# Real-world sleep phenotyping in autism: compliance, domain shift, and scalable wearable biomarkers

**Sanketh Vedula**[†] SVEDULA@PRINCETON.EDU
*Department of Electrical and Computer Engineering, Princeton University*
*Princeton Precision Health, Princeton University*
*Eric & Wendy Schmidt Center, Broad Institute of MIT and Harvard*

**Andrew Kim**[†] ANDREWKIM@PRINCETON.EDU
*Department of Chemical and Biological Engineering, Princeton University*
*Princeton Precision Health, Princeton University*

**Manoj Kumar** MK35@PRINCETON.EDU
*Princeton Precision Health, Princeton University*

**Olga Troyanskaya** OGT@PRINCETON.EDU
*Department of Computer Science, Princeton University*
*Lewis-Sigler Institute of Integrative Genomics, Princeton University*
*Princeton Precision Health, Princeton University*
*Center for Computational Biology, Flatiron Institute*

**Guillermo Sapiro** GUILLERMOS@PRINCETON.EDU
*Department of Electrical and Computer Engineering, Princeton University*
*Princeton Precision Health, Princeton University*
*Apple*

[†] *Equal contribution.*

## Abstract

Sleep is a scalable target for digital health, but real-world sleep phenotyping remains challenging, especially in autism spectrum disorder (ASD), where laboratory polysomnography (PSG) is difficult to deploy longitudinally. We study ambulatory sleep sensing in the Simons Sleep Project, a multimodal dataset of 200 children recorded over 14 nights using dry-electrode electroencephalography (EEG) headbands and wrist-worn sensors (photoplethysmography (PPG), inertial measurement units (IMU), temperature). We first show that EEG headband compliance declines systematically overnight and is associated with behavioral measures, introducing structured missingness and biasing sleep estimates. In contrast, wrist-worn sensors maintain near-complete adherence. We then demonstrate a substantial domain shift: sleep-staging models trained on clinical-grade data show limited off-the-shelf transfer to wearable recordings. However, this shift is learnable with limited supervision. For EEG, zero-shot models perform near chance, while fine-tuning on 30–50 participants (270-450 nights) recovers performance; fine-tuning the public SleepFM foundation model on 102 participants (∼9 nights per participant) achieves a macro F1 of 0.776. For PPG, the public PaPaGei foundation model achieves only 0.375 macro F1 out-of-the-box, whereas supervised models trained on PPG features reach 0.710, further indicating limited off-the-shelf transfer from clinical-grade pretraining to wearable recordings. Watch sensor ablations identify PPG as the most informative standalone peripheral signal, with PPG+IMU capturing most staging information from the wrist. EEG channel analysis further shows that frontal–occipital derivations outperform prefrontal channels, informing minimal-channel designs. Together, these results establish a modality hierarchy

for ambulatory sleep phenotyping: wrist-worn PPG+IMU provides a scalable backbone for longitudinal monitoring, while EEG remains valuable for micro-architectural analysis when tolerated. We anticipate these lessons on compliance-aware design, domain adaptation, and principled sensor selection to generalize to other vulnerable populations and at-home deployment of digital health studies.

**Keywords:** sleep phenotyping, ambulatory study, autism spectrum disorder, digital phenotyping, PPG, EEG

## 1. Introduction

Sleep is a fundamental component of human health, reflecting coordinated changes across neural, autonomic, metabolic, and thermoregulatory systems. Because it unfolds over extended periods in a quiescent state, it provides a uniquely informative window into resting physiology and overall health, making it well-suited for digital health and wearable-based biomarkers (Perez-Pozuelo et al., 2020). This opportunity is particularly important in autism spectrum disorder (ASD), where sleep disturbances affect 50–80% of individuals and are associated with behavioral difficulties, cognitive function, and quality of life (Miano et al., 2007; Devnani and Hegde, 2015). However, polysomnography (PSG), the gold standard for sleep assessment, is expensive, intrusive, and largely confined to laboratory settings, making multi-night recordings impractical. These limitations are exacerbated in neurodevelopmental populations, where sensory sensitivities and poor tolerance of obstructive sensors reduce participation and data quality. Together, these constraints limit reliable longitudinal sleep phenotyping and motivate the need for sensory-aware, machine learning–based approaches that enable multi-night assessment under real-world conditions.

Wearable sleep monitoring has shown promise, particularly through multimodal systems that combine motion and peripheral physiology (Mahato et al., 2024; Kucukosmanoglu et al., 2025). However, many sleep studies in ASD populations are conducted in laboratory settings and focus on aggregate staging accuracy (Hodge et al., 2012), with less attention to the practical factors that determine whether a sensing modality remains useful under repeated home use. At the same time, commercial always-on devices such as the Apple Watch (Apple Inc., 2025), Oura ring (Svensson et al., 2024), and Whoop band (Miller et al., 2020) illustrate the practicality of continuous sleep monitoring in real-world settings. In ambulatory studies of clinically heterogeneous populations, the value of a sensor depends not only on its theoretical information content, but also on whether it remains tolerable, available, and robust across the night and across repeated nights. Home-based sleep monitoring is therefore not simply a deployment of laboratory methods, but a distinct machine learning problem involving sensor degradation, selective missingness, and domain shift between clinical and wearable acquisition settings.

We study these challenges in the Simons Sleep Project (SSP) dataset (Hacohen et al., 2026), a unique ambulatory multimodal dataset of 200 participants with up to 14 overnight recordings per participant. SSP includes dry-electrode EEG from a DREEM headband (Beacon Biosignals, 2026), together with wrist-based photoplethysmography (PPG), inertial measurement unit (IMU), and peripheral temperature sensing from an Empatica wristband (Empatica, 2026). This setting enables a systematic study of wearable sleep phenotyping in a pediatric ASD cohort, spanning compliance, domain adaptation, and sensor selection. Our work makes the following contributions.

**Compliance is structured rather than random.** We show that headband EEG signal quality declines progressively over the course of the night, disproportionately underrepresenting late-night sleep stages, and that data loss is systematically associated with participant behavioral traits. In contrast, wrist-worn sensing maintains near-complete compliance. These findings motivate a compliance-aware filtering pipeline and show that nominally information-rich modalities can become selectively unavailable in precisely the participants for whom scalable monitoring is most needed.

**PSG-trained models do not generalize to wearable EEG, but the domain shift is learnable.** We evaluate standard PSG preprocessing pipelines, conventional sleep-staging models, and recently introduced sleep foundation models on DREEM EEG recordings, and find limited transfer from laboratory PSG to wearable EEG across all settings. PSG-oriented preprocessing pipelines often fail on DREEM recordings, while automatic staging models ranging from YASA (Vallat and Walker, 2021) and LUNA (Purcell, 2024) to SleepFM (Thapa et al., 2026) show poor zero-shot generalization. This gap can be reduced with modest in-domain supervision: fine-tuning on as few as 30–50 participants improves performance to approximately 0.710 macro F1, while end-to-end fine-tuning of SleepFM with 102 participants reaches 0.776. These results indicate that the lab-to-wearable gap reflects a structured, learnable domain shift rather than a complete absence of usable EEG signal. Per-channel analysis further shows that frontal-to-occipital derivations are more informative for sleep staging than prefrontal channels, offering practical guidance for minimal-channel wearable EEG design.

**Wrist-worn PPG supports DREEM-referenced sleep staging, but foundation models remain sensitive to acquisition setting.** Given the limited overnight compliance of headband EEG, we next examine whether sleep stages can be approximated from wrist-worn sensors alone. Supervised models trained on wearable PPG-derived features reach 0.710 macro F1, with performance already reaching 0.680 using data from only 30 participants, indicating that wrist-worn PPG contains substantial sleep-stage information in this ambulatory setting. However, PaPaGei (Pillai et al., 2025), a PPG foundation model pre-trained on clinical-grade transmission PPG, reaches only 0.375 macro F1 on wrist-worn reflective PPG. Fine-tuning partially recovers performance, suggesting that this gap reflects acquisition-domain mismatch rather than a fundamental absence of useful physiological signal. Together with the EEG results, these findings show that foundation models trained on laboratory or clinical-grade biosignals require explicit in-domain adaptation before deployment in ambulatory wearable settings.

**PPG is the most informative wrist-worn signal, defining a practical modality hierarchy.** We finally compare the information contributed by each wrist-worn sensor through systematic ablations. Among peripheral modalities, PPG is the strongest standalone signal for sleep staging, outperforming both IMU and peripheral temperature. Combining PPG with IMU captures most of the staging information available from the wrist, reaching 0.708 macro F1. Taken together with the compliance analysis, these results suggest that wrist-worn PPG+IMU provides the most scalable basis for longitudinal monitoring, while EEG remains valuable for micro-architectural analysis when tolerated.

Methodologically, we frame ambulatory sleep phenotyping as a deployment-centered ML problem rather than a standard sleep-staging benchmark. We evaluate each sensing and

modeling strategy along three practical axes: sensor availability, transfer across acquisition regimes, and the sleep information retained by reduced sensor sets. This perspective provides a framework for wearable ML studies in which missingness, tolerability, and domain shift are central rather than incidental challenges. Our study also benchmarks publicly released physiological foundation models (SleepFM, PaPaGei) on the SSP dataset, characterizing their failure modes under realistic sensing, different sensors and population shift. We further test whether these failures persist across cohorts through external validation on the DREAMT public wearable-PPG dataset (Wang et al., 2024). To assist the reader, following a notation summary in Table B.1, we provide a summary of the research questions, experimental designs, and key findings in Table C.1. A detailed discussion of related work on automated sleep staging, wrist-worn PPG sensing, and physiological foundation models is provided in Supplementary A.

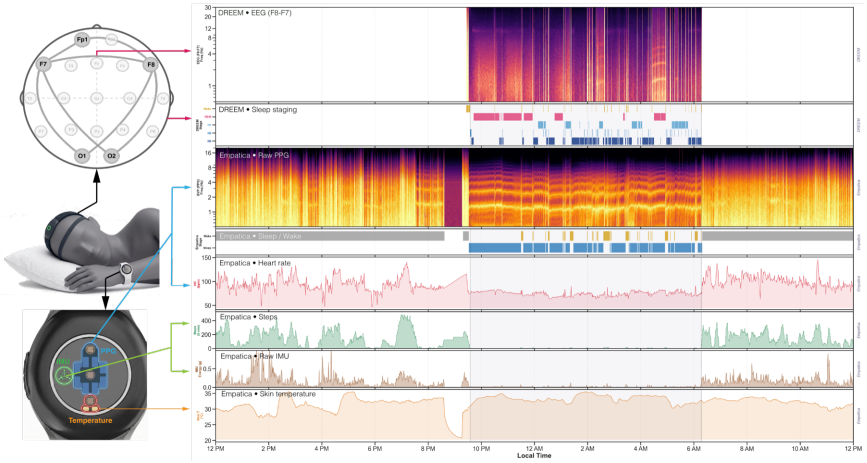



Figure 1: **Multimodal ambulatory setup for real-world sleep phenotyping in ASD.** The figure illustrates the Simons Sleep Project recording configuration, showing a participant equipped with a DREEM dry-electrode EEG headband and an Empatica wristband for overnight home monitoring. The EEG montage (Fp1, F7, F8, O1, O2) forms multiple bipolar derivations used for sleep staging, alongside representative EEG spectrograms and corresponding sleep-stage annotations. Complementary wrist-worn signals, including photoplethysmography (PPG), inertial measurement unit (IMU), and temperature, are visualized with their respective spectrograms and activity traces, highlighting the multimodal nature of the dataset used for scalable, real-world sleep analysis.

### Generalizable Insights about Machine Learning in the Context of Healthcare

This work offers insights that extend beyond ASD sleep monitoring. First, when deploying wearable-based ML systems in vulnerable and young populations, compliance is not merely a data quantity problem but a source of structured missingness. Participants who lose the most data may also be those with the most clinically relevant traits, creating a selection bias that standard missing-data assumptions do not address. Ignoring this structure can produce false discoveries. In our cohort, unfiltered wake-after-sleep-onset (WASO) appeared

to differ between autistic and typically developing participants, but this difference did not survive compliance-aware filtering, and both off-head propensity and the compliance score track behavioral dysregulation and autism-severity scores. Compliance-aware analysis is therefore a prerequisite for valid group-level inference in vulnerable cohorts, independent of the specific staging model used. Second, physiological ML models trained on laboratory or clinical-grade signals, such as PSG-based EEG or transmission PPG, can exhibit substantial domain shift when applied to consumer-grade wearable data. However, this shift is structured and recoverable with modest in-domain fine-tuning, suggesting that cross-acquisition validation is a prerequisite for deploying pre-trained physiological models at scale. Finally, sensor selection for ambulatory monitoring should be evaluated jointly on information content and real-world tolerability. In our setting, EEG is the richest modality but has the weakest compliance, whereas wrist PPG is highly tolerated and captures much of the staging information. This reframes the design problem from maximizing signal fidelity alone to maximizing deployable accuracy.

## 2. Simons Sleep Project (SSP) Dataset

The SSP dataset (Hacohen et al., 2026) is, to the best of our knowledge, the largest multimodal, multi-night sleep study in a pediatric autism cohort. It comprises 200 children aged 10–17 years, recruited as sibling pairs in which one sibling is diagnosed with autism and the other is either typically developing (TD) or diagnosed with attention-deficit/hyperactivity disorder (ADHD). ASD diagnosis was established using the Social Communication Questionnaire (SCQ), derived from the gold-standard Autism Diagnostic Interview-Revised (ADI-R) (Lord et al., 1994). The groups are age-matched ($p = 0.120$), but differ in sex ratio, consistent with the well-established male predominance in ASD (Zhang et al., 2020). In addition, 57% of autistic participants carry the frequently co-occurring ADHD diagnosis (Ghirardi et al., 2018). Each participant was recorded for up to 14 nights at home, enabling the study of night-to-night variability that single-night laboratory PSG cannot resolve. Recordings were synchronously collected using two wearable devices. The first was a Beacon Biosignals DREEM 3S dry-electrode EEG headband (Arnal et al., 2020; Beacon Biosignals, 2026), which provides 7 bipolar derivations from 5 scalp electrodes together with AASM-compliant automatic sleep staging. The second was an Empatica Embrace Plus wristband (Garbarino et al., 2014; Empatica, 2026), which provides photoplethysmography (PPG), 3-axis accelerometry (IMU), and peripheral temperature (Figure 1). Although the DREEM has FDA 510(k) clearance (K223539) and achieves agreement with expert-scored PSG comparable to human inter-scorer performance (83.5% accuracy) (Arnal et al., 2020), this validation was performed in adults aged 22–65 rather than in pediatric populations. Full cohort demographics are reported in Table D.1. The mismatch between the DREEM-validated cohort and the SSP cohort motivates the in-domain fine-tuning experiments presented in this paper. Because expert-scored PSG is unavailable for validating the SSP cohort, we further assess DREEM label reliability indirectly by comparing its stage proportions against normative pediatric AASM-PSG values (Scholle et al., 2011) and expected developmental trends. For example, N2 and REM proportions match closely, with small opposite-signed biases on N1 and N3, and the age-related N3 $\rightarrow$ N2 redistribution is reproduced (Supplementary E).

## 3. Wearable EEG compliance is structured and biases sleep estimates

A central challenge in home-based EEG monitoring is that signal availability cannot be taken for granted across the night. Unlike wrist-worn sensors, headband EEG is susceptible to electrode displacement, poor contact, and voluntary removal during sleep, making compliance a core determinant of data quality. We therefore begin by characterizing overnight EEG availability, asking both how signal quality changes over the sleep window and whether these losses are systematically associated with participant characteristics.

**Biased EEG signal quality declines over the sleep window.** The DREEM headband is susceptible to electrode displacement and voluntary removal during sleep, resulting in progressive signal degradation over the night. To quantify within-night signal loss, we define a per-participant compliance score as an exponential moving average of on-head epochs (Eq. (F.1)), summarized as a per-participant mean (Eq. (F.2)); see Supplementary F for full derivation. Of the 200 enrolled participants, 193 wore the DREEM headband and 196 wore the Empatica wristband for at least one valid night. Across the valid DREEM population, mean compliance declines to approximately 77% before the final wake period, whereas the Empatica wristband remains near 100% throughout (Figure 2(A)). This progressive loss of EEG signal has direct implications for sleep feature estimation. For example, Rapid eye movement (REM) sleep would be underestimated while slow-wave (N3) sleep overrepresented because their durations vary systematically across the night (Carskadon and Dement, 2011). These patterns suggest that sleep features derived from the early portion of the sleep window, where signal quality is most consistently preserved, are likely to be the most reliable for phenotyping. Although the compliance score is defined from DREEM's binary on-/off-head labels, independent DREEM-agnostic probes confirm that raw signal quality itself degrades across the night. Amplitude clipping, electrode-jump, and artifact rates increase from early to late sleep, and recordings with lower compliance carry systematically more raw contamination (Supplementary M). The compliance score is therefore a conservative lower bound on raw signal quality.

**Compliance is also associated with participant characteristics.** To assess whether compliance is structured rather than random, we examined its association with participant characteristics. Older children tolerate the headband better, with compliance scores positively associated with age ($\rho = 0.36$, $p < .001$), as shown in Figure 2(B). This positive correlation is present across the full cohort and within most diagnostic groups, but is weaker in the ASD subgroup (Figure 2(C)). Female participants show a weak negative trend in overnight compliance (Figure F.1) that may reflect signal degradation from dry-electrode contact with longer or thicker hair, but this trend is not statistically conclusive and no definitive conclusions can be drawn. Because younger participants lose more data, studies targeting early childhood are likely to face the greatest EEG attrition. Beyond these macro characteristics, the per-participant compliance score is also associated with autism core-severity measures spanning the social, repetitive, and sensory DSM-5 domains, most notably the Social Responsiveness Scale ($\rho = 0.21$, $p < .01$); these associations remain significant after adjusting for age and sex (Figure G.1). We emphasize that all such associations are correlational and, although adjusted for age and sex, are not fully adjusted for the other partially collinear behavioral and diagnostic variables in this cohort.

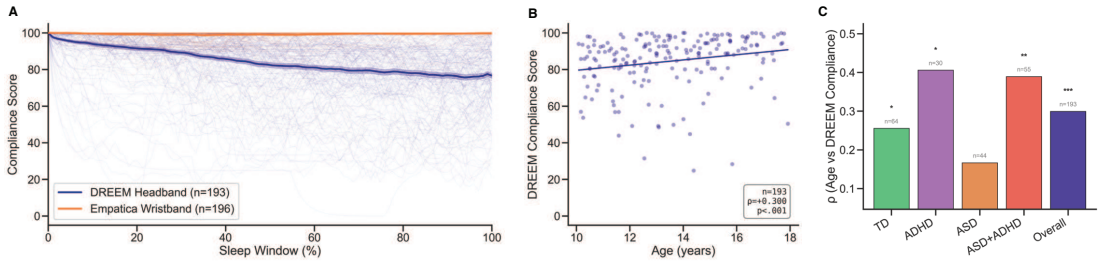



Figure 2: **EEG compliance declines overnight while wrist-worn sensing remains stable, with systematic variation across participants.** (A) Mean on-head compliance score (Eq. (F.1)) over the normalized time from sleep onset to final awakening for the DREEM headband ($n = 193$, blue) and the Empatica wristband ($n = 196$, orange). Thin lines show individual participants; solid bold lines show group means, with standard error of the mean shaded. Mean on-head compliance for the DREEM EEG headband decreases progressively across the normalized sleep window, whereas the Empatica wristband maintains near-constant adherence. (B) Average DREEM compliance score per-participant increases with age, indicating improved tolerance in older participants. (C) Spearman $\rho$ , p-values, and number of participants (n) are annotated for the cohort and each diagnostic group. Compliance varies across diagnostic groups, suggesting that data availability is structured and should be considered when deploying wearable sensors in real-world settings.

