NeuroImage 253 (2022) 119055

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Ongoing neural oscillations predict the post-stimulus outcome of closed loop auditory stimulation during slow-wave sleep

Miguel Navarrete[a,1,2,*], Steven Arthur[b,1], Matthias S. Treder[b], Penelope A. Lewis[a,*]

[a] Cardiff University Brain Research Imaging Center (CUBRIC), School of Psychology, Cardiff University, Maindy Rd, Cardiff CF24 4HQ, UK
[b] School of Computer Science and Informatics, Cardiff University, Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA, UK

ARTICLE INFO

Keywords:
Slow waves
Sleep
Spindles
Acoustic stimulation
CLAS
Classifiers
Feature importance

ABSTRACT

Large slow oscillations (SO, 0.5–2 Hz) characterise slow-wave sleep and are crucial to memory consolidation and other physiological functions. Manipulating slow oscillations may enhance sleep and memory, as well as benefitting the immune system. Closed-loop auditory stimulation (CLAS) has been demonstrated to increase the SO amplitude and to boost fast sleep spindle activity (11–16 Hz). Nevertheless, not all such stimuli are effective in evoking SOs, even when they are precisely phase locked. Here, we studied what factors of the ongoing activity patterns may help to determine what oscillations to stimulate to effectively enhance SOs or SO-locked spindle activity. Hence, we trained classifiers using the morphological characteristics of the ongoing SO, as measured by electroencephalography (EEG), to predict whether stimulation would lead to a benefit in terms of the resulting SO and spindle amplitude. Separate classifiers were trained using trials from spontaneous control and stimulated datasets, and we evaluated their performance by applying them to held-out data both within and across conditions. We were able to predict both when large SOs occurred spontaneously, and whether a phase-locked auditory click effectively enlarged them with good accuracy for predicting the SO trough (~70%) and SO peak values (~80%). Also, we were able to predict when stimulation would elicit spindle activity with an accuracy of ~60%. Finally, we evaluate the importance of the various SO features used to make these predictions. Our results offer new insight into SO and spindle dynamics and may suggest techniques for developing future methods for online optimization of stimulation.

1. Introduction

Slow wave sleep (SWS) is important for memory consolidation and crucial for metabolic regulation and neural recovery (Klinzing et al., 2019; Xie et al., 2013). This sleep state is characterized by slow wave activity (SWA), mainly associated with continuous epochs of high amplitude slow oscillations (SO, 0.5–2 Hz) (Iber et al., 2007). SOs are thought to be critical for memory consolidation because they provide the necessary neurophysiological conditions for hippocampal-cortical binding (Jiang et al., 2019; Klinzing et al., 2019). Evidence suggest that the coupling of SOs with fast sleep spindles (11–16 Hz) provides the ideal timing for transfer of information from the hippocampus to the cortex (Helfrich et al., 2019; Maingret et al., 2016; Peyrache et al., 2009). Therefore, several non-invasive methods have been proposed for boosting SO amplitudes. These suggest to be useful for improving cognitive and physiological characteristics related to SWA (Marshall et al., 2006; Massimini et al., 2007; Ngo et al., 2013).

Among techniques proposed for increasing the amplitudes of SOs, closed loop acoustic stimulation (CLAS) has proved to be promising in both young (Ngo et al., 2013; Ong et al., 2018) and older participants (Papalambros et al., 2017; Schneider et al., 2020). In CLAS, auditory clicks are applied on the peaks of SOs, increasing SO amplitudes of the next SO and associated phase-locked spindles (Ngo et al., 2013). CLAS has been suggested to improve post-sleep memory retention (Ngo et al., 2013; Papalambros et al., 2017) and to positively impact the immune and the autonomic function of sleep (Besedovsky et al., 2017; Grimaldi et al., 2019). Nevertheless, this technique is not without limitations. Previous studies showed that CLAS is a self-limited process, which increases neither SWA across the night, nor the density of slow waves during SWS (Ngo et al., 2015). Likewise, CLAS has limited ben-

Abbreviations: SO, Slow oscillations; SWS, Slow wave sleep; SWA, Slow wave activity; MCC, Matthews correlation coefficient; CLAS, closed loop acoustic stimulation; FS, Fast spindles; SS, Slow spindles; RF, Random Forest; FI, Feature importance.

[*] Corresponding authors.

E-mail addresses: mg.navarrete144@uniandes.edu.co (M. Navarrete), LewisP8@cardiff.ac.uk (P.A. Lewis).

[1] These authors contributed equally to this work.

[2] Present address: School of Biomedical Engineering, University of Los Andes, Bogotá, Colombia.

https://doi.org/10.1016/j.neuroimage.2022.119055.
Received 11 December 2021; Received in revised form 26 February 2022; Accepted 1 March 2022
Available online 8 March 2022.

efits in terms of improving the consolidation of visual and procedural memories previously associated with SOs (Leminen et al., 2017). Furthermore, increases in SO amplitude and power through this technique are apparently not beneficial enough to increase memory consolidation (Henin et al., 2019), and sensitivity to CLAS stimulation reduces with age (Schneider et al., 2020).

All this prior evidence suggests that the impact of CLAS is not exclusively determined by the effects of the sensory input, but is also influenced by ongoing neural processes in the SO (Navarrete et al., 2020a). Such ongoing processes therefore determine whether an auditory stimulus can increase SO amplitude. In theory, two principal neural processes could be determining the outcome of each auditory click on the SO. The first process relates to the type of wave that is stimulated. For instance, a recent study in rodents suggests dissociable functional roles for fast and slow oscillatory elements during SWA (Kim et al., 2019). Likewise, in humans, different memory mechanisms have been proposed for SO elements depending on their origin, spread and spindle locking (Bernardi et al., 2018; Helfrich et al., 2019; Siclari et al., 2014). Nevertheless, the auditory modulation of SOs by auditory clicks has typically been evaluated as if all waves originated from the same mechanisms, and as if every stimulus were equally efficient for increasing cognitive or physiological SO functions. A second process may be established by the ongoing neurophysiological processes during the SO that could modulate the response to the click. Previous work theorised that the outcome of the stimulus (measured as 'no response', 'increased SO amplitude' or 'sleep arousal') is determined by a sweet spot in stimulation that, if targeted appropriately, allows boosting of the SO without causing arousal (Bellesi et al., 2014). This work suggested that this sweet spot is determined by the SOs themselves, since SOs can supress inputs from the locus coeruleus (LC), thus preventing arousals in response to sound stimuli while increasing the SO amplitude. Unfortunately, the nature of the neural threshold determining this sweet spot is still unknown, leaving a lot of uncertainty about the outcome of CLAS stimulation, as little is known about the responsiveness of each individual SO. These two processes are not mutually exclusive, and these may set the fluctuating properties of SOs that could co-determine the responses to the stimulation. However, there are not a priori rules to determine when a SO will have a large amplitude spontaneously or when the optimal large-wave response will occur in response to the CLAS click.

Using pre-existing databases of CLAS responses, in this retrospective study we aimed to examine the factors predicting the spontaneous (non-stimulated) and click-evoked response on each SO. We hypothesize that if cortical activity modulates the response of each stimulus by differentiated SO processes, then it would be possible to predict subsequent spontaneous and induced cortical dynamics. To test this, we trained machine learning models to offline predict the outcome of cortical activity in both spontaneous and stimulation conditions. We used a series of features based on morphological characteristics of the ongoing SO and the estimated timing of the click stimulation. Our results show that the trained models can predict whether subsequent SO amplitude were high or low relative to the average, and whether an increment of spindle activity occurred. We did this in both unstimulated SWS and CLAS conditions. Furthermore, using a feature importance analysis we then evaluated the variables that allowed us to determine the strength of response of the ongoing wave to the auditory click. We found that ongoing SWS dynamics may predict the event amplitude of both spontaneous and stimulated SOs, and that the threshold of maximal responsiveness is determined by the level of cortical activation during the click as measured by the amplitude of SOs. From this, we argue that the response to stimulation is mediated by the high drive for cortico-thalamic activation and the reduced cortico- coerulear drive of large SO positive cycles. Our findings support the idea that SWS is comprised of several types of SOs. This analysis of the response to CLAS may help to unravel the function of the SO, as well as to pave the way for more targeted stimulation and enhancement of SWS.

2. Materials and methods

2.1. Datasets and experimental procedures

The data corresponds to the Young cohort from (Navarrete et al., 2020a). Briefly, polysomnographic data including EEG and hypnogram of 21 young adult individuals were analyzed (14 females and mean ± SD age = 25.7 ± 4.7 years). The participants spent two experimental nights in the laboratory undergoing one experimental stimulation (STIM) and one no-stimulation condition (SHAM). The order of experimental conditions was balanced across subjects and separated by at least one week. All protocols were approved by the appropriate ethic committees of local institutions (University of Lübeck – U.Lub – and University of Los Andes – U.And –) and written consents were obtained for each participant ($n = 11$, U.Lub; $n = 10$, U.And).

Acoustic stimuli for the STIM condition consisted of stereophonic clicks of pink noise (50 ms duration) with rising and falling slopes (5ms duration), and stimulation timestamps were recorded online when clicks were applied in STIM (stim-click) or for when they were predicted in SHAM (sham-click). For the SHAM condition, the detection protocol was identical to STIM, but the sound was muted. Briefly, the streamed signal was filtered in the SO frequency band (U.Lub: 0.25–4 Hz; U.And: 0.5–2 Hz) and negative EEG deflections that surpassed an adaptative threshold were identified as a SO down-state during SWS (U.Lub: -80 μV; U.And: -60 μV). These thresholds were independently updated: each 0.5 s to the largest negative amplitude lower than the default from the last 5 s for U.Lub whereas for U.And, this threshold was updated to half the amplitude of the detected SO if these values was larger than default. After trough detection, two consecutive clicks were applied on subsequent SOs aiming for the subsequent SO peaks (interstimulus trial interval: uLub = 1.076 s; uAnd = 1.053 s ± 0.06 s). After each trial, there was a pause for 2.5 s before trough detection was resumed in both detection methods. Stimulation was applied during sustained non-rapid eye movement sleep (NREM), including N2 and N3 stages, and this was manually halted if there was visual evidence of arousals or REM (Navarrete et al., 2020a).

2.2. Sleep and EEG analysis

For the analysis and pre-processing of sleep EEG, we closely followed the procedures from (Navarrete et al., 2020a). Sleep scoring was performed according to ASSM scoring criteria (Iber et al., 2007) by two trained experimenters blinded to stimulation conditions. All artefacts and arousals were marked in the hypnogram. We focused on events detected in Fz (referenced to linked mastoids) for SOs and sleep spindles activity (SA) because these are the locations where SO are more pronounced (Iber et al., 2007). The raw data were resampled at 200 Hz with linear interpolation after applying an antialiasing low pass FIR filter. After processing, only stimulations marked during N3 sleep stage were retained. SWA was obtained from the EEG signals filtered between 0.5 and 2 Hz using a zero-phase windowed equiripple FIR filter (3dB at 0.25 and 3.08 Hz; >37 dB at $f < 0.01$ Hz and $f > 4$ Hz). Waves were only considered as SOs when their negative deflection had consecutive zero crossings between 0.25 and 1.0 s, regardless of the wave amplitude (Riedner et al., 2007). Likewise, spindle activity was determined by applying a zero-phase bandpass FIR filter between 11 and 16 Hz (3 dB at 10.62 and 17.38 Hz; >40 dB at $f < 10.01$ Hz and $f < 18$ Hz). Then, the root mean squared (RMS) was computed by using a time window of 0.2 s. Rejection criteria included trials where stimulus was applied less than 2 s apart from any arousal, artefact, or changes to another sleep stage other than N3. The response to the stimulus was analyzed for the first click of each trial. Therefore, trials in which the subsequent stimuli were wrongly placed before the subsequent SO trough were excluded from the analysis. Hence, for analysis we included 459.19 ± 242.10 trials for SHAM and 289.33 ± 178.18 trials for STIM. For spindle event detection, the candidate events were first detected as discrete events

