bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Somnotate: A robust automated sleep stage classifier that exceeds human performance and identifies ambiguous states in mice

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], Vladyslav V. Vyazovskiy[2], Colin J. Akerman[1]

1 Department of Pharmacology, University of Oxford; Mansfield Road, Oxford, OX1 3QT, UK. ***2*** Department of Physiology, Anatomy and Genetics, University of Oxford; Parks Road, Oxford OX1 3PT, UK.

3 Nuffield Department of Clinical Neurosciences, University of Oxford; Level 6, West Wing, John Radcliffe Hospital, Oxford OX3 9DU, UK.

4 Sleep and Circadian Neuroscience Institute, University of Oxford; Parks Road, Oxford OX1 3PT, UK.

5 Institute for Neuroscience, Department of Health Sciences and Technology, ETH Zurich; Schorenstrasse 16, 8603 Schwerzenbach, Switzerland

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Manual sleep stage annotation is a time-consuming but often essential step in the analysis of sleep data. To address this bottleneck several algorithms have been proposed that automate this process, reporting performance levels that are on par with manual annotation according to measures of inter-rater agreement. Here we first demonstrate that inter-rater agreement can provide a biased and imprecise measure of annotation quality. We therefore develop a principled framework for assessing performance against a consensus annotation derived from multiple experienced sleep researchers. We then construct a new sleep stage classifier that combines automated feature extraction using linear discriminant analysis with inference based on vigilance state-dependent contextual information using a hidden Markov model. This produces automated annotation accuracies that exceed expert performance on rodent electrophysiological data. Furthermore, our classifier is shown to be robust to errors in the training data, robust to experimental manipulations, and compatible with different recording configurations. Finally, we demonstrate that the classifier identifies both successful and failed attempts to transition between vigilance states, which may offer new insights into the occurrence of short awake periods between REM and NREM sleep. We call our classifier 'Somnotate' and make an implementation available to the neuroscience community.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Long-term electrophysiological recordings from freely moving mice and other laboratory animals are a popular and powerful mode of investigation for neuroscientists, particularly sleep researchers (e.g. Schwierin et al. 1998, Leemburg et al 2010, Northeast et al. 2019). 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. Due to their continuous nature, these recordings generate large amounts of data, which places significant demands upon the analysis stages of the experimental process. As the vigilance state profoundly affects the behaviour and physiology of the animal, the first step in this analysis is typically sleep stage annotation. This involves the parcellation of the data into awake, rapid eye movement (REM) sleep and non-REM (NREM) sleep epochs.

Sleep stage annotation is typically performed manually by human experts. Classical criteria used for manual annotation of vigilance states in laboratory rodents include the amplitude of the EEG signal, and the presence of specific oscillations, such as slow waves (0.5-4Hz), spindles (10-15 Hz) and theta-frequency (5-10Hz) activity (Tobler et al. 1997, Ang et al. 2018). Conventionally, NREM sleep is defined by the presence of high amplitude slow waves and spindles, plus a low EMG tone. Wakefulness is defined by a low amplitude, activated EEG pattern, that is dominated by fast frequencies and often theta-activity, plus an EMG that displays elevated tone and phasic events associated with movement. Whereas REM sleep is defined by EEG signals that resemble wakefulness, but the EMG tone is generally low, except for brief muscle twitches.

Whilst manual annotation remains widely used, there are two principal motives for developing effective automated methods for annotating sleep data. The first motive is the substantial time-saving benefit and the resulting opportunities that automation affords. This enables experiments to be conducted at a scale that would be otherwise difficult to imagine (Funato at al. 2016) and can improve smaller-scale experiments by freeing up experimenters to focus upon points of interest, particularly if relevant parts of the data can be identified in a principled way. 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. This is especially the case around sleep stage transitions and during intermediate sleep states, when EEG signals can exhibit signatures of

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

multiple states. 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, both in humans (Nir et al. 2011, Bernardi et al. 2015) and in rodents (Vyazovskyi et al 2011, Funk et al. 2016, Soltani et al. 2019). When faced with such ambiguity, manual annotations often differ between sleep scorers. By standardising the annotation process, automated methods can remove inter-rater variance from the design and replication of experiments, and afford new opportunities to systematically describe these intermediate states.

The earliest approaches to automated sleep stage annotation relied upon hard-coded decision rules that were applied to subjectively selected features in the recordings (Martin et al. 1972; Benington, Kodali, and Heller 1994; Veasey et al. 2000; Stephenson et al. 2009). For example, Bennington et al. determined awake periods by identifying epochs in which the product of the power in the EEG beta frequency band (10-14 Hz) and theta band (4-9 Hz) was below a predetermined threshold. The remaining data was then partitioned into NREM and REM using the ratio between the power in the delta band (0.5-4 Hz) and upper theta band (6-9 Hz) (J. H. Benington, Kodali, and Heller 1994). Others soon realised that the accuracy of such classifiers could be improved by including more features. However, as features differ in their relevance to each vigilance state, systematic approaches are required to weigh different features and determine suitable decision boundaries. This motivated the use of more complex techniques such as linear discriminant analysis (LDA; (Brankačk et al. 2010), support vector machines (Crisler et al. 2008; T. Zeng et al. 2012), naive Bayes classifiers (Rytkönen, Zitting, and Porkka-Heiskanen 2011), artificial neural networks (reviewed in Ronzhina et al. 2012), and combinations thereof (e.g. Lajnef et al. 2015).

With most of these approaches however, each sample is classified independently. As the input data can be noisy, this can result in misclassifications and an overestimation of the number of state transitions. Human scorers can avoid these issues by using contextual information as they visually interrogate the data in a sequential manner, subjectively integrating the evidence within each time bin, with an estimate of the likelihood of each state based on the broader context. 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 genuinely 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 (Franken et al. 1991, Huang et al. 2006, Cui et al. 2014, dos Santos Lima et al. 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Artificial neural networks represent a powerful method for classifying data sets with large numbers of features that take the surrounding state sequence into account (Rumelhart et al. 1986; Werbos 1988; Hochreiter & Schmidhuber 1997). For instance, convolutional neural networks (CNNs) include in their inference the values of the samples immediately surrounding the sample of interest (Homma et al. 1988; Waibel et al. 1989; LeCun & Bengio 1995) and recurrent neural networks are capable of learning long-term dependencies (Hochreiter & Schmidhuber 1997). These principles have been combined to generate networks that are effective at inferring vigilance states (Yulita et al. 2017; Chambon et al. 2018; Malafeev et al. 2018; Phan et al. 2018; Phan et al. 2019; Sun et al. 2019). However, the power and flexibility of artificial neural networks comes at a cost in terms of complexity. Artificial neural networks are non-linear and typically have thousands of parameters that need to be learnt from labelled data. Overparameterisation leaves neural networks prone to overfitting, with the result that networks trained on certain datasets perform less well on data that has been collected under different conditions. In practice, the flexibility of a large number of free parameters poses challenges for sleep researchers that would like to adapt methods to their own experimental setup.

Conceptually simpler algorithms include hidden Markov models (HMMs), which have also been successfully employed for sleep stage annotation (Längkvist, Karlsson, and Loutfi 2012; Jiang et al. 2019). At the core of HMMs are estimates of the probability to change to a different vigilance state from one sample to the next. These probabilities can be learnt from labelled training data and then used to judge the likelihood of state sequences. For example, if the most likely state of the previous epoch is state A, and the state of the next epoch is also state A, the assignment of state B to the current epoch is probably erroneous, particularly if transitions between state A and B are rare. Conversely, if the state of the previous epochs is state A and the state of the following epochs is state B, the assignment of state B to the current epoch becomes more plausible. As well as applying HMMs post-hoc to the state sequence inferred by another algorithm, HMMs can also be applied directly to features derived from electrophysiological data (Doroshenkov, Konyshev, and Selishchev 2007; Pan et al. 2012; Fonseca et al. 2018). However, a crucial drawback of HMMs 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. HMMs are therefore only effective if the dimensionality of the input signal is relatively low. As a result, previous implementations have relied upon a small number of hand-crafted features as basis of their inference.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here we develop a sleep stage classifier that combines inference based on rich feature sets, with the low complexity of a HMM. To this end, we use LDA to perform targeted dimensionality reduction of the complex input signals, while retaining the maximal amount of (linearly decodable) information about vigilance states. This low dimensional representation is then passed to a HMM, which combines the information encoded independently by each sample, with the broader context given by the surrounding state sequence and estimated state transition probabilities. In demonstrating the performance of our classifier, we first draw upon human clinical studies (Danker-Hopfe et al. 2009; Magalang et al. 2013; Deng et al. 2019; Guillot et al. 2020) to establish an unbiased framework for assessing annotation accuracy, which uses the independent manual annotation of data by ten experienced rodent sleep researchers. Our classifier is shown to consistently exceed the accuracy of expert manual annotations on rodent electrophysiological data. We then demonstrate that our classifier is remarkably robust to errors in the training data and to experimental manipulations. Furthermore, our classifier maintains its performance with different numbers and sources of electrophysiological signals, which highlights the classifier's adaptability to different experimental scenarios. As the only free parameter is the time resolution or epoch length of the inference, the classifier is easy to use and retrain, even by an inexperienced user.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Unbiased and precise assessment of automated and manual sleep annotation

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, which is then compared to the level of agreement between the two manual annotations. This subtle difference in how manual and automated annotations are assessed can lead to systematic biases in favour of the automated annotation, as illustrated by the following example. Assume that one manual annotation is perfectly accurate but the other manual annotation misclassifies half of the data. The inter-rater agreement between manual annotations is calculated as 1 * 0.5 = 0.5. Now assume that the automated annotation has exactly the average accuracy of the two manual annotations, i.e. 0.75. The average agreement with the two manual annotations will be (1 * 0.75 + 0.5 * 0.75) / 2 = 0.5625. In other words, the automated annotation will appear to be more than 10% better than the manual annotations, even though its accuracy was exactly average. Conversely, to achieve an average agreement score of 0.5, the automated annotation would only need to have an accuracy of 0.667, i.e. it could be 10% less accurate than the mean manual accuracy, while still achieving the same agreement between manual annotations. The only condition under which this bias is zero, is if both manual annotations have the same exact accuracy. For example, if both manual annotations have an accuracy of 0.75, their agreement is 0.75 * 0.75 = 0.5625, and the average agreement of an automated annotation with both manual annotations yields the same value: (0.75 * 0.75 + 0.75 * 0.75) / 2 = 0.5625.

For these reasons, we were keen to compare automated annotations to a majority-vote consensus derived from multiple independent manual annotations. We asked ten experienced sleep researchers (**Figure 1 - Figure Supplement 1**) to annotate awake, NREM, and REM states during the same 12-hour period and based on simultaneous recordings of an anterior EEG, posterior EEG, and EMG in a freely behaving mouse (Materials and methods). This enabled us to generate consensus annotations based on multiple independent manual annotations (**Figure 1A-C**). First, we assessed the accuracy of each annotation against the consensus of the other nine annotations. This revealed that although the overall accuracy of the annotations was high, individual annotations varied in

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

terms of how closely they matched the consensus of the other annotations (**Figure 1C-D**). This variance would cause systematic bias if one was to rely upon the level of agreement between just two manual annotations (see above). We were also keen to assess how precisely the agreement between any two annotators is able to capture the mean accuracy of both manual annotations. We therefore compared the inter-rater agreement for each pair of annotations to the mean of their accuracies based on the majority-vote consensus of the remaining eight annotations (serving as a proxy for ground truth). Whilst there was a statistically significant linear relationship between inter-rater agreement and the mean accuracy of the two annotations, the relationship was weak (R^2 = 0.25; **Figure 1E**).

