PLOS COMPUTATIONAL BIOLOGY

METHODS

Somnotate: A probabilistic sleep stage classifier for studying vigilance state transitions

Paul J. N. Brodersen[1], **Hannah Alfonsa**[1], **Lukas B. Krone**[2], **Cristina Blanco-Duque**[2], **Angus S. Fisk**[3], **Sarah J. Flaherty**[2], **Mathilde C. C. Guillaumin**[3,4,5], **Yi-Ge Huang**[2], **Martin C. Kahn**[2], **Laura E. McKillop**[2], **Linus Milinski**[2], **Lewis Taylor**[3], **Christopher W. Thomas**[2], **Tomoko Yamagata**[3], **Russell G. Foster**[4], **Vladyslav V. Vyazovskiy**[2], **Colin J. Akerman**[1]*

1 Department of Pharmacology, University of Oxford; Mansfield Road, Oxford, United Kingdom, **2** Department of Physiology, Anatomy and Genetics, University of Oxford; Parks Road, United Kingdom, **3** Nuffield Department of Clinical Neurosciences, University of Oxford; John Radcliffe Hospital, Oxford, United Kingdom, **4** Sleep and Circadian Neuroscience Institute, University of Oxford; Oxford, United Kingdom, **5** Institute for Neuroscience, Department of Health Sciences and Technology, ETH Zurich; Schwerzenbach, Switzerland

* colin.akerman@pharm.ox.ac.uk

OPEN ACCESS

Citation: Brodersen PJN, Alfonsa H, Krone LB, Blanco-Duque C, Fisk AS, Flaherty SJ, et al. (2024) Somnotate: A probabilistic sleep stage classifier for studying vigilance state transitions. PLoS Comput Biol 20(1): e1011793. https://doi.org/10.1371/journal.pcbi.1011793

Editor: Daniel Bush, University College London, UNITED KINGDOM

Received: June 26, 2023

Accepted: January 2, 2024

Published: January 17, 2024

Copyright: © 2024 Brodersen et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Somnotate is distributed as open source software under the GPLv3 license. The repository is available at https://github.com/paulbrodersen/somnotate, and archived at Zenodo (DOI:10.5281/zenodo.8055723). The datasets including electrophysiological recordings, manual annotations, and automated annotations are also deposited at Zenodo (DOI:10.5281/zenodo.10200481.

Abstract

Electrophysiological recordings from freely behaving animals are a widespread and powerful mode of investigation in sleep research. These recordings generate large amounts of data that require sleep stage annotation (polysomnography), in which the data is parcellated according to three vigilance states: awake, rapid eye movement (REM) sleep, and non-REM (NREM) sleep. Manual and current computational annotation methods ignore intermediate states because the classification features become ambiguous, even though intermediate states contain important information regarding vigilance state dynamics. To address this problem, we have developed "Somnotate"—a probabilistic classifier based on a combination of linear discriminant analysis (LDA) with a hidden Markov model (HMM). First we demonstrate that Somnotate sets new standards in polysomnography, exhibiting annotation accuracies that exceed human experts on mouse electrophysiological data, remarkable robustness to errors in the training data, compatibility with different recording configurations, and an ability to maintain high accuracy during experimental interventions. However, the key feature of Somnotate is that it quantifies and reports the certainty of its annotations. We leverage this feature to reveal that many intermediate vigilance states cluster around state transitions, whereas others correspond to failed attempts to transition. This enables us to show for the first time that the success rates of different types of transition are differentially affected by experimental manipulations and can explain previously observed sleep patterns. Somnotate is open-source and has the potential to both facilitate the study of sleep stage transitions and offer new insights into the mechanisms underlying sleep-wake dynamics.

Funding: The research leading to these results received funding from the European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013 (ERG Grant Agreement 617670 to CJA), the Leverhulme Trust (RPG-2023-083 to CJA), the Medical Research Council UK (MR/S01134X/1 to VVV and CJA; Clinical Research Training Fellowship to YGH), the Wellcome Trust (106174/Z/14/Z to VVV; 206500/Z/17/Z to HA; 203971/Z/16/Z to LBK; 109059/Z/15/Z to CBD; WT106174/Z/14/ZMA to LT), the John Fell Fund (131/032 to VVV), the Royal Society (RG1204666 to VVV), St. John's College, University of Oxford (Junior research fellowship to HA), Hertford College, University of Oxford (Senior Scholarship to LBK), Radcliffe Department of Medicine, University of Oxford (Goodger and Schorstein Research Scholarship to LBK and MCK), the Clarendon Fund (Graduate Scholarship to CBD and MCCG), the Biotechnology and Biological Sciences Research Council (BB/J014427/1 to MCCG; BB/K011847/1 to LEM; BB/M011224/1 to CWT), Lincoln College, University of Oxford (Berrow Foundation Lord Florey Scholarship to MCK), the Novo Nordisk UK Research Foundation (Postdoctoral fellowship to LEM), Linacre College, University of Oxford (Sir Paul Nurse Junior Research Fellowship to LEM), Action on Hearing Loss (Studentship to LM), the Uehara Memorial Foundation (Overseas Postdoctoral Fellowship to TY), and the Naito Science and Engineering Foundation (Grant for Studying Overseas to TY). The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Typically, the three different vigilance states–awake, REM sleep, and non-REM sleep–exhibit distinct features that are readily recognised in electrophysiological recordings. However, particularly around vigilance state transitions, epochs often exhibit features from more than one state. These intermediate vigilance states pose challenges for existing manual and automated classification methods, and are hence often ignored. Here, we present 'Somnotate'—an open-source, highly accurate and robust sleep stage classifier, which supports research into intermediate states and sleep stage dynamics in mice. Somnotate quantifies and reports the certainty of its annotations, enabling the experimenter to identify abnormal epochs in a principled manner. We use this feature to identify intermediate states and to detect unsuccessful attempts to switch between vigilance states. This has the potential to provide new insights into the mechanisms of vigilance state transitions, and creates new opportunities for future experiments.

Introduction

Long-term electrophysiological recordings from freely behaving mice and other laboratory animals are a popular and powerful mode of investigation for neuroscientists, particularly sleep researchers [1–3]. This approach affords the study of a wide range of animal behaviours and associated neurophysiological activities, under experimentally controlled conditions. The recordings typically incorporate an electroencephalogram (EEG) signal recorded from the cortical surface at one or more locations, and may also include electromyogram (EMG) recordings from relevant muscle groups. As the vigilance state profoundly affects the behaviour and physiology of the animal, the first step in the analysis is typically sleep stage annotation. This involves the parcellation of the data into three vigilance states: awake, rapid eye movement (REM) sleep, and non-REM (NREM) sleep.

Whilst sleep stage annotation is typically performed by human experts, there are two principal motives for developing effective automated methods for annotating sleep data. First, manual annotation is time-consuming, as an experienced scorer typically requires several hours to annotate a single 24-hour data set. This places a significant burden on the analysis stage of most experiments, meaning that automation is required to conduct experiments at a scale that would be otherwise difficult to imagine [4,5]. To this end, several automated methods have been developed for sleep stage annotation, primarily in the context of human clinical data [6–24], but also in the context of laboratory animals [25–33].

The second principal motive for developing automated sleep stage annotation is that the underlying electrophysiological signals can be ambiguous with respect to the sleep stage [34]. This is especially the case during intermediate sleep states, when EEG signals can exhibit features of more than one vigilance state. Local slow wave activity for example, which is normally considered a hallmark of NREM sleep, has been observed during REM and awake states across different cortical regions in humans [35,36] and rodents [37–39]. Whilst the AASM guidelines [40] have established rules for managing such ambiguity for standardised human data in a clinical setting, no equivalent standard exists for animal research. Indeed, as the number and placement of electrodes often varies between experiments (and within a single study), it is difficult to formulate an analogous standard with similar precision. Consequently, manual annotations often differ between scorers [41,42]. Automated methods remove such inter-rater variance and afford new opportunities to systematically describe these intermediate states, and investigate their importance for sleep-wake dynamics. Whilst for human clinical data several methods have been designed with these aims in mind [43–47], no comparable approach exists for animal polysomnography.

Here, we present ‘Somnotate’—an open-source, highly accurate and robust sleep stage classifier, which supports research into intermediate states and sleep stage dynamics. Somnotate is shown to consistently exceed the accuracy of expert manual annotations on rodent electrophysiological data and is remarkably robust to errors in the training data and across a series of experimental manipulations. Somnotate is a probabistic classifier that quantifies the certainty of its predictions, thereby identifying epochs that present abnormal electrophysiological signals. This not only affords the opportunity to review the algorithm’s predictions, but also offers a principled method for describing intermediate sleep states across large data sets, where an exhaustive review of the data would be impractical. Here, we show that intermediate states are typically present around vigilance state transitions, but also identify failed transition attempts. We show that different types of state transitions have markedly different success rates, which may explain commonly observed sleep patterns, and we demonstrate that the success rates are altered by interventions such as sleep deprivation. Somnotate can thus facilitate the study of sleep stage transitions and has the potential to offer new insights into the mechanisms underlying sleep-wake dynamics.

Results

Establishing a probabilistic vigilance state classifier

In machine learning, classifiers typically learn a transformation that maps a sample consisting of a set of input values or features, onto a single output value or category. This forms the basis of most automated methods for sleep stage classification, such as decision trees, linear discriminant analysis (LDA), support vector machines, and neural networks [17,27,29–31,48]. A key weakness of deterministic machine learning approaches is that each input is mapped onto a single output. As a result, any ambiguity in the sample is ignored, and there is no indication how certain the classifier is in its estimate of the vigilance state. This problem can be solved by using a probabilistic classifier that maps each sample onto a probability distribution over vigilance states. The simplest probabilistic classifier is the naive Bayes classifier, which predicts the most likely state $\hat{s} \in S$ of each sample $d \in D$ in the test set based on the probability of the sample given each state $P(D|S)$ weighted by the frequency of states $P(S)$:

$$\hat{s} = \text{argmax}_s P(D|S)\ P(S)$$

A shortcoming of this approach is that each sample is classified independently and, as the input data can be noisy, this can result in misclassifications and an overestimation of the number of state transitions. Human scorers avoid these issues by using contextual information, subjectively integrating the evidence within each time bin, with an estimate of the likelihood of each state based on the broader context; the high performance of recurrent and convolutional neural networks compared to other methods is likely due to similar reasons [45,49,50]. Algorithmically, the simplest way to integrate predictions based on individual samples with contextual information is to smooth the inferred state sequence by determining the most common vigilance state surrounding the sample of interest. However, vigilance states can be short-lived, which poses a difficult problem for such an approach. For example, mice often transition from REM sleep to NREM sleep via a brief period in which their EEG and EMG activity reflect the awake state [51–54]. A probabilistic classifier that can systematically integrate contextual information without smoothing is the hidden Markov model (HMM). The HMM classifier is conceptually similar to the naive Bayes classifier, with the difference being that the prior probability of the state, P(S), is not approximated by its expected frequency, but rather depends on the overall most likely state sequence given the context of the surrounding samples.

A drawback of probabilistic classifiers is that they require an estimate of the multivariate probability distribution over input values for each state. With each additional feature, the number of samples needed to accurately estimate these probability distributions increases exponentially. Consequently, HMMs that have been used in vigilance state annotation previously used low dimensional, hand-crafted features [55–57] or state sequences inferred by other algorithms [7,8,20]. Due to the biases of human perception, manual feature engineering from complex signals carries the risk of discarding information that might be valuable for a downstream machine learning classifier. To avoid this, we set out to combine probabilistic classifiers with LDA, which can be trained to automatically extract low dimensional features from complex, high-dimensional inputs, whilst retaining the maximal amount of (linearly decodable) information about the target classes.

To benchmark the performance of our new approach, and to illustrate the advantage conferred by incorporating contextual information, we compared the performance of an HMM to a naive Bayes classifier and an LDA classifier on a data set of six 24-hour (i.e. 144 hours total) simultaneous recordings of anterior EEG, posterior EEG, and EMG in freely behaving mice (Materials and methods), which were independently scored by at least four experienced sleep researchers to annotate awake, NREM, and REM states (**Fig 1A**). For all classifiers, the data was prepared in the same way. The EEG and EMG traces (**Fig 1B**) were subsampled to 256 Hz and converted to multitaper spectrograms [58]. The power values in each frequency band were normalised by applying a log(x + 1) transformation and then converting the result to z-scores (**Fig 1C**). As power is exponentially distributed in EEG and EMG signals, a log(x + 1) transformation results in approximately normally distributed values. This facilitates the determination of robust linear discriminants and the targeted dimensionality reduction via linear discriminant analysis (LDA; **Fig 1D**), the final step. Thresholding the LDA representations yielded the LDA classification. For the naive Bayes classifier, multivariate Gaussian distributions (one for each state; **Fig 1G**) were fitted to the low dimensional representation of the samples in the training data set, and the corrsponding state annotations were used to determine the expected frequency of each state (**Fig 1H**). The states corresponding to samples in the test set were then predicted based on the probability of the sample given each state P(D|S) (**Fig 1E**), weighted by their frequency P(S). For the HMM, the probability given each state P(D|S) was computed in the same way as for the Bayes classifier. However, based on the expected transition frequencies observed in the state sequences in the training data set (**Fig 1I**), P(S) was computed using the Baum-Welch algorithm (**Fig 1F)** and the most likely state sequence through the test set was computed using the Viterbi algorithm [59].

