NeuroAtlas: Benchmarking Foundation Models for Clinical EEG and Brain-Computer Interfaces

Konstantinos Kontras[1,2,*], Trui Osselaer[1,*], Stylianos G. Mouslech[1,*], Angeliki-Ilektra Karaiskou[1,*], Guido Gagliardi[1,*], Thomas Strypsteen[1], Mohammad Hossein Badiei[1], Anku Rani[2], Maarten Vanmarcke[1], Miguel Bhagubai[1], Chanakya Ekbote[2], Jaedong Hwang[2], Christos Chatzichristos[1], Paul Pu Liang[2], Maarten De Vos[1]

[1]KU Leuven, [2]MIT, [*]Equal contribution

Foundation models (FMs) promise to extract unified representations that generalize across downstream tasks. They have emerged across fields including electroencephalography (EEG), but it is less clear how effective they are in this particular field. Published evaluations differ in datasets, in the EEG-specific preprocessing that might influence reported results, and in the reported metrics frequently obscuring the clinical relevance in EEG. We introduce NeuroAtlas, the largest EEG benchmark to date: 42 datasets and ∼260k hours covering clinical EEG (epilepsy, sleep medicine, brain age estimation) and brain-computer interfaces, and include multiple datasets per task along with bespoke clinical evaluation metrics. Besides evaluating EEG-FMs with respect to supervised baselines, we present results from generic time-series FMs. We report three findings. First, EEG-specific FMs do not consistently outperform time-series FMs which have neither EEG-focused architectures nor been pretrained on EEG. Second, standard machine learning metrics are insufficient to assess clinical utility: thus, we thoroughly evaluate more appropriate measures such as the quality of event-level decision-making, hypnogram-derived features, and the brain-age gap in the domains of epilepsy, sleep and brain age respectively. Third, model rankings and performance can vary substantially within domains. We conclude that pretrained models perform largely on par, with only narrow advantages for a few, and that current models do not yet deliver on the promise of an out-of-the-box unified EEG model. NeuroAtlas exposes this gap and provides the datasets and metrics for the next generation of unified EEG FMs.

Publication reference: This work is available as an arXiv preprint: arXiv:2605.14698.
Keywords: AI for Health; EEG, Foundation Models, Clinical Machine Learning, Brain-Computer Interfaces

Correspondence:
`konstantinos.kontras@kuleuven.be`

1 Introduction

Electroencephalography (EEG) provides a non-invasive and affordable assessment of brain function, used across a variety of clinical domains including epilepsy, sleep medicine, intensive care monitoring, and for steering brain-computer interfaces (BCIs). Each application is served today by a separate pipeline, typically tuned to a specific cohort, site, and device. Foundation models (FMs) aim to tackle this fragmentation. Recently, several EEG FMs have been proposed with the promise to derive cross-domain generalizable embeddings from a single backbone by training on massive and diverse EEG datasets. In other fields there have been notable successes in off-the-shelf performance [65, 14], raising the question: how close are current EEG-FMs to fulfilling their potential?

Existing evaluations give an incomplete answer. Most benchmarks reduce each domain to a single dataset, a single metric, and a single family of models [93, 35], leaving three central aspects of EEG-FM behavior untested. *First, intra-domain generalization.* Performance varies substantially across datasets within a single domain due to differences in acquisition device, electrode layout, referencing scheme, artifact profile, and population phenotype, and a credible benchmark should include multiple datasets per domain that span this variation. *Second, metrics that reflect clinical relevance.* Aggregate metrics such as AUROC or accuracy can hide failure modes that determine real-world usability, so each domain demands bespoke, clinically grounded evaluation. *Third, comparison against model alternatives.* Existing benchmarks typically compare EEG-FMs only against each other, leaving

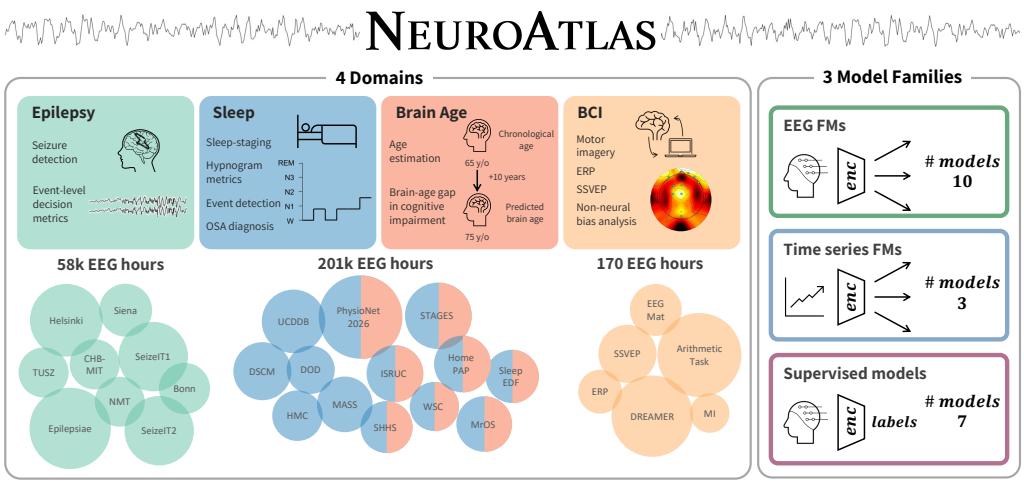

Figure 1: **NeuroAtlas overview.** NeuroAtlas spans four EEG domains (epilepsy, sleep, brain age, BCI) 42 datasets and $\sim 260k$ hours of EEG, and benchmarks three model families on each. Per-domain task list and dataset coverage, with total EEG hours per domain (epilepsy: 58k h; sleep: 201k h, shared with brain age; BCI: 170 h). Each domain is evaluated with both standard and clinically grounded tasks: seizure detection with event-level decision metrics (epilepsy); sleep staging, hypnogram-derived features, event detection, and OSA diagnosis (sleep); chronological age estimation and the brain-age gap as a marker of cognitive impairment (brain age); and motor imagery, ERP, SSVEP, and non-neural confound analysis (BCI). Three model families compared under a unified frozen-encoder probe: 10 EEG foundation models, 3 time-series foundation model families, and 7 task-specific supervised baselines.

the value of EEG-specific pretraining unverifiable against time-series FMs (TS-FMs) and models pretrained with supervision on other EEG datasets (referred to throughout as *Supervised-Pre* models). We introduce NeuroAtlas, advancing EEG-FM evaluation on three fronts:

1. **Comprehensive data coverage.** NeuroAtlas comprises 42 datasets and $\sim$260k hours of EEG across seizure detection, sleep staging, BCI, and brain-age prediction (a new benchmark domain), with multiple datasets per domain spanning varied acquisition protocols, populations, and sites to expose intra-domain variability.

2. **Comprehensive model coverage.** We evaluate 20 distinct models (29 size variants in total) spanning recent EEG foundation models, general-purpose time-series foundation models, and specialized supervised-pretrained baselines.

3. **Neurophysiologically and clinically grounded evaluation.** Each domain is evaluated both with standard metrics and bespoke domain-specific measures that reflect real-world clinical utility: event-level sensitivity at fixed false-alarm rates for seizure detection, hypnogram-derived macrostructure for sleep, per-subject reliability for brain age, and adjustment for non-neural confounds in BCIs.

