classification (100% sensitivity at 100% specificity), whereas a value of 0.5 signifies random performance. The AUC can be interpreted as the probability that a randomly selected instance from the first class receives a higher score than a randomly selected instance from the second class (Fawcett 2006). In addition, the AUC is closely related to the Wilcoxon test of ranks (Hanley and McNeil 1982). Thus, a higher AUC signifies a better separation of the classes, i.e., a better classification performance. The advantage of the AUC

is that it is a threshold-free way to assess classifier performance. We used the R-package ROCR (Sing et al. 2005) for AUC computations. In the multi-class classification setting we computed the accuracy (ACC), i.e., the fraction of correctly classified samples, due to the lack of an adequate ROC-based measure for multi-class classification. Further, due to the sleep stage SW not being available in the validation dataset, performance for the four-class setting could not be computed on the validation dataset.

Classification performance was first assessed by using a leave-one-subject-out cross-validation (LOSO-CV) procedure on the training dataset. Second, classifier performance was probed by applying the classifiers trained on the training dataset to the epochs of the validation dataset. Briefly, in LOSO-CV, all epochs from all-but one subject are used to train the classifier, and the epochs belonging to the remaining left-out subject are used as independent test data for measuring classification performance. This is carried out for each subject individually, thereby mimicking the application of the classifier to novel data. While for most applications standard cross-validation (CV) is an appropriate way to assess classifier performance, it may overestimate the performance when multiple instances originate from the same individual, as present in this study.

2.3. Classification Experiments

We examined the impact of using subsets of the whole 5-minute epoch on classifier performance. For training the classifier we first used the maximal usable length of the fMRI data in each epoch, representing 288 seconds (144 volumes at TR 2000 ms), followed by successively shorter windows of 144, 96, 72, 48, 36, and 24 seconds. Here, we followed two approaches: (i) use only the first *n* second fragment of each epoch which keeps the amount of training data constant (termed *single-window-variant*) or (ii) use all non-overlapping *n* second fragments of the epoch as separate training instances, thus generating relatively more training instances for shorter fragments (termed *multi-windows-variant*). For probing the impact of the fragment length when applying the classifier to new

unseen epochs, we used the first window of *n* seconds length (using the same window sizes as above). Performance measures were computed for all pairs of fragment lengths, using both the single-window and the multi-windows variant.

We trained one SVM classifier for each pair of sleep stages, i.e., six binary classifiers to separate between wakefulness (S0), sleep stage 1 (S1), sleep stage 2 (S2), and slow-wave (SW) sleep, combining sleep stages 3 and 4. In addition, we trained one classifier separating between S0 and any other sleep stage (S1, S2, SW) by pooling the data on all sleep stages (termed SX). In order to derive a multi-class classifier from the six binary classifiers we applied a standard voting scheme (Hsu and Lin 2002): each test epoch was scored by all six binary classifiers and the epoch was labeled with the behavioral stage receiving the most votes. Possible ties were broken randomly. For further validation the analysis was repeated with the roles of training and validation data swapped.

2.4. Model Decoding

Making use of the advantages of linear SVMs to interpret classification models more directly and to facilitate decoding of the connectivity patterns discriminating between states, we analyzed the models using the increased and reduced connectivity index (ref. to as ICI and RCI, respectively) (Richiardi *et al.* 2012), as explained below.

SVM decoding: From the linear SVMs we extracted weights for each rs-fMRI connection. These weights relate to the direction and importance of the single connection for the classification. Briefly, during training the SVM classifier assigns a weight $a_i \geq 0$ to each of the *N* training samples $\overrightarrow{x_i}$; the class for a new sample $\vec{x}$ is computed as:

$$f(\vec{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} a_i y_i K(\overrightarrow{x_i}, \vec{x})\right),$$

where *sgn* is the sign function (i.e., returns -1 or 1), *K* is the kernel function, and y_i is the class label (either 1 or -1). Trainings samples with $a_i \neq 0$ are called support vectors. In cases where we use a linear kernel, i.e., $K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \overrightarrow{x_i} \cdot \overrightarrow{x_j}$, we can rearrange the formula into:

$$f(\vec{x}) = \mathrm{sgn}\left(\vec{x} \cdot \sum_{i=1}^{N} a_i y_i \overrightarrow{x_i}\right),$$

hence the connection weights are derived by a linear combination of the support vectors:

$$\vec{w} = \sum_{i=1}^{N} a_i y_i \overrightarrow{x_i}\,.$$

Finally, in order to render the SVM connection weights between classifiers more comparable for visualization purposes, we rescaled the SVM connection weights by a standard Z-transformation (subtracting the mean weight and dividing by the standard deviation of the weights).

We used the permutation importance algorithm with 10,000 permutations (Altmann *et al.* 2010) to identify connections that exhibit significantly large SVM weights. Briefly, permutation importance works by randomly permuting the class labels N times and every time training an SVM with those labels. Each SVM produces a set of SVM weights. The fraction of random SVM weights for a given connection that is equal or larger than the unpermuted SVM weight is used as a p-value. This p-value signifies whether the SVM weight is significantly larger than can be expected by chance or whether it is within the area of expected variation.

Increased and Reduced Connectivity Index: We computed ICI and RCI by grouping overall functional connectivity into 'increasing and 'decreasing' connections based on the sign of the connection weights provided by the SVM (ref. to as SVM weight). Briefly, in the SVM model each connection between two ROIs receives either a positive or a negative SVM weight that indicate for which class (and how strongly) the corresponding connection votes. The ICI for a single epoch is defined as the weighted sum of all connection scores (i.e., scaled Fisher's z-scores) with a positive SVM weight. Correspondingly, the RCI is the weighted sum of all connection scores with a negative SVM weight. In both cases the absolute value of the SVM weight is used to compute the weighted sum. Hence, each epoch can be represented as a pair of its ICI and RCI, and allows visualization of discriminative connectivity alterations between sleep stages as estimated by the SVM. We computed the ICI and

RCI for epochs in the training dataset in the same LOSO-CV procedure that was also employed for performance assessment. Further, ICI and RCI for epochs in the validation dataset were computed using SVM trained on the entire training dataset.

2.5. SVM Sleep Stage Tracking

In order to exemplify the classifier's capability to track changes in consciousness, we selected one full-length scan (800 images; 26.7 min) based on the observation that the subject crossed all non-REM sleep stages during one recording. The full-length scan was processed in the exact same way as described above. In particular, the first six image volumes at the beginning of the fMRI recording were discarded due to potential incomplete T1-equilirbium. Based on our results on suitable application window sizes (Figure 1) we further analyzed the scan using a window of 96 s (48 volumes) shifted by 4 s (two volumes). Prediction for the sleep stage was obtained using the multiclass classifier. In addition, we used the S0|SX classifier to obtain a probability score for sleep (SX). Since the subject originated from the training dataset, we used classifiers trained on data from all-but that subject. Inter-rater agreement between predicted sleep stages and EEG derived stages was measured using Cohen's weighted kappa coefficient (κ) with squared weights for disagreements.

3. Results

3.1. Classifier Performance

The average performance of the classifiers for each combination of window lengths for training and applying the classifier is depicted in Figure 1. Training the classifiers with connectivity scores obtained from fragments of 24 s in the single-window-variant resulted in a performance close to chance, regardless of the fragment length used for applying the classifier (Figure 1A). Once training fragments were sufficiently long (≥72 s), performance mostly depended on the fragment length for which the classifier will be applied, with a peak found at a window length of 144 s. In contrast, using the multi-windows-variant with all non-overlapping fragments from one epoch for training the classifiers, there was no such drop in performance for short fragment lengths (Figure 1B). In fact, performance mainly depended on the fragment length used for applying the classifier. The strongest improvements of the multi-windows-variant over the single-window-variant were seen for the shortest fragment length. Sizes of the training fragments of longer than 72 s led to no further improvement but rather unchanged performance levels (Figure 1C).