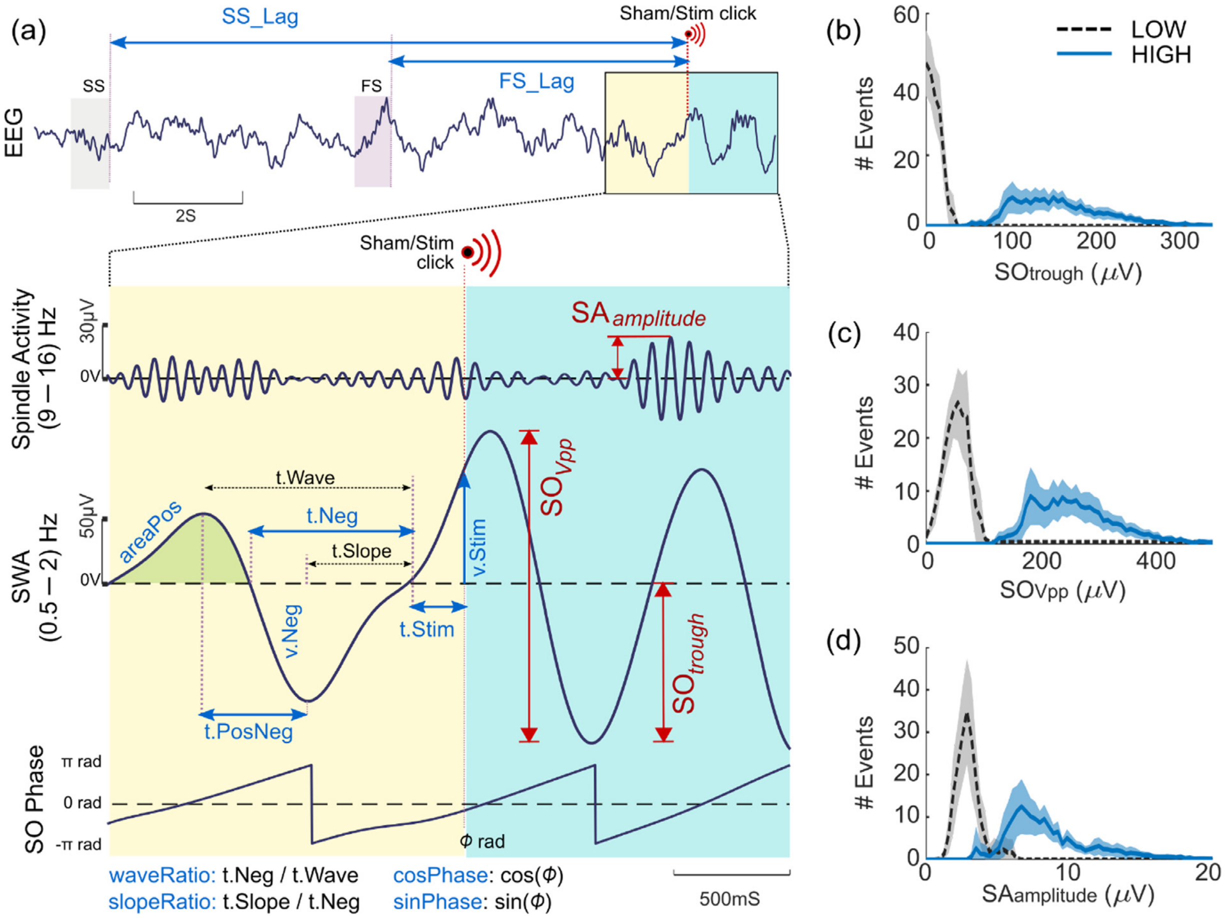

Fig. 1. Selection of features for training of classifiers and the post-stimulus measures evaluated as response to click stimulation together with between-subject average distribution of the events labelled LOW and HIGH. (a) Three post stimulus measures were assessed in the study (in red; SO_{Vpp}: peak-to-peak SO amplitude, SO_{trough}: SO's trough amplitude and $SA_{amplitude}$: spindle activity amplitude). The features used to train the machine learning model are marked in blue, and they include information from pre-stimulus fast spindles (FS: 12–16 Hz) or slow spindles (SS: 9–12 Hz) (e.g. FS_Lag: Time to the previous FS, SS_Lag: Time to the previous SS), as well as information of the pre-stimulus SO wave (e.g. v.Neg: negative voltage, t.Neg: time of negative wave) and information about the timing of stimulus (e.g. v.Stim: SO voltage during stimulation, sinPhase: sinus of phase of stimulation). At the stimulus time (Stim) an auditory click was applied in STIM trials whereas a muted sham-click was marked in SHAM trials. A meticulous description of all evaluated features is presented in Table 1. (b) Between-subject distribution of LOW and HIGH classes for SO trough amplitudes (SO_{trough}). (c) Between-subject distribution of LOW and HIGH classes for SO peak-to-peak amplitudes (SO_{Vpp}). (d) Between-subject distribution for LOW and HIGH classes for spindle activity amplitude ($SA_{amplitude}$). Dispersion areas represent mean ± 95% CI of the distribution.

where the RMS signal surpassed a threshold established as the 86.64 percentile (equivalent to 1.5 SD over the mean for a Gaussian distribution) of the spindle activity during N3. Then, spindles were identified as events with duration between 0.3 and 3 s (Warby et al., 2014), with at least five oscillations, a unimodal peak in the spindle frequency band and decreasing power for higher frequencies computed by the Morlet wavelet (Purcell et al., 2017). The frequency band for fast spindles (FS) was defined as 11–16 Hz, and in the 9–12 Hz for slow spindles (SS). Finally, to avoid differences due to phase distribution, ERP analyses (Section 3.5) were computed for events where sham and stim click were applied within the phase interval Φ: -π/4 to π/4 (Navarrete et al., 2020a). Therefore, the selection ratio for only the ERP analyses included 59.76% ± 10.98% of the trials for SHAM and 63.80% ± 9.78% of the trials for STIM.

2.3. Selection of features and labels for classification

For the selection of the model features, we selected a series of morphological wave features taken before and during the stimulation of the ongoing wave as shown in Fig. 1. These predictor features, further defined in Table 1, describe morphological characteristics of the pre-stimulus slow-wave structure (*v.Neg, t.Neg, t.PosNeg, areaPos, waveRatio* and *slopeRatio* in Table 1), structural characteristics of SO detection and peak estimation (*cosPhase, sinPhase, v.Stim* and *t.Stim* in Table 1) and long term dynamics of spindle activity (*FSonStim, SSonStim, FS_Lag* and *SS_Lag* in Table 1). Additionally, to control for random effects in the analysis of feature importance, two dummy variables were included. These were the features representing the subject ID and a random generated number (*sbjID* and *random* in Table 1).

Two post-stimulus measures were selected for classification of SO analysis and one for SS. All labels to be predicted were computed from wave measurements taken after stimulation. These correspond to the trough amplitude of the post-stimulus SO (SO_{trough}), the peak-to-peak amplitude of the SO after the stimulus (SO_{Vpp}), and the amplitude of the spindle activity locked to the subsequent positive SO after the stimulation ($SA_{amplitude}$). The selection of these variables was specified to include one measure with independent dynamics of the ongoing SO ($SA_{amplitude}$), one measure with reduced dynamics of the ongoing wave (SO_{trough}) and one measure including dynamics of the ongoing SO (SO_{Vpp}). Furthermore, all of these measures have been shown to be modulated by the CLAS click (Navarrete et al., 2020a; Ngo et al., 2013; Schneider et al., 2020). Nevertheless, as these evaluated measures represent a continuous distribution, post-stimulus measures were binarized for each recording.

Table 1
Selection of wave and stimulation characteristics used as predictors for classification.

Wave and stimulation characteristics used as predictors:		
Feature	Description	Variable
SbjID	Subject identification number	Ordinal
cosPhase	Cosine of the phase of auditory stimulation	Continuous
sinPhase	Sine of the phase of auditory stimulation	Continuous
v.Neg	Voltage of SO trough before the click	Continuous
v.Stim	Voltage in the moment of stimulation	Continuous
t.Neg	Time of duration for the negative wave before click	Continuous
t.PosNeg	Time of duration for the peak to trough before click	Continuous
t.Stim	Time between zero-crossing to click time	Continuous
areaPos	Area under curve for peak section before the click	Continuous
waveRatio	Duration ratio for the wave before click	Continuous
slopeRatio	Duration ratio for the negative wave before click	Continuous
FSonStim	Presence of fast spindle on stimulation (Yes: 1, No: 0)	Binary
SSonStim	Presence of slow spindle on stimulation (Yes: 1, No: 0)	Binary
FS_Lag	Time to the previous FS (12–16 Hz)	Continuous
SS_Lag	Time to the previous SS (9–12 Hz)	Continuous
random	Random dummy variable	Continuous
Wave characteristics used as labels:		
Feature	**Description**	
SO_{trough}	Trough amplitude of the post stimulus SO	
$\mathrm{SO}_{\mathrm{Vpp}}$	Peak to trough amplitude of the SO after the stimulus	
$\mathrm{SA}_{amplitude}$	Maximal amplitude of the spindle activity locked to the subsequent positive SO after the stimulation	

Our goal was to investigate whether stimulation results in a strong or a weak neural response. Predicting the post-stimulus values per se using regression would be an interesting alternative approach. However, the physiological function of SO and SS is determined mostly by the global dynamics of these rhythms rather than exclusively their scalp amplitude (Klinzing et al., 2019; Navarrete et al., 2020b). This is also supported by the variability of NREM activity (in frequency and EEG voltage) induced by interindividual differences such as age or sex (Bódizs et al., 2021). A specific SO trough voltage (e.g. 60 μV) may be considered to be a low amplitude wave in a young participant, but the same value may be within the 90th percentile of SO waves of an older participant. Therefore, we divided the measured amplitudes depending on the statistical distribution of these values in order to predict the response to auditory clicks. We focused on predicting strong vs weak responses for each participant rather than the exact prediction of event amplitudes in micro-volts. Thus, we thresholded post-stimulus values into LOW and HIGH categories and cast the analysis as a classification problem.

Labelling binarization was performed to determine Low and High values from the selected post-stimulus measures. For this, a thresholding process was applied for each subject. First, outlier values were eliminated by removing the highest and lowest 1%. Then, an empirical cumulative function was computed for each recording. For each subject, lower and higher limits were identified independently for each stimulation condition. Values lower than the 25th percentile were marked as the local low threshold, while values higher than the 75th percentile were determined as the high threshold. As CLAS increases the likelihood of larger amplitude events, it might be unclear whether the amplitude of the middle 50% of trials is caused by ongoing spontaneous activity, or whether they have been weakly boosted by the stimulus. Therefore, this middle 50% of trials were used for neither training nor testing. However, they were included for evaluating the average amplitudes when the classifiers were applied under more natural conditions (Generalization of classifier predictions). Subsequently, the low and high thresholds were averaged between conditions for each subject to determine subject-based thresholds across conditions. Then, values higher than the subject-based high threshold were marked as HIGH, and values lower than the subject-based low threshold were marked as LOW. An equal number of LOW and HIGH labels were randomly selected for each subject, and all LOW/HIGH trials of all subjects were used in the training process. Nevertheless, for the testing of each measure we excluded those subjects with a low number of selected trials ($N <$ 10th percentile for all subjects and measures). This led to have a total population of $N = 19$ for SO_{trough} and $\mathrm{SO}_{\mathrm{Vpp}}$, and a population of $N = 16$ for $\mathrm{SA}_{amplitude}$. From this, the SHAM dataset was built with trials identified in the non-stimulation condition whereas the STIM dataset was built with trials marked as auditory stimulated. After pre-processing and trial selection, an average of 132.2(86.6) events remained for each subject, condition, and measure (Total number of trials for SHAM dataset: SO_{trough}, $N = 2890$; $\mathrm{SO}_{\mathrm{Vpp}}$, $N = 3036$; $\mathrm{SA}_{amplitude}$, $N = 3874$. Total number of trials for STIM dataset: SO_{trough}, $N = 1988$; $\mathrm{SO}_{\mathrm{Vpp}}$, $N = 1986$; $\mathrm{SA}_{amplitude}$, $N = 1946$).

2.4. Training of the classifier

Classification models were trained using supervising learning. In supervising learning, these models try to find a good approximation of an unknown function *f* given paired examples (θ, $f(\theta)$), where θ is a set of features describing each dataset. Several classification algorithms were tested for this study. These included Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR). RF was found to give the highest or close-to-highest accuracies in most cases. SVM was found to give higher accuracies in some cases, but the results took significantly longer to compute whereas LR got lower accuracy in the classification. RF was therefore chosen because it presented the best trade-off between accuracy and computation time. Likewise, RF has also proved to be efficient in other problems based on sleep EEG (da Silveira et al., 2017; Dimitriadis et al., 2020). Consequently, we independently trained classification models for SHAM and STIM conditions using the RF algorithm.

A RF is an ensemble method or a meta-classifier that combines the outputs of a collection of decision trees. A decision tree is an algorithm that performs classification by performing a cascade of binary decisions (‘branch off left’ or ‘branch off right’). In each binary decision, a single feature is compared against a threshold value. The threshold is determined during training such that the classes are optimally separated using a metric called Gini impurity. A RF creates many such decision trees. To encourage different trees to focus on different aspects of the data, each tree is exposed to a different subset of the data. This is achieved by randomly selecting subsets of the samples. A RF performs classification by collecting ‘votes’ from all the decision trees and then selecting the majority vote. RF has been shown to be a good out-of-the-box classifier that is robust against overfitting, but it is not very sensitive to outliers and its hyperparameters are easy to set. In addition, there is no need for RF to prune the trees which helps in gaining higher accuracy (Breiman, 2001). We implemented decision trees grown with surrogate splits to optimize predictions, achieved by estimating tree splits using complementary features (Hapfelmeier et al., 2012). Hyperparameters were optimized before the learning process. The values were adjusted to optimize the performance of the algorithm. In our RF classifiers, the two hyperparameters were the number of trees in the forest and the maximal number of levels in each decision tree.