To assess the quality of manual and automated annotations in an unbiased and more precise way, we compared the annotations to the consensus of multiple independent manual annotations. This comparison is unbiased as both the manual and automated annotations are assessed in exactly the same way. We confirmed that it is also a more precise measure, as the spread of performance estimates of manual annotations was smaller when using the consensus of three independent manual annotations to assess the accuracy of a fourth annotation, than when using a single other annotation as a point of reference (**Figure 1F;** $p < 0.01$, Wilcoxon signed rank test). Finally, to estimate the minimum number of manual annotations required to achieve a high quality consensus sequence, we determined the consensus sequence of five annotations by majority-vote. Using either one, three or all five of the remaining unused annotations, we constructed a second consensus sequence, and computed the agreement between the two and then repeated this process for all possible combinations. On average, any individual manual annotation matched a consensus of five sequences for 92.5% ± 1.3% of the data (mean ± standard deviation), whereas a consensus of three annotations already significantly increased the agreement by 2.2% ± 1.5% (agreement 94.7% ± 0.8%, $p < 0.01$, Mann-Whitney rank test) **(Figure 1G**). There was a significant but more modest improvement by 0.5% ± 1.1% when the number of manual annotations was increased to five (agreement 95.3% ± 0.7%, $p < 0.01$, Mann-Whitney rank test). Another widely used measure of interrater agreement is Cohen's kappa, which accounts for the possibility of agreements occurring due to chance. When we repeated the analyses shown in **Figure 1D-G** using this performance measure, we obtained analogous results (**Figure 1 - Figure Supplement 1).**

In summary, a consensus derived from multiple independent manual annotations provides a less biased and more precise framework for assessing the quality of manual and automated annotations under comparable conditions. Based on these observations, we generated a larger test data set of six 24-hour EEG and EMG recordings (i.e. 144 hours total), which were independently scored by at

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

least four experienced sleep researchers. Unless noted, we used these six data sets throughout the rest of the study. This allowed us to compute the accuracy of manual and automated annotations using the majority-vote consensus of at least three other manual annotations for that recording. The recordings, individual manual annotations, and automated annotations are made freely available in standard formats at TBD.

Contextual information improves automated classification of sleep states

In machine learning, most classifiers learn a transformation that maps a sample consisting of a set of input values or features, onto an output value or category. This is the basis of the majority of automated methods for sleep stage classification, such as decision trees, linear discriminant analysis, support vector machines, and neural networks (without recurrence) (Brankačk et al. 2010; Crisler et al. 2008; T. Zeng et al. 2012; Rytkönen, Zitting, and Porkka-Heiskanen 2011; Ronzhina et al. 2012; Lajnef et al. 2015). A key weakness of these approaches is that each sample is classified independently of all other samples. In contrast, two methods that incorporate contextual information and have been successfully applied to sleep stage classification are recurrent neural networks and HMMs. Recurrent neural networks can perform exceptionally well at learning long-term dependencies from large feature sets in a variety of problems. However, their inherent flexibility poses problems in practice, as they require large amounts of labelled data to train, are prone to overfitting, and adapting their architecture to different inputs can be non-trivial. In a research setting, where changes to the experimental setup can be frequent and generating large, well curated training data sets is often impractical, recurrent neural networks can be a suboptimal choice. In contrast, HMMs have much fewer parameters and require much less data to train. Typically, however, HMMs have been applied either to low dimensional, hand-crafted features (Doroshenkov, Konyshev, and Selishchev 2007; Pan et al. 2012; Fonseca et al. 2018) or to state sequences inferred by other algorithms (Längkvist, Karlsson, and Loutfi 2012; Jiang et al. 2019). To incorporate more of the information present in the individual samples into the inference, we set out to combine HMMs with LDA, which automatically extracts low dimensional features from complex, high dimensional input signals while retaining the maximal amount of (linearly decodable) information about the labelled target classes.

Our first aim was to demonstrate the advantage conferred when the classifier incorporates contextual information. To that end, we split the data into a training set and a test set, and evaluated

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

three automated classifiers that differed only in the amount of contextual information that they incorporated into their inference (**Figure 2C-E**). The data preparation was the same for all classifiers: 1) data pre-processing (subsampling to 256 Hz; conversion to multitaper spectrograms), 2) normalisation (log(x + 1) transformation; conversion to z-scores), and 3) targeted dimensionality reduction using linear discriminant analysis (LDA).

In the first classifier, referred to as ‘LDA’, a set of thresholds computed as part of the LDA were applied to the low dimensional representation of the test data (**Figure 2C**). This is equivalent to selecting the state $\hat{s} \in S$ that best explains the values in the data sample $d \in D$, regardless of the prior probability of different states:

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

and thus represents a baseline performance for automated annotation without incorporating contextual information. In the second classifier, referred to as ‘Bayes’, we constructed a naive Bayes classifier by fitting multivariate Gaussian distributions (one for each state) to the low dimensional representation of the samples in the training data set, as well as computing the expected frequencies of the different states (**Figure 2D** and **2F**). The states corresponding to samples in the test set were then predicted based on the probability of the sample given each state weighted by the frequency of states:

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

In the third classifier, referred to as ‘HMM’, the states in the test set were predicted using a hidden Markov model. As above, multivariate Gaussian distributions were fitted to the low dimensional representation of samples in the training data. In addition however, the expected probabilities of transitioning from one state to another were estimated from the training data set (**Figure 2G**). The probability given each state P(D|S) was computed for each sample in the test set, combined with the expected state transition frequencies, and the most likely state sequence through the test set was computed using the Viterbi algorithm (Viterbi 1967) (**Figure 2E**). 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. Consequently, the HMM incorporates more contextual information into its inference than the Bayes classifier, as the transition probability matrix can be used to compute the expected state frequencies from the stationary distribution of a corresponding Markov model (whereas the state transition probabilities cannot be computed from the state frequencies alone). It was observed that incorporating contextual information more than halved the number of errors, as

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the accuracy significantly increased from the LDA classifier's accuracy of 92% ± 1%, to the HMM classifier, which achieved an accuracy of 97% ± 1% (errors correspond to standard deviations from the mean). These analyses confirmed the benefit of incorporating contextual information and established automated feature extraction using LDA, combined with context-aware state annotation using a HMM, as a promising solution for an automated sleep stage classifier. We refer to this classifier as 'Somnotate' and we set out to test its performance and robustness in the subsequent sections.

State annotation by Somnotate exceeds manual accuracy

Manual state annotation continues to be the yardstick by which any automated annotation is measured. 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 for that data set. The sleep researchers had a minimum of 2 years' experience (median of 5 years' experience) in manual vigilance state annotation and had annotated at least 768 hours of equivalent recordings (**Figure 3 – Figure Supplement 1**). Somnotate was trained and tested, in a hold-one-out fashion, on the same data set 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 (**Figure 3A**). 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 ($p < 0.001$, Wilcoxon signed rank). When we used Cohen's kappa instead of accuracy as a measure of performance, we obtained identical results (**Figure 3 - Figure Supplement 2**). 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 (**Figure 3B**). Somnotate identified more state transitions than were typically present in manual annotations, in particular if these transitions involved NREM states (**Figure 3D**). However, cumulatively, the differences between manual and automated state annotations resulted in minor differences in the overall state occupancy (**Figure 3C**). When partitioning the data by state according to the manual consensus or the automated annotation, there

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were no discernible differences between power spectra of the EEG activity (**Figure 3 – Figure Supplement 3**).

Approximately two thirds of the differences between the automated and manual consensus annotations had a duration that was shorter than the temporal resolution of the manual annotations (i.e. shorter than 4 s; **Figure 3E**). Many of these differences could therefore be resolved if 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 disagree with one another (**Figure 3F**). 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 (as in **Figure 1**). 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 $\rho = 0.80$, $p < 0.001$; **Figure 3G**). In other words, as one increases the number of experienced annotators, the manual consensus annotation converges on the automated annotation by Somnotate.

Somnotate is highly robust

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

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 4 experienced sleep researchers in a hold-one-out fashion, while randomly permuting an increasing proportion of the consensus state annotations. This experiment 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 (**Figure 4A-B**). Furthermore, classifier performance monotonically increased with increasing amounts of training data (**Figure 4 – Figure supplement**

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1), 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 drop when faced with test data collected under different conditions, particularly when features used by the classifier are consistently altered. One of the most common manipulations used in sleep research is sleep deprivation, which is known to result in changes in the EEG power spectrum after sleep onset. As the EEG power values are primary features used by Somnotate, it stood to reason that sleep deprivation could negatively impact its performance. To test this, 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' data recorded from the same animals when not experiencing sleep deprivation (**Figure 4C-D**). The annotation accuracy was found to be comparable and high across both the sleep deprivation and baseline conditions, confirming Somnotate's robustness to sleep deprivation induced changes in features used by the classifier.

Third, as Somnotate uses contextual information in the form of prior probabilities of the different vigilance states, changes in these probabilities could negatively impact the automated annotation. The values of the prior probabilities depend on how much time the animals spend in each state and how frequently they transition between states. To gauge the impact that variations in these priors might have on the performance, we trained and tested the HMM in a hold-one-out fashion on the six 24-hour EEG and EMG recordings with high quality consensus annotations, as before. However, we then evaluated the accuracy separately for the 12-hour light period and the 12-hour dark period, when state occupancies and state transition probabilities are very different (**Figure 4E-F**). Despite these large changes in vigilance state, the accuracy of Somnotate continued to match or exceed the accuracy of manual annotations by experienced sleep researchers across both the light and dark periods (**Figure 4G-H**). These observations demonstrate the robustness of Somnotate to changes in the values of the features used for inference, the state frequencies, and the state transition probabilities, which represent the main components that the classifier uses for contextual information.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A single EEG signal is sufficient for Somnotate to infer vigilance states with high accuracy

Although most electrophysiological signals show a dependence on vigilance state, some signals can be more informative of certain states. For example, REM sleep is often 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 (Huber et al. 2000). However, it is not always technically possible to record the 'ideal' combination of signals for the inference of vigilance states. We were therefore keen to assess Somnotate's ability to infer vigilance states from individual electrophysiological signals. A series of classifiers were trained and tested using only one electrophysiological signal as an input; either the anterior EEG, the posterior EEG, the LFP from primary motor cortex, or the EMG. This revealed that a single EEG signal was sufficient to infer the vigilance state with high accuracy (**Figure 5A**). In fact, 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). Underlying this was a small increase in the false negative detection rate for REM sleep, which was offset by an improved distinction between the awake and two sleep states (**Figure 5B**). The overall accuracy of the predictions based on the posterior EEG was on average 1 % lower ($p < 0.05$, Wilcoxon signed rank test), although in this case the identification of REM showed a similar accuracy to when all signals were provided (**Figure 5B**). 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; **Figure 5A-B**). 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. Overall, these data establish that Somnotate is able to accurately infer vigilance states from individual electrophysiological signals and in a manner that could be optimised depending on the experimental arrangement and objectives.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Somnotate identifies ambiguous states around successful and unsuccessful attempts to transition between states

HMMs belong to the category of Bayes classifiers that compute the likelihood of each state, for every data sample. This likelihood identifies samples where the classifier is certain in its prediction (i.e. samples where the likelihood of the predicted state is effectively one), and samples where the classifier is uncertain (i.e. samples where the likelihood of the predicted state is less than one). For our datasets, the cumulative distribution of likelihood values indicated a change point at 0.995 (**Figure 6 – Figure supplement 1**), with a minority of samples (5.5%) having a likelihood below this threshold. Notably, almost half of the periods for which Somnotate was uncertain (44%) coincided with instances where manual annotations disagreed with one another. This suggested that the difficulty in predicting the vigilance state for these time points was not an artefact of the inference method, but a result of ambiguity in the signals themselves. Human annotators often exclude such sections of the data from their analysis and, by analogy, the accuracy of our automated classifier increased when these ambiguous samples were excluded (**Figure 6 – Figure supplement 2**).

In contrast to manual annotation, Somnotate provides the opportunity to characterise these ambiguous samples in a principled way. An ambiguous signal can derive from measurement noise masking a true, unambiguous signal, or it can reflect an intermediate state that produces a mixed signal. Three lines of evidence support the idea that ambiguous samples derive from an intermediate state and are not a measurement artefact. First, the majority of ambiguous samples (76%) occur around state transitions (first two examples in **Figure 6A**), which represent relatively rare events in the recordings. Second, for nearly all ambiguous samples, the probability mass was concentrated in two states, rather than being randomly distributed across all three states (**Figure 6B**). Third, the power spectra for ambiguous samples showed elements of the power spectra of the two most likely states (**Figure 6C**). For example, samples where Somnotate was uncertain whether to assign the awake state or NREM sleep, showed a high power in the δ frequency band, characteristic of NREM sleep, but also an increased power in the γ frequency band, which is typically an indicator of the awake state (**Figure 6C**). Finally, as the periods during which the classifier was uncertain tended to be much longer than the duration of a single sample (**Figure 6 – Figure supplement 3**), these

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

intermediate states could not be an artefact of the temporal resolution, i.e. the result of a state transition occurring during a single one second sample.