A more differentiated view on the classifier performance for each task is provided in Figure 2 that zooms into performance levels for maximum fragment length of 288 s for applying the classifier. Here, in addition to the LOSO-CV accuracy the performance on the validation dataset is also displayed. The LOSO-CV performance for most of the tasks was almost perfect with AUCs close to 1.0, except for the discrimination between S1|S2 and S2|SW with AUCs of ~0.9 and ~0.8, respectively. As a general observation, performance levels were lower for separating adjacent sleep stages than to separate stages with at least one interposed stage. Moreover, the multi-windows-variant resulted in higher performance values and more stable classifiers as indicated by the slope of the curve. This benefit was most pronounced for those classifications tasks that showed poor performances at short training window length, such as S1|S2, S2|SW, and S1|SW. The task of separating wakefulness from any sleep stage (SX) achieved close to perfect performance in the LOSO-CV setting and on the validation dataset. Table 1 provides detailed information on the LOSO-CV

results of the binary classifiers for the training dataset using the longest window of 288 s. Based on a permutation test with 1,000 random relabelings of the training data, all classification results (AUC as well as accuracy) were significant.

In the multiclass scenario, accuracy was 78% for a three-class classifier (S0, S1, and S2) on both training (via LOSO-CV) and validation dataset (Figure 2). For comparison, random performance was estimated to be about 35%. In the four-class setting (S0, S1, S2, and SW) the accuracy dropped slightly to 70% on the training data. Notably, the performance of a random classifier was estimated with 22% in the four-class case. In both cases, we observed a notable benefit of using the multi-windows-variant over the single-window-variant for small window lengths up to 48 s.

In general, the performance on the validation dataset was almost indistinguishable from the performance estimated using LOSO-CV on the training dataset. In addition, we swapped the roles of the training dataset and the validation dataset and re-computed the classifier performance (Figure S1). Again, performance for most tasks was very good (AUCs above 0.9) with the exception being S1|S2 (peak LOSO-CV AUC of 0.8).

3.2. Model Decoding

SVM decoding: We computed the discriminative weights for each rs-fMRI connection for all six binary classifiers and the S0|SX classifier from SVMs trained using the multi-windows-variant (48-s-windows; Figure 3). The S0|S1 classifier has large weights for subcortical-cortical connections, in particular thalamocortical connections; thus, pointing toward the relative importance of these connections for the classification. To confirm this, we trained a classifier using only subcortical-cortical connections on the training dataset and it achieved an AUC of 0.95 on the validation dataset. Figure 4 provides a more detailed visualization of the discriminative connectivity pattern in anatomical space based on the 41 connections that achieve a $P \leq 0.01$ in the permutation importance test (Figure S2 shows all 200 connections with $P \leq 0.05$; Figures S3, S4 and S5 depict the connections that achieve a $P \leq 0.01$ in the permutation importance test for S0|S2, S0|SW and S1|SW,

respectively). Here it becomes further evident that frontotemporal and frontoparietal connections are stronger in S1 than in S0. Comparison among the discriminative patterns in the other classifiers revealed that both for S0|S2 and for S0|SW the subcortical-cortical connections including the thalamocortical connections lost their relative importance for the classification. Between S0 and S2 as well as between S0 and SW, connectivity of the middle and inferior occipital gyri to a group of areas comprising the calcarine gyrus, lingual gyrus and the cuneus, decreased (hot colored areas in OC-OC-fields in S0|S2 and S0|SW part of Figure 3). By contrast, increases of connectivity between S0 and S2 as well as between S0 and SW were observed between the calcarine gyrus and the lingual gyrus as well as the cuneus (cool colored areas in the respective OC-OC-fields). Further, specific parietal and limbic (mainly cingulate cortex) connections increased as well. The occipital lobe reorganization was also observed in the SVM connection weights for S1|SW with additional strong connections emerging from the calcarine gyrus to the contralateral calcarine gyrus and the bilateral cuneus and lingual gyrus. There was no strong focal discrimination pattern for S1|S2 and S2|SW. Finally, S0|SX combined the previous patterns between S0 and individual sleep stages, mainly strong thalamocortical connections and reorganization of connections within the occipital lobe.

Combined ICI/RCI plots for model decoding: A 2D-visualization of the SVM decision boundary as represented by ICI and RCI is shown in Figure 5 for all six binary classification tasks for the training dataset using LOSO-CV. Similar to the ROC-based performance measures, plots indicate that fewer misclassifications occurred for more distant sleep states in terms of a larger average distance of the individual epochs from the decision boundary. An exception is the classification between S2 and SW. Here, many instances were close to the decision boundary and many misclassifications occurred. A closer look at the distribution of ICI and RCI for each class and classifier helps interpreting these results (Figure S6 in Supplementary Material): While between S0 and either S1 or S2 the contribution of positively weighted discriminative edges (ICI) stayed almost constant, the average RCI showed a pronounced drop for the transitions from S0 to either S1 or S2. That is, a large portion of brain

connectivity increased when transitioning from wakefulness (S0) to light sleep (S1 or S2). When comparing S0 and SW there were changes of both, ICI and RCI, with the changes in positively weighted discriminative edges (ICI) being more pronounced. The same holds true for S1 and SW. Here the average of the negatively weighted discriminative edges (RCI) was found to be almost indistinguishable between sleep stages, whereas the main changes in brain connectivity occurred in ICI. That is, there was a decrease in large-scale brain connectivity when comparing S0 or S1 to SW. These general patterns were also observed on the single-subject level (Figure 5): As an example, subject 22 showed stronger RCI values, for S0 and S1 – which was even more pronounced for the comparison S0 and S2. When comparing S0 to SW (or S1 to SW), changes for subject 22 were mainly seen for ICI (reduction). Comparable ICI/RCI patterns were observed for the independent test instances (Figure S7 in Supplementary Material).

3.4. SVM Sleep Stage Tracking

To exemplify the feasibility of our method and a potential application field, we recorded and tracked sleep stages for a single subject over 27 min recordings across wakefulness and NREM sleep stages (Figure S8). For the S0|SX classifier we employed a sleep probability cutoff of 0.75 to decide between sleep and wakefulness. This cutoff was selected as it yielded both good sensitivity (0.86) and good specificity (0.89) in the LOSO-CV on the entire training data. Overall, as measured by Cohen's kappa for inter-rater agreement, the fMRI-based sleep staging captured the changes from wakefulness to the different sleep stages very well: κ=0.75. The fMRI classifier missed short changes in the EEG-based rating, but captured the overall trend. A short window at the beginning of the scan was two stages apart while the remainder or the scan was well captured. Periods of S0 were well recognized and, in general, phases of sleep were rarely missed by the probability SX-based classifier.

4. Discussion

Changes in consciousness are reflected in altered functional connectivity patterns of the brain as demonstrated by a growing literature (for reviews see: Boly et al. (2008); Larson-Prior *et al.* (2011); Picchioni *et al.* (2013)). Here we employed rs-fMRI based whole-brain connectivity data to train linear SVM classifiers to separate resting wakefulness from NREM sleep stages. Classification performance was estimated in two fMRI datasets with EEG-validated sleep ratings using the LOSO-CV scheme.