**Off-head events are structured and linked to behavioral traits.** By cross-referencing DREEM with concurrent Empatica signals, we observe cases in which a participant removes the headband while continuing to sleep, which we refer to as off-head periods (Figure 3(A)). Participants with greater off-head time tend to score higher on measures of behavioral dysregulation, including Child Behavior Checklist (CBCL) Attention Problems, CBCL Total Problems, Aberrant Behavior Checklist (ABC) Hyperactivity, and sensory processing scores, as shown in Figure 3(C). These associations are descriptive and should not be interpreted causally. In particular, behavioral scores, diagnosis, age, sex, and ADHD status are partially correlated in this cohort, and future work should quantify their independent contributions using multivariable models. Our main conclusion does not rely on attributing compliance loss to any single factor, but rather on the observation that missingness is structured and associated with participant characteristics. We note that off-head propensity and the compliance score capture the same underlying phenomenon through different filters. Off-head propensity isolates prolonged removals during which the concurrent Empatica signal indicates ongoing sleep (that is, volitional non-adherence), whereas the compliance score is also sensitive to transient dry-electrode slippage. Across the cohort the two are almost perfectly anti-correlated ($\rho \approx -0.997$), so analyzing either yields the same conclusion that adherence is trait-correlated (Supplementary G, Figure G.2).

**Poor compliance degrades EEG-derived sleep metrics.** Participants with lower compliance show greater disagreement between DREEM- and Empatica-derived estimates of total sleep time (TST) (Figure 3(B)). When nights with a high off-head percentage are excluded, the two modalities show improved agreement in their independently estimated sleep windows (Figure H.1). This observation motivates a compliance-aware filtering pipeline

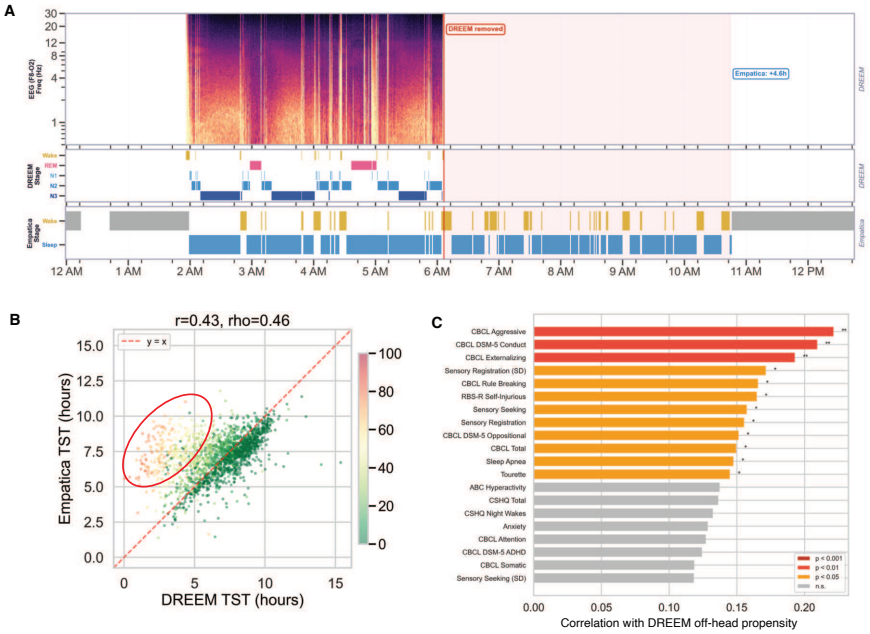



Figure 3: **Off-head events are systematic, linked to behavioral traits, and reduce the reliability of EEG-derived sleep estimates.** (A) A representative participant's EEG spectrogram (top), DREEM sleep stages (middle), and Empatica sleep/wake detection (bottom) show headband removal during ongoing sleep (an off-head period). (B) DREEM-derived total sleep time (TST) versus Empatica TST shows moderate agreement, with points colored by DREEM off-head %; larger deviations track higher off-head %, and the highlighted region marks nights where the headband came off, causing DREEM to underestimate sleep relative to Empatica. (C) Off-head propensity correlates more strongly with dysregulation-related scores (e.g., CBCL, ABC, sensory processing), indicating that off-head events are linked to behavioral traits rather than random.

and highlights that compliance in real-world settings is not merely a problem of reduced data quantity, but also a potential source of systematic bias that is seldom considered in machine learning for healthcare.

**The effect of EEG headband on sleep.** Beyond signal loss, the headband may itself affect sleep. Because participants wore DREEM on only a subset of Empatica-recorded nights, we compared Empatica-derived sleep metrics within participants between headband and Empatica-only nights, restricting to participants with at least three Empatica-only nights. TST was higher and restfulness lower on headband nights, whereas WASO and sleep efficiency showed no difference (Figure J.1); all effects were small to moderate ($|d_z| \leq 0.39$). Given the limited sample size and observational design, these results should be interpreted cautiously, but they highlight that sensing devices may influence the physiology they aim to measure.

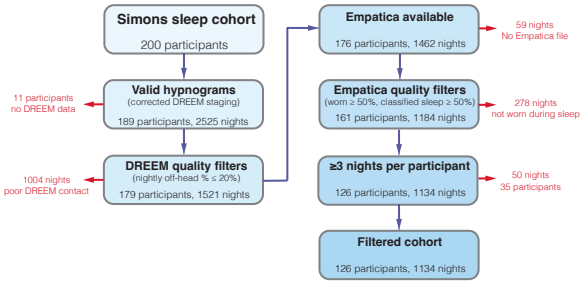



Figure 4: **Compliance-aware data filtering for downstream multimodal analysis.** The figure shows a sequential filtering pipeline that removes participants and nights based on EEG signal quality, wristband availability, and minimum data requirements. Counts at each stage illustrate how quality control reduces unreliable recordings.

## 4. Compliance-aware filtering for reliable cross-sensor analysis

The compliance issues described above affect not only EEG feature quality, but also the reliability of cross-sensor comparisons. Because this progressive signal loss cannot be reliably imputed without fabricating physiological state transitions, we first construct a compliance-aware filtering pipeline to retain nights for which EEG remains sufficiently available for meaningful multimodal evaluation.

**Principled quality filtering retains a balanced cohort for longitudinal analysis.** The compliance-aware filtering pipeline, summarized in Figure 4, was designed to retain recordings with sufficient overlapping EEG and wrist-sensor coverage for reliable epoch-level comparison, while preserving a longitudinal cohort with multiple nights per participant. These thresholds were chosen as pragmatic quality-control criteria and were not optimized to maximize downstream model performance. After filtering, the retained cohort comprised 128 participants and 1,147 nights across all four diagnostic groups, with a mean compliance score of 95% and balanced ASD versus TD representation (Table K.1). Cross-device validation showed that filtering substantially improved agreement between DREEM- and Empatica-derived sleep timing metrics: TST agreement increased from $r = 0.43$ to $r = 0.73$ by Pearson correlation and from $\rho = 0.46$ to $\rho = 0.73$ by Spearman correlation. Sleep onset and final awakening were also well aligned between the DREEM headband and Empatica wristband after filtering (Figure H.1). These results indicate that, after compliance-aware quality control, downstream analyses of sleep duration and timing can treat either device as a reliable reference, while also highlighting the importance of explicit compliance analysis in scalable real-world data acquisition.

**EEG captures sleep architecture that wrist actigraphy cannot resolve.** After compliance-aware filtering, the retained EEG data still preserve sleep-architecture information that the wristband cannot directly access. DREEM EEG differentiates N1, N2, N3, and REM, whereas the Empatica wristband natively provides only proprietary Wake/Sleep labels; the WASO estimates reported here are derived from those labels. Relative to EEG, Empatica overestimates WASO by 16.7 min and sleep fragmentation by 0.74 transitions/hr (Supplementary H). As shown in Figure 2(A), the wristband sometimes reports extended wake periods that DREEM resolves as brief arousals within sustained sleep. This is consistent with known limitations of actigraphy for resolving within-sleep wake boundaries in ASD cohorts (Yavuz-Kodat et al., 2019), and poor WASO agreement was also reported in the

initial SSP analysis (Hacohen et al., 2026). Notably, whereas prior actigraphy validations often found WASO underestimation due to quiet wakefulness being classified as sleep (Paquet et al., 2007), Empatica shows the opposite pattern here, likely inflating disruption by classifying brief motor restlessness as wake. These discrepancies motivate the two modeling analyses that follow: adapting PSG-trained sleep-staging models to dry-electrode wearable EEG (Section 5) and evaluating whether wrist sensors alone can support staging when headband EEG is impractical (Section 6).

## 5. PSG-based sleep models do not generalize to wearable EEG signals

Sleep staging algorithms are overwhelmingly developed on gold-standard PSG, which uses conventional clinical electrodes applied with conductive paste or gel in a full 10–20 montage with mastoid references and dedicated EOG and EMG channels, yielding signals typically in the $\pm$100–200 $\mu$V range (Berry et al., 2017; Keenan and Hirshkowitz, 2011). By contrast, the DREEM headband records 7-channel EEG from a reduced bipolar dry-electrode montage (Figure 1(B)). In our data, the raw DREEM signals also contain large clipped regions at $\pm$50,000 $\mu$V (Figure L.1(A)), likely reflecting contact instability and amplifier saturation associated with dry-electrode recording rather than physiological EEG amplitude. Together, these differences in sensor technology, montage, and signal characteristics create a substantial domain mismatch between PSG and wearable EEG. We therefore first characterize whether standard PSG-oriented preprocessing remains usable on DREEM recordings, and then assess whether the resulting domain mismatch can be mitigated with machine learning.

**Label provenance and training splits.** Because expert-scored PSG labels are not available for the SSP recordings, we use the DREEM headband's automatic sleep-stage annotations as a wearable-EEG reference throughout this work. Accordingly, our modeling experiments should be interpreted as predicting or approximating DREEM-derived sleep architecture, rather than as estimating absolute PSG ground truth. This distinction is especially important for the wrist-sensor experiments in the sequel, where models are trained to recover EEG-derived labels from a more scalable peripheral sensing configuration. All supervised experiments in this work use participant-level splits: nights from the same participant are assigned to only one of train, validation, or test folds. This prevents models from exploiting participant-specific sleep patterns or device artifacts across repeated nights. All reported fine-tuned metrics, including those in Section 5 and Section 6, are computed on held-out participant-level test folds under 5-fold cross-validation and never on training or validation data.

**Spectral artifact rejection indicates that most DREEM EEG remains usable.** To characterize the PSG-to-wearable mismatch at the signal-processing level, we applied LUNA's spectral artifact rejection procedure (Purcell, 2024), based on the method of Buckelmüller et al. (2006). This procedure relies on frequency-domain features and is therefore better suited to the atypical amplitude distributions observed in dry-electrode wearable EEG. Using this approach, mean artifact rates remained below 15% across all four diagnostic groups, with no substantial between-group differences (Figure L.1(B)). Detected artifacts were typically sparse and transient rather than clustered in sustained blocks, suggesting brief contact disruptions or electrode adjustment events rather than prolonged off-head

non-compliance. These results indicate that, despite clipped segments and dry-electrode artifacts, the DREEM EEG signal remains largely usable for downstream modeling after appropriate artifact handling. We also attempted to process the DREEM signals using HAPPE (Gabard-Durnam et al., 2018; Lopez et al., 2022), a widely used automated pipeline for EEG analysis in developmental and autism research. However, HAPPE's wavelet-based method did not produce artifact-free waveforms in this setting, likely because severe signal clipping (Figure L.1(A)) and the reduced bipolar montage violate assumptions inherited from standard PSG-style EEG preprocessing (further details are provided in Appendix L.1).

**Pretrained sleep staging models perform poorly on DREEM data.** We evaluated two widely used PSG-trained sleep staging models, YASA (Vallat and Walker, 2021) and LUNA (Purcell, 2024), together with the recently introduced sleep foundation model SleepFM (Thapa et al., 2026). YASA and LUNA classify 30 second epochs into sleep stages using engineered spectral and temporal EEG features, including band power, entropy, and Hjorth parameters, followed by gradient-boosted tree classifiers, although the specific feature sets differ between the two methods. SleepFM, by contrast, is a multimodal foundation model pre-trained on large-scale laboratory PSG recordings.

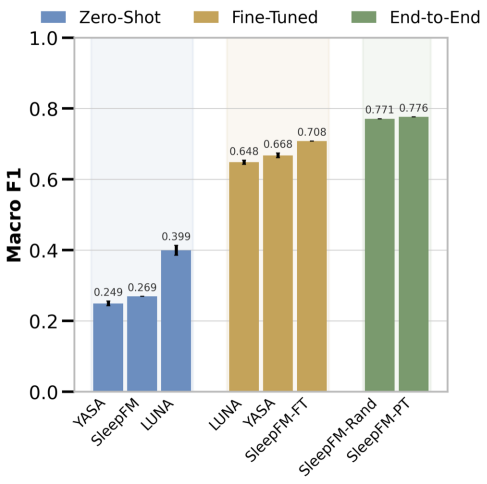



Figure 5: **Fine-tuning bridges the lab-to-wearable domain shift.** Macro F1 improves substantially after adapting PSG-trained sleep-staging models to DREEM EEG.

Zero-shot performance was poor across all models (Figure 5). On the standard 5-class staging task (Wake, N1, N2, N3, REM), all models scored below 0.400 macro F1: LUNA performed best at 0.399, followed by SleepFM at 0.269 and YASA at 0.249. Confusion matrices revealed systematic error patterns rather than random misclassification (Figure N.1). YASA tended to over-predict Wake, whereas LUNA was biased toward N3, consistent with prior observations that the two models show weak agreement in epoch-level predictions and may therefore fail in different ways (Luna, 2026). For YASA, many errors were concentrated in fine-grained sub-stage confusions, while the coarser wake-versus-sleep distinction was largely preserved. Consequently, summary metrics such as total sleep time (TST) and wake after sleep onset (WASO) remained reasonably well estimated, as also noted by Hacohen et al. (2026). By contrast, LUNA and SleepFM performed poorly even on the binary wake-versus-sleep distinction.

**Fine-tuning recovers performance.** To test whether poor zero-shot performance reflects a fundamental signal deficit or a correctable domain shift, we adapted each model to DREEM recordings using the headband's built-in staging labels. For SleepFM, we first considered a shallow adaptation setting, *Classifier Fine-tuned (SleepFM-FT)*, in which the pre-trained encoder and temporal encoder are frozen and only the downstream classifier is updated. Adaptation substantially improved performance (Figure 5): on the 5-class task,

LUNA reached 0.648 macro F1, YASA reached 0.668, and SleepFM-FT reached 0.708 (Table N.1). Similar gains were observed for both 4-class (Wake, Light, Deep, REM) and 6-class (Wake, N1, N2, N3, REM, Artifact) staging tasks (Figure N.2).

We next asked whether the remaining gap reflected limitations of shallow adaptation or of the signal itself. To address this, we evaluated two stronger SleepFM settings in addition to SleepFM-FT (Supplementary O). In *End-to-End Pretrained (SleepFM-PT)*, the encoder, temporal encoder, and classifier were jointly fine-tuned, allowing the representation itself to adapt to the DREEM domain. In *End-to-End Randomized (SleepFM-Rand)*, the same architecture was trained from random initialization using SSP data only. End-to-end adaptation achieved the best results: SleepFM-PT reached 0.776, while SleepFM-Rand reached 0.771. The small difference between these settings suggests that PSG pre-training provides only modest benefit here relative to training directly on in-domain DREEM data. Two observations reinforce this point. Handcrafted feature-based LUNA outperforms the pre-trained SleepFM zero-shot (0.399 vs. 0.269 macro F1), and end-to-end fine-tuning from random initialization (0.771) nearly matches fine-tuning from pre-trained weights (0.776). Together these indicate that a better in-domain representation exists and that the recovered performance stems from in-domain adaptation rather than from the pre-trained prior.

Despite these gains, N1 remained the most difficult stage, with recall of approximately 42–55% (Figure N.1), consistent with the known difficulty of N1 discrimination even in laboratory PSG settings (Vallat and Walker, 2021). Through sample size ablation, we observed that performance saturated quickly, with only 30–50 participants required to approach the final accuracy plateau (Figure N.3). Because sleep stages are intrinsically imbalanced and this imbalance shifts across adolescence (Table E.1), we report macro F1, the mean of per-class F1 scores, which weights every stage equally and is therefore not dominated by the majority N2 class. To avoid overfitting given the modest cohort, the foundation models are fine-tuned rather than trained from scratch, and the sample-size ablations show that 30–50 participants already suffice, well below the full training pool. These results indicate that DREEM recordings contain sufficient information for sleep staging and that the PSG-to-wearable gap is substantial and is learnable. In addition, Supplementary P shows that prefrontal channels contribute relatively little discriminative information compared with frontal-to-occipital derivations, suggesting that a more comfortable headband with fewer, strategically selected channels may achieve similar performance.

## 6. DREEM-referenced sleep staging from wrist-worn sensors

Wrist-worn sensors offer a potential alternative to headband EEG for sleep staging in populations with limited headband tolerance. Although such devices are becoming increasingly common in consumer settings, including the Apple Watch, Fitbit, and Oura, their use for sleep staging in ASD and other vulnerable populations remains relatively understudied. To the best of our knowledge, the SSP dataset is the largest public ambulatory corpus with aligned EEG and PPG recordings, enabling direct comparison between EEG-derived and peripheral-sensor-based staging. It also provides a unique opportunity to train models that predict EEG-level sleep stages from wrist-worn physiological signals under real-world conditions. If such sensors can approximate EEG-based staging, they may offer a practical

path to retaining participants who cannot tolerate the headband, thereby mitigating the compliance-related loss described in Section 3.