To investigate the new analysis opportunities this generates, we focused upon the 24% of ambiguous samples that were associated with an incomplete state transition (such as the third example in **Figure 6A**) 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 (**Figure 6D**). 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). However, 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 was adopted for ambiguous time points (state probability below 0.95; **Figure 6 – Figure supplement 4**). These observations may explain why animals often appear to enter a brief awake state after a period of REM sleep, before they resume with NREM sleep. In summary, these analyses represent an example of the additional opportunities generated by using an automated classifier that computes the likelihood of vigilance states at each time point.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Here we present a novel sleep stage classifier that achieves performance levels that surpass human experts, is robust, easy to use and provides quantification of ambiguous states. We call the classifier 'Somnotate' and we make this available to the neuroscientific community. Somnotate combines optimal feature extraction by linear discriminant analysis (LDA), with state-dependent contextual information derived via a hidden Markov model (HMM). Through a series of systematic benchmarking tests against expert manual annotations, we demonstrate that Somnotate outperforms the accuracy achieved by experienced sleep researchers. The classifier is shown to be robust to errors in the training data, able to operate across experimental manipulations, and compatible with different electrophysiological signals. Finally, we demonstrate that Somnotate is well-placed to quantify ambiguous states, which can be used to investigate putative failed transitions between vigilance states.

Despite the development of multiple algorithms for sleep stage classification, many sleep researchers and clinicians continue to manually score their data. We believe that several barriers have prevented widespread adoption of automated solutions, which include issues relating to the true performance levels of automated classifiers, their robustness, adaptability, accessibility, and whether they offer the potential for new insights. We will discuss Somnotate in the context of each of these aspects.

In terms of performance levels, reports of human-like performance by automated methods may fall short in practice. We believe 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 (Danker-Hopfe et al. 2009; Deng et al. 2019). Using this improved assessment of performance, we showed that, on average, Somnotate matched the consensus more closely than any individual manual annotation. Further, the more manual annotations that were used to generate the consensus sequence, the more closely this consensus matched the automated annotation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Key to Somnotate's performance is its incorporation of contextual information. Human experts continually use contextual information as they interrogate such 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 (Längkvist, Karlsson, and Loutfi 2012; Jiang et al. 2019) and recurrent neural networks (Yulita et al. 2017; Chambon et al. 2018; Malafeev et al. 2018; Phan et al. 2018; Phan et al. 2019; Sun et al. 2019). As HMMs are much easier to adapt and optimise by a non-expert, we concentrated our efforts on improving the state-of-the-art for HMM-based inference of vigilance states. Our classifier represents an advance upon previous work, 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.

A key advantage to Somnotate is its robustness. 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 show such overfitting to a standard or control data set, that their performance is diminished in other settings when the probabilities of key features vary (Veasey et al. 2000; Khalighi et al. 2013; Malafeev et al. 2018, Sun et al. 2019; Guillot et al. 2020). In contrast, we saw no statistically significant drop in the performance of Somnotate when used with data collected from sleep deprived animals, even though sleep deprivation elicits significant changes in the EEG spectrogram, and thus changes in key features used by Somnotate. Similarly, dividing the data into light and dark periods produced pronounced changes in the state transition frequencies, yet Somnotate continued to perform at least as well as expert sleep researchers.

In terms of adaptability, some automated methods have been optimised for a specific input signal (Lefort et al. 2018). In other cases, an automated method can be recalibrated to accommodate changes in the experimental setup. For example, support vector machines and neural networks can be retrained, or linear discriminants and decision trees can be re-evaluated. However, as most previous methods have several free parameters, adapting them to a different experimental arrangement can be time-consuming, with uncertain returns. Here we showed that Somnotate performs well with a variety of different input signals, which are typically available in an experimental setting. Furthermore, as there are no free parameters other than the desired time resolution of the state prediction, re-training requires no optimisation, and is straightforward and

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fast. Training Somnotate takes approximately one second per 24 hours of data on a standard desktop computer.

A further barrier to the widespread adoption of automated solutions is the issue of accessibility. Available software implementations for automated sleep stage classification can be expensive (Taguchi et al. 2004; Alloca et al. 2019). We provide an open source implementation 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.

Finally, previous descriptions of automated methods for sleep stage classification have understandably focused on the savings in person-hours (Khalighi et al. 2013; Sun et al. 2019). In our opinion, it is also important to recognise that some analyses may *require* 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 certain areas of investigation difficult. For example, EEG traces show signatures of multiple states, particularly around state transitions (Glin et al. 1991, Gottesmann 1996, Emrick et al. 2016, Funk et al. 2016), which we show is where most disagreements between manual annotations occur. And whilst humans are very good at determining the most likely state at any given time point, they struggle to quantify intermediate states. In contrast, there is no difference between these two tasks for our classifier. In a subset of cases, Somnotate indicated intermediate states in the absence of a state transition, which we defined as failed transitions. Interestingly, the distribution of failed transitions was highly non-random. Notably, REM-to-awake transitions were nearly always successful, whereas failures were much more common for REM-to-NREM transitions. The differential failure rate may explain the preponderance of short bouts of wakefulness 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 transition directly from REM to NREM. This highlights a potential direction for future investigations, which could lead to a richer description of the neurophysiological mechanisms of vigilance state transitions.

More broadly, we have assessed Somnotate's performance upon a number of electrophysiological signals recorded in mice. The use of targeted feature extraction via LDA means that Somnotate is agnostic with respect to the exact nature of the input signal. In principle, the method could be applied to any high frequency time series data that contains information about an organism's vigilance state. Hence we plan to expand this approach to other types of signals, such as surface

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

EEG, actigraphy, or respiratory activity (Zeng et al. 2012; Khalighi et al. 2016; Boe et al. 2019; Guillot et al. 2020). In conclusion, we present Somnotate - a method for automated sleep stage classification from long-term electrophysiological recordings in freely moving animals. Somnotate achieves performance levels that exceed human experts and meets important criteria in terms of robustness, ease of use, accessibility, and the potential for new biological insights.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and methods

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). Animals were maintained under a 12-h:12-h light-dark (LD) cycle. For the subset of animals that underwent a sleep deprivation (SD) protocol, the animal was pre-exposed to novel objects to encourage exploratory behaviour. The SD protocol then consisted of delivering novel objects for the first six hours 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 micrometre 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 (Krone et al., 2021). EEG and LFP signals were referenced to a cerebellum screw. For EMG recordings, wire electrodes were inserted into the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 a sampling rate of 256 Hz using custom-made 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.

Manual vigilance state annotation

Manual annotation of vigilance states was performed offline, based on 4 s epochs using SleepSign 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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; Fisher 1936) as implemented in the scikit-learn python library (Pedregosa et al. 2011). 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 an HMM in a supervised fashion with multivariate Gaussian state emissions using the python library pomegranate (Schreiber 2018; all optional parameters at default values). If the annotations were not based on the consensus of multiple manual annotations, mislabelled samples in the training data resulted in non-zero probabilities for disallowed state transitions, specifically awake-to-REM transitions. These were pruned by removing all state transitions with probability below 0.0001 per second. The accuracy of the trained LDA and HMM models were ascertained by applying them to held out test data. 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author contributions

PJNB designed and wrote the Somnotate software. HA tested and provided feedback on Somnotate during its early development. PJNB and CJA designed the validation experiments and PJNB carried out the analysis. HA, LBK, and CBD performed the in-vivo recordings. LBK organised the manual annotation effort. HA, LBK, CBD, ASF, SF, MCCG, YGH, MCK, LEK, LEM, LM, LT, CWT, TY, and VVV contributed manual annotations. VVV and CJA supervised the work. PJNB and CJA wrote the manuscript, with input from all authors.

Acknowledgements

We would like to thank the Akerman lab for advice and comments. The research leading to these results received funding from the European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013, ERC Grant Agreement 617670 and also Medical Research Council (UK) project MR/N026039/1 and MR/S01134X/1, and Wellcome Trust 106174/Z/14/Z. HA was funded by a Sir Henry Wellcome Postdoctoral Fellowship and St John's College Junior Research Fellowship; LBK by a Wellcome Trust PhD studentship (203971/Z/16/Z), Hertford College Senior Scholarship, Goodger and Schorstein Research Scholarships in Medical Sciences; CBD by a Wellcome Trust PhD studentship (109059/Z/15/Z) and Clarendon Fund Scholarship; MCCG by a BBSRC DTP grant (BB/J014427/1) and Clarendon Scholarship; YGH by a Medical Research Council/Stroke Association Clinical Research Training Fellowship; MCK by a Berrow Foundation Lord Florey Scholarship, plus Goodger and Schorstein Research Scholarship in Medical Sciences; LEM by a Biotechnology and Biological Sciences Research Council Industrial CASE Grant with Eli Lilly & Company Ltd (BB/K011847/1), a Novo Nordisk Postdoctoral Fellowship and a Sir Paul Nurse Junior Research Fellowship at Linacre College; LM by an Action on Hearing Loss studentship; CWT by a BBSRC DTP Grant (BB/M011224/1); TY by a Uehara Memorial Foundation Overseas Postdoctoral Fellowship and a Naito Grant for Studying Overseas.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Allocca, G. et al. (2019) ‘Validation of “somnivore”, a machine learning algorithm for automated scoring and analysis of polysomnography data’, Frontiers in Neuroscience, 13(March), pp. 1–18. doi: 10.3389/fnins.2019.00207.

Benington, J. H., Kodali, S. K. and Heller, H. C. (1994) ‘Scoring transitions to REM sleep in rats based on the EEG phenomena of pre-REM sleep: An improved analysis of sleep structure’, Sleep, 17(1), pp. 28–36. doi: 10.1093/sleep/17.1.28.

Boe, A. J. et al. (2019) ‘Automating sleep stage classification using wireless, wearable sensors’, npj Digital Medicine. Springer US, 2(1), pp. 1–9. doi: 10.1038/s41746-019-0210-1.

Brankačk, J. et al. (2010) ‘EEG gamma frequency and sleep-wake scoring in mice: Comparing two types of supervised classifiers’, Brain Research, 1322, pp. 59–71. doi: 10.1016/j.brainres.2010.01.069.

Chambon, S. et al. (2018) ‘A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series’, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(4), pp. 758–769. doi: 10.1109/TNSRE.2018.2813138.

Crisler, S. et al. (2008) ‘Sleep-stage scoring in the rat using a support vector machine’, Journal of Neuroscience Methods, 168(2), pp. 524–534. doi: 10.1016/j.jneumeth.2007.10.027.

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

Danker-Hopfe, H. et al. (2009) ‘Interrater reliability for sleep scoring according to the Rechtschaffen \& Kales and the new AASM standard’, Journal of Sleep Research, 18(1), pp. 74–84. doi: 10.1111/j.1365-2869.2008.00700.x.

Deng, S. et al. (2019) ‘Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard’, Sleep and Breathing. Sleep and Breathing, 23(2), pp. 719–728. doi: 10.1007/s11325-019-01801-x.

Doroshenkov, L., Konyshev, V. A. and Selishchev, S. (2007) ‘Classification of Human Sleep Stages Based on EEG Processing Using Hidden Markov Models’, Meditsinskaia tekhnika, 41, pp. 24–28. doi: 10.1007/s10527-007-0006-5.

dos Santos Lima, G. Z. et al. (2019) ‘Hippocampal and cortical communication around micro-arousals in slow-wave sleep’, Scientific Reports, 9(1), pp. 1–13. doi: 10.1038/s41598-019-42100-5.

Emrick, J. J. et al. (2016) ‘Different simultaneous sleep states in the hippocampus and neocortex’, Sleep, 39(12), pp. 2201–2209. doi: 10.5665/sleep.6326.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fonseca, P. et al. (2018) ‘A comparison of probabilistic classifiers for sleep stage classification’, Physiological Measurement. IOP Publishing, 39(5). doi: 10.1088/1361-6579/aabbc2.