In brief, classification performance was close to perfect for separating wakefulness (S0) from any sleep stage (SX) (AUC values ~0.95). Performance based on the LOSO-CV scheme in the original dataset differed only marginally from performance on the validation dataset, which points out good generalizability of the linear SVM classifier to novel instances. As second probe of the generalization capability the roles of the training and validation dataset were swapped. This procedure had practically no impact on the classification performance, further corroborating the good overall classification performance and generalizability for the S0|SX classification and most of the binary classifications. For specific binary classifications, such as S1|S2, classification accuracy was lower (AUC values 0.8-0.9), and subsequently, multiclass classification performance was around 75% which, however, is still considered good in terms of multiclass problems. As expected from the rs-fMRI literature on sleep, a better discrimination of more 'remote' sleep stages reflects the gradual connectivity reorganizations that accompany the transitions from wakefulness to deep sleep, although some processes such as thalamocortical connectivity might show a non-linear (e.g., U-shaped) relation with sleep stages: here, transient loss of thalamocortical connectivity has been observed in light sleep with (partial) recurrence in SW (Spoormaker *et al.* 2010; Picchioni *et al.* 2014). A main reason for lower S1|S2 discrimination may be the actual presence of only weak to moderate connectivity changes between these two sleep stages, as reported before on the first dataset using different analysis approaches (Spoormaker *et al.* 2010; Sämann *et al.* 2011) and as similarly observed in the recent study of Tagliazucchi and Laufs (2014). In addition, classical scoring rules assessing epochs of 20 to 30 s do not capture faster changes of neural activity. For example, transient EEG

waveforms, such as K-complexes and sleep spindles, define occurrence of sleep stage 2, with the Rechtschaffen and Kales rules stating: "(…) relatively long periods may intervene between these events without the occurrence of a stage change. If less than 3 min of the recording which would ordinarily meet the requirements for Stage 1 intervene between sleep spindles and/or K complexes, these intervening epochs are to be scored Stage 2 (…)" (Rechtschaffen and Kales 1968). Some connectivity changes, e.g., those accompanying sleep spindles (Andrade et al. 2011), may therefore not be sustained enough to alter the average connectivity of a whole epoch of stage 2. Further, this rule may foster misclassifications as the underlying brain state may in fact more resemble 'intermittent stage 1' (Himanen and Hasan 2000). Though this limitation is intrinsic to the accepted gold standard chosen for this study, it is noteworthy that our growing-window-approach demonstrated a minimum of about 1 minute of fMRI data needed until most classifiers stabilized (Figure 1).

Our data confirm the very good reported classification accuracy for predicting sleep stages by an SVM approach (Tagliazucchi *et al.* 2012; Tagliazucchi and Laufs 2014). One apparent performance difference is a medium S2|SW classification accuracy in our work (AUC ~0.8) compared to an accuracy of 0.95 for classifying between N2 and N3 in Tagliazucchi *et al.* (2012). Several factors may play a role for this: First, Tagliazucchi *et al.* (2012) used mainly (20 of 22) functionally defined ROIs based on temporal ICA whereas we used a predefined anatomical parcellation (Tzourio-Mazoyer *et al.* 2002). Using functionally defined ROIs could indeed protect from non-discriminating features within an atlas ROI that reduce its discrimination power (Shirer *et al.* 2012). Later, Tagliazucchi and Laufs (2014) used a whole brain parcellation with feature selection based on univariate comparisons, yet, pairwise discrimination accuracies between sleep stages were not reported in detail. Further, as cerebellar connections seem to play a role in the descent to sleep (Tagliazucchi and Laufs 2014), the exclusion of these due to incomplete coverage in some subjects in our sample may have impacted classifier performance to some degree. Second, the applied CV scheme differed in that we used the

more realistic LOSO-CV while Tagliazucchi *et al.* (2012) relied on a standard 5-fold CV scheme, which may result in instances from the same individual being used for building the classifier and for assessing its performance. As this can lead to artificially increased performance values, we consider the approach presented here more conservative. Third, the exact origin of the validation dataset: Tagliazucchi *et al.* (2012) set aside the last 18 scans of a recording series for being used as validation dataset. In contrary, the validation scheme described here likely involves less dependencies as at it was acquired 2 years after the first dataset by different operators, after a substantial hardware upgrade of the scanner (GE Signa to GE Signa Excite), and in the context of a different experimental design with different instructions (see methods). The use of a different kernel in the SVM is unlikely to be of large relevance, as the feature space is already large and should allow for linear decision boundary (LaConte et al. 2005). In fact, Tagliazucchi *et al.* (2012) report little differences between a polynomial kernel function and a radial basis function kernel. Further, the use of linear classifiers is preferable as it facilitates a direct decoding of the model, which we did in extracting discriminative weights for each connection and demonstrating respective ICI and RCI differences across sleep stages.

We evaluated the effects of the (functional connectivity) window length used for training the classifier and its application. For this, we screened all possible combinations of window lengths during training and application; typically window lengths for training and application are set to be identical (representing the diagonal of the matrix shown in Figure 1) (Shirer *et al.* 2012; Tagliazucchi *et al.* 2012). Generally, the lack of a steep drop in performance when moving to rather short (58–72 s) time windows was noted. In addition, the asymmetry of the result matrix (B) compared with (A) points out that the use of multiple non-overlapping shorter windows seems superior compared to the use of a single larger time window of the same total length for classifier training. As the EEG-validated sleep stage was assumed to be the same for all non-overlapping time windows of that epoch, this pattern could indirectly point towards outlier features of some parts of the epoch that reduce the classification performance of the whole epoch. Simply put, the benefit of the multi-

window-variant over the single-window-variant of the same length is clearly due to the increased sample size, whereas the benefit of the multiple short windows over the single full-length window is more complex and may be due to sampling of connectivity changes that are non-stationary throughout the 5 min epoch.

After having established a good classification performance of the classifier, we extracted discriminative weights for each connection in each classification condition. This method allowed us to reconstruct the discriminatory patterns in anatomical space which facilitates both a comparison of our results with other rs-fMRI connectivity studies on sleep and a comparison across different types of analyses such as ICA, cross-correlation analysis and graph theory analysis (Spoormaker *et al.* 2010; Andrade *et al.* 2011; Sämann *et al.* 2011). In the present hypothesis-free approach, several discriminatory patterns were identified: First, a decrease of thalamocortical and generally subcortical-cortical (all lobes) connectivity emerged as strong carrier of the discrimination between S0 and S1, with thalamocortical connections regaining strength in deeper sleep stages, as observed previously (Spoormaker *et al.* 2010; Picchioni *et al.* 2013). Concurrently, widespread weak to moderate connectivity increases were observed some of which were more transient (such as frontoparietal connections) and some were lasting into deep sleep (such as frontolombic and parietolimbic connections; compare S0|S1, S0|S2 and S0|SW). Generally, these observations are in line with previous reports (Larson-Prior et al. 2009; Spoormaker *et al.* 2010; Spoormaker et al. 2012; Tagliazucchi and Laufs 2014). A major impact for the discrimination between wakefulness and sleep was found for the reorganization of connectivity within the occipital lobe that comprised parallel decreases and increases. Decreasing connectivity of the calcarine cortex to contralateral occipital regions in sleep (note blueish areas within OC-OC-fields of the S0|SW comparison, Figure 3) was similarly reported in a seed-based connectivity analysis of our group (Spoormaker *et al.* 2012) and parallels reports of increasing BOLD fluctuations in visual and somato-motor regions during light sleep (Fukunaga *et al.* 2006; Horovitz *et al.* 2009). These latter reports focused on voxelwise

fluctuation amplitudes or variance, yet, a full integration of these measures with connectivity changes has not been achieved so far.

Despite the reorganization within the occipital lobe, decoupling of the occipital lobe from the subcortical system as a whole was observed. Both observations are likely unrelated to the eyes-open vs. -closed status as subjects were resting with their eyes closed from the start. Notably, signature and distribution of functional connectivity changes of the visual cortex may also depend on the pre-processing scheme (see supplemental of Tagliazucchi and Laufs (2014)). Either way, specific occipito-temporal and temporoparietal connections showed gradual decline of connectivity from S0 to deep sleep, matching previous observations (Horovitz *et al.* 2009; Spoormaker *et al.* 2010; Sämann *et al.* 2011). Frontal connectivity changes appeared as complex and heterogeneous, with sleep-associated connectivity decreases to subcortical areas and (selected) occipital, limbic and parietal areas (e.g., angular gyrus).