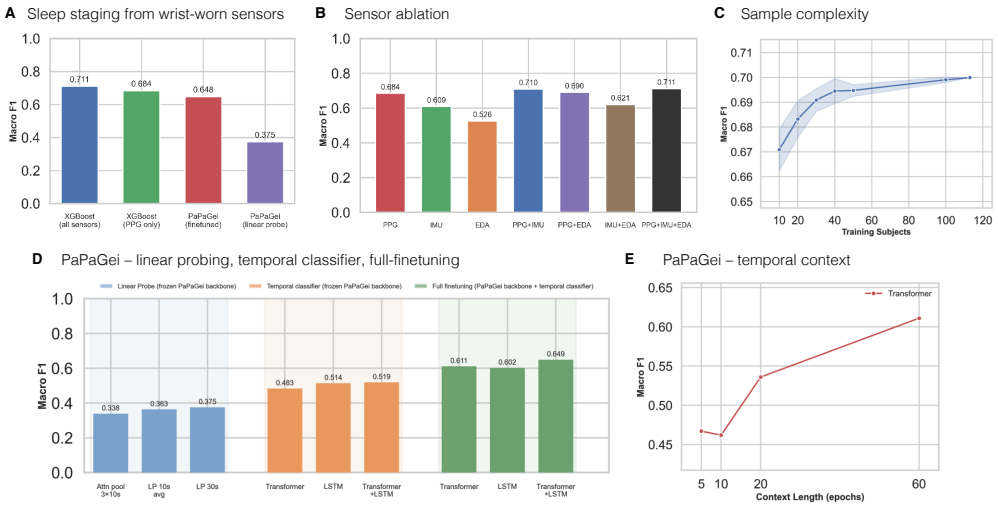



Figure 6: **Wrist-worn sensors support DREEM-referenced sleep staging, but PPG foundation models require domain adaptation and temporal context.** (A) Wrist-worn sensors achieve strong DREEM-referenced performance, with PPG-based models approaching full-sensor performance; however, PPG foundation models do not generalize off-the-shelf and, even after fine-tuning, struggle to match hand-crafted features derived from Empatica signals, highlighting the need for careful PPG processing. (B) Ablation shows PPG as the most informative sensor, with PPG+IMU close to optimal. (C) Training on as few as 10–20 participants yields reasonable performance, with variance decreasing as data increases. (D) Linear probing on PaPaGei performs poorly, suggesting domain shift between gold-standard transmission PPG (on which PaPaGei is trained) and wearable PPG, as well as limitations from the short native embedding context; incorporating temporal classifiers improves performance, and full fine-tuning with temporal context yields the best results. (E) Increasing context length further boosts performance under full fine-tuning, emphasizing the importance of long-range temporal structure for sleep staging.

**PPG is the most informative wrist-worn signal for sleep staging.** Identifying which wrist-worn sensors carry the most sleep-stage information is important for designing recording configurations that are minimal, affordable, and well tolerated. Using the filtered cohort described in Section 4, we trained XGBoost classifiers to predict DREEM-derived sleep stages from wrist-based features (Figure Q.1) and performed a systematic sensor ablation. Specifically, we trained models using PPG alone, IMU alone, EDA alone, and all pairwise and three-way combinations. Among individual modalities, PPG was the strongest standalone signal, achieving a macro F1 of 0.683 and outperforming even the combined IMU+EDA configuration (0.620; Figure 6(A,B)). This distinction was especially clear for REM sleep, which was captured more accurately by PPG than by IMU, whereas IMU contributed chiefly to sleep-wake discrimination. Combining PPG with IMU yielded the best two-sensor configuration (0.708), and adding EDA provided only a marginal additional

gain (0.710). These results indicate that cardiovascular and motion signals capture most of the DREEM-referenced sleep-stage information available from the wrist, and that EDA can be omitted with little loss in performance (full metrics in Table Q.1).

Wrist-based models also exhibited rapid sample-efficiency saturation, with performance plateauing at around 40 participants (Figure 6(C)), suggesting that modest sample sizes are sufficient to learn effective feature-based staging models in this setting. We further evaluated variants of the PPG foundation model PaPaGei (Pillai et al., 2025) (Figure Q.2) to test whether large-scale PPG pretraining transfers to wrist-worn reflective PPG. Across linear probing, temporal classification, and full fine-tuning, PaPaGei underperformed the feature-based XGBoost baseline: frozen linear probing reached 0.375, while full fine-tuning with a temporal classifier reached 0.647, compared with 0.710 for XGBoost (Figure 6(A,D)). Longer temporal context improved fine-tuned performance (Figure 6(E)), indicating the importance of long-range structure for PPG-based sleep staging. However, stage-dependent degradation remained (Figure Q.3), suggesting that off-the-shelf PPG foundation models remain difficult to deploy for clinically meaningful sleep characterization without acquisition-specific adaptation. To test whether such in-domain adaptation actually generalizes, we evaluated cross-cohort transfer to DREAMT (Wang et al., 2024), a second public wearable-PPG dataset with expert-scored PSG. In-domain performance did not translate into reliable external transfer. An SSP fine-tuned PaPaGei transferred to DREAMT with a Cohen's $\kappa$ of only 0.069, collapsing onto deep sleep (assigning 63.5% of epochs to deep sleep against a true rate of 3.4%). This indicates that fine-tuning on one wearable cohort does not, by itself, solve the domain-shift problem (Supplementary R). This comparison should be interpreted as a deployment-pipeline comparison rather than a pure architecture comparison. The XGBoost model uses one-minute Empatica summaries, including heart rate, heart-rate variability, motion, temperature, and related features derived from proprietary preprocessing. In contrast, PaPaGei operates on raw PPG waveforms, where denoising, beat detection, artifact rejection, and temporal aggregation must be learned by the model or handled separately. Thus, the stronger feature-based performance likely reflects both the information content of wrist-worn physiological signals and the importance of robust preprocessing for noisy reflective PPG. Full PaPaGei analyses are provided in Appendix Q.

## 7. Discussion

Ambulatory wearable sensing promises scalable, longitudinal measurement of sleep physiology outside the clinic. However, our findings show that the central challenges of real-world deployment differ from those emphasized in laboratory-based benchmarking, where the focus is often on dataset scale and marginal gains in model accuracy. A summary of our findings is provided in Supplementary Table C.1.

Compliance is a major factor to consider in designing and analyzing ambulatory studies with wearables. In our study, DREEM headband compliance declined to approximately 77% before the final wake period, whereas the Empatica wristband remained near 100%. Importantly, this data loss was not random, but systematically associated with behavioral measures, meaning that participants for whom scalable monitoring may be most clinically important may also be more likely to lose EEG data. To support more reliable downstream analyses, we developed a compliance-aware filtering pipeline that retained an enriched co-

hort of 128 of 200 participants. These lessons extend beyond sleep and ASD: any ambulatory digital health study that deploys sensors with differing tolerability should expect and quantify compliance-trait confounding before building predictive models. Ignoring this issue risks developing tools that perform best in the populations that need them least.

This structured missingness has concrete implications for downstream modeling, and its effect depends on the type of task. Because distinguishing missing-at-random (MAR) from missing-not-at-random (MNAR) mechanisms is difficult in ambulatory neurodevelopmental data, the safe default is to treat the missingness as informative. *Local* tasks, such as per-epoch sleep staging, operate within retained segments and depend only on a short surrounding window, so they are comparatively robust to progressive dropout. *Global* tasks, such as night-level aggregates and cross-group comparisons, are far more vulnerable, because missingness skews the aggregate before any comparison is made. Since compliance declines steadily across the night, this bias is also non-uniform. Early-night measures such as sleep-onset latency are relatively protected, whereas late-night and within-sleep measures such as WASO and late-night REM are more exposed. This mirrors the original SSP analysis (Hacohen et al., 2026), in which the objective measure separating autistic children from siblings was the early-night sleep-onset latency, while objective WASO showed no reliable group difference. Respecting this structure motivated our decision to predict DREEM-derived stages from the higher-adherence wrist signals (Section 6) rather than impute missing EEG or restrict phenotyping to the compromised modality.

A consistent finding across both EEG and PPG experiments is that physiological foundation models pre-trained on laboratory or clinical-grade data show limited off-the-shelf transfer to wearable-acquired data of the same nominal modality. Using macro F1 as the evaluation metric, SleepFM, pre-trained on over 100,000 hours of PSG, scored below 0.270 on DREEM EEG, while PaPaGei, pre-trained on clinical transmission PPG, reached only 0.375 on wrist reflective PPG. These failures arise from structured domain gaps, including differences in electrode technology, montage configuration, and signal conditioning, rather than from model architecture limitations alone. Fine-tuning recovered substantial performance in both cases. End-to-end adaptation of SleepFM on the full cohort reached 0.776, while supervised wrist-worn models using PPG- and IMU-derived features reached 0.710; the EEG domain shift was bridgeable with as few as 30–50 participants. These results suggest that the effective deployment domain of physiological foundation models is strongly conditioned on acquisition regime, including sensor type, montage, signal conditioning, and population. Cross-device evaluation should therefore accompany deployment on hardware different from that used during pre-training. Although device- and domain-specific adaptation is possible, as demonstrated here, this requirement may limit the deployment of current foundation models in real-world clinics and studies that lack the technical infrastructure or expertise to adapt models to their own acquisition settings. More fundamentally, the core foundation-model promise of label-efficient, broad generalization is not yet met in this setting. A randomly initialized SleepFM nearly matches its pre-trained counterpart given the same labels, and an SSP fine-tuned PaPaGei fails to transfer to the external DREAMT cohort. We therefore position SSP, together with cross-cohort evaluations such as DREAMT, as a benchmark for testing whether physiological foundation models genuinely generalize to pediatric, wearable, and clinically relevant sleep data, rather than only after within-cohort fine-tuning.

Combining the compliance and modeling results yields a practical modality hierarchy for ambulatory phenotyping. For macro-sleep measures that are robustly captured by peripheral sensing, such as total sleep time and broad sleep-wake timing, wrist-worn PPG+IMU may be preferable to EEG in large longitudinal studies because it avoids compliance-driven bias. However, wrist-worn wearables remain less reliable than EEG for estimating wake after sleep onset (WASO), an important metric in its own right and a contributor to derived measures such as TST. Improving WASO estimation in compliant peripheral modalities is therefore an important direction for future work. For micro-architectural phenotyping (spindle density and stage-specific spectral features), EEG remains necessary but researchers and clinicians must accept a biased subsample, and bias due to compliance issues must be accounted for. A practical deployment strategy would therefore use PPG+IMU as the backbone for longitudinal monitoring across all participants when using current technology, reserving headband EEG for targeted sleep analysis in participants who tolerate it. More broadly, translating staging models into clinical tools will require robustness to modality dropout, traceable label provenance, and evidence that improved staging accuracy leads to improved clinical outcomes.

## 8. Limitations

Several limitations should be considered when interpreting these results. First, all EEG staging labels come from the DREEM headband's built-in algorithm, which has been validated against expert-scored PSG in adults aged 22–65 (Arnal et al., 2020), but not in pediatric populations. All staging results should therefore be read as DREEM-referenced rather than as expert-PSG ground truth; we partially mitigate this by validating DREEM labels against normative pediatric stage proportions and developmental trends (Supplementary E), but direct expert-PSG validation in this cohort remains unavailable. All results are specific to the DREEM 3S headband and Empatica Embrace Plus wristband. Different dry-electrode EEG devices or PPG wristbands may exhibit different domain shifts, compliance patterns, and signal quality profiles. Our findings therefore emphasize the need for acquisition-specific validation and adaptation before deploying pretrained physiological models on new wearable systems. Finally, the compliance-trait associations reported here are correlational and cross-sectional. It is possible that compliance improves with habituation over the 14-night recording protocol, and that the observed behavioral associations reflect first-night effects rather than stable trait-level predictors. Although we additionally verified that the associations between compliance and severity persist after adjusting for age and sex (Supplementary G), the behavioral, diagnostic, age, and sex variables remain partially collinear in this cohort, so residual confounding cannot be excluded and these associations should not be interpreted causally.

**Data and Code Availability** This study uses the Simons Sleep Project (SSP) dataset (Hacohen et al., 2026), a multimodal ambulatory sleep dataset available to qualified researchers through SFARI Base. All code, hyperparameters, environment specifications, and replication details used to produce our analyses and figures are available at https://github.com/sankethvedula/sleep-mlhc2026.

## Acknowledgments

This work was partially supported by ONR, Simons Foundation, NSF, and gifts from Apple and Google. This investigation was supported in part by the Simons Foundation grant AR-Director-00007393 awarded to O.G.T, and by a grant that supported S.V. from the Schwab Charitable Fund made possible by the generosity of Eric and Wendy Schmidt. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding and gifting agencies and companies

## References

Apple Inc. Estimating sleep stages from Apple Watch. https://www.apple.com/health/pdf/Estimating_Sleep_Stages_from_Apple_Watch_Oct_2025.pdf, October 2025. (Cited on pages 2 and 22.)

Pierrick J. Arnal, Valentin Thorey, Eden Debellemaniere, Michael E. Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, Mounir Chennaoui, and Fabien Sauvet. The Dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, 43(11), 2020. doi: 10.1093/sleep/zsaa097. (Cited on pages 5, 16, 22, 24, and 27.)

Beacon Biosignals. DREEM 3S. https://beacon.bio/dreem-3s, 2026. Accessed April 2026. (Cited on pages 2 and 5.)

Richard B. Berry, Rita Brooks, Charlene Gamaldo, Susan M. Harding, Robin M. Lloyd, Stuart F. Quan, Matthew T. Troester, and Bradley V. Vaughn. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, version 2.4 edition, 2017. (Cited on pages 10 and 22.)

Vera Birrer, Mohamed Elgendi, Olivier Lambercy, and Carlo Menon. Evaluating reliability in wearable devices for sleep staging. *npj Digital Medicine*, 7:74, 2024. doi: 10.1038/s41746-024-01016-9. (Cited on page 22.)

Jeannine Buckelmüller, Hans-Peter Landolt, Hans H. Stassen, and Peter Achermann. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience*, 138 (1):351–356, 2006. doi: 10.1016/j.neuroscience.2005.11.005. (Cited on page 10.)

Ashura W. Buckley, Ashanti J. Rodriguez, Kara Jennison, Jennifer Buckley, Audrey Thurm, Sachio Sato, and Susan Swedo. Rapid eye movement sleep percentage in children with autism compared with children with developmental delay and typical development. *Archives of Pediatrics & Adolescent Medicine*, 164(11):1032–1037, 2010. doi: 10.1001/archpediatrics.2010.202. (Cited on page 24.)

Kimberly L. H. Carpenter, Jordan Hashemi, Kathleen Campbell, Steven J. Lippmann, Jeffrey P. Baker, Helen L. Egger, Steven Espinosa, Saritha Vermeer, Guillermo Sapiro, and Geraldine Dawson. Digital behavioral phenotyping detects atypical pattern of facial expression in toddlers with autism. *Autism Research*, 14(3):488–499, 2021. doi: 10.1002/aur.2391. (Cited on page 23.)

Mary A. Carskadon and William C. Dement. Normal human sleep: an overview. In Meir H. Kryger, Thomas Roth, and William C. Dement, editors, *Principles and Practice of Sleep Medicine*, pages 16–26. Elsevier Saunders, 5th edition, 2011. (Cited on pages 6 and 24.)

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6): 877–888, 2015. doi: 10.5665/sleep.4732. (Cited on pages 22 and 47.)

Loris Constantin, Christian M. Horvath, Florent Baty, Clémentine Aguet, Jérôme Van Zaen, Alia Lemkaddem, Loïc Jeanningros, Martin Proença, Xiaoli Yang, Kurt De Jaegere, Sebastian R. Ott, João Jorge, Jean-Philippe Thiran, Theo A. Meister, Rodrigo Soria, Hildegard Tanner, Emrush Rexhaj, Mathieu Lemay, Anne-Kathrin Brill, and Fabian Braun. Towards long-term sleep staging via wearable reflective photoplethysmography. *Sleep*, 49(3), 2026. doi: 10.1093/sleep/zsaf246. (Cited on page 22.)

Massimiliano de Zambotti, Cathy Goldstein, Jesse Cook, Luca Menghini, Marco Altini, Philip Cheng, and Rebecca Robillard. State of the science and recommendations for using wearable technology in sleep and circadian research. *Sleep*, 47, 2024. doi: 10.1093/sleep/zsad325. (Cited on page 22.)

Preeti A. Devnani and Anaita U. Hegde. Autism and sleep disorders. *Journal of Pediatric Neurosciences*, 10(4):304–307, 2015. doi: 10.4103/1817-1745.174438. (Cited on pages 2 and 22.)

Empatica. Empatica Embrace Plus. https://www.empatica.com/embraceplus/, 2026. Accessed April 2026. (Cited on pages 2 and 5.)

Pedro Fonseca, Tim Weysen, Maaike S. Goelema, Els I. S. Møst, Mustafa Radha, Charlotte Lunsingh Scheurleer, Leonie van den Heuvel, and Ronald M. Aarts. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*, 40(7):zsx097, 2017. doi: 10.1093/sleep/zsx097. (Cited on page 22.)

Laurel J. Gabard-Durnam, Adriana S. Mendez Leal, Carol L. Wilkinson, and April R. Levin. The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, 12:97, 2018. doi: 10.3389/fnins.2018.00097. (Cited on pages 11, 25, 37, and 38.)

Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W. Picard, and Simone Tognetti. Empatica E3: a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *4th International Conference on Wireless Mobile Communication and Healthcare (MobiHealth)*, 2014. doi: 10.4108/icst.mobihealth.2014.257418. (Cited on pages 5, 24, and 27.)

Laura Ghirardi, Isabell Brikell, Ralf Kuja-Halkola, Christine M. Freitag, Barbara Franke, Philip Asherson, Paul Lichtenstein, and Henrik Larsson. The familial co-aggregation of

ASD and ADHD: a register-based cohort study. *Molecular Psychiatry*, 23(2):257–262, 2018. doi: 10.1038/mp.2017.17. (Cited on pages 5 and 27.)

Márton Á. Goda, Peter H. Charlton, and Joachim A. Behar. pyPPG: a Python toolbox for comprehensive photoplethysmography signal analysis. *Physiological Measurement*, 45(4): 045001, 2024. doi: 10.1088/1361-6579/ad33a2. (Cited on page 48.)

Micha Hacohen, Adam Levy, Hadas Kaiser, LeeAnne Snyder, Alpha Amatya, Brigitta B. Gundersen, John E. Spiro, and Ilan Dinstein. An open science resource for accelerating scalable digital health research in autism and other neurodevelopmental conditions. *Nature Neuroscience*, 29:467–478, 2026. doi: 10.1038/s41593-025-02146-3. (Cited on pages 2, 5, 10, 11, 15, 16, 22, and 27.)

Daniel Hodge, Ashley M. N. Parnell, Charles D. Hoffman, and Dwight P. Sweeney. Methods for assessing sleep in children with autism spectrum disorders: a review. *Research in Autism Spectrum Disorders*, 6(4):1337–1344, 2012. doi: 10.1016/j.rasd.2012.05.009. (Cited on page 2.)

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. (Cited on pages 22 and 47.)

Sharon A. Keenan and Max Hirshkowitz. EEG recording and signal processing. In Meir H. Kryger, Thomas Roth, and William C. Dement, editors, *Principles and Practice of Sleep Medicine*, pages 1602–1609. Elsevier Saunders, 5th edition, 2011. (Cited on page 10.)

Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar. SleepPPG-Net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 27(2):924–932, 2022. doi: 10.1109/JBHI.2022.3225363. (Cited on page 22.)

Harun Onder Kucukosmanoglu, Yucel Cimtay, Gulay Tasci Yilmaz, and Cem Ersoy. Detection of sleep disorders using wearable devices and machine learning. *Scientific Reports*, 15:44089, 2025. doi: 10.1038/s41598-025-27739-7. (Cited on pages 2 and 22.)

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022. doi: 10.1038/s41597-022-01411-5. (Cited on pages 22 and 47.)