Franken P, Dijk DJ, Tobler I, Borbély AA. Sleep deprivation in rats: effects on EEG power spectra, vigilance states, and cortical temperature. Am J Physiol. 1991 Jul;261(1 Pt 2):R198-208. doi: 10.1152/ajpregu.1991.261.1.R198. PMID: 1858947.

Franken, P. (2002) ‘Long-term vs. short-term processes regulating REM sleep’, Journal of Sleep Research, 11(1), pp. 17–28. doi: 10.1046/j.1365-2869.2002.00275.x.

Funato, H. et al. (2016) ‘Forward-genetics analysis of sleep in randomly mutagenized mice’, Nature, 539(7629), pp. 378–383. doi: 10.1038/nature20142.

Funk, C. M. et al. (2016) ‘Local slow waves in superficial layers of primary cortical areas during REM sleep’, Current Biology, 26(3), pp. 396–403. doi: 10.1016/j.cub.2015.11.062.

Glin, L. et al. (1991) ‘The intermediate stage of sleep in mice’, Physiology and Behavior, 50(5), pp. 951–953. doi: 10.1016/0031-9384(91)90420-S.

Gottesmann, C. (1996) ‘The transition from slow-wave sleep to paradoxical sleep: Evolving facts and concepts of the neurophysiological processes underlying the intermediate stage of sleep’, Neuroscience and Biobehavioral Reviews, 20(3), pp. 367–387. doi: 10.1016/0149-7634(95)00055-0.

Guillot, A. et al. (2020) ‘Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging’, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(9), pp. 1955–1965. doi: 10.1109/TNSRE.2020.3011181.

Hansson-Sandsten, M. (2011) ‘Optimal multitaper wigner spectrum estimation of a class of locally stationary processes using Hermite functions’, Eurasip Journal on Advances in Signal Processing, 2011. doi: 10.1155/2011/980805.

Hochreiter, S. and Schmidhuber, J. (1997) ‘Long short-term memory’, Neural computation, 9(8), pp. 1–32.

Homma, T., Atlas, L. E. and Marks, R. J. (1987) ‘Artificial Neural Network for Spatio-Temporal Binary Patterns: Application To Phoneme Classification.’, p. 21.

Huang, Z. L. et al. (2006) ‘Altered sleep-wake characteristics and lack of arousal response to H 3 receptor antagonist in histamine H1 receptor knockout mice’, Proceedings of the National Academy of Sciences of the United States of America, 103(12), pp. 4687–4692. doi: 10.1073/pnas.0600451103.

Huber, R., Deboer, T. O. M. and Tobler, I. (2000) ‘Topography of EEG dynamics after sleep deprivation in mice’, Journal of Neurophysiology, 84(4), pp. 1888–1893. doi: 10.1152/jn.2000.84.4.1888.

Jiang, D. et al. (2019) ‘Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement’, Expert Systems with Applications, 121, pp. 188–203. doi: 10.1016/j.eswa.2018.12.023.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Khalighi, S. et al. (2013) 'Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels', Expert Systems with Applications. Elsevier Ltd, 40(17), pp. 7046–7059. doi: 10.1016/j.eswa.2013.06.023.

Khalighi, S. et al. (2016) 'ISRUC-Sleep: A comprehensive public dataset for sleep researchers', Computer Methods and Programs in Biomedicine. Elsevier Ireland Ltd, 124(November), pp. 180–192. doi: 10.1016/j.cmpb.2015.10.013.

Krone, L. B. et al. (2021) 'A role for the cortex in sleep–wake regulation', Nature Neuroscience. Springer US, 24(9), pp. 1210–1215. doi: 10.1038/s41593-021-00894-6.

Lajnef, T. et al. (2015) 'Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines', Journal of Neuroscience Methods. Elsevier B.V., 250, pp. 94–105. doi: 10.1016/j.jneumeth.2015.01.022.

Längkvist, M., Karlsson, L. and Loutfi, A. (2012) 'Sleep Stage Classification Using Unsupervised Feature Learning', Advances in Artificial Neural Systems, 2012, pp. 1–9. doi: 10.1155/2012/107046.

LeCun, Y. and Bengio, Y. (1995) 'Convolutional networks for images, speech, and time series', The handbook of brain theory and neural networks, 3361(May 2013), pp. 255–258.

Lefort, J. M. et al. (2018) Harnessing olfactory bulb oscillations to perform fully brain-based sleep-scoring and real-time monitoring of anaesthesia depth. doi: 10.1371/journal.pbio.2005458.

Magalang, U. J. et al. (2013) 'Agreement in the scoring of respiratory events and sleep among international sleep centers', Sleep, 36(4), pp. 591–596. doi: 10.5665/sleep.2552.

Malafeev, A. et al. (2018) 'Automatic human sleep stage scoring using deep neural networks', Frontiers in Neuroscience, 12(NOV), pp. 1–15. doi: 10.3389/fnins.2018.00781.

Martin, W. B. et al. (1972) 'Pattern recognition of EEG-EOG as a technique for all-night sleep stage scoring', Electroencephalography and Clinical Neurophysiology, 32(4), pp. 417–427. doi: 10.1016/0013-4694(72)90009-0.

Nir, Y. et al. (2011) 'Regional Slow Waves and Spindles in Human Sleep', Neuron, 70(1), pp. 153–169. doi: 10.1016/j.neuron.2011.02.043.

Northeast, R. C. et al. (2019) 'Sleep homeostasis during daytime food entrainment in mice', Sleep, 42(11), pp. 1–13. doi: 10.1093/sleep/zsz157.

Pan, S. T. et al. (2012) 'A transition-constrained discrete hidden Markov model for automatic sleep staging', BioMedical Engineering Online, 11, pp. 1–19. doi: 10.1186/1475-925X-11-52.

Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in {P}ython', Journal of Machine Learning Research, 12, pp. 2825–2830.

Penzel, T. and Conradt, R. (2000) 'Computer based sleep recording and analysis', Sleep Medicine Reviews, 4(2), pp. 131–148. doi: 10.1053/smrv.1999.0087.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Phan, H. et al. (2019) ‘Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification’, IEEE Transactions on Biomedical Engineering, 66(5), pp. 1285–1296. doi: 10.1109/TBME.2018.2872652.

Phan, H. et al. (2018) ‘Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks’, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2018-July, pp. 1452–1455. doi: 10.1109/EMBC.2018.8512480.

Ronzhina, M. et al. (2012) ‘Sleep scoring using artificial neural networks’, Sleep Medicine Reviews, 16(3), pp. 251–263. doi: 10.1016/j.smrv.2011.06.003.

Rumelhart, D. E., Hinton, G. E. and Williams, R. (1986) ‘Learning representations by back-propagating errors’, Nature, 323(9).

Rytkönen, K. M., Zitting, J. and Porkka-Heiskanen, T. (2011) ‘Automated sleep scoring in rats and mice using the naive Bayes classifier’, Journal of Neuroscience Methods, 202(1), pp. 60–64. doi: 10.1016/j.jneumeth.2011.08.023.

Schreiber, J. (2018) ‘pomegranate: Fast and flexible probabilistic modeling in python’, Journal of Machine Learning Research, 18, pp. 1–6.

Schwierin, B., Borbély, A. A. and Tobler, I. (1998) ‘Sleep homeostasis in the female rat during the estrous cycle’, Brain Research, 811(1–2), pp. 96–104. doi: 10.1016/S0006-8993(98)00991-3.

Stephenson, R. et al. (2009) ‘Automated analysis of sleep-wake state in rats’, Journal of Neuroscience Methods, 184(2), pp. 263–274. doi: 10.1016/j.jneumeth.2009.08.014.

Sun, C. et al. (2019) ‘A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation’, IEEE Access, 7, pp. 109386–109397. doi: 10.1109/ACCESS.2019.2933814.

Taguchi, Y. et al. (2004) ‘Accuracy evaluation of sleep-wake stage analysis with SleepSign Ver2.0’, Sleep and Biological Rhythms, 2(SUPPL. 1), p. 92352004. doi: 10.1111/j.1479-8425.2004.00117.x.

Veasey, S. C. et al. (2000) ‘An automated system for recording and analysis of sleep in mice.’, Sleep, 23(8), pp. 1025–1040.

Viterbi, A. J. (1967) ‘Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm’, IEEE Transactions on Information Theory, 13(2), pp. 260–269. doi: 10.1109/TIT.1967.1054010.

Vyazovskiy, V. V. et al. (2011) ‘Local sleep in awake rats’, Nature. Nature Publishing Group, 472(7344), pp. 443–447. doi: 10.1038/nature10009.

Waibel, A. et al. (1989) ‘Phoneme Recognition Using Time-Delay Neural Networks’, IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3), pp. 328–339. doi: 10.1109/29.21701.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Werbos, P. J. (1988) ‘Generalization of backpropagation with application to a recurrent gas market model’, Neural Networks, 1(4), pp. 339–356. doi: 10.1016/0893-6080(88)90007-X.

Yulita, I. N., Fanany, M. I. and Arymurthy, A. M. (2017) ‘Combining deep belief networks and bidirectional long short-term memory case study: Sleep stage classification’, International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017-Decem(September), pp. 19–21. doi: 10.1109/EECSI.2017.8239089.

Zeng, T. et al. (2012) ‘Automated determination of wakefulness and sleep in rats based on non-invasively acquired measures of movement and respiratory activity’, Journal of Neuroscience Methods. Elsevier B.V., 204(2), pp. 276–287. doi: 10.1016/j.jneumeth.2011.12.001.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure legends

Figure 1. 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 moving 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 of the mean.*

Figure 2. Contextual information improves automated sleep stage classification. *(**A**) A fifteen-minute segment of the consensus of manual annotations by four independent experienced sleep researchers, based on recordings from a freely behaving mouse. (**B**) Annotation was based on the anterior EEG, posterior EEG and EMG traces (top), and corresponding multi-taper spectrograms (bottom). (**C**) Two-dimensional representation of the segment after targeted dimensionality reduction via LDA. Small values in the first component (LDA1) indicate the awake state, large values indicate either REM or NREM. Small values in the second component (LDA2) indicate the awake or NREM state, large values indicate REM. (**D**) Probability of each state when fitting two-dimensional Gaussian distributions to the values in C. (**E**) Likelihood of each state given the probability of each state (as shown in 'D') and all possible state sequences, weighted by their likelihood given the state transition probabilities (as shown in 'G'). (**F**) The state occupancy based on the time spent in each state across six 24-hour data sets, according to at least four manual annotations. (**G**) The corresponding state transition probabilities. (**H**) Comparison of the LDA, naive Bayes, and HMM classifiers. Without any contextual information, applying linear thresholds to the values in 'C' yields the LDA classifier. Combining the probability of the data given the state with the prior state probability based on state occupancy shown in 'F' yields a naive Bayes classifier. If instead the prior probability of each state is derived from the state transition*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

probabilities shown in 'G', the classifier becomes a HMM. The accuracy of the three classifiers was evaluated across six 24-hour data sets in a hold-one-out fashion. Error bars indicate the standard deviation of the mean. P-values are based on a Wilcoxon signed rank test with a Bonferroni-Holm correction for multiple comparisons.