Three findings emerge consistently across NeuroAtlas. (i) Model rankings and absolute performance shift substantially across datasets within the same domain, making single-dataset comparisons unreliable predictors of broader behavior. (ii) EEG-FMs do not consistently outperform generic TS-FMs, raising the question of how much of the reported gain stems from EEG-specific architectural choices or pretraining on neurophysiological data. (iii) Aggregate metrics translate only partially into clinical usability, with gaps revealed by hypnogram structure, event-level decision quality, and individual-level reliability. We release NeuroAtlas in full, including dataset access scripts, preprocessing pipelines, model configurations, and evaluation code, to support reproducible progress.

2 Related Work

2.1 EEG Benchmarks and Evaluation Protocols

A number of recent benchmarks have delivered tremendous efforts to evaluate EEG-FMs across multiple domains: AdaBrain-Bench [91], (13 datasets, 4 EEG-FMs) , EEG-Bench [35] (14 datasets, 3 EEG-FMs), EEG-FM-Benchmark [49] (13 datasets, 12 EEG-FMs), EEG-FM-Bench [93] (14 datasets, 7 EEG-FMs, 2 general-purpose time-series FMs), ST-EEGFormer [95] (10 datasets, 5 EEG-FMs) and Brain4FMs [72] (18 datasets, 15 brain FMs). A recurring limitation among many of these benchmarks is that, while they aim for a wide array of tasks, the performance within distinct clinical domains is typically represented by a single cohort or dataset. In addition, as observed by Kuruppu et al. [44], the evaluation protocols employed by the EEG-FM papers themselves make direct comparison impossible. With the exception of TUAB and TUEV, there is a significant heterogeneity in the selected downstream tasks across these works. Plus, only a limited number of these works extensively investigated generalization through out-of-distribution evaluation without fine-tuning, using e.g. linear probing or few-shot evaluation. Furthermore, the EEG-FM benchmarks typically report a fixed set of evaluation metrics (e.g., balanced accuracy and AUROC for classification tasks), disregarding the clinical meaning of different metrics across domains.

2.2 Foundation Models

EEG-FMs such as BIOT [94], LaBraM [31], EEGPT [86], CBraMod [87], NeuroLM [30], and REVE [59] aim to learn generalizable representations or embeddings from large-scale, heterogeneous, unlabeled EEG data. They are typically pretrained through self-supervised learning, with the pretext tasks mainly consisting of either masked signal reconstruction in the temporal and/or spectral domain or contrastive learning. On the other hand, general-purpose TS-FMs such as Chronos [4], MOMENT [24], and Moirai [89] are pretrained on diverse temporal corpora and target forecasting and related tasks. They use no EEG-specific architecture or pretraining data, but recent evidence shows EEG-FMs might be competitive on specific clinical tasks: TimesNet [90] on sleep staging in the FoME [73] evaluation, MOMENT on seizure detection in the BrainWave [100] evaluation, and PatchTST [56] on forecasting and imputation in the Brant evaluation [44]. This raises the question of how much of the reported gain on EEG-FM benchmarks comes from EEG-specific design rather than from generic large-scale temporal pretraining, a question NeuroAtlas is designed to answer.

3 Benchmark Pipeline

NeuroAtlas asks whether the latent representations learned by pretrained models, regardless the pretraining strategy, encode the information required for clinically meaningful tasks across diverse EEG domains, in a sufficiently disentangled form to be recovered by a simple downstream readout. To answer this directly, we evaluate every model under a unified protocol: the pretrained backbone is treated as a fixed feature extractor, and a linear probe is fit per fold of each downstream task, with classifier, and evaluation held identical across model families. This isolates the quality of the pretrained representation itself from confounds such as fine-tuning, architecture-specific tuning, or task-specific heads.

Models and datasets NeuroAtlas evaluates three groups of models: EEG-FMs, TS-FMs and Supervised-Pre EEG models. This setup enables comparison between EEG self-supervised pretraining, generic temporal representation learning, and task-specific supervised training. A full of the models is in Appendix A.2. On top of these pretrained models, we evaluate CBraMod [87] with randomly initialized weights as a unified random-init reference, isolating the contribution of pretraining from architecture and probe. Furthermore, the benchmark spans 42 datasets distributed over the domains of epileptic seizure detection, sleep medicine, brain age prediction and BCI, an overview of which can be found in Appendix B.

Preprocessing Harmonizing preprocessing in EEG is a challenging task. Not only have the different application domains typically operated under different electrode configurations, channel referencing, sampling rates and frequency filtering, but even the models have been pretrained under different specific preprocessing conditions. In NeuroAtlas, we have maintaining the broaded preprocessing by

only applying a highpass filter at 0.5Hz and a notch filter at the relevant power line noise to mitigate power line interference. Then, to match the in-distribution data of each EEG model, the recordings are resampled to the training sampling rate of each model. individual EEG-FMs. Dataset-model channel matching information and further details on for this step are in Appendix A.2.

Evaluation We evaluate EEG-FMs, TS-FMs, and supervised models based on their ability to encode generalizable, task-relevant features both within and across domains. To this end, we consistently apply a linear probing approach on the extracted, frozen embeddings across all the domains, without any fine-tuning of the encoding backbones (neither FMs nor supervised models). In the rare cases that a downstream dataset overlaps with a dataset used for pretraining the supervised models, this is clearly indicated in the associated results section. Furthermore, every evaluation was carried out in a purely cross-subject manner, without any fine-tuning on specific subject characteristics. The measured performance metrics are unique to each domain to ensure their clinical relevance to the specific applications.

4 Experiments

4.1 Epilepsy

Epileptic seizure detection is among the most frequently reported clinical tasks in the EEG-FM literature, yet evaluations share two limitations that obscure what is being measured. (i) Typically, the task is performed by segmenting each recording into non-overlapping 10-second windows [30]; the model produces a binary classification per window (seizure vs. non-seizure), and performance is summarized by aggregating across windows. Window-level AUROC measured in cases of extreme class imbalance can remain high even while the model is generating a significant amount of false alarms, which is unacceptable under the strict false-alarm constraints of clinical seizure detection [19]. (ii) The value of pretraining is realized when learned representations transfer to new tasks with minimal adaptation; current EEG-FM evaluations, however, do not isolate the contribution of pretraining from that of the downstream task-specific training, leaving the quality of the representation itself difficult to assess. NeuroAtlas addresses both by evaluating across seven datasets ($\sim$1,410 patients, $\sim$58,374 hours, $\sim$9,019 annotated seizures).

Evaluation. All models receive identical preprocessing and channel selection. We report event-level sensitivity over 0.1–100 FA/h as the area under the median (5-fold) sensitivity vs. $\log_{10}$(FA/h) curve [10] and for completeness the AUROC and balanced accuracy in App. D.2. For every model we perform the same steps in terms of preprocessing, channel selection and window cropping (10s). We report (i) window-level AUC and balanced accuracy and (ii) event-level sensitivity averaged over the clinically relevant 0.1–100 FA/h range under any-overlap event scoring [19, 10], computed as the area under the mean (5-fold) sensitivity vs. $\log_{10}$(FA/h) curve.

