Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When accurate sleep scoring misses brief awakenings: event-level consequences of temporal decoding under experimental perturbation

Amine BOUGARNE

amine.bougarne@um5r.ac.ma

Mohammed V University

Article

Keywords: automated sleep scoring, mouse polysomnography, hidden Markov model, temporal decoding, brief awakenings, sleep fragmentation, event-based evaluation, optogenetics

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11486001/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When accurate sleep scoring misses brief awakenings: event-level consequences of temporal decoding under experimental perturbation

Amine Bougarne[1*]

[1]*Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco.*

*Corresponding author. E-mail: amine.bougarne@um5r.ac.ma

ORCID: 0009-0000-6896-6926 (A. Bougarne)

Abstract

Automated sleep scoring in rodents is usually validated by epoch-wise agreement with a human expert. Brief awakenings occupy few epochs and barely move that figure, yet their number is a standard read-out of sleep fragmentation. I re-analysed the Oxford Mouse Polysomnography Benchmark and compared 13 scoring pipelines sharing the same spectral features but differing in classifier and temporal decoder. Models were developed on seven multi-scorer recordings, frozen, and applied once to 11 mice recorded at baseline and under optogenetic stimulation and to 6 mice before and after sleep deprivation. On the optogenetic recordings, epoch macro-F1 spanned 0.87 to 0.95, whereas brief-wake event F1 spanned 0.12 to 0.68 and the estimated brief-awakening rate 2.5 to 45.6 per hour of sleep (reference 36.1). Adding a standard hidden Markov decoder to a convolutional network changed macro-F1 by less than one point but erased most single-epoch awakenings (event F1 0.67 to 0.20). A decoder whose transition penalty is relaxed by a learned change score kept the epoch-level benefit and recovered these events. No pipeline reproduced the per-animal change between conditions, which the reference itself did not resolve. These exploratory results argue for event-level, duration-stratified validation when automated scoring is used to measure fragmentation.

Keywords: automated sleep scoring · mouse polysomnography · hidden Markov model · temporal decoding · brief awakenings · sleep fragmentation · event-based evaluation · optogenetics

1 Introduction

Sleep and wakefulness in mammals are organised by mutually inhibitory networks in the brainstem and hypothalamus that favour rapid, all-or-none switching between states.[1–3] Experimental sleep research reads the output of these circuits from the electroencephalogram (EEG) and electromyogram (EMG), classified epoch by epoch into wakefulness, non-rapid-eye-movement (NREM) sleep and REM sleep.[4] In mice the epoch is conventionally 4 s long.[5,6] A single 4-s wake epoch inside a long NREM bout is, in this vocabulary, the smallest observable awakening, and counts of such events are among the outcomes that circuit manipulations are designed to change.

Manual scoring of multi-day recordings is slow, and automated alternatives have multiplied. Unsupervised and feature-based approaches[7–9] have been joined by convolutional and recurrent networks trained end to end[10–16] and by probabilistic classifiers that model state transitions explicitly.[17] Human sleep staging went through the same transition, from single-channel networks to sequence models and large multi-cohort systems.[18–22] A recent review weighs the advantages and limitations of automated scoring in humans and rodents side by side.[23]

Most of these tools are judged by how often their epoch labels agree with an expert, reported as accuracy, Cohen's κ or macro-averaged F1. In mice, several methods now report agreement close to that between experts.[9,10,17] Human scorers disagree too, more on some states and transitions than on others,

and this caps what any comparison with a single reference can show.[24–27] A recent multi-laboratory study in mice found that expert agreement was lowest for REM sleep and that networks trained in one laboratory transferred poorly to others.[28]

Epoch agreement weights every 4-s epoch equally. Missing a one-epoch awakening costs almost nothing in macro-F1, yet the number of such awakenings is what an experimenter reports when the question is fragmentation. In human polysomnography, arousals as short as 3 s are scored as distinct events.[29] In mice, brief awakenings of up to 16 s have long served as an index of sleep continuity.[5] Wake bouts follow a heavy-tailed duration distribution across mammals,[30] hypnogram summary statistics depend on how bouts and transitions are segmented,[31] and transition dynamics change with disease.[32] Brief arousals during NREM sleep are also biologically structured: they follow infraslow noradrenergic fluctuations, relate to memory consolidation and, in mice, often last only 2–4 s.[33–36] Other fields learned long ago that frame-wise and event-wise scores can disagree,[37,38] and sleep-spindle detection is routinely assessed at the level of events.[39] Rodent sleep-stage classification has not generally adopted this practice.

Temporal context is where the two views diverge most. Many scorers pass their frame-wise outputs through a hidden Markov model (HMM) or an equivalent constraint so that the hypnogram respects plausible transition statistics.[10,17,40] Such constraints remove isolated errors. By construction, they also penalise short excursions into another state, and the penalty is learned from training data that need not share the transition statistics of an experimental condition. Brodersen and colleagues assembled the benchmark used here partly with this concern in mind: it contains recordings in which optogenetic stimulation increased the number of brief awakenings and the probability of NREM-to-wake and REM-to-wake transitions.[17,41,42] Their LDA–HMM classifier, Somnotate, was already evaluated on intermediate states, transition dynamics and robustness to both manipulations,[17] and I do not revisit those questions.

The question here is narrower. When scorers reach similar epoch-level performance, how far do their temporal constraints change the brief awakenings they detect, and does the difference propagate to the estimated effect of an experimental manipulation? I compared otherwise identical scorers with and without temporal decoding, evaluated brief wake episodes by one-to-one event matching, estimated the baseline-to-stimulation change per animal, and tested a decoder whose transition penalty follows a learned change score, in the tradition of input-dependent HMMs[43] (Fig. 1). The analysis protocol, including a primary end point on the preservation of the experimental contrast, was fixed before any test-set result was computed. That end point did not separate the models. The event-level findings that give this paper its title came from secondary analyses, and I present them as exploratory.

2 Results

2.1 Data audit and quality of the reference

All 41 recordings listed in the dataset documentation were found, and every signal matched its annotation to the second (Table 1). The redistributed copy names the six multi-scorer test files SampleFile_A to SampleFile_F and spells "artefact" where the documentation writes Sample_A to Sample_F and "artifact"; the pipeline resolved these 24 discrepancies explicitly and logged each one. Signals were stored in μV. No recording met the pre-declared exclusion rule, and the largest fraction of flat epochs in any recording was 4% (sleep-deprivation baseline of mouse To).

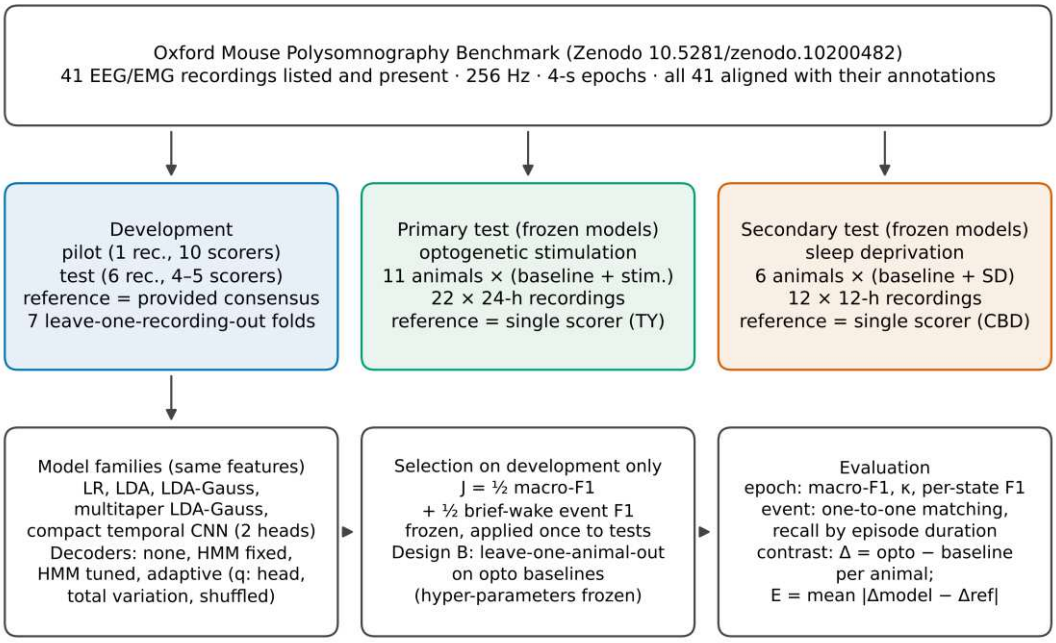


Fig. 1 Study design and data flow. Recordings from the Oxford Mouse Polysomnography Benchmark were split into a development set (pilot and multi-scorer test recordings) and two test sets used only after the models were frozen. All scorers share the same 76 spectral features and differ in classifier and temporal decoder. Hyper-parameters were selected on out-of-fold development predictions with a pre-declared objective. Brief wake episodes were defined as one to four consecutive wake epochs (4–16 s) flanked by sleep.

The audit found one duplicate. The pilot annotation of scorer CBD is byte-identical to the CBD annotation of the baseline sleep-deprivation recording of mouse RR1; both cover 12 h from ZT0. The two EDF files differ in channel montage, and I could not establish whether they come from the same session. Because the pilot recording belonged to the development set, the sleep-deprivation results for RR1 may not be fully independent of training.

Table 1 Data inventory after the audit.

Group	Condition	Recordings	Duration	Animals identified	Scorers per recording	Role
Pilot	reference	1	12 h (ZT0–12)	not documented	10	development
Test	reference	6	24 h	not documented	4–5	development
Optogenetic stimulation	baseline	11	24 h	11 (file prefix)	1 (TY)	primary test
Optogenetic stimulation	stimulation	11	24 h	11 (file prefix)	1 (TY)	primary test
Sleep deprivation	baseline	6	12 h (ZT0–12)	6 (file prefix)	1 (CBD)	secondary test
Sleep deprivation	deprivation	6	12 h (ZT0–12)	6 (file prefix)	1 (CBD)	secondary test

All 41 recordings were present, readable and aligned with their annotations (difference 0 s). No recording met the pre-declared exclusion rule. Animal identity in the test groups was taken from the shared file prefix together with the README statement that baselines come from the same animals. The pilot CBD annotation is byte-identical to the CBD annotation of the RR1 sleep-deprivation baseline (see Results).

Agreement on the multi-scorer recordings was high. Fleiss' κ ranged from 0.82 to 0.92, 73–91% of epochs were scored identically by every expert, and 0.1–3.5% were ties (Supplementary Table S1). The distributed consensus matched the recomputed majority on every non-tied epoch. Against the majority of the other experts, the scorer of the optogenetic recordings (TY) reached macro-F1 of 0.925–0.952 and brief-wake event F1 of 0.69–0.84 while annotating 15–62% more brief wake episodes per hour of sleep

(Table 2). The optogenetic reference therefore comes from an experienced scorer who marks brief awakenings more readily than the consensus used for training, a point that matters for every comparison below.