***Figure 3. 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 3 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 classifier performed best at a time resolution of 1 s. (**F**) Venn-diagram of the time points at which the automated annotation and manual consensus differed. (**G**) Somnotate was trained on six 24-hour data sets and then tested on a 12-hour data set, independently annotated by ten experienced sleep researchers (as in **Figure 1**). The accuracy of the annotation by Somnotate was compared to consensus annotations generated from different numbers of manual annotations. Error bars indicate the standard deviation of the mean. P-values are derived from a Wilcoxon signed rank test.*

***Figure 4. Automated sleep stage classification by Somnotate is robust.** (**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. The accuracy was evaluated based on an individual manual annotation on six 12-hour data sets acquired during a normal sleep cycle, and compared to the accuracy on six 12-hour data sets acquired from the same animals after sleep deprivation. (**C**) The accuracy of a pre-trained classifier was evaluated against a manual annotation of normal sleep-wake cycle data (six 12-hour data sets), and compared to its accuracy on data from the same animals after undergoing a sleep deprivation protocol (six 12-hour data sets). Confusion matrices are shown for the normal sleep-wake cycle (left, 'baseline'), 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. (**E**) To assess the*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

*impact of a change in state transition probabilities, the accuracy of a classifier trained on 24-hour data sets (i.e. full-day data sets) was evaluated on either data sets acquired during the light period (six 12-hour data sets) or during the dark period (six 12-hour data sets). The state occupancy is shown during the light (left) and dark period (right). (**F**) The state transition probabilities, normalised to their corresponding values over the full-day, for the light (left) and dark periods (right). (**G**) Confusion matrices for the light (left) and dark periods (right), with entries corresponding to the difference after subtracting the corresponding values for the full-day. (**H**) Somnotate's accuracy was compared to the accuracy of individual manual annotations (n=25) during the light (left) and dark periods (right). Values throughout indicate mean ± standard deviation. P-values are derived from a Wilcoxon signed rank test.*

Figure 5. A single EEG signal is sufficient for Somnotate to infer vigilance state with high accuracy. *(**A**) The accuracy of Somnotate's sleep stage classification using a single input signal. Classifiers were 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.*

Figure 6. Somnotate identifies time intervals with ambiguous states. *(**A**) Three examples of ambiguous 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, 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 plots of the state probabilities assigned to each sample in six 24-hour data sets. In the vast majority of samples, the probability mass was concentrated in one or two states (left). 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.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Supplement legends

Figure 1 – Figure Supplement 1. The consensus of manual annotations yields a better estimate of annotation accuracy, as measured by Cohen's kappa. *The analyses in Figure 1D-G 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 the standard deviation of the mean.*

Figure 3 – Figure Supplement 1. Manual annotations were carried out by experienced sleep researchers. *All authors who provided manual annotations reported their task-relevant experience in years, as well as determined the number of 12-hour and 24-hour data sets they had manually annotated previously.*

Figure 3 – Figure Supplement 2. Sleep stage classification with Somnotate exceeds manual performance*. The analysis in Figure 3A was repeated using Cohen's kappa as a measure of performance. 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).*

Figure 3 – Figure Supplement 3. EEG power spectra by state according to Somnotate (solid lines) or manual consensus annotations (dotted lines). *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. Solid lines indicate the median EEG power according to Somnotate's annotation; dotted lines correspond to the median power according to the manual consensus annotation.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.06.463356; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4 – Figure Supplement 1. 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 to us was twenty 24-hour EEG and EMG recordings under baseline conditions. The line indicates the median. Error bars demarcate the 5th and 95th percentile.*

Figure 6 – Figure Supplement 1. 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.*

Figure 6 – Figure Supplement 2. Excluding samples where the classifier is not certain improves the accuracy of automated annotation. *(**A**) Classifier accuracy was compared between cases when all time points were included ('baseline') and when 5.5% of samples were removed because the likelihood of the predicted state dropped below 0.995 ('refined'). The plot indicates mean ± standard deviation and p-values are derived from a Wilcoxon signed rank test. (**B**) Confusion matrices when including all time points (left), when excluding time points where the automated annotation was uncertain (middle) and the difference between these (right).*

Figure 6 – Figure Supplement 3. Duration of ambiguous intervals around successful and failed state transitions. *The probability of the predicted state was computed for each sample in six 24-hour data sets, and ambiguous samples were identified as having a likelihood below 0.995. Each interval of consecutive ambiguous samples was checked for the presence of state transitions.*

Figure 6 – Figure Supplement 4. Frequencies of successful and failed transitions. *The analysis in Figure 6D) was repeated using a more conservative threshold to identify ambiguous samples (P(predicted state) < 0.95). Relative frequencies of successful state transitions (left), failed state transitions (middle) and the ratio between these (right) are shown. Values indicate mean ± standard deviation.*

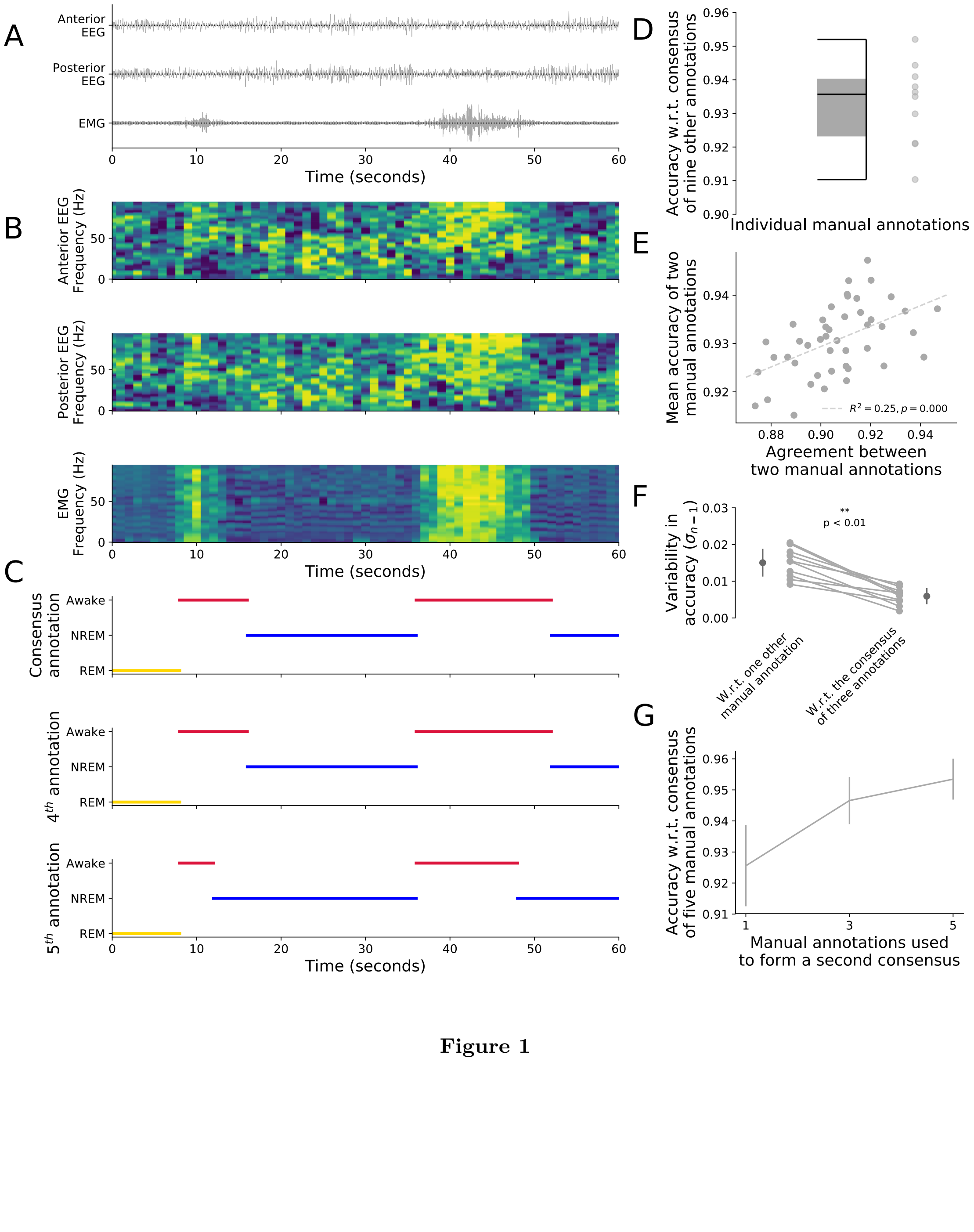

A
Anterior EEG
Posterior EEG
EMG
0
10
20
30
40
50
60
Time (seconds)
B
Anterior EEG Frequency (Hz)
Posterior EEG Frequency (Hz)
EMG Frequency (Hz)
0
50
Time (seconds)
C
Consensus annotation
4th annotation
5th annotation
Awake
NREM
REM
Time (seconds)
D
Accuracy w.r.t. consensus of nine other annotations
0.90
0.91
0.92
0.93
0.94
0.95
0.96
Individual manual annotations
E
Mean accuracy of two manual annotations
0.92
0.93
0.94
0.88
0.90
0.92
0.94
R^2 = 0.25, p = 0.000
Agreement between two manual annotations
F
Variability in accuracy (σ_{n-1})
0.00
0.01
0.02
0.03
**
p < 0.01
W.r.t. one other manual annotation
W.r.t. the consensus of three annotations
G
Accuracy w.r.t. consensus of five manual annotations
0.91
0.92
0.93
0.94
0.95
0.96
1
3
5
Manual annotations used to form a second consensus

Figure 1

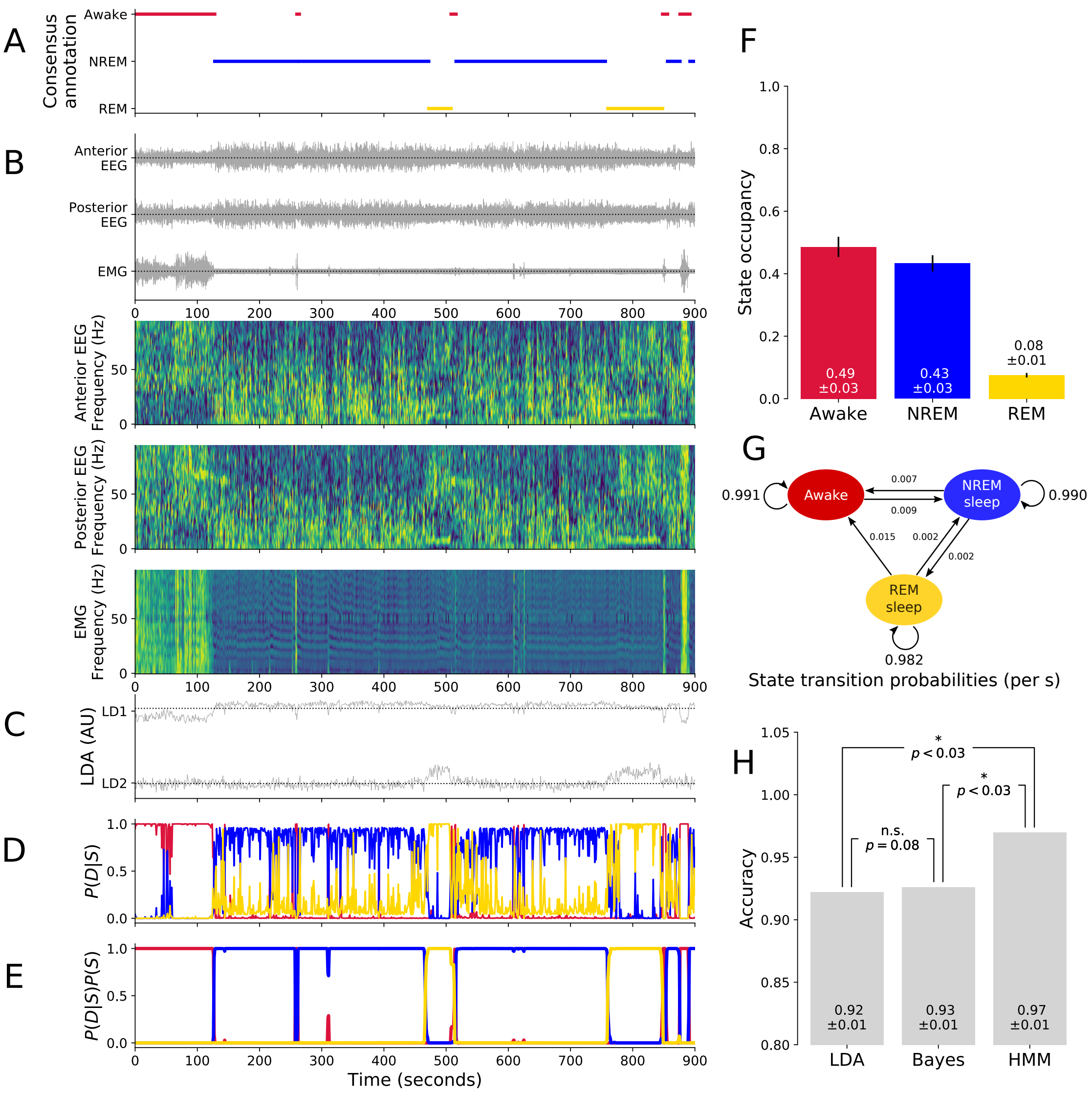

A
Consensus annotation
Awake
NREM
REM
B
Anterior EEG
Posterior EEG
EMG
0
100
200
300
400
500
600
700
800
900
Anterior EEG Frequency (Hz)
0
50
Posterior EEG Frequency (Hz)
0
50
EMG Frequency (Hz)
0
50
0
100
200
300
400
500
600
700
800
900
C
LDA (AU)
LD1
LD2
D
$P(D|S)$
0.0
0.5
1.0
E
$P(D|S)P(S)$
0.0
0.5
1.0
0
100
200
300
400
500
600
700
800
900
Time (seconds)
F
State occupancy
0.0
0.2
0.4
0.6
0.8
1.0
0.49
±0.03
0.43
±0.03
0.08
±0.01
Awake
NREM
REM
G
Awake
NREM sleep
REM sleep
0.991
0.990
0.982
0.007
0.009
0.015
0.002
0.002
State transition probabilities (per s)
H
Accuracy
0.80
0.85
0.90
0.95
1.00
1.05
*
$p < 0.03$
*
$p < 0.03$
n.s.
$p = 0.08$
0.92
±0.01
0.93
±0.01
0.97
±0.01
LDA
Bayes
HMM