Results. Event-level performance varies strongly across datasets (Figure 2): the dataset-best Event-Sens@FA AUC ranges from ≈ 0.50 on Helsinki to ≈ 0.74 on SeizeIT1, the leading model alternates between NeuroLM, REVE, and SleepFM, and within-dataset spreads of ≈ 0.2–0.4 AUC make these rank changes consequential. All three pretrained families exceed the CBraMod random-init reference on every dataset, but per-model the margin lies within one fold-standard-deviation of random init on the majority of datasets. A one-sided binomial top-3 win test (Appendix D.1.4, Table 3) identifies only NeuroLM ($6/7$ datasets, $p < 10^{-3}$) and REVE ($5/7$ overall, $6/7$ within EEG-FMs; both $p \leq 0.002$) as exceeding chance; the remaining EEG-FMs do not, even within their own family. Pretraining therefore captures something beyond architecture and probe, but the gain is concentrated in two models and is small enough (1 std) that similar performance to new dataset is not yet warranted. Three further analyses refine these conclusions. First, metric choice matters: AUROC and Event-Sens@FA produce different rankings on every dataset (Spearman $\rho \in [0.61, 0.81]$; Appendix D.1.2, Figure 7), most visibly on TUSZ where NeuroLM is rank-1 by AUROC but is displaced by Seizure-TF on Event-Sens@FA. Second, ILAE-2017-type-stratified [5] probability distributions (App. D.1.1, Figure 6) expose dataset-dependent failure modes: the best EEG-FM concentrates near 1 on focal, generalised and focal-to-bilateral events on TUSZ and SeizeIT2 but shows long tails on Epilepsiae focal events. Third, on three recording-level datasets (Bonn, TUAB, NMT; App. D.1.3- Figure 8), REVE leads TUAB and NMT, and TS-FMs reach near-parity with EEG-FMs on NMT.

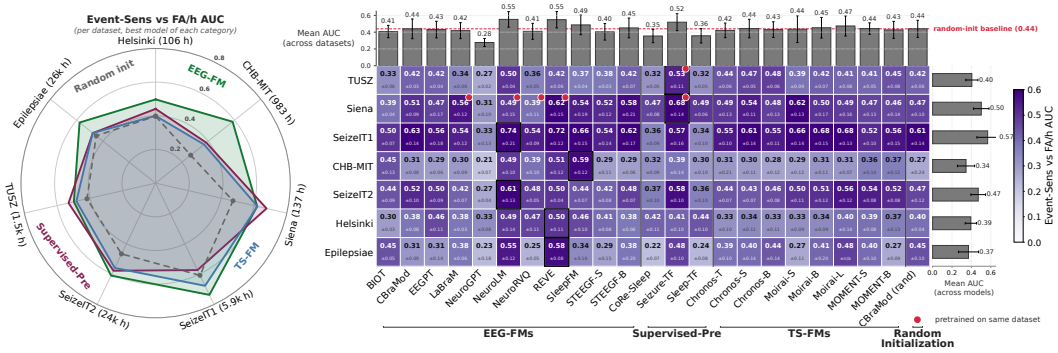

Figure 2: **Epilepsy: no model or family is consistently best across datasets, and most models perform on par with random initialization. (a)** Family-best Event-Sens@FA AUC across seven continuously-labeled datasets (labels include total EEG hours), as the area under the mean (5-fold) sensitivity vs. $\log_{10}$(FA/h) curve over 0.1–100 FA/h under any-overlap event scoring [19]. The leading EEG-FM differs across datasets and no family consistently dominates. **(b)** Per-(model, dataset) AUC for 23 models: EEG-FMs, Supervised-Pre, TS-FMs, CBraMod random-init. Cells show mean AUC (bold) and 5-fold std; black borders mark the dataset-best model; red dots flag pretraining/evaluation overlap. Grey bars: per-model means across datasets (top, red dashed: random-init baseline at 0.44) and per-dataset means across models (right). Most per-model means lie within ≈ 0.05 AUC of the 0.44 baseline; only NeuroLM and REVE exceed it at approx. 1 std (both 0.55), and the dataset-best alternates between NeuroLM, REVE and SleepFM.

4.2 Sleep-staging

Current EEG-FM sleep-staging benchmarks remain limited in both scale and clinical relevance. First, typically evaluations are conducted on small amounts of data in terms of both hours and number of patients [35, 72, 93, 96]. The previous most extensive benchmark included 558 patients from 3 datasets totaling 7.5k hours [91], making it difficult to assess the cross-dataset performance and to disentangle transferable representations from dataset-specific effects. Second, in clinical practice, sleep is interpreted through the hypnogram, the ordered sequence of stages (wake, REM, N1–N3) staged according to the American Academy of Sleep Medicine (AASM) criteria, from which clinical macrostructural features like REM latency and total sleep time (TST) are derived [84]. Furthermore, diagnosis is supported by event annotations such as microarousals, respiratory events, and limb movements. Beyond primary sleep disorders, such markers have also been linked to cognitive impairment and neurodegenerative diseases [16, 48]. However, EEG-FM's performance is predominantly reported as epoch-wise agreement with expert labels, which does not directly assess agreement on the sequential organization underlying macrostructural hypnogram features, used in clinical practice [30, 59]. Overall, this highlights the need for more comprehensive evaluation in terms of downstream cohorts and evaluated metrics.

Evaluation protocol We evaluate sleep representations across 15 datasets, representing $\sim$15.8k patients and $\sim$201k EEG hours, by splitting recordings in 30-second epochs, and extracting embeddings per epoch. Evaluation is conducted along four complementary axes. First, epoch-wise sleep staging is evaluated by window-level Cohen's κ. Second, we assess whether predicted hypnograms preserve macrostructural hypnogram features used in downstream clinical tasks (for a full overview see App. C.2). Third, to assess representation beyond the primary stages labels, we evaluate epoch-level detection of microarousals, detection of respiratory events and periodic limb movements. Finally, we evaluate diagnosis, we assess recording-level diagnosis of the primary sleep disorder obstructive sleep apnea (OSA) and cognitive impairment from sleep EEG.

Results We first examine epoch-wise sleep-staging agreement (Cohen's κ) across families and datasets (Figure 3A), where three patterns emerge. Performance varies strongly across datasets, with the best per-dataset κ ranging from 0.48 to 0.82 (per-model results in App. D.3). Across families, Supervised-Pre models achieve the strongest overall performance (binomial test, Table 5), but this advantage is largely attributable to their sequence-based architectures, which encode inter-epoch

