Sleep Advances, 2025, **6**, zpaf025
https://doi.org/10.1093/sleepadvances/zpaf025
Advance access publication 10 May 2025
Original Article

Original Article

Probability estimation of narcolepsy type 1 in DTA mice using unlabeled EEG and EMG data

Laura Rose[1], Alexander Neergaard Zahid[2], Louise Piilgaard[1], Christine Egebjerg[1], Frederikke Lynge Sørensen[3], Mie Andersen[3], Tessa Radovanovic[3], Anastasia Tsopanidou[3], Stefano Bastianini[4], Chiara Berteotti[4], Viviana Lo Martire[4], Micaela Borsa[5,6], Ryan K. Tisdale[7,11], Yu Sun[7], Maiken Nedergaard[3], Alessandro Silvani[4], Giovanna Zoccoli[4], Antoine Adamantidis[5,6], Thomas S. Kilduff[7], Noriaki Sakai[8], Seiji Nishino[8], Sébastien Arthaud[9], Christelle Peyron[9], Patrice Fort[9], Morten Mørup[2], Emmanuel Mignot[10,¥] and Birgitte Rahbek Kornum[1,*,¥]

[1]Department of Neuroscience, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[3]Division of Glial Disease and Therapeutics, Center for Translational Neuromedicine, University of Copenhagen, Copenhagen, Denmark
[4]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy
[5]Zentrum für Experimentelle Neurologie, Department of Neurology, University Hospital Bern, Bern, Switzerland
[6]Department of Biomedical Research, University of Bern, Bern, Switzerland
[7]Center for Neuroscience, Biosciences Division, SRI International, Menlo Park, CA, USA
[8]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA
[9]Center for Neuroscience Research in Lyon (CRNL), SLEEP Team, Université de Lyon, Lyon, France
[10]Center for Narcolepsy, Department of Psychiatry, Stanford University School of Medicine, Stanford, CA, USA
[11]Present address: F. Hoffmann-La Roche Ltd, Basel, Switzerland

[¥]The authors contributed equally to this manuscript.
[*]Corresponding author: Birgitte R. Kornum, Department of Neuroscience, University of Copenhagen, Blegdamsvej 3B, 24-6-14, 2200 Copenhagen N, Denmark. Email: kornum@sund.ku.dk

Abstract

The manual evaluation of mouse sleep studies is labor-intensive and time-consuming. Although several approaches for automatic sleep stage classification have been proposed, no automatic pipeline for detecting a specific mouse phenotype has yet been developed. Here, we present a fully automated pipeline for estimating the probability of Narcolepsy Type 1 (NT1) in the hypocretin-tTA; TetO-Diphteria toxin A (DTA) mouse model using unlabeled electroencephalographic (EEG) and electromyographic (EMG) data. The pipeline is divided into three modules: (1) automatic sleep stage classification, (2) feature extraction, and (3) phenotype classification. We trained two automatic sleep stage classifiers, Usleep$_{EEG}$ and Usleep$_{EMG}$, using data from 83 wild-type (WT) mice. We next computed features such as EEG spectral power bands, EMG root mean square, and bout metrics from 11 WT and 19 DTA mice. The features were used to train an L1-penalized logistic regression classifier in a Leave-One-Subject-Out approach, achieving an accuracy of 97%. Finally, we validated the pipeline in a held-out dataset of EEG/EMG recordings at four different timepoints during disease development in seven DTA mice, finding that the pipeline captured disease progression in all mice. While our pipeline generalizes well to data from other laboratories, it is sensitive to artifacts, which should be considered in its application. With this study, we present a pipeline that facilitates a fast assessment of NT1 probability in the DTA model and thus can accelerate large-scale evaluations of NT1 treatments.

Key words: automated phenotype detection; mouse model; narcolepsy; sleep; EEG; EMG

Statement of Significance

In this study, we developed the first fully automated pipeline to estimate the probability of NT1 in DTA mice using unlabeled EEG and EMG data. Our pipeline facilitates a fast and accurate assessment of the NT1 probability, with the potential to accelerate large-scale evaluations. This advancement will significantly enhance preclinical research and improve the efficiency of treatment evaluations. The work establishes a new benchmark in the field and could inspire others to extend this approach to the study of other diseases potentially driving advancements in drug development across various disease models.

Submitted for publication: January 19, 2025; **Revised:** April 5, 2025

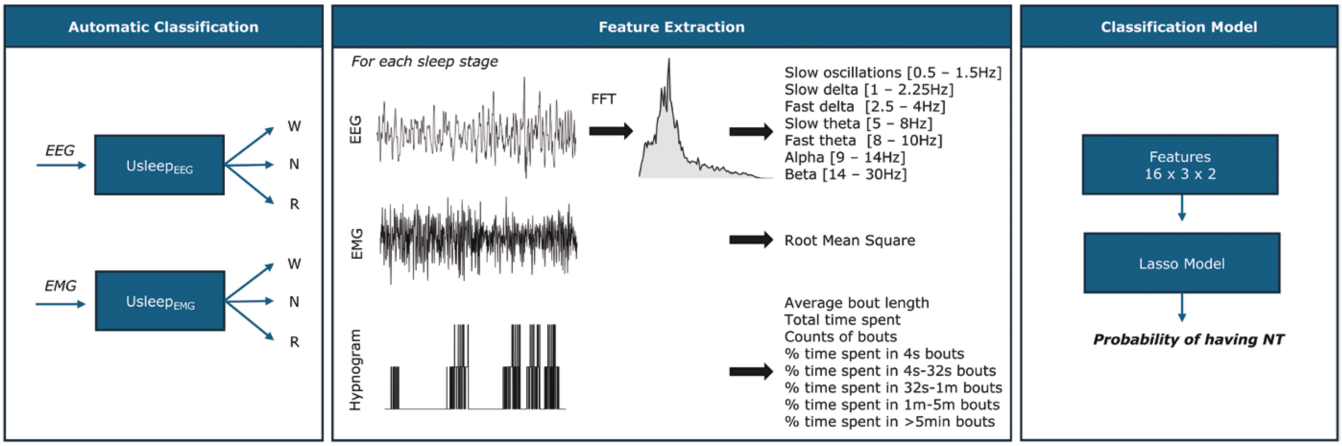


Figure 1. A fully automated pipeline for probability estimation of NT1 in mice from unlabeled EEG and EMG data. Flowchart of pipeline. In the first step, raw EEG and EMG data are fed into two automatic classifiers UsleepEEG and UsleepEMG. Each classifier outputs the probability of either wakefulness, NREM sleep and REM sleep, yielding two sets of predicted hypnograms yEEG and yEMG. Next, for each y and each sleep stage, we extract 16 features. Seven of these features are relative EEG power bands (slow oscillations, slow delta, fast delta, slow theta, fast theta, alpha, and beta). Additionally, the root mean square reflects the amplitude of the EMG signal. The remaining eight features come from the hypnogram of y and include: average bout length [s], total time spent in the sleep stage [%], counts of bouts per hour, total time spent in 4-second bouts [%], total time spent in bouts with duration from 4-s to –32s [%], total time spent in bouts with duration 32-s to –1-min [%], total time spent in bouts with duration 1 min to 5 min [%] and total time spent in > 5-min bouts [%]. In total, 16 × 3 × 2 features are obtained and are then fed into an NT1 classifier model that outputs the probability of having NT1.

Introduction

Narcolepsy type 1 (NT1) is a chronic sleep disorder characterized by excessive daytime sleepiness, fragmented sleep-wake patterns, symptoms of dissociated REM sleep, and cataplexy (i.e. a sudden loss in muscle tone triggered by strong positive emotions) [1, 2]. NT1 is hypothesized to be an autoimmune disease caused by the loss of hypocretin/orexin neurons in the lateral hypothalamus [3–5]. The prevalence of NT1 is 25–50 cases pr. 100 000 individuals. Symptomatic treatment options exist, however, none of the approved treatments result in full symptom relief, and no treatment is yet developed to hinder or slow down the disease progression [6, 7].

Preclinical mouse models have been used to study sleep disorders and their response to treatment. Current mouse models for NT1 include the genetic *Hcrt*-knockout (HCRT-KO) model lacking the HCRT peptides [8], the transgenic neuron-ablated Ataxin-3 model that gradually loses HCRT neurons from birth [9], and the double transgenic conditional neuron-ablated *orexin-tTA;TetO diphtheria toxin A* (DTA) model [10]. In the DTA model, HCRT neuron-specific DTA expression is controlled by dietary doxycycline (DOX), with DOX withdrawal initiating a gradual loss of HCRT neurons [10]. Although these NT1 mouse models exhibit high face validity, with a phenotype that includes sleep-wake fragmentation and cataplexy-like attacks, the DTA model shows better construct validity, with presumed HCRT neuron loss [4, 6, 11] and a typical post-pubertal or adult-onset [6, 11].

In mouse sleep studies, EEG, EMG, and video signals are recorded to measure brain activity, muscle activity, and behavior, respectively. Conducting sleep studies is highly labor-intensive and time-consuming, as they involve surgery for EEG/EMG recordings and manual evaluation of data to classify sleep/wake states to further determine if the mouse exhibits an NT1 phenotype. To increase efficiency, several automatic sleep stage classification models have been proposed for mouse studies (see review by Rayan et al. for more detailed information [12]). The time-consuming aspect of manual evaluation does not only apply to mouse sleep but also human sleep. In the human domain automatic sleep stage classification and automatic detection of the narcolepsy phenotype in polysomnography (PSG) data has successfully been demonstrated by Stephansen and colleagues [13]. The paper presents a fast and automated approach for narcolepsy diagnosis through features extracted from what was then coined hypnodensity.