Figure 2

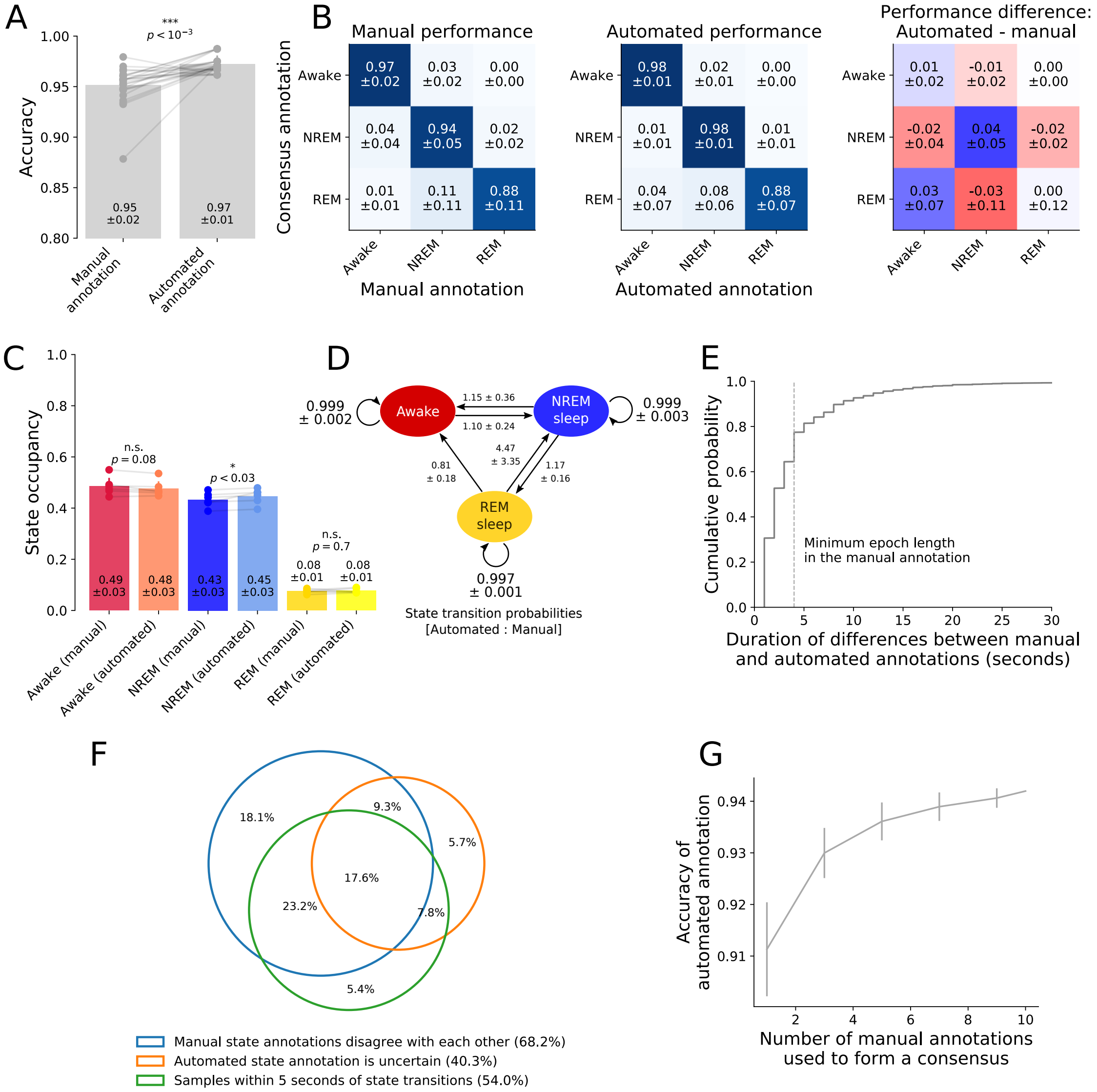

A
***
$p < 10^{-3}$
Accuracy
1.00
0.95
0.90
0.85
0.80
0.95 ±0.02
0.97 ±0.01
Manual annotation
Automated annotation
B
Manual performance
Automated performance
Performance difference: Automated - manual
Consensus annotation
Awake
NREM
REM
0.97 ±0.02
0.03 ±0.02
0.00 ±0.00
0.04 ±0.04
0.94 ±0.05
0.02 ±0.02
0.01 ±0.01
0.11 ±0.11
0.88 ±0.11
Manual annotation
0.98 ±0.01
0.02 ±0.01
0.00 ±0.00
0.01 ±0.01
0.98 ±0.01
0.01 ±0.01
0.04 ±0.07
0.08 ±0.06
0.88 ±0.07
Automated annotation
0.01 ±0.02
-0.01 ±0.02
0.00 ±0.00
-0.02 ±0.04
0.04 ±0.05
-0.02 ±0.02
0.03 ±0.07
-0.03 ±0.11
0.00 ±0.12
C
State occupancy
1.0
0.8
0.6
0.4
0.2
0.0
n.s.
$p = 0.08$
*
$p < 0.03$
n.s.
$p = 0.7$
0.49 ±0.03
0.48 ±0.03
0.43 ±0.03
0.45 ±0.03
0.08 ±0.01
0.08 ±0.01
Awake (manual)
Awake (automated)
NREM (manual)
NREM (automated)
REM (manual)
REM (automated)
D
Awake
NREM sleep
REM sleep
0.999 ± 0.002
0.999 ± 0.003
0.997 ± 0.001
1.15 ± 0.36
1.10 ± 0.24
4.47 ± 3.35
0.81 ± 0.18
1.17 ± 0.16
State transition probabilities [Automated : Manual]
E
Cumulative probability
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Minimum epoch length in the manual annotation
Duration of differences between manual and automated annotations (seconds)
F
18.1%
9.3%
5.7%
17.6%
23.2%
7.8%
5.4%
Manual state annotations disagree with each other (68.2%)
Automated state annotation is uncertain (40.3%)
Samples within 5 seconds of state transitions (54.0%)
G
Accuracy of automated annotation
0.94
0.93
0.92
0.91
2
4
6
8
10
Number of manual annotations used to form a consensus