One potential application of fMRI-based sleep staging is to screen resting state fMRI datasets to exclude epochs during which subjects were not awake. Only recently, SVM based classifiers from combined EEG-fMRI data of 30 subjects were applied to datasets of the 1,000 Functional Connectomes Project (Biswal *et al.* 2010; Tagliazucchi and Laufs 2014): In brief, results point to sleep occurring in a high proportion of cases which corroborates a long suspected confound for clinical (or genomic) rs-fMRI studies (Sämann *et al.* 2010; Sämann *et al.* 2011). Such an application needs particularly good discrimination between wakefulness and light sleep: here, we report AUC-values of ~0.9 for S0|S1 discrimination with sensitivity and specificity values being both above 0.85. Overall, this performance seems sufficient to screen larger datasets without losing too many data points due to false classification as S1. While an influence of the instruction (eyes closed, eyes open [with or without fixation]) on the resulting behavioral state is likely (Tagliazucchi and Laufs 2014), it should be considered that connectivity patterns *during* wakefulness already differ between the various eyes-states (Patriat *et al.* 2013). Therefore, extrapolation from classifiers built on eyes-closed datasets to datasets with a different obtained with eyes-open instructions might lead to uncontrolled false

classifications. Here we demonstrated by predicting the instances in the validation dataset that our classifier generalizes well across hardware changes and different operators given the eyes-closed instruction as well as identical postprocessing. Ideally, classifiers have to be rigorously validated (i.e., with available ground truth) on data from different instructions and hardware prior to their application to a wide variety of acquisition settings as can be found in the 1,000 Functional Connectomes Project.

Given the relatively short required window length for applying the classifier (about 1 minute), another application is the construction of hypnograms from fMRI data (see Figure S8 for one long epochs of 27 minutes). In this exploratory analysis our rs-fMRI-based sleep showed a good agreement with the EEG-based gold standard (κ=0.75). Thus, such hypnograms may provide a more objective estimate of subjects' vigilance states compared to self-reported vigilance. By employing such an intermediate step, epochs of a specific vigilance state of interest can be extracted and merged from longer recordings to ensure a homogeneous vigilance background.

Several limitations of this study need to be discussed: first, we built our classifiers on combinations of fMRI data with 5 minute-epochs of one prevailing sleep stage. However, using continuous and uninterrupted EEG epochs is not a prerequisite for training the classifier. Theoretically, the data basis could have been larger if discontinuous but shorter time windows (e.g., 1-minute EEG-ratings) would have been employed. Second, our validation dataset did not cover slow wave sleep (SW), which rendered the independent validation of three of the six binary classifiers impossible. Third, the optimal handling of global signal regression and other nuisance variables is still under discussion; recent reports even suggest that individual decisions for global regression per epoch/recording may be preferable (Chen et al. 2012). Here, we refrained from including the global signal as regressor due to its close correlation with thalamic activity (Zhang et al. 2008) that may alter thalamocortical connectivity patterns. Indeed, when applied to our data, the regression of global signal makes the differences in thalamocortical connectivity between S0 and S1 disappear (Figure S9 in the Supplementary Material). Further, when conducting the LOSO-CV for S0|S1 the AUC of 0.93 (as

gained by our original processing) drops to 0.84 after global signal regression. This indicates, that the well-known thalamocortical disconnection in the transition from S0 to S1 is an important carrier of the classification performance, although the remaining information after global signal regression is still useful. In conclusion, we would therefore not recommend a global regression step for rs-fMRI-connectivity based sleep stage classification. Fourth, the behavioral instructions differed between the training dataset, in which subjects were encouraged to fall asleep, and the validation dataset, in which subjects were asked to try to stay awake. In the latter case, cognitive mechanisms may have been activated that prevent falling asleep and that overlap with sleep promoting changes. Overall, however, the results of the validation across samples suggest that the influence of specific brain activity during wakefulness on the sleep stage classification is only minor. Fifth, our study protocol did not aim at collecting REM sleep data, which limits a full description of sleep stage specific connectivity. Still, for applications that aim at detecting intrusions of sleep into recordings of presumed wakefulness, the lack of REM sleep in our data is less critical, as direct transitions from wakefulness to REM are considered rare or even pathological events. Last, we used a broad band-pass filter (0.01–0.1 Hz), accepting that classification results may be different for smaller bandwidths, e.g., as observed for graph-theory analysis (Achard et al. 2006; Spoormaker *et al.* 2010).

Clearly, the classification approach presented here is only validated for rs-fMRI data from a single research center. Further studies are required to elucidate which parameters are essential to a successful transfer of such a classifier to other sites with different hardware and acquisition parameters. Required changes may be trivial, such as adjusting the threshold between two sleep stages (here a default of 0.5 was used), or essential, such as retraining the classifier under consideration of the data distribution at the new site (e.g., transductive learning or learning under covariate shift) or even requiring additional training data from the new site. Although done so previously (Tagliazucchi and Laufs 2014), in our opinion, the application of a classifier trained in one site to other sites with different hardware (field strength, coil) and sequence parameters still contains an unknown risk of systematic classification shift or more noisy predictions. It should be

emphasized, however, that in our work a successful transfer of the classifier to a second dataset (and vice versa) was observed that differed in (i) details of the EPI-sequence parameters, (ii) instructions to the subjects before the rs-fMRI acquisition and (iii) differences in the team supervising the EEG-fMRI acquisition. Further encouraging is the successful use of the classifier to detect propofol-induced loss of consciousness from whole brain functional connectivity (AUC 0.82-1.0; Figure S10 in the Supplementary Material). Since this dataset was acquired on a different hardware platform (Schröter et al. 2012), this is a first strong hint towards generalizability of the classifier to 3 Tesla platforms. Still, further validation on sleep EEG-fMRI recordings acquired on different platforms and field strength are needed to warrant broad generalizability. Based on the current data we were able to reliably detect periods of sleep as short as 72 s (Figure 1). The lower limit of the length of the temporal window is determined by the quality of the estimated functional connections. Typically, low frequency BOLD signal fluctuations < 0.1 Hz are considered for functional connectivity analyses. These are well sampled using standard MRI repetition times (in the order of seconds). However, when reducing the length of the sampling window, according to the convolution theorem, the frequency spectrum will be convolved with a sinc function, which results in a blurred representation of individual fluctuations and a potentially distorted correlation analysis. In our data, the lower limit for reliable sleep stage detection was about 1 minute. Shorter fMRI time windows worsened the performance. This is in line with observations showing that correlation strength within and between networks of resting state fMRI data stabilize for data acquisition windows longer than about 1-2 minutes (Van Dijk et al. 2010; Whitlow et al. 2011). However, new methods such as multiband fMRI (Feinberg and Setsompop 2013) that enable effective TRs of less than 1 s will bring the required acquisition time closer to 30 s or even 20 s, that is into the range of EEG based sleep scoring.

In conclusion, we report that an SVM-based multivariate classification algorithm trained by EEG-validated rs-fMRI acquisitions of defined behavioral states allows for robust automated discrimination of wakefulness from sleep and decoding of separate NREM-sleep stages. We extend previous observations in that our hypothesis-free, atlas-based parcellation system revealed major discrimi-

nating patterns such as subcortical-cortical connectivity including thalamocortical connectivity as well as strong and spatially concise connectivity re-configuration within the occipital lobe that converge with a growing literature on sleep rs-fMRI analysis.

Acknowledgements

We thank Jonas Richiardi for helpful discussions and for help with using the Flexible Brain Graph Visualizer for visualizing connectivity patterns and Renate Wehrle for evaluation of the EEG recordings. Ines Eidner and Rosa Schirmer are acknowledged for their assistance in data acquisition.

References

AASM. 2007. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Chicago: American Academy of Sleep Medicine.

Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. 2006. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. The Journal of neuroscience : the official journal of the Society for Neuroscience 26:63-72.

Altmann A, Tolosi L, Sander O, Lengauer T. 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26:1340-1347.

Andrade KC, Spoormaker VI, Dresler M, Wehrle R, Holsboer F, Samann PG, Czisch M. 2011. Sleep spindles and hippocampal functional connectivity in human NREM sleep. The Journal of neuroscience : the official journal of the Society for Neuroscience 31:10331-10339.

Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski AM, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kotter R, Li SJ, Lin CP, Lowe MJ, Mackay C, Madden DJ, Madsen KH, Margulies DS, Mayberg HS, McMahon K, Monk CS, Mostofsky SH, Nagel BJ, Pekar JJ, Peltier SJ, Petersen SE, Riedl V, Rombouts SA, Rypma B, Schlaggar BL, Schmidt S, Seidler RD, Siegle GJ, Sorg C, Teng GJ, Veijola J, Villringer A, Walter M, Wang L, Weng XC, Whitfield-Gabrieli S, Williamson P, Windischberger C, Zang YF, Zhang HY, Castellanos FX, Milham MP. 2010. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences of the United States of America 107:4734-4739.

Boly M, Phillips C, Tshibanda L, Vanhaudenhuyse A, Schabus M, Dang-Vu TT, Moonen G, Hustinx R, Maquet P, Laureys S. 2008. Intrinsic brain activity in altered states of consciousness: how conscious is the default mode of brain function? Annals of the New York Academy of Sciences 1129:119-129.

Braun AR, Balkin TJ, Wesenten NJ, Carson RE, Varga M, Baldwin P, Selbie S, Belenky G, Herscovitch P. 1997. Regional cerebral blood flow throughout the sleep-wake cycle. An H2(15)O PET study. Brain : a journal of neurology 120 (Pt 7):1173-1197.

Cecchi G, Rish I, Thyreau B, Thirion B, Plaze M, Paillere-Martinot M-L, Martelli C, Martinot J-L, Poline J-B. 2009. Discriminative Network Models of Schizophrenia. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, editors. Advances in Neural Information Processing Systems 22 p 252-260.

Chang C, Glover GH. 2010. Time-frequency dynamics of resting-state brain connectivity measured with fMRI. NeuroImage 50:81-98.

Chang C, Lin C. 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:1-27.

Chen G, Chen G, Xie C, Ward BD, Li W, Antuono P, Li SJ. 2012. A method to determine the necessity for global signal regression in resting-state fMRI studies. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 68:1828-1835.

Christoff K. 2012. Undirected thought: neural determinants and correlates. Brain research 1428:51-59.

Christoff K, Gordon AM, Smallwood J, Smith R, Schooler JW. 2009. Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. Proceedings of the National Academy of Sciences of the United States of America 106:8719-8724.

Cole DM, Smith SM, Beckmann CF. 2010. Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. Frontiers in systems neuroscience 4:8.

De Havas JA, Parimal S, Soon CS, Chee MW. 2012. Sleep deprivation reduces default mode network connectivity and anti-correlation during rest and task performance. NeuroImage 59:1745-1751.

Dosenbach NU, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, Nelson SM, Wig GS, Vogel AC, Lessov-Schlaggar CN, Barnes KA, Dubis JW, Feczko E, Coalson RS, Pruett JR, Jr., Barch DM, Petersen SE, Schlaggar BL. 2010. Prediction of individual brain maturity using fMRI. Science 329:1358-1361.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recogn Lett 27:861-874.

Feinberg DA, Setsompop K. 2013. Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. Journal of magnetic resonance 229:90-100.

Fornito A, Harrison BJ, Goodby E, Dean A, Ooi C, Nathan PJ, Lennox BR, Jones PB, Suckling J, Bullmore ET. 2013. Functional Dysconnectivity of Corticostriatal Circuitry as a Risk Phenotype for Psychosis. JAMA psychiatry.

Fox MD, Raichle ME. 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature reviews Neuroscience 8:700-711.

Fukunaga M, Horovitz SG, van Gelderen P, de Zwart JA, Jansma JM, Ikonomidou VN, Chu R, Deckers RH, Leopold DA, Duyn JH. 2006. Large-amplitude, spatially correlated fluctuations in BOLD fMRI signals during extended rest and early sleep stages. Magnetic resonance imaging 24:979-992.

Greicius MD, Kiviniemi V, Tervonen O, Vainionpaa V, Alahuhta S, Reiss AL, Menon V. 2008. Persistent default-mode network connectivity during light sedation. Human brain mapping 29:839-847.

Guyon A, Cathala L, Paupardin-Tritsch D, Eugene D. 2002. Furosemide modulation of GABA(A) receptors in dopaminergic neurones of the rat substantia nigra. Neuropharmacology 43:750-763.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.

Himanen SL, Hasan J. 2000. Limitations of Rechtschaffen and Kales. Sleep medicine reviews 4:149-167.

Horovitz SG, Braun AR, Carr WS, Picchioni D, Balkin TJ, Fukunaga M, Duyn JH. 2009. Decoupling of the brain's default mode network during deep sleep. Proceedings of the National Academy of Sciences of the United States of America 106:11376-11381.

Hsu CW, Lin CJ. 2002. A comparison of methods for multiclass support vector machines. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council 13:415-425.

Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, Della Penna S, Duyn JH, Glover GH, Gonzalez-Castillo J, Handwerker DA, Keilholz S, Kiviniemi V, Leopold DA, de Pasquale F, Sporns O, Walter M, Chang C. 2013. Dynamic functional connectivity: promise, issues, and interpretations. NeuroImage 80:360-378.

Kiem SA, Andrade KC, Spoormaker VI, Holsboer F, Czisch M, Samann PG. 2013. Resting state functional MRI connectivity predicts hypothalamus-pituitary-axis status in healthy males. Psychoneuroendocrinology 38:1338-1348.

LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26:317-329.

Larson-Prior LJ, Power JD, Vincent JL, Nolan TS, Coalson RS, Zempel J, Snyder AZ, Schlaggar BL, Raichle ME, Petersen SE. 2011. Modulation of the brain's functional network architecture in the transition from wake to sleep. Progress in brain research 193:277-294.

Larson-Prior LJ, Zempel JM, Nolan TS, Prior FW, Snyder AZ, Raichle ME. 2009. Cortical network functional connectivity in the descent to sleep. Proceedings of the National Academy of Sciences of the United States of America 106:4489-4494.

Maquet P. 2000. Functional neuroimaging of normal human sleep by positron emission tomography. Journal of sleep research 9:207-231.

Meyer-Lindenberg A, Weinberger DR. 2006. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nature reviews Neuroscience 7:818-827.

Patriat R, Molloy EK, Meier TB, Kirk GR, Nair VA, Meyerand ME, Prabhakaran V, Birn RM. 2013. The effect of resting condition on resting-state fMRI reliability and consistency: a comparison between resting with eyes open, closed, and fixated. NeuroImage 78:463-473.

Pereira F, Mitchell T, Botvinick M. 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45:S199-209.

Picchioni D, Duyn JH, Horovitz SG. 2013. Sleep and the functional connectome. NeuroImage 80:387-396.

Picchioni D, Pixa ML, Fukunaga M, Carr WS, Horovitz SG, Braun AR, Duyn JH. 2014. Decreased connectivity between the thalamus and the neocortex during human nonrapid eye movement sleep. Sleep 37:387-397.

Platt J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10:61-74.

Rechtschaffen A, Kales A. 1968. A Manual of Standardized Terminology, Techniques, and Scoring Systems for Sleep Stages of Human Subjects: Brain Information/Brain Research Institute.

Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. 2011. Decoding brain states from fMRI connectivity graphs. NeuroImage 56:616-626.

Richiardi J, Gschwind M, Simioni S, Annoni JM, Greco B, Hagmann P, Schluep M, Vuilleumier P, Van De Ville D. 2012. Classifying minimally disabled multiple sclerosis patients from resting state functional connectivity. NeuroImage 62:2021-2033.

Sämann PG, Tully C, Spoormaker VI, Wetter TC, Holsboer F, Wehrle R, Czisch M. 2010. Increased sleep pressure reduces resting state functional connectivity. Magma 23:375-389.

Sämann PG, Wehrle R, Hoehn D, Spoormaker VI, Peters H, Tully C, Holsboer F, Czisch M. 2011. Development of the brain's default mode network from wakefulness to slow wave sleep. Cerebral cortex 21:2082-2093.