Here, we developed the first fully automated pipeline for estimating the probability of NT1 in DTA mice based on unlabeled EEG and EMG data (Figure 1). As narcolepsy is characterized by symptoms of dissociated REM sleep (muscle atonia appearing in wakefulness or high muscle tone in REM sleep), a novel aspect of our approach was to train a sleep stage classification model using either only the EEG or the EMG, employing the resulting probabilities for feature extraction and NT1 prediction. For comparison, we included a sleep stage classifier that was trained on the combined EMG and EEG signals. We hope that this approach will facilitate pre-clinical research into NT1 pathogenesis and symptomatology and importantly allow for faster evaluation of novel drug treatments in the DTA mouse model.

Methods

Data collection overview

We collected mouse EEG and EMG data from seven different laboratories, resulting in eight cohorts labeled A to H (Table 1). Cohorts A to E were used to train three automatic classifiers (Usleep$_{EEG}$, Usleep$_{EMG,}$ Usleep$_{EEG,EMG}$) for sleep stage classification. Cohort F was used to train an L1-penalized logistic regression classifier for NT1 probability estimation, while cohorts G to H served as held-out datasets for external validation of the pipeline. Each sleep recording includes at least one EEG and one EMG channel, with electrode placement detailed in Table 1. The recording lengths vary across mice and some mice have multiple recordings. For cohorts A to F, epochs were manually annotated by trained experts from the corresponding sleep laboratories as either one of the three states: wakefulness, NREM sleep, or REM sleep. Since the automatic classifiers are trained on healthy mice, they can only detect wakefulness, NREM sleep, and REM stages. Consequently, when applied to NT1 mice, the model cannot detect cataplexy or delta attacks. These epochs were masked during the validation analysis (Figure 2), but retained in the training of the NT1 classifier, with the models classifying them as one of the three main stages.

Experimental data acquisition

Details on animal procedures and data collection of cohorts A-H from the corresponding labs are separately described below. All animal studies have been approved by the respective national

Table 1. Overview of data collection

Cohort	Lab	WT	DTA	EEG	EMG	Usleep	Lasso	Test
A	1	10	0	1 ipsilateral fronto-parietal differential	Neck	X		
B	2	17	0	2 parietal 2 frontal	Neck	X		
C	3	23	0	1 parietal 1 frontal	Neck	X		
D	4	28	0	1 parietal / 1frontal or 1 cerebellum / 1 frontal	Neck	X		
E	5	5	0	1 parietal 1 frontal	Neck	X		
F	3	11	19	1 parietal 1 frontal	Neck		X	
G	6	0	7	1 Parietal- Interparietal differential	Neck			X
H	7	8	6	2 parietal 2 frontal	Neck			X

All data were downsampled to 128 Hz and had sleep annotations in four second windows. Cohort A-E is used to train $\text{Usleep}_{\text{EEG}}$ and $\text{Usleep}_{\text{EMG}}$, cohort F is used to train a NT1 classifier, while cohort G-H are used for validation.

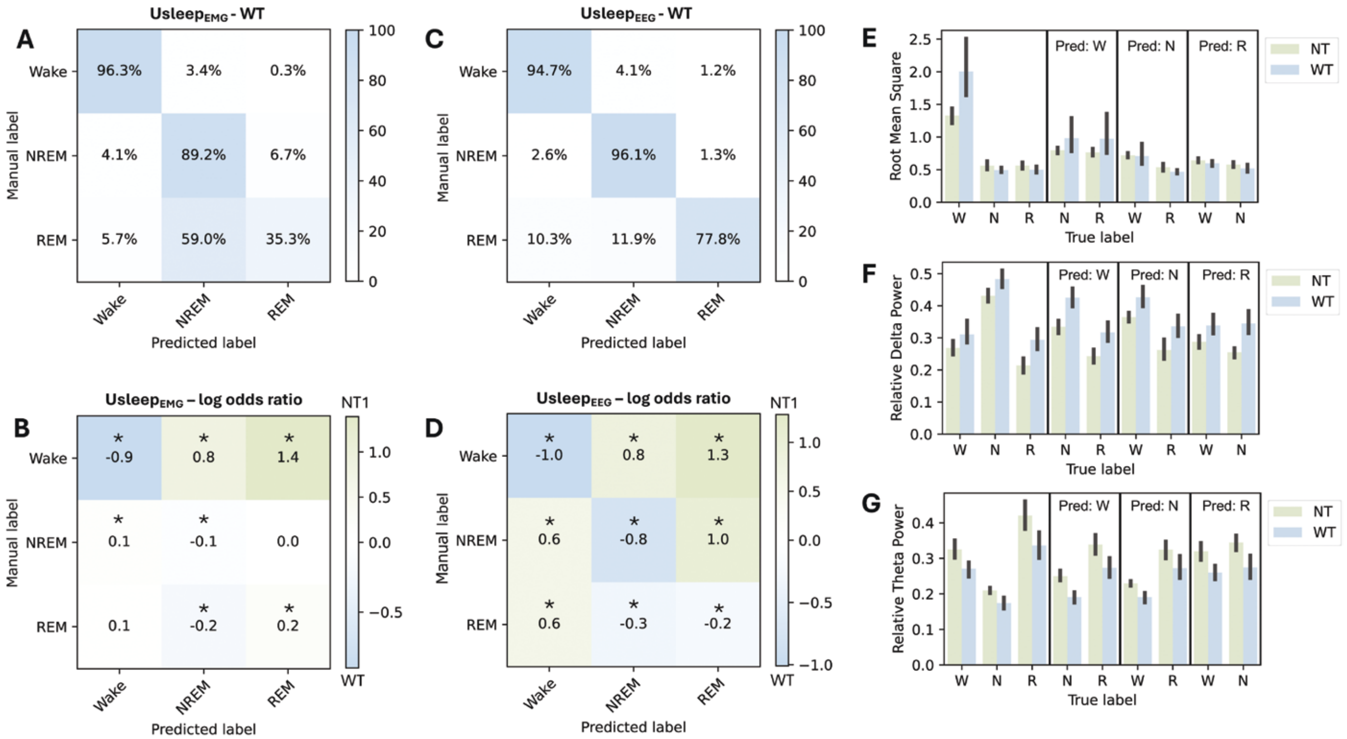


Figure 2. Models trained on one modality can be used to predict wakefulness and sleep in WT and NT1 mice. (A) Confusion matrix of UsleepEMG with row-wise normalization (recall) tested in a held-out test set of 11 WT mice. (B) The log-odds ratio of the confusion matrices of UsleepEMG from 11 WT and 19 NT mice. Blue color indicates a higher likelihood of the event occurring in WT mice and green color indicates a higher likelihood of the event occurring in NT mice. * Significant difference between WT and NT (CI of the log-odds ratio (LOR) does not include 0). (C) Confusion matrix of UsleepEEG with row-wise normalization (recall) tested in the held-out WT mice cohort. (D) The log-odds ratio of the confusion matrix from WT and NT mice for UsleepEEG. (E) Average root mean square for NT (green) and WT (blue) mice and W = Wakefulness, N = NREM sleep and R = REM sleep. First window includes epochs where the model and the expert agree, the remaining windows are disagreement epochs where manual label and predicted label are not the same. (F) Bar plot of the average relative delta power across genotypes (G) Bar plot of the average relative theta power across genotypes.

authorities and carried out according to ethical guidelines (European Communities Council Directive (86/609/EEC)) and ARRIVE (Animal Research: Reporting In Vivo Experiments).

Dataset from Cohort A. Cohort A consists of 10 WT male mice (C57BL/6J background, 15.0 ± 0.4 weeks of age at surgery). The data and experimental procedures of this cohort have previously been published [14] (WT control group). The study protocol was approved by the Bologna University ethics committee. For cohort A sleep scoring was performed on 4-second epochs by expert investigators using a validated semi-automated procedure (SCOPRISM [15]) on raw EEG and EMG data. Investigators corrected the automated scoring result if needed based on the visualization of raw EEG and EMG recordings.

Dataset from Cohort B: Cohort B consists of 17 WT male mice (6–15 weeks of age, C57BL/6JRj background, Janvier Labs, Le Genest-Saint-Isle, France). The mice underwent intracerebral injections of a viral vector expressing a calcium sensor, GCaMP6, under control of the HCRT promoter two weeks prior to EEG/EMG implantation, and an optical fiber had been implanted just above the lateral hypothalamus intended for calcium imaging. This data is not used for the present study. The experimental procedure for EEG/EMG recordings was similar to what has previously been published in [16]. All experimental procedures were approved by the Veterinary Office of the Canton of Bern, Switzerland (License number BE 45/18). For cohort B scoring of the different vigilance stages (wakefulness, NREM sleep, REM sleep) was conducted in 1-second epochs using custom-written MATLAB scripts.

Dataset from Cohort C and F. Cohort C consists of 23 wildtype (WT) mice (12 females, 4-15 weeks of age, C57BL/6 background) and cohort F consists of 11 WT (10 females, 4-15 weeks of age, C57BL/6 background) and 19 DTA mice (7 females, 10-15 weeks of age, double transgenic C57BL/6-Tg (Hcrt/tTA; TetO DTA background). Mice were either purchased from Taconic Biosciences (C57BL6/J6NTac; Ejby, Denmark), Janvier Labs (C57BL/6JRj; Le Genest-Saint-Isle, France), or bred in-house as part of a transgenic breeding program.

The experimental procedures of these cohorts have previously been published in [17–19]. All experiments were approved by the Danish Animal Experiments Inspectorate (license #2019-15-0201-00016). Wakefulness, NREM sleep, and REM sleep were determined in four seconds epochs according to standard criteria.

Dataset from Cohort D. Cohort D consists of 28 male mice (WT or TH-Cre mice,12-24 weeks of age at the time of recording; C57BL/6 background; Janvier Labs or bred in-house). The experimental procedures of this cohort have previously been published [20] and a subset of the collected data has previously been published [21]. All experiments were approved by the Danish Animal Experiments Inspectorate. For cohort D sleep state scoring (wakefulness, NREM sleep, and REM sleep) was performed manually using SleepScore in either one or four-second epochs based on standard criteria for EEG and EMG recordings with the assistance of video.