K. L. Lopez, A. D. Monachino, S. Morales, S. C. Leach, M. E. Bowers, and L. J. Gabard-Durnam. HAPPILEE: HAPPE in low electrode electroencephalography, a standardized pre-processing software for lower density recordings. *NeuroImage*, 260:119390, 2022. doi: 10.1016/j.neuroimage.2022.119390. (Cited on pages 11 and 38.)

Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism Diagnostic Interview–Revised: a revised version of a diagnostic interview for caregivers of individuals with

possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5):659–685, 1994. doi: 10.1007/BF02172145. (Cited on page 5.)

Luna. SOAP/POPS staging. https://zzz.nyspi.org/luna/vignettes/soap-pops/, 2026. Luna documentation, accessed April 2026. (Cited on page 11.)

Kuldeep Mahato, Tamoghna Saha, Shichao Ding, Samar S. Sandhu, An-Yi Chang, and Joseph Wang. A hybrid system for continuous multi-parameter monitoring of vital signs using transdermal optical sensing. *Nature Electronics*, 7:735–750, 2024. doi: 10.1038/s41928-024-01247-4. (Cited on pages 2 and 22.)

Silvia Miano, Oliviero Bruni, Maurizio Elia, Antonella Trovato, Arianna Smerieri, Elisabetta Verrillo, Michele Roccella, Mario Giovanni Terzano, and Raffaele Ferri. Sleep in children with autistic spectrum disorder: a questionnaire and polysomnographic study. *Sleep Medicine*, 9(1):64–70, 2007. doi: 10.1016/j.sleep.2007.01.014. (Cited on pages 2 and 22.)

Andrew C. Miller, Joseph Futoma, Salar Abbaspourazad, Christina Heinze-Deml, Saba Emrani, Ian Shapiro, and Guillermo Sapiro. A wearable-based aging clock associates with disease and behavior. *Nature Communications*, 16:9264, 2025. doi: 10.1038/s41467-025-64275-4. (Cited on page 23.)

Dean J. Miller, Michele Lastella, Aaron T. Scanlan, Clint Bellenger, Shona L. Halson, Gregory D. Roach, and Charli Sargent. A validation study of the WHOOP strap against polysomnography to assess sleep. *Journal of Sports Sciences*, 38(22):2631–2636, 2020. doi: 10.1080/02640414.2020.1797448. (Cited on page 2.)

Achille Nazaret and Guillermo Sapiro. A large-scale observational study of the causal effects of a behavioral health nudge. *Science Advances*, 9(40), 2023. doi: 10.1126/sciadv.adi1752. (Cited on page 23.)

Jean Paquet, Anna Kawinska, and Julie Carrier. Wake detection capacity of actigraphy during sleep. *Sleep*, 30(10):1362–1369, 2007. doi: 10.1093/sleep/30.10.1362. (Cited on pages 10 and 22.)

Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michael Aupetit, Juan M. Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital Medicine*, 3(1):42, 2020. doi: 10.1038/s41746-020-0244-4. (Cited on page 2.)

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, April 2025. doi: 10.48550/arXiv.2410.20542. (Cited on pages 3, 14, 22, 26, 47, and 50.)

Shaun M Purcell. 0321 Luna: Merging Open-source Tools, Data and Models for Sleep Signal Analyses. *Sleep*, 47(Supplement_1):A138, May 2024. doi: 10.1093/sleep/zsae067.0321. (Cited on pages 3, 10, 11, 22, and 25.)

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digital Medicine*, 4(1):135, 2021. doi: 10.1038/s41746-021-00510-8. (Cited on page 22.)

Jessica Vensel Rundo and Ralph Downey. Polysomnography. In Kerry H. Levin and Patrick Chauvel, editors, *Clinical Neurophysiology: Basis and Technical Aspects*, volume 160 of *Handbook of Clinical Neurology*, chapter 25, pages 381–392. Elsevier, 2019. doi: 10.1016/B978-0-444-64032-1.00025-4. (Cited on page 22.)

Sabine Scholle, Ulrike Beyer, Michael Bernhard, Simone Eichholz, Thomas Erler, Peter Graness, Barbara Goldmann-Schnalke, Katrin Heisch, Frank Kirchhoff, Kathrin Klementz, Grit Köch, Andreas Kramer, Detlef Schnalke, Alfred Wiater, and Robert Wittmann. Normative values of polysomnographic parameters in childhood and adolescence: Quantitative sleep parameters. *Sleep Medicine*, 12(6):542–549, 2011. doi: 10.1016/j.sleep.2010.11.011. (Cited on pages 5, 26, and 28.)

Torbjorn Svensson, Un-Il Chung, Isao Tokuda, and Akiko Kishi Svensson. A validation study of a consumer wearable sleep tracker compared to a portable EEG system in naturalistic conditions. *Sleep Medicine*, 115:251–263, 2024. doi: 10.1016/j.sleep.2024.01.020. (Cited on page 2.)

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore Iv, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2):752–762, 2026. doi: 10.1038/s41591-025-04133-4. (Cited on pages 3, 11, 22, 25, and 44.)

Raphael Vallat and Matthew P. Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10, 2021. doi: 10.7554/eLife.70092. (Cited on pages 3, 11, 12, 22, 25, and 41.)

Kaiwen Wang, Yifan Yang, Aritra Saha, et al. DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology. *Proceedings of Machine Learning Research (CHIL)*, 2024. Wearable PPG with expert-scored PSG; also available on PhysioNet. (Cited on pages 4, 14, 26, and 53.)

E. Yavuz-Kodat, E. Reynaud, M.-M. Geoffray, N. Limousin, P. Franco, P. Bourgin, and C. M. Schroder. Validity of actigraphy compared to polysomnography for sleep assessment in children with autism spectrum disorder. *Frontiers in Psychiatry*, 10:551, 2019. doi: 10.3389/fpsyt.2019.00551. (Cited on page 9.)

Yan Zhang, Nana Li, Chao Li, Zhaoqing Zhang, Haiying Teng, Yao Wang, Tingting Zhao, Lei Shi, Kun Zhang, Kun Xia, Jun Li, and Zhongshang Sun. Genetic evidence of gender difference in autism spectrum disorder supports the female-protective effect. *Translational Psychiatry*, 10(1):4, 2020. doi: 10.1038/s41398-020-0699-8. (Cited on pages 5 and 27.)

# Supplementary Material

## Contents

## Supplementary A. Related work

Sleep staging, the classification of sleep into discrete physiological states, is a core task in both clinical sleep medicine and sleep research. The American Academy of Sleep Medicine (AASM) standard (Berry et al., 2017) defines five stages (Wake, N1, N2, N3, and REM) scored from polysomnography (PSG), which integrates EEG, electrooculography (EOG), and electromyography (EMG). Because manual scoring is time-consuming and subject to inter-scorer variability (Rundo and Downey, 2019), numerous automated approaches have been proposed, including feature-based machine learning classifiers such as YASA (Vallat and Walker, 2021) and LUNA (Purcell, 2024), to recent deep learning foundation models such as SleepFM (Thapa et al., 2026), which is pre-trained on more than 100,000 hours of PSG. Dry-electrode EEG headbands such as DREEM have shown staging accuracy comparable to inter-scorer agreement under controlled laboratory conditions (Arnal et al., 2020). However, most automated sleep staging methods are developed and validated on laboratory-grade data, leaving open whether they generalize to the distinct signal characteristics and recording conditions of wearable EEG collected outside the lab.

Wrist-worn sensing offers a scalable alternative, with consumer devices typically estimating sleep stages from accelerometry and photoplethysmography (PPG) (Mahato et al., 2024; Kucukosmanoglu et al., 2025; Apple Inc., 2025). Earlier work has shown that sleep staging from transmission PPG is feasible in clinical or otherwise controlled settings (Kotzen et al., 2022; Fonseca et al., 2017; Radha et al., 2021). At the same time, although consumer devices can reliably distinguish sleep from wake, their ability to resolve individual NREM and REM stages, which are critical for characterizing sleep quality, often remains limited relative to EEG-based approaches (Paquet et al., 2007; de Zambotti et al., 2024; Birrer et al., 2024). More recently, Constantin et al. (2026) showed that reflective PPG can support sleep staging in naturalistic home environments, although validation there was performed against actigraphy rather than EEG-based staging. Our setting differs in two important respects: we study wrist-worn reflective PPG under real-world home conditions, and we validate against dry-electrode EEG recorded from the DREEM headband. This provides a rare evaluation of wrist-worn reflective PPG sleep staging against concurrently recorded wearable EEG in a pediatric cohort. In parallel, recent foundation models such as PaPaGei (Pillai et al., 2025) have been trained on large clinical-grade PPG datasets, including MIMIC-III (Johnson et al., 2016), VitalDB (Lee et al., 2022), and MESA (Chen et al., 2015), with the goal of learning general-purpose physiological representations. In this work, we further assess the transfer of such representations to wrist-worn reflective PPG in this setting.

Prior ASD sleep studies (Miano et al., 2007; Devnani and Hegde, 2015) have relied primarily on parent-reported questionnaires or single-night laboratory PSG, both of which have important limitations. Questionnaires are subjective and coarse, whereas single-night PSG cannot capture night-to-night variability and may itself alter sleep through the first-night effect and the unfamiliar recording environment. The SSP dataset (Hacohen et al., 2026), which we use here, was designed to address this gap through multi-night home recordings that combine EEG with wrist-worn sensors in a large ASD cohort. More broadly, this work lies at the intersection of digital phenotyping and wearable-derived biomarkers, both of which have shown promise in ASD as complements to traditional clinical assessment.

For example, computer vision applied to home videos can identify behavioral markers of ASD (Carpenter et al., 2021), and mobile and wearable sensors can capture clinically meaningful behavioral signatures in everyday environments (Nazaret and Sapiro, 2023). Beyond behavior, wrist-worn PPG features have been shown to predict biological age and cardiovascular risk (Miller et al., 2025), underscoring that peripheral signals captured by consumer wearables can reflect broader aspects of systemic health. Within this broader context, the present study examines sleep as a target for wearable phenotyping in ASD, focusing on which modalities support staging, whether pre-trained models transfer across acquisition settings, and how compliance shapes both.

## Supplementary B. Glossary of Acronyms

Table B.1: **Glossary of acronyms.** Abbreviations used throughout the paper.

| Acronym | Definition |
|---|---|
| AASM | American Academy of Sleep Medicine |
| ABC | Aberrant Behavior Checklist |
| ADHD | Attention Deficit/Hyperactivity Disorder |
| ASD | Autism Spectrum Disorder |
| BVP | Blood Volume Pulse |
| CBCL | Child Behavior Checklist |
| CSHQ | Children's Sleep Habits Questionnaire |
| EDA | Electrodermal Activity |
| EEG | Electroencephalography |
| EMA | Exponential Moving Average |
| EMG | Electromyography |
| EOG | Electrooculography |
| FA | Final Awakening |
| FDA | Food and Drug Administration |
| FT | Fine-Tuning |
| HAPPE | Harvard Automated Processing Pipeline for EEG |
| IMU | Inertial Measurement Unit |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| NREM | Non-Rapid Eye Movement (sleep) |
| PPG | Photoplethysmography |
| PSG | Polysomnography |
| RBS-R | Repetitive Behavior Scale–Revised |
| REM | Rapid Eye Movement (sleep) |
| SCQ | Social Communication Questionnaire |
| SE | Sleep Efficiency |
| SO | Sleep Onset |
| SRS-2 | Social Responsiveness Scale, Second Edition |
| SSP | Simons Sleep Project |

Table B.1 (continued)

| Acronym | Definition |
|---|---|
| TD | Typically Developing |
| TIB | Time in Bed |
| TST | Total Sleep Time |
| WASO | Wake After Sleep Onset |

## Supplementary C. Summary of Experiments

Table C.1: **Experimental overview.** Summary of key research questions, experimental designs, main takeaways, and supporting evidence.

| Question | Experiment Design | Takeaway | Evidence |
|---|---|---|---|
| **1. Does EEG compliance decline overnight?** | Compute EMA-based compliance score across the normalized sleep window for DREEM (Arnal et al., 2020) ($n$=193) and Empatica (Garbarino et al., 2014) ($n$=196). | DREEM declines to ∼77% before final wake; Empatica remains near 100%. Late-night loss disproportionately underrepresents stages like REM that change as the night progresses (Buckley et al., 2010; Carskadon and Dement, 2011). | Fig. 2(A) |
| **2. Is compliance associated with participant characteristics?** | Correlate per-participant DREEM compliance with age, sex, and diagnostic group via Spearman rank correlation. | Compliance increases with age ($\rho$=0.36, $p$<.001) across most groups but weakly in ASD. No conclusive sex difference; should be investigated again with a larger cohort. | Fig. 2(B,C), Fig. F.1 |
| **3. Are off-head events linked to behavioral traits?** | Correlate off-head fraction at varying removal thresholds with CBCL, ABC, and sensory scores. | Behavioral dysregulation (attention problems, hyperactivity, sensory scores) predicts earlier headband removal. | Fig. 3 |
| **4. Does wearing the EEG headband alter sleep?** | Within-subject comparison of Empatica-derived TST, WASO, SE, and restfulness on DREEM+Empatica vs. Empatica-only nights (≥3 nights per condition). | TST was higher on headband nights ($d_z$=0.39, $p$<.001) and restfulness lower ($d_z$=−0.27, $p$=0.037). WASO and SE showed no difference. Effect sizes small to moderate ($\lvert d_z \rvert \leq 0.39$); limited sample precludes strong conclusions. | Fig. J.1 |
| **5. Can compliance-aware filtering improve data quality?** | Sequential 5-step filtering pipeline based on off-head rate, concurrent Empatica availability, and minimum night count. | Retains 128 participants and 1,147 nights. TST cross-device correlation improves from $r$=0.43 to $r$=0.73 with average compliance score of 95%. | Fig. 4, Table K.1 |

Table C.1 (continued)

| Question | Experiment Design | Takeaway | Evidence |
|---|---|---|---|
| **6. Does wrist actigraphy agree with EEG on within-sleep metrics?** | Compare DREEM and Empatica sleep metrics (TST, SO, FA, WASO, SE, fragmentation) across $n$=1,141 concurrent nights. | Strong agreement on gross timing (TST $r$=0.73, SO $\rho$=0.89). Weak agreement on within-sleep disruption: Empatica overestimates WASO by 16.7 min and fragmentation by 0.74 transitions/hr | Fig. H.1 |
| **7. Do standard PSG pre-processing pipelines work on wearable EEG?** | Apply HAPPE (Gabard-Durnam et al., 2018) wavelet-based pipeline and LUNA (Purcell, 2024) spectral-method artifact detection to DREEM recordings. | HAPPE fails likely due to signal clipping and unfamiliar montage. LUNA artifact detection succeeds; median artifact rates 11–13%, below 15% across all groups. | Fig. L.1 |
| **8. Do PSG-trained staging models generalize to wearable EEG?** | Evaluate YASA (Vallat and Walker, 2021), LUNA (Purcell, 2024), and SleepFM (Thapa et al., 2026) zero-shot on DREEM 5-class staging against built-in DREEM labels. | All models achieve macro F1 below 0.400 zero-shot (YASA 0.249, SleepFM 0.269, LUNA 0.399). Systematic biases: YASA over-predicts Wake; LUNA biased toward N3. After fine-tuning, SleepFM End-to-End Pretrained achieves the best macro F1 of 0.776. Domain gap, not signal deficit. | Fig. 5, Fig. N.1 |
| **9. Can fine-tuning recover EEG staging performance?** | Fine-tune YASA, LUNA, and SleepFM (Classifier FT, End-to-End Pretrain, and End-to-End Random) on DREEM with 5-fold CV. Vary training set size. | Fine-tuned models converge to macro F1 ∼0.710–0.730. SleepFM End-to-End achieves macro F1 of 0.776 (pretrained) and 0.771 (random init). Performance plateaus at 30–50 participants. | Fig. 5, Figs. N.2, N.3 |
| **10. Which EEG channels are most informative for staging?** | Apply fine-tuned YASA to each of 7 DREEM bipolar channels independently on 6-class staging. Evaluate per-channel macro F1 and off-head rate. | Frontal-to-occipital derivations (F7–O1, F7–O2, F8–O1, F8–O2) consistently outperform prefrontal channels. Off-head rates uniformly low across all channels, underpredicted relative to the DREEM headband's own off-head detection. | Fig. P.1 |
| **11. Which wrist-worn sensor is most informative for sleep staging?** | Train XGBoost classifiers on all single, pairwise, and three-way Empatica sensor combinations (PPG, IMU, EDA) to predict DREEM 4-class stages. Vary training set size. | PPG is the strongest standalone signal (macro F1 0.683). PPG+IMU is optimal (macro F1 0.708). Adding EDA yields only marginal improvement ( 0.710). Performance saturates at ∼40 participants. | Fig. 6(A–C), Fig. Q.1 |

Table C.1 (continued)

| Question | Experiment Design | Takeaway | Evidence |
|---|---|---|---|
| **12. Do PPG foundation models transfer to wrist-worn reflective PPG?** | Evaluate PaPaGei (Pillai et al., 2025) with linear probes and end-to-end fine-tuning with temporal classifiers on Empatica reflective PPG. Vary temporal context length. | Linear probing achieves only macro F1 0.375. Full fine-tuning with temporal context recovers to 0.647 but remains below the supervised baseline (macro F1 0.710). Longer context improves performance, confirming that long-range temporal structure matters. Domain mismatch between clinical transmission and wearable reflective PPG. | Fig. 6(D,E), Fig. Q.3 |
| **13. Does raw EEG signal quality decline independently of DREEM's labels?** | DREEM-agnostic probes (amplitude clipping, electrode-jump rate) and four artifact detectors (YASA variance/covariance, LUNA time/spectral) computed per 30 s epoch across the sleep window and by compliance quartile. | All probes worsen across the night and within on-head epochs; lowest-compliance nights carry 3.9× more clipping and 1.7× more jumps, and more LUNA artifact ($\rho$=−0.25). Compliance is a conservative lower bound on signal quality. | Suppl. M |
| **14. Is compliance associated with nuanced behavioral/severity traits?** | Correlate per-participant compliance score with autism core-severity scores (e.g., SRS-2), with and without adjusting for age and sex; relate off-head propensity to compliance. | Compliance tracks severity (SRS $\rho$=0.21, $p$<.01), surviving age/sex adjustment; off-head propensity and compliance are anti-correlated ($\rho\approx-0.997$). | Suppl. G |
| **15. Are DREEM labels reliable in this pediatric ASD cohort?** | Compare DREEM stage proportions against normative pediatric AASM-PSG values (Scholle et al., 2011) and expected age trends. | N2/REM match norms; N1 under- and N3 over-estimated (opposite-signed); the N3→N2 redistribution with age is reproduced. | Suppl. E, Table E.1 |
| **16. Does in-domain fine-tuning generalize across cohorts?** | Cross-cohort train/test transfer between SSP and DREAMT (Wang et al., 2024) (wearable PPG with expert PSG); evaluate the SSP fine-tuned PaPaGei on DREAMT. | In-domain performance does not transfer. SSP fine-tuned PaPaGei reaches $\kappa$=0.069 on DREAMT, collapsing onto deep sleep. Fine-tuning alone does not solve domain shift. | Suppl. R, Table R.1 |

## Supplementary D. SSP Cohort Characteristics and Recording Setup

The SSP dataset (Hacohen et al., 2026) comprises 200 children aged 10–17 years recruited as sibling pairs, one with autism (ASD; $n = 102$, 86M/16F, $13.94 \pm 2.13$ yr) and one either typically developing (TD) or with attention deficit/hyperactivity disorder (ADHD) ($n = 98$, 52M/46F, $13.46 \pm 2.16$ yr), age-matched across groups ($p = 0.120$) but imbalanced in sex ratio consistent with male predominance in ASD (Zhang et al., 2020); 57% of autistic participants carry a co-occurring ADHD diagnosis (Ghirardi et al., 2018), and ASD severity scores (SRS-2: $73.71 \pm 10.01$ vs. $47.89 \pm 7.14$; RBS-R: $24.52 \pm 14.05$ vs. $4.15 \pm 6.76$) separate the two groups (Table D.1). Each participant was recorded over up to 14 nights at home with a Beacon Biosignals DREEM 3S (Arnal et al., 2020) dry-electrode EEG headband and an Empatica Embrace Plus wristband (Garbarino et al., 2014). The DREEM 3S provides 7 bipolar derivations from 5 dry scalp electrodes (Fp1, F7, F8, O1, O2), comprising four frontal-to-occipital channels (F7–O1, F7–O2, F8–O1, F8–O2), two prefrontal channels (Fp1–F7, Fp1–F8), and one cross-hemispheric channel (F8–F7), with AASM-compliant automatic scoring of 30-second epochs; it has FDA 510(k) clearance (K223539) and achieves 83.5% staging accuracy against expert-scored PSG in adults aged 22–65 (Arnal et al., 2020), motivating the in-domain fine-tuning experiments reported in this paper. The Empatica Embrace Plus records 64 Hz photoplethysmography (PPG), 64 Hz 3-axis accelerometry (IMU), and peripheral temperature. The wearing configuration and bipolar montage are shown in Figure 1 of the main text.