For training, all selected trials of all subjects were pooled. To obtain an unbiased estimate of generalization performance, leave-one-subject-out cross-validation (LOOOCV) was performed: in each iteration, one subject was left out and the classifier was trained using trials from the other subjects. This was then repeated, leaving each subject out in turn. To tune the model's hyperparameters, nested cross-validation was performed: in each iteration, the training data were again split into a training set (70%) and a validation set (30%). The validation set was used to find the best hyperparameters. The optimal model was then taken forward and tested on the held-out subject. This resulted in a total of N classifiers being trained for each measure in both SHAM and STIM cases, where N corresponds to the total of subjects evaluated.

Hence, considering θ_{SH} the set of features for the SHAM dataset, and θ_{ST} the set of features for the STIM dataset, we will denote as f_{SH} the RF classifier trained in the SHAM dataset *(θ_{SH}, $f(\theta_{SH})$)*. Similarly, we will denote as f_{ST} the RF classifier trained in the STIM dataset *(θ_{ST}, $f(\theta_{ST})$)*.

To evaluate the performance of the binary classification we used the Matthews correlation coefficient (MCC) (Boughorbel et al., 2017) which is a robust test for binary decisions. These metrics were calculated according with the predictive forecasting of the binary classification model. Therefore, defining HIGH instances as positive (P) and marked LOW instances as negatives (N), every classification response can be grouped within four cases: (i) True positives (TP): positive values that are predicted as positive; (ii) False negatives (FN): positive values wrongly marked as negatives; (iii) False positives (FP): Negative values classified as positives, and (iv) True negatives (TN): actual negative values correctly classified as negatives. Hence, the unbiased performance evaluated in the MCC sense is determined by:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

This value ranges from -1 (perfect misclassification) to 1 (perfect classification), where 0 indicates random labelling.

Accuracy (ACC) accounts for the proportion of correctly classified samples on all classes:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

As well as ACC, the MCC gives high scores for correctly labelled classes. Nevertheless, MCC is a stricter measure with high scores only if the model can predict a high percentage of true positives and a high percentage of true negatives on any balanced or imbalanced dataset (Chicco and Jurman, 2020). Therefore, we used MCC as the main metric for evaluating training performance whereas ACC was used as a descriptive metric.

Finally, to discern the performance differences in STIM or SHAM datasets, we compared the performance of each model applied within and cross-condition (i.e., SHAM model predicting SHAM labels $f_{SH}(\theta_{SH})$ vs STIM model predicting SHAM labels $f_{ST}(\theta_{SH})$ as well as SHAM model predicting STIM labels $f_{SH}(\theta_{ST})$ vs STIM model predicting STIM labels $f_{ST}(\theta_{ST})$.

2.5. Feature importance

It is possible to achieve high performance in both f_{SH} and f_{ST} classifier models. However, the features important for making the prediction may differ between the classifiers, with the importance of some features common to both models. Here we were interested in how the models differed, and which features were uniquely important for one model or the other.

We assessed the relevance of each feature by computing an unbiased measure of feature importance. We implemented a heuristic metric based on the permutation importance of each feature evaluated in the holdout subject (Altmann et al., 2010). For each feature, the algorithm permutes the values of the evaluated feature *S* times and applies the trained algorithm to predict the original labels. The permutation makes a feature uninformative by destroying its relationship with the class labels. The permutation increases the model's prediction error that derives in a distribution of *S* null performances. The importance of each feature is thus determined by the change in the model's performance. Hence, big changes in the error are expected for important features whereas none or small changes are expected for less important features.

In our analysis, we first estimated the MCC performance of the original model (*H*). Subsequently, we randomly permute the observations of x_j to estimate the performance of the model using the altered features using one hundred repetitions ($S = 100$) for each feature (x_j, $j = 1, 2, \ldots$ *N* features). For each permuted feature, this resulted in a distribution ($H^*{}_j$) of one hundred performance values. Then, we took the differences $d_j = H - H^*{}_j$ and computed the mean D_j and standard deviation σ_j. We then defined the feature importance (FI) by permutation for x_j as the *z*-value $FI_j = D_j / \sigma_j$. Changes in performance of the SHAM model in the SHAM dataset ($FI_{SH} = FI(f_{\mathrm{SH}}(\theta_{SH}))$) and changes of the STIM model in the STIM dataset ($FI_{ST} = FI(f_{\mathrm{ST}}(\theta_{ST}))$) represent within-condition predictor weights. Likewise, changes in performance of the SHAM model in the STIM dataset ($FI_{STx} = FI(f_{\mathrm{SH}}(\theta_{ST}))$) and changes of the STIM model in the SHAM dataset ($FI_{SHx} = FI(f_{\mathrm{ST}}(\theta_{SH}))$) represent cross-condition predictor weights. Finally, we relied on the assumption that highly important features in both models (SHAM and STIM) should be significantly greater than the dummy *random* feature included in the model.

For determining features describing similar dynamics between conditions, we relied on the assumption that features equally important in both models (SHAM and STIM) have similar changes in performance when applied within the same dataset. This rationale is further described in Table S1. For each model, the predictor weights represented by the *FI* values represent the vector space of random variables in each dataset. Then, we determined the correlations in between *FI* values when SHAM and STIM were evaluated in the same dataset using the cross-condition models (e.g. FI correlation in SHAM dataset $\rho_{\mathrm{SH,ST}}(\theta_{SH}) = \mathrm{corr}(FI_{SH}, FI_{SHx})$, and FI correlations in STIM dataset: $\rho_{\mathrm{SH,ST}}(\theta_{ST}) = \mathrm{corr}(FI_{STx}, FI_{ST})$). Hence, correlated *FI* variables between models are not independent in the random space (Papoulis and Pillai, 2002), and highly correlated features suggest that these predictors may describe similar processes in SHAM and STIM datasets (Table S1).

2.6. Statistics

To compare classifier performances within and between conditions, we implemented a Wilcoxon signed rank test across holdout subjects. This is equivalent to implementing a 21-fold cross-validated signed-rank test (Thomas, 1998). Also, we evaluated whether the performance of the trained models were statistically different by computing the McNemar's test which evaluates the misclassification rates between models (Thomas, 1998).

A one-way analysis of variance (ANOVA) was calculated on FI; then the features were compared against the dummy random feature to check for random effects using a Tukey's Honestly Significant Difference procedure. For comparing FI ranks, we compared the Spearman's rank correlation within and between conditions. Briefly, we split the STIM and SHAM datasets into half of the subjects. This resulted in four groups: sham group 1 ('SH-G1′), sham group 2 ('SH-G2′), stim group 1 ('ST-G1′), and stim group 2 ('ST-G2′). For this, we use a Monte Carlo method to sample the subjects uniformly random in each half in 1600 non-repeated combinations. For each combination of half subjects, we computed Spearman's rank correlation of the FIs. These correlations were subsequently normalized using a Fisher *z*-transformation. Then, the median Fisher z value of each permutation was obtained, generating thus a distribution of permuted correlations Fisher *z*-transformed for each of the four groups. After obtaining the Fisher z distributions for each group, we applied a Kruskall-Wallis ANOVA to test the null hypothesis that the sample data from each group comes from the same distribution.

For comparing EEG trials, we computed significant differences between STIM vs. SHAM within subjects using the Welch unequal variance t-test with the Moser–Stevens correction for degrees of freedom (Moser and Stevens, 1992).

Further differences between groups were computed using paired t-tests. The Benjamini & Hochberg false discovery correction (FDR) for multiple comparisons was applied to control for the family-wise error rate (Benjamini and Hochberg, 1995).

3. Results

3.1. Overview of datasets and variability of selected measures

The use of two different datasets using slightly different methods for wave detection in CLAS may have induced some variability in our data. We did not find significant differences of sleep macro-structure between stimulation conditions either when the datasets were evaluated individually or when all participants were summed up (Table S2). Previous analyses have also shown that both CLAS methods were able to detect online SOs with amplitudes larger than 90% of the amplitudes within the same period of stimulation (Navarrete et al., 2020b). Thus, these analyses suggest that trials from both datasets are indeed comparable.

On the other hand, the binarization process determined subject-specific threshold values that differentiated low and high amplitude events Fig. 1.b, c show the average distribution of labelled LOW and HIGH events. A paired t-test showed that the differences between HIGH and LOW values were significant on a group level for SO_{trough} (mean $\mathrm{LOW}_{\mathrm{Thres}}$ = 24.7 μV, 95%CI = 21.7 μV to 27.7 μV; mean $\mathrm{HIGH}_{\mathrm{Thres}}$ = 105.6 μV, 95%CI = 92.5 μV to 118.7 μV; t(22.2) = -11.8, p < .001) as well as for $\mathrm{SO}_{\mathrm{Vpp}}$ (mean $\mathrm{LOW}_{\mathrm{Thres}}$ = 79.9 μV, 95%CI = 73.5 μV to 86.3 μV; mean $\mathrm{HIGH}_{\mathrm{Thres}}$ = 188.0 μV, 95%CI = 170.5 μV to 205.5 μV; t(25.3) = -11.4, p < .001) and $\mathrm{SA}_{amplitude}$ (mean $\mathrm{LOW}_{\mathrm{Thres}}$ = 3.7 μV, 95%CI = 3.5 μV to 3.9 μV; mean $\mathrm{HIGH}_{\mathrm{Thres}}$ = 6.5 μV, 95%CI = 6.1 μV to 6.9 μV; t(21.5) = -12.2, p < .001).

The fact that slightly different methods were used for online wave thresholding during CLAS in the two datasets means that some additional variability may have been induced into the recordings of the raw signal. This could have some spurious effects on the values used for thresholding the LOW and HIGH classes from the raw data. We therefore compared LOW and HIGH threshold between the two evaluated datasets and found no significant differences between them (Table S3). From these results, we highlight that the $\mathrm{LOW}_{\mathrm{Thres}}$ for both SO measures are typical for events that are not considered SOs while the $\mathrm{HIGH}_{\mathrm{Thres}}$ are typical for SOs considered as high amplitude (Iber et al., 2007). The same characteristic can be considered for SA amplitude LOW and HIGH thresholds (Purcell et al., 2017). Hence, it is important to note that the LOW/HIGH values do not differentiate between no-event/event trials, but rather between no-event/high-event trials.

3.2. Classification performance

Overall, our trained classifiers achieved good accuracies for both SHAM and STIM models and these were comparable when evaluated on within conditions datasets. Specifically, f_{SH} and f_{ST} models had comparable performance in all three evaluated measures: SO_{trough} (SHAM:ACC= 0.71, 95%CI = 0.69 to 0.74; STIM:ACC= 0.71, 95%CI = 0.67 to 0.74, p = .601), $\mathrm{SO}_{\mathrm{Vpp}}$ (SHAM:ACC= 0.84, 95%CI = 0.82 to 0.86; STIM:ACC= 0.83, 95%CI = 0.81 to 0.85, p = .295) and $\mathrm{SA}_{amplitude}$ (SHAM:ACC= 0.60, 95%CI = 0.56 to 0.64; STIM:ACC= 0.59, 95%CI = 0.55 to 0.62, p = .691). This indicates that it is equally possible to predict SO amplitude outcomes from SO morphology in spontaneous and stimulation conditions. However, the amplitude of spindles locked to SOs may be harder to predict as the accuracy was lower when compared to the prediction of SO amplitudes.

We were especially interested in determining whether characteristics of the ongoing spontaneous SO and spindle signal could predict if the auditory stimulation would result in a LOW or HIGH SO outcome. To test this, we evaluated the performance of the trained classifier in a cross-classification paradigm. Thus, classifiers trained on the STIM dataset were applied to the SHAM dataset to predict trial labels in SHAM, and classifiers trained on the SHAM dataset were applied to the STIM dataset to predict STIM trial labels. This process helped us to determine the level at which information from unstimulated SWS (SHAM dataset) can be used to predict SO characteristics after an auditory stimulation (STIM dataset), thus providing an idea of how much the auditory stimulation disrupts or alters the ongoing SO pattern. Similar performance of f_{SH} and f_{ST} when applied within condition indicates a similar degree of generalization for the two models within their own dataset. Equivalent performance of these two classifiers across conditions (cross-classification) would indicate that both classifiers have similar mappings of the feature hyperspace (Thomas, 1998). In other words, similar performance in the cross-condition denotes that the prediction capabilities of these models were mostly based on the characteristics of the spontaneous signal. Conversely, changes in the performance of classifiers in the cross-condition indicate that the mapping of the feature hyperspace is different, and these differences are caused by the auditory click.

Like for the ACC comparisons, a Wilcoxon signed rank test on the MCC revealed no significant differences in within dataset classification for trough, peak-to-peak voltage, or amplitude measures ($f_{\mathrm{SH}}(\theta_{SH})$ vs $f_{\mathrm{ST}}(\theta_{ST})$), for SO_{trough} (Z = 0.56, p = .573), $\mathrm{SO}_{\mathrm{Vpp}}$ (Z = 1.41, p = .159) and $\mathrm{SA}_{amplitude}$ (Z = 0.62, p = .535) (Fig. 2a). Hence, we accept the null hypothesis regarding comparisons of performance of all three measures within the trained datasets suggesting that f_{SH} and f_{ST} perform similarly when applied within condition.