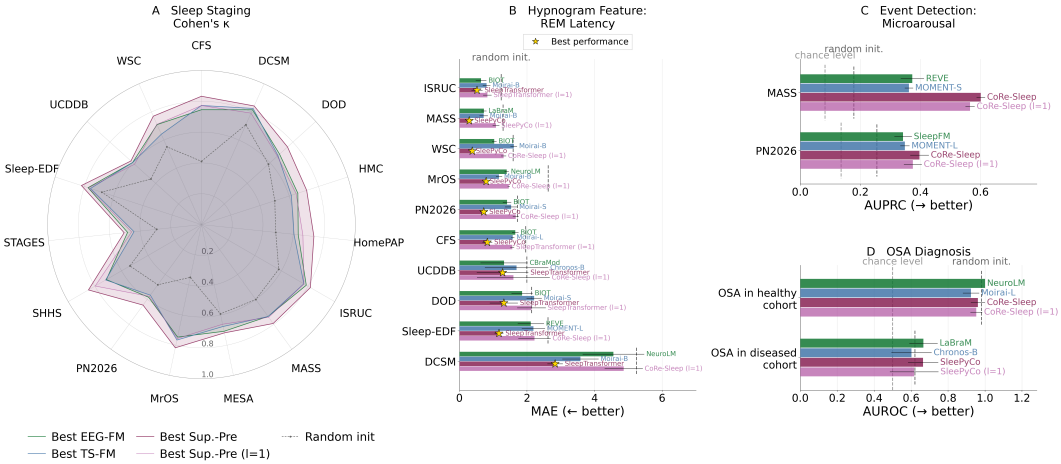

Figure 3: **Sleep: Performance is strongly dataset-dependent across tasks.** Metrics are shown for the best-performing model per category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models with and without temporal context (sequence length l=1), and a random-initialized reference. **(a)** Cohen's κ sleep staging agreement with ground-truth labels. Model categories achieve similar performance when excluding the effects of sequence modeling. **(b)** MAE between ground truth and predicted REM latency. Performance on sleep staging agreement does not always correspond to performance on hypnogram features. **(c)** AUPRC for microarousal detection, for events with a minimal duration of 3 seconds. Chance baseline is indicated by prevalence. While all model categories perform above baseline, FMs do not reach Supervised-Pre performance. **(d)** AUROC for diagnosis of OSA compared to healthy patients, and in a cohort with other disorders. Chance baseline is indicated at 0.5. While OSA is detected in a healthy cohort, performance degrades in the presence of other disorders.

temporal context. An ablation with sequence length 1 (Sup-Pre $(\ell{=}1)$ in Figure 3) removes most of this gap, consistent with the well-known role of sequence modeling in sleep staging [61], a property not explicitly encoded in current FM architectures. Once sequence context is removed, Supervised-Pre $(\ell{=}1)$ match TS-FMs and EEG-FMs, suggesting that EEG-specific pretraining is less decisive for sleep staging than architectural design factors. Within EEG-FMs, NeuroLM and REVE consistently rank among the top three across datasets (binomial test, Table 5).

Beyond epoch-wise staging, we evaluate three clinical axes: hypnogram-derived features, event detection, and recording-level diagnosis. *Hypnogram features* (Figure 3B shows REM latency, used for the diagnosis of hypersomnolence [7]; full feature list in App. D.3) show substantial dataset variability, and crucially do not track Cohen's κ: as illustrated by DCSM (comparing Figure 3A and B), strong staging agreement does not imply accurate feature recovery. Reporting both metrics offers a more clinically grounded view of model performance. For *event detection* (Figure 3C, microarousals), all model categories exceed the prevalence baseline in AUPRC, but inter-dataset variability remains pronounced; even without inter-epoch context, Supervised-Pre models outperform FMs, indicating that current FMs' representations encode microarousal events incompletely. *Respiratory and limb-movement detection* are reported in App. D.3. Finally, recording-level *OSA diagnosis* succeeds when patients are compared to healthy controls but degrades when other conditions such as snoring or affective disorders are included; EEG-only embeddings therefore insufficiently encode disorder-specific features, and accurate OSA diagnosis should incorporate respiratory and oxygen-saturation channels. Cognitive-impairment diagnosis is reported in App. D.3.

4.3 Brain Aging Estimation for Neurodegenerative Assessment

Chronological age influences the spectral and temporal dynamics in EEG, particularly during sleep, where changes in micro- and macrostructure provide a window into neurophysiological ageing [80]. Predicting age from sleep EEG can therefore reveal whether learned representations capture broad, physiological patterns rather than only task-specific discriminative cues. In this sense, sleep-EEG age prediction provides a continuous regression task tied to known neurophysiological changes across the lifespan[22], but it has not been included in most existing EEG-FM benchmarks. Age prediction

itself is primarily useful as a representation-level probe. Its clinical relevance appears more promising through the "Brain Age Gap" (BAG) modelling framework, where a model learns the expected relationship between EEG and age from healthy individuals. The brain-age gap (BAG) is defined as the difference between predicted and chronological age. Positive values indicate older-appearing EEG activity and is commonly interpreted as accelerated brain ageing BAG provides a general, label-agnostic marker of brain health and has been associated with outcomes such as cognitive impairment, dementia risk, and mortality [77]. This makes brain age a useful benchmark for EEG-FMs: it tests whether representations capture continuous neurophysiological patterns and whether this signal can be translated into a clinically motivated digital biomarker.

Evaluation We evaluate brain ageing in two complementary ways. First, we test age encoding by predicting chronological age from sleep EEG, reporting Pearson's r for subject-level age ordering and MAE for error in years. Because raw MAE depends on cohort age range, we also report improvement over a train-mean baseline, which predicts the training-set mean age for all subjects. Improvement over this baseline aims to show improvement beyond learning the mean age of the cohort. Second, we assess whether the learned age signal can be used as a clinically meaningful marker. For this, models are trained on healthy subjects to learn a reference pattern of healthy ageing, and then applied to an independent holdout set containing age-matched healthy and cognitively impaired (CI) subjects. We use the resulting brain-age gap to test whether the models indicate whether CI subjects show older-appearing EEG than healthy subjects of the same chronological age. We evaluate this both by comparing mean BAG between groups, and by testing whether BAG separates CI from healthy participants . This analysis is applied to all our model categories. More details on the metrics and the models can be found in App. C.3

Results Across cohorts, the feature-based brain-age model of Sun et al. [79] provides the strongest task-specific reference, achieving the largest MAE improvement over the dummy baseline and the highest Pearson correlation in most datasets (Figures 4A–B). To quantify whether this apparent consistency exceeds chance, we counted how often each model ranked among the top three within each dataset and compared this count against a chance top-3 rate using a one-sided exact binomial test 11, 12. Across all 26 models, Sun et al. is the only model significant for both MAE improvement and Pearson r. Among EEG-FMs specifically, performance is more fragmented. In the EEG-FM-only analysis, NeuroRVQ-EEG shows significant top-3 recurrence for MAE, while EEGPT and NeuroRVQ-EEG have the highest Pearson r recurrence without reaching significance. Thus, while EEG-FMs encode useful age-related information, no single EEG-FM emerges as a stable winner across both metrics and cohorts. The benchmark instead reveals a heterogeneous landscape: task-specific brain-age supervision is the most consistent reference, supervised sleep-staging models remain competitive, and different foundation models perform well in different datasets. Further ablations on the performance is discussed in App. D.4

FM-derived BAG Shows Limited Clinical Transfer PhysioNet Benchmark We next evaluate BAG on an age-matched PhysioNet 2026 healthy/cognitive impairment (CI) holdout. BAG-based separation remains weak overall. The supervised Sun et al. model shows the clearest group-level shift ($\Delta\mathrm{BAG} = +1.99$ years; Figure S19) and reaches AUROC $= 0.592$ (Figure 4C). NeuroLM-VQ is the only EEG-FM, matching this AUROC with a similar positive BAG shift ($+1.72$ years), but both remain far from robust healthy and CI separation. Sleep-staging models and smaller MOMENT variants show weak positive separation, whereas most remaining models are close to chance. Thus, BAG transfer to healthy–CI discrimination remains limited. The unclear transfer from models fitting age to BAG remain consistent with literature, where the relationship between fitting the age distribution and the residual error of prediction as a digital biomarker remains debatable. [32]