Table D.1: Participant characteristics of the SSP dataset (Hacohen et al., 2026). Values are mean ± SD or $n$ (%).

| **Variable** | **Autism** ($n$=102) | **Siblings** ($n$=98) |
|---|---|---|
| *Demographics* | | |
| Age (years) | $13.94 \pm 2.13$ | $13.46 \pm 2.16$ |
| Sex (M/F) | 86 / 16 | 52 / 46 |
| *Autism severity* | | |
| SRS-2 total | $73.71 \pm 10.01$ | $47.89 \pm 7.14$ |
| RBS-R total | $24.52 \pm 14.05$ | $4.15 \pm 6.76$ |
| *Sleep (parent-reported)* | | |
| CSHQ total | $46.19 \pm 8.05$ | $43.31 \pm 6.67$ |
| *Co-occurring diagnoses* | | |
| ADHD | 58 (56.86%) | 30 (29.41%) |
| Depression | 33 (32.35%) | 23 (23.47%) |
| Anxiety | 16 (15.68%) | 11 (11.22%) |

## Supplementary E. DREEM label reliability against pediatric norms

Expert-scored PSG is unavailable for the SSP cohort, so we assess DREEM label reliability indirectly by comparing DREEM-derived stage proportions against normative pediatric AASM-PSG values (Scholle et al., 2011) and against expected developmental trends. Table E.1 reports the stage distribution (percent of total sleep time) for each diagnostic group alongside the Scholle normative ranges.

In the neurotypical (TD) subgroup, N2 and REM proportions match the age-matched norms almost exactly, while N1 is modestly underestimated and N3 modestly overestimated (opposite-signed biases consistent with some lighter-sleep epochs being scored as deeper). The proportions are stable across TD, ASD, ADHD, and ASD+ADHD, indicating the imbalance reflects normal adolescent sleep architecture rather than a diagnosis- or device-specific artifact. Aggregating to one value per participant, the proportions move in the expected developmental directions with age (N3 declining, $\rho = -0.23$; N2 rising, $\rho = +0.26$; N1 and REM flat), reproducing the known N3 $\rightarrow$ N2 redistribution. That DREEM reproduces both the normative composition and the developmental trajectory gives confidence that its labels are physiologically plausible in this pediatric cohort.

Table E.1: **Distribution of sleep-stage proportions by diagnostic group versus pediatric norms.** Percent of total sleep time per stage. Normative bases from Scholle et al. (2011) (Scholle et al., 2011) are the observed group means (ages 11.5–16.9) and regression predictions (ages 10–17). DREEM-derived proportions are shown for each diagnostic group and the full filtered cohort.

| Group / Source | Participants | N1 (%) | N2 (%) | N3 (%) | REM (%) |
|---|---|---|---|---|---|
| Scholle (2011), observed | ages 11.5–16.9 | 6.9–7.9 | 42.2–49.4 | 22.9–28.9 | 19.0–20.3 |
| Scholle (2011), regression | ages 10–17 | 7.1–7.5 | 40.2–48.0 | 24.5–30.1 | 18.6–21.1 |
| TD | 37 | 5.2 | 43.9 | 30.5 | 20.4 |
| ASD | 34 | 5.2 | 42.3 | 31.5 | 21.0 |
| ADHD | 17 | 4.9 | 41.4 | 32.3 | 21.4 |
| ASD+ADHD | 40 | 5.4 | 40.4 | 33.6 | 20.6 |
| All | 128 | 5.3 | 42.0 | 32.0 | 20.8 |

## Supplementary F. Compliance Score

We define a per-participant compliance score to quantify how EEG signal availability evolves across the sleep window. Rather than computing a simple proportion of on-head epochs (which would treat all gaps equally regardless of timing), we use an exponential moving average (EMA) that gives greater weight to recent epochs. This captures the trajectory of signal availability over the night: a headband that stays on for the first half and then falls off produces a qualitatively different compliance profile than one with intermittent brief gaps throughout.

Let $h_i \in \{0, 1\}$ denote whether epoch $i$ is on-head, where $i = 1, \ldots, N$ indexes 30-second epochs within the sleep window. The compliance score at epoch $i$ is defined as,

$$C_i = \alpha\, h_i + (1 - \alpha)\, C_{i-1}, \quad C_0 = 1, \tag{F.1}$$

where $\alpha \in (0, 1]$ controls the smoothing window. The initialization $C_0 = 1$ assumes the headband is on-head at sleep onset. Smaller $\alpha$ produces a smoother trajectory reflecting sustained trends, while larger $\alpha$ tracks rapid fluctuations.

The per-participant mean compliance score summarizes overall signal availability,

$$\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{F.2}$$

This statistic captures both total on-head time and the temporal pattern of data loss, since a participant whose headband falls off abruptly mid-night will have a lower $\bar{C}$ than one with the same total on-head time distributed as brief intermittent gaps, reflecting the greater impact of sustained signal loss on sleep staging continuity.

**Compliance shows no conclusive sex difference.** Compliance shows no conclusive difference by sex for the DREEM headband, stratified by diagnostic group (Figure F.1). Male and female participants have comparable DREEM compliance (Male: $n = 99$, mean $= 81.5$; Female: $n = 93$, mean $= 81.5$; $p = 0.12$). The age-related compliance decline reported in the main text is therefore not confounded by sex.

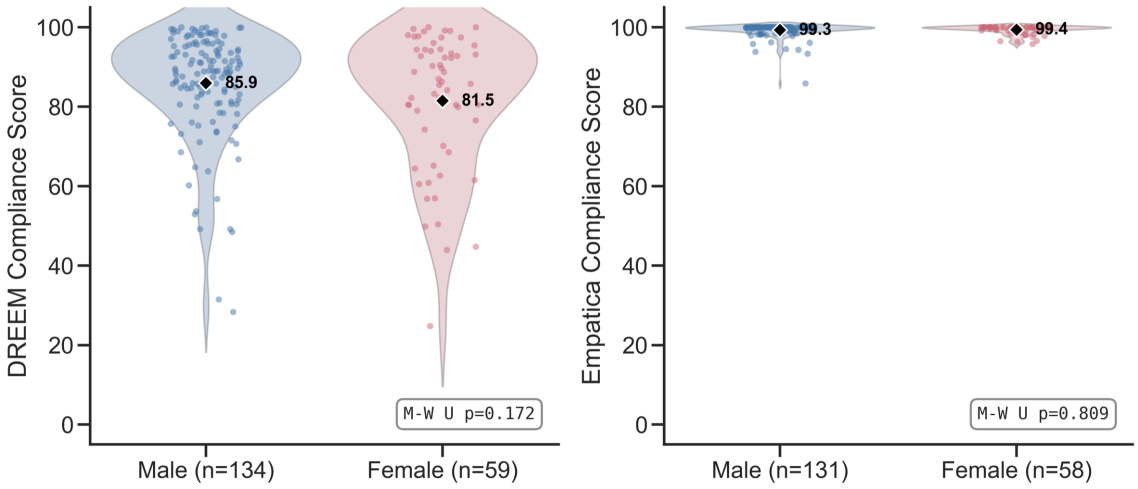



Figure F.1: **DREEM and Empatica compliance show no conclusive difference by sex.** Mean per-participant compliance score by sex, with DREEM compliance on the left and Empatica compliance on the right. The statistics are annotated in the text box below the plot.

## Supplementary G. Compliance versus behavioral and diagnostic severity

Our main-text analysis relates compliance to macro characteristics (age, sex, diagnostic group). Here we relate the per-participant compliance score directly to nuanced behavioral and autism-severity measures across the 193 participants who wore the headband. The compliance score is significantly associated with several autism core-severity measures spanning the social, repetitive, and sensory DSM-5 domains, most notably the Social Responsiveness Scale (SRS-2; $\rho = 0.21$, $p < .01$). These associations remain significant after adjusting for age and sex (Figure G.1), indicating that behavioral variation contributes to compliance beyond the macro characteristics analyzed in the main text.

We also relate the compliance score to off-head propensity, a physiologically interpretable form of non-compliance that isolates prolonged headband removal while the concurrent Empatica signal indicates ongoing sleep (i.e., volitional non-adherence). The two quantities are almost perfectly anti-correlated across the cohort ($\rho \approx -0.997$; Figure G.2), so analyzing off-head propensity is effectively an analysis of compliance expressed on a more behaviorally interpretable scale. Both lenses yield the same conclusion. Device adherence is trait-correlated and must be treated as potential structured missingness. All associations are correlational and are not fully adjusted for the collinear behavioral, diagnostic, age, and sex variables.

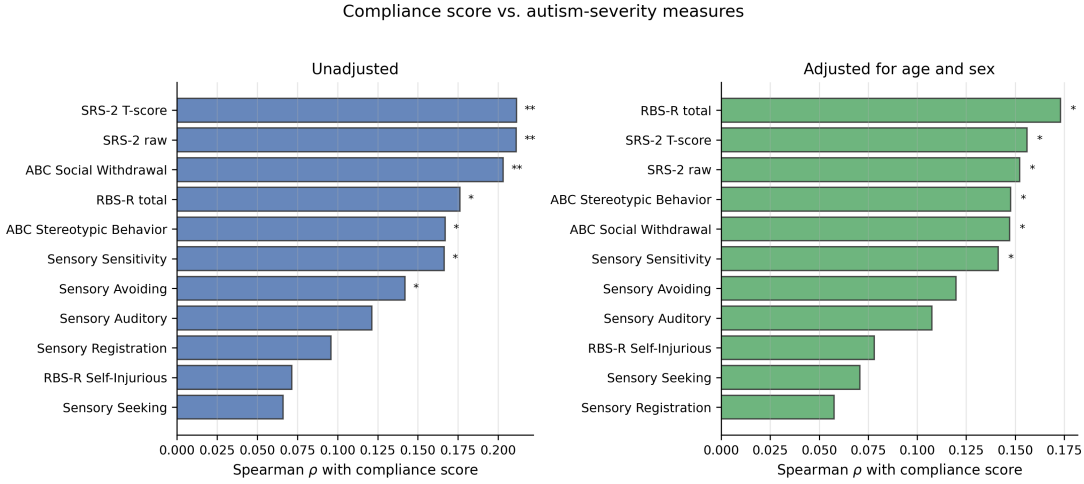



Figure G.1: **Compliance score versus autism-severity measures.** Spearman correlations between the per-participant DREEM compliance score and behavioral/severity scores (left), and the same associations after adjusting for age and sex (right). The Social Responsiveness Scale shows the strongest association ($\rho = 0.21$, $p < .01$), which persists after adjustment.

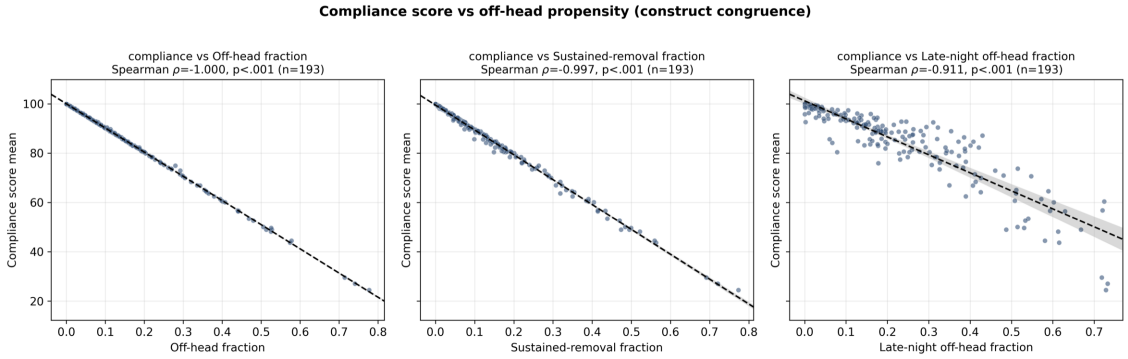



Figure G.2: **Compliance score and off-head propensity are near-perfectly anti-correlated.** Per-participant compliance score ($n = 193$) versus three off-head propensity definitions: off-head fraction (Spearman $\rho = -1.000$), sustained-removal fraction ($\rho = -0.997$), and late-night off-head fraction ($\rho = -0.911$); all $p < .001$. Dashed lines are OLS fits with 95% confidence bands. The near-perfect anti-correlation shows the compliance score and off-head propensity capture the same underlying non-adherence through different filters.

## Supplementary H. Cross-Device Sleep Metric Validation

We compared sleep metrics from the DREEM headband and Empatica wristband across $n = 1{,}141$ concurrent recording nights from 128 participants to validate cross-device agreement after compliance-driven filtering (Figure H.1). Both devices agree well on gross sleep timing, with TST ($r = 0.73$, $\rho = 0.73$), sleep onset ($r = 0.65$, $\rho = 0.89$), and final awakening ($r = 0.79$, $\rho = 0.82$) all showing strong agreement, confirming that both sensors reliably detect when participants fall asleep and wake up. Disagreement emerges for within-sleep disruption metrics. WASO ($r = 0.23$, $\rho = 0.47$), sleep efficiency ($r = 0.30$, $\rho = 0.33$), and sleep fragmentation ($r = 0.33$, $\rho = 0.37$) show substantially weaker agreement with a directional bias, namely Empatica overestimates WASO by 16.7 minutes on average and detects 0.74 more sleep-to-wake transitions per hour. This is consistent with IMU-based actigraphy misclassifying brief motor restlessness during sleep as wakefulness, inflating apparent disruption.

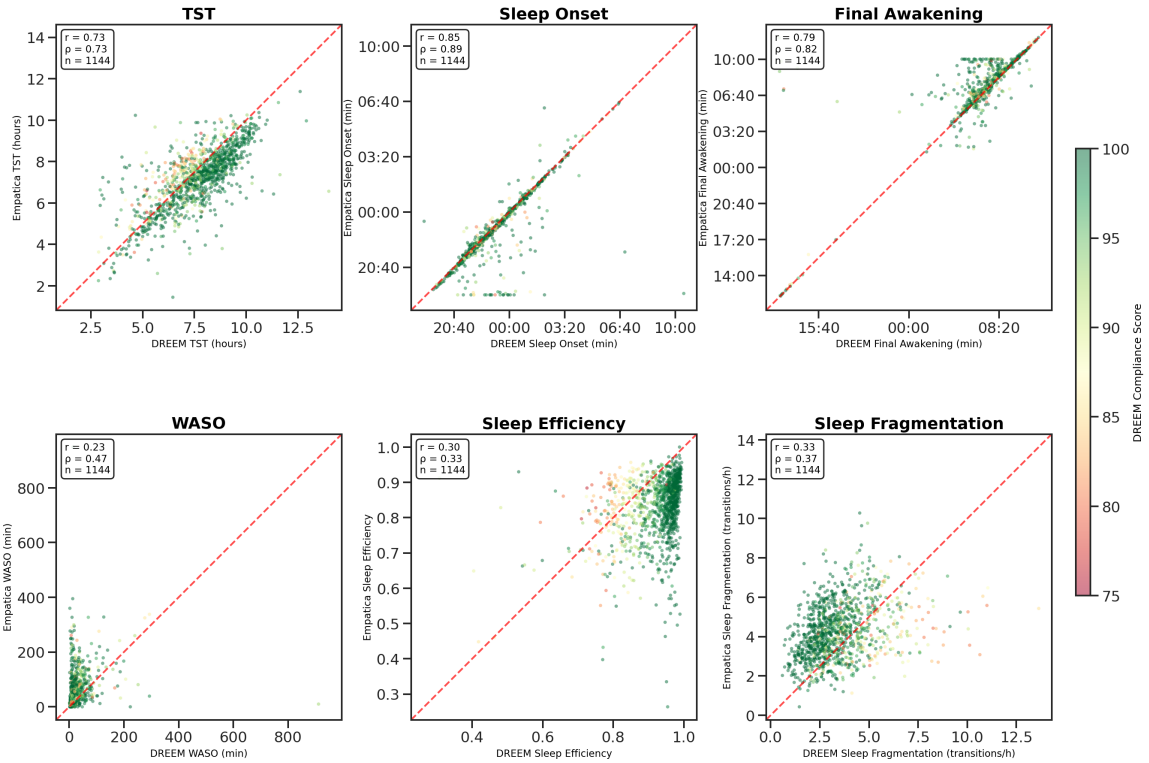



Figure H.1: **DREEM and Empatica agree on gross sleep timing but diverge on within-sleep disruption metrics.** DREEM versus Empatica sleep metrics across $n = 1{,}141$ concurrent recording nights (128 participants) prior to compliance-based filtering. Each point is one night, colored by DREEM compliance score (lighter = lower compliance); the red line is the identity. Top row (total sleep time (TST), sleep onset, final awakening) shows strong agreement. Bottom row (wake after sleep onset (WASO), sleep efficiency, sleep fragmentation) shows weak, systematically biased agreement, with Empatica overestimating wake.

## Supplementary I. Sleep Metric Definitions

**Sleep Efficiency.** Sleep efficiency (SE) is defined as the proportion of time in bed spent asleep,

$$\mathrm{SE} = \frac{\mathrm{TST}}{\mathrm{TIB}}, \tag{I.1}$$

where Total Sleep Time (TST) is the total duration of epochs scored as sleep and Time in Bed (TIB) is the total duration of the recording window, both in minutes. For the DREEM headband, TST is taken from the EEG-based staging and $\mathrm{TIB} = t_{\mathrm{end}} - t_{\mathrm{start}}$ from the recorded sleep window. For the Empatica watch, TST is the count of epochs classified as sleep between sleep onset (SO) and final awakening (FA), and TIB is the total number of epochs in the SO→FA window (1 min/epoch).