Dataset from Cohort E. Cohort E consists of five WT male mice (8–26 weeks of age at recording; C57BL/6J genetic background, three from Charles River Laboratories, Les Oncins, France and two donated by Pr. Miquel Vila). Experimental procedures and data from two WT male mice have previously been published in [22]. All experiments were approved by either the Université Claude Bernard Lyon 1 Ethic Committee (C2EA-055; #DR-2015-42) or the CELYNE Ethics Research Committee (C2EA-042; APAFIS#20701). For cohort E, vigilance states were visually scored using a 5-second sliding window frame and assigned as either wakefulness, NREM sleep, or REM sleep.

Dataset from Cohort G: Seven NT1 DTA male mice were double transgenic offspring of Hcrt/tTA mice (C57BL/6-Tg(Hcrt/tTA)/Yamanaka) and B6.Cg-Tg(tetO DTA) 1Gfi/J mice (JAX #008468). Both parental strains were from a C57BL/6J genetic background. Parental strains and offspring used for EEG/EMG recording were maintained on a diet (Envigo T-7012, 200 DOXycycline) containing DOX (DOX(+) condition) to repress transgene expression until neurodegeneration was desired. Mice were maintained on normal chow for six weeks. The data and experimental procedures of this cohort have previously been published [23]. All experimental procedures were approved by the Institutional Animal Care and Use Committee at SRI International.

Dataset from Cohort H: Cohort H consists of eight WT males (males, 15 weeks of age, C57BL/6 background) and six NT1 male mice (15 weeks of age, double transgenic C57BL/6-Tg (hcrt/tTA;TetO DTA background)). Mice were transferred from Taconic Biosciences (Hudson, NY, USA). Detailed information about EEG/EMG surgery and data acquisition can be found in Sakai et al. [24]. All experiments were approved by the Stanford University Administrative Panel on Laboratory Animal Care and were conducted in accordance with the Stanford University Administrative Panel on Laboratory Animal Care Guidelines (APLAC-#21,646).

Fully automated pipeline for NT1 probability estimation

As illustrated in Figure 1, the full pipeline is divided into three modules: (1) automatic sleep stage classification, (2) feature extraction, and (3) NT1 classification. The automatic sleep stage classifiers (Usleep$_{EEG}$ and Usleep$_{EMG}$) are used to obtain annotated labels from unlabeled EEG and EMG signals. Utilizing both models yields two sets of annotations for each mouse (y_eeg and y_emg). Sixteen features per sleep stage model are computed resulting in 16 × 3 × 2 features per mouse. In the final part, an L1-penalized logistic regression (NT1 classifier) is used to obtain the probability of having NT1 based on features from 11 WT mice and 19 NT1 mice. Each part is explained in detail below. Moving forward, the term pipeline refers to the full pipeline of all three modules, while Usleep$_{EEG}$ and Usleep$_{EMG}$ refer to the automatic sleep stage classifiers and the NT1 classifier refers to the L1-penalized logistic regression classification model for NT1 prediction.

Automatic sleep stage classification. For automatic sleep stage classification, we fine-tuned three different versions of U-Sleep: Usleep$_{EEG}$ (model trained on a single EEG channel), Usleep$_{EMG}$ (model trained on a single EMG channel), and Usleep$_{EEG,EMG}$ (model trained on both channels, used for comparison). Cohorts A–E are used for training the three models (see Table 1). We used Python version 3.7.16 for all data preprocessing and training of Usleep$_{EEG}$, Usleep$_{EEG}$ and Usleep$_{EEG,\ EMG}$. Specific packages are detailed in Supplementary Data S1–S2.

U-Sleep

U-Sleep. U-Sleep is a state-of-the-art model developed for human sleep stage classification [25]. It was trained on 15 660 participants coming from 16 different clinical studies. The architecture of the model is based on a preceding U-Time model [26] and has a form inspired by U-Net [27] which originally was developed for image segmentation. U-Sleep is a fully convolutional neural network and consists of an encoder block, decoder block, and a segment classifier. The encoder is mapping the input signal into a feature representation, the decoder projects it back to the input space and the segment classifier is making sleep stage predictions for a chosen time resolution. The original prediction resolution is in 30-second windows, matching the human sleep annotation resolution. However, the segment classifier is capable of predictions at any frequencies at test time, making U-sleep very flexible.

Data preprocessing. The data were preprocessed according to the pipeline for human PSG data described in Perslev et al. [25]. First, all EEG and EMG signals were resampled to 128 Hz using polyphase filtering. Next, each signal was individually scaled to have a median of 0 and an interquartile range (IQR) of one. Noisy bouts were clipped if they had an absolute deviation of more than 20 times the IQR of that specific channel from the median. The signals were further bandpass filtered with the cut-off frequencies of [0.3 Hz–35 Hz]. As each laboratory scores the data differently (e.g. cohort A scores intermediate states and artifact-specific stages, while others only score wakefulness, NREM sleep, and REM sleep), we used the standard classes wakefulness, NREM sleep, and REM sleep, and relabeled everything else as wakefulness unless the next stage was not wakefulness, in which case they were replaced with the previous sleep stage. To ensure consistency in the resolution of sleep scorings across different labs, all data was converted to 4-second epochs. The 1-second sleep stage data was reshaped into 4-second epochs, where each epoch contained four consecutive 1-second sleep stage labels. For each 4-second epoch, the most frequent sleep stage was selected. In cases where two sleep stages were equally represented, we computed the probability distribution of each stage across all 1-second epochs and randomly assigned a stage based on these probabilities. For datasets with 5-second epochs, we first upsampled the labels to a 1-second resolution before downsampling to 4-second epochs, using the same method as described above.

Transfer learning. The U-Sleep models pre-trained for human sleep staging were fine-tuned to fit mouse sleep data. There are two different versions of the original U-Sleep model used for this paper. The first one is trained to use any EEG channel which is utilized for fine-tuning Usleep$_{EEG}$ and Usleep$_{EMG}$. The second one is originally trained on two signals (an EEG and an EOG channel) and used for fine-tuning Usleep$_{EEG,EMG}$. While Usleep$_{EEG}$ and Usleep$_{EMG}$ are a part of the pipeline, Usleep$_{EEG,EMG}$ are trained for comparison purposes that are presented in the Methods section "Test of Alternative Pipelines."

We replaced the last layer such that the model learns classification for three stages (wakefulness, NREM sleep, REM sleep) instead of the five sleep stages in human sleep. To address the fact that one of our models Usleep$_{EMG}$ only relies on the EMG signal (and that the pre-trained model was trained on a single EEG channel), we trained the convolutional layer of the first encoder block from scratch along with the new classification head while freezing the rest of the parameters. For the fine-tuning phase, we unfreeze all layers and trained the model again with a lower learning rate. For this application, we have set the prediction window to four seconds, resembling the length of the manual sleep stage resolution.

Model training. Each batch consisted of 128 sequences of 44 seconds (11 × 4-second epochs), and batch elements were sampled as proposed by Perslev et al. [25]. The learning rate was set in the first part of the training, and in the fine-tuning phase. We used an unweighted cross-entropy cost function and an Adam optimizer. The model was trained for 1200 epochs, unless 200 consecutive epochs showed no improvement in the validation loss. The learning curves as well as the model choice are illustrated in Figure S3-S5 respectively for Usleep$_{EEG,}$ Usleep$_{EMG}$ and Usleep$_{EEG,EMG}$.

Feature extraction. As illustrated in Figures 1, 16 features were extracted from each sleep stage. Since there were three sleep stages and two sets of hypnograms (i.e. one from Usleep$_{EEG}$ and one from Usleep$_{EMG}$) 16 × 3 × 2 features were extracted in total per mouse. The EEG band power frequency features were obtained by computing the power spectral density using Welch's method (window size of four seconds) from the SciPy 1.7.3 package in Python 3.7.16 [28]. For each frequency band of interest, the integral of that area was calculated using the composite Simpson's rule. To account for inter-subject variability, the relative power was computed by normalizing with the integral of the entire power spectrum. The following frequency bands were calculated: slow-oscillations [0.5-1.5 Hz], slow-delta [1-2.25 Hz], fast-delta [2.5-4 Hz], slow-theta [5-8 Hz], fast-theta [8-10 Hz], alpha [9-14 Hz], and beta [14-30 Hz]. The root mean square (RMS) was used as a representation of the EMG signal's amplitude. For each sleep stage, we calculated the average bout length in seconds, the total time spent in that stage as a fraction of the total time spent in all stages, and the count of bouts in the specific stage per hour. Additionally, for each stage, we estimated the time spent in bouts of 4 seconds, 4–32 seconds, 32–60 seconds, 60–300 seconds, and greater than 300 seconds as a fraction of time spent in all bouts in that stage. All features were standardized prior to model training in each fold.

Model training. A total of 11 WT mice and 19 DTA mice (Cohort F) were used to train the NT1 classifier. The effect of the weighted L1 term is to reduce the number of variables by setting irrelevant parameters to zero. The larger the value of λ, the greater the emphasis on the regularization term [29]. The model is optimized using a Coordinate Descent algorithm with the "liblinear" setting in SciPy. We trained the classification model with an outer leave-one-subject-out (LOSO) approach and an inner 5-fold cross-validation approach allowing us to find the optimal value for λ while considering different splits. In each fold, we tested 10 equidistant values of λ on a log scale between 10^{-4} and 10^{4}, and determined the best λ. Finally, we trained a new NT1 classification model on the entire dataset using the optimal value of λ. The NT1 classification pipeline was implemented using the LogisticRegressionCV function from Scikit-learn version 1.0.2 [30].