4.4 Brain Computer Interfaces (BCI)

BCI paradigms are well represented in EEG-FM benchmarks [94, 18, 30, 49, 95] with performance being reported on multiple datasets. The most extensive benchmark, EEG-FM-Benchmark [49], reports performance on 7 BCI datasets spanning Motor Imagery (MI), Event-Related Potentials (ERP) and Steady-State Visually Evoked Potential (SSVEP). This work further expands on these efforts by evaluating EEG-FMs on a diverse set of 17 datasets in total, adding also two cognitive tasks and one emotion recognition task. Additionally, we draw the attention to the fact that eye movements and gaze orientation can correlate with task conditions, causing models to rely on oculomotor activity instead

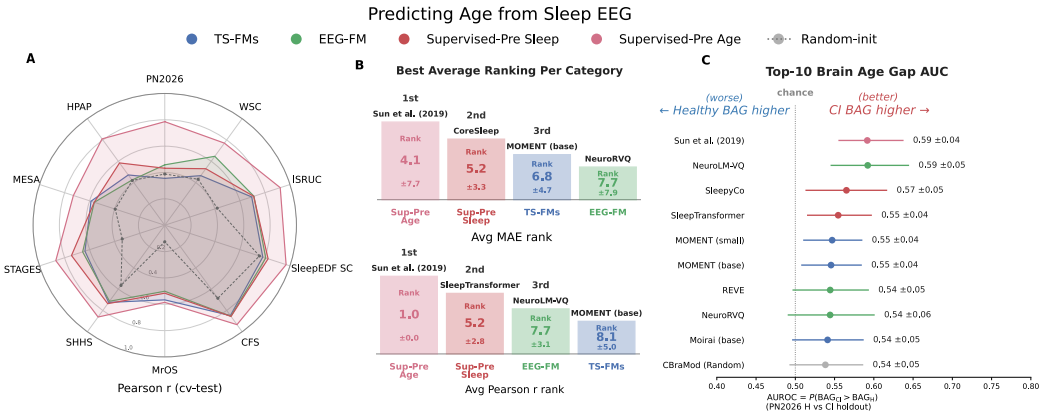

Figure 4: **Supervised Models outperform EEG-FMs for Brain Age Prediction, but performance does not translate to neurodegenerative assessment biomarker.** (A) Radar plot of performance (Pearson's r) of best model per category of models in different cohorts. Supervised Brain Age prediction outperforms the rest. (B) Best average ranking of the best model of each category for MAE and r metrics. (C) Top-10 AUROC for cognitive-impairment separation, using BAG (predicted minus chronological age).AUROC measures whether cognitively impaired subjects have higher BAG than age-matched healthy subjects. Although several models show weak positive separation, performance remains modest across model categories, indicating that accurate age prediction does not necessarily yield a robust BAG-based biomarker.

of the intended neural signal. Low-frequency drifts can further contribute to signal non-stationarity over time [66] and post-cue saccades might obscure neural activity [6]. We investigate to what degree peformance is driven by such non-neural confounding factors.

Evaluation protocol The FM backbones remain frozen, and evaluation follows a leave-one-subject-out (LOSO) protocol to assess cross-subject generalization. The effect of non-neural confounding is assessed by comparing performance metrics with and without artifact correction preprocessing. We do so by applying a 4Hz highpass filter, a 40Hz lowpass and removing the first second post-cue in each trial, as proposed by [6]. Because decoding accuracies in finite-sample neural classification settings can substantially exceed theoretical chance levels purely by chance [17], we additionally report chance-corrected metrics and interpret performance relative to statistically meaningful baselines rather than raw accuracy alone. Further details can be round in App.C.4.

Results As shown in Fig. 5a, our investigation suggests that FMs are still prone to capturing non-neural confounding sources, such as eye gaze. In MI paradigms, not filtering or removing the first post-cue second significantly increases performance above statistically significant chance level for most EEG FMs, while confound filtering drops the performance below chance level. Similar analyses on the remaining paradigms did not further affect model performance, and are reported in the App.D.5. Furthermore, as shown in Fig. 5b, although linear probing often exceeds the normalized balanced accuracy chance level of 50%, it frequently remains below the statistically significant chance threshold under LOSO evaluation, suggesting that FMs still do not reliably extract generalizable task-relevant information across subjects. In terms of EEG-FM variability, the results demonstrate substantial differences in how well EEG-FMs generalize across BCI paradigms and datasets. While EEGPT and ST-EEG-S achieve the strongest performance on MI tasks, other models such as REVE and NeuroLM show superior performance on ERP and cognitive benchmarks, indicating that different EEG-FMs capture complementary neural representations. Importantly, no single model consistently outperforms others across all paradigms, suggesting that current unsupervised EEG representation learning approaches remain paradigm-dependent rather than universally transferable.

5 Discussion

Are EEG-FM models generally consistent within domains? The first goal of this benchmark was to provide an extensive analysis of cross-dataset performance within the domains of epileptic

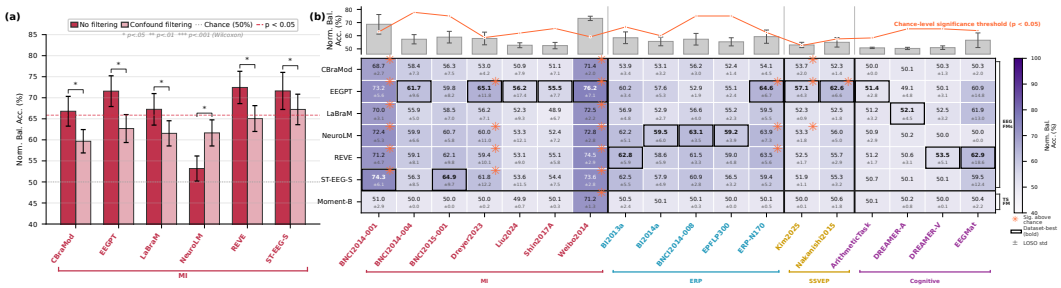

Figure 5: **BCI: MI performance is strongly driven by confounds** Linear probing performance of frozen EEG and time-series foundation models across 17 BCI datasets spanning four paradigms (MI, ERP, SSVEP, Cognitive). Metrics are reported as normalized balanced accuracy (%), where 50% equals chance regardless of the number of classes. The per-dataset significance threshold ($p < 0.05$, binomial test [17]) is shown as a dashed red line or orange markers. **(a)** Normalized balanced accuracy on Motor Imagery datasets comparing embeddings extracted without confound filtering and with confound filtering bars) for 6 EEG-FMs. Error bars denote the standard error of the mean across MI datasets; (Wilcoxon signed-rank test; $*\,p < .05$, $**\,p < .01$, $***\,p < .001$). Confound filtering significantly affect performance, suggesting that the evaluated models rely primarily on artifacts. **(b)** Per-dataset heatmap for 6 EEG-FMs and 1 TS-FM evaluated on confound filtered data. Cell color encodes normalized balanced accuracy (40–100%; darker = higher); each cell reports the score (bold for dataset-best) with LOSO standard deviation ($\pm$). Orange stars ($*$) mark results significantly above chance. Note that due to the varying number of classes and the low sample counts, the significance level is much higher than chance level and varies across datasets.

seizure detection, sleep staging, brain age prediction and BCI. While performances for all models naturally varied significantly between datasets of every domain, we investigated to what degree different domains have a consistent set of best performing EEG-FMs. In epilepsy, sleep and BCIs, Neuro-LM and REVE tend to end up consistently in the top-3 ranked EEG-FMs, while in the novel task (brain age), no clear top competitors emerge.