**Restfulness (IMU-based mean power).** Restfulness quantifies nocturnal motor activity from the Empatica wristband accelerometer, where lower values indicate more restful sleep. First, for each accelerometer axis $k \in \{x, y, z\}$, a 4th-order Butterworth high-pass filter with cutoff $f_c = 0.5\,\mathrm{Hz}$ removes the gravitational component,

$$\hat{a}_k = \mathrm{HPF}_{f_c = 0.5\,\mathrm{Hz}}(a_k). \tag{I.2}$$

The instantaneous power is the sum of squared filtered accelerations,

$$P(t) = \hat{a}_x(t)^2 + \hat{a}_y(t)^2 + \hat{a}_z(t)^2, \tag{I.3}$$

and the restfulness metric is defined as the mean power over all samples, namely

$$\bar{P} = \frac{1}{N} \sum_{t=1}^{N} P(t), \tag{I.4}$$

where $N$ is the total number of accelerometer samples at 64 Hz, and $\bar{P}$ is in units of $g^2$.

## Supplementary J. Effect of Headband on Sleep

Because participants wore the Empatica watch on all study nights but the DREEM headband on only a subset, we use Empatica-derived sleep metrics to compare headband versus non-headband nights within the same participant. We restricted the analysis to participants with at least 3 nights in each condition ($n = 526$ DREEM+Empatica nights, $n = 334$ Empatica-only nights).

TST was higher on DREEM nights ($\mu = 5.22$ vs. 4.79 hrs, $d_z = 0.39$, $p < .001$) and restfulness was lower ($\mu = 50.48$ vs. 54.16, $d_z = -0.27$, $p = 0.037$), while WASO ($d_z = 0.09$, $p = 0.305$) and sleep efficiency ($d_z = -0.12$, $p = 0.325$) showed no difference. All effect sizes are small to moderate ($|d_z| \leq 0.39$). The limited sample size and the non-randomized assignment of conditions preclude definitive conclusions about whether the headband causally affects sleep (Figure J.1).

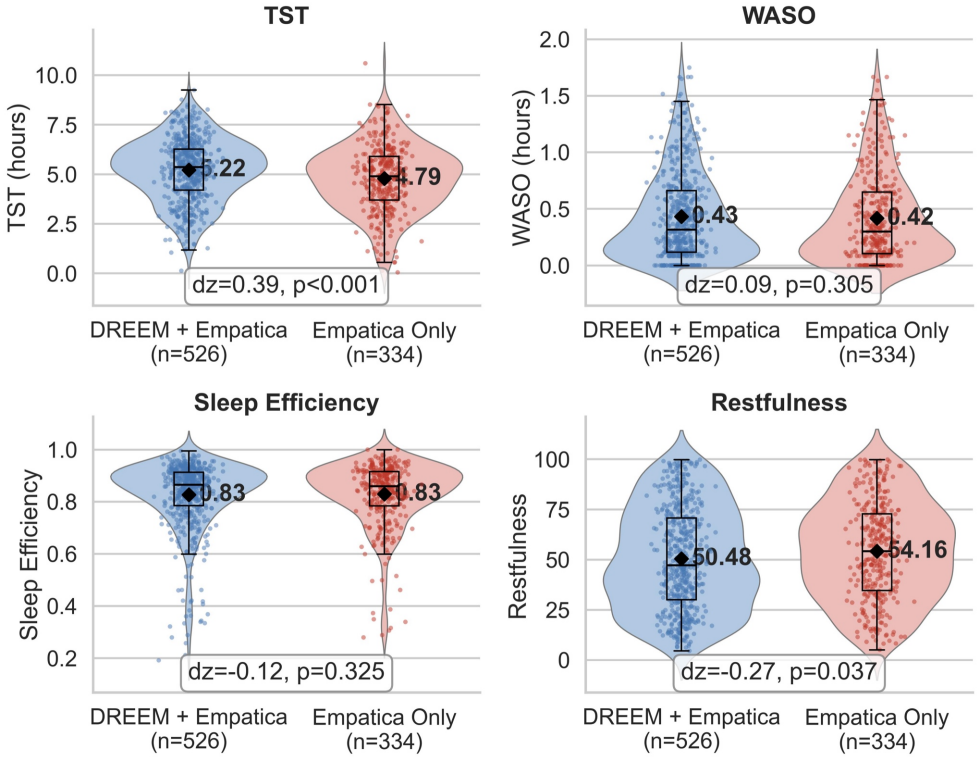



Figure J.1: **Within-subject comparison of Empatica-derived sleep metrics on headband versus non-headband nights.** Four Empatica-derived sleep metrics, namely total sleep time (TST), wake after sleep onset (WASO), sleep efficiency, and restfulness, compared between DREEM+Empatica nights and Empatica-only nights, restricted to participants with at least 3 nights in each condition ($n = 526$ DREEM+Empatica, $n = 334$ Empatica-only). The Cohen's effect size ($d_z$) and the two-sided Wilcoxon signed-rank test $p$-value are annotated in each panel.

## Supplementary K. Compliance-aware cohort filtering

This supplementary section describes the compliance-driven filtering pipeline introduced in Section 4 and illustrated in Figure 4, along with the demographic, sleep, and compliance characteristics of the resulting cohort. The pipeline was designed to balance signal quality against sample size, and proceeds in five sequential steps:

1. Participants with no DREEM recordings on any night are excluded.
2. Nights with more than 20% off-head epochs are excluded.
3. Nights lacking concurrent Empatica data are excluded.
4. Nights failing Empatica artifact and sleep-classification checks are excluded.
5. Participants with fewer than three surviving nights are excluded.

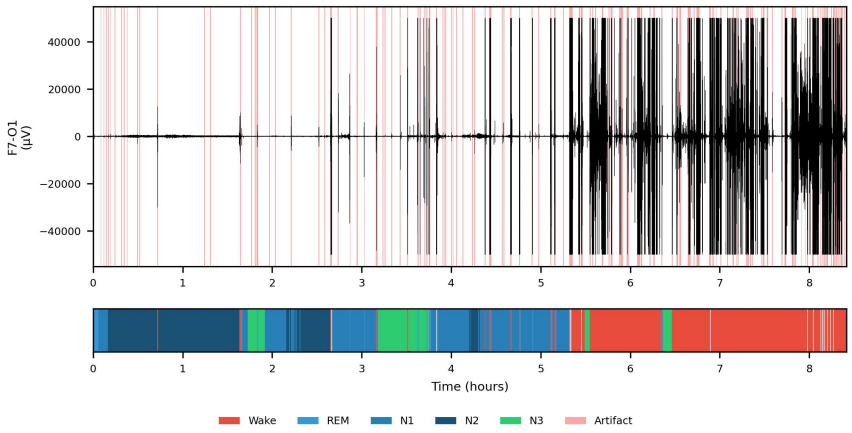
After applying these criteria, 128 of the 200 enrolled participants and 1,147 nights are retained for downstream analysis. Table K.1 reports the characteristics of this filtered cohort, stratified by the four diagnostic groups used throughout the paper.

Table K.1: Filtered cohort characteristics (128 participants, 1,147 nights). Values are mean ± SD or $n$ (%).

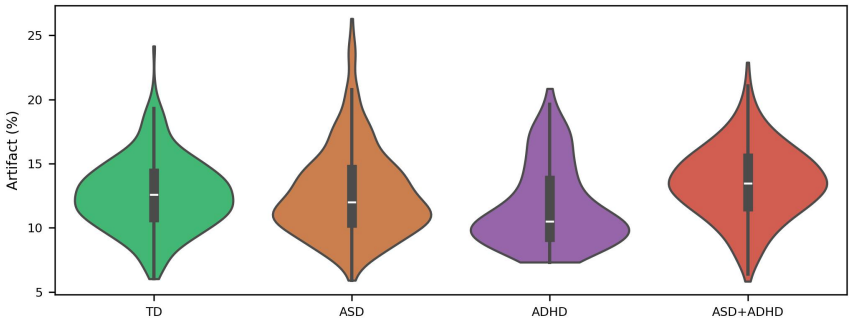
| | **ASD** ($n$=34) | **ASD+ADHD** ($n$=40) | **ADHD** ($n$=17) | **TD** ($n$=37) |
|---|---|---|---|---|
| Age (years) | $12.9 \pm 2.0$ | $13.4 \pm 2.2$ | $13.4 \pm 2.4$ | $12.7 \pm 2.3$ |
| Sex (M/F) | 28/6 | 38/2 | 10/7 | 19/18 |
| CSHQ total | $42.6 \pm 7.8$ | $44.1 \pm 6.8$ | $44.4 \pm 5.0$ | $39.5 \pm 5.7$ |
| Nights per participant | $10.1 \pm 4.5$ | $8.9 \pm 4.9$ | $9.1 \pm 3.1$ | $8.0 \pm 3.4$ |
| Total nights | 343 | 355 | 154 | 295 |
| DREEM Compliance (%) | $95.5 \pm 4.0$ | $94.2 \pm 4.0$ | $95.5 \pm 2.5$ | $95.5 \pm 3.4$ |

## Supplementary L. EEG signal quality and artifact analysis

Raw DREEM recordings contain large regions clipped at approximately ±50,000 $\mu$V, far exceeding the ±100–200 $\mu$V amplitude range assumed by standard laboratory-grade PSG pipelines. These clipped segments constitute approximately 4–5% of each recording and arise from the high-impedance characteristics of dry-electrode contact without conductive gel. Because HAPPE (Gabard-Durnam et al., 2018) assumes PSG-grade amplitude distributions, its internal thresholds and artifact-rejection criteria are incompatible with these signals, causing the pipeline to fail on DREEM data.




(A)




(B)

Figure L.1: **EEG artifacts are sparse and remain below 15% across all diagnostic groups.** (A) Raw EEG amplitude traces for a representative night with LUNA-detected artifact epochs shaded in red, showing sparse temporal distribution consistent with transient electrode adjustments rather than sustained off-head events. (B) Distribution of LUNA-detected artifact percentages per recording night, stratified by diagnostic group (ASD, ASD+ADHD, ADHD, TD); the horizontal line marks the 15% threshold. All groups have median artifact rates near 11–13%, well below 15%.

### L.1. HAPPE preprocessing on DREEM EEG

We additionally evaluated whether DREEM EEG recordings could be processed using HAPPE (Gabard-Durnam et al., 2018; Lopez et al., 2022), an automated preprocessing pipeline widely used for developmental and autism EEG studies. HAPPE is designed to handle high-artifact EEG and low-density recordings, making it a natural candidate for pediatric wearable EEG data. However, in our setting, HAPPE did not produce artifact-free waveforms from the DREEM recordings when using its wavelet-based artifact correction procedure.

The main failure mode was driven by the mismatch between DREEM dry-electrode recordings and the assumptions of standard EEG preprocessing pipelines. The raw DREEM signals contain large clipped regions at $\pm$50,000 $\mu$V, likely reflecting contact instability or amplifier saturation rather than physiological EEG activity. In addition, the DREEM montage consists of a reduced set of bipolar dry-electrode derivations, rather than a conventional referential PSG or 10–20 EEG montage. These properties make amplitude-based and wavelet-based artifact correction difficult, because large non-physiological excursions dominate the decomposition and can propagate into the reconstructed signal.

For this reason, we did not use HAPPE-preprocessed waveforms for downstream sleep staging experiments. Instead, we used LUNA's spectral artifact rejection procedure, which operates on frequency-domain features and was more robust to atypical amplitude distributions in the DREEM recordings.

## Supplementary M. Independent signal-quality analysis of DREEM EEG

The compliance score is derived from DREEM's binary on-/off-head labels. To verify that it reflects genuine signal-quality degradation rather than a coarse binary heuristic, we characterize raw EEG quality using probes that are independent of DREEM's proprietary algorithm.

**DREEM off-head status reflects electrode-level degradation.** We trained YASA-based staging predictors to classify each of the seven raw bipolar channels independently as sleep-stage or off-head. Each channel independently flagged approximately 1.5% of epochs as off-head, and a simple "any-channel" aggregation rule recovered DREEM's headband-level off-head labels at 83% balanced accuracy and 83% sensitivity. Representing each epoch by YASA's 65 staging features per channel, concatenated across all seven channels (455 dimensions) and projected to 2D with UMAP, epochs move progressively away from the clean-sleep manifold as more channels degrade, with the most degraded epochs (those with 6–7 channels flagged) forming a distinct lobe (Figure M.1), and a classifier on these features distinguishes DREEM off-head epochs at AUC $\approx 0.80$ at the true $\sim$4% prevalence. DREEM's binary off-head label thus tracks graded, electrode-level degradation rather than a single global signal.

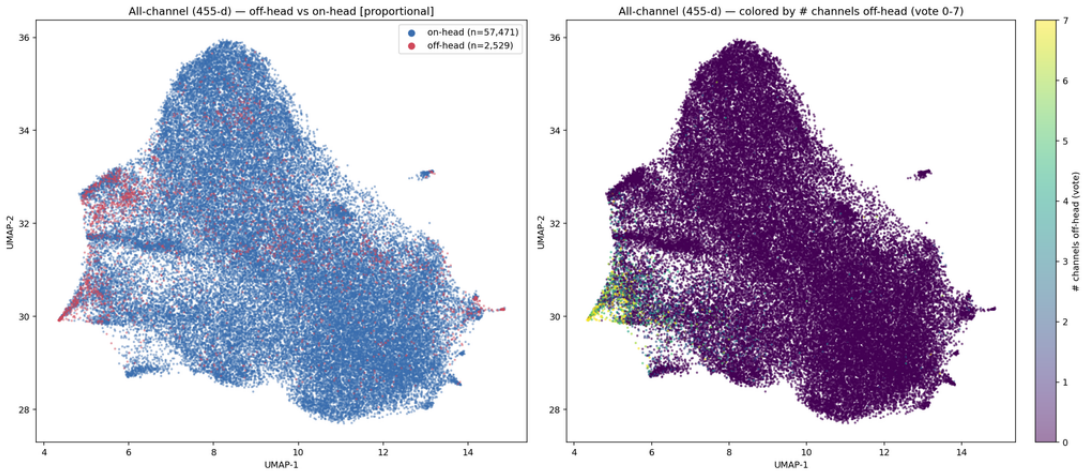



Figure M.1: **Per-channel feature geometry tracks graded electrode-level degradation.** UMAP of per-epoch 455-dimensional per-channel YASA features. (Left) Epochs colored by DREEM's binary on-/off-head label ($n$ = 57,471 on-head, $n$ = 2,529 off-head). (Right) Epochs colored by the number of channels independently flagged off-head (0–7), revealing a graded dose-response lobe. Epochs move progressively away from the clean-sleep manifold as more channels degrade.

**DREEM-agnostic quality metrics decline across the night.** We define two raw-signal probes. The first is an amplitude-clipping rate (fraction of each 30 s epoch clipped at $\pm$50,000 $\mu$V), and the second is an electrode-jump rate (sample-to-sample jumps above 300 $\mu$V, capturing dry-electrode slippage). Both worsen as the night progresses (amplitude clipping 4.1×, electrode jumps 1.7×), and the same upward trend holds within on-head

epochs alone (Figure M.2), so the decline reflects genuine degradation rather than the headband simply coming off.

**Low compliance coincides with high raw contamination.** Nights in the lowest compliance quartile carry an amplitude-clipping rate of 1.16% versus 0.30% in the highest quartile (3.9×), and an electrode-jump rate of 2.46% versus 1.48% (1.7×). Across all 2,497 nights, compliance is negatively correlated with clipping (Spearman $\rho = -0.58$) and jumps ($\rho = -0.29$), so low compliance and high raw contamination identify the same recordings.

**Artifact detectors flag contamination even within on-head epochs.** Applying YASA variance (single-channel) and covariance (multi-channel) detectors and LUNA time-domain and spectral detectors, within scored on-head epochs the LUNA spectral detector flags 14.3% of epochs on average, LUNA time-domain 4.5%, YASA variance 2.9%, and YASA covariance 0.6%. Contamination grows across the sleep window for all four detectors, and participants with lower compliance carry more LUNA-detected on-head artifact (LUNA spectral $\rho = -0.25$, $p < 0.001$, $n = 192$). The compliance score therefore captures a conservative lower bound on the true signal quality, yet still reveals a systematic, structured decline across the night that is routinely overlooked in machine-learning studies of wearable physiological data.

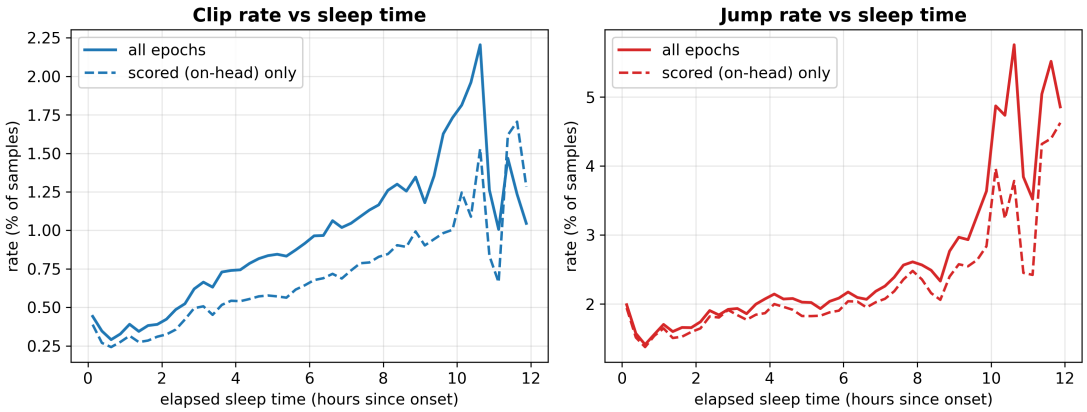



Figure M.2: **Raw signal quality degrades across the night, independently of DREEM's off-head labels.** Amplitude-clipping rate (left) and electrode-jump rate (right) as a function of elapsed sleep time, for all epochs (solid) and for on-head ("scored") epochs only (dashed). Both rates rise across the night and the trend persists within on-head epochs alone, so the decline reflects genuine signal degradation rather than the headband coming off; the binary compliance score is therefore a conservative lower bound on raw signal quality.

## Supplementary N. EEG-based Staging Details

**Confusion matrices.** Per-stage confusion matrices for YASA, LUNA, and SleepFM in both zero-shot and fine-tuned settings (5-class) are shown in Figure N.1. In the zero-shot setting, YASA over-predicts Wake across all true stages, LUNA shows a strong bias toward N3, and SleepFM exhibits distributed misclassification. After fine-tuning, all three models correctly classify over 74% of N2 epochs and over 86% of N3 epochs. N1 remains the most difficult stage (∼42–55% recall), consistent with known difficulties in N1 discrimination even in PSG settings (Vallat and Walker, 2021).