The picture was different in the crossover analysis, where classifiers were trained on one dataset, then applied on both, and classification performance was compared. Firstly, we found significant differences between MCC rates when both classifiers (trained on Stim f_{ST} and Sham f_{SH}) were applied to the SHAM dataset (i.e., $f_{\mathrm{SH}}(\theta_{SH})$ vs $f_{\mathrm{ST}}(\theta_{SH})$). Thus, as shown in Fig. 2b, the classification rate in SHAM differed for SO_{trough} (Z = 2.55, p = .011) and $\mathrm{SO}_{\mathrm{Vpp}}$ (Z = 2.77, p = .006), but not $\mathrm{SA}_{amplitude}$ (Z = 1.34, p = .179). We also found a difference in classification rate when both classifiers were applied to the STIM dataset (i.e., $f_{\mathrm{SH}}(\theta_{ST})$ vs $f_{\mathrm{ST}}(\theta_{ST})$), Fig. 3c. Thus, f_{ST} outperforms f_{SH} when predicting the LOW/HIGH levels of SO_{trough} in the stimulation condition (Z = -2.70, p = .007). However, there was no difference in classification performance for $\mathrm{SO}_{\mathrm{Vpp}}$ (Z = -0.54, p = .586), or $\mathrm{SA}_{amplitude}$ (Z = -0.57, p = .570).

These crossover results suggests that the classifiers (SO_{trough}; $\mathrm{SO}_{\mathrm{Vpp}}$; $\mathrm{SA}_{amplitude}$) differ in how they obtain the information from the data to predict each evaluated measure. Primarily, because we evaluated the performance within condition and found no difference between classifiers, we know that they have similar performance in their respective condition (Fig. 2a). This means that any differences between conditions cannot be attributed to imbalance in the performance of *within* classification. Similarly, looking at the spindle measure, the performance for predicting $\mathrm{SA}_{amplitude}$ is similar for both classifiers in the crossover analysis (Fig. 2b & c). This indicates that there is not enough information for the f_{ST} classifier to predict $\mathrm{SA}_{amplitude}$ responses beyond the information provided by the spontaneous ongoing SO features in the STIM dataset. Turning to the SO measures (SO_{trough} and $\mathrm{SO}_{\mathrm{Vpp}}$), the SHAM classifier f_{SH} outperformed the stimulation classifier f_{ST} on both measures in the SHAM dataset θ_{SH}, (Fig. 2B), but the STIM classifier f_{ST} only outperformed the SHAM classifier f_{SH} for SO_{trough} in the STIM dataset θ_{ST} vs θ_{SH} (Fig. 2c). The fact that the STIM classifier did not outperform the SHAM classifier f_{SH} for SO_{Vpp} in the STIM dataset may indicate that the auditory click does not significantly alter the positive SO amplitude, increasing the complexity of the $\mathrm{SO}_{\mathrm{Vpp}}$ detection. However, the fact that the STIM classifier outperformed the SHAM classifier on SO_{trough} indicates that the pre-stimulus information used to predict SO_{trough} is dif-

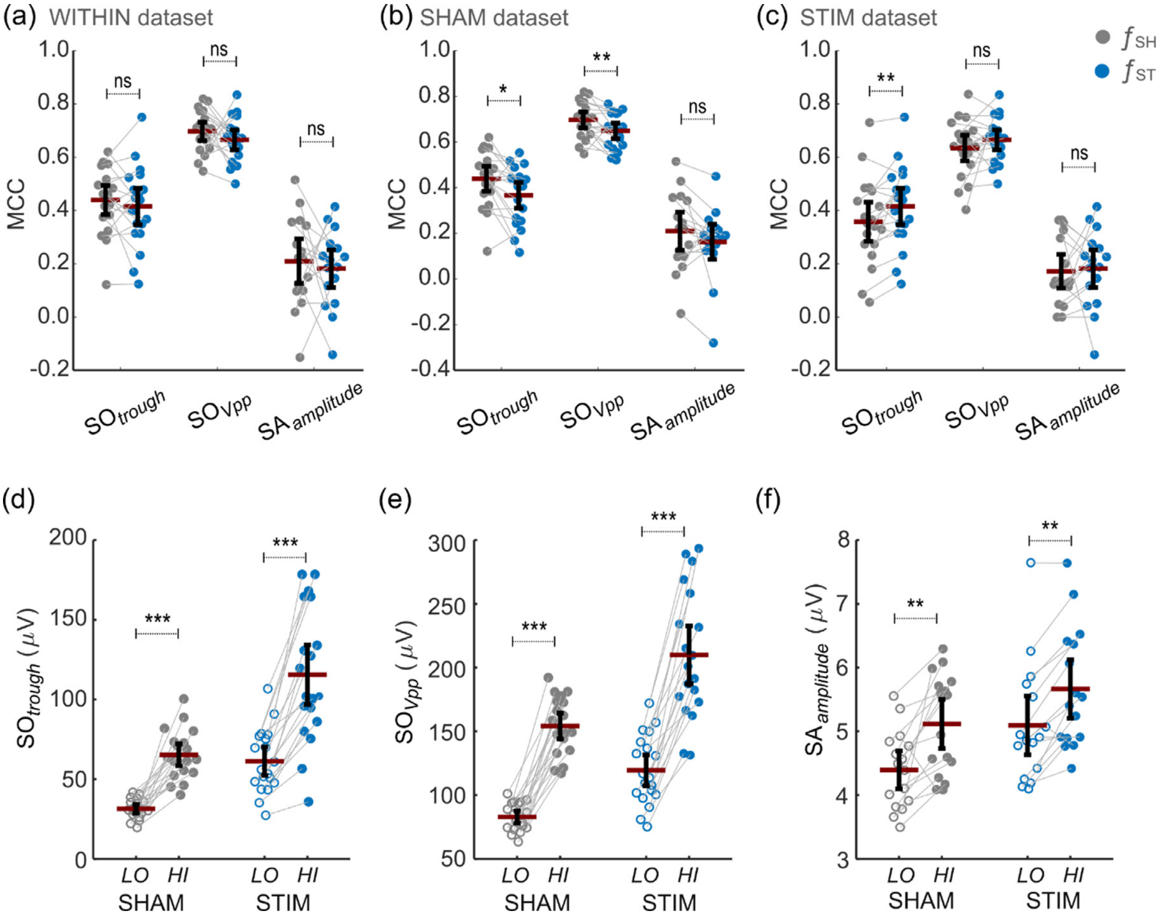

Fig. 2. Within and cross-classification performance of trained models and average LOW and HIGH values for post-classification labels. (a) Classifier differences in MCC performance for within dataset tests. (b) Classifier differences in MCC performance when SHAM classifier (f_{SH}) and STIM classifier (f_{ST}) were tested in the SHAM dataset. (c) Classifier differences in MCC performance when f_{SH} and f_{ST} were tested in the STIM dataset. (d) Average amplitude of SO_{trough} for trials classified as LOW and HIGH in SHAM and STIM datasets. We can notice a larger increment of SO amplitudes in those trials marked as HIGH. The same effect was observed for SO_{Vpp} (e) and $SA_{amplitude}$ (f). (*) for $p < .05$, (**) for $p < .01$, (**) for $p < .001$, (n.s) for not significant. All error bars represent mean ± 95%CI.

ferent for the two classifiers. This suggests that the stimulation strongly influences the post-stimulus SO trough, since that trough is no longer predicted by the same pre-stimulus features as it would have been if no click had occurred.

To further corroborate our results, we compared the trained classifiers in each dataset using a more robust method based on their classification errors. Hence, we applied a non-parametric McNemar test which compares the incorrectly labelled trials generated by each classifier. Under the null hypothesis, the two classifiers should have the same error rate, so for instance, the f_{SH} and f_{ST} classifiers would not be different (Thomas, 1998). After FDR correction, our McNemar's tests showed that the error rate of SO measures differed significantly between f_{SH} and f_{ST}. This was the case in both the SHAM dataset $f_{SH}(\theta_{SH})$ vs $f_{ST}(\theta_{SH})$: SO_{trough} ($p < .001$), SO_{Vpp} ($p < .001$); and the STIM dataset $f_{SH}(\theta_{ST})$ vs $f_{ST}(\theta_{ST})$: SO_{trough} ($p = .005$); SO_{Vpp}, ($p < .001$). Furthermore, we did not find differences between classifiers for spindle activity when both the f_{SH} and f_{ST} were applied in either SHAM ($SA_{amplitude}$, $p = .308$) or STIM datasets ($SA_{amplitude}$, $p = .138$). Overall, these results are perfectly in keeping with the MCC analysis presented above.

Thus, both the analysis of differences in MCC performance and the analysis of differences in the error rate confirm that training classifiers in SHAM and STIM datasets results in two models with different feature mapping (Thomas, 1998). This is particularly true for the SO measures (SO_{trough} and SO_{Vpp}) which learn different dynamics from the evaluated datasets imposed by the auditory stimulus. The lack of forecasting of stimulus-related changes for spindle activity might indicate that the EEG information that predicts these effects is mainly associated with post-stimulus processes.

3.3. Generalization of classifier predictions

Next, we wanted to check how accurately our classifiers had predicted whether subsequent SOs would fit the HIGH or LOW amplitude classes by examining the actual amplitudes which occurred. Therefore, we evaluated the response of the classification for all trials. These included trials labelled trials within LOW and HIGH groups as well as unlabelled trials falling between the [LOW, HIGH] interval thresholds, which were initially discarded during training of classifiers. The amplitude of post-stimulus responses was analyzed with a 2 (Condition: SHAM vs STIM) x 2 (Classification Label: LOW vs HIGH) ANOVA as seen in Fig. 2d–f. This showed a main effect of classification label for SO_{trough} (F(73,1) = 59.37, $p < .001$), SO_{Vpp} (F(73,1) = 126.03, $p < .001$) and $SA_{amplitude}$ (F(61,1) = 9.86, $p = .003$). In all conditions, a Tukey post hoc test revealed that HIGH labels identified higher amplitude events for SO_{trough} (SHAM: t(18.0) = -9.1, $p < .001$; Cohen's d = 2.88, 95%CI [1.95, 3.82]; STIM: t(18.0) = -8.7, $p < .001$; Cohen's d = 1.66, 95%CI [0.90, 2.42], Fig. 2d), as well as for SO_{Vpp} (SHAM: t(18.0) = -18.3, $p < .001$; Cohen's d = 4.04, 95%CI [2.90, 5.18]. STIM: t(18.0) = -12.3, $p < .001$; Cohen's d = 2.23, 95%CI [1.40, 3.07], Fig. 2e) and $SA_{amplitude}$ (SHAM: t(15.0) = -3.0, $p = .008$; Cohen's d = 1.03, 95%CI [0.27, 1.79]; STIM: t(15.0) = -3.3, $p = .005$; Cohen's d = 0.61, 95%CI [-0.13, 1.34], Fig. 2f).