Figure 3

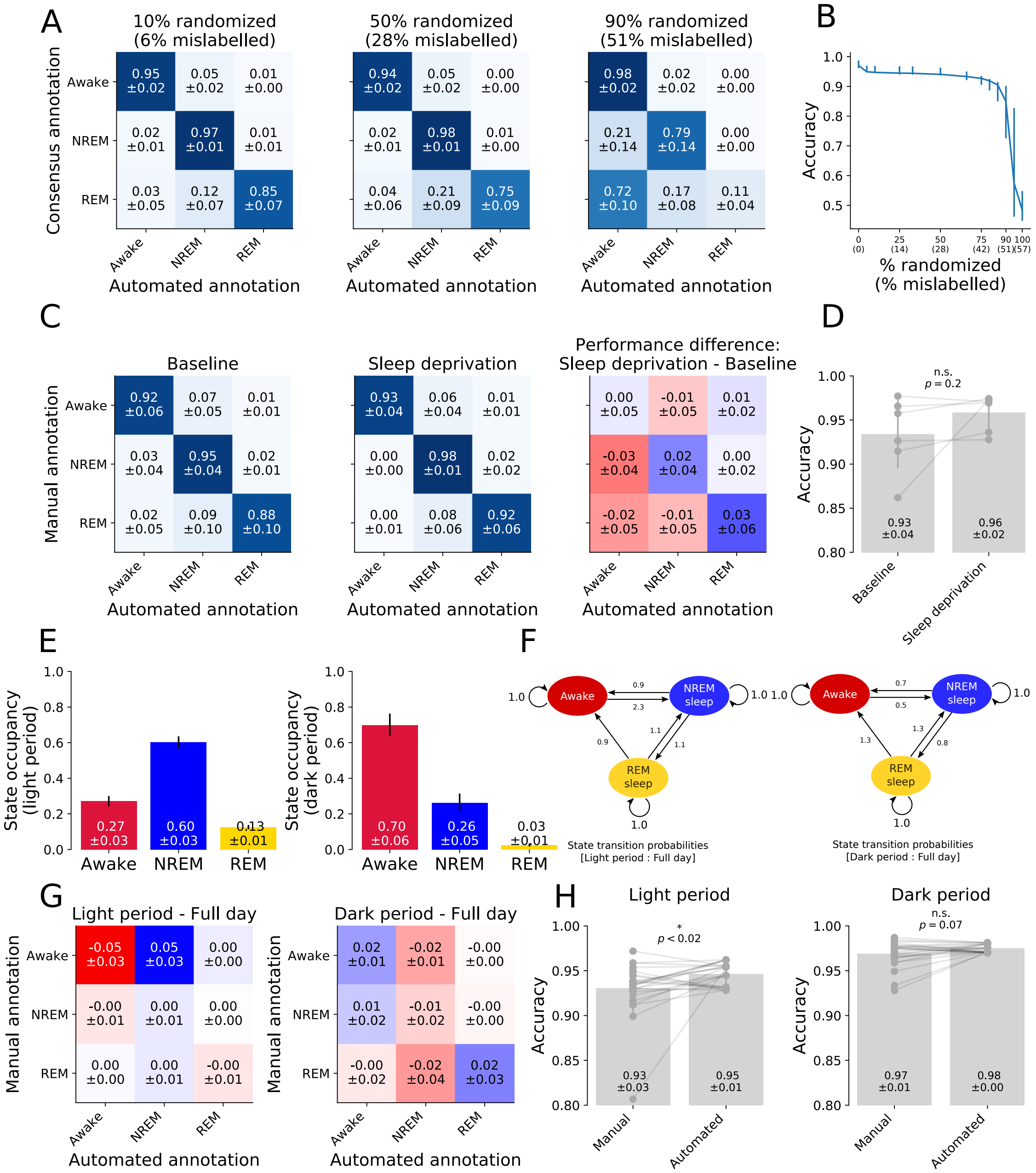

A
10% randomized
(6% mislabelled)
Consensus annotation
Awake
NREM
REM
0.95 ±0.02
0.05 ±0.02
0.01 ±0.00
0.02 ±0.01
0.97 ±0.01
0.01 ±0.01
0.03 ±0.05
0.12 ±0.07
0.85 ±0.07
Automated annotation
50% randomized
(28% mislabelled)
0.94 ±0.02
0.05 ±0.02
0.00 ±0.00
0.02 ±0.01
0.98 ±0.01
0.01 ±0.00
0.04 ±0.06
0.21 ±0.09
0.75 ±0.09
90% randomized
(51% mislabelled)
0.98 ±0.02
0.02 ±0.02
0.00 ±0.00
0.21 ±0.14
0.79 ±0.14
0.00 ±0.00
0.72 ±0.10
0.17 ±0.08
0.11 ±0.04
B
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0 (0)
25 (14)
50 (28)
75 (42)
90 (51)
100 (57)
% randomized
(% mislabelled)
C
Baseline
Manual annotation
0.92 ±0.06
0.07 ±0.05
0.01 ±0.01
0.03 ±0.04
0.95 ±0.04
0.02 ±0.01
0.02 ±0.05
0.09 ±0.10
0.88 ±0.10
Sleep deprivation
0.93 ±0.04
0.06 ±0.04
0.01 ±0.01
0.00 ±0.00
0.98 ±0.01
0.02 ±0.02
0.00 ±0.01
0.08 ±0.06
0.92 ±0.06
Performance difference:
Sleep deprivation - Baseline
0.00 ±0.05
-0.01 ±0.05
0.01 ±0.02
-0.03 ±0.04
0.02 ±0.04
0.00 ±0.02
-0.02 ±0.05
-0.01 ±0.05
0.03 ±0.06
D
n.s.
p = 0.2
1.00
0.95
0.90
0.85
0.80
0.93 ±0.04
0.96 ±0.02
Baseline
Sleep deprivation
E
State occupancy
(light period)
0.27 ±0.03
0.60 ±0.03
0.13 ±0.01
State occupancy
(dark period)
0.70 ±0.06
0.26 ±0.05
0.03 ±0.01
0.0
0.2
0.4
0.6
0.8
1.0
F
Awake
NREM sleep
REM sleep
1.0
0.9
2.3
1.1
1.1
0.9
State transition probabilities
[Light period : Full day]
0.7
0.5
1.3
0.8
1.3
State transition probabilities
[Dark period : Full day]
G
Light period - Full day
-0.05 ±0.03
0.05 ±0.03
0.00 ±0.00
-0.00 ±0.01
0.00 ±0.01
0.00 ±0.00
0.00 ±0.00
0.00 ±0.01
-0.00 ±0.01
Dark period - Full day
0.02 ±0.01
-0.02 ±0.01
-0.00 ±0.00
0.01 ±0.02
-0.01 ±0.02
-0.00 ±0.00
-0.00 ±0.02
-0.02 ±0.04
0.02 ±0.03
H
Light period
*
p < 0.02
0.93 ±0.03
0.95 ±0.01
Manual
Automated
Dark period
n.s.
p = 0.07
0.97 ±0.01
0.98 ±0.00

Figure 4

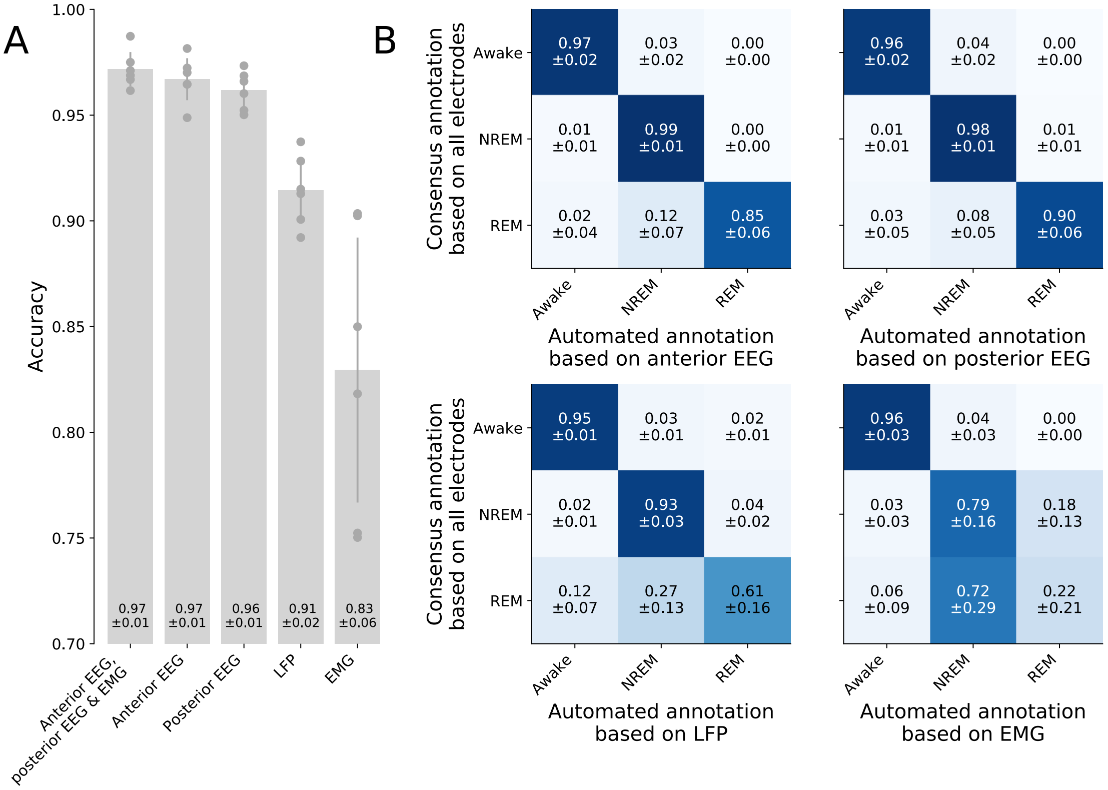

A
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
0.97 ±0.01
0.97 ±0.01
0.96 ±0.01
0.91 ±0.02
0.83 ±0.06
Anterior EEG, posterior EEG & EMG
Anterior EEG
Posterior EEG
LFP
EMG
B
Consensus annotation based on all electrodes
Awake
NREM
REM
0.97 ±0.02
0.03 ±0.02
0.00 ±0.00
0.01 ±0.01
0.99 ±0.01
0.00 ±0.00
0.02 ±0.04
0.12 ±0.07
0.85 ±0.06
Awake
NREM
REM
Automated annotation based on anterior EEG
0.96 ±0.02
0.04 ±0.02
0.00 ±0.00
0.01 ±0.01
0.98 ±0.01
0.01 ±0.01
0.03 ±0.05
0.08 ±0.05
0.90 ±0.06
Awake
NREM
REM
Automated annotation based on posterior EEG
Consensus annotation based on all electrodes
Awake
NREM
REM
0.95 ±0.01
0.03 ±0.01
0.02 ±0.01
0.02 ±0.01
0.93 ±0.03
0.04 ±0.02
0.12 ±0.07
0.27 ±0.13
0.61 ±0.16
Awake
NREM
REM
Automated annotation based on LFP
0.96 ±0.03
0.04 ±0.03
0.00 ±0.00
0.03 ±0.03
0.79 ±0.16
0.18 ±0.13
0.06 ±0.09
0.72 ±0.29
0.22 ±0.21
Awake
NREM
REM
Automated annotation based on EMG

Figure 5

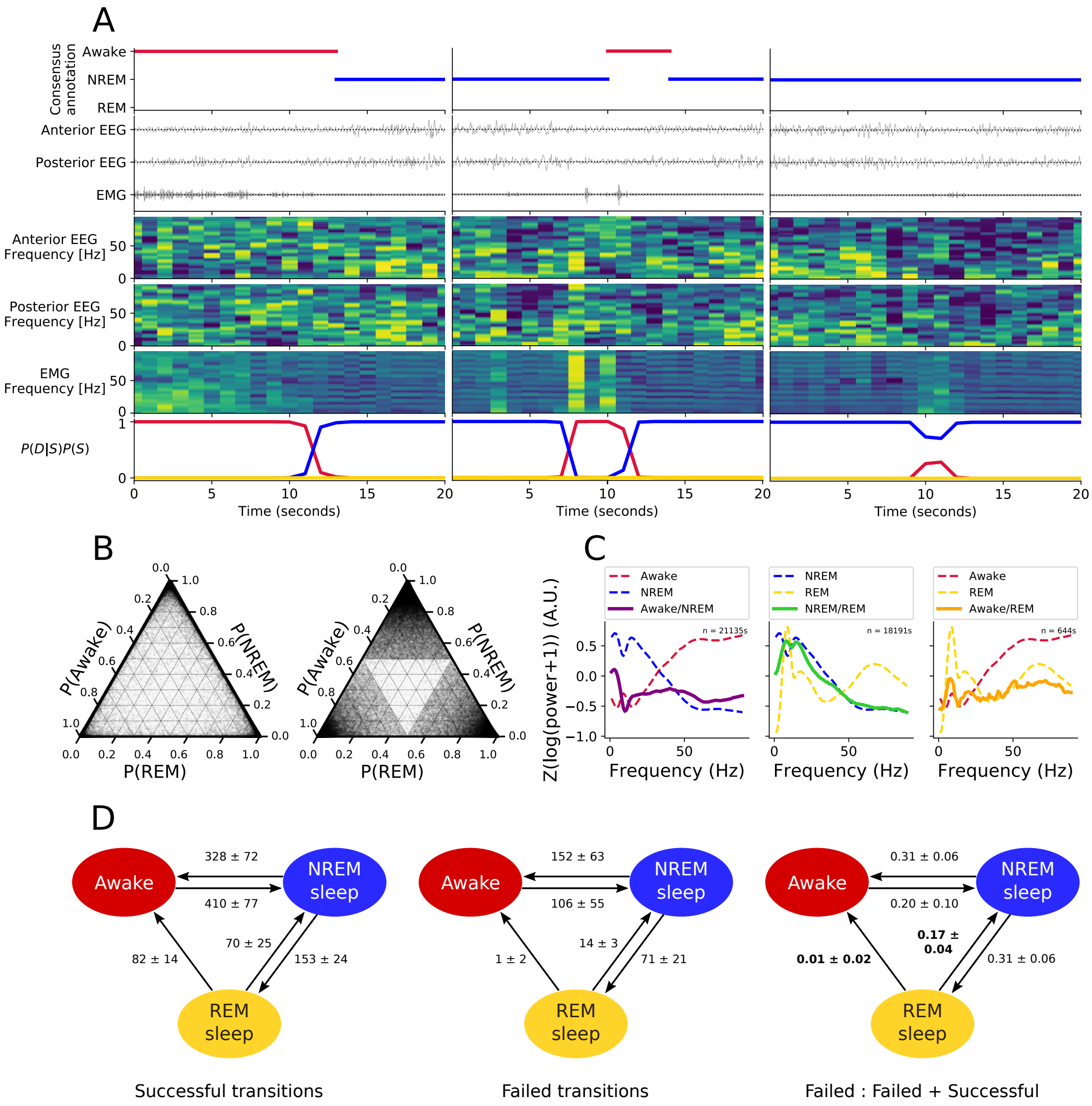

A
Consensus annotation
Awake
NREM
REM
Anterior EEG
Posterior EEG
EMG
Anterior EEG Frequency [Hz]
Posterior EEG Frequency [Hz]
EMG Frequency [Hz]
0
50
P(D|S)P(S)
1
0
0
5
10
15
20
Time (seconds)
B
P(Awake)
P(NREM)
P(REM)
0.0
0.2
0.4
0.6
0.8
1.0
C
Awake
NREM
Awake/NREM
REM
NREM/REM
Awake/REM
n = 21135s
n = 18191s
n = 644s
Z(log(power+1)) (A.U.)
0.5
0.0
−0.5
−1.0
0
50
Frequency (Hz)
D
Awake
NREM sleep
REM sleep
328 ± 72
410 ± 77
82 ± 14
70 ± 25
153 ± 24
Successful transitions
152 ± 63
106 ± 55
1 ± 2
14 ± 3
71 ± 21
Failed transitions
0.31 ± 0.06
0.20 ± 0.10
0.01 ± 0.02
0.17 ± 0.04
0.31 ± 0.06
Failed : Failed + Successful