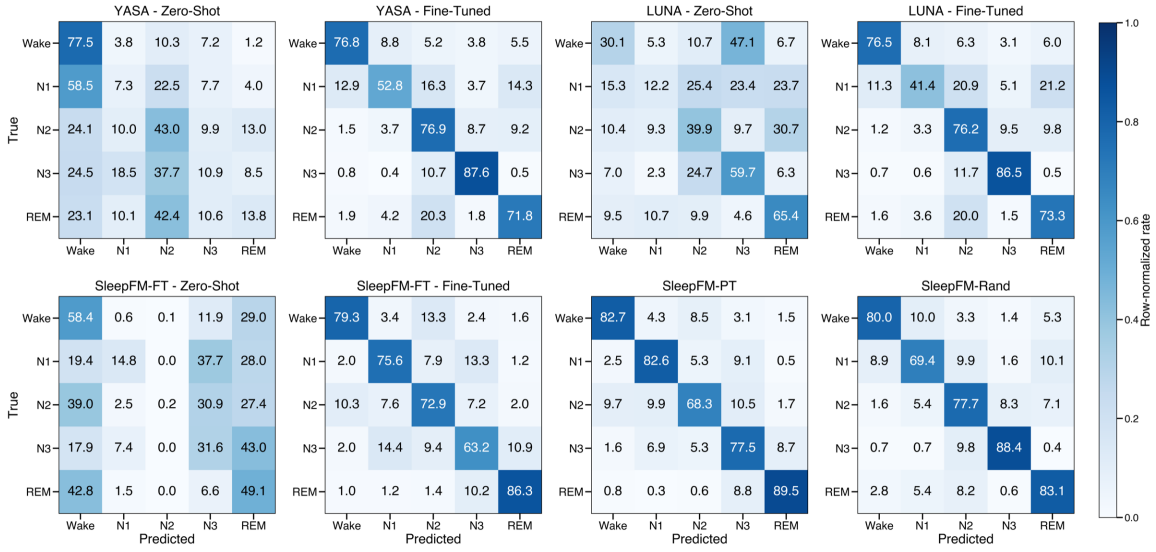



Figure N.1: **Fine-tuning resolves systematic zero-shot staging biases across all three EEG models.** Row-normalized 5-class EEG staging confusion matrices for YASA, LUNA, and SleepFM-FT against DREEM ground-truth stages (Wake, N1, N2, N3, REM). For each pair of figures, the left column shows zero-shot performance, while the right column shows performance after fine-tuning (FT) on DREEM data. YASA zero-shot over-predicts Wake and LUNA zero-shot is biased toward N3; fine-tuning resolves these biases across all three models.

**YASA and LUNA accuracy across staging granularities.** Fine-tuned YASA and LUNA accuracy and Cohen's $\kappa$ across three staging granularities (4-class, 5-class, 6-class) alongside zero-shot results are shown in Figure N.2. Both models improve substantially after fine-tuning at every granularity. Performance decreases modestly as the number of classes increases, consistent with the greater difficulty of resolving finer-grained stage boundaries.

**Sample efficiency of DREEM fine-tuning.** We varied the number of participants used for fine-tuning from 1 up to the full training pool and measured 5-class staging accuracy and Cohen's $\kappa$ on held-out participants. Both models achieve rapid performance gains with small training sets, reaching near-plateau accuracy and $\kappa$ with approximately 30–50 participants (Figure N.3). Despite LUNA's superior zero-shot performance, YASA overtakes it after fine-tuning on as few as 10 participants, and maintains higher accuracy across all

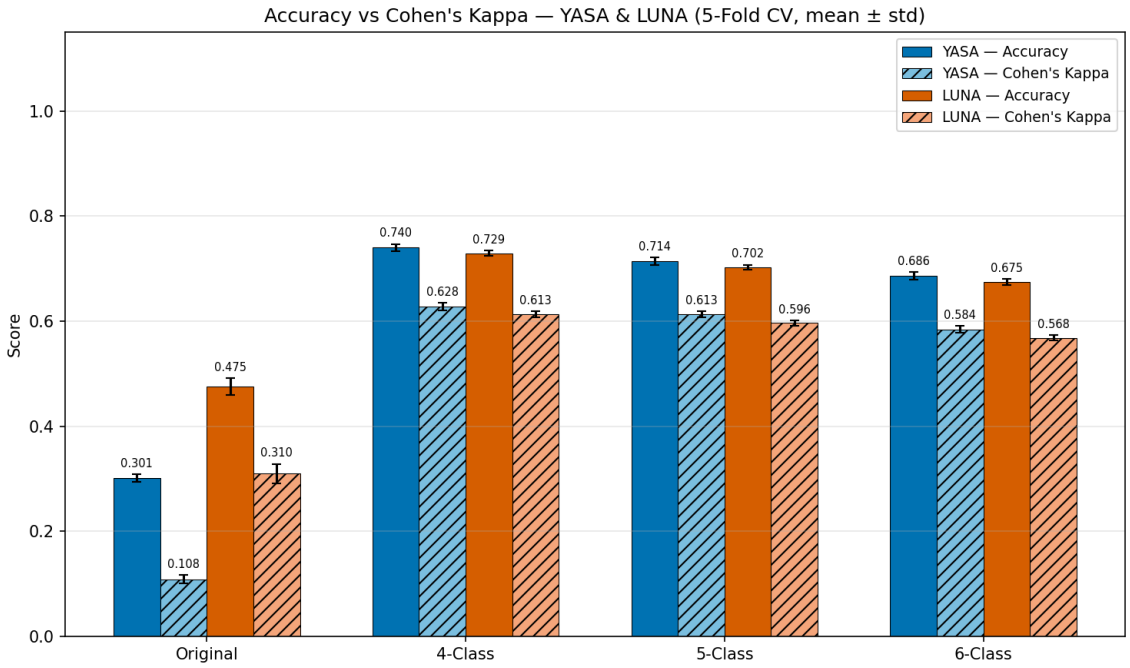



Figure N.2: **Fine-tuning improves EEG staging accuracy at every classification granularity.** YASA and LUNA accuracy and Cohen's $\kappa$ across three staging granularities (4-class = Wake/Light/Deep/REM; 5-class = Wake/N1/N2/N3/REM; 6-class = Wake/N1/N2/N3/REM/Artifact), comparing zero-shot and fine-tuned (FT) variants of each model. Bars show the mean across 5-fold cross-validation; error bars are 1 std. Both models improve substantially after fine-tuning at every granularity, and performance decreases modestly as class count increases.

subsequent training set sizes. These results indicate that the domain shift can be bridged with a modest pilot cohort, a practical consideration for deploying PSG-trained models in new wearable EEG settings.

**Summary of zero-shot and fine-tuned EEG staging metrics.** Table N.1 summarizes the 5-fold cross-validation performance of YASA, LUNA, and SleepFM in both zero-shot and fine-tuned settings on the 4-class, 5-class, and 6-class staging tasks. Values are reported as mean ± half-range $(\max - \min)/2$ across folds. SleepFM is reported on the 5-class task only, as fine-tuning was not performed at the other granularities. Zero-shot metrics at 4-class and 6-class evaluate the pre-trained 5-class outputs under the corresponding label space (4-class via label mapping; 6-class with the Artifact class treated as a never-predicted positive, which lowers Macro F1).

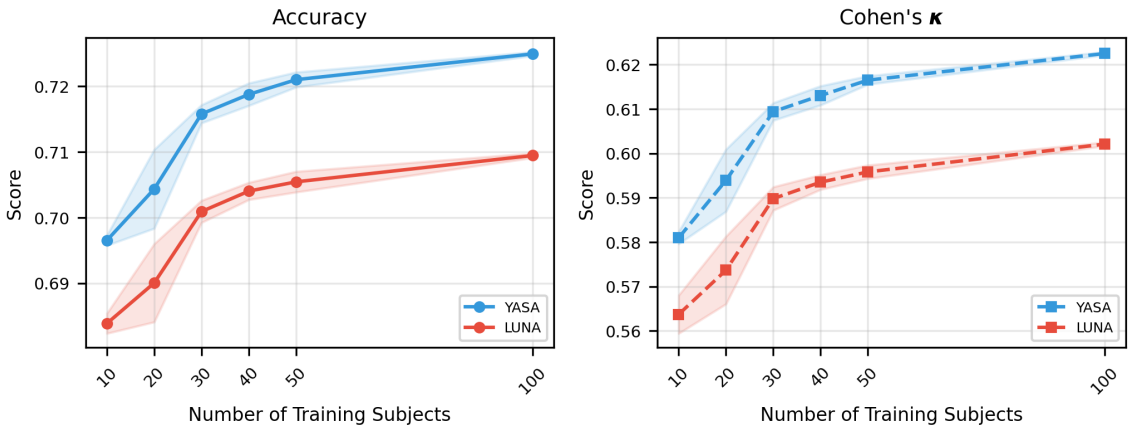



Figure N.3: **EEG staging performance plateaus with as few as 30–50 fine-tuning participants.** Sample efficiency of DREEM fine-tuning: 5-class staging accuracy (circle) and Cohen's $\kappa$ (square) as a function of the number of training participants for YASA and LUNA. Both models plateau around 30–50 participants.

Table N.1: **Accuracy, Cohen's $\kappa$, and Macro F1 for zero-shot and fine-tuned EEG staging models on DREEM.** Mean ± half-range across 5-fold cross-validation. Class definitions: 4-class = (Wake, Light Sleep, Deep Sleep, REM); 5-class = (Wake, N1, N2, N3, REM); 6-class = (Wake, N1, N2, N3, REM, Artifact). Zero-shot rows apply the pre-trained model (YASA defaults, LUNA POPS s2, SleepFM) without DREEM fine-tuning; fine-tuned rows correspond to the models after fine-tuning on DREEM. SleepFM was evaluated only on the 5-class task.

| Task | Model | Setting | Accuracy | Cohen's $\kappa$ | Macro F1 |
|---|---|---|---|---|---|
| 4-class | YASA | Zero-shot | $0.362 \pm 0.012$ | $0.108 \pm 0.015$ | $0.310 \pm 0.012$ |
| | YASA | Fine-tuned | $0.740 \pm 0.011$ | $0.628 \pm 0.011$ | $0.737 \pm 0.009$ |
| | LUNA | Zero-shot | $0.521 \pm 0.020$ | $0.336 \pm 0.025$ | $0.491 \pm 0.020$ |
| | LUNA | Fine-tuned | $0.729 \pm 0.008$ | $0.613 \pm 0.008$ | $0.729 \pm 0.008$ |
| 5-class | YASA | Zero-shot | $0.315 \pm 0.013$ | $0.115 \pm 0.013$ | $0.249 \pm 0.010$ |
| | YASA | Fine-tuned | $0.714 \pm 0.010$ | $0.613 \pm 0.010$ | $0.668 \pm 0.010$ |
| | LUNA | Zero-shot | $0.475 \pm 0.022$ | $0.310 \pm 0.024$ | $0.399 \pm 0.019$ |
| | LUNA | Fine-tuned | $0.702 \pm 0.008$ | $0.596 \pm 0.007$ | $0.648 \pm 0.009$ |
| | SleepFM | Zero-shot | $0.351 \pm 0.029$ | $0.141 \pm 0.033$ | $0.269 \pm 0.021$ |
| | SleepFM | Classifier FT | $0.743 \pm 0.006$ | $0.659 \pm 0.006$ | $0.708 \pm 0.006$ |
| | SleepFM | End-to-End Pretrained | $0.820 \pm 0.016$ | $0.757 \pm 0.019$ | $0.776 \pm 0.014$ |
| | SleepFM | End-to-End Random | $0.815 \pm 0.017$ | $0.752 \pm 0.019$ | $0.771 \pm 0.017$ |
| 6-class | YASA | Zero-shot | $0.301 \pm 0.011$ | $0.108 \pm 0.012$ | $0.203 \pm 0.007$ |
| | YASA | Fine-tuned | $0.686 \pm 0.008$ | $0.584 \pm 0.008$ | $0.585 \pm 0.003$ |
| | LUNA | Zero-shot | $0.475 \pm 0.022$ | $0.310 \pm 0.024$ | $0.333 \pm 0.015$ |
| | LUNA | Fine-tuned | $0.675 \pm 0.006$ | $0.568 \pm 0.007$ | $0.573 \pm 0.003$ |

## Supplementary O. Fine-tuning SleepFM

We adapted SleepFM (Thapa et al., 2026) to DREEM data using three configurations (Figure O.1). *Classifier FT* freezes the pre-trained encoder and trains only an LSTM classification head on DREEM embeddings. *End-to-End Pretrain* jointly updates the encoder and LSTM classifier, allowing the learned representations to adapt to the dry-electrode DREEM domain. *End-to-End Random* uses the same architecture but with randomly initialized weights trained only on SSP data, serving as a control for whether PSG pre-training is necessary.

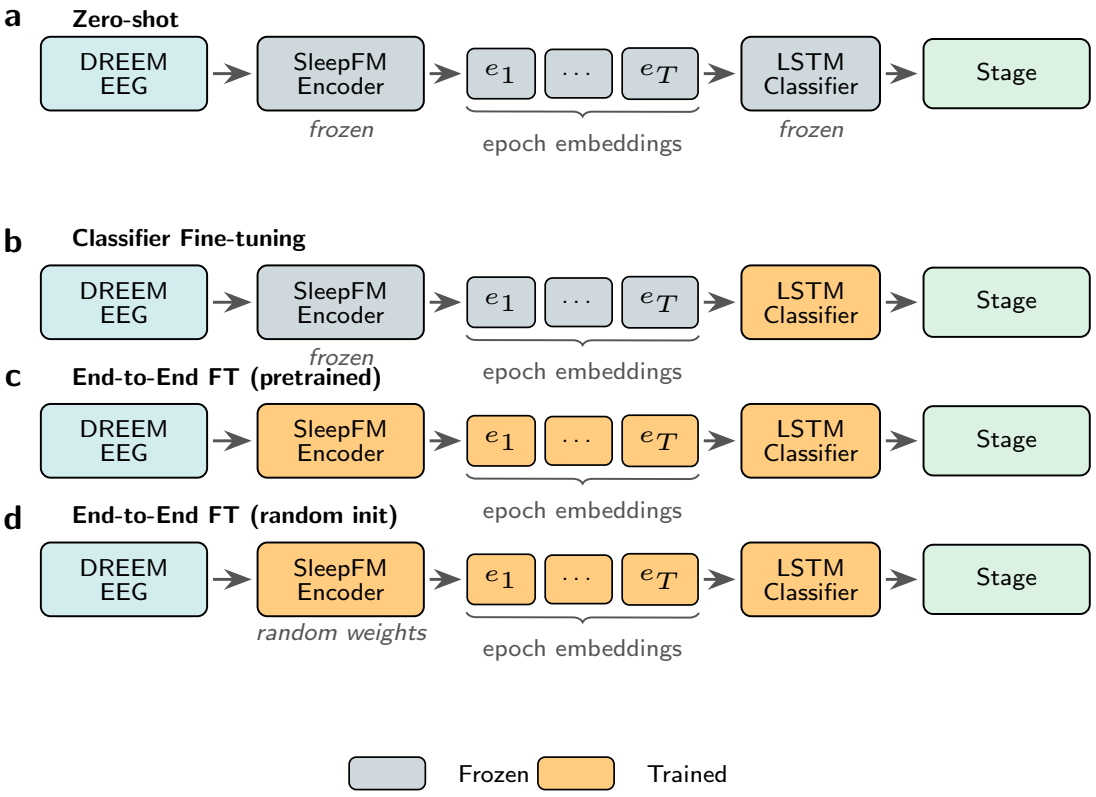



Figure O.1: **SleepFM adapts PSG-pre-trained representations to wearable EEG via three configurations.** The model is pre-trained via contrastive learning on multimodal PSG data and adapted to DREEM recordings using either a frozen encoder with an LSTM head (Classifier FT), end-to-end updates to both encoder and classifier with pre-trained weights (End-to-End Pretrain), or the same end-to-end architecture trained from randomly initialized weights (End-to-End Random).

## Supplementary P. Per-channel EEG staging accuracy

This supplementary section presents the per-channel staging analysis that informs the argument in Section 5 that prefrontal channels contribute relatively little discriminative information compared to frontal-to-occipital derivations.

Because we fine-tuned YASA to operate on DREEM data, we applied the model to all 7 bipolar channels independently, enabling a per-channel evaluation of staging accuracy and off-head compliance. The 5 electrodes making up the 7 channels connect different regions of the head (Figure 1(B)), carrying different brain signals. This opens the possibility for some channels to be more informative with respect to sleep staging.

Channel configuration has a marked effect on classification accuracy. Frontal-to-occipital derivations (F7–O1, F7–O2, F8–O1, F8–O2) outperform prefrontal derivations (Fp1–F7, Fp1–F8) on 6-class staging (Figure P.1). In contrast to the difference in accuracy, off-head rates remain uniformly low across all channels, with no single channel standing out as notably worse. While the compliance-based filtering (Section 4) reduced the overall off-head percentage, the off-head predictions are underpredicted relative to the DREEM headband's own off-head detection, suggesting that the EEG-derived features used by YASA do not fully capture device-level off-head events, which likely cross-validate signals across multiple channels. Taken together, these results suggest that prefrontal channels contribute relatively little discriminative information compared to frontal-to-occipital derivations, raising the possibility that a more comfortable headband with fewer, strategically chosen channels could achieve comparable accuracy.

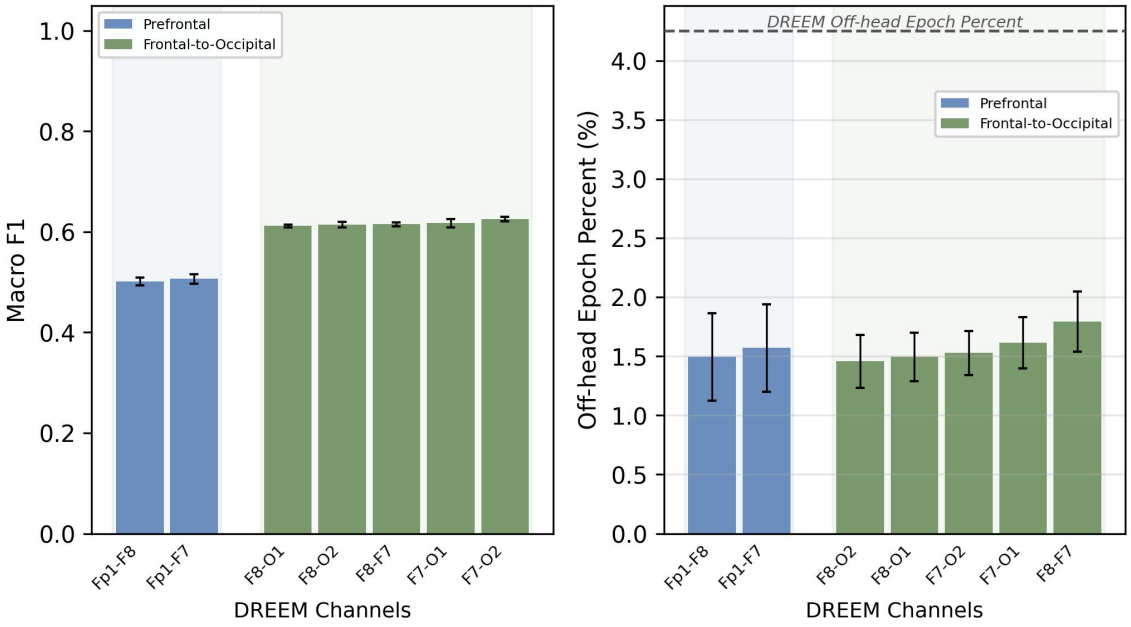



Figure P.1: **Frontal-to-occipital EEG channels consistently outperform prefrontal derivations for sleep staging.** Per-channel macro F1 (left) and off-head epoch percent (right) for fine-tuned YASA on 6-class staging across the seven DREEM bipolar derivations. Frontal-to-occipital channels (F7–O1, F7–O2, F8–O1, F8–O2) consistently outperform prefrontal channels (Fp1–F7, Fp1–F8), while off-head rates are uniformly low and well below the DREEM system's overall off-head rate (dashed line) across all channels.