The LDA classifier was found to achieve an accuracy of 92% ± 1% on the test data set (**Fig 1J, "LDA"**), which was comparable to previous work that used an LDA classifier to predict vigilance states from rodent experimental EEG data (89% ± 1% accuracy [29]). The Bayes classifier achieved an accuracy of 93% ± 1% on the test data set (**Fig 1J, "Bayes"**), which was consistent with previously reported values using a similar approach to predict vigilance states from rodent EEG data (94% ± 1% [30]). Finally, the HMM had an accuracy of 97% ± 1%, which was significantly more accurate than either the LDA or the Bayes classifiers, reducing the number of errors by more than half in both cases (**Fig 1J, "HMM"**). These analyses confirmed the benefit of incorporating contextual information. They also established automated feature extraction using LDA, combined with context-aware state annotation using a HMM, as a highly effective strategy for achieving very accurate automated sleep stage classification of experimental data. We refer to this improved classifier as 'Somnotate'.

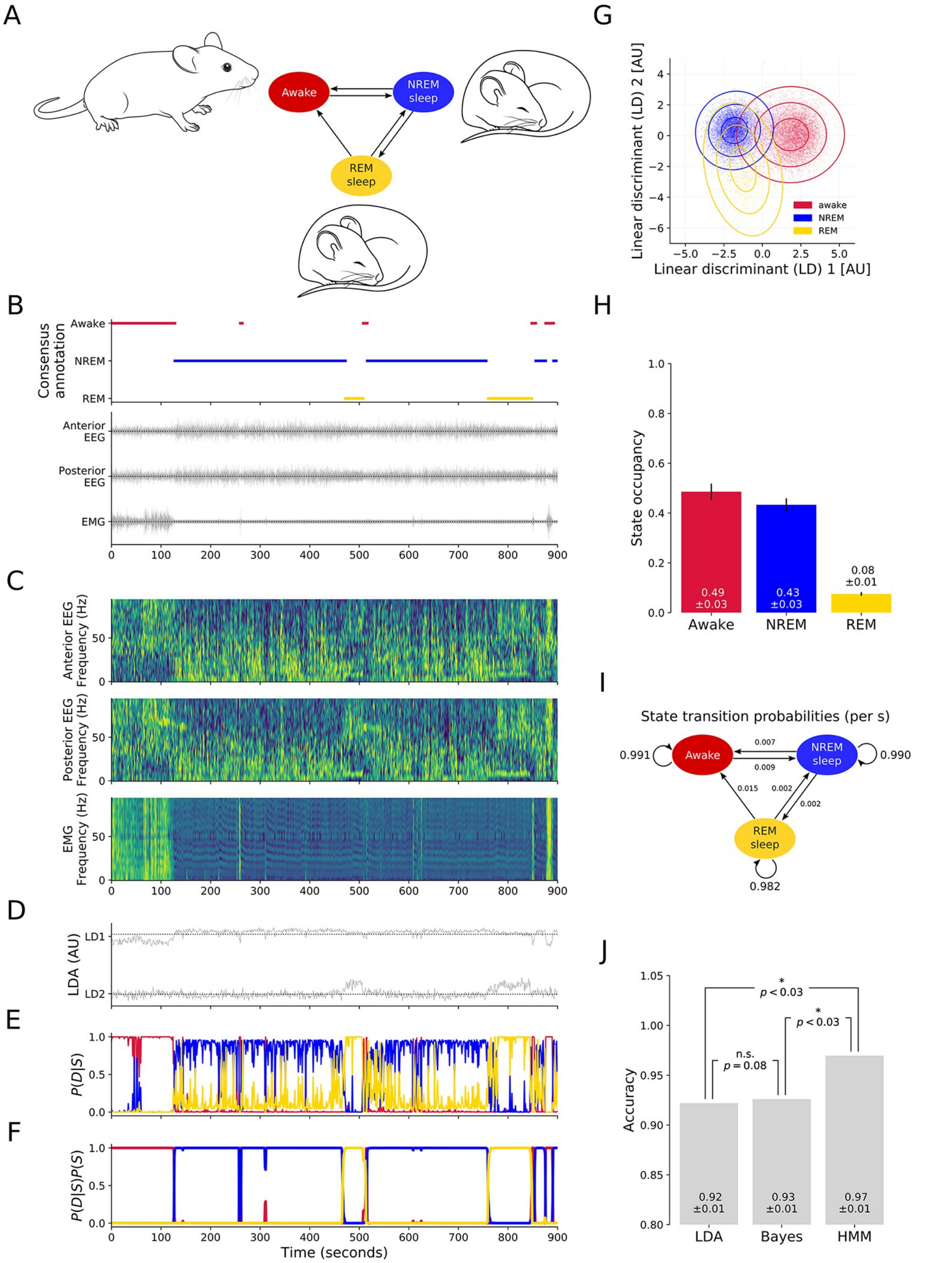

Fig 1. Establishing a probabilistic sleep stage classifier. (**A**) Continuous EEG and EMG recordings were made across a full sleep-wake cycle from freely behaving mice. (**B**) A fifteen-minute segment of the consensus of manual annotations by four independent experienced sleep researchers (top) and the corresponding anterior EEG, posterior EEG, and EMG recording. (**C**) Anterior EEG, posterior EEG and EMG multi-taper spectrograms. (**D**) Two-dimensional representation of the segment after targeted dimensionality reduction via LDA. Negative values in the first component ('LD1') and in the second component ('LD2') indicate the awake state; positive LD1 with negative LD2 indicates NREM; negative LD1 with positive LD2 indicates REM. (**E**) Probability of each state when fitting two-dimensional Gaussian distributions to the values in 'D'. (**F**) Likelihood of each state given the probability of each state (as shown in 'E') and all possible state sequences, weighted by their likelihood given the state transition probabilities (as shown in 'I'). (**G**) Distribution of values after dimensionality reduction by LDA. Each dot corresponds to a randomly chosen 1-second epoch. Colour indicates the state assigned in the manual consensus annotation. Lines indicate the standard deviations of multivariate Gaussian distributions, one for each state, fitted to all samples in the data set. (**H**) The state occupancy based on the time spent in each state across six 24-hour data sets, according to at least four

manual annotations. (**I**) The corresponding state transition probabilities. (**J**) Accuracy of the LDA classifier, the naive Bayes classifier, and the HMM classifier (i.e. Somnotate). Accuracy was evaluated across six 24-hour data sets in a hold-one-out fashion. Error bars indicate standard deviation. P-values are derived from a Wilcoxon signed rank test with a Bonferroni-Holm correction for multiple comparisons.

https://doi.org/10.1371/journal.pcbi.1011793.g001

Sleep stage annotation by Somnotate exceeds manual accuracy and compares favourably to other automated solutions

Manual annotation continues to be the gold standard by which any automated annotation is measured. The performance of sleep stage classifiers is typically measured by computing their agreement with two independent manual annotations. Performance is evaluated as the average agreement of the automated annotation with each of the two manual annotations, and this average is then compared to the level of agreement between the two manual annotations. This subtle difference in how manual and automated annotations are compared can lead to systematic biases in favour of the automated annotation (see **S1 Appendix**).

For these reasons, we were keen to compare manual and automated annotations to a majority-vote consensus derived from multiple independent manual annotations. We generated a data set of six 24-hour (i.e. 144 hours total) simultaneous recordings of anterior EEG, posterior EEG, and EMG in freely behaving mice (Materials and methods), which were independently scored by at least four experienced sleep researchers to annotate awake, NREM, and REM states. The sleep researchers had a median of 5 years' experience in manual vigilance state annotation (minimum of 2 years' experience), and had manually annotated a median of 1272 hours (minimum of 768 hours) of equivalent recordings (**S1 Table**). The recordings, individual manual annotations, and automated annotations are made freely available in standard formats (see Data Availability Statement).

To determine the accuracy of manual annotations by experienced sleep researchers, we compared individual manual annotations for any of the six 24-hour recordings to the consensus of the remaining three or more annotations of the same recording. Somnotate was trained and tested in a hold-one-out fashion on the same six recordings, and its accuracy was determined by comparison to the consensus of the manual annotations. This revealed that the accuracy of Somnotate exceeded the accuracy of manual annotations by 13 experienced sleep researchers (**Fig 2A**). Out of a total of 25 manual annotations, 22 were less accurate than the automated annotation. Twelve out of the thirteen annotators had a lower average accuracy than the automated classifier on the same data sets. Thus Somnotate significantly exceeded human performance in terms of accuracy ($p < 0.001$, Wilcoxon signed rank), and we obtained identical results for other measures of performance, such as Cohen's kappa and the weighted F1 score (**Tables 1** and **S2** and **S6 Fig**).

The difference between the confusion matrices for the manual and automated annotations indicated that the performance difference between manual and automated annotation was mainly driven by a more accurate annotation of NREM states (**Fig 2B**). Somnotate identified more state transitions than were typically present in manual annotations, in particular if these transitions involved NREM states. Cumulatively however, the differences between manual and automated state annotations resulted in minor differences in the overall state occupancy (**Fig 2C and 2D**). When partitioning the data by state according to the manual consensus or the automated annotation, there were no discernible differences between power spectra of the EEG activity (**S1 Fig**).

Approximately two thirds of the differences between Somnotate and the manual consensus annotations had a duration that was shorter than the temporal resolution of the manual annotations (i.e. shorter than 4 s; **Fig 2E**). Many of these differences could therefore be resolved if