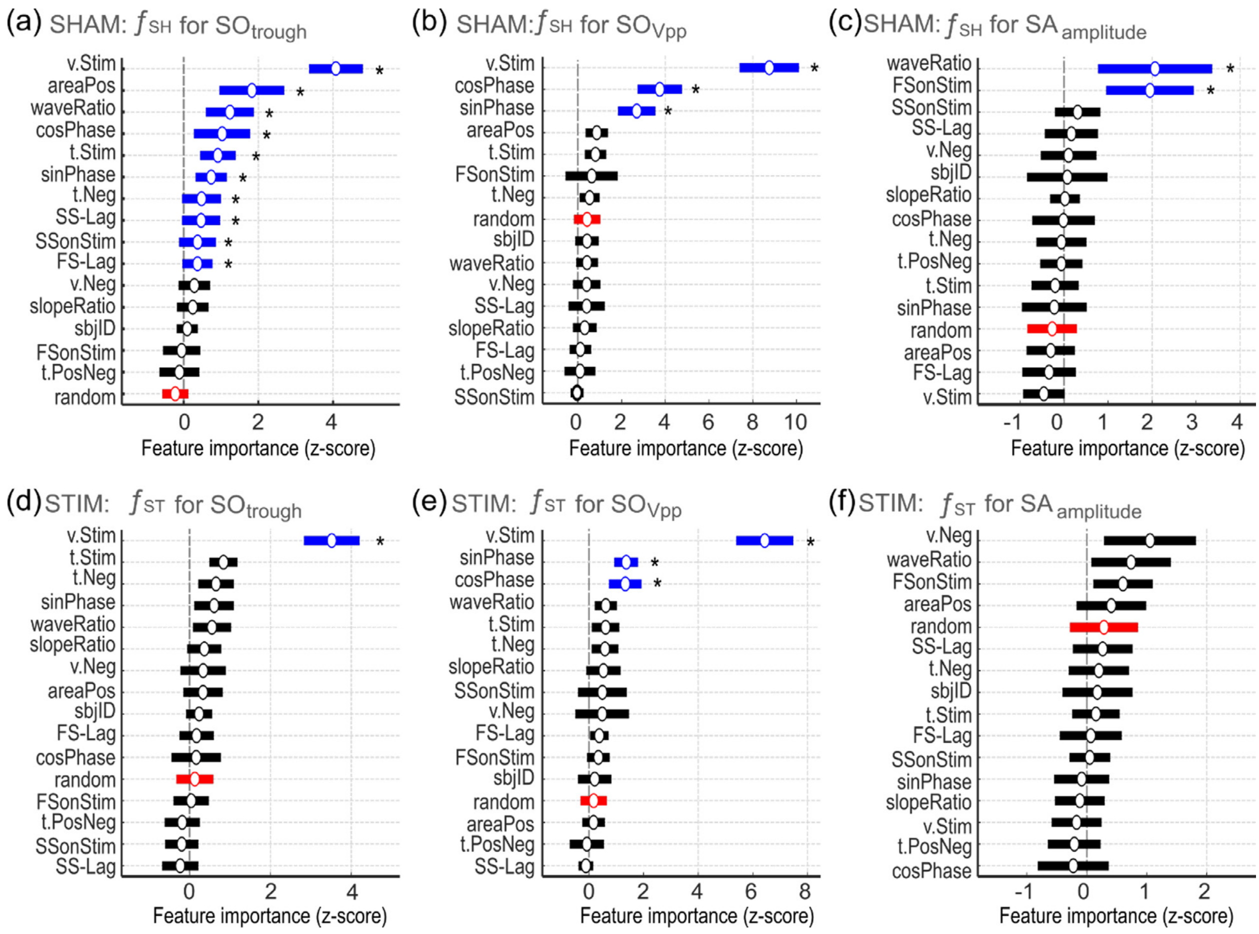

Fig. 3. Feature importance of SHAM and STIM classifiers. (a) Feature importance by permutation for classification of SO_{trough} in the SHAM dataset. (b) Feature importance for classification of SO_{Vpp} in the SHAM dataset. (c) Feature importance for classification of $SA_{amplitude}$ in the SHAM dataset. (d) Feature importance for classification of SO_{trough} in the STIM dataset. (e) Feature importance for classification of SO_{Vpp} in the STIM dataset. (f) Feature importance by permutation (*FI*) for classification of $SA_{amplitude}$ in the STIM dataset. Vertical dashed lines at *z*-score 0 indicate the level of random variability for FI. Feature importance values are provided as the *z*-score of the permuted MCC performance, and bars in blue indicate those features that are significantly larger than the random dummy feature (in red). (*) for corrected $p < .05$. All error bars represent mean ± 95%CI.

These differences were corroborated when we computed only labelled as well as only unlabelled trials (Fig. S1). Hence, for labelled trials we found differences for SO_{trough} (SHAM: Cohen's d = 2.16, 95%CI [1.34, 2.99], $p < .001$; STIM: Cohen's d = 1.56, 95%CI [0.82, 2.31], $p < .001$), as well as for SO_{Vpp} (SHAM: Cohen's d = 6.04, 95%CI [4.50, 7.58], $p < .001$; STIM: Cohen's d = 3.05, 95%CI [2.09, 4.01], $p < .001$) and $SA_{amplitude}$ (SHAM: Cohen's d = 1.27, 95%CI [0.49, 2.06], $p = .001$; Cohen's d = 0.65, 95%CI [-0.08, 1.39], $p = .004$). Likewise, for unlabelled trials we found differences for SO_{trough} (SHAM: Cohen's d = 0.83, 95%CI [0.15, 1.52], $p < .001$; STIM: Cohen's d = 0.41, 95%CI [-0.25, 1.07], $p < .001$), as well as for SO_{Vpp} (SHAM: Cohen's d = 1.14, 95%CI [0.43, 1.84], $p < .001$; STIM: Cohen's d = 0.67, 95%CI [0.00, 1.34], $p < .001$) and $SA_{amplitude}$ (SHAM: Cohen's d = 0.35, 95%CI [-0.37, 1.08], $p = .008$; Cohen's d = 0.30, 95%CI [-0.42, 1.02], $p = .005$).

As the classifiers did not see any pre-stimulus dynamics for events between [LOW, HIGH] thresholds during training, the classification of such trials could potentially be random. If that were the case, no difference in mean amplitude for trials classifies as LOW and HIGH would have been evident. The fact that such differences appeared even when all trials are considered confirms that the classifiers were able to generalize to the novel trials. We thus conclude that the classifier accurately predicted whether SOs and sleeep spindles will have comparatively large or small amplitudes when it categorises them into LOW or HIGH classes

3.4. Feature importance of evaluated classifiers

The feature importance (FI), as determined for f_{SH} and f_{ST} classifiers, showed how strongly each pre-stimulus feature weight was when the model was applied to unseen data (holdout subject). Firstly, for each condition and measure, we applied a one-way ANOVA to test whether features differed in importance. Secondly, we identified the features with FI larger than the FI estimated for a dummy random feature using Tukey's honesty test.

Interestingly, our analysis revealed that permuting features arising from the wave structure and estimated click timing ('sham-click') in spontaneous SO dynamics (SHAM condition) caused significant changes in the *z*-score of MCC (Fig. 3a and b). We found that the *z*-scored FI differed between features in the SHAM classifier for SO_{trough} (one-way ANOVA, F(15,280) = 14.27, $p < .001$, Fig. 3a). Tukey post hoc tests revealed that several features for which FI was significantly higher than the FI for the dummy random feature (-0.22, 95%CI: -0.57 to 0.12). These included features from the estimated sham-click (*v.Stim* (4.07, 95%CI: 3.35 to 4.79; $p < .001$), *cosPhase* (1.03, 95%CI: 0.28 to 1.78; $p = .008$), *t.Stim* (0.92, 95%CI: 0.44 to 1.39; $p = .001$), *sinPhase* (0.74, 95%CI: 0.32 to 1.16; $p = .002$)) as well as features of the slow-wave structure (*areaPos* (1.83, 95%CI: 0.96 to 2.69; $p = .001$), *waveRatio* (1.24, 95%CI: 0.60 to 1.88; $p = .001$), *t.Neg* (0.48, 95%CI: -0.04 to 1.00; $p = .030$), *SS-Lag* (0.47, 95%CI: -0.04 to 0.98; $p = .030$), *SSonStim* (0.37, 95%CI: -0.12 to 0.87; p = .047) and *FS-Lag* (0.37, 95%CI: -0.03 to 0.77; p = .030)).

We also found differences of FI between the various features in the SHAM classifier for SO_{Vpp} (one-way ANOVA, F(15,225) = 36.18, p < .001, Fig. 3b). A Tukey post hoc test revealed that FI for *v.Stim* (8.75, 95%CI: 7.39 to 10.11; p = .001), *cosPhase* (3.74, 95%CI: 2.72 to 4.76; p < .001), and *sinPhase* (2.69, 95%CI: 1.84 to 3.54; p < .001) were higher than chance and larger than the dummy random feature (0.42, 95%CI: -0.18 to 1.03).

Similarly, we found differences of FI between features in the SHAM classifier for $SA_{amplitude}$ (one-way ANOVA, F(15,240) = 4.40, p < .001, Fig. 3c). A Tukey post hoc test revealed that *waveRatio* (2.07, 95%CI: 0.77 to 3.36; p = .015) and *FSonStim*(1.95, 95%CI: 0.96 to 2.94; p = .006) were higher than chance and larger than the random dummy feature (-0.27, 95%CI: -0.83 to 0.29).

Interestingly, three features that depend entirely upon the sham-click time (*v.Stim, cosPhase, sinPhase*) have consistently high FI in SHAM. Indeed, the placement of the Sham-click after the detected SO trough is non-random, and the above features were based on this time. Therefore, these features may strongly predict the magnitude of the next oscillation. Specifically, *v.Stim* (amplitude at sham-click time) is highly correlated with the peak amplitude of the sham-stimulated SO (SHAM dataset: r = 0.932, p < .001). Similarly, *cosPhase* (the cosine of the phase at sham-click time) correlates with the negative-to-positive slope of the SO (SHAM dataset: r = 0.207, p < .001) and the area of the negative SO deflection before the sham-click (SHAM dataset: r = 0.122, p < .001). Likewise, *sinPhase* correlates with the negative-to-positive slope of the SO before the sham-click (SHAM dataset: r = 0.123, p < .001). Overall, these results show that features that seem related only to the timing of the sham-click (i.e. *v.Stim, cosPhase, sinPhase*) also predict characteristics of the SO structure.

In contrast to SHAM, in the STIM dataset the feature importance of only a few variables were significantly higher than the feature importance of the dummy random feature. Thus, FI differed between features for the STIM classifier on SO_{trough} (one-way ANOVA, F(15,278) = 12.943, p < .001, Fig. 3d). The Tukey post hoc test revealed that *v.Stim* (3.52, 95%CI: 2.83 to 4.20; p < .001) was the only feature higher than the dummy *random* variable (0.13, 95%CI: -0.33 to 0.59). Other click-related FI such as for *tStim* and *sinPhase* were also higher than chance, but they were not statistically different from the dummy *random* variable.

As with the SHAM dataset, the FI differed between features for the STIM classifier on SO_{Vpp} (one-way ANOVA, F(15,241) = 27.958, p < .001, Fig. 3e). The Tukey post hoc test revealed *v.Stim* (6.44, 95%CI: 5.39 to 7.48; p < .001), *sinPhase* (1.37, 95%CI: 0.94 to 1.80; p = .004), and *cosPhase* (1.33, 95%CI: 0.73 to 1.93; p = .014) were higher than the *random* dummy feature (0.17, 95%CI: -0.30 to 0.65).

For $SA_{amplitude}$, although there was a significant difference between groups as determined by a one-way ANOVA (F(15,239) = 1.805, p = .035, Fig. 3f), there were no significant differences between the FI of the *random* dummy variable and the remainder of features.

As each of the models was trained on data from different subjects, this could lead to cases where testing features could have values outside the range of the training features. This would cause a model drift usually known as covariate shift (Shimodaira, 2000). This effect is usually solved using variable transformation or batch normalization prior to training and classification (Kawakita and Takeuchi, 2014; Sugiyama et al., 2007), but this is hard to implement in online acquisition and difficult to interpret in the physiological context. Therefore, we used the feature values as computed and evaluated differences in distribution between training and testing sets (Table S4). We found no differences between training and testing feature distributions across subjects for within condition experiments ($f_{SH}(\theta_{SH})$ and $f_{ST}(\theta_{ST})$). Similarly, we found no differences between training and testing for most of the features for between condition experiments ($f_{SH}(\theta_{ST})$ and $f_{ST}(\theta_{SH})$). However, we found only one significant effect of covariate shift for *v.Neg* when evaluating the SO_{Vpp} response in between condition experiments (p < .01 FDR corrected). Interestingly, *v.Neg* was not highlighted by our analyses as a variable with high FI although this is commonly used as the main characteristic for SO detection for CLAS (using negative wave thresholding).

We also compared the FI weights between conditions. As seen in the Fig. S2, we only found differences in *areaPos* when evaluating the SO_{trough} ($\Delta FI_{(ST-SH)}$ = -1.50, 95%CI: 0.91 to 2.95; p = .047) as well as for the top three features in SO_{Vpp} (*v.Stim* $\Delta FI_{(ST-SH)}$ = -2.32, 95%CI: 1.04 to 2.65; p = .030; *sinPhase* $\Delta FI_{(ST-SH)}$ = -1.33, 95%CI: 0.89 to 2.73; p = .030; *cosPhase* $\Delta FI_{(ST-SH)}$ = -2.41, 95%CI: 0.98 to 4.00; p = .003). However, as we were interested in evaluating the "importance" of the features for prediction, we evaluated whether the FI ranks are more similar within conditions but more dissimilar between conditions. For this, using 1600 Monte Carlo permutations, we divided STIM and SHAM datasets in half subjects and compared FI ranks between the four groups (sham group 1 (SH-G1), sham group 2 (SH-G2), stim group 1 (ST-G1), and stim group 2 (ST-G2)). A Kruskall-Wallis ANOVA test rejected the null hypothesis that the sample data from each group comes from the same distribution (Table S5) and post hoc analyses indicated significant differences of FI ranks between but not within conditions (Fig. S2). Therefore, these analyses further suggest that the rank of the heuristic FI are more correlated within datasets than between datasets.