Table 2 Brief-wake annotation by the optogenetic reference scorer (TY) compared with the majority of the other experts.

Recording	Other scorers	Macro-F1	Cohen's κ	Brief-wake event F1	TY rate (/h sleep)	Majority rate (/h sleep)	Ratio
pilot	9	0.925	0.896	0.693	39.8	24.5	1.62
Sample B	3	0.952	0.945	0.769	29.7	25.8	1.15
Sample F	4	0.948	0.956	0.835	20.8	17.4	1.20

Majority recomputed from the other scorers' hypnograms (ties excluded). Brief wake = 4–16 s flanked by sleep.

2.2 Similar epoch-level performance, different brief awakenings

On the optogenetic recordings, all CNN variants reached macro-F1 between 0.930 and 0.947 and the LDA-based variants between 0.874 and 0.932 (Table 3, Fig. 2b). Brief-wake detection was far more variable. Event F1 ranged from 0.12 to 0.68, and the estimated rate of brief awakenings from 2.5 to 45.6 episodes per hour of sleep, against 36.1 in the reference (Fig. 2d). Frame-wise scorers overestimated the rate (LR 45.0, LDA 45.6) and HMM-decoded scorers underestimated it (LDA–HMM 12.4, LDA-Gauss–HMM 13.6, tuned CNN–HMM 13.3, fixed CNN–HMM 4.1). The same ordering held in development (Fig. 2a) and in the sleep-deprivation recordings (Fig. 2c), where the reference rate was 26.2 per hour of sleep and the undecoded CNN came closest (26.4).

Table 3 Epoch-level and brief-wake performance by data split.

Scorer	Dev macro-F1	Dev event F1	Dev rate	Opto macro-F1	Opto event F1	Opto rate	SD macro-F1	SD event F1	SD rate
LR	0.910	0.570	54.3	0.889	0.587	45.0	0.843	0.348	74.9
LDA	0.909	0.559	51.0	0.874	0.558	45.6	0.843	0.345	74.9
LDA+HMM	0.948	0.592	13.7	0.931	0.463	12.4	0.888	0.462	12.9
LDA+adaptive	0.943	0.677	21.7	0.925	0.576	20.0	0.880	0.479	23.4
LDA-Gauss+HMM*	0.946	0.611	14.4	0.932	0.500	13.6	0.879	0.466	13.4
Somnotate-like (multitaper)*	0.802	0.379	8.3	0.873	0.384	10.8	0.814	0.301	7.7
CNN	0.949	0.716	26.1	0.944	0.674	23.9	0.898	0.570	26.4
CNN+HMM fixed	0.935	0.360	6.4	0.936	0.201	4.1	0.884	0.332	5.8
CNN+HMM tuned	0.948	0.607	14.3	0.945	0.514	13.3	0.901	0.526	12.9
CNN+adaptive	0.951	0.724	23.2	0.947	0.677	21.8	0.904	0.602	21.1
CNN+adaptive (q TV)	0.949	0.719	23.3	0.945	0.670	21.9	0.899	0.581	22.2
CNN+adaptive (q shuffled)	0.930	0.232	4.1	0.930	0.123	2.5	0.896	0.250	4.6
Somnotate (provided)†	0.944	0.671	22.9	0.925	0.594	23.2	0.890	0.506	23.8
Human reference	–	–	27.6	–	–	36.1	–	–	26.2

Means over recordings. Development values are out-of-fold but come from non-nested selection and are optimistic. Rate = brief wake episodes (4–16 s) per hour of sleep of the source. * Re-implementation inspired by Somnotate, not the original software. † Hypnograms distributed with the dataset; training data undocumented, out-of-sample status not guaranteed.

The CNN makes the point most clearly. With and without a fixed HMM, the same network gave epoch macro-F1 of 0.936 and 0.944 on the optogenetic recordings, a difference below one percentage point. Its event F1 dropped from 0.674 to 0.201, and the estimated rate from 23.9 to 4.1 episodes per hour of sleep. Across the 13 scorers, the ranking by macro-F1 was a poor guide to the ranking by event F1 (Spearman $\rho = 0.36$; descriptive only, since the scorers are not independent).

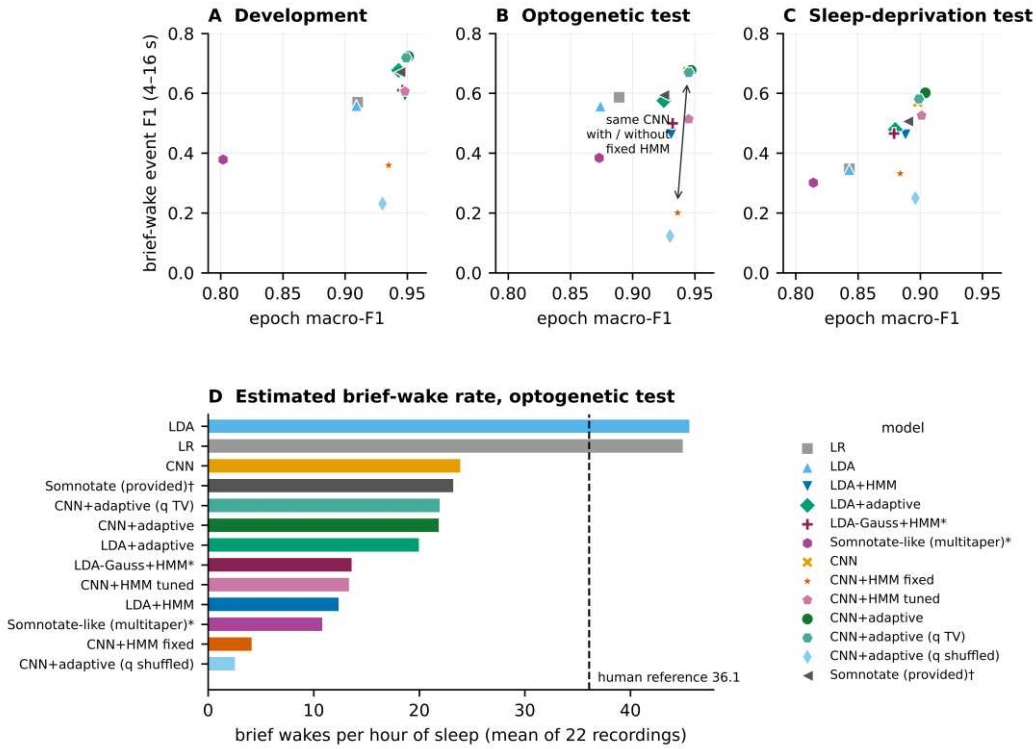


Fig. 2 Similar epoch-level performance, different brief awakenings. a–c, Epoch macro-F1 against brief-wake event F1 (4–16 s, one-to-one matching) for the 13 scorers, averaged over recordings, in development (7 recordings, out-of-fold), the optogenetic test (22 recordings, 11 mice) and the sleep-deprivation test (12 recordings, 6 mice). The arrow in **b** links the same CNN with and without a fixed HMM. **d**, Mean estimated brief-wake rate per hour of sleep on the optogenetic test; dashed line, human reference. *Re-implementation inspired by Somnotate. †Hypnograms distributed with the dataset; training data undocumented.

2.3 The loss depends on episode duration

Recall of reference wake episodes depended on their duration and, above all, on the decoder (Fig. 3). Single-epoch episodes (4 s) made up 5,297 of the 11,306 reference wake episodes flanked by sleep in the optogenetic recordings. For these, recall was roughly 0.55–0.58 for LR and LDA, 0.43–0.45 for the undecoded and adaptive CNNs, 0.35 for the distributed Somnotate hypnograms, 0.15–0.20 for the HMM-decoded LDA models and the tuned CNN–HMM, and close to zero for the CNN with a fixed HMM. At 16 s the fixed CNN–HMM still recovered only about 60% of episodes, while most other scorers exceeded 85%. Beyond 20 s the curves converged, and episodes longer than 64 s were recovered by every scorer. The hypnograms in Fig. 4, from the animal with the median primary error, show the same thing at a glance: the fixed HMM removes most of the single-epoch awakenings that the reference and the undecoded network share, and leaves long wake bouts untouched.

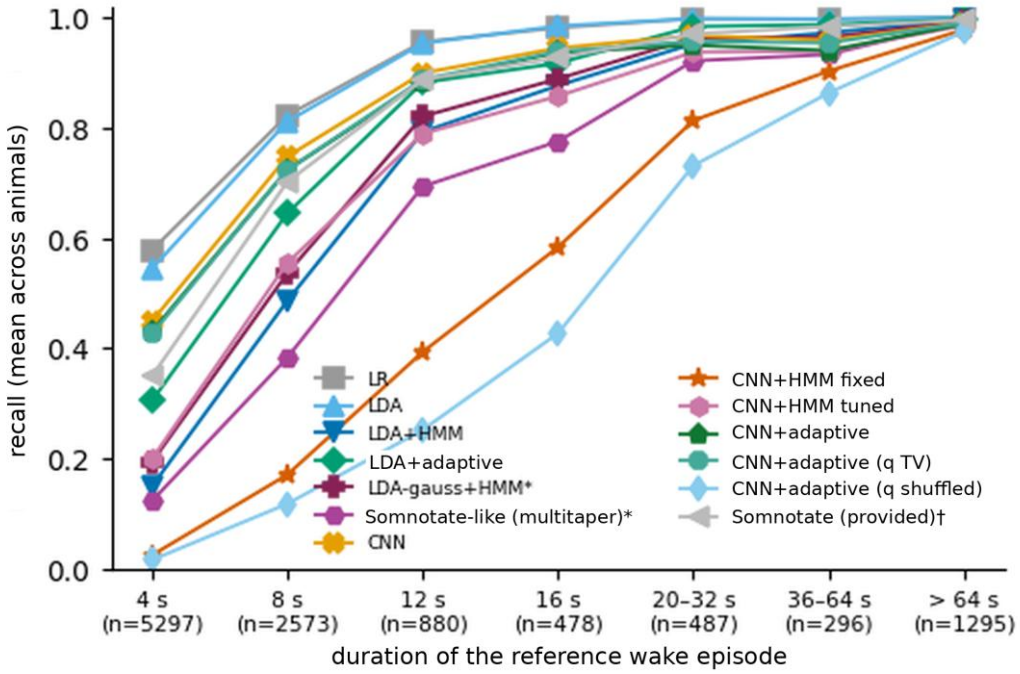


Fig. 3 Recall of reference wake episodes by duration. Optogenetic test, 22 recordings. Reference wake episodes flanked by sleep were binned by duration; *n* is the number of reference episodes per bin. Recall was computed per recording with one-to-one matching (tolerance one epoch) and averaged within, then across, the 11 mice. Rendered from the analysis notebook; axis labels translated, data unchanged.