Schröter MS, Spoormaker VI, Schorer A, Wohlschlager A, Czisch M, Kochs EF, Zimmer C, Hemmer B, Schneider G, Jordan D, Ilg R. 2012. Spatiotemporal reconfiguration of large-scale brain functional networks during propofol-induced loss of consciousness. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:12832-12840.

Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD. 2009. Neurodegenerative diseases target large-scale human brain networks. Neuron 62:42-52.

Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD. 2012. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cerebral cortex 22:158-165.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21:3940-3941.

Spoormaker VI, Gleiser PM, Czisch M. 2012. Frontoparietal Connectivity and Hierarchical Structure of the Brain's Functional Network during Sleep. Frontiers in neurology 3:80.

Spoormaker VI, Schroter MS, Gleiser PM, Andrade KC, Dresler M, Wehrle R, Samann PG, Czisch M. 2010. Development of a large-scale functional brain network during human non-rapid eye movement sleep. The Journal of neuroscience : the official journal of the Society for Neuroscience 30:11379-11387.

Supekar K, Menon V, Rubin D, Musen M, Greicius MD. 2008. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. PLoS computational biology 4:e1000100.

Tagliazucchi E, Laufs H. 2014. Decoding Wakefulness Levels from Typical fMRI Resting-State Data Reveals Reliable Drifts between Wakefulness and Sleep. Neuron 82:695-708.

Tagliazucchi E, von Wegner F, Morzelewski A, Borisov S, Jahnke K, Laufs H. 2012. Automatic sleep staging using fMRI functional connectivity data. NeuroImage 63:63-72.

Tagliazucchi E, von Wegner F, Morzelewski A, Brodbeck V, Jahnke K, Laufs H. 2013. Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. Proceedings of the National Academy of Sciences of the United States of America 110:15419-15424.

Team RC. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Tolosi L, Lengauer T. 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27:1986-1994.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15:273-289.

van den Heuvel MP, Hulshoff Pol HE. 2010. Exploring the brain network: a review on resting-state fMRI functional connectivity. European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology 20:519-534.

Van Dijk KR, Hedden T, Venkataraman A, Evans KC, Lazar SW, Buckner RL. 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. Journal of neurophysiology 103:297-321.

Van Dijk KR, Sabuncu MR, Buckner RL. 2012. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage 59:431-438.

Vanable PA, Aikens JE, Tadimeti L, Caruana-Montaldo B, Mendelson WB. 2000. Sleep latency and duration estimates among sleep disorder patients: variability as a function of sleep disorder diagnosis, sleep history, and psychological characteristics. Sleep 23:71-79.

Wehrle R, Kaufmann C, Wetter TC, Holsboer F, Auer DP, Pollmacher T, Czisch M. 2007. Functional microstates within human REM sleep: first evidence from fMRI of a thalamocortical network specific for phasic REM periods. The European journal of neuroscience 25:863-871.

Whitlow CT, Casanova R, Maldjian JA. 2011. Effect of resting-state functional MR imaging duration on stability of graph theory metrics of brain network connectivity. Radiology 259:516-524.

Zhang D, Snyder AZ, Fox MD, Sansbury MW, Shimony JS, Raichle ME. 2008. Intrinsic functional relations between human cerebral cortex and thalamus. Journal of neurophysiology 100:1740-1748.

Figure Captions

Figure 1: **Performance overview.** Each entry in the matrix corresponds to the average AUC across the six binary classification tasks (S0|S1, S0|S2, S0|SW, S1|S2, S1|SW, and S2|SW) for pairs of window lengths of the training window (x-axis) and the application window on which the trained classifiers were used in a LOSO-CV manner (y-axis). (A) Performance for use of the single-window-variant. Note decreases in performance for smallest training window size. (B) Performance for use of the multi-windows-variant. (C) Difference of AUC values between approaches shown in (A) and (B). Note strong performance improvement in (B) compared with (A) for the smallest training window size.

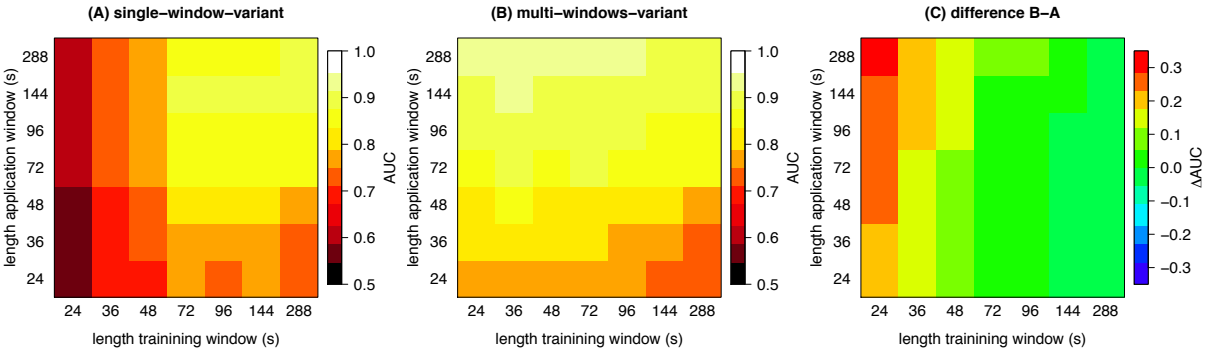


Figure 2: **Classifier performance for individual tasks.** Each panel depicts the performance of one classifier trained on the training dataset using either LOSO-CV (blue) or the validation dataset (orange). X-axis denotes the training window length using either the single-window-variant (down facing triangles) or the multi-windows-variant (up facing triangles). All results refer to a test window length of 288 s (i.e., all 144 image volumes) and are provided as AUC. Dashed horizontal lines mark random performance. The eighth and ninth panel shows the performance of a three-class classifier (S0, S1, S2) and four-class-classifier (S0, S1, S2, SW), respectively. Here results are given in accuracy (ACC).