3.5. Similarity of feature importance between SHAM and STIM models

Finally, we wanted to determine whether the features in SHAM and STIM datasets are evaluated in the same way by f_{SH} and f_{ST} classifiers. We therefore studied how feature permutation changed the classifier performance within and across conditions. Considering that dynamics learnt by f_{SH} are based only on the drift of SWA in the SHAM dataset, features evaluated similarly by both classifiers may indicate that these are mostly related to spontaneous EEG activity. From a mathematical perspective, when applied to the same dataset, f_{SH} and f_{ST} classifiers can be considered as vectorial functions in the same vectorial space. The predictor weights estimated as values of feature importance thus represent the vector space of random variables in the evaluated dataset. Features with correlated feature importance in both f_{SH} and f_{ST} classifiers are therefore not independent in the multidimensional feature space, suggesting that these predictors may describe similar processes in SHAM and STIM conditions (Table S1). The correlation of cross-condition changes in performance in the SHAM dataset during the crossover analysis were defined by $\rho_{SH,ST}(\theta_{SH}) = \mathrm{corr}(FI_{SH}, FI_{SHx})$, interpreted as the correlations of heuristic FI computed from f_{SH} and f_{ST} classifiers on the SHAM dataset. The correlations of cross-condition changes in performance in the SHAM dataset were defined by $\rho_{SH,ST}(\theta_{ST}) = \mathrm{corr}(FI_{STx}, FI_{ST})$, interpreted as the correlations of heuristic FI computed from f_{SH} and f_{ST} classifiers on the STIM dataset.

In Fig. 4 we show the cross-correlogram of FI for within vs cross-condition when the classifiers were applied in SHAM and STIM datasets. Our data suggest that the amplitude of the stimulated SO (*v.Stim*) is similar in importance for predicting SO activity in both datasets. After FDR correction, we found the correlations of feature importance were significant for *v.Stim* (within condition) vs *v.Stim* (cross condition) evaluated under SO_{trough} (*v.Stim*: $\rho_{SH,ST}(\theta_{SH})$ = 0.92, p < .001 and $\rho_{SH,ST}(\theta_{ST})$ = 0.88, p < .001, Fig. 4a and b) and SO_{Vpp} (*v.Stim*: $\rho_{SH,ST}(\theta_{SH})$ = 0.92, p < .001 and $\rho_{SH,ST}(\theta_{ST})$ = 0.88, p < .001, not shown). This suggests that data dynamics mapped by this feature in both models may describe similar information in SHAM and STIM datasets. Hence, the *v.Stim* predictor is not an independent variable between spontaneous and stimulation conditions. This shared dependence between conditions suggests that the classifier effect of this feature on the amplitude of the SO outcome is mainly generated from spontaneous neural dynamics rather than dynamics associated with the stimulation.

Similarly, for $SA_{amplitude}$ we found that correlations between feature importance from within and cross condition classifiers were significant

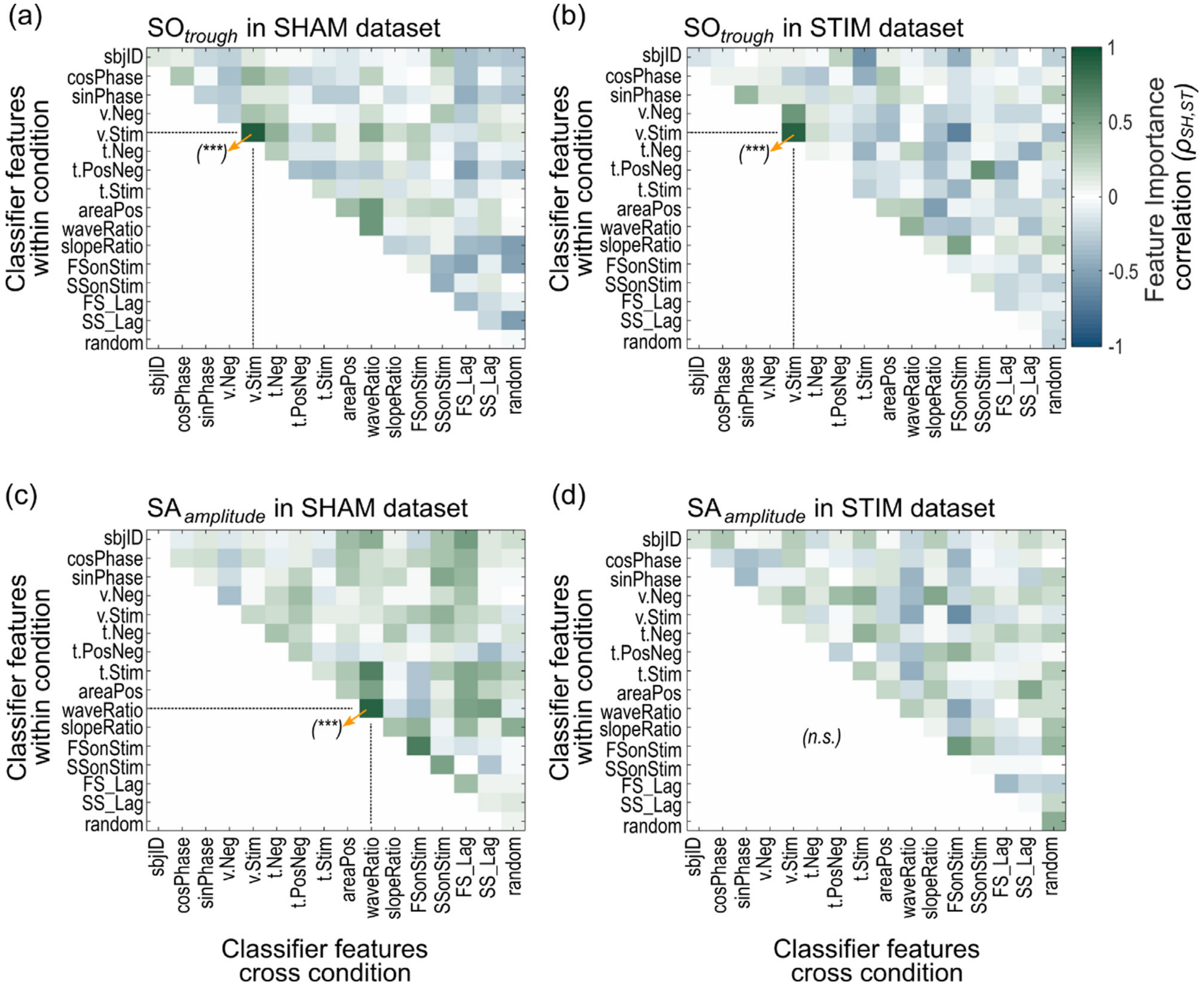

Fig. 4. Cross-correlogram for changes in feature importance between within and cross classification models applied to SHAM ($\rho_{\mathrm{SH,ST}}(\theta_{\mathrm{SH}})$) or STIM ($\rho_{\mathrm{SH,ST}}(\theta_{\mathrm{ST}})$) datasets. (a, b) Correlations of feature importance between classifiers applied to SHAM (a) and STIM (b) datasets evaluated on SO$_{trough.}$ (c-d) Correlations of feature importance between classifiers applied to SHAM (c) and STIM (d) datasets evaluated on SA$_{amplitude.}$ Significant correlations were highlighted by an arrow and (***) for p < .001 after FDR correction, (n.s) stands for not significant.

for correlation of *waveRatio* (within condition) with *waveRatio* (cross condition) in SHAM ($\rho_{\mathrm{SH,ST}}(\theta_{SH}) = 0.90$, $p < .001$, Fig. 4c), but this was not true in the STIM dataset (Fig. 4d). This further suggests that the auditory click disturbs the pre-stimulus slow-wave dynamics that predict SA$_{amplitude}$. Specifically for the SHAM dataset, we found a high correlation of feature importance between within and cross condition classifiers for *waveRatio* vs *t.Stim* ($\rho_{\mathrm{SH,ST}}(\theta_{\mathrm{SH}}) = 0.72$, non-corrected $p = .011$) and *FSonStim* vs *FSonStim* ($\rho_{\mathrm{SH,ST}}(\theta_{\mathrm{SH}}) = 0.75$, non-corrected $p = .006$) but these did not survive after FDR correction. This further suggests that information from the shape of the SO wave predicts the post-click spontaneous spindle activity.

Further electrophysiological analysis of classified trials shows how predicted SO amplitudes (HIGH and LOW classes) mediate the post-click SO or spindle ERP amplitudes in SHAM and STIM datasets. Fig. 5a depicts the average ERP for trials predicted as HIGH vs trials predicted as LOW for SO$_{trough}$ in the SHAM dataset. Note the apparent increase of SO trough amplitude before the estimated time of stimulation. This could suggest that this trough amplitude (*v.Neg*) is a main predictor for spontaneous SO$_{trough}$. However, *v.Neg* was not identified as a top predictor by feature importance as indicated previously in Fig. 3a. Hence, bigger troughs do not predict a bigger post-click SOs, but rather the apparent increasing of trough amplitude may correspond to averaging the structure of the negative deflection of SOs, represented by the most important features (*areaPos, waveRatio, cosPhase, t.Stim, sinPhase* as in Fig. 3a and Section 3.3). As expected, the amplitude of the wave at the estimated click (*v.Stim*) shows larger positive deflections for trials predicted as HIGH SO$_{trough}$.

Similarly, Fig. 5c depicts the averages of trials predicted as HIGH outcome vs trials predicted as LOW for SO$_{trough}$ in the STIM dataset. Unlike predictions in SHAM, in STIM the structure of the pre-stimulus SO did not predict changes in the averaged SO between HIGH and LOW trials. Only the SO amplitude during the click (*v.Stim*) provided a representative feature for SO$_{trough}$ prediction.

Conversely, EEG trials representing LOW and HIGH SA$_{amplitude}$ were not different for either SHAM (Fig. 5b) or STIM datasets (Fig. 5d). These findings are in keeping with reduced importance of SO amplitudes for the prediction of post-event SA$_{amplitude}$ in both conditions.

Importantly, features should not simply be considered in isolation since they interact together in terms of their impact on the subsequent oscillatory structure. To look at this, we performed a more detailed analysis of the combined contribution of the highest ranked features in the prediction of LOW/HIGH for SO$_{trough}$ outcome (Fig. 5e). First, to distinguish between the feature distributions, we calculated the difference between normalized event distributions (HIGH - LOW) for SHAM and STIM