Figure 6

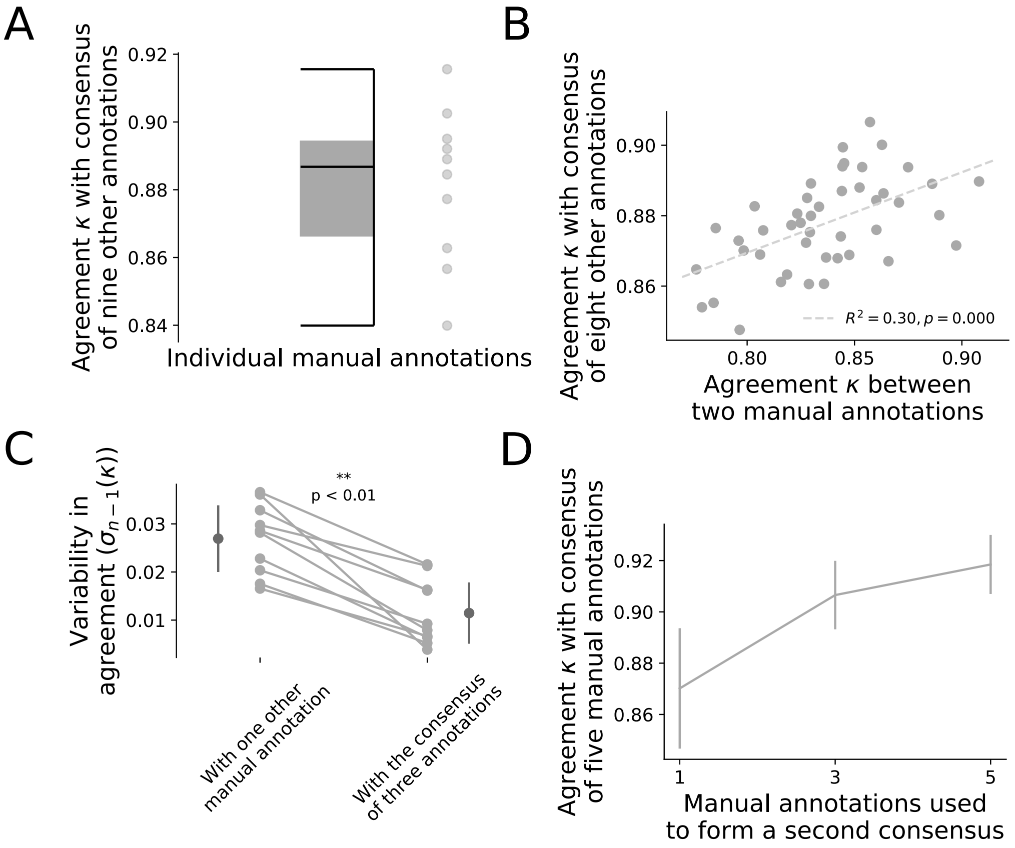
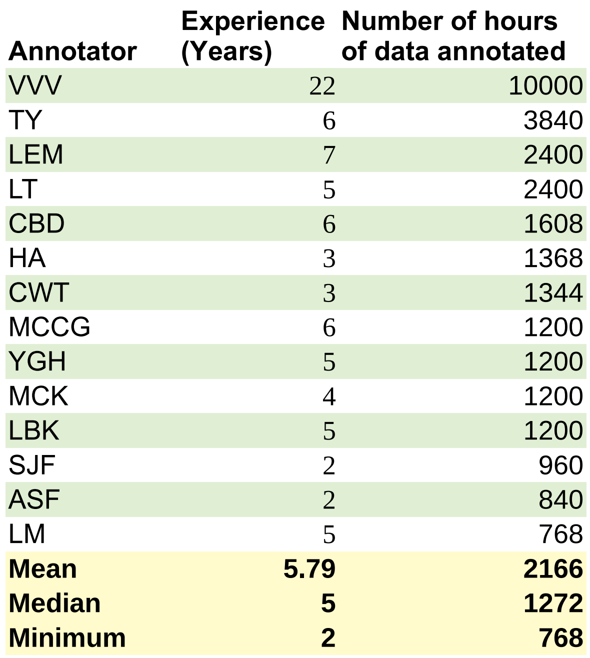


Figure 1 - Figure Supplement 1



Annotator	Experience (Years)	Number of hours of data annotated
VVV	22	10000
TY	6	3840
LEM	7	2400
LT	5	2400
CBD	6	1608
HA	3	1368
CWT	3	1344
MCCG	6	1200
YGH	5	1200
MCK	4	1200
LBK	5	1200
SJF	2	960
ASF	2	840
LM	5	768
Mean	**5.79**	**2166**
Median	**5**	**1272**
Minimum	**2**	**768**

Figure 3 - Figure Supplement 1

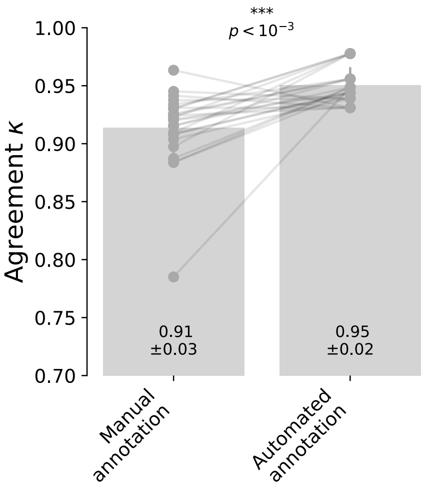


Figure 3 - Figure Supplement 2

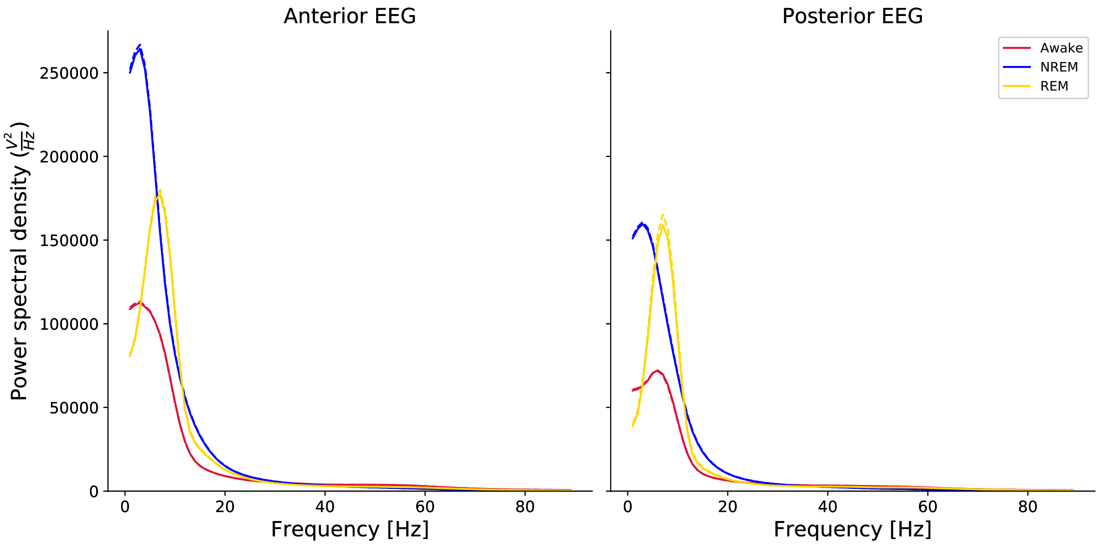


Figure 3 - Figure Supplement 3

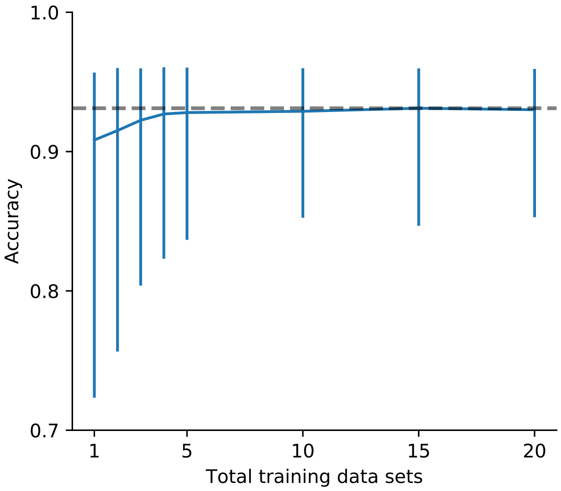


Figure 4 - Figure Supplement 1

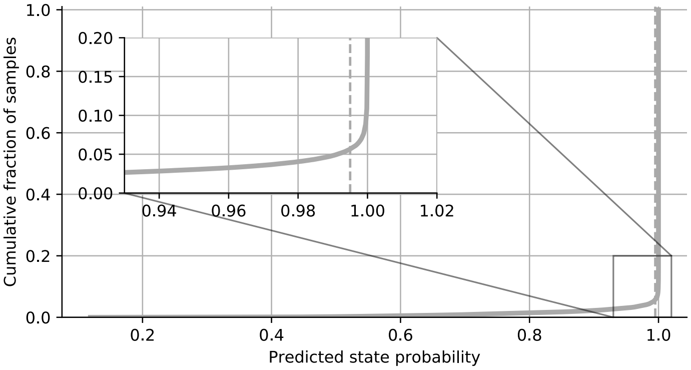


Figure 6 - Figure Supplement 1

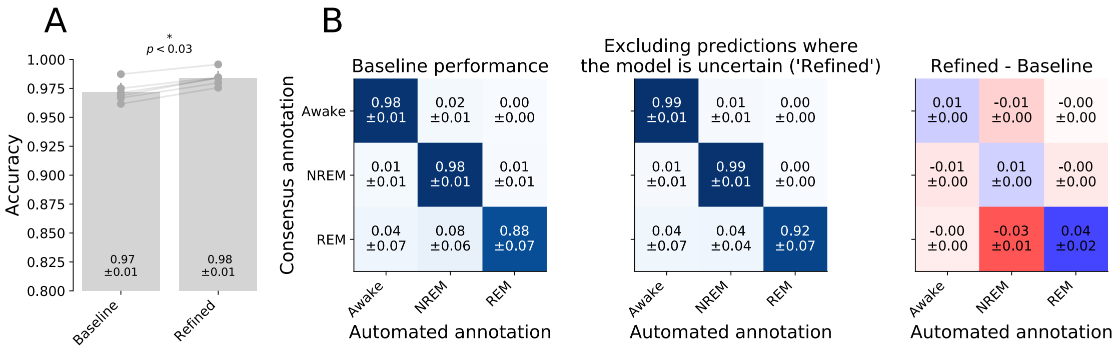


Figure 6 - Figure Supplement 2

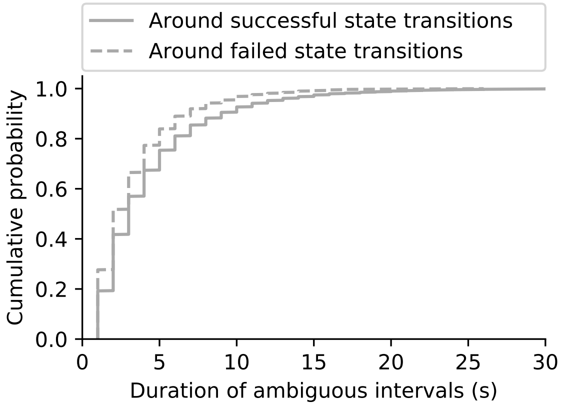


Figure 6 - Figure Supplement 3

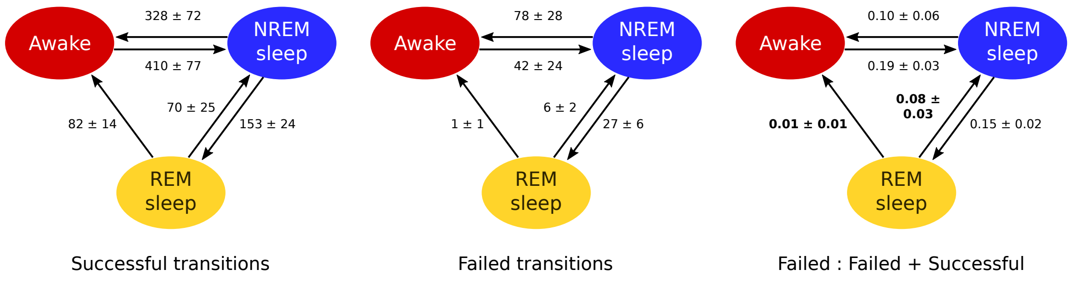


Figure 6 - Figure Supplement 4