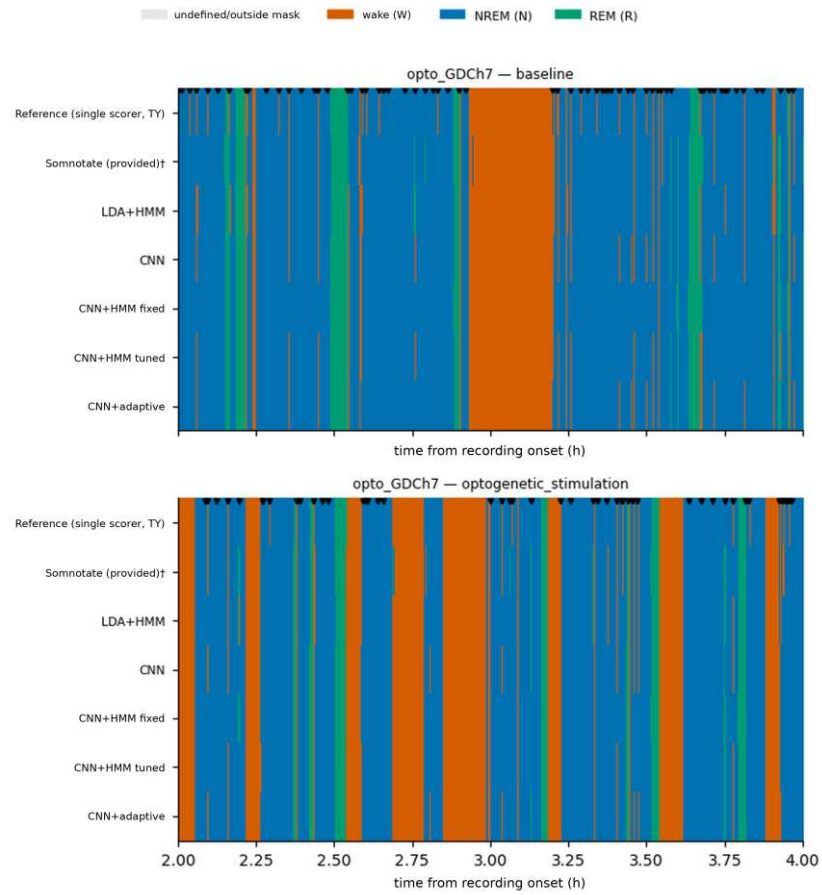


Fig. 4 Example hypnograms. Mouse GDCh7, the animal with the median primary error of the pre-declared main model, chosen by a rule fixed before inspection; hours 2–4 after recording onset, baseline (top) and stimulation (bottom). Black triangles mark reference brief wake episodes (4–16 s). The fixed HMM removes most single-epoch awakenings, whereas long wake bouts are preserved by all scorers. Rendered from the analysis notebook; labels translated.

2.4 What the adaptive decoder changes

The decoder selected on development data used the learned change score with $\lambda_{\max} = 2$, $\rho = 0$ and $\gamma = 4$ (Table 4). In words, it smooths twice as strongly as the standard HMM within a state and imposes almost no transition penalty where the change head predicts a boundary. On the optogenetic recordings it reached macro-F1 0.947 and event F1 0.677, essentially the values of the undecoded network (0.944 and 0.674). The tuned constant-weight decoder chose the weakest smoothing on offer ($\lambda_{\max} = 0.25$) and reached event F1 0.514.

Two controls show where the benefit comes from. Replacing the learned q_t by a within-recording permutation of the same values brought event F1 down to 0.123 and the estimated rate to 2.5 per hour of sleep, so the gain requires the change score to be aligned in time with the signal. Using the total variation between consecutive posteriors instead of the learned head gave almost the same result (event F1 0.670). For LDA, the adaptive decoder also outperformed the fixed HMM on events (0.576 versus 0.463). The adaptive decoder kept the cleaning effect of a strong temporal prior on epoch labels and removed most of its cost on brief awakenings. It did not, however, find more brief awakenings than the network without any decoder.

Table 4 Decoders selected on development data.

Scorer	Decoder	Selected parameters	Development objective J
CNN+HMM fixed	standard hybrid HMM	λmax = 1, ρ = 1, β = 1	0.647
CNN+HMM tuned	constant weight, tuned	λmax = 0.25, ρ = 1, β = 0	0.778
CNN+adaptive	learned change score	λmax = 2.0, ρ = 0, γ = 4, β = 0	0.837
CNN+adaptive (q TV)	total-variation change score	λmax = 0.25, ρ = 0, γ = 4, β = 0	0.834
CNN+adaptive (q shuffled)	learned score, permuted in time	as CNN+adaptive	–
LDA+HMM	standard hybrid HMM	λmax = 1, ρ = 1, β = 1	0.770
LDA+adaptive	total-variation change score	λmax = 1.0, ρ = 0.5, γ = 4, β = 1	0.810
LR	none	C = 0.01	–

λt = λmax[ρ + (1 − ρ)(1 − qt)^γ]; ρ = 1 gives a constant weight. J = 0.5 × macro-F1 + 0.5 × brief-wake event F1, averaged over the seven development recordings (out-of-fold). The tuned decoder selected the lowest λmax in the grid.

2.5 The experimental contrast

According to the reference, the brief-wake rate rose from baseline to stimulation by 4.4 episodes per hour of sleep on average (95% CI −1.7 to 10.0; sign-flip $p = 0.18$; $n = 11$ mice; Fig. 5a and Supplementary Fig. S1). Within the pre-declared 4–16-s window, the reference therefore does not establish a change in brief-awakening rate at the level of the animal.

Most scorers estimated a small decrease instead (mean Δ between −3.4 and −0.6); only LR and LDA estimated small increases (+0.6 and +1.5). The mean signed bias was negative for every scorer (−2.8 to −7.7), and its confidence interval excluded zero for eight of them (Table 5). The primary end point E ranged from 9.6 to 11.1 episodes per hour of sleep, more than twice the mean reference change, with overlapping confidence intervals (Fig. 5b). None of the four pre-declared comparisons was significant (Holm-adjusted $p = 1.0$ throughout). The interval for the difference between the adaptive and the fixed decoder (−1.2 to 2.7) supports neither as the better preserver of the contrast.

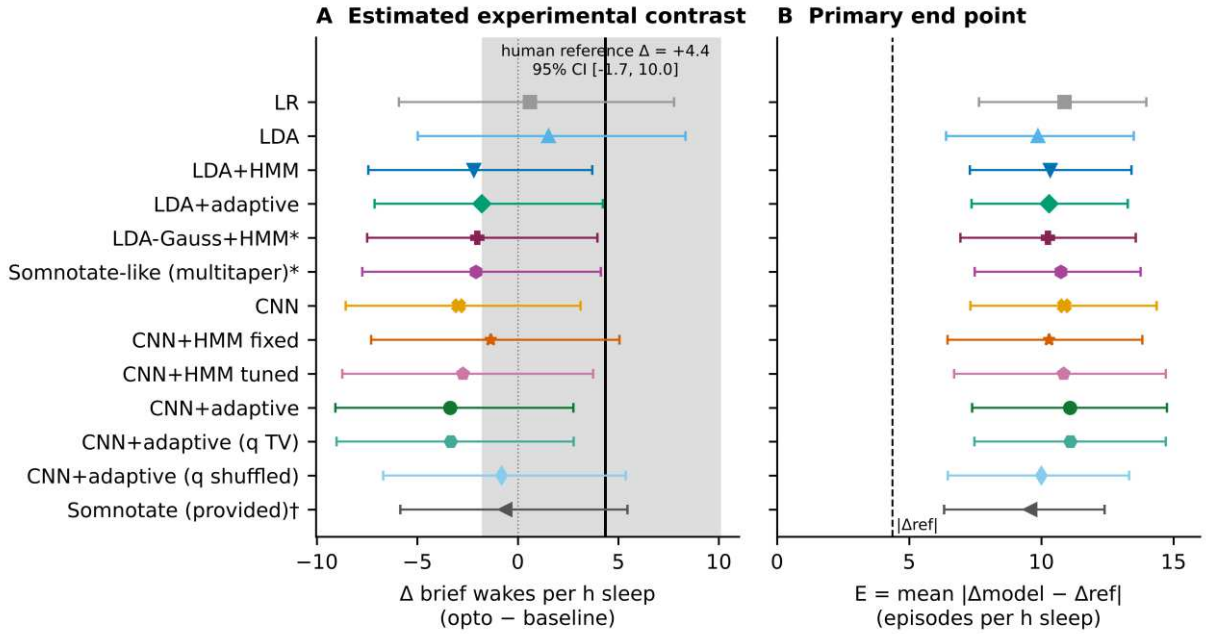


Fig. 5 Experimental contrast and primary end point. a, Black line and grey band, mean change in reference brief-wake rate from baseline to stimulation and its 95% bootstrap CI (n = 11 mice). Symbols, mean change estimated by each scorer; error bars, 95% bootstrap CI of the per-animal bias, plotted around the estimate. **b**, Primary end point E, the mean absolute difference between model and reference changes, with 95% bootstrap CI; dashed line, absolute mean reference change.

Table 5 Experimental contrast (baseline → stimulation) and primary end point, optogenetic test (n = 11 mice).

Scorer	Mean Δ rate (/h sleep)	Bias Δmodel − Δref [95% CI]	E = mean |Δmodel − Δref| [95% CI]
Human reference (TY)	+4.36 [-1.73, 10.05]	–	–
LR	+0.60	-3.76 [-10.28, 3.41]	10.88 [7.64, 13.97]
LDA	+1.53	-2.83 [-9.33, 3.98]	9.87 [6.40, 13.50]
LDA+HMM	-2.18	-6.54 [-11.80, -0.65]	10.34 [7.29, 13.41]
LDA+adaptive	-1.79	-6.15 [-11.49, -0.14]	10.29 [7.36, 13.27]
LDA-Gauss+HMM*	-2.02	-6.38 [-11.85, -0.40]	10.25 [6.94, 13.57]
Somnotate-like (multitaper)*	-2.09	-6.45 [-12.10, -0.23]	10.74 [7.48, 13.75]
CNN	-2.94	-7.31 [-12.92, -1.24]	10.87 [7.32, 14.36]
CNN+HMM fixed	-1.34	-5.70 [-11.65, 0.70]	10.29 [6.45, 13.82]
CNN+HMM tuned	-2.73	-7.09 [-13.10, -0.62]	10.85 [6.70, 14.70]
CNN+adaptive	-3.37	-7.73 [-13.44, -1.60]	11.08 [7.38, 14.75]
CNN+adaptive (q TV)	-3.34	-7.70 [-13.37, -1.57]	11.09 [7.47, 14.71]
CNN+adaptive (q shuffled)	-0.81	-5.17 [-11.06, 1.01]	10.00 [6.46, 13.32]
Somnotate (provided)†	-0.63	-4.99 [-10.21, 1.09]	9.58 [6.32, 12.39]

Pre-declared comparison (difference in |error|)	Mean diff.	Hodges–Lehmann	dz	95% CI	p (sign-flip)	p (Holm)
CNN+adaptive vs CNN+HMM fixed	+0.79	+0.73	+0.22	[-1.24, 2.72]	0.470	1.00
CNN+adaptive vs CNN+HMM tuned	+0.24	+0.11	+0.09	[-1.15, 1.85]	0.797	1.00
CNN+HMM fixed vs CNN	-0.57	-0.67	-0.13	[-2.89, 1.82]	0.657	1.00
LDA+HMM vs LDA	+0.47	-0.36	+0.07	[-2.82, 4.31]	0.823	1.00

Reference change: sign-flip p = 0.176. Confidence intervals are percentile bootstrap over animals (10,000 resamples); sign-flip tests are exact (2^{11} patterns). Negative differences favour the first scorer. A non-significant result does not establish equivalence.