How do EEG-FMs compare to other model families? The linear probing performances on EEG-FMs and pre-trained supervised models vary significantly between different domains. In seizure detection, EEG-FMs consistently outperform their supervised counterparts with the only exceptions being the datasets that the supervised models were actually pretrained on. In sleep, EEG-FMs perform similarly to pretrained supervised models on a window-per-window level, but fall significantly behind when the supervised models are trained to operate in a sequence-to-sequence manner, which fall outside of the capabilities of linear probing. In brain age prediction, EEG-FMs perform worse on the task of chronological age prediction, yet perform on par when it comes to the clinically more relevant task of estimating the BAG. Finally, EEG-FMs struggle to reach motor imagery performance significantly above chance level. The comparison with general-purpose time-series FMs is more consistent across domains: we do not find evidence that EEG-FMs are able to consistently outperform time-series FMs, despite their apparent advantage in only having to model EEG.

The importance of clinically relevant metrics! Across domains, we observe that standard metrics insufficiently address clinical utility. In epilepsy, we observe that relying on AUROC for model selection leads to diverging results from using the more application-oriented EventSens@FA/h (Figure 7). In sleep, models with similar window-level κ-score show large differences in REM-latency (Figure 3B). Similarly in brain age, models with substantially different MAE on age estimation do not necessarily differ in BAG. Finally, in BCI, applying appropriate preprocessing to prevent non-neural confounding often makes the difference between a model performing above significance level or not.

Limitations and conclusions This benchmark still has some limitations. We have focused our efforts on a single evaluation protocol in linear probing, explicitly not including other paradigms such as fine-tuning, and few-shot learning. In addition, we were not able to include promising models such as BrainWave [100] due to their implementation not being fully available. Despite these limitations, NeuroAtlas significantly advances the state of EEG-FM benchmarking both in scale and rigorous evaluation metrics.

References

[1] Minkyu Ahn and Sung Chan Jun. Performance variation in motor imagery brain–computer interface: A brief review. *Journal of Neuroscience Methods*, 243:103–110, 2015.

[2] Diego Alvarez-Estevez and Roselyne Rijsman. Haaglanden Medisch Centrum sleep staging database. *PhysioNet*, March 2022. Version 1.1.

[3] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[4] Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

[5] GN Avakyan, DV Blinov, AV Lebedeva, SG Burd, and GG Avakyan. Ilae classification of the epilepsies: the 2017 revision and update. *Epilepsy and paroxysmal conditions*, 9(1):6–25, 2017.

[6] Stylianos Bakas, Siegfried Ludwig, Dimitrios A Adamos, Nikolaos Laskaris, Yannis Panagakis, and Stefanos Zafeiriou. Latent alignment in deep learning models for EEG decoding. *Journal of neural engineering*, 22(1):016047, 2025.

[7] Lucie Barateau, Fabio Pizza, Giuseppe Plazzi, and Yves Dauvilliers. Narcolepsy. *Journal of Sleep Research*, 31(4):e13631, August 2022.

[8] Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models. *arXiv preprint arXiv:2510.13068*, 2025.

[9] Miguel Bhagubai, Christos Chatzichristos, Lauren Swinnen, Jaiver Macea, Jingwei Zhang, Lieven Lagae, Katrien Jansen, Andreas Schulze-Bonhage, Francisco Sales, Benno Mahler, et al. Seizeit2: Wearable dataset of patients with focal epilepsy. *Scientific Data*, 12(1):1228, 2025.

[10] Miguel Bhagubai, Lauren Swinnen, Evy Cleeren, Wim Van Paesschen, Maarten De Vos, and Christos Chatzichristos. Towards automated seizure detection with wearable eeg–grand challenge. *IEEE Open Journal of Signal Processing*, 5:717–724, 2024.

[11] Miguel Bhagubai, Kaat Vandecasteele, Lauren Swinnen, Jaiver Macea, Christos Chatzichristos, Maarten De Vos, and Wim Van Paesschen. The power of ecg in semi-automated seizure detection in addition to two-channel behind-the-ear eeg. *Bioengineering*, 10(4), 2023.

[12] Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *J. Am. Geriatr. Soc.*, 59(12):2217–2225, December 2011.

[13] Robyn M Busch, Olivia Hogue, Abagail F Postle, and Darlene P Floden. Automated detection of cognitive impairment in clinical practice. *J. Neurol.*, 271(8):5187–5196, August 2024.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[15] 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, June 2015.

[16] Julie Anja Engelhard Christensen, Poul Jennum, Henriette Koch, Rune Frandsen, Marielle Zoetmulder, Lars Arvastson, Søren Rahn Christensen, and Helge Bjarrup Dissing Sorensen. Sleep stability and transitions in patients with idiopathic REM sleep behavior disorder and patients with Parkinson's disease. *Clinical Neurophysiology*, 127(1):537–543, January 2016.

[17] Etienne Combrisson and Karim Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 250:126–136, 2015. Cutting-edge EEG Methods.

[18] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[19] Jonathan Dan, Una Pale, Alireza Amirshahi, William Cappelletti, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Adriano Bernini, Luca Benini, Sándor Beniczky, David Atienza, and Philippe Ryvlin. Szcore: Seizure community open-source research evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms. *Epilepsia*, 66(S3):14–24, 2024.

[20] Silvia López de Diego. *Automated interpretation of abnormal adult electroencephalograms*. Temple University, 2017.

[21] Paolo Detti. Siena Scalp EEG Database. *PhysioNet*, August 2020. Version 1.0.0.

[22] Denis A. Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from m/eeg resting-state signals. *NeuroImage*, 262:119521, 2022.

[23] Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online erd-bci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012.

[24] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

[25] Pierre Guetschel, Théodore Papadopoulo, and Michael Tangermann. Neural network transfer learning with fast calibration for mental imagery decoding. In *Proceedings of the 10th International Brain-Computer Interface Meeting 2023*, June 2023.

[26] Antoine Guillot, Fabien Sauvet, Emmanuel H. During, and Valentin Thorey. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(9):1955–1965, 2020.

[27] John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, June 2010. Version 1.0.0.

[28] Elisabeth R M Heremans, Astrid Devulder, Pascal Borzée, Rik Vandenberghe, François-Laurent De Winter, Mathieu Vandenbulcke, Maarten Van Den Bossche, Bertien Buyse, Dries Testelmans, Wim Van Paesschen, and Maarten De Vos. Wearable sleep recording augmented by artificial intelligence for alzheimer's disease screening. *NPJ Aging*, 11(1):34, May 2025.

[29] Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient p300-based brain–computer interface for disabled subjects. *Journal of Neuroscience methods*, 167(1):115–125, 2008.