Test of Alternative Pipelines

As a final step, we explored alternative pipelines for potential improvement of our method. Our original pipeline (i.e. Ensemble pipeline; Figure 1) was based on using both the Usleep$_{EEG}$ and Usleep$_{EMG}$ models to obtain two sets of predictions per subject, resulting in 16 × 3 × 2 features for training the NT1 classifier in an ensemble manner. The global pipeline was tested to determine if an NT1 classifier could capture differences between the two genotypes regardless of the sleep stages. To achieve this, we computed seven power band features from the EEG signal and the RMS from the EMG signal, averaging them across all sleep stages for each mouse. We used these features to train an NT1 classifier in a LOSO manner as explained in the section "Model Classification." We then moved on to examine the effect of using single modality classifiers from either the EEG (single-EEG pipeline) or the EMG channel (single-EMG pipeline). We used the sleep stage predictions from Usleep$_{EEG}$ to compute 16 features from each sleep stage, which then were used for training an NT1 classifier. We repeated this process with the respective output predictions from the Usleep$_{EMG}$ model. The multi-channel pipeline utilized the Usleep$_{EEG,EMG}$ classifier to generate predictions. Features computed from the single annotation set were subsequently used to train an NT1 classifier.

Statistics

Hypnograms obtained from the automatic sleep stage classifiers were evaluated against manual labels in confusion matrices. Two confusion matrices were obtained for each model as the models were tested in a cohort of NT1 mice and WT mice. Differences in the performance of the models in NT1 mice and WT mice were computed by taking the log odds ratio (LOR) between the two confusion matrices. The LOR were iteratively applied across all fields of the matrices resulting in a likelihood for a certain (mis)classification that is more likely to happen for an NT1 mouse.

$$LOR = \ln\left(\frac{x_1 \cdot x_2}{x_3 \cdot x_2}\right)$$

where x_1, x_2, x_3 and x_4 represents elements in the contingency table (see Supplementary Data S3 for details).

The confidence intervals (CI) were further computed as

$$\mathrm{CI_{up}} = \exp\left(\ln\left(\frac{x_1 \cdot x_4}{x_3 \cdot x_2}\right) + 1.96 \cdot \sqrt{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}}\right)$$

$$\mathrm{CI_{in}} = \exp\left(\ln\left(\frac{x_1 \cdot x_4}{x_3 \cdot x_2}\right) - 1.96 \cdot \sqrt{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}}\right)$$

The computation of the ci enabled the evaluation of the statistical significance of the LOR; if the ci does not include zero, the LOR is considered statistically significant [31]. The ci have not been corrected for multiple comparisons.

To evaluate the performance of all classifiers (method section "Test of Alternative Pipelines"), we tested all five pipelines against each other using a McNemar test. We tested the hypothesis that classifiers have different proportions of errors on the test set against the null hypothesis that they have similar proportions of errors. The McNemar test is suitable for evaluating differences in classification performance when the data are paired, such as predictions on the same test set. We assumed a binomial distribution and conducted the test using the statsmodels Python package.

Results

Models trained on one modality can be used to predict wakefulness and sleep in WT and NT1 mice

To evaluate the performance of the automatic sleep stage classifiers, Usleep$_{\mathrm{EEG}}$ and Usleep$_{\mathrm{EMG}}$ were tested in cohort F, a held-out dataset of 11 WT mice and 19 NT1 mice (see Table 1). Our results showed that Usleep$_{\mathrm{EMG}}$ was good at separating wakefulness from sleep; however, it did not perform well in differentiating between sleep stages (Figure 2A). Of the true REM sleep episodes, 59% were classified as NREM sleep, and 6.7% of the true NREM sleep episodes were predicted as REM sleep. Usleep$_{\mathrm{EMG}}$ correctly identified 96.3% of the wakefulness epochs and 89.2% of the NREM sleep epochs. The precision for Usleep$_{\mathrm{EMG}}$ in wakefulness and NREM sleep was 96.7% and 85.4%, (Supplementary Figure S1C) indicating that among all the epochs that were classified as wakefulness, 96.7% was correctly classified as wakefulness and out of all NREM sleep predictions 85.4% were correctly classified as NREM sleep. While the performance for both wakefulness and NREM sleep was good, Usleep$_{\mathrm{EMG}}$ performed less well for detecting REM sleep. Only 35.3% of the true REM sleep epochs were correctly identified (Figure 2A), and only 45.5% of the epochs classified as REM sleep were manually labeled as REM sleep (Supplementary Figure S1C). This indicated that Usleep$_{\mathrm{EMG}}$ was conservative in predicting REM sleep, and when it predicted REM sleep, it was less precise.

Usleep$_{\mathrm{EEG}}$ achieved a high recall for both wakefulness (94.7%), NREM sleep (96.1%), and to a lesser extent REM sleep (77.8%; Figure 2C). The same pattern was seen for the precision (Supplementary Figure S1A). Similar to Usleep$_{\mathrm{EMG}}$, REM sleep was the most difficult sleep stage to predict: 10.3% of the REM sleep was misclassified as wakefulness and 11.9% of the REM sleep was misclassified as NREM sleep. However, as opposed to Usleep$_{\mathrm{EMG}}$, the precision (Supplementary Figure S1A) for REM sleep was 80.7%, indicating that most of the classified REM sleep was also manually labeled as REM sleep.

When using the Usleep$_{\mathrm{EMG}}$ model, NT1 mice were significantly more likely to have an increased number of wakefulness-NREM sleep misclassifications (manual *versus* predicted; 0.8 LOR; Supplementary Table S1 for statistics) and wakefulness-REM sleep misclassifications (1.4 LOR; Supplementary Table S1). A decreased EMG amplitude for these wakefulness-NREM sleep and wakefulness-REM sleep epochs was found across genotypes (expressed as RMS, Figure 2E), explaining why the models predict these as NREM sleep or REM sleep rather than wakefulness. Wakefulness-NREM sleep and wakefulness-REM sleep misclassifications were more likely to occur in NT1 mice (compared predictions by Usleep$_{\mathrm{EEG}}$ and manual labels; LOR of 0.8 and 1.3, respectively; see Figure 2D; Supplementary Table S2 for statistics). An increased delta power was found in wakefulness-NREM sleep epochs compared to correctly identified wakefulness (Figure 2F), as in contrast to little or no relative difference in theta power between wakefulness and wakefulness-REM sleep (Figure 2G). Our results further showed that misclassifications of NREM sleep-wakefulness, NREM sleep-REM sleep, and REM sleep-wakefulness were more likely to occur in NT1 mice (LOR of 0.6, 1.0, and 0.6, respectively; see Figure 2D; Supplementary Table S2 for statistics). A decreased delta power was observed for NREM sleep-wakefulness relative to NREM sleep (Figure 2F). Increased theta power in NREM sleep-REM sleep relative to NREM sleep and decreased theta power for REM sleep-wakefulness relative to REM sleep was found across genotypes (Figure 2G).

Collectively, misclassifications were observed in WT and NT1 mice with both models, Usleep$_{\mathrm{EEG}}$ and Usleep$_{\mathrm{EMG}}$. Although our data suggest that they occur for the same reason, the misclassifications are more frequently occurring in NT1 mice.

Misclassifications by Usleep$_{\mathrm{EEG}}$ and Usleep$_{\mathrm{EMG}}$ can be used to identify abnormalities in NT1 mice sleep-wake patterns

To examine when misclassifications occurred, we looked at sequences that covered some of the misclassification periods of an NT1 mouse. While some of the misclassifications occurred around transitions from wakefulness to sleep (Figure 3A), other misclassifications appeared around microevents such as microarousals (Figure 3B). Specifically, for this NT1 mouse, there were sequences where both models predicted sleep with high certainty while the expert labeled it as wakefulness (Figure 3C, Figure 3D). For both cases, the power spectrum showed either high delta or high theta power during the periods of high certainty for sleep, as determined by the hypnodensity from Usleep$_{\mathrm{EEG}}$. A low EMG amplitude was observed when Usleep$_{\mathrm{EMG}}$ predicted sleep (Figure 3C), whereas a high EMG amplitude was noted when Usleep$_{\mathrm{EMG}}$ indicated wakefulness (Figure 3D). Thus, such episodes could be short sleep or cataplexy episodes missed by the expert.

Usleep$_{\mathrm{EEG}}$ and Usleep$_{\mathrm{EMG}}$ were solely trained on WT mice restricting the models to only make wakefulness, NREM sleep and REM sleep predictions. Hence, when disease-specific behaviors