2.6 Sensitivity analyses and ablations

The ordering of E changed little across the pre-declared sensitivity analyses (Fig. 6a, Supplementary Table S2). Excluding artefact-tagged or quality-flagged epochs, or moving the window to 4–8 s or 4–32

s, shifted *E* by at most 1.4 episodes per hour. Expressing rates per hour of analysed recording roughly halved all values (3.9–5.2); this reflects the larger denominator rather than better performance, and the two scales should not be compared. In design B, where models were trained on baselines scored by the same expert as the test data, *E* did not fall for the decoded or CNN models (9.7–11.8) and the bias stayed negative. The undecoded LR and LDA had the lowest *E* in this design (7.7 and 7.3).

Removing channels lowered epoch-level performance, as expected, and hurt event detection more (Fig. 6b, Supplementary Table S3). Without the EMG, event F1 fell from 0.677 to 0.461 for the adaptive CNN and from 0.463 to 0.330 for LDA–HMM. With the frontal EEG alone, the fixed CNN–HMM detected almost no brief awakenings (event F1 0.047). Training-set instead of per-recording standardisation lowered LDA–HMM macro-F1 from 0.931 to 0.872 with little change in *E* (10.1 versus 10.3).

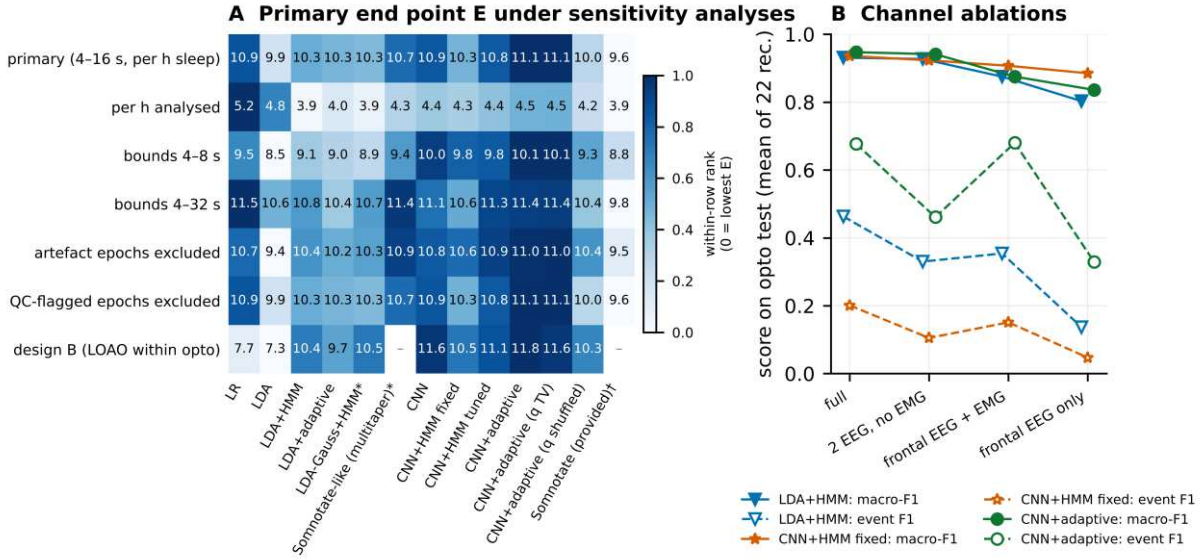


Fig. 6 Sensitivity analyses and channel ablations. a, Primary end point *E* for every scorer under the pre-declared sensitivity analyses (episodes per hour; colour encodes the rank within each row because the per-analysed-hour row has a different scale). Design B was run only for scorers trained within the notebook. **b**, Macro-F1 (solid) and brief-wake event F1 (dashed) on the optogenetic test for three scorers when channels are removed.

2.7 Human ceiling, disagreement and calibration

On the development recordings, each expert was left out in turn and compared, like the scorers, with the majority of the remaining experts. The left-out expert reached macro-F1 0.930 and brief-wake event F1 0.674. The adaptive and undecoded CNNs reached comparable values (macro-F1 0.942–0.944, event F1 0.689–0.701), whereas the fixed CNN–HMM matched on macro-F1 (0.925) but not on events (0.354; Supplementary Table S4). These scorer values come from non-nested selection and are optimistic. Accuracy fell from about 0.97–0.99 on unanimous epochs (0.91 for the multitaper model) to 0.45–0.56 on epochs where fewer than 75% of experts agreed (Supplementary Fig. S2), so much of the residual epoch error lies where experts also disagree. HMM-decoded posteriors were more over-confident than those of LR and of the undecoded CNN (Supplementary Fig. S3).

3 Discussion

Thirteen scorers that differ mainly in temporal decoding reached epoch macro-F1 values within a few points of one another and of a held-out human expert. The brief awakenings they detected differed by an

order of magnitude. A standard hybrid HMM added to a well-performing network erased most single-epoch awakenings while moving macro-F1 by less than one point. The sign of the error depended on the decoder: frame-wise classifiers produced too many brief awakenings and HMM decoding too few, so no single correction factor could be applied across tools. The pattern held in development, under optogenetic stimulation and after sleep deprivation, and it did not depend on the bounds of the brief-wake window. I regard it as the most dependable result of this study.

The pre-declared primary end point, preservation of the baseline-to-stimulation change in brief awakenings, did not separate the models. Every scorer missed the per-animal changes by more than the average change itself, and the reference change could not be distinguished from zero within the 4–16-s window. The emphasis of this paper on event-level consequences of decoding was chosen after these results were known. It should be read as a hypothesis generated in one dataset and tested in another, not as a confirmed finding.

3.1 Why a fixed hidden Markov model erases brief awakenings

The mechanism is simple. A single wake epoch inside NREM sleep requires two transitions, NREM to wake and back again. With 4-s epochs the off-diagonal entries of a transition matrix estimated from training hypnograms are small, so the evidence for wake at that epoch has to outweigh the product of two small probabilities. A network with a ±16-s receptive field has already smoothed its posteriors over neighbouring epochs; adding an HMM counts temporal context twice, and short excursions are the first to go. Three observations fit this account. The tuned constant-weight decoder chose the weakest smoothing available. Removing the EMG, which carries much of the short-lived arousal signal, left the fixed HMM almost blind to brief awakenings. And the loss was milder for LDA–HMM, whose emissions are not context-dependent, than for the CNN–HMM.

The adaptive decoder shows that the trade-off is not inevitable. Relaxing the transition penalty where a learned change score predicts a boundary preserved strong within-state smoothing and recovered the brief awakenings seen by the network alone. The permutation control indicates that the timing of the change score matters, not only its distribution. Decoders of this kind belong to the family of input-dependent transition models,[43] and I make no claim of novelty for the idea. Nor did the adaptive decoder outperform a network without any decoder, and on the primary end point it was no better than the alternatives.

3.2 Relation to previous work

The benchmark was assembled to ask whether scorers withstand experimental manipulations, including the change in transition statistics produced by the optogenetic protocol.[17,42] Somnotate treats transitions probabilistically and reports intermediate states.[17] SPINDLE combined a network with HMM-style constraints and showed that its output supported downstream analyses of bout number, bout duration and transitions across cohorts.[10] The present results do not contradict these reports. They suggest instead that the summaries usually checked (accuracy, state fractions, mean bout durations) can look reassuring while the shortest wake episodes, which account for nearly half of the wake episodes flanked by sleep in these recordings, are treated very differently from one tool to the next. The Somnotate hypnograms distributed with the dataset fell between the undecoded and fixed-HMM scorers on this axis.

Annotation quality frames all of this. Inter-scorer variability limits what any comparison with a single reference can show,[24–27] and in mice it is substantial across laboratories.[28] Agreement on the multi-scorer recordings was high overall, yet every scorer fell close to chance on epochs where experts were split, and the expert who scored the optogenetic recordings marked more brief awakenings than the consensus. Part of the negative bias is therefore probably a difference in scoring style between training and test references

rather than an error of the algorithms alone. Design B, in which training and test labels came from the same expert, did not remove the bias. That argues against style as the only explanation without excluding it.

3.3 Biological interpretation

Optogenetic activation of arousal-promoting circuits can wake mice within seconds.[44] Photoactivation of GAD2 neurons of the lateral preoptic area wakes mice from sleep and produces an activated EEG,[41] and the dataset documentation describes more (brief) awakenings during stimulation.[42] Even with the human reference, I did not detect a clear change in the rate of 4–16-s awakenings per hour of sleep. Several explanations remain open and the present data cannot separate them. Stimulation-evoked awakenings may often outlast 16 s and fall outside the window. The effect may be confined to stimulation periods, which could not be located because the files carry no stimulation timestamps. Normalising by hours of sleep may also absorb part of the effect if sleep time itself changes. I therefore draw no new conclusion about the biology of the manipulation, and the null contrast should not be read as evidence that brief awakenings were unaffected.

The 4-s grid of the reference sets a further limit. Mouse micro-arousals frequently last 2–4 s[36] and are coupled to infraslow noradrenergic dynamics.[33–35] Many such events are partly or wholly invisible at the resolution of the manual scoring used here. The "brief wake episodes" of this study are annotated wake runs of one to four epochs, not physiological micro-arousals. Methods that output finer-grained labels[21] would need a reference at matching resolution to be evaluated on such events.

3.4 Practical implications

When automated scoring is used to quantify fragmentation, event-level metrics for brief awakenings belong next to epoch agreement, with recall stratified by episode duration and with the rates estimated by both the scorer and the reference. Temporal decoders should be reported with and without decoding on the same posteriors, because the decoder alone can change the number of detected brief awakenings several-fold. When a manipulation is expected to alter transition statistics, the automated estimate of the experimental contrast should be checked per animal against a human reference on at least a subset of animals before it is used.