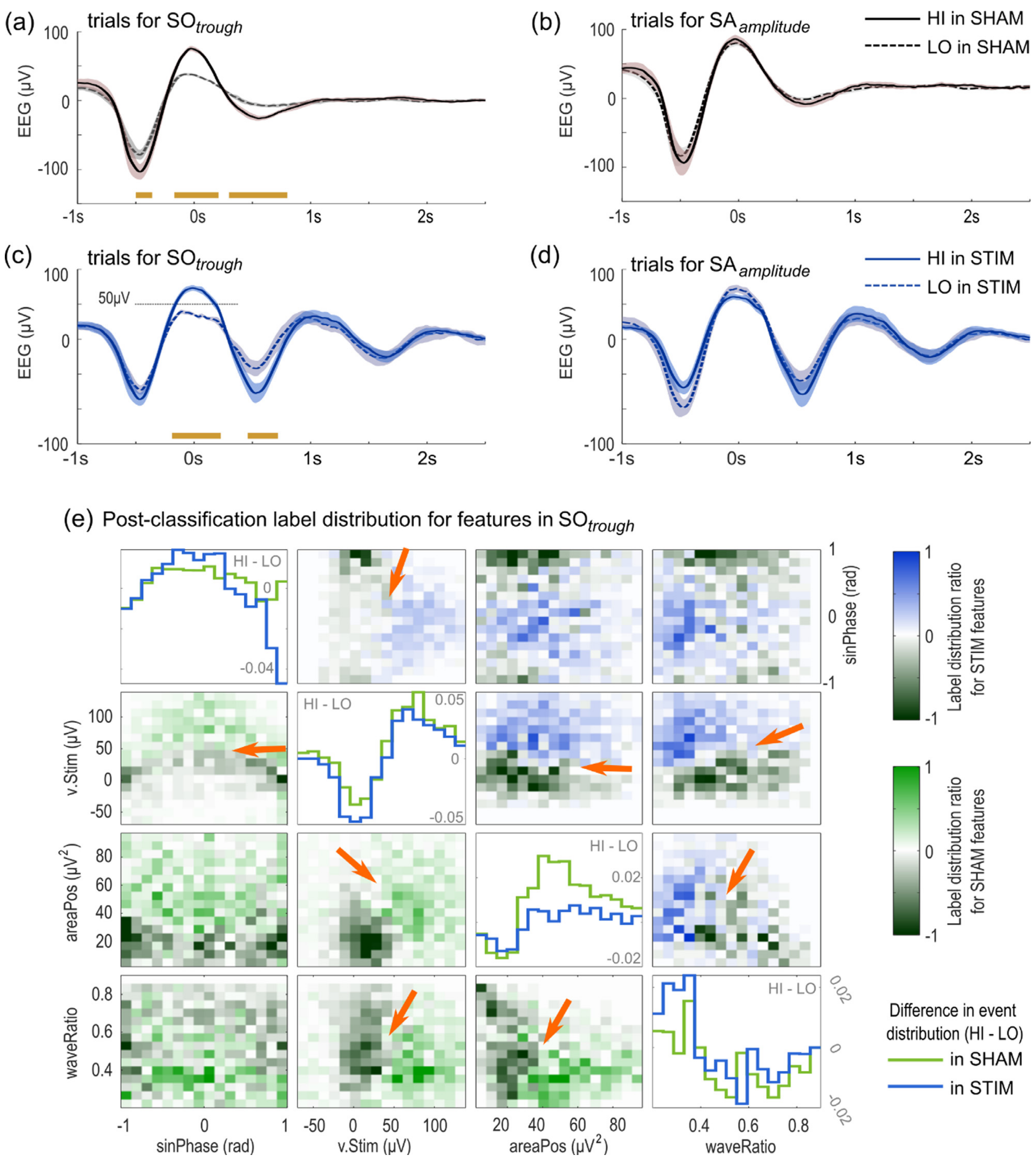

Fig. 5. Wave amplitude and feature importance for post-classification labels. (a) Post-classification HIGH (HI) vs LOW (LO) EEG average trials locked to the estimated click for SO_{trough} classification evaluated in the SHAM dataset. (b) Post-classification HI vs LO EEG average trials locked to the estimated click for $SA_{amplitude}$ classification evaluated in the SHAM dataset. (c) Post-classification HI vs LO EEG average trials locked to the applied click for SO_{trough} classification evaluated in the STIM dataset. (d) Post-classification HI vs LO EEG average trials locked to the applied click for $SA_{amplitude}$ classification evaluated in the STIM dataset. Orange bars indicate clusters of significant differences between classified LO and HI trials ($p < .05$) after FDR correction. Error areas represent mean ± 95%CI. (e) Pair plots for some of the top features by FI for SO_{trough} indicating the difference of normalized histograms (HIGH - LOW) for SHAM (lower diagonal) and STIM (upper diagonal) datasets. Plots in the diagonal shows the histogram difference between selected features. Positive values indicate larger concentration of HIGH events whereas negative values indicate a larger concentration of LOW values. Orange arrows indicate apparent boundaries between HIGH and LOW events for STIM (black-to-blue) and SHAM (black-to-green) conditions. v.Stim: SO voltage during stimulation; sinPhase: sinus of phase of estimated/applied stimulation; areaPos: Area under curve for the peak of the SO wave before the click, and waveRatio: Duration ratio for the wave before click.

datasets. The resultant distribution indicates that large *v.Stim* predict larger subsequent SO_{trough} amplitudes. Notably, trials having a *v.Stim* > 50 µV are associated with a larger concentration of HIGH events for both STIM and SHAM (Fig. 5e). Interestingly, this 50µV threshold is also evident in the averaged SWA in which ERPs were locked to the time of stimulation for HIGH and LOW trials in SHAM (Fig. 5a) and STIM (Fig. 5c) conditions. Furthermore, the features from the wave structure (*areaPos* and *waveRatio*) also hint at larger effects of these dynamics in the distribution of LOW and HIGH events in both datasets (orange arrows in Fig. 5e).

4. Discussion

We were able to accurately predict whether application of an auditory click to a SO will lead to a large enhancement of the subsequent SO. This was achieved by using features of an ongoing slow oscillation in a machine learning classifier. Thus, we were able to predict the categories (LOW/HIGH) of the trough amplitude (SO_{trough}), the peak-to-trough amplitude of the SO (SO_{Vpp}), and the maximal amplitude of the spindle activity locked to the subsequent positive SO ($SA_{amplitude}$) after each click-time in both STIM and SHAM conditions by analysing oscillation features from before the click or sham-click. We also correctly predicted whether either spontaneous oscillatory activity or a phase-locked sound would lead to a large neural response (classification accuracy > 67% for SO amplitude and > 55% for spindle activity) in both STIM and SHAM conditions. Finally, we performed a feature importance analysis to identify the most important EEG features for prediction of SOs and spindles. To this end, we permuted features to assess how they influenced the generalization of classifiers to unseen trials.

4.1. Classification using pre-stim SO wave features differentiates SO click-response in SHAM and STIM

Although our classifiers performed similarly in SHAM and STIM datasets, performance was decreased in cross-classification for both datasets. This is likely due to disruption in the ongoing EEG pattern caused by the auditory click, since the click is the only thing that distinguishes the two conditions. This pattern revealed that STIM and SHAM classifiers represent non-dependent non-orthogonal mathematical models that derive different information from the ongoing signal in these two conditions, at least when classifying for SO_{trough} and SO_{Vpp}. These results demonstrate that the EEG features prior to the presentation of the sound stimulus carry information about the magnitude of the subsequent brain response following the sound. Conversely, we found no differences between classifiers for SHAM and STIM when predicting $SA_{amplitude}$, possibly indicating that pre-stimulus information is not very helpful when predicting this measure in the STIM dataset.

4.2. Spontaneous pre-stim features predict post-stimulus SHAM and STIM spindle dynamics

Overall, we found no differences in the model prediction of spindle activity amplitude between SHAM and STIM models. Previous studies indicated a boosting effect of CLAS on spindle activity. This enhancing effect is specifically related to an increase in amplitude of spindle activity locked to the SO directly after the click (Ngo et al., 2013; Schneider et al., 2020). We found that the presence of spindle events in the ongoing wave and the SO structure were the main variables predicting spindle amplitude. Furthermore, we found evidence that CLAS disturbed the spontaneous relationship between post-stimulus spindle activity and ongoing SO dynamics. However, the ongoing SWA appears to retain the same level of information for SHAM and STIM classifiers to predict the amplitude of post-stimulus spindle activity. Consistent with previous literature (Antony et al., 2019; Ngo et al., 2015), we found that changes in spindle related activity (e.g. FSonStim) were better predicted by refractory periods of spindle rebound. Our results further support the idea that responses to external stimuli such as clicks diverge for innate SWA because of the distinctive thalamocortical dynamics of SO and spindles (Navarrete et al., 2020a). Hence, the pre-stimulus conditions evaluated here in both STIM and SHAM models only conveyed information relating to spontaneous spindle dynamics, and information relevant to the spindle response to CLAS might be mainly concealed within post-stimulus SWS dynamics.

4.3. Feature importance in classification of the SHAM dataset

Analysis of feature importance for the SHAM classifier allowed us to determine which features best characterize the level of cortical activity of the SO cycle. We found that variables relating to the timed structure of the SO were important for predicting spontaneous SO amplitudes. Indeed, because the timing of stimulation depends on the wave detection algorithm (Navarrete et al., 2020a; Ngo et al., 2015), the stimulation-related variables were given high feature importance accordingly. Our results thus suggest that sham-click stimulation related features in the SHAM dataset also convey information about subsequent spontaneous cortical activation. Specifically, the voltage during the sham-click (*v.Stim*) may indirectly reveal the level of spontaneous neural depolarization, which is then measured via scalp amplitude (Crunelli et al., 2018; Siclari et al., 2014). Meanwhile, global neural synchrony may be determined by the timing and phase of the estimated stimulus (*cosPhase, sinPhase* and *t.Stim*) along with the information about SO wave structure (*areaPos, waveRatio, slopeRatio*, etc) (Riedner et al., 2007). Therefore, the features we identified as important for the prediction of spontaneous SO activity (amplitude: *v.Stim*; and SO structure: *areaPos, waveRatio, t.Neg, cosPhase, sinPhase*) may also index the strength of thalamocortical drive on individual SOs.

Following this, we hypothesize that the prediction of spontaneous SO activity by the SHAM classifier may help to discriminate thalamocortical from cortico-cortical SOs. Recent studies proposed that synchronizing factors of the SO may assist in discerning different thalamic and cortical mechanisms involving the dynamics of the ongoing SWA (Bernardi et al., 2018; Siclari et al., 2014). In this direction, thalamocortical circuits trigger Up-states on each SO wave, contributing to cortical synchronization across the cortex and driving the structure of the SO cycle (Amzica and Steriade, 1995; Crunelli et al., 2018) and thus may be indicated by those features describing the SO structure. Likewise, the high importance of the scalp amplitude predictor (*v.Stim*) agrees with previous work indicating that increased levels of cortical depolarization may lead to large synchronous hyperpolarizations (Neske, 2016), and therefore this could highlight intrinsic characteristics of the thalamocortical dynamics (Siclari et al., 2014). Nevertheless, this hypothesis should be tested by adding a larger spatial sample and further intracortical recordings.

4.4. Feature importance in the STIM classifier indexes SO dynamics modulating response to the auditory click

Our feature importance analyses for the STIM classifier indicated that the main variable predicting the post-stimulus SO-amplitude outcome is the wave amplitude during stimulation. This result follows previous work which has suggested that the peak amplitude is the optimal timing for the auditory click to enhance the ongoing waves (Ngo et al., 2015, 2013). Our findings build on this by indicating that considering the stimulus phase alone is not enough to optimally enhance the SO. Instead, the wave amplitude during stimulation better predicts the outcome of the acoustic click. Consistent with this, we recently demonstrated that the window of opportunity for the click to effectively enhance SOs is limited to a wide phase interval around the wave peak (Navarrete et al., 2020a), challenging the hypothesis of a particular SO phase as the optimal target for SO enhancement (Santostasi et al., 2015). Neglecting the importance of wave amplitude could explain the failure of many auditory clicks to boost SO memory consolidation function while still increasing the post-stimulus response (Henin et al., 2019). In the current report, we show that the average amplitude of post-stimulus trials increases considerably when our classification method is used to select the SO cycles to stimulate (Cohen's d > 1.6 for SO amplitude and Cohen's d > 0.6 for spindle activity). We therefore suggest that future detection algorithms may consider inclusion of decision rules that evaluate amplitude characteristics to boost the effectiveness of the acoustic click for enhancing the SO response.

Our results also suggest that auditory clicks during elevated neural excitability within SO peaks set the arousal threshold which contributes to the cortical synchronization process of slow-waves. The predictor importance analyses for the STIM classifier showed that stimulation disturbs the effect of the SO wave structure on the prediction of SO troughs. However, wave amplitude during the click still drives the click outcome as it does in the spontaneous generation of large SO troughs in SHAM. Consistent with this, previous studies have indicated that cortical excitability is maximal during periods of neural activation within SO peaks (Massimini, 2002; Rosanova and Timofeev, 2005). Likewise, it has been proposed that the diffusely projected "matrix" thalamic system imposes a sensory threshold upon the arousal system, allowing auditory stimulation to effectively enhance or decrease SOs, or provoke cortical arousals (Bellesi et al., 2014). We suggest that this arousal threshold is determined by the amplitude of the SO. This could be partly mediated by thalamocortical activity, since this drives synchronous firing in the cortex (Neske, 2016; Siclari et al., 2014). Importantly, cortical activity also suppresses the locus coeruleus such that it is harder for it to trigger arousal during large cortical up-states (Bellesi et al., 2014; Eschenko et al., 2012). Consequently, because of the increased excitability of the cortical up-states in SO peaks and the reduced cortico-coerulear interactions during large SO peaks, the stimulus during high amplitude waves induces a global hyperpolarizing effect which manifests in large SO troughs.

4.5. *Study limitations*

We would also like to caution the reader about some particularities of our study. Firstly, our definition of SA amplitude is broader than the definition of spindle events as we did not evaluate detected spindles. Nevertheless, SA amplitude and spindle detection are closely entwined. Likewise, we did not rule out the possibility that other SO or spindle features may also contribute to the prediction of spontaneous or boosted SO oscillations. Nevertheless, in this study we included the most prominent morphological features of the SO wave, and those that are less likely to be perturbed by the auditory click while preventing feature multicollinearity. Secondly, we must make the reader aware that computing all measures described here may be very slow, and therefore inefficient. Even aiming for a specific positive *v.Stim* amplitude, the most important feature, could be difficult and induce delays in the stimulation algorithm. Instead, computing *v.Stim* as an online hard threshold (such as in the SO detection process e.g. +50 μV) might also improve the amplitude responses for auditory clicks. Thirdly, our analysis process focussed on the global dynamics of SO activity across a sample of healthy subjects, therefore excluding any within-subject variability that may modify the response to the stimulus. Likewise, these parameters could change for CLAS responses in subjects with variable levels of cortical excitability such as CAPs and ultraslow oscillations (Bernardi et al., 2018; Lecci et al., 2017). Lastly, the real effects on functional or cognitive performance related to the enhancement of SO and spindles by CLAS should be further evaluated by measuring pre-stimulus metrics. Indeed, the increasing of SO amplitudes by CLAS could be related to underlying neurophysiological mechanisms that may or may not subserve memory consolidation. Hence, caution must be taken when associating the studied features with memory processes as this relationship cannot be predicted by the offline and retrospective analysis of this study.

5. Conclusions

To summarize, we trained machine learning algorithms to predict spontaneous and post-stimulus SO and SA amplitudes. We found that spontaneously generated SO trough amplitude can be predicted by the ongoing structure of the previous SO wave. By contrast, stimulus-related increase in a SO is mainly predicted by SO amplitude at the time of the click. Therefore, we suggest that SO amplitude may work as a kind of cortical threshold to prevent the click from causing arousal, while maintaining maximal cortical activity.