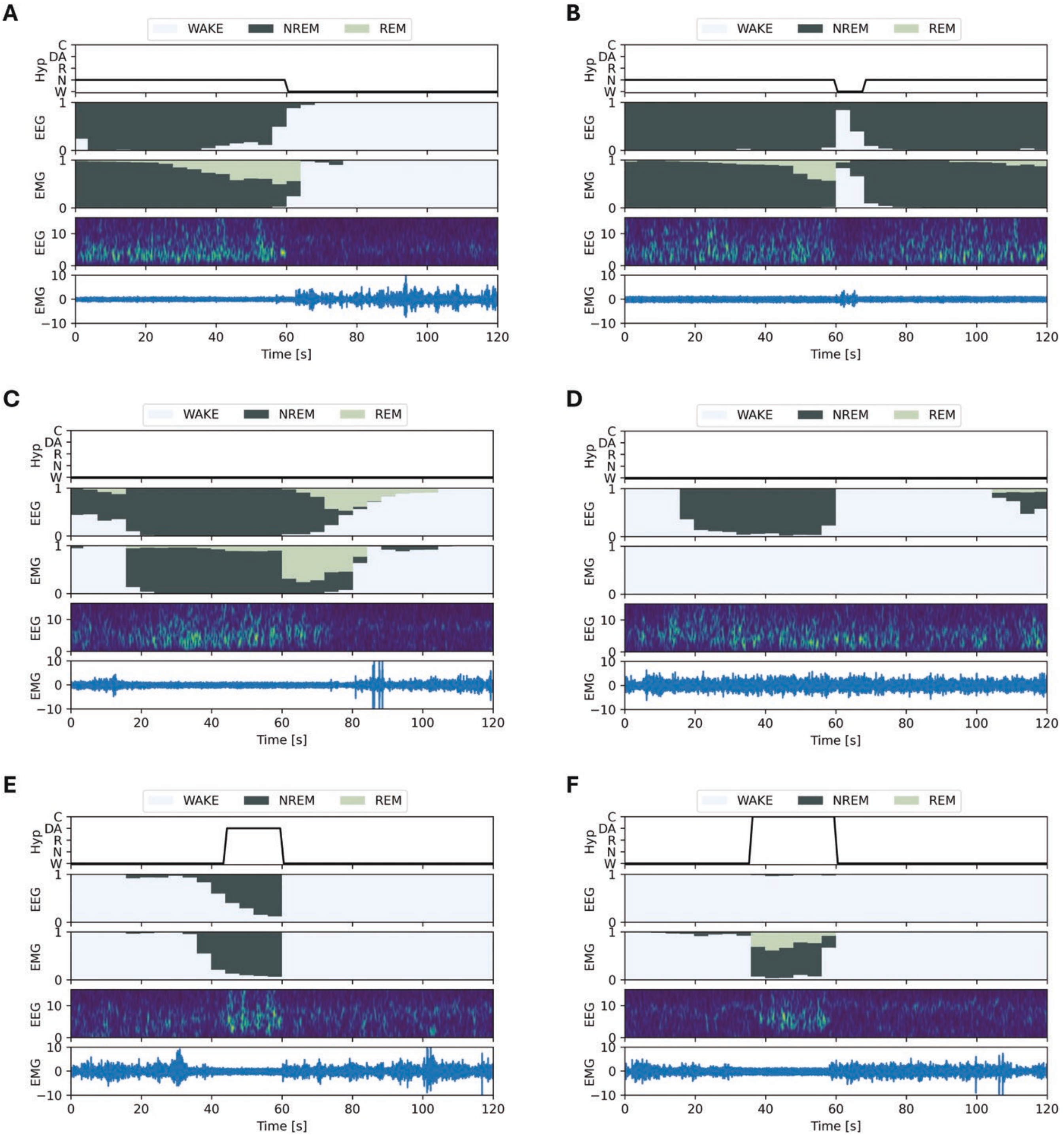


Figure 3. Misclassifications by UsleepEEG and UsleepEMG can be used to identify altered behavior in NT1 mice. Sequences of 120 seconds are plotted from one NT1 mouse. The first row shows hypnograms from manual scores (i.e. Hyp; W = Wakefulnss, N = NREM sleep, R = REM sleep, DA = Delta Attacks, C = Cataplexy). The second and third rows show hypnodensity from UsleepEEG and UsleepEMG. Below is a spectrogram of the EEG sequence where the frequency is on the y-axis, and yellow represents an increase in power. The last row represents the raw EMG signal. (A) Sequence covering a transition from NREM sleep to wakefulness. (B) Sequence covering a microarousal event. (C) Sequence with a wakefulness-NREM sleep and wakefulness—REM sleep misclassification (i.e., manual label indicates wakefulkness while the models indicate sleep). (D) Sequence of a wakefulness-NREM sleep misclassification. (E) Sequence covering delta attack. (F) Sequence covering a cataplexy attack.

such as cataplexy or delta attacks [19] appeared it would be classified into one of these three main stages. To investigate how the models would classify these periods, we visualized scored sequences of delta attack (Figure 3E) and cataplexy (Figure 3F). For delta attacks, both models predicted a high probability of NREM sleep, and the data exhibited the characteristics of NREM sleep (Figure 3E; high delta power in the EEG power spectrum and a low EMG amplitude). Although there was an increase in theta power during the period of cataplexy, Usleep$_{EEG}$ still labeled it as wakefulness. Cataplexy was detected as NREM sleep/REM sleep by Usleep$_{EMG,}$ because a clear loss in muscle tone was present during this period. Essentially the misclassifications reflect the altered behavioral phenotypes reported in NT1 mice [19].

Feature extraction from each sleep stage can be used to develop a model for probability estimation of mouse NT1

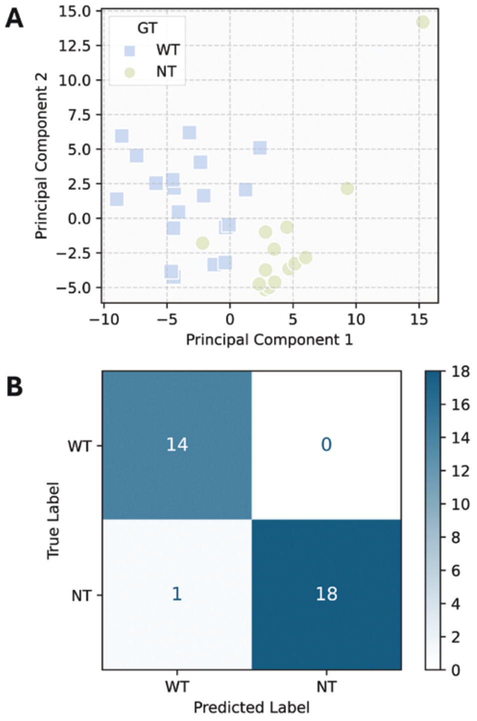
Following feature extraction (Figure 1), we decomposed these using principal component analysis (PCA) for visualization in two dimensions (Figure 4A). A clear separation of WT mice and NT1 mice was observed suggesting that the computed features could differentiate the two genotypes. The features were thus used to train the NT1 classifier using 5-fold cross-validation, which achieved an accuracy of 0.97 (Figure 4B). Some of the most important features as determined by the model coincide with well-known characteristics of the NT1 phenotype, such as the average bout length of wakefulness and NREM sleep while other features, such as increased slow-delta power in REM sleep, might be additional features to characterize the NT1 phenotype in mice (Figure 4C).

Performance of different pipelines for probability estimation of NT1 in mice.

Once we established that the features we selected worked well for probability estimation, we explored if the pipeline could be simplified. To do so, we first tested if the global signal across sleep stages was strong enough to distinguish WT from NT1 mice. Using the global signal as the main pipeline would make the automatic sleep staging step redundant and therefore greatly simplify the approach. Eight features were extracted and used in the global pipeline (Figure 5A; seven power band features from the EEG signal and the RMS from the EMG signal). Although the pipeline is much simpler, there was a clear drop in performance (0.79 accuracy; Figure 5B), when features were not extracted for each sleep stage. Second, we investigated the effect of only using the hypnogram from either the EEG (i.e. Single-EEG pipeline) or the EMG (i.e. Single-EMG pipeline; Figure 5A), as opposed to the Ensemble Pipeline that uses both individual models. We found that the EEG pipeline provided most of the performance gain and performed similarly to the Ensemble Pipeline (Figure 5B). Finally, we tested how a sleep stage model that relied on both EEG and EMG signals (i.e. a multi-channel pipeline) compared to the Ensemble approach, and found that it resulted in one additional incorrect prediction (Figure 5B). All the models were tested against each other with a McNemar test (Supplementary Table S3). Although none of the pipeline comparisons yielded significant differences, we chose to proceed with the ensemble approach that used the two separate models, Usleep$_{EEG}$ and Usleep$_{EMG}$, expecting this would make the pipeline more robust. In cases where one channel is noisy, the other can still be used to generate a useful hypnogram, effectively reducing the risk of poor performance due to single-channel noise. While the ensemble increases model complexity, this is a trade-off to gain robustness against channel-specific artifacts.

The automated pipeline successfully captures disease progression in a held-out dataset of 7/7 DTA mice

To validate our pipeline, we tested our final model in a held-out dataset of seven DTA mice (Cohort G; Table 1). As shown in Figure 6B, the pipeline correctly captured disease progression across weeks for all mice. To further examine robustness towards noise, we tested the pipeline in a group of WT and NT1 mice with and without artifacts (Figure 6C-D). Since our pipeline heavily relies on power spectrum features, the performance of the model was compromised when noise was present in both signals. The



C

Model	Feature	Sleep Stage	Importance	NT compared to WT
Usleep$_{EEG}$	Average bout length	NREM	2.38	Decreased
Usleep$_{EEG}$	4sec bouts	Wake	1.72	Decreased
Usleep$_{EEG}$	Slowdelta	REM	1.50	Decreased
Usleep$_{EEG}$	Average bout length	Wake	1.31	Decreased
Usleep$_{EMG}$	4sec bouts	Wake	1.19	Decreased
Usleep$_{EMG}$	>5min bouts	NREM	1.15	Increased
Usleep$_{EMG}$	Alpha	REM	1.11	Increased
Usleep$_{EMG}$	32sec-1min bouts	REM	0.92	Decreased
Usleep$_{EEG}$	1min-5min bouts	REM	0.75	Decreased
Usleep$_{EMG}$	1min-5min bouts	Wake	0.73	Increased
Usleep$_{EMG}$	>5min bouts	REM	0.72	Decreased
Usleep$_{EEG}$	Slowtheta	Wake	0.62	Increased
Usleep$_{EMG}$	>5min bouts	Wake	0.62	Decreased
Usleep$_{EEG}$	>5min bouts	Wake	0.61	Decreased
Usleep$_{EEG}$	1min-5min bouts	NREM	0.61	Decreased

Figure 4. Feature extraction from each sleep state can be used to develop a model for probability estimation of mouse NT1. (A) Scatter plot of principal component one and principal component two, WT mice and NT1 mice. (B) Confusion matrix showing the performance of the NT1 classifier trained to predict the probability of having NT1 (C) In table C, the most important features are listed in descending order.