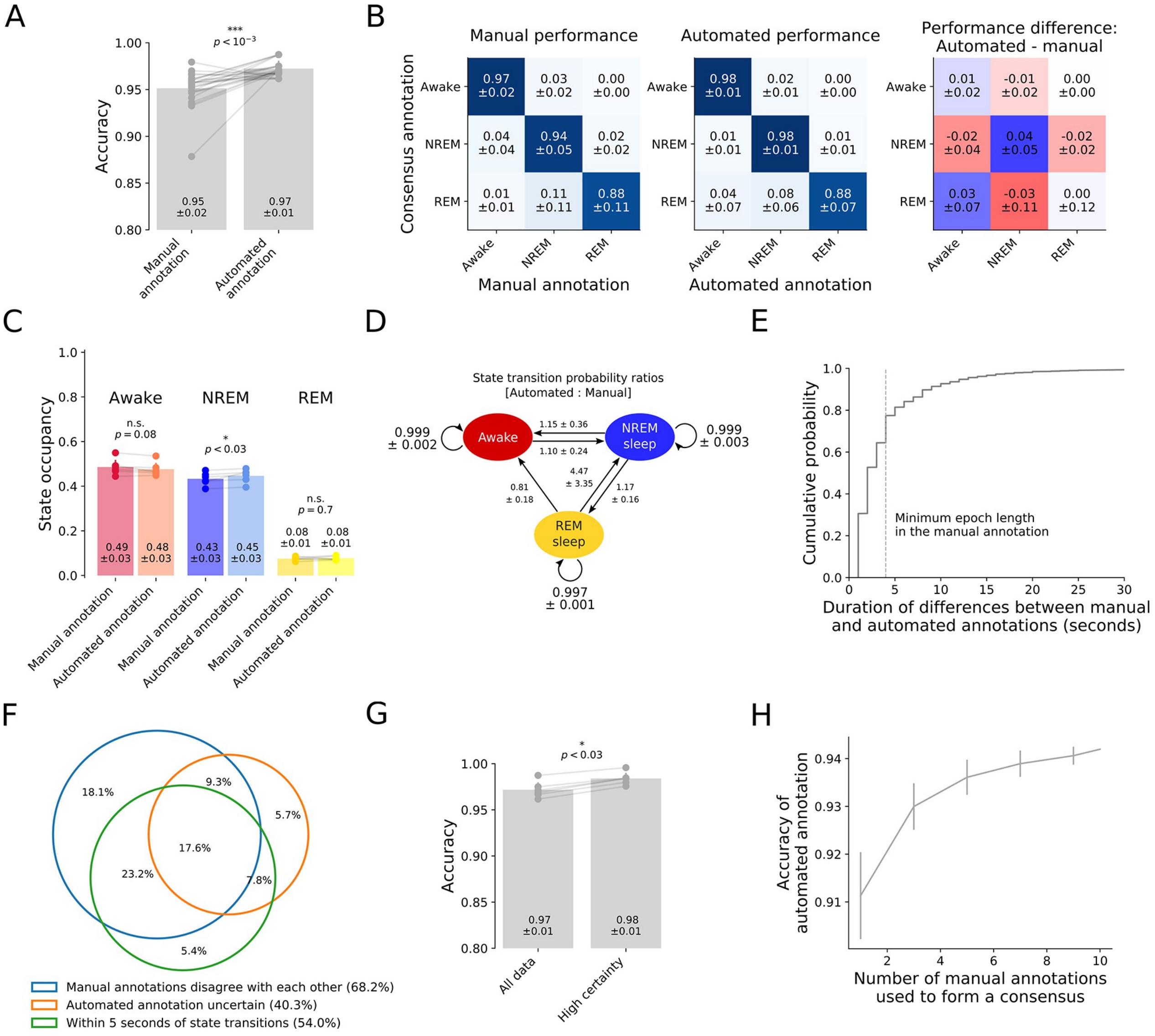

Fig 2. Automated sleep stage classification by Somnotate exceeds manual accuracy. (**A**) Somnotate was trained and tested, in a hold-one-out fashion, on six 24-hour data sets. Using a consensus annotation based on at least three manual annotations, the accuracy of the classifier was compared to the accuracy of individual manual annotations (n = 25 manual annotations from 13 experienced sleep researchers). (**B**) The confusion matrix for individual manual annotations compared to the manual consensus (left), for the automated classifier compared to the manual consensus (middle), and the difference between these two confusion matrices (right). (**C**) Comparison of state occupancies between the automated and manual consensus annotations. (**D**) State transition probabilities in the automated annotation, normalised to the state transition probabilities in the manual consensus annotation. (**E**) Cumulative frequency plot shows the duration of the differences between the automated annotation and the manual consensus. Note that the manual annotation had a temporal resolution of 4 s (vertical dashed line), whereas the automated classification was performed at a time resolution of 1 s. (**F**) Venn-diagram of the time points at which the automated annotation and manual consensus differed. (**G**) Excluding samples where Somnotate is not certain improves accuracy. Classifier accuracy was compared between cases when all samples were included ('All data') and when 5.5% of samples were removed because the likelihood of the predicted state dropped below 0.995 ('High certainty'). The plot indicates mean ± standard deviation and p-values are derived from a Wilcoxon signed rank test. (**H**) Somnotate was trained on six 24-hour data sets and then tested on a 12-hour data set, which had been independently annotated by ten experienced sleep researchers (as in Fig 1). The accuracy of the annotation by Somnotate was compared to consensus annotations generated from different numbers of manual annotations. Error bars indicate standard deviation. P-values are derived from a Wilcoxon signed rank test.

https://doi.org/10.1371/journal.pcbi.1011793.g002

the data had been manually annotated at a higher temporal resolution, albeit at the expense of a greater investment of time. In other cases however, annotating a definitive state may not have been possible. For instance, the animal may have been transitioning from one state to another, resulting in ambiguous EEG and EMG waveforms that reflect an ‘intermediate’ state. Consistent with this scenario, 54% of differences between the consensus and the automated annotation occurred within 5 seconds of a state transition, where manual annotations also often disagreed with one another (Fig 2F). As Somnotate is a probabilistic classifier, it can automatically identify intermediate states as epochs with non-zero probabilities for more than one state (denoted by "Automated state annotations uncertain" in Fig 2F). Excluding these epochs reduces the differences between the consensus annotation and the automated annotation by a third (Fig 2G), and provides a convenient way to clean up the EEG data for further downstream analysis. As a final validation, we trained Somnotate on the six 24-hour test data sets, but then tested performance on the 12-hour data set that had been annotated by ten experienced sleep researchers. This revealed that the more manual annotations that were used to generate a consensus sequence of the test data, the more closely this manual consensus matched the automated annotation (Spearman’s rank correlation $\rho = 0.80$, $p < 0.001$; Fig 2H). In other words, as one increases the number of experienced annotators, the manual consensus converges on the automated annotation by Somnotate.

Finally, we were keen to compare Somnotate to other solutions for automated vigilance state annotation in experimental animal models, in particular mice. To that end, we assessed the performance of IntelliSleepScorer [60] and SPINDLE [49] on the six 24-hour EEG & EMG recordings annotated by multiple experts. We choose these two programs as reference points based on the following criteria: (1) both programs and all of their dependencies are open source software, freely available, and thus highly accessible. (2) Both support re-training, and can, in principle, be adapted to different recording configurations and experimental setups. (3) Both come with fully parameterised models that have been optimised by their authors based on large corpuses of annotated recordings. (4) Analogously to Somnotate, both use three electrophysiological signals by default, specifically frontal EEG, parietal EEG, and neck EMG, to annotate each epoch (SPINDLE: 4 seconds; IntelliSleepScorer: 10 seconds) with one of three vigilance states (Awake, NREM, REM); (5) Both are well-documented and include example data, such that we could be confident that we prepared our test data appropriately, and invoked the software correctly. Beyond these objective criteria, we were particularly interested in SPINDLE, as its underlying approach is conceptually similar to Somnotate: both, SPINDLE and Somnotate, use HMMs for classification. As HMMs require low dimensional signals to be trained effectively, SPINDLE and Somnotate employ sophisticated dimensionality reduction techniques. However, whereas Somnotate uses linear discriminant analysis (LDA), requiring optimisation of only four parameters, SPINDLE uses a deep convolutional neural network (CNN). This leverages recent advances in deep learning, but also requires the optimisation of hundreds of thousands of parameters and hence substantially more training data. We were interested to what degree this trade-off could be justified by the model’s performance. IntelliSleepScorer, on the other hand, is based on LightGBM, a decision tree ensemble model developed by Microsoft, and seemed a reasonable representative of the current state-of-the-art, being to our knowledge the most recently released automated solution for animal polysomnography. Interestingly, SPINDLE—at least in our hands—performed much better than the more recently published IntelliSleepScorer across all evaluated metrics (accuracy, Cohen’s kappa, and weighted F1 score; Table 1). However, though the differences were much smaller, Somnotate still performed better than SPINDLE. This indicates that for animal polysomnography, a combination of targeted dimensionality reduction with hidden Markov models yields

Table 1. Performance of Somnotate and other state-of-the-art algorithms for automated mouse polysomnography compared to manual annotation by experienced experts. Manual and automated annotations of six 24 hour datasets were evaluated based on the consensus of multiple expert annotations. Values represent mean ± standard deviation.

Method	Accuracy	Cohen's kappa	Weighted F1 score
Somnotate	0.97 ± 0.01	0.95 ± 0.01	0.97 ± 0.01
Manual	0.95 ± 0.02	0.91 ± 0.03	0.95 ± 0.02
SPINDLE	0.94 ± 0.01	0.90 ± 0.01	0.94 ± 0.01
IntelliSleepScorer	0.75 ± 0.04	0.54 ± 0.07	0.73 ± 0.05

https://doi.org/10.1371/journal.pcbi.1011793.t001

comparable results, and that linear, low-parameter approaches to dimensionality reduction such as linear discriminant analysis are sufficient to perform well.

Somnotate is highly robust

Machine learning algorithms can be uniquely sensitive to patterns in the data. This sensitivity is often desirable, but can also be problematic. We were therefore keen to examine Somnotate's performance under conditions in which the training data contained errors, or where the test data reflected different experimental conditions, or where the data was acquired using a different experimental recording configuration.

First, to test sensitivity to errors in the training data, we evaluated Somnotate's accuracy on six 24-hour EEG and EMG recordings annotated by at least four experienced sleep researchers in a hold-one-out fashion, while randomly permuting an increasing proportion of the consensus state annotations. This revealed that Somnotate is extremely robust to errors in the training data, as the accuracy on the test set only displayed a notable drop in performance when more than half of the training samples were misclassified (**Fig 3A and 3B**). Furthermore, classifier performance monotonically increased with increasing amounts of training data (**S2 Fig**), consistent with the idea that the classifier does not overfit the training data and, as a result, does not learn patterns that are present due to chance.

Second, classifiers are susceptible to being fine-tuned to a standard training data set, such that performance levels can drop when faced with test data collected under different conditions, particularly when features used by the classifier are altered in a consistent manner. To assess robustness to changes in features, we tested Somnotate's performance on data from a sleep deprivation experiment. Sleep deprivation is a common experimental manipulation that is known to change the EEG power spectrum after sleep onset, and EEG power values are primary features used by Somnotate. We evaluated the accuracy of our pre-trained classifier on six 3-hour data sets recorded after sleep onset in sleep-deprived mice, and compared this to the accuracy on matched 'baseline' (i.e. without sleep deprivation) data recorded from the same animals (**Fig 3C and 3D**). The annotation accuracy was found to be comparable and high across both the sleep deprivation and baseline conditions, consistent with Somnotate being robust to experimentally-induced changes in relevant features (**Fig 3D**).

Third, Somnotate uses contextual information in the form of prior probabilities of the different vigilance states. These probabilities depend on how much time animals spend in each state and how frequently they transition between states, both of which can change under experimental conditions. To assess the classifier's robustness to variations in these prior probabilities, we trained and tested Somnotate in a hold-one-out fashion on the six 24-hour EEG and EMG recordings with high quality consensus annotations, as before. We then evaluated Somnotate's accuracy on data sets from mice that had experienced repeated, experimentally-induced awakenings throughout the day. These awakenings were achieved via optogenetic stimulation of channelrhodopsin-2 (ChR2) expressing inhibitory neurons in the lateral