## Supplementary Q. Sleep staging from wrist-worn sensors

**Sample efficiency.** A practical question for deploying wrist-based sleep staging is how many labeled participants are needed to train a reliable model. We varied the number of training participants for the PPG+IMU model and measured accuracy and $\kappa$ as a function of cohort size. Performance increased with the number of participants and began to saturate around 40 subjects, with diminishing returns beyond that point (panel (C) of Figure 6 in the main text). This suggests that a modest labeled cohort is sufficient to train effective feature-based wrist-staging models in this setting, and that the SSP dataset provides enough labeled data for benchmarking wrist-worn sleep staging models.

**Additional PaPaGei adaptation experiments.** We next evaluated whether PaPaGei (Pillai et al., 2025), a PPG foundation model, could provide a transferable representation for wrist-worn sleep staging. This is a challenging transfer setting because PaPaGei was pretrained on clinical-grade transmission PPG from MIMIC-III (Johnson et al., 2016), VitalDB (Lee et al., 2022), and MESA (Chen et al., 2015), whereas our target data consist of wrist-worn reflective PPG from Empatica devices. Compared with clinical PPG, wrist reflective PPG is more affected by motion artifacts, variable sensor contact, and differences in acquisition geometry.

We evaluated both frozen-backbone and end-to-end fine-tuned PaPaGei variants. In the *frozen-backbone* setting, PaPaGei was used as a fixed feature extractor. We tested linear probing on 30-second PPG windows, average pooling of embeddings from three 10-second windows, and attention pooling over the same within-epoch embeddings. In these models, only the pooling and classification layers were trained. These frozen variants performed poorly, suggesting that PaPaGei embeddings do not directly provide a sleep-stage-separable representation for wrist PPG. The 10-second variants were particularly limited, indicating that such short PPG windows provide insufficient temporal context for reliable sleep-stage discrimination.

We then evaluated end-to-end fine-tuned variants in which the PaPaGei backbone was updated jointly with a temporal classifier. Because sleep staging depends on temporal structure beyond a single epoch, these models used sequences of PaPaGei embeddings from neighboring windows. We considered three temporal heads: a Transformer encoder, an LSTM, and a Transformer followed by an LSTM. Fine-tuning improved performance, especially when longer temporal context was included, indicating that temporal structure is important for PPG-based sleep staging. However, even the best fine-tuned PaPaGei model remained below the XGBoost model trained on Empatica-derived feature summaries.

One reason for this gap is the difference in input representation. PaPaGei operates directly on raw wrist PPG waveforms, where denoising, beat detection, artifact rejection, and temporal aggregation must be learned or handled by the model. In contrast, the XGBoost baseline uses one-minute Empatica summaries, including heart rate, heart-rate variability, motion, temperature, and related features, which are derived from proprietary preprocessing pipelines. Thus, the stronger performance of feature-based wrist models likely reflects both the informativeness of wearable physiology and the value of robust preprocessing for noisy wrist PPG.

**PaPaGei transfer and per-stage performance.** Paralleling the EEG domain-shift analysis (Section 5), PaPaGei transfer was limited across adaptation strategies. Frozen linear probes, within-epoch pooling, and end-to-end fine-tuning all underperformed the supervised PPG+IMU feature-based model (panel (D) of Figure 6 in the main text). These results likely reflect several compounding factors: the mismatch between transmission PPG pre-training data and Empatica reflective PPG, the limited temporal context of short raw PPG windows, the difficulty of extracting reliable morphology from 64 Hz wrist PPG, and the known susceptibility of raw wrist PPG to motion artifacts. In particular, feature-extraction tools such as pyPPG (Goda et al., 2024) are designed around the recovery of morphological landmarks, which may be difficult to estimate robustly from lower-resolution wearable PPG.

The degradation was not uniform across sleep stages. The supervised PPG+IMU model achieved per-stage F1 scores of 0.760 for Wake, 0.650 for Light, 0.650 for Deep, and 0.780 for REM. Fine-tuned PaPaGei, by contrast, reached 0.470 for Wake, 0.190 for Light, 0.220 for Deep, and 0.440 for REM. The largest declines occurred for Light and Deep sleep, whereas Wake and REM retained roughly half of the supervised model's performance. One possible explanation is that Wake and REM produce larger physiological changes in heart rate, movement, and autonomic state that remain partially detectable from noisy wrist PPG, while distinguishing NREM sub-stages depends on subtler cardiac signatures that may require higher-fidelity preprocessing, longer temporal context, or complementary modalities such as EDA or EEG.

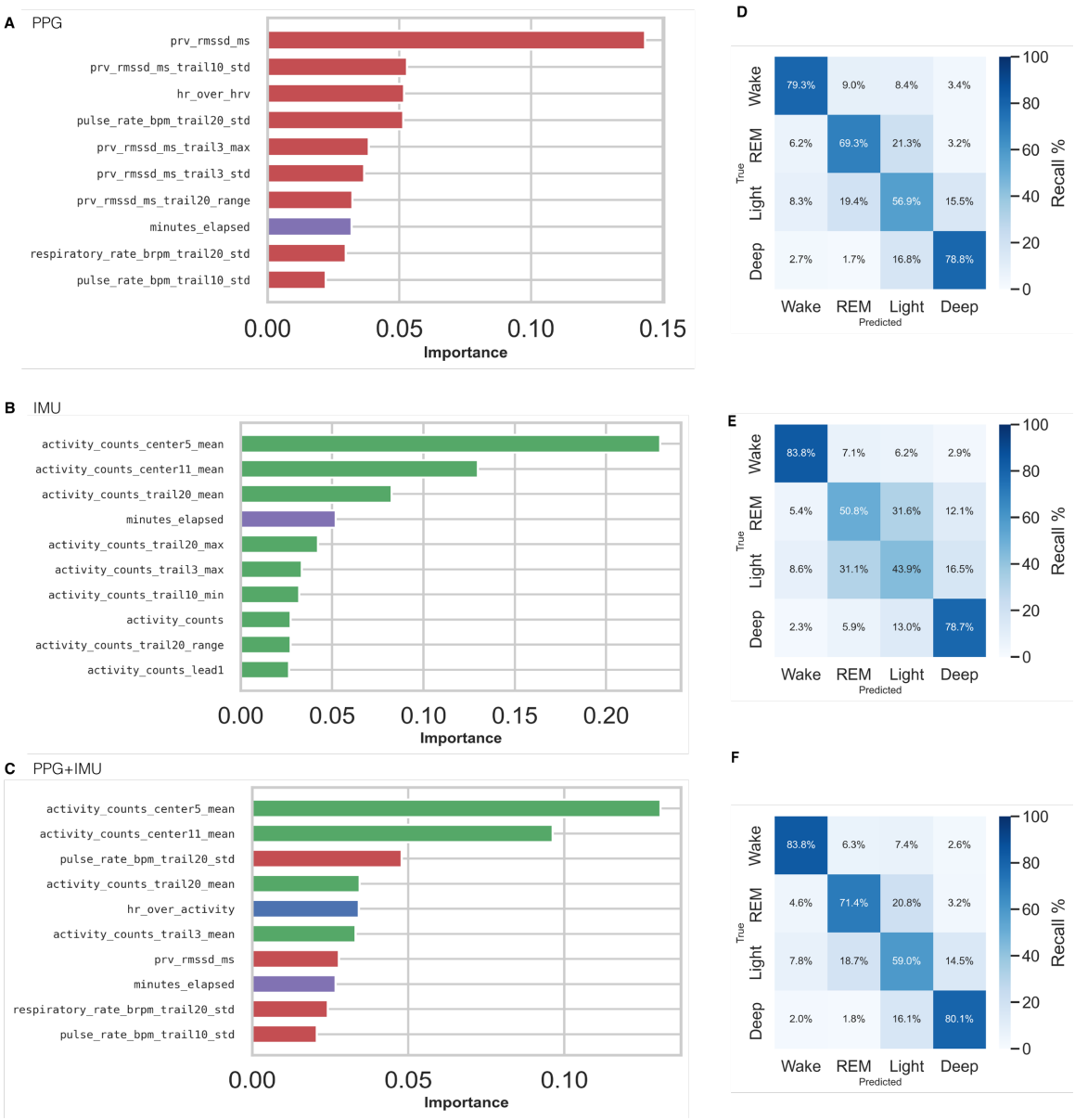



Figure Q.1: **PPG and IMU capture complementary aspects of sleep staging, with comparable overall performance.** (A–C) Feature importance shows that PPG emphasizes heart rate variability and pulse dynamics, while IMU captures activity-driven signals; in the combined model, both contribute meaningfully. (D–F) Confusion matrices for PPG-only, IMU-only, and PPG+IMU models show similar overall performance, with PPG better resolving REM and IMU excelling at sleep–wake discrimination. The combined model leverages both strengths, yielding improved balance across sleep stages.

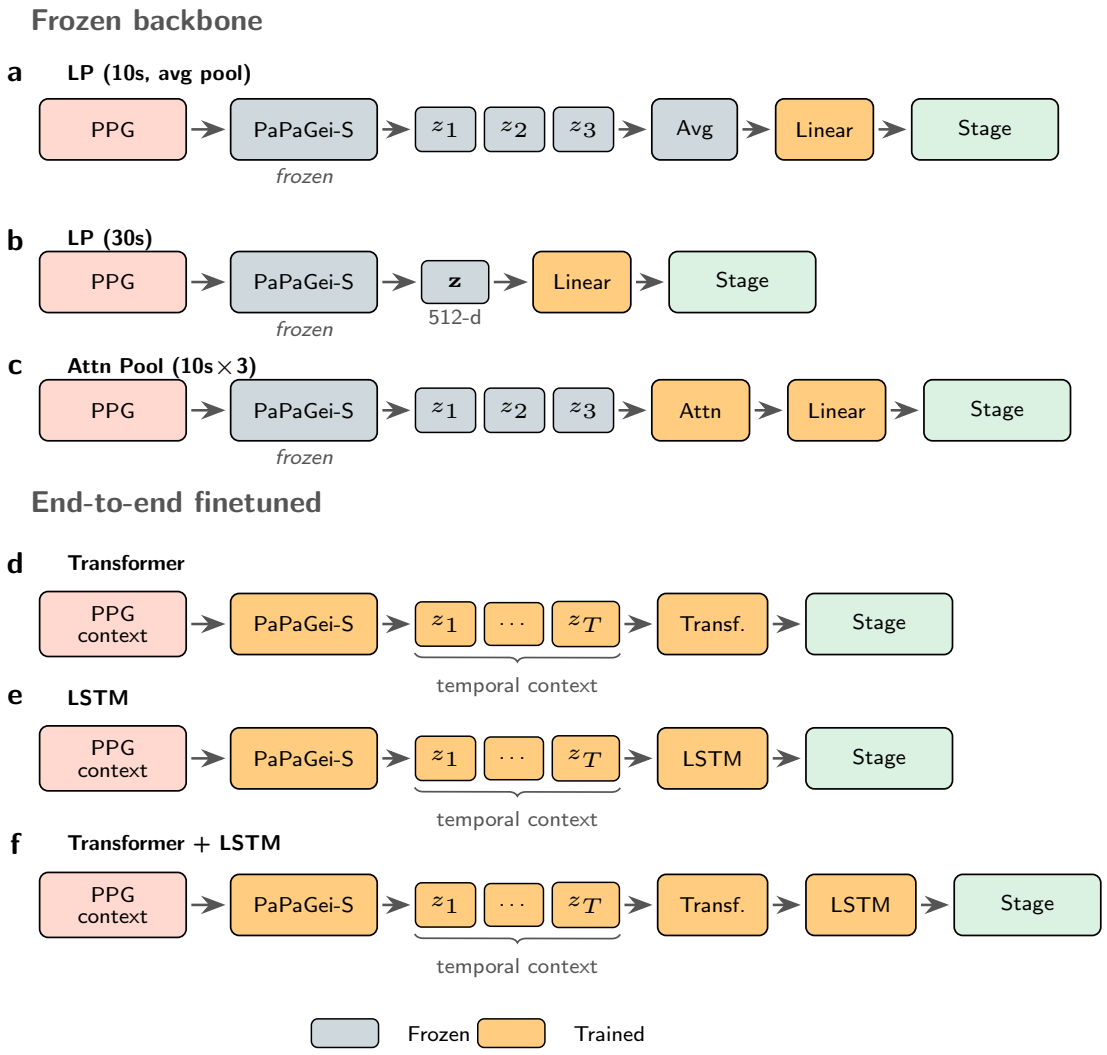



Figure Q.2: **PaPaGei PPG foundation model adapted for wrist-worn sleep staging via fine-tuning.** Modification of the PaPaGei foundation model (Pillai et al., 2025) for fine-tuning.

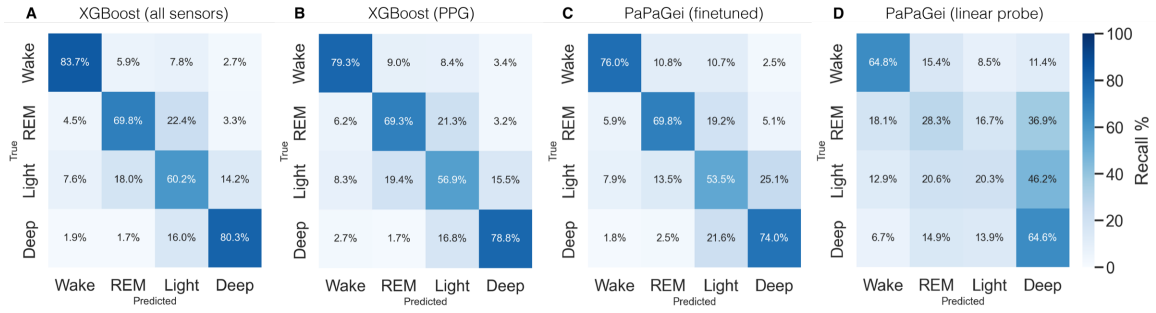



Figure Q.3: **PPG-based models approach multimodal performance, while PPG foundation models require adaptation.** Confusion matrices compare (A) XGBoost with all sensors, (B) XGBoost with PPG-only, (C) fine-tuned PaPaGei, and (D) linear-probe PaPaGei. PPG-only models achieve performance close to multimodal models, indicating that PPG captures most of the sleep staging signal. Fine-tuned PaPaGei partially recovers performance, while linear probing performs poorly, highlighting both domain shift and the need for task-specific adaptation.

Table Q.1: **Accuracy, Cohen's $\kappa$, and Macro F1 for wrist-worn sleep staging models.** Mean ± std across 5 leave-subjects-out folds. XGBoost uses handcrafted features from Empatica summarized data (1-min resolution) aligned to 30-s DREEM epochs. PaPaGei-S is an end-to-end finetuned PPG foundation model with Transformer+LSTM temporal aggregation over 60-epoch context windows. Classes: Wake, Light (N1+N2), Deep (N3), REM.

| Method | Modality | Accuracy | Cohen's $\kappa$ | Macro F1 |
|---|---|---|---|---|
| XGBoost | PPG + IMU + EDA | $0.703 \pm 0.013$ | $0.586 \pm 0.016$ | $0.710 \pm 0.013$ |
| | PPG + IMU | $0.700 \pm 0.012$ | $0.584 \pm 0.015$ | $0.708 \pm 0.012$ |
| | PPG + EDA | $0.685 \pm 0.015$ | $0.561 \pm 0.018$ | $0.689 \pm 0.014$ |
| | PPG | $0.678 \pm 0.014$ | $0.554 \pm 0.018$ | $0.683 \pm 0.015$ |
| | IMU + EDA | $0.611 \pm 0.004$ | $0.462 \pm 0.009$ | $0.620 \pm 0.008$ |
| | IMU | $0.595 \pm 0.007$ | $0.446 \pm 0.012$ | $0.608 \pm 0.009$ |
| | EDA | $0.532 \pm 0.009$ | $0.352 \pm 0.012$ | $0.525 \pm 0.012$ |
| PaPaGei-S | PPG | $0.640 \pm 0.015$ | $0.493 \pm 0.015$ | $0.647 \pm 0.015$ |

## Supplementary R. External validation on the DREAMT cohort

To test whether in-domain fine-tuning solves the domain-shift problem or merely fits the adaptation cohort, we evaluated cross-cohort transfer between SSP and DREAMT (Wang et al., 2024). DREAMT contains 100 adults (median age 59 years) with one overnight gold-standard PSG recording per participant, expert-scored sleep-stage labels, and simultaneous wearable PPG from the Empatica E4. The task remains sleep staging from physiological signals, but the population, acquisition setting, and wearable data distribution differ substantially from SSP.

As shown in Table R.1, in-domain performance does not translate into reliable external transfer. Models evaluated within the same cohort substantially outperform models transferred across cohorts. SSP→SSP reaches $\kappa = 0.555$ whereas SSP→DREAMT falls to 0.220, and DREAMT→DREAMT reaches 0.325 whereas DREAMT→SSP falls to 0.109. The SSP fine-tuned PaPaGei transfers worst of all ($\kappa = 0.069$), with a dominant failure mode of severely overpredicting deep sleep, assigning 63.5% of DREAMT epochs to deep sleep against a true rate of only 3.4%. A model can therefore improve on its adaptation cohort while still failing under external validation, which is why the benchmark is not invalidated by the observation that in-domain fine-tuning recovers performance; rather, it quantifies when such adaptation fails to generalize.

Table R.1: **Cross-cohort sleep-staging transfer between SSP and DREAMT.** Within-cohort cross-validation versus cross-cohort transfer. In-domain performance does not translate into reliable external transfer; the SSP fine-tuned PaPaGei collapses onto deep sleep on DREAMT.

| Train → Test | Accuracy | Balanced Acc. | Macro F1 | Cohen's $\kappa$ |
|---|---|---|---|---|
| SSP → SSP *(within-cohort CV)* | 0.680 | 0.710 | 0.685 | 0.555 |
| DREAMT → DREAMT *(within-cohort CV)* | 0.665 | 0.404 | 0.421 | 0.325 |
| SSP → DREAMT *(multimodal)* | 0.449 | 0.532 | 0.388 | 0.220 |
| DREAMT → SSP *(multimodal)* | 0.423 | 0.365 | 0.277 | 0.109 |
| SSP → DREAMT *(PaPaGei fine-tuned)* | 0.164 | 0.348 | 0.188 | 0.069 |