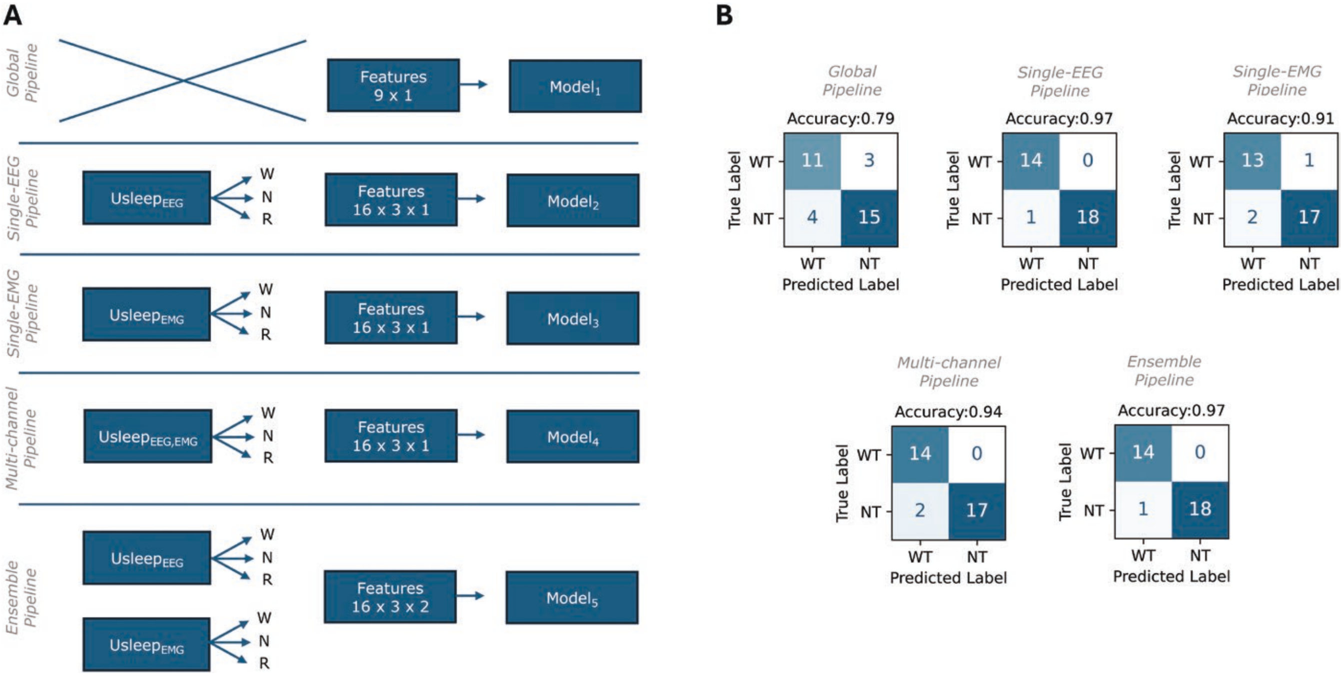


Figure 5. Pipeline performances for probability estimation of NT1 (A) Different pipelines tested for developing the best fully automated pipeline. The first approach tested was the Global Pipeline, where features of each mouse are averaged across all epochs regardless of the sleep stages. Only nine features are extracted: seven power bands from the EEG signal and the root mean square feature from the EMG signal. Next, we used single modality classifiers (Single-EEG Pipeline, Single-EMG Pipeline). Finally, we further tested an automatic sleep stage classifier that takes two modalities (multi-channel) as input instead of one. The Ensemble Pipeline resembles the pipeline represented in Figure 1. (B) Confusion matrices for each pipeline. Both the Single-EEG Pipeline and the Ensemble approach achieve a performance of 0.97, with only one mouse being misclassified.

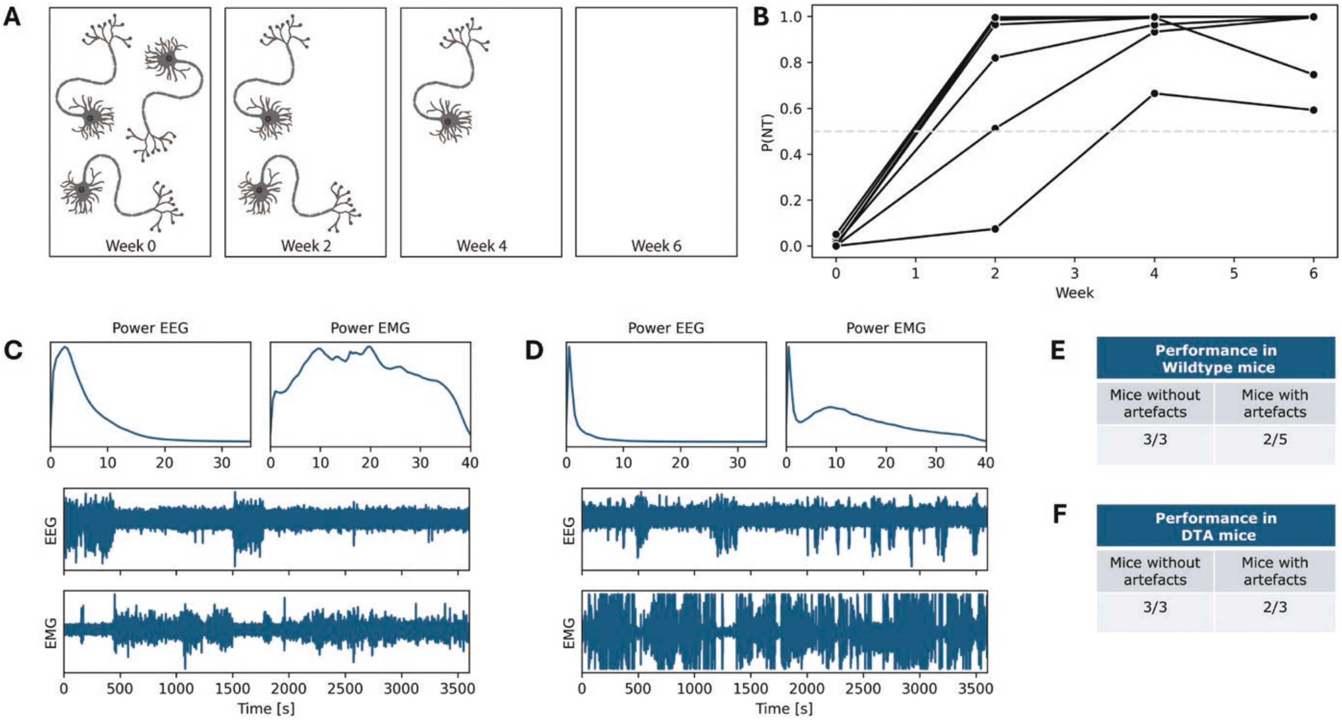


Figure 6. Automated pipeline captures disease progression in a held-out dataset of 7/7 DTA mice (A) In the DTA NT1 model, orexin neurons degenerate across time. In week 0 all orexin neurons exist similar to WT mice, across weeks 2–6 there is a progressive appearance of NT1 symptoms. (B) A line plot reflecting the probability of having NT across weeks for seven DTA mice (i.e., mouse model presented in A). (C), An example of an artifact-free power plot of the EEG and the EMG and the raw traces below. (D), signals with artifacts. (E-F) Performance of the pipeline when tested in WT and NT1 mice with and without artifacts.

pipeline accurately detected WT and NT1 mice when the data was artifact free but failed when recordings contained excessive noise (Figure 6E–F).

Discussion

In this study, we propose a fully automated pipeline for probability estimation of NT1 in mice based on unlabeled EEG and EMG recordings. The pipeline achieves an accuracy of 97% (Figure 4) 100% on hold-out cohort G (Figure 6), and 71.4% on cohort H (Figure 6). Subject-wise average on both these cohorts results in an accuracy of 80%, a sensitivity of 92%, and a specificity of 62%. While there is a performance decrease in the presence of noise (Figure 6), overall results indicate that the pipeline generalizes well to other laboratories.

Automatic Sleep Stage Classifiers

In our pipeline, Usleep$_{EMG}$ and Usleep$_{EEG}$ generate hypnograms for bout analysis, spectral features, and RMS, essential for characterizing phenotypic behavior. Thus, the whole pipeline relies on the accuracy of the automatic sleep stage classifiers. Our results show that REM sleep is the most difficult sleep stage to score (Figure 2), which is in line with current research [32–35]. One of the reasons for this could be that REM sleep is the minority class. From a modeling perspective, a model can achieve a high accuracy by favoring the majority class and neglecting the minority class. Hence, techniques such as adjusting the loss function, for instance by using weighted cross-entropy, or stratifying the dataset to ensure equal representation of each class, are often used to balance the data and help the model learn the minority class. In this study, we accommodated the imbalanced classes with uniform sampling of a class (i.e. all classes are sampled with equal probability) during training. However, our results still show a lower performance for REM sleep prediction for both models compared to NREM sleep and wakefulness (Figure 2A-D). Although additional methods could be explored to improve REM sleep prediction, numerous studies [32, 34] have previously investigated various techniques with limited gains, suggesting that this stage might inherently have more variability making it harder to learn. Comparing Usleep$_{EEG}$ to other state-of-the art models in the field [32, 35], Usleep$_{EEG}$ performs on a similar scale particularly for wakefulness and NREM sleep (Supplementary Table S4) confirming that our approach reaches state-of-the-art performance.

Many models in the field use post-hoc methods to correct “wrong” predictions in terms of sleep physiology, which inevitably improves accuracy during evaluation. As an example, SPINDLE [32] uses an HMM to suppress implausible transitions through a transition matrix. Although this improves model performance, it also constrains the model to only work for WT mice, as some theoretically impossible transitions, such as going from wakefulness directly to REM sleep, can be observed in NT1 mice [36]. For this reason, we did not make any post-hoc corrections, allowing for unrestricted sleep transitions to be modeled in NT1 mice, even if it meant sacrificing performance.

Misclassifications

For both Usleep$_{EMG}$ and Usleep$_{EEG,}$ we observed that some misclassifications were more likely to occur in NT1 mice. Compared to WT mice, NT1 mice exhibit more fragmented sleep/wake patterns [10]. Since there often is inconsistency in scoring sleep stage transitions both among experts and between experts and models, this might increase the number of misclassifications.