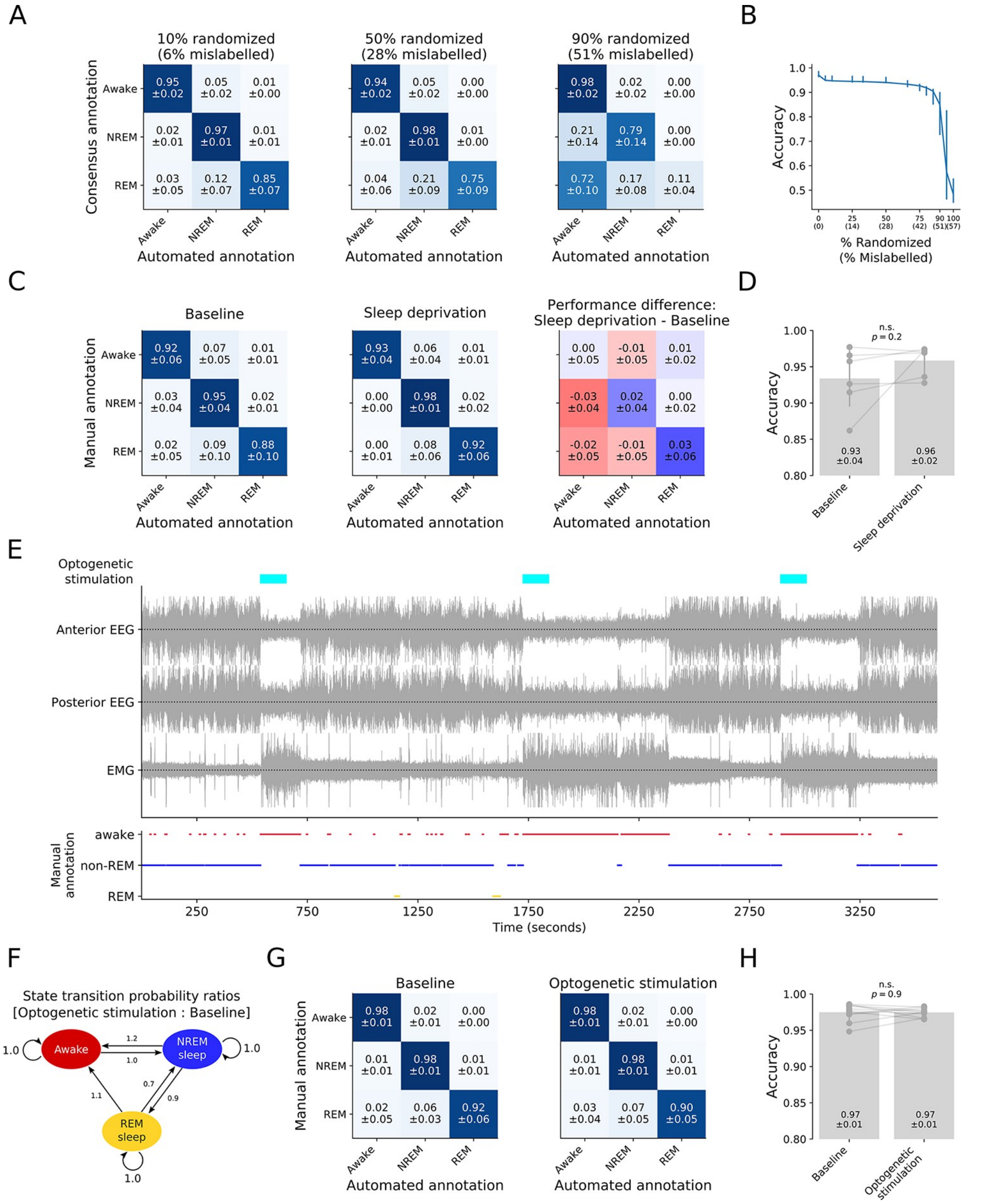

Fig 3. Somnotate is robust to errors in the training data, changes in the features of the data, and changes in the vigilance state transition probabilities. (**A**) Somnotate's accuracy was evaluated on six 24-hour data sets, in a hold-one-out fashion, while permuting an increasing fraction of annotations in the training data. Confusion matrices show the results when permuting 10% of the training data annotations (resulting in 6% mislabelled time points; left), permuting 50% of the training data annotations (resulting in 28% mislabelled time points; middle), or permuting 90% of the training data annotations (resulting in 51% mislabelled time points; right). Values represent mean ± standard deviation. (**B**) Somnotate's accuracy as a function of the percentage of permuted training data annotations. (**C**) The accuracy of Somnotate, pre-trained on 24h standard sleep-wake cycle datasets, was evaluated against a manual annotation of baseline control data (six 12-hour light cycle-only data sets), and compared to its accuracy on data from the same animals after undergoing a sleep deprivation protocol (six 12-hour light cycle-only data sets). Confusion matrices are shown for the baseline (left), following sleep deprivation (middle), and as the difference between these two confusion matrices (right). (**D**) Comparison of Somnotate's overall accuracy on baseline data and data collected after sleep deprivation. P-value is derived from a Wilcoxon signed rank test. (**E**) Mice experienced experimentally-induced awakenings via optogenetic stimulation of ChR2-expressing inhibitory neurons in the lateral preoptic hypothalamus. (**F**) State transition probabilities during optogenetic stimulation, normalised to the state transition probabilities during the baseline condition. The optogenetic manipulation increased the probability that the animals transitioned from NREM sleep and REM sleep, to the awake state. (**G**) The accuracy of Somnotate, pre-trained on 24h standard sleep-wake cycle datasets, was evaluated against a manual annotation for eleven 24-hour data sets recorded during

optogenetic stimulation, and for baseline recordings from the same animals on days when optogenetic stimulation was not performed. Confusion matrices are shown for the baseline recordings (left) and the recordings with optogenetic stimulation (right). (**H**) The accuracy of Somnotate's annotations was near identical for both data sets and not significantly different ($p > 0.9$, Wilcoxon signed rank test).

https://doi.org/10.1371/journal.pcbi.1011793.g003

preoptic hypothalamus (**Fig 3E**; as described in [61]), which affected the probabilities with which the animals transitioned between states (**Fig 3F**), but did not affect the overall time spent in each state ($p > 0.15$ for all states, Wilcoxon signed rank test). Despite the changes in transition probabilities, there was no detectable change in Somnotate's performance (**Fig 3G and 3H**).

Finally, although most electrophysiological signals show a dependence on vigilance state, some signals are thought to be more informative of certain states. For example, REM sleep is indicated by a high power in the theta frequency band of an EEG recording, which is typically more apparent for a posterior electrode than an anterior electrode [62]. However, due to experimental requirements and technical limitations, it is not always possible to record the 'ideal' combination of signals for the inference of vigilance states. We were hence keen to assess Somnotate's robustness to changes in recording configurations, and trained and tested Somnotate in a hold-one-out fashion on the six 24-hour recordings with high quality consensus annotations, while using only data from a single EEG, EMG, or LFP electrode. Our tests revealed that a single EEG signal was sufficient to infer the vigilance state with high accuracy (**S3 Fig**).

Taken together, these observations demonstrate the robustness of Somnotate to changes in the features used for inference, the state frequencies, and the state transition probabilities. Somnotate is therefore able to perform highly accurate sleep stage annotation under a variety of experimental conditions.

Somnotate identifies failed state transitions

HMMs belong to the category of Bayes classifiers that compute the likelihood of each state, for every data sample (i.e. time point). This means that Somnotate is able to distinguish samples where it is certain in its prediction (i.e. where the likelihood of the predicted state is effectively one), and samples where it is uncertain (i.e. where the likelihood of the predicted state is less than one). For our data sets, the cumulative distribution of likelihood values indicated a change point at 0.995 (**S4 Fig**), with a minority of samples (5.5%) having a likelihood below this threshold. Notably, for samples where Somnotate was uncertain, almost half (44%) coincided with instances where manual annotations disagreed with one another. This suggested that the difficulty in predicting the vigilance state at these time points was not an artefact of the inference method, but a result of ambiguity within the signals. Human annotators often exclude such sections of the data from their analysis and, by analogy, the accuracy of our automated classifier further increased when these ambiguous samples were excluded **(Fig 2G)**. By reporting the certainty of its predictions, Somnotate greatly facilitates the identification of ambiguous samples that the experimentalist may wish to review or examine further. Somnotate also offers the chance to characterise such ambiguous samples in a principled way.

An ambiguous sample could result from measurement noise masking a true, unambiguous signal, or it could stem from a mixed signal reflecting an intermediate state. Four lines of evidence support the idea that ambiguous samples reflect an intermediate state and are not a measurement artefact. First, the majority of ambiguous samples (76%) occur around state transitions, which represent relatively rare events in the recordings (first two examples in **Fig 4A**). Second, for nearly all ambiguous samples, the probability mass was concentrated in two states, rather than being randomly distributed across all three states (**Fig 4B**). Third, the power

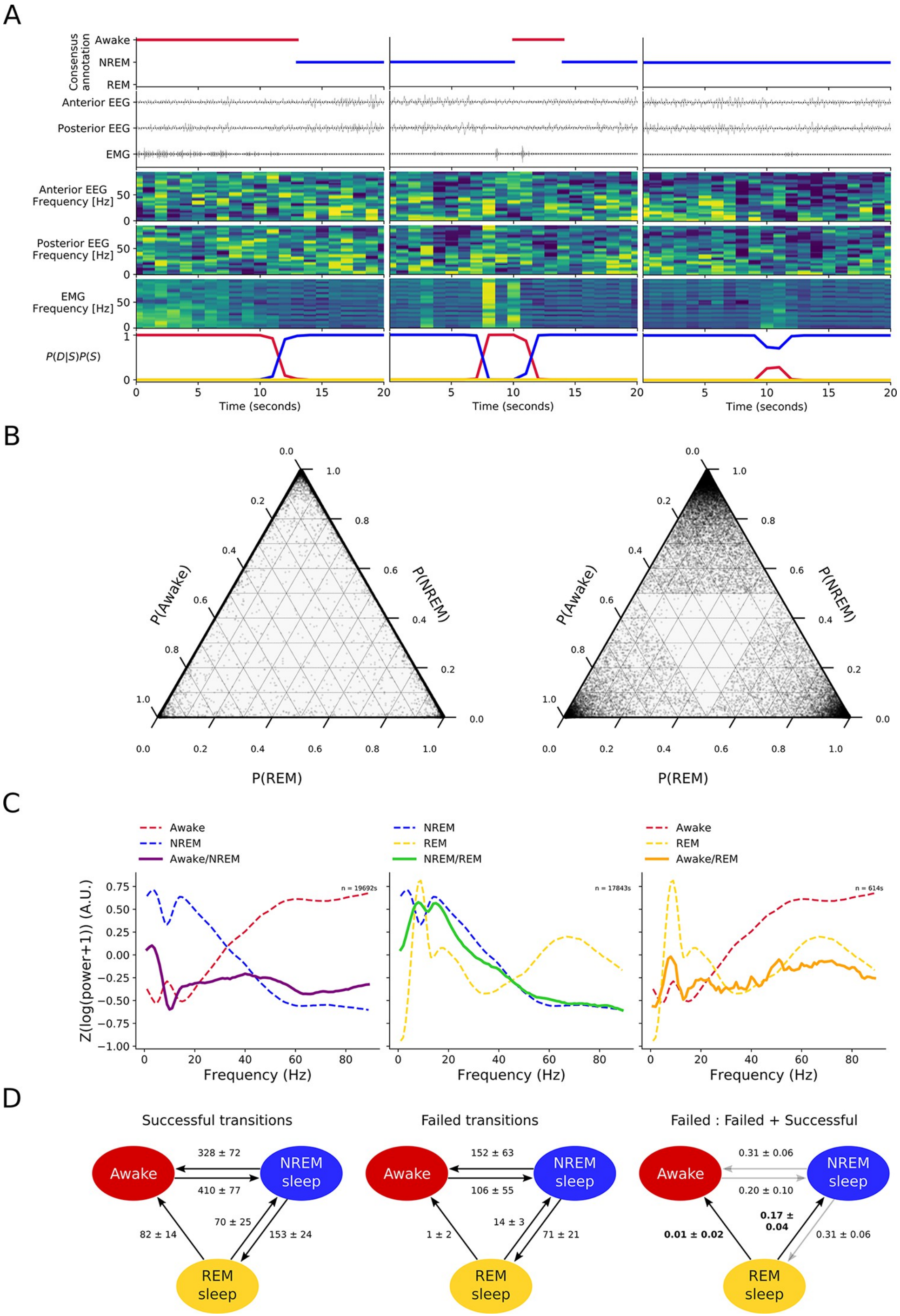

Fig 4. Somnotate identifies intermediate states associated with successful and failed vigilance state transitions. (**A**) Three examples of intermediate states identified by Somnotate, in which the probability of the most likely state dropped below 0.995. In each case, the consensus annotation, input signals, power spectra and likelihood of each state assigned by Somnotate, are shown. The first example (left) shows a successful state transition from awake to NREM sleep. Just before the transition, Somnotate identifies time points with intermediate states in which the probability of being awake has decreased and NREM sleep has increased. The second example (middle) shows a brief state transition from NREM sleep to awake, and then back to NREM sleep, which includes time points with intermediate states. The third example (right) shows a failed transition from NREM sleep to awake, which includes a series of time points with intermediate states in which there is a partial decrease in the probability of NREM sleep and partial increase in the probability of being awake. (**B**) Ternary plot of the state probabilities assigned to each time point with an intermediate

state in six 24-hour data sets (left). In the vast majority of cases, the probability mass was concentrated in one or two states. This was different to a theoretical distribution in which the probability mass outside the most likely state was randomly assigned to the other two states (right). (**C**) Power spectra extracted for time points with intermediate states (solid lines). For reference, the power spectra for the "pure" states are also shown (dashed lines). (**D**) Relative frequencies of successful state transitions (per day; left), failed state transitions (middle) and the ratio between these (right). Values indicate mean ± standard deviation. The failure rates of REM transitions were statistically significantly different from one another ($p < 0.001$; χ^2 contingency test), and hence indicated with black arrows; the failure rates of NREM transitions were not significantly different ($p > 0.05$; χ^2 contingency test), and hence indicated by grey arrows.

https://doi.org/10.1371/journal.pcbi.1011793.g004

spectra for ambiguous samples showed elements of the power spectra of the two most likely states (**Fig 4C**). For example, samples that Somnotate was uncertain whether to assign as awake or NREM sleep showed a high power in the δ frequency band, characteristic of NREM sleep, but also high power in the γ frequency band, which is typically an indicator of the awake state (**Fig 4C**). Fourth, the periods during which the classifier was uncertain tended to be much longer than the duration of a single sample and up to a maximum of 30 seconds, such that these intermediate states could not simply reflect the temporal resolution of the sampling (i.e. the result of a state transition occurring during a single one second sample).

To demonstrate the opportunities afforded by Somnotate's identification of intermediate states, we focused upon the 24% of ambiguous samples that were associated with an incomplete state transition (such as the third example in **Fig 4A**) and referred to these as 'failed transitions', to distinguish them from successful state transitions. We computed the frequencies of these different transition types and expressed the failed transitions as a proportion of all transitions (**Fig 4D**). This revealed that the probability of failing to transition was not random. The overall probability of failed transitions was higher when moving out of NREM sleep, than when moving out of REM sleep ($p < 0.001$, χ^2 contingency test). Furthermore, whereas the two state transitions out of NREM sleep exhibited a similar probability of failing ($p = 0.98$, χ^2 contingency test), the state transitions out of REM sleep differed, with a transition from REM-to-NREM showing a higher probability of failing than a transition from REM-to-awake (17% versus 1%; $p < 0.001$, χ^2 contingency test). The same pattern of failed transitions remained when a more conservative threshold criterion was adopted such as a state probability below 0.95. These observations may explain why animals often appear to enter a brief awake state between REM sleep and NREM sleep [51–54], as transitions from REM to awake, and from awake to NREM, fail much more rarely than transitions from REM to NREM.