3.5 Limitations

The test sets are small (11 and 6 mice) and come from one laboratory and one recording system. The optogenetic and sleep-deprivation references are single-scorer hypnograms, and the reference contrast in the primary window was uncertain. Animal identity was undocumented for the development recordings, and one pilot annotation duplicates a sleep-deprivation baseline annotation. The original Somnotate software could not be run in the analysis environment; the Somnotate-like multitaper model performed poorly (macro-F1 0.87 on the optogenetic test) and is not a reproduction. Hyper-parameter selection was not nested. CNN results are averages of three networks, and between-seed variability was not reported separately. Raw EEG and EMG have not yet been inspected around a systematic sample of matched, missed and spurious events, and rate errors have not been decomposed into errors on episode counts and on sleep duration (Supplementary Note 4). Finally, the framing of the paper changed after the primary analysis, and the event-level findings need confirmation in an independent dataset with its own pre-specified end point.

In this public mouse dataset, epoch-level agreement did not predict how many brief awakenings an automated scorer would find. Temporal decoding was the main source of the difference: a fixed HMM removed most single-epoch awakenings at almost no cost in macro-F1, and an adaptive decoder restored

them without surpassing an undecoded network. No scorer reproduced the per-animal change in brief awakenings between baseline and optogenetic stimulation, a change that the human reference itself did not establish in the pre-specified window. Studies that use automated scoring to measure sleep fragmentation would gain from reporting event-level, duration-stratified metrics alongside epoch agreement.

4 Methods

4.1 Dataset

I used the Oxford Mouse Polysomnography Benchmark (Zenodo record 10200482, version 1.0.0, CC BY 4.0),[42] accessed through a Kaggle redistribution in which some archives had been repacked. It contains EEG and EMG recordings from freely behaving mice, sampled at 256 Hz and scored in 4-s epochs.[17] Four groups are provided: a 12-h pilot recording scored by 10 experts; six 24-h recordings scored by four or five experts each; six mice recorded during a 12-h sleep-deprivation protocol and on an adjacent baseline day, scored by one expert;[45] and 11 mice recorded for 24 h during optogenetic stimulation of GAD2-expressing neurons of the lateral preoptic area and on a baseline day, also scored by one expert.[41] Somnotate hypnograms are distributed for all recordings. No new animal experiments were performed.

4.2 Signals and annotations

I used the frontal EEG, parietal EEG and EMG channels designated in the dataset documentation. Annotations follow the Visbrain stage-duration format, in which each line gives a state and the cumulative end time of the interval.[46] The format was parsed explicitly; end times were checked to increase monotonically and to match the declared duration, and intervals were projected onto a 4-s grid aligned to recording onset. Sleep-movement labels were mapped to wake, following the Somnotate convention. Artefact-tagged epochs kept their underlying state in the primary analysis, and undefined epochs were masked. Events shorter than one epoch were not evaluated.

4.3 Audit and quality control

Before any modelling, an automated audit listed every file, matched recordings to annotations, compared signal and annotation durations, checked channel assignments and searched for duplicate files. Signals were read directly from the EDF headers and data records. For each channel I computed the fraction of non-finite samples, of flat 1-s blocks and of saturated samples, together with amplitude percentiles. A recording was to be excluded if a used channel was more than 50% flat, contained non-finite values, was discontinuous (EDF+D) or had a file size inconsistent with its header. High-amplitude segments were kept, and epoch-level flat or saturated flags were reserved for a sensitivity analysis.

4.4 Human reference

For multi-scorer recordings I used the consensus hypnogram distributed with the dataset and checked it against a majority vote recomputed from the individual hypnograms, with ties set to undefined. I computed Fleiss' κ, the fraction of unanimous epochs, pairwise Cohen's κ and, for each expert, agreement with the majority of the remaining experts. The optogenetic recordings were scored by a single expert (initials TY) and the sleep-deprivation recordings by another (CBD). These hypnograms served as the available reference and are not described as consensus. Because TY also scored three multi-scorer recordings (pilot, test recordings B and F), those recordings were used to compare this expert's brief-wake annotations with the majority of the others. The expert hypnograms were produced by the investigators

of the original studies and are distributed with the dataset under a CC BY 4.0 licence, with scorers identified only by the initials used in the file names. No scorer was recruited, contacted or asked to perform any task for the present study, and no information about the scorers was used beyond their released annotations.

4.5 Features

The analysis is offline and non-causal. For each 4-s epoch and each EEG channel, a Welch periodogram (1-s Hann windows, 50% overlap) gave the $\log_{10}$ power in 1-Hz bins from 1 to 30 Hz, the $\log_{10}$ band power in the delta (1–4 Hz), theta (6–9 Hz), sigma (10–15 Hz), beta (15–30 Hz) and gamma (30–45 Hz) ranges, and the log theta/delta ratio. EMG features were the log power in 10–45 Hz and 55–100 Hz (avoiding mains interference), the log RMS after 10-Hz high-pass filtering and the standard deviation of the log RMS across eight 0.5-s sub-windows, for 76 features per epoch. Each feature f of recording r was standardised with its median and inter-quartile range,

$$\tilde{x}_{t,f}^{(r)} = \frac{x_{t,f}^{(r)} - \mathrm{med}_t\left(x_{t,f}^{(r)}\right)}{\mathrm{IQR}_t\left(x_{t,f}^{(r)}\right)} \tag{1}$$

This transductive normalisation uses the unlabelled distribution of the test recording. Per-recording standardisation can bias estimates when state proportions differ between conditions,[11] so an ablation used statistics learned on the training set instead. A second branch, intended to approximate the spectral representation of Somnotate, used multitaper log spectra (DPSS, $NW = 2$, three tapers; 1–44 Hz in 1-Hz bins for all three channels) standardised per frequency and per recording.

4.6 Classifiers

Every scorer returns, for each epoch t, a posterior vector $\mathbf{p}_t = (p_t(\mathrm{W}), p_t(\mathrm{N}), p_t(\mathrm{R}))$. Logistic regression (LR) was L2-regularised, with the inverse penalty C chosen in {0.01, 0.1, 1, 10} on the development folds. Linear discriminant analysis (LDA) used shrinkage covariance estimates. A Somnotate-inspired model (LDA-Gauss) projected the features onto two discriminant axes and fitted one Gaussian emission per state in that space; the same model was fitted to the multitaper branch. These are re-implementations of the general approach and not the original software, which requires pomegranate 0.14.4 and could not be installed in the execution environment.[46] The Somnotate hypnograms distributed with the dataset were evaluated as an external comparator. Their training data are undocumented, so their out-of-sample status cannot be assumed.

The temporal convolutional network (CNN) took the sequence of epoch features as input through three one-dimensional convolutional layers (64 channels, kernel size 3, dilations 1, 2 and 1, batch normalisation, ReLU, dropout 0.2), giving a symmetric receptive field of ±4 epochs (±16 s). A state head produced $\mathbf{p}_t$ and a change head produced q_t, the probability that the state at epoch t differs from that at $t-1$. With labels y_t and change targets $c_t = \mathbf{1}[y_t \neq y_{t-1}]$, the network minimised

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \log p_t(y_t) \; - \; \frac{\eta}{T-1}\sum_{t=2}^{T}\left[w_{\mathrm{pos}}\, c_t \log q_t + (1-c_t)\log(1-q_t)\right] \tag{2}$$

where $\eta = 0.5$, $\bar{c}$ is the frequency of state changes in the training labels and the positive class is up-weighted by $w_{\mathrm{pos}} = \sqrt{(1-\bar{c})/\bar{c}}$. Networks were trained with AdamW (learning rate 10^{-3}, weight decay 10^{-4}) on random 256-epoch segments with early stopping on a held-out training recording. Three networks with pre-specified seeds were averaged; seeds were treated as computational repetitions, not as additional observations.

4.7 Temporal decoding

Let π denote the state frequencies and A the 3 × 3 transition matrix estimated from the training hypnograms with a pseudocount of 1. All decoders share one family. The weight given to the transition prior at epoch t is

$$\lambda_t = \lambda_{\max}\,[\rho + (1-\rho)\,(1-q_t)^{\gamma}], \qquad 0 \le \rho \le 1 \tag{3}$$

and the corresponding time-varying transition matrix is the row-wise tempered version of A,

$$\tilde{A}_t(i,j) = \frac{\exp\big(\lambda_t \log A_{ij}\big)}{\sum_k \exp(\lambda_t \log A_{ik})} \tag{4}$$

The decoded hypnogram maximises the sum of tempered emission and transition scores,

$$\hat{\mathbf{s}} = \arg\max_{s_1,\ldots,s_T} \Bigg\{ \log \pi(s_1) + \sum_{t=1}^{T} [\log p_t(s_t) - \beta \log \pi(s_t)] + \sum_{t=2}^{T} \log \tilde{A}_t(s_{t-1}, s_t) \Bigg\} \tag{5}$$

Setting $\beta = 1$ corresponds to the scaled-likelihood convention of hybrid network–HMM systems.[47] With $\gamma = 1$, Eq. (3) reduces to the linear form $\lambda_t = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})(1 - q_t)$ with $\lambda_{\min} = \rho\lambda_{\max}$. Four decoders were compared. "HMM fixed" is the standard hybrid HMM ($\lambda_{\max} = 1$, $\rho = 1$, $\beta = 1$). "HMM tuned" keeps a constant weight ($\rho = 1$) but selects $\lambda_{\max}$ and β, so it can only strengthen or weaken smoothing uniformly; it is the control for any benefit of adaptation. "Adaptive" allows $\rho < 1$, relaxing the constraint where q_t signals a change, with q_t taken from the change head. Two ablations replaced q_t by the total variation between consecutive posteriors,

$$q_t^{\mathrm{TV}} = \frac{1}{2}\,\|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1 \tag{6}$$

or by a within-recording random permutation of the learned q_t. LR and LDA were decoded with the same family; LDA-Gauss used its Gaussian log-likelihoods as emissions with a fixed HMM. Sequences were obtained with the Viterbi algorithm[48] and marginal posteriors with a normalised forward–backward recursion.[40] Both were checked against exhaustive enumeration on random problems, and each decoder was checked in its limiting cases (Supplementary Note 3).

4.8 Validation design

The analysis protocol was written and fingerprinted (SHA-256) before any test result was produced. Development used the pilot and the six multi-scorer recordings. Because the dataset does not document whether these come from distinct animals, I used leave-one-recording-out validation (seven folds) and do not assume independence between development animals. Hyper-parameters were selected on out-of-fold development predictions with the pre-declared objective

$$J = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{2}\,\mathrm{F1}_r^{\mathrm{macro}} + \frac{1}{2}\,\mathrm{F1}_r^{\mathrm{event}}\right] \tag{7}$$

over the $R = 7$ development recordings. The decoder grid covered $\lambda_{\max} \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$, $\rho \in \{0, 0.5, 1\}$, $\gamma \in \{0.5, 1, 2, 4\}$ and $\beta \in \{0, 1\}$, with ties broken in favour of the simpler decoder. Selection was not nested, so development scores are optimistic and are reported only to describe selection.

After selection, all models were refitted on the seven development recordings, frozen, and applied once to the 22 optogenetic recordings (primary test) and the 12 sleep-deprivation recordings (secondary test). In a pre-declared secondary design (design B), models with the frozen hyper-parameters were

trained on the baseline recordings of ten optogenetic mice and tested on both recordings of the eleventh, in turn for each mouse. This design removes the shift between development and test scorers.