[30] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

[32] Robert J. Jirsaraie, Alexander J. Gorelik, M. M. Gatavins, Denis A. Engemann, Ryan Bogdan, Deanna M. Barch, and Aristeidis Sotiras. A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility. *Patterns*, 4(4):100712, 2023.

[33] Emily S Kappenman, Jaclyn L Farrens, Wendy Zhang, Andrew X Stewart, and Steven J Luck. Erp core: An open resource for human event-related potential research. *NeuroImage*, 225:117465, 2021.

[34] Emily S. Kappenman and Steven J. Luck, editors. *The Oxford Handbook of Event-Related Potential Components*. Oxford Library of Psychology. Oxford University Press, 2011. Online edition, Oxford Academic, 18 September 2012.

[35] Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A Benchmark for EEG Foundation Models in Clinical Applications, 2025. Version Number: 1.

[36] Stamos Katsigiannis and Naeem Ramzan. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

[37] Bastiaan Kemp, Aeilko Zwinderman, Bert Tuk, Hilbert Kamphuisen, and Josefien Oberyé. The sleep-EDF database [expanded], 2018.

[38] Sirvan Khalighi, Teresa Sousa, José Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124, 11 2015.

[39] Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The nmt scalp eeg dataset: An open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

[40] Heegyu Kim, Kyungho Won, Minkyu Ahn, and Sung Chan Jun. A 40-class ssvep speller dataset: Beta range stimulation for low-fatigue bci applications. *Scientific Data*, 12(1):1751, 2025.

[41] Juliane Klatt, Hinnerk Feldwisch-Drentrup, Matthias Ihle, Vincent Navarro, Markus Neufang, Cesar Teixeira, Claude Adam, Mario Valderrama, Catalina Alvarado-Rojas, Adrien Witon, Michel Le Van Quyen, Francisco Sales, Antonio Dourado, Jens Timmer, Andreas Schulze-Bonhage, and Bjoern Schelter. The epilepsiae database: An extensive electroencephalography database of epilepsy patients. *Epilepsia*, 53(9):1669–1676, 2012.

[42] Konstantinos Kontras, Christos Chatzichristos, Huy Phan, Johan Suykens, and Maarten De Vos. Coresleep: A multimodal fusion framework for time series robust to imperfect modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:840–849, 2024.

[43] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI using dry EEG electrodes Dataset (bi2014a)*. PhD thesis, GIPSA-lab, 2019.

[44] Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. Eeg foundation models: a critical review of current progress and future directions. *Journal of neural engineering*, 23(2):021001, 2026.

[45] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[46] Seongju Lee, Yeonguk Yu, Seunghyeok Back, Hogeon Seo, and Kyoobin Lee. SleePyCo: Automatic sleep scoring with feature pyramid and contrastive learning. *Expert Systems with Applications*, 240:122551, 2024.

[47] Yun Ji Lee, Jae Yong Lee, Jae Hoon Cho, and Ji Ho Choi. Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine*, 18(1):193–202, January 2022.

[48] Sijie Li, Bian Liu, Qing-hao Li, Yan Zhang, Haihua Zhang, Shan Gao, Longcai Wang, Tao Wang, Zhifa Han, Guiyou Liu, and Kun Wang. Evaluating the Bidirectional Causal Association Between Daytime Napping and Alzheimer's Disease Using Mendelian Randomization. *Journal of Alzheimer's Disease*, 89(4):1315–1322, October 2022.

[49] Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems. *arXiv preprint arXiv:2601.17883*, 2026.

[50] Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, Yuheng Chen, Shaojie Hou, Xiaolian Zhu, and Dongrui Wu. MIRepNet: A pipeline and pre-trained model for EEG-based motor imagery classification. *Knowledge-Based Systems*, 343:115966, 2026.

[51] Haijie Liu, Penghu Wei, Haochong Wang, Xiaodong Lv, Wei Duan, Meijie Li, Yan Zhao, Qingmei Wang, Xinyuan Chen, Gaige Shi, et al. An eeg motor imagery dataset for brain computer interface in acute stroke patients. *Scientific Data*, 11(1):131, 2024.

[52] Deeksha M. Shama, Jiasen Jing, and Archana Venkataraman. Deepsoz: A robust deep model for joint temporal and spatial seizure onset localization from multichannel eeg data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–194. Springer, 2023.

[53] Walter McNicholas, Liam Doherty, Silke Ryan, John Garvey, Patricia Boyle, and Eric Chua. St. vincent's university hospital / university college dublin sleep apnea database, 2004.

[54] Masaki Nakanishi, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PloS one*, 10(10):e0140703, 2015.

[55] Wassim Nasreddine. Epileptic EEG Dataset, March 2021.

[56] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[57] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[58] Christian O'Reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, 23(6):628–635, 2014.

[59] Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg–adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

[60] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. DCSM sleep staging dataset, 2021.

[61] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, March 2019.

[62] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, August 2022. arXiv:2105.11043 [cs].

[63] Léa Pillette, Aline Roc, Bernard N'kaoua, and Fabien Lotte. Experimenters' influence on mental-imagery based brain-computer interface user training. *International Journal of Human-Computer Studies*, 149:102603, 2021.

[64] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, December 1997.

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[66] Haider Raza, Dheeraj Rathee, Shang-Ming Zhou, Hubert Cecotti, and Girijesh Prasad. Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related eeg-based brain-computer interface. *Neurocomputing*, 343:154–166, 2019. Learning in the Presence of Class Imbalance and Concept Drift.

[67] S Redline, P V Tishler, T D Tosteson, J Williamson, K Kump, I Browner, V Ferrette, and P Krejci. The familial aggregation of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.*, 151(3 Pt 1):682–687, March 1995.

[68] Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and p300-based bci performance in people with amyotrophic lateral sclerosis. *Frontiers in human neuroscience*, 7:732, 2013.

[69] Julio Rodriguez-Larios and Kaat Alaerts. Tracking transient changes in the neural frequency architecture: harmonic relationships between theta and alpha peaks facilitate cognitive performance. *Journal of Neuroscience*, 39(32):6291–6298, 2019.

[70] Carol L Rosen, Dennis Auckley, Ruth Benca, Nancy Foldvary-Schaefer, Conrad Iber, Vishesh Kapur, Michael Rueschman, Phyllis Zee, and Susan Redline. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. *Sleep*, 35(6):757–767, June 2012.

[71] Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

[72] Fanqi Shen, Enhong Yang, Jiahe Li, Junru Hong, Xiaoran Pan, Zhizhang Yuan, Meng Li, and Yang Yang. Brain4FMs: A Benchmark of Foundation Models for Electrical Brain Signal, 2026. Version Number: 1.

[73] Enze Shi, Kui Zhao, Qilong Yuan, Jiaqi Wang, Huawen Hu, Sigang Yu, and Shu Zhang. Fome: A foundation model for eeg using adaptive temporal-lateral attention scaling. *arXiv preprint arXiv:2409.12454*, 2024.

[74] Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Do-Won Kim, Jichai Jeong, Han-Jeong Hwang, and Klaus-Robert Müller. Open access dataset for eeg+ nirs single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2016.

[75] Gaganpreet Singh, Caroline PC Chanel, and Raphaëlle N Roy. Mental workload estimation based on physiological features for pilot-UAV teaming applications. *Frontiers in Human Neuroscience*, 15:692878, 2021.