Finally, we investigated if the state transition failure rates were sensitive to an animal's recent sleep-wake history. The recovery after sleep deprivation is known to be dominated by long periods of NREM and REM sleep (**Fig 5A** & [2,51]). We asked whether this change in state occupancy is also associated with a change in the failure rate of state transitions. Indeed, we observed an increase in the failure rate of transitions out of NREM, as well as an increase in the failure rate of REM-to-NREM transitions, increasing the likelihood of remaining in either sleep state for a longer period of time ($p < 0.05$, Mann-Whitney rank test, **Fig 5B**). In contrast, there was no difference in the failure rate for awake-to-NREM transitions. Together, these changes in the failure rate of the different state transitions are consistent with the observed vigilance state occupancy following sleep deprivation. More generally, these analyses confirm the additional opportunities afforded by Somnotate's ability to identify intermediate states and to provide a more complete and richer characterisation of sleep-wake dynamics.

Discussion

We present Somnotate—a purpose-built probabilistic sleep stage classifier that quantifies the certainty of its annotations. This enables the experimenter to assign vigilance states (awake,

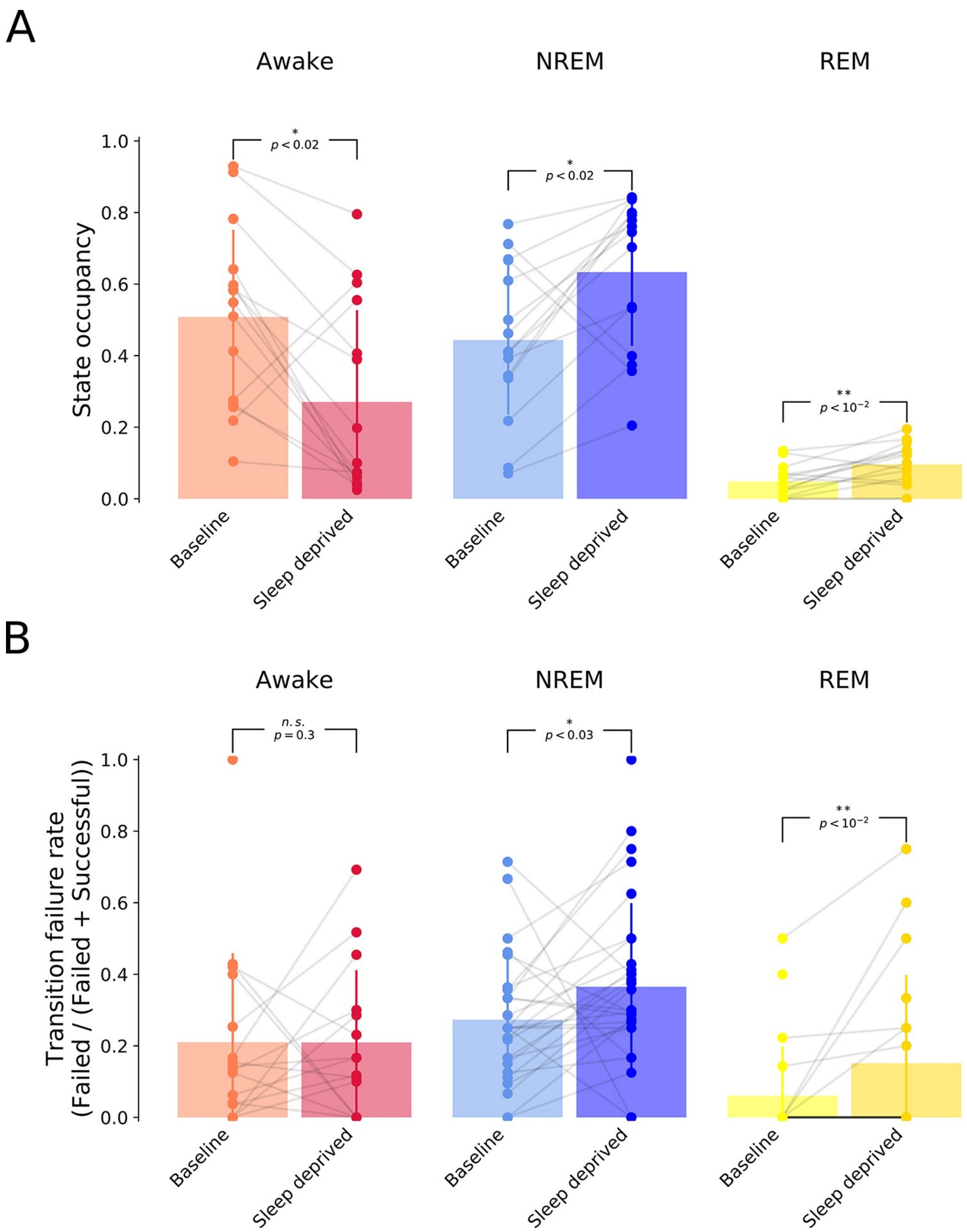

Fig 5. Vigilance state transition failure rates depend upon sleep-wake history. (**A**) State occupancy during the first two hours of recovery from a period of sleep deprivation versus baseline recordings performed in the same animals over an equivalent period but not following sleep deprivation. (**B**) Failure rates for transitions during the awake state (i.e. awake-to-NREM transitions; left), transitions during NREM (NREM-to-awake and NREM-to-REM transitions; middle), and transitions during REM (REM-to-awake and REM-to-NREM transitions; right). P-values are derived from a Wilcoxon signed rank tests with a Bonferroni-Holm correction for multiple comparisons.

https://doi.org/10.1371/journal.pcbi.1011793.g005

REM sleep, and NREM sleep) with high accuracy, whilst also identifying ambiguous epochs that represent intermediate states. Somnotate combines optimal feature extraction by linear discriminant analysis (LDA), with state-dependent contextual information derived via a hidden Markov model (HMM). Our benchmarking tests demonstrate that this approach optimises the use of contextual information, outperforming other classifiers that are compatible with data collected under experimental conditions. Furthermore, through systematic comparisons against expert manual annotations, we demonstrate that Somnotate surpasses the accuracy achieved by experienced sleep researchers. Somnotate is shown to be robust to errors in the training data, able to operate across different types of experimental manipulations, and compatible with different electrophysiological signals.

Somnotate's ability to quantify the certainty of its predictions creates new opportunities for the researcher, by systematically identifying epochs in which the vigilance state is ambiguous. With its dual output at every time point–a prediction and a certainty for each vigilance state–Somnotate offers users the chance to rapidly identify and confirm the classifier's performance in a targeted manner. Furthermore, this feature affords new analysis opportunities that are only feasible with an automated approach. Whilst the gold standard for sleep stage classification remains human experts, there is an element of subjectivity to all manual annotations that makes areas of investigation difficult. For example, EEG signals often show signatures of multiple states, particularly around state transitions [38,63–65], where most disagreements between manual annotations occur. And although humans are very good at determining the most likely state at a given time point, they struggle to quantify intermediate states [66].

To demonstrate the opportunities that result from this feature, we used Somnotate to identify intermediate states associated with incomplete state transitions, which we define as failed transitions. Interestingly, the probability of failed transitions was highly non-random and depended upon the type of transition, with REM-to-awake transitions showing relatively low failure rates, but REM-to-NREM transitions showing high failures. This differential failure rate may explain the preponderance of brief awake periods between REM and NREM sleep, as it may be easier for the underlying neuronal networks to transition from REM to awake, and then to NREM, rather than to transition directly from REM to NREM [51–54]. In addition, we revealed that the animal's recent sleep-wake history affects the probability of observing different types of failed transitions. More specifically, recovery sleep following a period of sleep deprivation exhibited an increase in the probability of failed transitions from NREM to REM, and from NREM to awake. These observations suggest that the underlying neuronal networks find it more difficult to transition out of the NREM state following sleep deprivation, which is consistent with the well-established finding that sleep deprivation leads to longer and more intense periods of NREM sleep [2,51]. These applications highlight the power of Somnotate's ability to determine intermediate states and reveal new opportunities to investigate the neurophysiological mechanisms of state transitions.

Despite previous efforts to develop algorithms for sleep stage classification, many sleep researchers continue to manually score their data. We believe that several barriers have prevented widespread adoption of automated solutions and our findings allow us to evaluate Somnotate in terms of these potential stumbling blocks. A first barrier to the widespread adoption of automated solutions is the issue of accessibility. Commercially available software for automated sleep stage classification can be expensive and details of their operation are often not made available to the user [33,67]. We provide an open source implementation of our algorithm written in Python. The code comes with extensive documentation including detailed installation instructions and a comprehensive tutorial. The modules of the code base can be integrated into an existing workflow. Alternatively, we also provide a fully-fledged pipeline as a standalone command line application.

In terms of performance levels, reports of human-like performance by automated methods may fall short in practice. Our findings suggest that the choice of performance metric may have contributed to this, as we show that inter-rater agreement can be an imprecise measure of annotation accuracy and is typically used in a manner that favours automated annotation. To improve upon this standard, we evaluated the quality of automated annotations against the consensus of at least three manual annotations. An annotation based on the consensus by majority vote will be more accurate than any individual annotation, whenever manual errors show some degree of independence from one another [68,69]. Using this improved assessment of performance, we showed that Somnotate matched the consensus more closely than any individual manual annotation. What was particularly encouraging was that the more manual annotations that were used to generate the consensus sequence, the more closely this consensus matched the automated annotation by Somnotate.

Human experts are aware that they continually use contextual information whilst interrogating time series data, relating information at a time point of interest, with information that they infer over longer timescales. This is often overlooked in automated classifiers, although a subset have used algorithms that incorporate contextual information, including those that have used HMMs [7,8,49] and recurrent or convolutional neural networks [18,19,21–24]. The inherent flexibility of neural networks can pose problems, as they require large amounts of labelled data to train [50], are prone to overfitting, and adapting their architecture to different inputs can be non-trivial. For these reasons neural networks can be a suboptimal choice in an animal research setting, where changes to the experimental arrangement are common and generating the large, well-curated training data sets required by neural networks, is usually impractical [50,70]. As HMMs are easier to optimise by non-experts, place significantly less demands on training data, and are thus more compatible with the constraints of experimental settings, we concentrated our efforts on improving the state-of-the-art for HMM-based inference of vigilance states. Somnotate represents an advance upon previous work on related classifiers, by first using LDA to automatically extract features that carry the maximum amount of linearly decodable information about vigilance states, and then incorporating state-dependent contextual information through the application of a HMM.

Another concern with automated approaches is how annotation is affected by differences in the training data or the features used for classification. We found that Somnotate is remarkably robust to errors in the training data, with test performance only dropping significantly when more than half of the training samples had been deliberately misclassified. Furthermore, automated scoring methods can exhibit overfitting to standard or control data sets, which means that their performance is diminished in other settings when the probabilities of key features vary [21,24,26,71,72]. This was not the case with Somnotate. We saw no drop in performance when Somnotate annotated data collected under different experimental conditions in which the features used for inference, the state transition probabilities, and/or the state frequencies varied. Somnotate's high performance was maintained on data from sleep-deprived mice in which the EEG spectrogram is significantly altered, and on data from optogenetic manipulation experiments in which the state transition probabilities are changed. These applications establish that Somnotate is well-suited to perform accurate sleep stage annotation under a variety of experimental conditions.

In terms of adaptability, the use of targeted feature extraction via LDA means that Somnotate is agnostic with respect to the exact nature of the input signal, and our data suggests that Somnotate retains a high accuracy with different EEG recording configurations and even other data sources such as local field potential (LFP) recordings. In principle, Somnotate could be applied to any high frequency time series data that contains information about an animal's vigilance state, and hence we plan to expand this approach to other types of signals, such as surface EEG, actigraphy, or respiratory activity [31,72–74]. While some automated approaches depend on a specific input signal [32], others are also adaptable in the sense that they can be recalibrated to changes in the experimental setup [17,27,29–31,48]. However, as these automated approaches tend to have several free parameters, adaptation to a different setup can be time-consuming, with uncertain returns. As there are no free parameters in Somnotate other than the desired time resolution of the state prediction, re-training requires no optimisation, and is straightforward and fast. Training Somnotate takes approximately one second per 24 hours of data on a standard desktop computer.