Particularly for Usleep$_{EMG,}$ there was an increased likelihood for wakefulness-NREM sleep and wakefulness-REM sleep misclassifications (Figure 2B), due to a decrease in EMG amplitude in some of the epochs labeled as wakefulness by manual scoring (Figure 2E). As the models only score wakefulness, NREM sleep, and REM sleep, all epochs with delta attacks and cataplexy annotations were left out of the performance evaluation. Such behaviors are therefore not contributing to the misclassifications unless they have not been labeled by the expert. Instead, the increased wakefulness-NREM sleep and wakefulness-REM sleep misclassifications with the Usleep$_{EMG}$ model could be due to inactive wakefulness and an increased number of microsleep episodes or they could be unscored delta attacks or cataplexy.

Similar to Usleep$_{EMG,}$ Usleep$_{EEG}$ also has an increased likelihood of wakefulness-NREM sleep and wakefulness-REM sleep misclassifications in NT1 mice suggesting alterations in the EEG during wakefulness (Figure 2D). One possible reason for the wakefulness-NREM sleep misclassification may be the presence of increased delta power during wakefulness (Figure 2F). These episodes could be unscored delta attacks or perhaps a sign of sleepiness. Increased delta power during wakefulness is indeed a commonly observed marker in sleep-deprived humans [37, 38], but is not as established in mice [39]. Vyazovskiy et al. [39] found an increase in delta (1.5–4Hz) and low theta (5-6.5Hz) power in wakefulness EEG during 6-h of sleep deprivation in rats. Thus, the delta-dominated wakefulness that is misclassified as NREM sleep might serve as a marker for sleepiness in NT1 mice. With Usleep$_{EEG}$ NREM sleep-wakefulness and REM sleep-wakefulness misclassifications are also seen. These indicate that NT1 mice exhibit sleep alterations with a drop in delta power in NREM sleep (Figure 2F) and theta power in REM sleep (Figure 2G), respectively. These alterations are also observed in a study by Christensen et al. [40] in humans and could potentially serve as a translational diagnostic marker.

The increased likelihood of NREM sleep—REM sleep misclassification in NT1 (Figure 2D) and these epochs showing characteristics of REM sleep (high theta peak; Figure 2G) suggest that some epochs scored as NREM sleep might indeed be REM sleep. This might in part be caused by the scoring criteria used by manual experts. Since direct transitions into REM sleep is basically absent in WT mice, early and short REM sleep episodes in NT1 [36] mice might be overlooked and scored as NREM sleep.

All these different misclassifications raise the question of whether it is the manual expert or the model that is correct. In this context, it should be considered that human experts often rely on video to score sleep and wake states, especially in mouse models of diseases. While the video can enable more precise scoring, it also contributes to greater variation between experts, as some rely more on video and others more on EEG/EMG signals. Further, both manual scoring and video analysis are highly repetitive tasks that lead to fatigue errors. Sleep/wake transitions (Figure 3A) and microevents (Figure 3B) are also a source of variation, as these can be difficult to label manually. One reason is that the transition does not always happen at exactly the same time according to the EEG power spectrum and the EMG signal (Figure 3A). Transitions can also show elements of two different stages in one epoch, thus forcing them into discrete wakefulness and sleep stage categories that cause inconsistencies.

The fact that Usleep$_{EMG}$ and Usleep$_{EEG}$ are trained on healthy mice means that altered behaviors such as delta attacks might end up classified as NREM sleep (Figure 3D). As a result, it may be difficult to compare delta and theta power in wakefulness, NREM sleep, and REM sleep with previous work because much of the altered behavior captured by manual experts may end up being classified in other stages. Naturally, the misclassifications will affect evaluation of sleep stages in terms of computing the average bout length, EMG amplitude, and spectral features. Thus, our model cannot be used for accurate quantification of features such as sleep stage bout length. However, the pipeline can still effectively distinguish between phenotypes which was our goal.

Feature extraction and classification model

Handcrafted feature extraction for NT1 detection has previously been proposed and used in humans [13]. However, to the best of our knowledge, this is the first fully automated tool for probability estimation in NT1 mice. Given the known signs of NT1 (i.e. altered EEG power spectra, altered REM sleep amplitude, and increased sleep/wake fragmentation [5]), we selected specific features that would capture these changes. We used spectral power features to capture altered EEG, altered EMG amplitude, and average bout length and counts of bouts, along with time spent in the different sleep stages, to measure sleep/wake fragmentation. We could have expanded our feature space by incorporating more features similar to Stephansen et al. [13], however, we chose to keep the feature space small due to the smaller sample size, and due to the scope of this study being to develop a simple pipeline for probability estimation of NT1 rather than identifying biomarkers.

Validation and Limitations of the Framework

We found that the pipeline does not perform well in the presence of noisy data (Figure 6C). Since movement artifacts appear as low-frequency noise in the signal, they directly mask the signal of interest and affect the computation of oscillations in the low and high delta range, which are variables the model relies on for the classification (Figure 4). This could suggest that the pipeline would benefit from a noise detection step prior to feature extraction, where methods such as Independent Component Analysis [41] could be used to remove artifacts. While we demonstrated that the pipeline is sensitive to noise, we have shown that it generalizes to multiple laboratories.

In this study, we focused on the DTA model, but a similar approach could be taken in future studies with other NT1 mouse models, such as the HCRT-KO and Ataxin-3 mouse models. Similarly, extending the work to the study of HcrtR1 and HcrtR2 knock-out mice could help better define various aspects of the phenotypes. Although HcrtR2 may play a greater role in the NT1 phenotype, both HCRT receptors are likely to be involved in the full NT1 phenotype [42–44]. Additionally, it would be interesting to explore whether a similar pipeline could be applied to mouse models in other domains, such as Alzheimer disease, depression, epilepsy, rapid eye movement sleep behavior disorder, or restless legs syndrome. This could potentially lead to faster and more unbiased tools for testing different treatment options across various mouse models.

Our pipeline focuses on sleep architecture specifically NREM and REM sleep rather than the direct detection of cataplexy. This aligns with real-world diagnostic practices, where abnormalities in sleep structure and REM sleep onset play a central role in supporting an NT1 diagnosis in the absence of observed cataplexy. While further work is needed to adapt the framework to human data, this study can serve as a foundation for developing similar automated diagnostic tools for clinical use. By providing a fully automated and scalable method for sleep stage classification and phenotypic profiling, our approach has the potential to assist in early or ambiguous NT1 diagnoses, particularly when cataplexy is not clearly present.

Supplementary material

Supplementary material is available at *SLEEP Advances* online.

Acknowledgments

Natural Language Process (NLP) models were used for this project to improve readability of the manuscript and correct grammar during the preparation of this work. The text has been reviewed after application of the model.

Author Contributions

Laura Rose (Conceptualization [Equal], Data curation [Lead], Formal analysis [Lead], Methodology [Lead], Validation [Lead], Visualization [Lead], Writing - original draft [Lead]), Alexander Zahid (Supervision [Equal], Writing - review & editing [Equal]), Louise Piilgaard (Data curation [Equal], Writing - original draft [Supporting], Writing - review & editing [Equal]), Christine Egebjerg (Data curation [Equal], Writing - review & editing [Supporting]), Frederikke Sørensen (Data curation [Equal], Writing - review & editing [Supporting]), Mie Andersen (Data curation [Equal], Writing - review & editing [Supporting]), Tessa Radovanovic (Data curation [Equal], Writing - review & editing [Supporting]), Anastasia Tsopanidou (Data curation [Equal], Writing - review & editing [Supporting]), Stefano Bastianini (Data curation [Equal], Writing - review & editing [Supporting]), Chiara Berteotti (Data curation [Supporting], Writing—review & editing [Supporting]), Viviana Lo Martire (Data curation [Supporting], Writing—review & editing [Supporting]), Micaela Borsa (Data curation [Equal], Writing—review & editing [Supporting]), Ryan Tisdale (Data curation [Equal], Writing—review & editing [Supporting]), Yu Sun (Data curation [Equal], Writing—review & editing [Supporting]), Maiken Nedergaard (Data curation [Equal], Writing—review & editing [Supporting]), Alessandro Silvani (Data curation [Equal], Writing—review & editing [Equal]), Giovanna Zoccoli (Data curation [Equal], Writing—review & editing [Supporting]), Antoine Adamantidis (Data curation [Equal], Writing—review & editing [Supporting]), Thomas Kilduff (Data curation [Equal], Writing—review & editing [Supporting]), Noriaki Sakai (Data curation [Equal], Writing— review & editing [Supporting]), Seiji Nishino (Data curation [Equal], Writing—review & editing [Supporting]), Sebastien Arthaud (Data curation [Equal], Writing—review & editing [Supporting]), Christelle Peyron (Data curation [Equal], Writing—review & editing [Supporting]), Patrice FORT (Data curation [Equal], Writing - review & editing [Supporting]), Morten Mørup (Supervision [Equal], Writing - review & editing [Equal]), Emmanuel Mignot (Conceptualization [Equal], Project administration [Equal], Supervision [Equal], Writing—review & editing [Equal]), and Birgitte Kornum (Conceptualization [Equal], Funding acquisition [Lead], Project administration [Equal], Resources [Lead], Supervision [Equal], Writing—original draft [Equal], Writing—review & editing [Lead])

Funding

This research has been supported by the Lundbeck Foundation (grant R344-2020-749).

Conflict of Interest

B.R.K. has consulted for UCB Pharma, H. Lundbeck A/S, Gubra, and Orexia Therapeutics. B.R.K. have submitted patent applications within the field of narcolepsy and is a founder of the University of Copenhagen spin-out company Ceremedy ApS.

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author. The code for our model is available at https://github.com/laulaurose/usleep-mouse.