4.9 Events and metrics

A brief wake episode was a maximal run of wake epochs lasting 4–16 s (one to four epochs) immediately preceded and followed by NREM or REM sleep. Runs touching a recording boundary or an undefined epoch were treated as censored and excluded. Model hypnograms were masked with the reference mask so that every source was evaluated over the same epochs.

Predicted and reference episodes were matched one to one. A pair was a candidate if the two intervals overlapped or if their onsets differed by at most one epoch. Among candidates, the assignment maximising the summed intersection-over-union,

$$\mathrm{IoU}(a,b) = \frac{|a \cap b|}{|a \cup b|} \tag{8}$$

was found with the Hungarian algorithm.[49] With M matched pairs, N_{ref} reference and N_{pred} predicted episodes, event precision is M/N_{pred}, recall is M/N_{ref} and

$$\mathrm{F1}^{\mathrm{event}} = \frac{2M}{N_{\mathrm{ref}} + N_{\mathrm{pred}}} \tag{9}$$

Onset and duration errors were computed on matched pairs. Recall was also computed for all reference wake episodes flanked by sleep, stratified by duration (4, 8, 12, 16, 20–32, 36–64 and > 64 s). Transitions between consecutive defined epochs were matched in the same way to obtain transition recall, precision, latency and false transitions per hour of sleep. Epoch-level metrics were macro-F1, per-state F1, balanced accuracy and Cohen's κ. Probabilistic outputs were assessed with the Brier score, log-loss and expected calibration error. All definitions were covered by unit tests on synthetic sequences, including boundaries, undefined epochs, censoring and one-to-one matching.

4.10 Statistics

The animal was the unit of inference. For mouse a and source m (reference or model), the brief-wake rate was the number of episodes divided by the hours of sleep of that source, and the experimental change was

$$\Delta_a^m = \frac{N_a^{m,\mathrm{stim}}}{H_a^{m,\mathrm{stim}}} - \frac{N_a^{m,\mathrm{base}}}{H_a^{m,\mathrm{base}}} \tag{10}$$

The pre-declared primary end point and the accompanying signed bias were

$$E = \frac{1}{n}\sum_{a=1}^{n} \left|\Delta_a^{\mathrm{model}} - \Delta_a^{\mathrm{ref}}\right|, \qquad b = \frac{1}{n}\sum_{a=1}^{n} \left(\Delta_a^{\mathrm{model}} - \Delta_a^{\mathrm{ref}}\right) \tag{11}$$

with $n = 11$. Confidence intervals were percentile bootstrap intervals over animals (10,000 resamples).[50] Four pre-declared paired comparisons of absolute errors (adaptive versus fixed HMM, adaptive versus tuned HMM, fixed HMM versus no decoding for the CNN, and LDA–HMM versus LDA) were tested with exact two-sided sign-flip permutation tests. For paired differences d_a,

$$p = 2^{-n} \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \mathbf{1}\left[\left|\frac{1}{n}\sum_{a=1}^{n} \sigma_a d_a\right| \geq \left|\frac{1}{n}\sum_{a=1}^{n} d_a\right|\right] \tag{12}$$

which enumerates all $2^{11} = 2{,}048$ sign patterns. *P* values were adjusted with the Holm procedure.[51] Hodges–Lehmann estimates and paired effect sizes are reported alongside. A *P* value above 0.05 was not read as equivalence, and no power analysis was performed.

The pre-declared sensitivity analyses changed the upper bound of the brief-wake window (8 or 32 s), the rate denominator (hours of analysed recording instead of hours of sleep), the handling of artefact-tagged epochs and of epochs flagged by quality control, and the training design (design B). Ablations removed the EMG, the parietal EEG or both, and replaced per-recording by training-set standardisation.

4.11 Software and reproducibility

The complete pipeline is a single Kaggle notebook (Python 3.13, NumPy 2.1, pandas 2.3, SciPy 1.16, scikit-learn 1.6, PyTorch 2.11). It detects the dataset, reads EDF and stage-duration files with its own readers, writes an audit report, a manifest, the partitions, the configuration and protocol fingerprint, all out-of-sample predictions, per-recording metrics and event matches, and regenerates every table and figure. The full run reported here (run identifier 20261021T222435) completed in under 1.5 h on a single GPU, with stage timings logged by the notebook.

4.12 Use of AI-assisted tools

An AI assistant (Claude, Anthropic) was used to help write the analysis code, prepare figures and edit the manuscript. The author reviewed the code, checked every reported value against the analysis outputs and takes full responsibility for the content.

Data availability

The recordings and annotations are publicly available from Zenodo (https://doi.org/10.5281/zenodo.10200482) under a CC BY 4.0 licence. All derived files (audit report, manifest, partitions, configuration and protocol fingerprint, out-of-sample predictions, per-recording metrics, event matches and figure source data) are generated by the analysis notebook; they are available from the author on reasonable request and will be deposited in a public repository with a persistent identifier before publication.

Code availability

The analysis notebook and the script that regenerates it are available from the author on request and will be released in a public repository with a persistent identifier before publication.

References

1. Saper, C. B., Fuller, P. M., Pedersen, N. P., Lu, J. & Scammell, T. E. Sleep state switching. *Neuron* **68**, 1023–1042 (2010). https://doi.org/10.1016/j.neuron.2010.11.032
2. Weber, F. & Dan, Y. Circuit-based interrogation of sleep control. *Nature* **538**, 51–59 (2016). https://doi.org/10.1038/nature19773
3. Scammell, T. E., Arrigoni, E. & Lipton, J. O. Neural circuitry of wakefulness and sleep. *Neuron* **93**, 747–765 (2017). https://doi.org/10.1016/j.neuron.2017.01.014
4. Rechtschaffen, A. & Kales, A. (eds) *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects* (US Government Printing Office, 1968).
5. Tobler, I., Deboer, T. & Fischer, M. Sleep and sleep regulation in normal and prion protein-deficient mice. *J. Neurosci.* **17**, 1869–1879 (1997).
6. Franken, P., Chollet, D. & Tafti, M. The homeostatic regulation of sleep need is under genetic control. *J. Neurosci.* **21**, 2610–2621 (2001). https://doi.org/10.1523/JNEUROSCI.21-08-02610.2001
7. Sunagawa, G. A., Sei, H., Shimba, S., Urade, Y. & Ueda, H. R. FASTER: an unsupervised fully automated sleep staging method for mice. *Genes Cells* **18**, 502–518 (2013). https://doi.org/10.1111/gtc.12053
8. Exarchos, I. et al. Supervised and unsupervised machine learning for automated scoring of sleep–wake and cataplexy in a mouse model of narcolepsy. *Sleep* **43**, zsz272 (2020). https://doi.org/10.1093/sleep/zsz272

9. Wang, L. A., Kern, R., Yu, E., Choi, S. & Pan, J. Q. IntelliSleepScorer, a software package with a graphic user interface for automated sleep stage scoring in mice based on a light gradient boosting machine algorithm. *Sci. Rep.* **13**, 4275 (2023). https://doi.org/10.1038/s41598-023-31288-2
10. Miladinović, Đ. et al. SPINDLE: end-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species. *PLoS Comput. Biol.* **15**, e1006968 (2019). https://doi.org/10.1371/journal.pcbi.1006968
11. Barger, Z., Frye, C. G., Liu, D., Dan, Y. & Bouchard, K. E. Robust, automated sleep scoring by a compact neural network with distributional shift correction. *PLoS ONE* **14**, e0224642 (2019). https://doi.org/10.1371/journal.pone.0224642
12. Yamabe, M. et al. MC-SleepNet: large-scale sleep stage scoring in mice by deep neural networks. *Sci. Rep.* **9**, 15793 (2019). https://doi.org/10.1038/s41598-019-51269-8
13. Grieger, N., Schwabedal, J. T. C., Wendel, S., Ritze, Y. & Bialonski, S. Automated scoring of pre-REM sleep in mice with deep learning. *Sci. Rep.* **11**, 12245 (2021). https://doi.org/10.1038/s41598-021-91286-0
14. Tezuka, T. et al. Real-time, automatic, open-source sleep stage classification system using single EEG for mice. *Sci. Rep.* **11**, 11151 (2021). https://doi.org/10.1038/s41598-021-90332-1
15. Jha, P. K., Valekunja, U. K. & Reddy, A. B. SlumberNet: deep learning classification of sleep stages using residual neural networks. *Sci. Rep.* **14**, 4797 (2024). https://doi.org/10.1038/s41598-024-54727-0
16. Zaid, Y. W. K. et al. A single-channel EEG classification system for multiscale characterization of mouse vigilance state. Preprint at *bioRxiv* https://doi.org/10.64898/2026.08.14.744824 (2026).
17. Brodersen, P. J. N. et al. Somnotate: a probabilistic sleep stage classifier for studying vigilance state transitions. *PLoS Comput. Biol.* **20**, e1011793 (2024). https://doi.org/10.1371/journal.pcbi.1011793
18. Supratak, A., Dong, H., Wu, C. & Guo, Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1998–2008 (2017). https://doi.org/10.1109/TNSRE.2017.2721116
19. Phan, H., Andreotti, F., Cooray, N., Chén, O. Y. & De Vos, M. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**, 400–410 (2019). https://doi.org/10.1109/TNSRE.2019.2896659
20. Stephansen, J. B. et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **9**, 5229 (2018). https://doi.org/10.1038/s41467-018-07229-3
21. Perslev, M. et al. U-Sleep: resilient high-frequency sleep staging. *npj Digit. Med.* **4**, 72 (2021). https://doi.org/10.1038/s41746-021-00440-5
22. Vallat, R. & Walker, M. P. An open-source, high-performance tool for automated sleep staging. *eLife* **10**, e70092 (2021). https://doi.org/10.7554/eLife.70092
23. Rayan, A., Szabo, A. B. & Genzel, L. The pros and cons of using automated sleep scoring in sleep research. *Sleep* **47**, zsad275 (2024). https://doi.org/10.1093/sleep/zsad275
24. Danker-Hopfe, H. et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J. Sleep Res.* **18**, 74–84 (2009).
25. Rosenberg, R. S. & Van Hout, S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J. Clin. Sleep Med.* **9**, 81–87 (2013). https://doi.org/10.5664/jcsm.2350
26. Lee, Y. J., Lee, J. Y., Cho, J. H. & Choi, J. H. Interrater reliability of sleep stage scoring: a meta-analysis. *J. Clin. Sleep Med.* **18**, 193–202 (2022). https://doi.org/10.5664/jcsm.9538
27. de Chazal, P., Mazzotti, D. R. & Cistulli, P. A. Automated sleep staging algorithms: have we reached the performance limit due to manual scoring? *Sleep* **45**, zsac159 (2022). https://doi.org/10.1093/sleep/zsac159
28. Rose, L. et al. Lack of consensus for manual mouse sleep scoring limits implementation of automatic deep learning models. Preprint at *bioRxiv* https://doi.org/10.64898/2026.03.27.714381 (2026).
29. American Sleep Disorders Association Atlas Task Force. EEG arousals: scoring rules and examples. A preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. *Sleep* **15**, 174–184 (1992). https://doi.org/10.1093/sleep/15.2.174
30. Lo, C.-C. et al. Common scale-invariant patterns of sleep–wake transitions across mammalian species. *Proc. Natl Acad. Sci. USA* **101**, 17545–17548 (2004). https://doi.org/10.1073/pnas.0408242101
31. Swihart, B. J., Caffo, B., Bandeen-Roche, K. & Punjabi, N. M. Characterizing sleep structure using the hypnogram. *J. Clin. Sleep Med.* **4**, 349–355 (2008).
32. Bianchi, M. T., Cash, S. S., Mietus, J., Peng, C.-K. & Thomas, R. Obstructive sleep apnea alters sleep stage transition dynamics. *PLoS ONE* **5**, e11356 (2010). https://doi.org/10.1371/journal.pone.0011356
33. Lecci, S. et al. Coordinated infraslow neural and cardiac oscillations mark fragility and offline periods in mammalian sleep. *Sci. Adv.* **3**, e1602026 (2017). https://doi.org/10.1126/sciadv.1602026
34. Osorio-Forero, A. et al. Noradrenergic circuit control of non-REM sleep substates. *Curr. Biol.* **31**, 5009–5023 (2021). https://doi.org/10.1016/j.cub.2021.09.041
35. Kjaerby, C. et al. Memory-enhancing properties of sleep depend on the oscillatory amplitude of norepinephrine. *Nat. Neurosci.* **25**, 1059–1070 (2022). https://doi.org/10.1038/s41593-022-01102-9