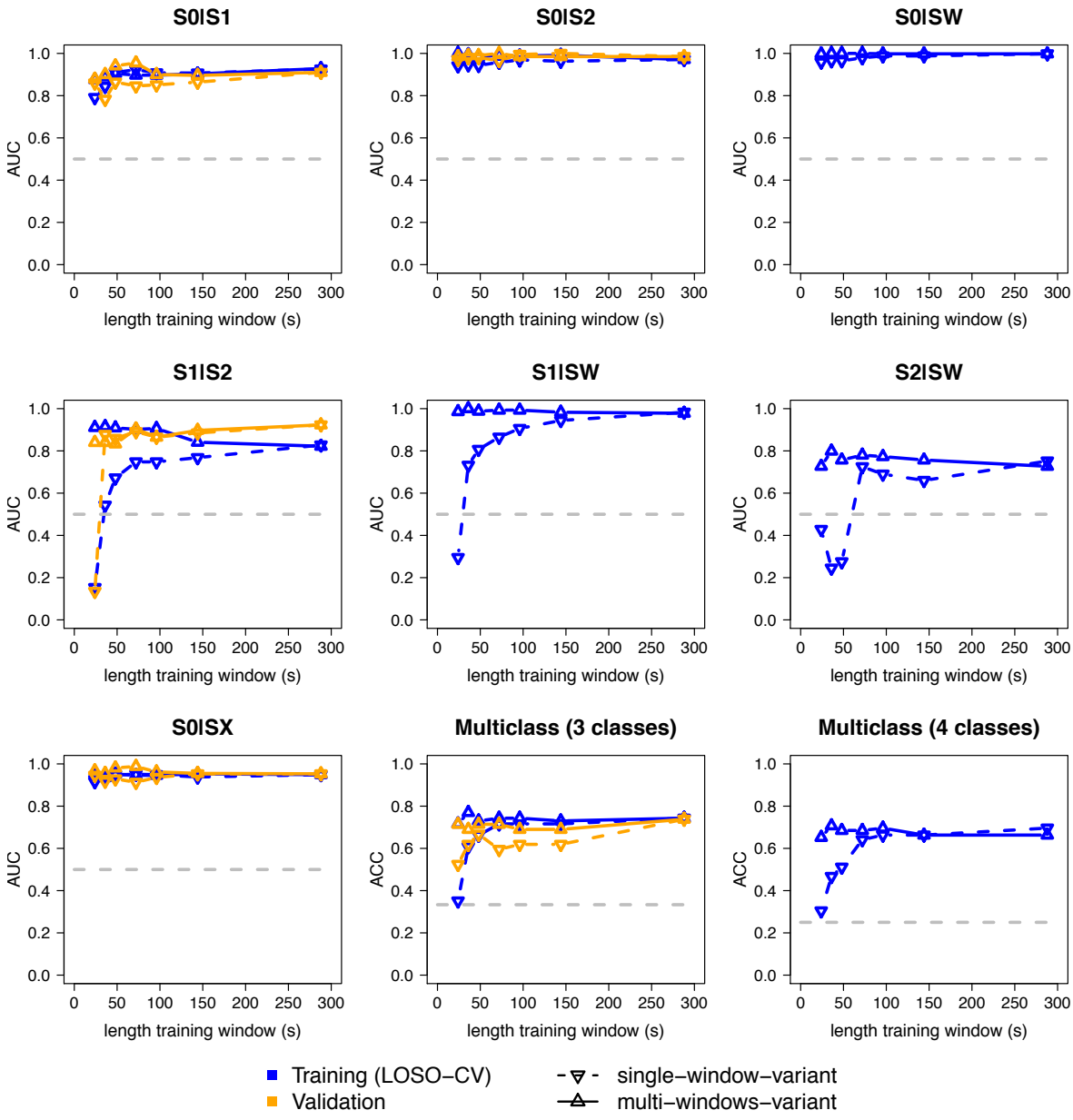


Figure 3: **Discriminative weights derived from linear SVM classifiers.** The discriminative weights for each connection are color coded with warm colors indicating stronger connections during the first stage and cold colors indicating stronger connections during the second stage. E.g., in S0|S1 warm and cold refer to stronger connections in S0 and S1, respectively. Weights are represented as Z-scores and capped at 3.0 for better comparison. The axes are labeled with the lobe or gross anatomical region of the ROI: FR (frontal), CE (central), LI (limbic), OC (occipital), PA (parietal), SC (subcortical), and TM (temporal). ROIs on the y-axis are ordered in the same way as the ones on the x-axis; for details on ROIs see Supplementary Table S2.

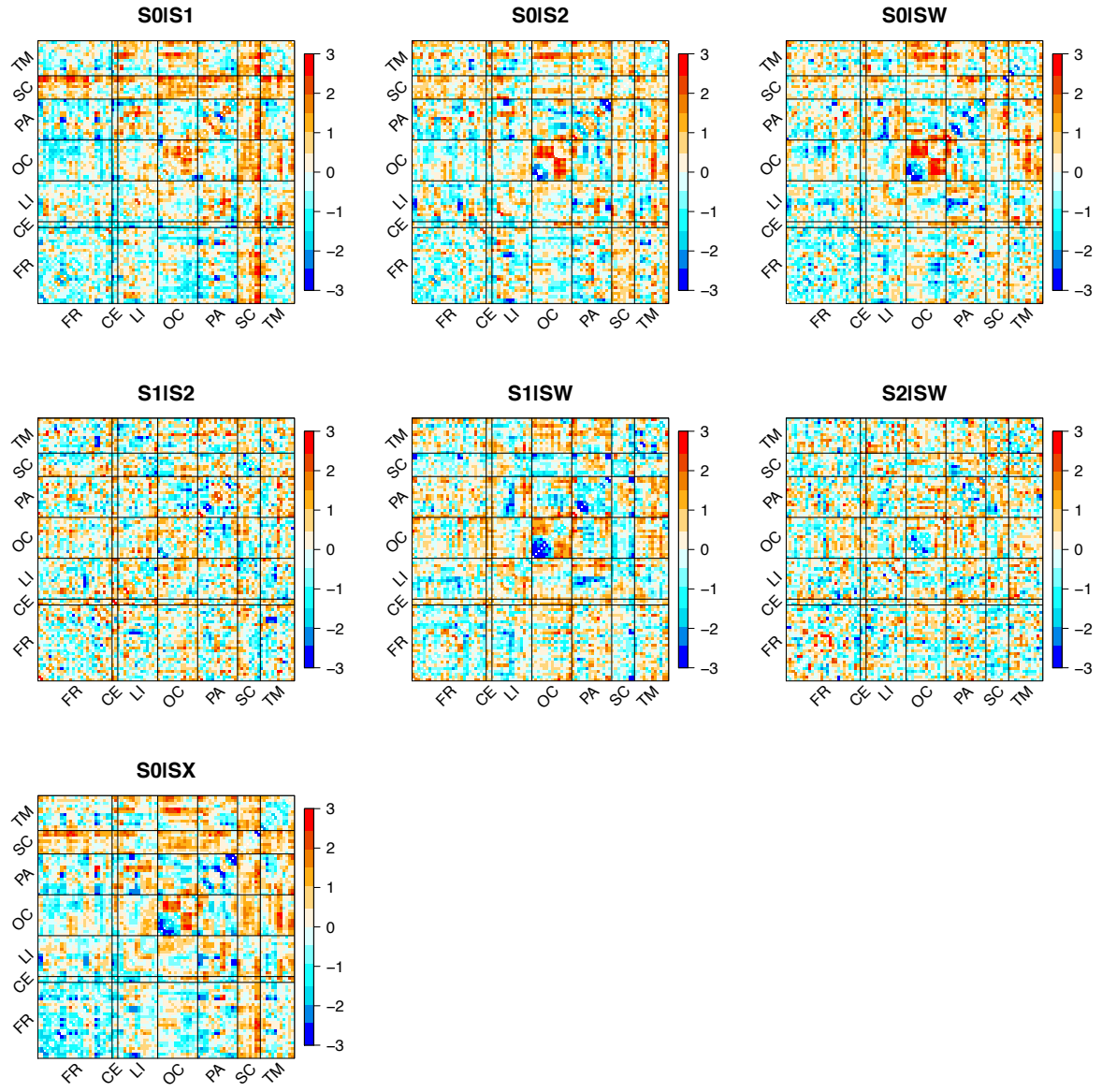


Figure 4: **Visualization of the most discriminant connections for S0|S1.** The matrix highlights the most discriminant connections for S0|S1 (in Z-scores). These 41 connections achieved a p-value of 0.01 or less in the permutation importance test. The connections voting for S0 are coded in red and the connections voting for S1 are coded blue. Surrounding the matrix are anatomical back-reconstructions of these connections in axial (left) and sagittal view (bottom) generated with the Flexible Brain Graph Visualizer (https://sourceforge.net/projects/flexbgv/). Yellow circles represent ROIs, which are scaled according to the number of connections. Edges are colored red and blue according to the coloring in the matrix.