References

1. Kornum BR, Knudsen S, Ollila HM, *et al.* Narcolepsy. *Nat Rev Dis Primers.* 2017;**3**:16100. doi:10.1038/nrdp.2016.100
2. Sorensen GL, Knudsen S, Jennum P. Sleep transitions in hypocretin-deficient narcolepsy. *Sleep.* 2013;**36**(8):1173–1177. doi:10.5665/sleep.2880
3. Peyron C, Faraco J, Rogers W, *et al.* A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat Med.* 2000;**6**(9):991–997. doi:10.1038/79690
4. Thannickal TC, Moore RY, Nienhuis R, *et al.* Reduced number of hypocretin neurons in human narcolepsy. *Neuron.* 2000;**27**(3):469–474. doi:10.1016/s0896-6273(00)00058-1
5. Diniz Behn CG, Klerman EB, Mochizuki T, Lin SC, Scammell TE. Abnormal sleep/wake dynamics in orexin knockout mice. *Sleep.* 2010;**33**(3):297–306. doi:10.1093/sleep/33.3.297
6. Silber MH, Krahn LE, Olson EJ, Pankratz VS. The epidemiology of narcolepsy in Olmsted County, Minnesota: A population-based study. *Sleep.* 2002;**25**(2):197–202. doi:10.1093/sleep/25.2.197
7. Ohayon MM, Priest RG, Zulley J, Smirne S, Paiva T. Prevalence of narcolepsy symptomatology and diagnosis in the European general population. *Neurology.* 2002;**58**(12):1826–1833. doi:10.1212/wnl.58.12.1826
8. Chemelli RM, Willie JT, Sinton CM, *et al.* Narcolepsy in orexin knockout mice: Molecular genetics of sleep regulation. *Cell.* 1999;**98**(4):437–451. doi:10.1016/s0092-8674(00)81973-x
9. Hara J, Beuckmann CT, Nambu T, *et al.* Genetic ablation of orexin neurons in mice results in narcolepsy, hypophagia, and obesity. *Neuron.* 2001;**30**(2):345–354. doi:10.1016/s0896-6273(01)00293-8
10. Tabuchi S, Tsunematsu T, Black SW, *et al.* Conditional ablation of orexin/hypocretin neurons: A new mouse model for the study of narcolepsy and orexin system function. *J Neurosci.* 2014;**34**(19):6495–6509. doi:10.1523/jneurosci.0073-14.2014
11. Dauvilliers Y, Montplaisir J, Molinari N, *et al.* Age at onset of narcolepsy in two large populations of patients in France and Quebec. *Neurology.* 2001;**57**(11):2029–2033. doi:10.1212/wnl.57.11.2029
12. Rayan A, Agarwal A, Samanta A, Severijnen E, van der Meij J, Genzel L. Sleep scoring in rodents: Criteria, automatic approaches and outstanding issues. *Eur J Neurosci.* 2024;**59**(4):526–553. doi:10.1111/ejn.15884
13. Stephansen JB, Olesen AN, Olsen M, *et al.* Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun.* 2018;**9**(1):5229. doi:10.1038/s41467-018-07229-3
14. Bastianini S, Silvani A, Berteotti C, *et al.* Sleep related changes in blood pressure in hypocretin-deficient narcoleptic mice. *Sleep.* 2011;**34**(2):213–218. doi:10.1093/sleep/34.2.213
15. Bastianini S, Berteotti C, Gabrielli A, *et al.* SCOPRISM: A new algorithm for automatic sleep scoring in mice. *J Neurosci Methods.* 2014;**235**:277–284. doi:10.1016/j.jneumeth.2014.07.018
16. Gent TC, Bandarabadi M, Herrera CG, Adamantidis AR. Thalamic dual control of sleep and wakefulness. *Nat Neurosci.* 2018;**21**(7):974–984. doi:10.1038/s41593-018-0164-7
17. Lie MEK, Falk-Petersen CB, Piilgaard L, Griem-Krey N, Wellendorph P, Kornum BR. GABAA receptor β1-subunit knock-out mice show increased delta power in NREM sleep and decreased theta power in REM sleep. *Eur J Neurosci.* 2021;**54**(2):4445–4455. doi:10.1111/ejn.15267
18. Poulie CBM, Chan CB, Parka A, *et al.* In Vitro and In Vivo Evaluation of Pellotine: A Hypnotic Lophophora Alkaloid. *ACS Pharmacol Transl Sci.* 2023;**6**(10):1492–1507. doi:10.1021/acsptsci.3c00142
19. Piilgaard L, Rose L, Gylling Hviid C, Kohlmeier KA, Kornum BR. Sex-related differences within sleep-wake dynamics, cataplexy, and EEG fast-delta power in a narcolepsy mouse model. *Sleep.* 2022;**45**(7). doi:10.1093/sleep/zsac058
20. Andersen M, Tsopanidou A, Radovanovic T, *et al.* Using Fiber Photometry in Mice to Estimate Fluorescent Biosensor Levels During Sleep. *Bio Protoc.* 2023;**13**(15):e4734. doi:10.21769/BioProtoc.4734
21. Kjaerby C, Andersen M, Hauglund N, *et al.* Memory-enhancing properties of sleep depend on the oscillatory amplitude of norepinephrine. *Nat Neurosci.* 2022;**25**(8):1059–1070. doi:10.1038/s41593-022-01102-9
22. Laguna A, Peñuelas N, Gonzalez-Sepulveda M, *et al.* Modelling human brain-wide pigmentation in rodents recapitulates age-related multisystem neurodegenerative deficits. *bioRxiv.* 2023;**15**:8819. doi:10.1101/2023.08.08.552400
23. Sun Y, Tisdale R, Park S, *et al.* The development of sleep/wake disruption and cataplexy as hypocretin/orexin neurons degenerate in male vs. female Orexin/tTA; TetO-DTA Mice. *Sleep.* 2022;**45**(12). doi:10.1093/sleep/zsac039
24. Sakai N, Nishino S. Comparison of solriamfetol and modafinil on arousal and anxiety-related behaviors in narcoleptic mice. *Neurotherapeutics.* 2023;**20**(2):546–563. doi:10.1007/s13311-022-01328-2
25. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel CU. Resilient high-frequency sleep staging. *NPJ Digit Med.* 2021;**4**(1):72. doi:10.1038/s41746-021-00440-5
26. Perslev M, Jensen MH, Darkner S, Jennum PJ, Igel CU. A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems.* 2019;**32**:4415.
27. Ronneberger O, Fischer P, Brox TU-net. Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 2015;**9351**:234. doi:10.1007/978-3-319-24574-4_28
28. Virtanen P, Gommers R, Oliphant TE, *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;**17**(3):352. doi:10.1038/s41592-020-0772-5
29. Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. *J R Stat Soc Series B Stat Methodol.* 2011;**73**(3):273–282. doi:10.1111/j.1467-9868.2011.00771.x
30. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;**12**:2825.
31. Norton EC, Dowd BE, Maciejewski ML. Odds ratios-current best practice and use. *JAMA - Journal of the American Medical Association.* 2018;**320**(1):84. doi:10.1001/jama.2018.6971

32. Miladinović D, Muheim C, Bauer S, *et al*. SPINDLE: End-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species. *PLoS Comput Biol.* 2019;**15**(4):e1006968. doi:10.1371/journal.pcbi.1006968
33. Kam K, Rapoport DM, Parekh A, Ayappa I, Varga AW. WaveSleepNet: An interpretable deep convolutional neural network for the continuous classification of mouse sleep and wake. *J Neurosci Methods.* 2021;**360**:109224. doi:10.1016/j.jneumeth.2021.109224
34. Yamabe M, Horie K, Shiokawa H, Funato H, Yanagisawa M, Kitagawa H. MC-SleepNet. Large-scale sleep stage scoring in Mice by deep neural networks. *Sci Rep.* 2019;**9**(1):1. doi:10.1038/s41598-019-51269-8
35. Grieger N, Schwabedal JTC, Wendel S, Ritze Y, Bialonski S. Automated scoring of pre-REM sleep in mice with deep learning. *Sci Rep.* 2021;**11**(1):12245. doi:10.1038/s41598-021-91286-0
36. Fujiki N, Cheng T, Yoshino F, Nishino S. Specificity of direct transition from wake to REM sleep in orexin/ataxin-3 transgenic narcoleptic mice. *Exp Neurol.* 2009;**217**(1):46–54. doi:10.1016/j.expneurol.2009.01.015
37. Aeschbach D, Matthews JR, Postolache TT, Jackson MA, Giesen HA, Wehr TA. Two circadian rhythms in the human electroencephalogram during wakefulness. *Am J Physiol Regul Integr Comp Physiol.* 1999;**277**(6 46-6):R1771–R1779. doi:10.1152/ajpregu.1999.277.6.R1771
38. Aeschbach D, Matthews JR, Postolache TT, Jackson MA, Giesen HA, Wehr TA. Dynamics of the human EEG during prolonged wakefulness: Evidence for frequency-specific circadian and homeostatic influences. *Neurosci Lett.* 1997;**239**(2-3):121–124. doi:10.1016/s0304-3940(97)00904-x
39. Vyazovskiy VV, Tobler I. Theta activity in the waking EEG is a marker of sleep propensity in the rat. *Brain Res.* 2005;**1050**(1-2):64–71. doi:10.1016/j.brainres.2005.05.022
40. Christensen JAE, Munk EGS, Peppard PE, *et al*. The diagnostic value of power spectra analysis of the sleep electroencephalography in narcoleptic patients. *Sleep Med.* 2015;**16**(12):1516–1527. doi:10.1016/j.sleep.2015.09.005
41. Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Netw.* 2000;**13**(4-5):411–430. doi:10.1016/s0893-6080(00)00026-5
42. Willie JT, Chemelli RM, Sinton CM, *et al*. Distinct narcolepsy syndromes in orexin receptor-2 and orexin null mice: Molecular genetic dissection of non-REM and REM sleep regulatory processes. *Neuron.* 2003;**38**(5):715–730. doi:10.1016/s0896-6273(03)00330-1
43. Sakurai T. The neural circuit of orexin (hypocretin): Maintaining sleep and wakefulness. *Nat Rev Neurosci.* 2007;**8**(3):171–181. doi:10.1038/nrn2092
44. Dashti HS, Daghlas I, Lane JM, *et al*.; 23andMe Research Team. Genetic determinants of daytime napping and effects on cardiometabolic health. *Nat Commun.* 2021;**12**(1):900. doi:10.1038/s41467-020-20585-3