Based on these findings, we suggest that the online detection of a few salient features and the application of our random forest classifier could be used in the optimization of future CLAS algorithms. Using this method, future studies could evaluate the effect of less stimuli but improved SO response on memory and sleep architecture. Applying optimized version of CLAS may also facilitate a better understanding of the dynamics of both spontaneous SOs and SOs that have been boosted through CLAS.

Credit authorship contribution statement

Miguel Navarrete: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Steven Arthur:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Matthias S. Treder:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision. **Penelope A. Lewis:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

MN and PL were supported by the European Research Council (681607), SA was supported by a generous donation from Mr. Bernard Cheek. The authors would like to thank L. Santamaria-Covarrubias and M. Rakowska for their helpful comments and discussion on this manuscript.

Data and code availability

The data used in the study as well as the software and scripts used to compute the analysis have been made publicly available via the Open Science Framework and can be accessed at: https://osf.io/j9vka/

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119055.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347. doi:10.1093/bioinformatics/btq134.

Amzica, F., Steriade, M., 1995. Short- and long-range neuronal synchronization of the slow (<1 Hz) cortical oscillation. J. Neurophysiol. 73, 20–38. doi:10.1152/jn.1995.73.1.20.

Antony, J.W., Schönauer, M., Staresina, B.P., Cairney, S.A., 2019. Sleep spindles and memory reprocessing. Trends Neurosci. 42, 1–3. doi:10.1016/j.tins.2018.09.012.

Bellesi, M., Riedner, B.A., Garcia-Molina, G.N., Cirelli, C., Tononi, G., 2014. Enhancement of sleep slow waves: underlying mechanisms and practical consequences. Front. Syst. Neurosci. 8, 1–17. doi:10.3389/fnsys.2014.00208.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300.

Bernardi, G., Siclari, F., Handjaras, G., Riedner, B.A., Tononi, G., 2018. Local and widespread slow waves in stable NREM sleep: evidence for distinct regulation mechanisms. Front. Hum. Neurosci. 12, 1–13. doi:10.3389/fnhum.2018.00248.

Besedovsky, L., Ngo, H.V.V., Dimitrov, S., Gassenmaier, C., Lehmann, R., Born, J., 2017. Auditory closed-loop stimulation of EEG slow oscillations strengthens sleep and signs of its immune-supportive function. Nat. Commun. 8, 1984. doi:10.1038/s41467-017-02170-3.

Bódizs, R., Szalárdy, O., Horváth, C., Ujma, P.P., Gombos, F., Simor, P., Pótári, A., Zeising, M., Steiger, A., Dresler, M., 2021. A set of composite, non-redundant EEG measures of NREM sleep based on the power law scaling of the Fourier spectrum. Sci. Rep. 11, 1–18. doi:10.1038/s41598-021-81230-7.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One 12, 1–17. doi:10.1371/journal.pone.0177678.

Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324.

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 21, 1–13. doi:10.1186/s12864-019-6413-7.

Crunelli, V., Larincz, M.L., Connelly, W.M., David, F., Hughes, S.W., Lambert, R.C., Leresche, N., Errington, A.C., 2018. Dual function of thalamic low-vigilance state oscillations: rhythm-regulation and plasticity. Nat. Rev. Neurosci. 19, 107–118. doi:10.1038/nrn.2017.151.

da Silveira, T.L.T., Kozakevicius, A.J., Rodrigues, C.R., 2017. Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain. Med. Biol. Eng. Comput. 55, 343–352. doi:10.1007/s11517-016-1519-4.

Dimitriadis, S., Salis, C., Liparas, D., 2020. A sleep disorder detection model based on EEG cross-frequency coupling and random forest. medRxiv 1–26. 10.1101/2020.06.10.20126268

Eschenko, O., Magri, C., Panzeri, S., Sara, S.J., 2012. Noradrenergic neurons of the locus coeruleus are phase locked to cortical up-down states during sleep. Cereb. Cortex 22, 426–435. doi:10.1093/cercor/bhr121.

Grimaldi, D., Papalambros, N.A., Reid, K.J., Abbott, S.M., Malkani, R.G., Gendy, M., Iwanaszko, M., Braun, R.I., Sanchez, D.J., Paller, K.A., Zee, P.C., 2019. Strengthening sleep–autonomic interaction via acoustic enhancement of slow oscillations. Sleep 42, 1–11. doi:10.1093/sleep/zsz036.

Hapfelmeier, A., Hothorn, T., Ulm, K., 2012. Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Comput. Stat. Data Anal. 56, 1552–1565. doi:10.1016/j.csda.2011.09.024.

Helfrich, R.F., Lendner, J.D., Mander, B.A., Guillen, H., Paff, M., Mnatsakanyan, L., Vadera, S., Walker, M.P., Lin, J.J., Knight, R.T., 2019. Bidirectional prefrontal-hippocampal dynamics organize information transfer during sleep in humans. Nat. Commun. 10, 3572. doi:10.1038/s41467-019-11444-x.

Henin, S., Borges, H., Shankar, A., Sarac, C., Melloni, L., Friedman, D., Flinker, A., Parra, L.C., Buzsaki, G., Devinsky, O., Liu, A., 2019. Closed-loop acoustic stimulation enhances sleep oscillations but not memory performance. eNeuro doi:10.1523/ENEURO.0306-19.2019, ENEURO.0306-19.2019.

Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S.F., 2007. The AASM Manual for the Scoring of Sleep and Associated Events: Rules Terminology and Technical Specifications, 1st ed. American Academy of Sleep Medicine, Westchester, IL.

Jiang, X., Gonzalez-Martinez, J., Halgren, E., 2019. Posterior hippocampal spindle ripples co-occur with neocortical theta bursts and downstates-upstates, and phase-lock with parietal spindles during NREM sleep in humans. J. Neurosci. 39, 8949–8968. doi:10.1523/JNEUROSCI.2858-18.2019.

Kawakita, M., Takeuchi, J., 2014. Safe semi-supervised learning based on weighted likelihood. Neural Netw. 53, 146–164. doi:10.1016/j.neunet.2014.01.016.

Kim, J., Gulati, T., Ganguly, K., 2019. Competing roles of slow oscillations and delta waves in memory consolidation versus forgetting. Cell 179, 514–526. doi:10.1016/j.cell.2019.08.040, .e13.

Klinzing, J.G., Niethard, N., Born, J., 2019. Mechanisms of systems memory consolidation during sleep. Nat. Neurosci. 22, 1598–1610. doi:10.1038/s41593-019-0467-3.

Lecci, S., Fernandez, L.M.J., Weber, F.D., Cardis, R., Chatton, J.Y., Born, J., Lüthi, A., 2017. Coordinated infraslow neural and cardiac oscillations mark fragility and offline periods in mammalian sleep. Sci. Adv. 3. doi:10.1126/sciadv.1602026.

Leminen, M., Virkkala, J., Saure, E., Paajanen, T., Zee, P., Santostasi, G., Hublin, C., Müller, K., Porkka-Heiskanen, T., Huotilainen, M., Paunio, T., 2017. Enhanced memory consolidation via automatic sound stimulation during non-REM sleep. Sleep 40. doi:10.1093/sleep/zsx003.

Maingret, N., Girardeau, G., Todorova, R., Goutierre, M., Zugaro, M., 2016. Hippocampo-cortical coupling mediates memory consolidation during sleep. Nat. Neurosci. 19, 959–964. doi:10.1038/nn.4304.

Marshall, L., Helgadóttir, H., Mölle, M., Born, J., 2006. Boosting slow oscillations during sleep potentiates memory. Nature 444, 610–613. doi:10.1038/nature05278.

Massimini, M., 2002. EEG slow (~1 Hz) waves are associated with nonstationarity of thalamo-cortical sensory processing in the sleeping human. J. Neurophysiol. 89, 1205–1213. doi:10.1152/jn.00373.2002.

Massimini, M., Ferrarelli, F., Esser, S.K., Riedner, B.a., Huber, R., Murphy, M., Peterson, M.J., Tononi, G., 2007. Triggering sleep slow waves by transcranial magnetic stimulation. Proc. Natl. Acad. Sci. U. S. A. 104, 8496–8501. doi:10.1073/pnas.0702495104.

Moser, B.K., Stevens, G.R., 1992. Homogeneity of variance in the two-sample means test. Am. Stat. 46, 19–21.

Navarrete, M., Schneider, J., Ngo, H.V, Valderrama, M., Casson, A.J., Lewis, P.A., 2020a. Examining the optimal timing for closed-loop auditory stimulation of slow-wave sleep in young and older adults. Sleep 43, 1–14. doi:10.1093/sleep/zsz315.

Navarrete, M., Valderrama, M., Lewis, P.A., 2020b. The role of slow-wave sleep rhythms in the cortical-hippocampal loop for memory consolidation. Curr. Opin. Behav. Sci. 32, 102–110. doi:10.1016/j.cobeha.2020.02.006.

Neske, G.T., 2016. The slow oscillation in cortical and thalamic networks: mechanisms and functions. Front. Neural Circuits 9, 1–25. doi:10.3389/fncir.2015.00088.

Ngo, H.V.V, Martinetz, T., Born, J., Mölle, M., 2013. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. Neuron 78, 545–553. doi:10.1016/j.neuron.2013.03.006.

Ngo, H.V.V, Miedema, A., Faude, I., Martinetz, T., Molle, M., Born, J., 2015. Driving sleep slow oscillations by auditory closed-loop stimulation–a self-limiting process. J. Neurosci. 35, 6630–6638. doi:10.1523/JNEUROSCI.3133-14.2015.

Ong, J.L., Patanaik, A., Chee, N.I.Y.N., Lee, X.K., Poh, J.H., Chee, M.W.L., 2018. Auditory stimulation of sleep slow oscillations modulates subsequent memory encoding through altered hippocampal function. Sleep 1–11. doi:10.1093/sleep/zsy031.

Papalambros, N.A., Santostasi, G., Malkani, R.G., Braun, R., Weintraub, S., Paller, K.A., Zee, P.C., 2017. Acoustic enhancement of sleep slow oscillations and concomitant memory improvement in older adults. Front. Hum. Neurosci. 11, 1–14. doi:10.3389/fnhum.2017.00109.

Papoulis, A., Pillai, U., 2002. Probability, Random Variables and Stochastic Processes, 4th ed. McGraw-Hill.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., Battaglia, F.P., 2009. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat. Neurosci. 12, 919–926. doi:10.1038/nn.2337.

Purcell, S.M., Manoach, D.S., Demanuele, C., Cade, B.E., Mariani, S., Cox, R., Panagiotaropoulou, G., Saxena, R., Pan, J.Q., Smoller, J.W., Redline, S., Stickgold, R., 2017. Characterizing sleep spindles in 11,630 individuals from the national sleep research resource. Nat. Commun. 8, 15930. doi:10.1038/ncomms15930.

Riedner, B.a., Vyazovskiy, V.V, Huber, R., Massimini, M., Esser, S., Murphy, M., Tononi, G., 2007. Sleep homeostasis and cortical synchronization: III. A high-density EEG study of sleep slow waves in humans. Sleep 30, 1643–1657. doi:10.1093/sleep/30.12.1643.

Rosanova, M., Timofeev, I., 2005. Neuronal mechanisms mediating the variability of somatosensory evoked potentials during sleep oscillations in cats. J. Physiol. 562, 569–582. doi:10.1113/jphysiol.2004.071381.

Santostasi, G., Malkani, R., Riedner, B., Bellesi, M., Tononi, G., Paller, K.A., Zee, P.C., 2015. Phase-locked loop for precisely timed acoustic stimulation during sleep. J. Neurosci. Methods 259, 101–114. doi:10.1016/j.jneumeth.2015.11.007.

Schneider, J., Lewis, P.A., Koester, D., Born, J., Ngo, H.V.V, 2020. Susceptibility to auditory closed-loop stimulation of sleep slow oscillations changes with age. Sleep 1–10. doi:10.1093/sleep/zsaa111.

Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference 90, 227–244. doi:10.1016/s0378-3758(00)00115-4.

Siclari, F., Bernardi, G., Riedner, B.a., LaRocque, J.J., Benca, R.M., Tononi, G., 2014. Two distinct synchronization processes in the transition to sleep: a high-density electroencephalographic study. Sleep 37, 1621–1637. doi:10.5665/sleep.4070.

Sugiyama, M., Krauledat, M., Müller, K.R., 2007. Covariate shift adaptation by importance weighted cross validation. J. Mach. Learn. Res. 8, 985–1005.

Thomas, D., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1924.

Warby, S.C., Wendt, S.L., Welinder, P., Munk, E.G.S., Carrillo, O., Sorensen, H.B.D., Jennum, P., Peppard, P.E., Perona, P., Mignot, E., 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat. Methods 11, 385–392. doi:10.1038/nmeth.2855.

Xie, L., Kang, H., Xu, Q., Chen, M.J., Liao, Y., Thiyagarajan, M., O'Donnell, J., Christensen, D.J., Nicholson, C., Iliff, J.J., Takano, T., Deane, R., Nedergaard, M., 2013. Sleep drives metabolite clearance from the adult brain. Science 342, 373–377. doi:10.1126/science.1241224, (80-.).