36. Hauglund, N. L. et al. Cortical firing dynamics during micro-arousals vary with sleep/wake history and micro-arousal duration. *Sci. Rep.* **16**, 15391 (2026). https://doi.org/10.1038/s41598-026-45192-y
37. Ward, J. A., Lukowicz, P. & Gellersen, H. W. Performance metrics for activity recognition. *ACM Trans. Intell. Syst. Technol.* **2**, 6 (2011). https://doi.org/10.1145/1889681.1889687
38. Tatbul, N., Lee, T. J., Zdonik, S., Alam, M. & Gottschlich, J. Precision and recall for time series. In *Advances in Neural Information Processing Systems 31* (NeurIPS, 2018).
39. Warby, S. C. et al. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nat. Methods* **11**, 385–392 (2014). https://doi.org/10.1038/nmeth.2855
40. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989). https://doi.org/10.1109/5.18626
41. Yamagata, T. et al. The hypothalamic link between arousal and sleep homeostasis in mice. *Proc. Natl Acad. Sci. USA* **118**, e2101580118 (2021). https://doi.org/10.1073/pnas.2101580118
42. Brodersen, P. J. N. & Akerman, C. J. The Oxford Mouse Polysomnography Benchmark Data Set, version 1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.10200482 (2023).
43. Bengio, Y. & Frasconi, P. Input–output HMMs for sequence processing. *IEEE Trans. Neural Netw.* **7**, 1231–1249 (1996).
44. Adamantidis, A. R., Zhang, F., Aravanis, A. M., Deisseroth, K. & de Lecea, L. Neural substrates of awakening probed with optogenetic control of hypocretin neurons. *Nature* **450**, 420–424 (2007). https://doi.org/10.1038/nature06310
45. Blanco-Duque, C. et al. Oscillatory-Quality of sleep spindles links brain state with sleep regulation and function. *Sci. Adv.* **10**, eadn6247 (2024). https://doi.org/10.1126/sciadv.adn6247
46. Brodersen, P. J. N. Somnotate: automated polysomnography for experimental animal research. *GitHub* https://github.com/paulbrodersen/somnotate (accessed 21 October 2026).
47. Bourlard, H. A. & Morgan, N. *Connectionist Speech Recognition: A Hybrid Approach* (Kluwer Academic, 1994). https://doi.org/10.1007/978-1-4615-3210-1
48. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**, 260–269 (1967). https://doi.org/10.1109/TIT.1967.1054010
49. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**, 83–97 (1955). https://doi.org/10.1002/nav.3800020109
50. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, 1993).
51. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

Acknowledgements

I thank P. J. N. Brodersen, C. J. Akerman and all contributors to the Oxford Mouse Polysomnography Benchmark for making the recordings, annotations and Somnotate code openly available.

Author contributions

A.B. conceived the study, designed and implemented the analysis, interpreted the results, prepared the figures and wrote the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The author declares no competing interests.

Ethics declarations

This study re-analysed publicly available data and involved no new experiments on animals or humans. Ethical approval and licensing for the original animal experiments are described in the source publications.

Consent statement

Informed consent was not applicable. The human experts mentioned in this work are the sleep scorers whose annotations form part of the published Oxford Mouse Polysomnography Benchmark. They did not take part in the present study as research participants, no data were collected from or about them, and their annotations were used only as the reference labels that the dataset releases as a public benchmark under a CC BY 4.0 licence.

Supplementary Information

When accurate sleep scoring misses brief awakenings: event-level consequences of temporal decoding under experimental perturbation

Amine Bougarne

Supplementary Note 1. Model and decoder details. The CNN had 76 input channels (features), three convolutional layers with 64 filters (kernel 3; dilations 1, 2, 1), batch normalisation, ReLU and dropout 0.2, a 1 × 1 state head and a kernel-2 change head operating on adjacent time steps. Training used batches of 32 random 256-epoch segments, 200 steps per epoch, at most 30 epochs and early stopping with patience 5 on a held-out training recording. Three seeds (20261021, 20261022, 20261023) were trained and their posteriors averaged. The positive class of the change head was weighted by w_{pos} as defined in Eq. (2). LR used $C = 0.01$, selected on development macro-F1. Transition matrices and state priors were re-estimated on the training recordings of each fold and, for the frozen models, on all development recordings.

Supplementary Note 2. Pre-declared protocol. The primary end point, the brief-wake window and its sensitivity bounds, the rate denominators, the event tolerance, the selection objective, the four primary comparisons, the label conventions and the normalisation were fingerprinted (SHA-256 5d8c10c9…dd38d) before the first full run. The notebook refuses to run in full mode if these parameters change without a written justification.

Supplementary Note 3. Verification. Readers, grid projection, episode extraction, censoring, one-to-one matching, contrast computation, the Holm procedure and the sign-flip test were covered by unit tests on synthetic sequences, all of which passed in the reported run. Viterbi decoding and forward–backward marginals were compared with exhaustive enumeration on random 6-epoch, 3-state problems. The decoders were checked in their limiting cases: vanishing smoothing reduces to the arg max of the posteriors; $\rho = 1$ reproduces the fixed matrix; $q_t = 1$ with $\rho = 0$ gives uniform transitions; and $\gamma = 1$ reproduces the linear $\lambda_{\min}/\lambda_{\max}$ form. All decoders for a given network operated on the same averaged posteriors.

Supplementary Note 4. Verification analyses not yet performed. (i) Visual inspection of EEG and EMG for a random, pre-specified sample of 4-s and 8–16-s reference episodes, of episodes found by the CNN but removed by the fixed HMM, and of CNN false positives, in several animals; (ii) decomposition of rate errors into errors on episode counts and on hours of sleep; (iii) per-seed variability of CNN results with seeds fixed in advance; (iv) evaluation of the multi-scorer recordings against TY alone versus the consensus, to estimate the share of bias attributable to scoring style; (v) confirmation in an independent multi-scorer dataset.

Supplementary Table S1 Agreement among human scorers on the development recordings.

Recording	Scorers	Tied epochs	Fleiss' κ	Unanimous epochs
pilot	10	1.5%	0.838	72.6%
Sample A	4	3.5%	0.819	80.8%
Sample B	4	2.3%	0.910	90.9%
Sample C	4	2.3%	0.904	89.9%
Sample D	4	3.3%	0.888	87.4%
Sample E	4	1.7%	0.920	90.2%
Sample F	5	0.1%	0.905	87.6%

The distributed consensus matched the recomputed majority on every non-tied epoch of all seven recordings.

Supplementary Table S2 Primary end point E (and mean signed bias) under the pre-declared sensitivity analyses.

Scorer	Primary	Per h analysed	4–8 s	4–32 s	Artefacts excl.	QC excl.	Design B
LR	10.88 (-3.76)	5.22 (-1.13)	9.54 (-3.82)	11.49 (-3.27)	10.67 (-3.58)	10.87 (-3.75)	7.69 (-5.45)
LDA	9.87 (-2.83)	4.78 (-0.66)	8.45 (-3.03)	10.59 (-2.22)	9.35 (-2.46)	9.86 (-2.82)	7.26 (-4.44)
LDA+HMM	10.34 (-6.54)	3.88 (-2.10)	9.06 (-5.67)	10.77 (-6.63)	10.38 (-6.67)	10.34 (-6.54)	10.42 (-6.59)
LDA+adaptive	10.29 (-6.15)	4.05 (-2.03)	8.96 (-5.24)	10.40 (-5.82)	10.20 (-6.11)	10.28 (-6.14)	9.71 (-6.54)
LDA-Gauss+HMM*	10.25 (-6.38)	3.94 (-2.06)	8.89 (-5.54)	10.75 (-6.70)	10.28 (-6.50)	10.25 (-6.38)	10.49 (-6.89)
Somnotate-like (multitaper)*	10.74 (-6.45)	4.27 (-2.08)	9.41 (-5.72)	11.43 (-6.37)	10.92 (-6.67)	10.74 (-6.45)	–
CNN	10.87 (-7.31)	4.36 (-2.72)	10.00 (-7.09)	11.07 (-7.36)	10.76 (-7.29)	10.87 (-7.31)	11.61 (-8.82)
CNN+HMM fixed	10.29 (-5.70)	4.32 (-1.60)	9.79 (-5.62)	10.65 (-5.85)	10.65 (-6.01)	10.29 (-5.70)	10.52 (-5.95)
CNN+HMM tuned	10.85 (-7.09)	4.41 (-2.46)	9.82 (-6.66)	11.34 (-7.39)	10.93 (-7.22)	10.85 (-7.09)	11.14 (-7.86)
CNN+adaptive	11.08 (-7.73)	4.51 (-2.96)	10.09 (-7.35)	11.39 (-7.96)	11.02 (-7.72)	11.08 (-7.73)	11.75 (-9.04)
CNN+adaptive (q TV)	11.09 (-7.70)	4.52 (-2.90)	10.10 (-7.38)	11.40 (-7.85)	11.04 (-7.74)	11.09 (-7.70)	11.61 (-8.78)
CNN+adaptive (q shuffled)	10.00 (-5.17)	4.17 (-1.31)	9.34 (-5.36)	10.45 (-5.21)	10.39 (-5.46)	10.00 (-5.17)	10.27 (-5.80)
Somnotate (provided)†	9.58 (-4.99)	3.86 (-1.61)	8.83 (-4.80)	9.77 (-4.94)	9.53 (-5.00)	9.57 (-4.98)	–

Values are E (bias) in episodes per hour; the per-analysed-hour column uses a different denominator and its scale is not comparable with the others. 95% CIs are given in the notebook output file sensitivity_analyses.csv. Design B was not applicable to the externally provided hypnograms or to the multitaper model.