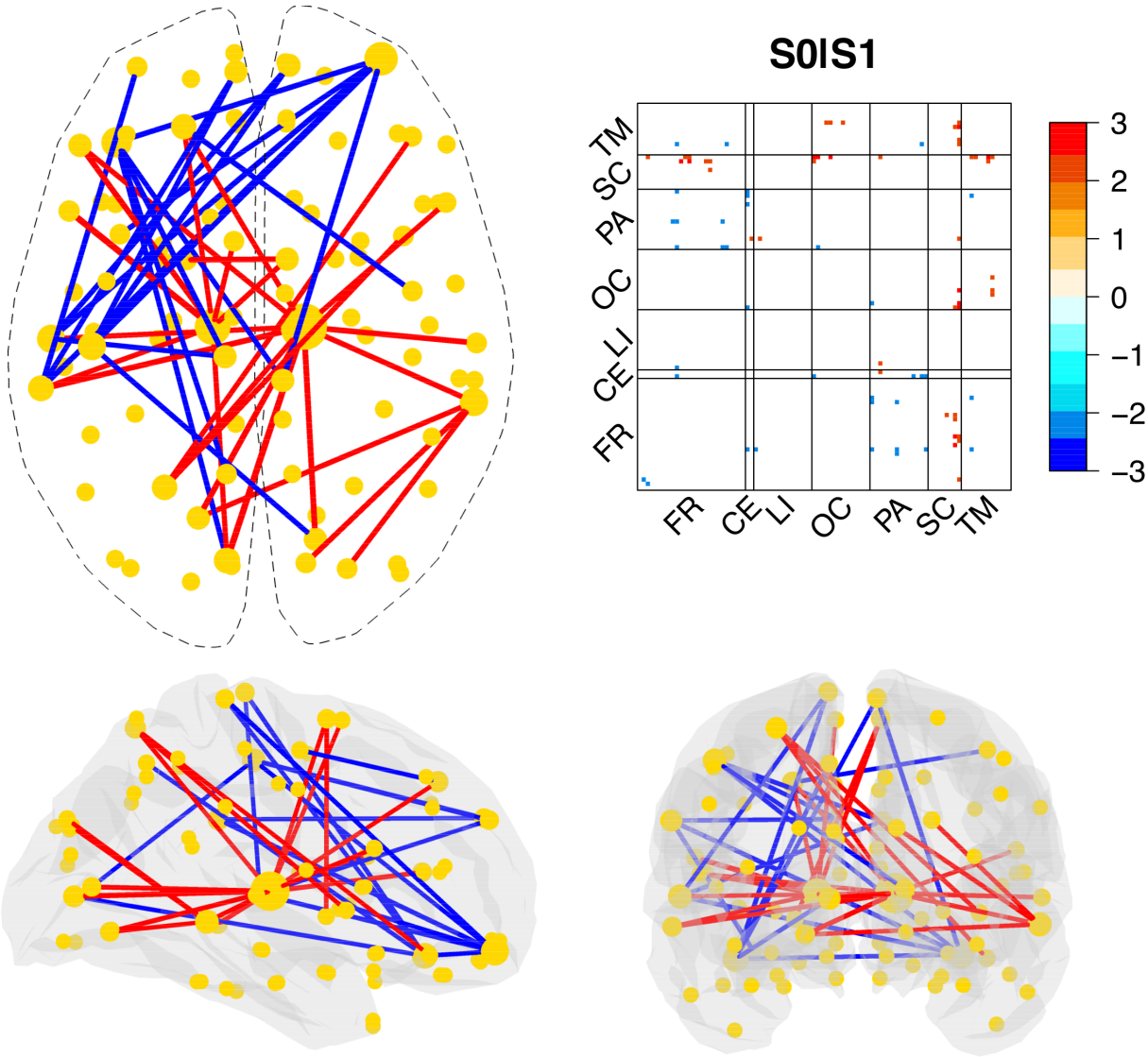
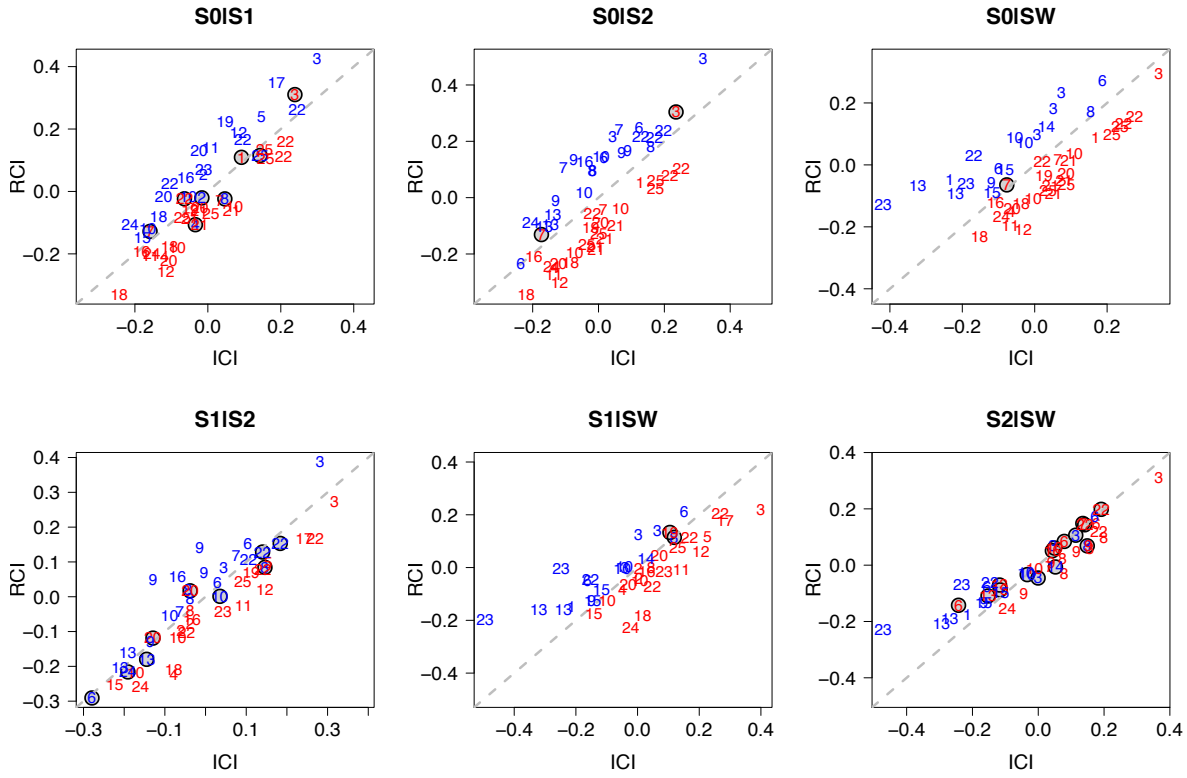


Figure 5: **Increased and Reduced connectivity index for all classification tasks**. Each panel represents one binary classifier with the ICI plotted on the x-axis and the RCI on the y-axis. Each epoch is represented by an integer; epochs originating from the same individual are represented by the same integer across all classification tasks. Misclassified epochs are encircled in black. The dashed grey line represents the decision boundary.



Tables

Table 1: Detailed LOSO-CV classification results on the training dataset using the 288 s window. AUC: area under the ROC curve based on the predicted SVM scores; Wilcoxon P-value: two-sided test based on the predicted SVM scores; ACC: accuracy, i.e., fraction of correct classifications; PPV: positive predictive value, i.e., fraction of correctly classified subjects that were predicted to sleep. For ACC, PPV, Sensitivity and Specificity a threshold of 0.5 was used. The permutation P-value for AUC and ACC was derived by randomly permuting the class labels 1,000 times and then conducting the LOSO-CV. The fraction of instances with equal or larger AUC/ACC values than the unpermuted labels serves as a P-value.

Task	AUC	Wilcoxon P-value	Permutation P-value (AUC)	ACC	Permutation P-value (ACC)	Sensitivity	Specificity	PPV
S0|S1	0.93	6.8e-9	<0.001	0.88	<0.001	0.87	0.89	0.87
S0|S2	0.97	1.8e-11	<0.001	0.92	<0.001	0.88	0.96	0.95
S0|SW	0.99	2.3e-12	<0.001	0.98	<0.001	0.94	1.00	1.00
S1|S2	0.82	7.9e-5	0.001	0.72	0.007	0.54	0.91	0.87
S1|SW	0.99	6.9e-11	<0.001	0.90	<0.001	0.94	0.87	0.85
S2|SW	0.75	5.4e-3	0.005	0.69	0.012	0.67	0.71	0.63
S0|SX	0.95	1.8e-11	<0.001	0.92	<0.001	0.95	0.85	0.94