In summary, Somnotate's development as a probabilistic sleep stage classifier affords new biological insights through the identification and characterization of intermediate states in polysomnography, whilst simultaneously setting new standards in terms of performance, robustness, ease of use, and accessibility.

Materials and methods

Ethics statement

All electrophysiological recordings were performed in accordance with the United Kingdom Animal Scientific Procedures Act 1986 under personal and project licences granted by the United Kingdom Home Office. Ethical approval for the animal experimentation was granted by the Animal Welfare and Ethical Review Body at the University of Oxford.

Animal husbandry and sleep deprivation

All experiments were performed on adult male C57BL/6 wild-type mice, which were bred, housed and used in accordance with the UK Animals (Scientific Procedures) Act (1986). The data collected were part of projects examining the role of cortex in sleep-wake regulation [75] and sleep homeostasis [61,76]. Animals were maintained under a 12-hour:12-hour light-dark (LD) cycle. For the subset of animals that underwent a sleep deprivation (SD) protocol, each animal was pre-exposed to novel objects to encourage exploratory behaviour. The SD protocol then consisted of delivering novel objects for the first three (Fig 3) or for the first six hours (Fig 5) of the light cycle, under the continuous observation of an experimenter. Once an animal had stopped exploring an object, a new object was presented.

Surgical procedures and electrode configuration

For chronic electroencephalogram (EEG) and electromyogram (EMG) recordings, custom-made headstages were constructed by connecting three stainless steel screw electrodes (Fine Science Tools), and two stainless steel wires, to an 8-pin surface mount connector (8415-SM, Pinnacle Technology Inc., Kansas). For LFP recordings, a 16-channel silicon probe (NeuroNexus Technologies Inc., Ann Arbor, MI, USA; model: A1x16- 3mm-100-703-Z16) with a spacing of 100 μm between individual channels was used. Device implantation was performed using stereotactic surgery, aseptic technique, isoflurane anaesthesia (3–5% for induction and 1–2% for maintenance) and constant body temperature monitoring. Analgesia was provided at the beginning of surgery and during recovery (buprenorphine and meloxicam). A craniotomy was performed over the right frontal cortex (AP +2 mm, ML +2 mm from Bregma), right occipital cortex (AP +3.5 mm, ML +2.5 mm from Bregma), and the cerebellum (-1.5 mm posterior from Lambda, ML 0). A subset of animals were further implanted with a bipolar concentric electrode (PlasticsOne Inc., Roanoke, VA, USA) in the right primary motor cortex, anterior to the frontal EEG screw. To accommodate this additional implant, the frontal EEG screw was typically implanted 0.2–1.6 mm posterior to the target coordinates. For EEG recordings, a screw was fixed over both the right frontal and occipital cortex. For LFP and multi-unit activity recording in a subset of animals, a 16-channel silicon probe was implanted into primary motor cortex (+1.1 mm AP (anterior), -1.75 mm ML (left), tilt -15° (left)) under microscopic control, as reported previously [75]. EEG and LFP signals were referenced to a cerebellum screw. For EMG recordings, wire electrodes were inserted into the left and right neck muscles, and one signal acted as reference to the other. All implants were secured using a non-transparent dental cement (SuperBond from Prestige Dental Products Ltd, Bradford, UK). Animals were allowed to recover for at least 1 week before recordings.

In vivo data acquisition

Animals were moved to a recording chamber and housed individually in a Plexiglas cage (20.3 x 32 x 35 cm). Recordings were performed using a 128-channel Neurophysiology Recording System (Tucker-Davis Technologies Inc., Alachua, FL, USA), acquired using the electrophysiological recording software, Synapse (Tucker-Davis Technologies Inc., Alachua, FL, USA), and stored locally for offline analysis. EEG, EMG, and LFP signals were continuously recorded, filtered between 0.1–100 Hz, and stored at a sampling rate of 305 Hz. EEG, EMG and LFP signals were resampled at 256 Hz using custom code in MATLAB (MathWorks, v2017a), and converted into the European Data Format. The first and/or last 30 seconds of recordings could contain missing values as this corresponded to the period when the electrodes were being connected/disconnected from the recording system. These epochs were excluded from all subsequent analyses.

Optogenetic stimulation

We employed a protocol previously described in detail in [77,42]. Briefly, channelrhodopsin-2 (ChR2) was expressed in glutamate decarboxylase 2 expressing (GAD2^{+}) interneurons by injection of an adenovirus construct (UNC vector core, AAV5-EF1a-DIO-ChR2-eYFP) into the lateral preoptic area (LPO) of the hypothalamus in adult Gad2-IRES-Cre mice (Jackson Laboratory 019022; B6N.Cg-Gad2$^{tm2(cre)Zjh}$/J). For optical stimulation, either an optic fiber (400 μm diameter, Doric Lenses Inc, Quebec) or a custom made optrode, consisting of an optic fiber glued with tungsten wires, was inserted to 0.2 mm above the virus injection site. All electrophysiological recordings were made 4 to 7 weeks post virus injection. To optogenetically stimulate GAD2^{+} neurons, we applied 10 ms pulses of light from a blue LED (470 nm, 10.8–13.2mW at fiber tip) at various frequencies (20, 10, 5, 2 or 1 Hz), for a duration of 2 minutes, every 20 ± 2 minutes. On baseline days, no optogenetic stimulation was provided.

Manual vigilance state annotation

Manual annotation of vigilance states was performed offline, based on 4 s epochs using Sleep-Sign software (Kissei Comtec). The anterior EEG channel, the posterior EEG channel, and the EMG channel were displayed on-screen simultaneously and visually inspected for vigilance state scoring. Three vigilance states were identified, as is typical in laboratory rodent studies. Waking was defined by a low-voltage, high-frequency EEG signal, with a high level or phasic EMG activity. During active, exploratory waking, a transient increase in theta-activity (5–10 Hz) was typically observed in the occipital derivation, overlying the hippocampus. NREM sleep was defined by an overall higher amplitude signal, dominated by slow waves (<4 Hz) and spindle oscillations (10–15 Hz) that were especially prominent in the anterior EEG channel, while the EMG signal was typically low. REM sleep was characterised by low-voltage, high-frequency EEG, dominated by theta activity especially in the posterior EEG channel, with a low level of EMG activity.

Data pre-processing for automated annotation

We first computed the spectrograms of the anterior EEG, the posterior EEG, and the EMG traces. To reduce sensitivity to noise present in electrophysiological recordings, we used a multitaper approach, as this results in more robust estimates of the power than the more conventional Baum-Welch algorithm. Specifically, we used the implementation in the lspopt python library (1 second long segments with no overlap, other parameters at default values). We then discarded parts of the power spectrum that are strongly influenced by signals not related to changes in vigilance states. We discarded signals in the 0–0.5 Hz frequency range in the EEG and EMG recordings, as these are dominated by drift due to animal locomotion. Furthermore, we discarded signals between 45–55 Hz and above 90 Hz, as these were strongly affected by 50 Hz electrical noise. We then applied a log(x+1) transformation to map the heavy-tailed distribution of power values to a distribution that is more normally distributed. The normal distribution is the maximum entropy distribution for continuous distributions on unbounded domains, and as such, samples are maximally far apart from one another (compared to other distributions with the same variance). This facilitates downstream classification into separable groups. The re-mapped power values were then normalised by converting them to Z-scores (mean subtraction followed by rescaling to unit variance). Normalisation ensures that all frequencies are weighted equally in the downstream feature extraction. Finally, the normalised spectrograms were concatenated, resulting in a high-dimensional signal.

Automated feature extraction

Features for downstream classification were then extracted from the concatenated spectrograms in a targeted manner using linear discriminant analysis (LDA [78]), as implemented in the scikit-learn python library [79]. LDA determines a linear projection of high dimensional data to a low dimensional representation, such that samples belonging to different classes are optimally linearly separated in the low dimensional space. Thus, information in the signal about the different classes is preserved, while non-informative components of the signal are discarded. This has two further effects. Firstly, training of any classifier is accelerated, which implicitly or explicitly fits a joint probability distribution to the components of the training data. The number of samples required to accurately fit a joint probability distribution increases exponentially with the number of dimensions. As the dimensionality of the data is reduced, fewer samples are required to escape the under-sampled regime and accurately determine the shape of the data distribution. This is enhanced by the fact that the components of the LDA are largely independent of one another–unlike the original signal, in which many frequencies are highly correlated with each other. Secondly, as much of the original signal is effectively discarded, artefacts that contaminate the signal are also removed.

Automated vigilance state annotation

Given three target states (awake, NREM sleep, and REM sleep), dimensionality reduction with LDA results in two-dimensional signals. These two-dimensional signals together with the corresponding manual annotations were used to train a HMM in a supervised fashion, with multivariate Gaussian state emissions using the python library pomegranate with all optional parameters at default values [80]. Manual annotations occasionally contain mislabelled samples, which when parameterising the HMM result in non-zero probabilities for state transitions that are thought to be absent in wildtype/control data, notably awake-to-REM transitions in mice [49]. Somnotate facilitates pruning of these spurious transitions by exposing a threshold parameter, such that state transitions below this threshold are removed. For the purpose of this study, state transitions with probability below 0.0001 per second were removed. The accuracy of the trained LDA and HMM models were ascertained by applying them to held out test data, which had not been used to fit the linear discriminants or to parameterize the HMM. For each sample, the probability of each state was computed using the Baum-Welch algorithm, and the most likely state sequence was determined using the Viterbi algorithm. Unless specified otherwise, training and testing occurred in a hold-one-out fashion, i.e. the model was trained using all but one of the specified datasets and then tested on the remaining dataset. This process was repeated for all possible train-test splits of the data. SPINDLE and IntelliSleepScorer were both evaluated based on the predictions made by extensively pre-trained models provided by the authors, and with all other parameters at default values. As both models contain thousands of model parameters, we opted to evaluate them without re-training, as the limited number of data sets available to us with high-quality manual annotations based on the consensus of multiple experts would potentially have been insufficient to do so.

Recording artefacts

Samples containing artefacts associated with the animal's gross body movements were identified during manual annotations, but were still included in the analysis of vigilance states and in the data used to train Somnotate. Such artefacts represented 1.0% ± 1.0% of the consensus manual annotations (mean ± standard deviation; 3.8% ± 2.8% in the individual manual annotations) and did not influence the automated feature extraction by LDA, so did not impact the quality of the automated annotations. However, such artefacts could affect downstream analyses in future applications, such as spectral analysis of the recorded signals. For this reason, Somnotate includes two features to facilitate the detection and removal of artefacts. First, Somnotate detects and demarcates gross movements that generate voltage deflections outside of the dynamic range of the recording system (with an optional padding to also remove voltage deflections preceding and following such events), so that they are not included in downstream analyses. Second, Somnotate has the option to present samples to the user where the classifier was uncertain about state assignment. Intervals consisting of consecutive samples in which the probability of the inferred state is below one are scored according to the sum of the residual probabilities (i.e. one minus the probability of the inferred state) and presented to the user in descending order. Movement artefacts associated with prolonged voltage deflections, or that strongly affect the spectral features identified by LDA, result in a high score and can be excluded by the user.

Supporting information

S1 Appendix. Unbiased and precise assessment of automated and manual sleep annotation.
(DOCX)

S1 Table. Manual annotations were performed by experienced sleep researchers. All authors who provided manual annotations reported their task-relevant experience in years, plus the approximate number of hours that they had previously manually annotated.
(DOCX)

S2 Table. Performance of Somnotate and other state-of-the-art algorithms for automated mouse polysomnography compared to manual annotation by experienced experts. Manual and automated annotations of six 24 hour datasets were evaluated based on the consensus of multiple expert annotations.
(DOCX)

S1 Fig. EEG power spectra by state according to Somnotate (A) or manual consensus annotations (B). Somnotate was trained and tested, in a hold-one-out fashion, on six 24-hour data sets. Spectrograms were computed for the anterior and posterior EEG and partitioned according to the predicted state. The process was repeated using the manual consensus annotations.