Supplementary Table S3 Channel and normalisation ablations, optogenetic test.

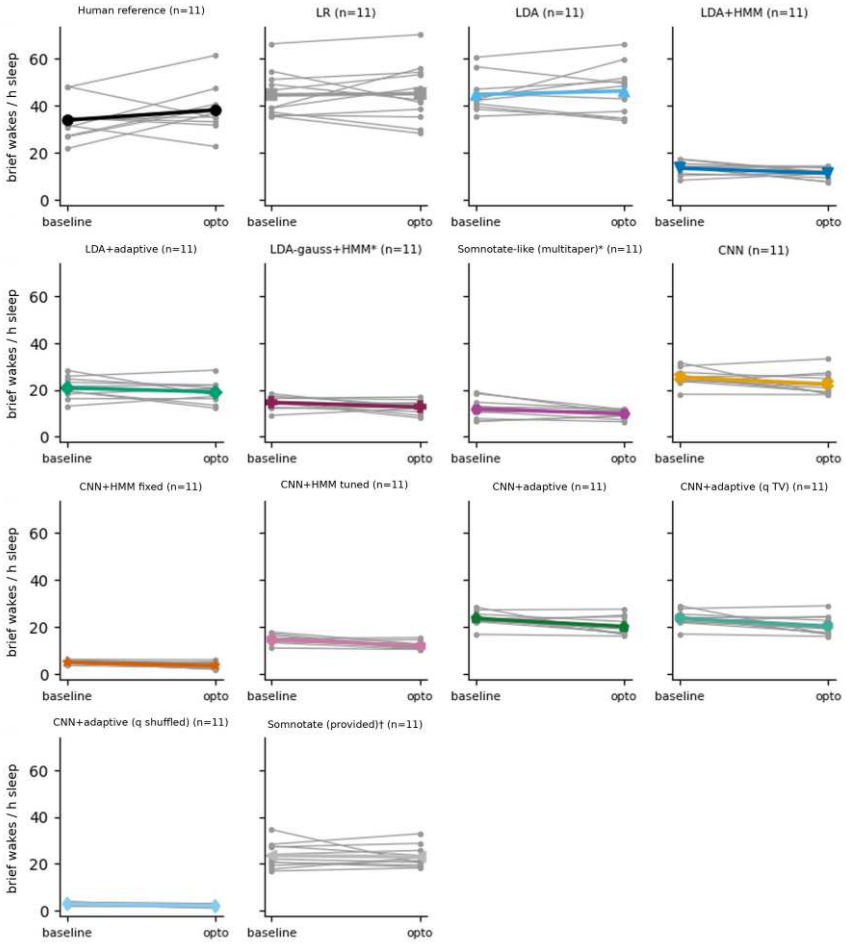
Scorer	Input / normalisation	Macro-F1	Brief-wake event F1	E [95% CI]
LDA+HMM	full	0.931	0.463	10.34 [7.29, 13.41]
LDA+HMM	2 EEG, no EMG	0.927	0.330	9.66 [5.98, 13.08]
LDA+HMM	frontal EEG + EMG	0.875	0.354	10.59 [7.30, 13.71]
LDA+HMM	frontal EEG only	0.803	0.137	9.98 [6.34, 13.34]
LDA+HMM	train-set standardisation	0.872	0.394	10.08 [6.40, 13.48]
CNN+HMM fixed	full	0.936	0.201	10.29 [6.45, 13.82]
CNN+HMM fixed	2 EEG, no EMG	0.923	0.105	9.85 [6.11, 13.24]
CNN+HMM fixed	frontal EEG + EMG	0.907	0.151	10.22 [6.60, 13.55]
CNN+HMM fixed	frontal EEG only	0.885	0.047	9.58 [6.00, 12.84]
CNN+adaptive	full	0.947	0.677	11.08 [7.38, 14.75]
CNN+adaptive	2 EEG, no EMG	0.941	0.461	9.97 [5.83, 13.87]
CNN+adaptive	frontal EEG + EMG	0.875	0.680	11.27 [7.67, 14.63]
CNN+adaptive	frontal EEG only	0.836	0.329	10.28 [6.26, 14.11]

Decoder hyper-parameters were frozen at the values selected with the full feature set.

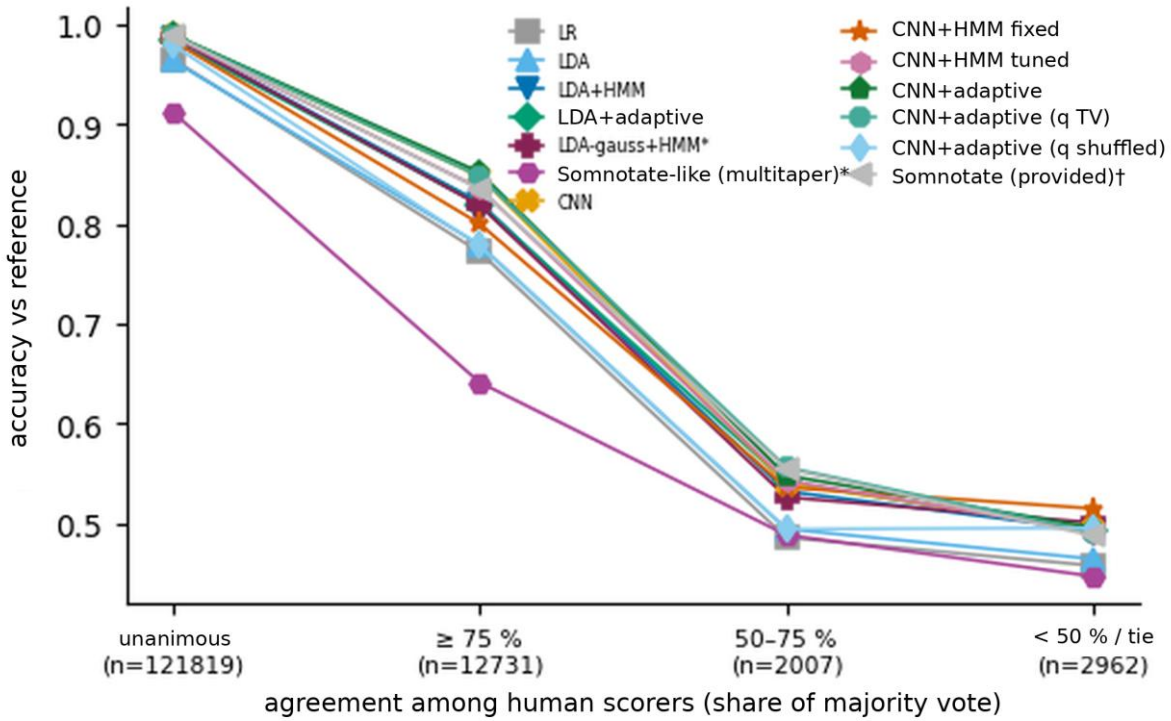
Supplementary Table S4 Development recordings: comparison with the majority of the other experts (leave-one-expert-out).

Source	Macro-F1	Cohen's κ	Brief-wake event F1
human (left-out)	0.930	0.904	0.674
CNN+adaptive	0.944	0.917	0.701
CNN	0.942	0.915	0.689
CNN+HMM fixed	0.925	0.895	0.354
CNN+HMM tuned	0.943	0.912	0.594
LDA+HMM	0.941	0.917	0.589
Somnotate (provided)†	0.939	0.921	0.642
LR	0.898	0.861	0.541

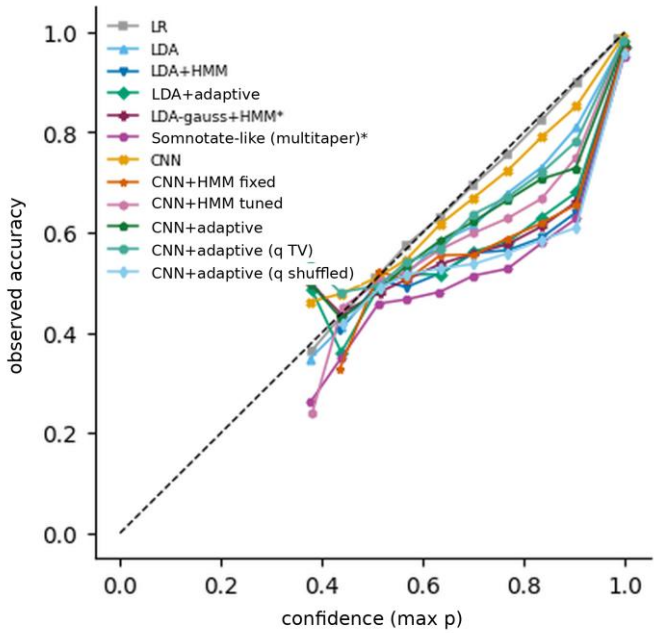
Recordings with at least three scorers. For each left-out expert, the reference is the majority of the remaining experts; scorers are evaluated against the same reference. Scorer values come from non-nested selection and are optimistic.



Supplementary Fig. S1 Per-animal brief-wake rates at baseline and under stimulation. Grey lines, individual mice (n = 11); coloured line, mean. Rates are per hour of sleep of the source shown. Rendered from the analysis notebook; labels translated.



Supplementary Fig. S2 Accuracy by level of agreement between human scorers. Development recordings with at least three scorers, out-of-fold predictions. Epochs were grouped by the share of experts voting for the majority state; *n* is the number of epochs per group. Rendered from the analysis notebook; labels translated.



Supplementary Fig. S3 Reliability diagram on the optogenetic test. Epochs pooled over the 22 recordings, 15 confidence bins (bins with fewer than 20 epochs hidden). The summary is descriptive because epochs are not independent; per-recording expected calibration errors are in the metrics file. Rendered from the analysis notebook; labels translated.

Supplementary Data. Output files. The following files are generated by the notebook and will be deposited with the code: manifest.csv, splits.csv, exclusions.csv, configuration.json, software_versions.txt, selected_hyperparameters.json, hyperparameter_search_dev.csv, metrics_per_recording.csv, event_matches.csv, oos_predictions_*.csv.gz, primary_endpoint_per_animal.csv, secondary_effect_preservation.csv, sleep_deprivation_effect_preservation.csv, ablations.csv, abstention_secondary.csv,

performance_by_human_agreement.csv, calibration_bins_test_opto.csv, annotator_leave_one_out_profiles.csv, human_ceiling_leave_one_out.csv, data_dictionary.csv, figure_legends.csv and the audit folder.