The lines indicate the median EEG power.
(EPS)

S2 Fig. Somnotate's performance as a function of the amount of training data. Somnotate was trained and tested, in a hold-one-out fashion, using different numbers of 24-hour data sets. The maximum amount of training data available was twenty 24-hour EEG and EMG recordings under baseline conditions. The line indicates the median. Error bars demarcate the 5th and 95th percentile.
(EPS)

S3 Fig. A single EEG signal is sufficient for Somnotate to infer vigilance states with high accuracy. (**A**) The accuracy of Somnotate's sleep stage classification using a single input signal. Somnotate was trained and tested, in a hold-one-out fashion, on six 24-hour data sets. Only one signal was provided as an input signal: either the anterior EEG, the posterior EEG, the LFP from primary somatosensory cortex, or the EMG. (**B**) Confusion matrices when using only the anterior EEG (top left), the posterior EEG (top right), an LFP (bottom left), or the EMG (bottom right). Values indicate mean ± standard deviation. The overall accuracy of the predictions based on the anterior EEG alone did not differ from the overall accuracy when the anterior EEG, posterior EEG and EMG were provided simultaneously ($p > 0.24$, Wilcoxon signed rank test), as the small increase in the false negative detection rate for REM sleep was offset by an improved distinction between the awake and two sleep states. The overall accuracy of the predictions based on the posterior EEG was 1% lower on average ($p < 0.05$, Wilcoxon signed rank test), although in this case the identification of REM showed a similar accuracy to when all signals were provided. The accuracy of predictions based on either an LFP recorded from primary somatosensory cortex, or only the EMG signal, was in both cases worse (by 6% and 14%, respectively), largely due to the performance on REM sleep episodes ($p < 0.05$ in both cases, Wilcoxon signed rank test). However, each individual signal was still sufficient to distinguish between awake and asleep states with high accuracy (~95%), indicating that either signal would be sufficient in experiments that do not need to distinguish between REM and NREM sleep states.
(EPS)

S4 Fig. Selecting a state probability threshold to identify ambiguous samples. The probability of the predicted state was computed for each sample in six 24-hour data sets. As the cumulative distribution of probabilities exhibits an elbow at 0.995, this value was chosen as a threshold below which samples were classified as ambiguous.
(EPS)

S5 Fig. The consensus of manual annotations yields a better estimate of annotation accuracy. (**A**) The annotation of vigilance states was based on recordings of the anterior EEG, posterior EEG and EMG from a freely behaving mouse. A one-minute segment of the recordings is shown. (**B**) Multi-taper spectrograms for each of the recorded signals in 'A'. (**C**) The majority-vote consensus of manual annotations by three independent experienced sleep researchers (top), which discriminates the vigilance states of 'awake' (red), 'NREM' sleep (blue) and 'REM' sleep (yellow). A fourth (middle) and fifth (bottom) independent individual manual annotation of the same segment. (**D**) A total of ten experienced sleep researchers independently annotated the same 12-hour recording and the accuracy of each annotation was assessed by using the consensus of the other nine annotations as a proxy for the ground truth. (**E**) For each possible pair of annotations, the inter-rater agreement was plotted against the mean accuracy of the pair of annotations, when judged against a consensus based on the remaining other eight annotations. (**F**) There was greater variability in the accuracy of an annotation when

judged against a single other manual annotation, than when judged against the consensus of three randomly selected annotations (without replacement). Plot shows the variability in accuracy estimates (standard deviation with Bessel correction), which was significantly lower when using the consensus of three annotations ($p < 0.01$, Wilcoxon signed rank test). (**G**) A consensus was constructed from five of the ten independent annotations based on majority vote. A second consensus annotation was then constructed using either one, three or all five of the remaining annotations. The plot shows the mean agreement between the two consensus annotations. Error bars represent the standard deviation.
(EPS)

S6 Fig. The consensus of manual annotations yields a better estimate of annotation accuracy, as measured by Cohen's kappa. The analyses in Figs 2A and S5D–S5G were repeated using Cohen's kappa as a measure of performance. (**A**) A total of ten experienced sleep researchers independently annotated the same 12-hour recording and the accuracy of each annotation was assessed by using the consensus of the other nine annotations as a proxy for the ground truth. (**B**) For each possible pair of annotations, the inter-rater agreement was plotted against the mean accuracy of the pair of annotations, when judged against a consensus based on the remaining other eight annotations. (**C**) There was greater variability in the accuracy of an annotation when judged against a single other manual annotation, than when judged against the consensus of three randomly selected annotations (without replacement). Plot shows the variability in accuracy estimates (standard deviation with Bessel correction), which was significantly lower when using the consensus of three annotations ($p < 0.01$, Wilcoxon signed rank test). (**D**) A consensus was constructed from five of the ten independent annotations based on majority vote. A second consensus annotation was then constructed using either one, three or all five of the remaining annotations. The plot shows the mean agreement between the two consensus annotations. Error bars represent standard deviation. (**E**) Somnotate was trained and tested, in a hold-one-out fashion, on six 24-hour data sets. Using a consensus annotation based on at least 3 manual annotations, the Cohen's kappa score of the automated annotation was compared to the Cohen's kappa score of individual manual annotations ($n = 25$ manual annotations from 13 experienced sleep researchers). P-value is derived from a Wilcoxon signed rank test.
(EPS)

Acknowledgments

We would like to thank the Akerman lab for advice and comments. We thank M. Andina, E. Tyler, and L. Kravitz for providing the mouse drawings.

Author Contributions

Conceptualization: Paul J. N. Brodersen, Colin J. Akerman.

Data curation: Paul J. N. Brodersen, Hannah Alfonsa, Lukas B. Krone, Cristina Blanco-Duque, Angus S. Fisk, Sarah J. Flaherty, Mathilde C. C. Guillaumin, Yi-Ge Huang, Martin C. Kahn, Laura E. McKillop, Linus Milinski, Lewis Taylor, Christopher W. Thomas, Tomoko Yamagata, Vladyslav V. Vyazovskiy.

Formal analysis: Paul J. N. Brodersen.

Funding acquisition: Russell G. Foster, Vladyslav V. Vyazovskiy, Colin J. Akerman.

Investigation: Hannah Alfonsa, Lukas B. Krone, Cristina Blanco-Duque, Tomoko Yamagata.

Methodology: Paul J. N. Brodersen.

Project administration: Lukas B. Krone, Vladyslav V. Vyazovskiy, Colin J. Akerman.

Software: Paul J. N. Brodersen.

Supervision: Russell G. Foster, Vladyslav V. Vyazovskiy, Colin J. Akerman.

Validation: Paul J. N. Brodersen.

Visualization: Paul J. N. Brodersen.

Writing – original draft: Paul J. N. Brodersen.

Writing – review & editing: Paul J. N. Brodersen, Colin J. Akerman.

References

1. Schwierin B, Borbély AA, Tobler I. Sleep homeostasis in the female rat during the estrous cycle. Brain Res. 1998; 811(1–2):96–104. https://doi.org/10.1016/s0006-8993(98)00991-3 PMID: 9804908
2. Leemburg S, Vyazovskiy V V., Olcese U, Bassetti CL, Tononi G, Cirelli C. Sleep homeostasis in the rat is preserved during chronic sleep restriction. Proc Natl Acad Sci U S A. 2010; 107(36):15939–44. https://doi.org/10.1073/pnas.1002570107 PMID: 20696898
3. Northeast RC, Huang Y, McKillop LE, Bechtold DA, Peirson SN, Piggins HD, et al. Sleep homeostasis during daytime food entrainment in mice. Sleep. 2019; 42(11):1–13. https://doi.org/10.1093/sleep/zsz157 PMID: 31329251
4. Funato H, Miyoshi C, Fujiyama T, Kanda T, Sato M, Wang Z, et al. Forward-genetics analysis of sleep in randomly mutagenized mice. Nature. 2016 Nov 2; 539(7629):378–83. https://doi.org/10.1038/nature20142 PMID: 27806374
5. Sun H, Jia J, Goparaju B, Bin Huang G, Sourina O, Bianchi MT, et al. Large-scale automated sleep staging. Sleep. 2017; 40(10). https://doi.org/10.1093/sleep/zsx139 PMID: 29029305
6. Martin WB, Johnson LC, Viglione SS, Naitoh P, Joseph RD, Moses JD. Pattern recognition of EEG-EOG as a technique for all-night sleep stage scoring. Electroencephalogr Clin Neurophysiol. 1972; 32 (4):417–27. https://doi.org/10.1016/0013-4694(72)90009-0 PMID: 4111497
7. Längkvist M, Karlsson L, Loutfi A. Sleep Stage Classification Using Unsupervised Feature Learning. Adv Artif Neural Syst. 2012; 2012:1–9.
8. JIANG D, LU Y nan, MA Y, WANG Y. Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement. Expert Syst Appl. 2019; 121:188–203.
9. Vallat R, Walker MP. An open-source, high-performance tool for automated sleep staging. Elife. 2021 Oct 14; 10:1–24. https://doi.org/10.7554/eLife.70092 PMID: 34648426
10. Olesen AN, Jørgen Jennum P, Mignot E, Sorensen HBD. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. Sleep. 2021; 44(1):1–12.
11. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. npj Digit Med. 2021; 4(1):1–12.
12. Şen B, Peker M, Çavuşoğlu A, Çelebi F V. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J Med Syst. 2014; 38(3). https://doi.org/10.1007/s10916-014-0018-0 PMID: 24609509
13. Radha M, Garcia-Molina G, Poel M, Tononi G. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. 2014 36th Annu Int Conf IEEE Eng Med Biol Soc EMBC 2014. 2014;1876–80. https://doi.org/10.1109/EMBC.2014.6943976 PMID: 25570344
14. Aboalayon KAI, Faezipour M, Almuhammadi WS, Moslehpour S. Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation. Entropy. 2016;18(9).
15. Boostani R, Karimzadeh F, Nami M. A comparative review on sleep stage classification methods in patients and healthy individuals. Comput Methods Programs Biomed. 2017; 140:77–91. https://doi.org/10.1016/j.cmpb.2016.12.004 PMID: 28254093
16. Fiorillo L, Puiatti A, Papandrea M, Ratti PL, Favaro P, Roth C, et al. Automated sleep scoring: A review of the latest approaches. Sleep Med Rev. 2019; 48:101204. https://doi.org/10.1016/j.smrv.2019.07.007 PMID: 31491655

17. Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P, Provazník I. Sleep scoring using artificial neural networks. Sleep Med Rev. 2012; 16(3):251–63. https://doi.org/10.1016/j.smrv.2011.06.003 PMID: 22030383

18. Yulita IN, Fanany MI, Arymurthy AM. Combining deep belief networks and bidirectional long short-term memory case study: Sleep stage classification. Int Conf Electr Eng Comput Sci Informatics. 2017;2017-Decem(September):19–21.

19. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. IEEE Trans Neural Syst Rehabil Eng. 2018; 26(4):758–69. https://doi.org/10.1109/TNSRE.2018.2813138 PMID: 29641380

20. Li X, Cui L, Tao S, Chen J, Zhang X, Zhang GQ. HyCLASSS: A Hybrid Classifier for Automatic Sleep Stage Scoring. IEEE J Biomed Heal Informatics. 2018; 22(2):375–85. https://doi.org/10.1109/JBHI.2017.2668993 PMID: 28222004

21. Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, et al. Automatic human sleep stage scoring using deep neural networks. Front Neurosci. 2018; 12(NOV):1–15. https://doi.org/10.3389/fnins.2018.00781 PMID: 30459544

22. Phan H, Andreotti F, Cooray N, Chen OY, Vos M De. Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS. 2018;2018-July:1452–5.

23. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. IEEE Trans Biomed Eng. 2019; 66(5):1285–96. https://doi.org/10.1109/TBME.2018.2872652 PMID: 30346277

24. Sun C, Fan J, Chen C, Li W, Chen W. A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation. IEEE Access. 2019; 7:109386–97.

25. Benington JH, Kodali SK, Heller HC. Scoring transitions to REM sleep in rats based on the EEG phenomena of pre-REM sleep: An improved analysis of sleep structure. Sleep. 1994; 17(1):28–36. https://doi.org/10.1093/sleep/17.1.28 PMID: 8191200

26. Veasey SC, Valladares O, Fenik P, Kapfhamer D, Sanford L, Benington J, et al. An automated system for recording and analysis of sleep in mice. Sleep. 2000; 23(8):1025–40. PMID: 11145318

27. Crisler S, Morrissey MJ, Anch AM, Barnett DW. Sleep-stage scoring in the rat using a support vector machine. J Neurosci Methods. 2008; 168(2):524–34. https://doi.org/10.1016/j.jneumeth.2007.10.027 PMID: 18093659

28. Stephenson R, Caron AM, Cassel DB, Kostela JC. Automated analysis of sleep-wake state in rats. J Neurosci Methods. 2009; 184(2):263–74. https://doi.org/10.1016/j.jneumeth.2009.08.014 PMID: 19703489

29. Brankačk J, Kukushka VI, Vyssotski AL, Draguhn A. EEG gamma frequency and sleep-wake scoring in mice: Comparing two types of supervised classifiers. Brain Res. 2010; 1322:59–71. https://doi.org/10.1016/j.brainres.2010.01.069 PMID: 20123089

30. Rytkönen KM, Zitting J, Porkka-Heiskanen T. Automated sleep scoring in rats and mice using the naive Bayes classifier. J Neurosci Methods. 2011; 202(1):60–4. https://doi.org/10.1016/j.jneumeth.2011.08.023 PMID: 21884727

31. Zeng T, Mott C, Mollicone D, Sanford LD. Automated determination of wakefulness and sleep in rats based on non-invasively acquired measures of movement and respiratory activity. J Neurosci Methods. 2012; 204(2):276–87. https://doi.org/10.1016/j.jneumeth.2011.12.001 PMID: 22178621

32. Lefort JM, Laville D, Id SB, Lacroix MM, Benchenane K. Harnessing olfactory bulb oscillations to perform fully brain-based sleep-scoring and real-time monitoring of anaesthesia depth. 2018. 1–32 p. https://doi.org/10.1371/journal.pbio.2005458 PMID: 30408025

33. Allocca G, Ma S, Martelli D, Cerri M, Del Vecchio F, Bastianini S, et al. Validation of 'somnivore', a machine learning algorithm for automated scoring and analysis of polysomnography data. Front Neurosci. 2019; 13(March):1–18. https://doi.org/10.3389/fnins.2019.00207 PMID: 30936820

34. Van Gorp H, Huijben IAM, Fonseca P, Van Sloun RJG, Overeem S, Van Gilst MM. Certainty about uncertainty in sleep staging: A theoretical framework. Sleep. 2022; 45(8):1–8. https://doi.org/10.1093/sleep/zsac134 PMID: 35675746

35. Nir Y, Staba RJ, Andrillon T, Vyazovskiy V V., Cirelli C, Fried I, et al. Regional Slow Waves and Spindles in Human Sleep. Neuron. 2011 Apr; 70(1):153–69. https://doi.org/10.1016/j.neuron.2011.02.043 PMID: 21482364

36. Bernardi G, Betta M, Ricciardi E, Pietrini P, Tononi G, Siclari F. Regional delta waves in human rapid eye movement sleep. J Neurosci. 2019; 39(14):2686–97. https://doi.org/10.1523/JNEUROSCI.2298-18.2019 PMID: 30737310

37. Vyazovskiy V V., Olcese U, Hanlon EC, Nir Y, Cirelli C, Tononi G. Local sleep in awake rats. Nature. 2011; 472(7344):443–7. https://doi.org/10.1038/nature10009 PMID: 21525926

38. Funk CM, Honjoh S, Rodriguez A V., Cirelli C, Tononi G. Local slow waves in superficial layers of primary cortical areas during REM sleep. Curr Biol. 2016; 26(3):396–403. https://doi.org/10.1016/j.cub.2015.11.062 PMID: 26804554

39. Soltani S, Chauvette S, Bukhtiyarova O, Lina JM, Dubé J, Seigneur J, et al. Sleep–Wake Cycle in Young and Older Mice. Front Syst Neurosci. 2019; 13(September):1–14. https://doi.org/10.3389/fnsys.2019.00051 PMID: 31611779

40. Iber C, Ancoli-Israel S, Chesson A, Quan S. The AASM Manual for Scoring of Sleep and Associated Events. AASM Manual for Scoring Sleep. 2007. p. 3–49.

41. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. J Clin Sleep Med. 2022; 18(1):193–202. https://doi.org/10.5664/jcsm.9538 PMID: 34310277

42. Fiorillo L, Pedroncelli D, Agostini V, Favaro P, Faraci FD. Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring. Sleep. 2023; 46(5):1–12. https://doi.org/10.1093/sleep/zsad028 PMID: 36762998

43. Bakker JP, Ross M, Cerny A, Vasko R, Shaw E, Kuna S, et al. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. Sleep. 2023; 46(2):1–12. https://doi.org/10.1093/sleep/zsac154 PMID: 35780449

44. Phan H, Mikkelsen K, Chen OY, Koch P, Mertins A, De Vos M. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. IEEE Trans Biomed Eng. 2022; 69 (8):2456–67. https://doi.org/10.1109/TBME.2022.3147187 PMID: 35100107

45. Fiorillo L, Favaro P, Faraci FD. DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model with Uncertainty Estimates. IEEE Trans Neural Syst Rehabil Eng. 2021; 29:2076–85. https://doi.org/10.1109/TNSRE.2021.3117970 PMID: 34648450

46. Fiorillo L, Monachino G, van der Meer J, Pesce M, Warncke JD, Schmidt MH, et al. U-Sleep's resilience to AASM guidelines. npj Digit Med. 2023; 6(1). https://doi.org/10.1038/s41746-023-00784-0 PMID: 36878957

47. Anderer P, Ross M, Cerny A, Vasko R, Shaw E, Fonseca P. Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing. Front Sleep. 2023;2(Version 3).

48. Lajnef T, Chaibi S, Ruby P, Aguera PE, Eichenlaub JB, Samet M, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods. 2015; 250:94–105. https://doi.org/10.1016/j.jneumeth.2015.01.022 PMID: 25629798

49. Miladinović D, Muheim C, Bauer S, Spinnler A, Noain D, Bandarabadi M, et al. SPINDLE: End-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species. PLoS Comput Biol. 2019; 15(4):1–30. https://doi.org/10.1371/journal.pcbi.1006968 PMID: 30998681

50. Yamabe M, Horie K, Shiokawa H, Funato H, Yanagisawa M, Kitagawa H. MC-SleepNet: Large-scale Sleep Stage Scoring in Mice by Deep Neural Networks. Sci Rep. 2019; 9(1):1–12.

51. Franken P, Dijk DJ, Tobler I, Borbely AA. Sleep deprivation in rats: effects on EEG power spectra, vigilance states, and cortical temperature. Am J Physiol Integr Comp Physiol. 1991; 261(1):R198–208. https://doi.org/10.1152/ajpregu.1991.261.1.R198 PMID: 1858947

52. Huang ZL, Mochizuki T, Qu WM, Hong ZY, Watanabe T, Urade Y, et al. Altered sleep-wake characteristics and lack of arousal response to H 3 receptor antagonist in histamine H1 receptor knockout mice. Proc Natl Acad Sci U S A. 2006; 103(12):4687–92. https://doi.org/10.1073/pnas.0600451103 PMID: 16537376

53. dos Santos Lima GZ, Lobao-Soares B, Corso G, Belchior H, Lopes SR, de Lima Prado T, et al. Hippocampal and cortical communication around micro-arousals in slow-wave sleep. Sci Rep. 2019; 9(1):1–13.

54. Cui N, Mckillop LE, Fisher SP, Oliver PL, Vyazovskiy V V. Long-term history and immediate preceding state affect EEG slow wave characteristics at NREM sleep onset in C57BL/6 mice. Arch Ital Biol. 2014; 152(2–3):156–68. https://doi.org/10.12871/0002982920142310 PMID: 25828687

55. Doroshenkov L., Konyshev VA, Selishechev S. Classification of Human Sleep Stages Based on EGG Processing Using Hidden Markov Models. 2007; 41(1):25–8.

56. Pan ST, Kuo CE, Zeng JH, Liang SF. A transition-constrained discrete hidden Markov model for automatic sleep staging. Biomed Eng Online. 2012; 11:1–19.

57. Fonseca P, Den Teuling N, Long X, Aarts RM. A comparison of probabilistic classifiers for sleep stage classification. Physiol Meas. 2018; 39(5). https://doi.org/10.1088/1361-6579/aabbc2 PMID: 29620019

58. Hansson-Sandsten M. Optimal multitaper wigner spectrum estimation of a class of locally stationary processes using Hermite functions. EURASIP J Adv Signal Process. 2011;2011.

59. Viterbi AJ. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Trans Inf Theory. 1967; 13(2):260–9.

60. Wang LA, Kern R, Yu E, Choi S, Pan JQ. IntelliSleepScorer, a software package with a graphic user interface for automated sleep stage scoring in mice based on a light gradient boosting machine algorithm. Sci Rep. 2023; 13(1):1–11.

61. Yamagata T, Kahn MC, Prius-Mengual J, Meijer E, Sabanovi M, Guillaumin MCC, et al. The hypothalamic link between arousal and sleep homeostasis in mice. Proc Natl Acad Sci U S A. 2021; 118(51):1–12. https://doi.org/10.1073/pnas.2101580118 PMID: 34903646

62. Huber R, Deboer TOM, Tobler I. Topography of EEG dynamics after sleep deprivation in mice. J Neurophysiol. 2000; 84(4):1888–93. https://doi.org/10.1152/jn.2000.84.4.1888 PMID: 11024081

63. Glin L, Arnaud C, Berracochea D, Galey D, Jaffard R, Gottesmann C. The intermediate stage of sleep in mice. Physiol Behav. 1991; 50(5):951–3. https://doi.org/10.1016/0031-9384(91)90420-s PMID: 1805286

64. Gottesmann C. The transition from slow-wave sleep to paradoxical sleep: Evolving facts and concepts of the neurophysiological processes underlying the intermediate stage of sleep. Neurosci Biobehav Rev. 1996; 20(3):367–87. https://doi.org/10.1016/0149-7634(95)00055-0 PMID: 8880730

65. Emrick JJ, Gross BA, Riley BT, Poe GR. Different simultaneous sleep states in the hippocampus and neocortex. Sleep. 2016; 39(12):2201–9. https://doi.org/10.5665/sleep.6326 PMID: 27748240

66. de Chazal P, Mazzotti DR, Cistulli PA. Automated sleep staging algorithms: have we reached the performance limit due to manual scoring? Sleep. 2022; 45(9):1–3. https://doi.org/10.1093/sleep/zsac159 PMID: 35866932

67. Taguchi Y, Hando S, Sakata M, Eguchi N, Urade Y. Accuracy evaluation of sleep-wake stage analysis with SleepSign Ver2.0. Sleep Biol Rhythms. 2004; 2(SUPPL. 1):92352004.

68. Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, et al. Interrater reliability for sleep scoring according to the Rechtschaffen \& Kales and the new AASM standard. J Sleep Res. 2009; 18(1):74–84.

69. Deng S, Zhang X, Zhang Y, Gao H, Chang EIC, Fan Y, et al. Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard. Sleep Breath. 2019; 23(2):719–28. https://doi.org/10.1007/s11325-019-01801-x PMID: 30783913

70. Van Der Donckt J, Van Der Donckt J, Rademaker M, Vandewiele G, Van Hoecke S. Do Not Sleep on Linear Models: Simple and Interpretable Techniques Outperform Deep Learning for Sleep Scoring. SSRN Electron J. 2022;1–20.

71. Khalighi S, Sousa T, Pires G, Nunes U. Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. Expert Syst Appl. 2013; 40(17):7046–59.

72. Guillot A, Sauvet F, During EH, Thorey V. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. IEEE Trans Neural Syst Rehabil Eng. 2020; 28(9):1955–65. https://doi.org/10.1109/TNSRE.2020.3011181 PMID: 32746326

73. Khalighi S, Sousa T, Santos JM, Nunes U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. Comput Methods Programs Biomed. 2016; 124(November):180–92. https://doi.org/10.1016/j.cmpb.2015.10.013 PMID: 26589468

74. Boe AJ, McGee Koch LL, O'Brien MK, Shawen N, Rogers JA, Lieber RL, et al. Automating sleep stage classification using wireless, wearable sensors. npj Digit Med. 2019; 2(1):1–9. https://doi.org/10.1038/s41746-019-0210-1 PMID: 31886412

75. Krone LB, Yamagata T, Blanco-Duque C, Guillaumin MCC, Kahn MC, van der Vinne V, et al. A role for the cortex in sleep–wake regulation. Nat Neurosci. 2021; 24(9):1210–5. https://doi.org/10.1038/s41593-021-00894-6 PMID: 34341585

76. Alfonsa H, Burman RJ, Brodersen PJN, Newey SE, Mahfooz K, Yamagata T, et al. Intracellular chloride regulation mediates local sleep pressure in the cortex. Nat Neurosci. 2023; 26(1):64–78. https://doi.org/10.1038/s41593-022-01214-2 PMID: 36510112

77. Chung S, Weber F, Zhong P, Tan CL, Nguyen TN, Beier KT, et al. Identification of preoptic sleep neurons using retrograde labelling and gene profiling. Nature. 2017; 545(7655):477–81. https://doi.org/10.1038/nature22350 PMID: 28514446

78. Fisher R. A., FRS Sc.D. The use of multiple measurements in taxonomic problems. Ann Eugen. 1936; 7 (2):179–88.

79. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res. 2011; 12:2825–30.

80. Schreiber J. pomegranate: Fast and flexible probabilistic modeling in python. J Mach Learn Res. 2018; 18:1–6.