[76] Nathan J Stevenson, Karoliina Tapani, Leena Lauronen, and Sampsa Vanhatalo. A dataset of neonatal eeg recordings with seizure annotations. *Scientific data*, 6(1):190039, 2019.

[77] Haoqi Sun, Sarah Milton, Yi Fang, et al. Machine learning–based sleep electroencephalographic brain age index and dementia risk: An individual participant data meta-analysis. *JAMA Network Open*, 9(3):e261521, 2026.

[78] Haoqi Sun, Luis Paixao, Jefferson T Oliva, Balaji Goparaju, Diego Z Carvalho, Kicky G van Leeuwen, Oluwaseun Akeju, Robert J Thomas, Sydney S Cash, Matt T Bianchi, et al. Brain age from the electroencephalogram of sleep. *Neurobiology of aging*, 74:112–120, 2019.

[79] Haoqi Sun, Luis Paixao, Jefferson T. Oliva, Balaji Goparaju, Diego Z. Carvalho, Kicky G. van Leeuwen, Oluwaseun Akeju, Robert J. Thomas, Sydney S. Cash, Matt T. Bianchi, and M. Brandon Westover. Brain age from the electroencephalogram of sleep. *Neurobiology of Aging*, 74:112–120, February 2019.

[80] Haoqi Sun, Elissa Ye, Luis Paixao, Wolfgang Ganglberger, Catherine J. Chu, Can Zhang, Jonathan Rosand, Emmanuel Mignot, Sydney S. Cash, David Gozal, Robert J. Thomas, and M. Brandon Westover. The sleep and wake electroencephalogram over the lifespan. *Neurobiology of Aging*, 124:60–70, 2023.

[81] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

[82] Rahul Thapa, Bryan He, Magnus Ruud Kjær, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. *arXiv preprint arXiv:2405.17766*, 2024.

[83] Valentin Thorey. Dreem open datasets, 2025.

[84] Matthew M. Troester, Stuart F. Quan, Richard B. Berry, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Darien, IL, version 3 edition, 2023.

[85] Erwan Vaineau, Alexandre Barachant, Anton Andreev, Pedro C Rodrigues, Grégoire Cattan, and Marco Congedo. Brain invaders adaptive versus non-adaptive p300 brain-computer interface dataset. *arXiv preprint arXiv:1904.09111*, 2019.

[86] Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

[87] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.

[88] M Brandon Westover, Valdery Moura Junior, Robert Thomas, Sydney Cash, Samaneh Nasiri, Haoqi Sun, Aditya Gupta, Jonathan Rosand, Manohar Ghanta, Wolfgang Ganglberger, Umakanth Katwa, Katie Stone, Zhiyong Zhang, Gauri Ganjoo, Thijs E Nassi PhD Candidate, Ruoqi Wei, Dennis Hwang, Lynn Marie Trotti, Ankit Parekh, Erikjan Meulenbrugge, Emmanuel Mignot, Rhoda Au, Gari Clifford, and David Rapoport. The human sleep project, 2023.

[89] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

[90] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[91] Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications. *arXiv preprint arXiv:2507.09882*, 2025.

[92] Kerui Wu, Ziyue Zhao, and Bülent Yener. Large eeg-u-transformer for time-step level detection without pre-training. *arXiv preprint arXiv:2504.00336*, 2025.

[93] Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A Comprehensive Benchmark for the Systematic Evaluation of EEG Foundation Models, February 2026. arXiv:2508.17742 [eess].

[94] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

[95] Liuyin Yang, Qiang Sun, Ang Li, and Marc M Van Hulle. Are eeg foundation models worth it? comparative evaluation with traditional decoders in diverse bci tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

[96] Yuxuan Yao, Hongbo Wang, Li Chen, Yiheng Peng, and Jingjing Luo. Foundation models for EEG decoding: current progress and prospective research. *Journal of Neural Engineering*, 22(6):061002, December 2025.

[97] Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PloS one*, 9(12):e114853, 2014.

[98] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla. Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study. *WMJ*, 108(5):246–249, August 2009.

[99] Han Yuan and Bin He. Brain–Computer Interfaces Using Sensorimotor Rhythms: Current State and Future Perspectives. *IEEE Transactions on Biomedical Engineering*, 61(5):1425–1435, 2014.

[100] Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2024.

[101] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inform. Assoc.*, 25(10):1351–1358, October 2018.

[102] Danhua Zhu, Jordi Bieger, Gary Garcia Molina, and Ronald M Aarts. A survey of stimulation methods used in SSVEP-based BCIs. *Computational intelligence and neuroscience*, 2010(1):702357, 2010.

[103] Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

Supplementary Material

A Model Description

A.1 Input-space specifications

EEG-FM substantially vary in the specific way they handle EEG signals as input. Axes of variation include sampling rate, normalization, re-referencing, flexibility in number of channels or time windows, a fixed set of allowed channels versus interpolating spherical channel coordinates,... . For the models evaluated in NEUROATLAS, we describe their input specifications under which they were designed in Table 1.

A.2 Models Included in NEUROATLAS

NEUROATLAS evaluates models from three complementary groups: supervised EEG models, EEG foundation models, and general time-series foundation models. The goal is to compare task-specific supervised representations, EEG-oriented pretrained representations, and generic temporal representations under a unified frozen-encoder evaluation protocol.

EEG Foundation Models We evaluate the following EEG foundation models (EEG-FMs): BIOT [94], CBraMod [87], EEGPT [86], LaBraM [31], NeuroGPT [18], NeuroLM [30], NeuroRVQ [8], REVE [59], SleepFM [82], ST-EEGFormer-S/B/L [95], and MIRepNet [50]. These models are pretrained on EEG or neurophysiological data using objectives such as masked reconstruction, contrastive learning, autoregressive modeling, discrete token prediction, or multimodal alignment. EEG-FMs differ in preprocessing requirements and channel configurations, summarized in Table 1.

General Timeseries Foundation Models We evaluate the following timeseries foundation models (TS-FMs): Chronos-T/S/B/L [4], Moirai-S/B/L [89], and MOMENT-S/B/L [24]. These models are pretrained on heterogeneous non-EEG temporal datasets and are designed to capture generic timeseries features. Chronos models univariate time series through tokenized numerical sequences. MOMENT uses patch-based temporal representations with instance normalization and supports several downstream modes such as forecasting, classification, imputation, and representation extraction. Moirai supports multivariate forecasting through any-variate sequence modeling. Unlike EEG-FMs, these models do not explicitly encode electrode topology, montage geometry, or EEG-specific referencing assumptions, detailed specifications can be found in Table 1. They therefore provide a domain-agnostic baseline for testing whether generic timeseries pretraining can transfer to EEG tasks when used as a frozen feature extractor.

Supervised EEG Models We evaluate supervised pretrained (Supervised-Pre) from all domains: we use epilepsy-trained Seizure-TF [92] and DeepSOZ [52] and sleep-trained SleepTransformer [62], SleePyCo [46], and CoRe-Sleep [42]. On brain age, we evaluate the sleep EEG Brain Age feature-based model described in Sun et al. [78]; On BCI, we evaluate EEGNet [45]. These models are pretrained to perform a specific task and their input assumptions are often closely tied to the target task. For example, EEGNet is used in the BCI setting, where channel selection and spatial filtering are central to motor imagery decoding. As such, these models provide task- or domain-specific reference baselines for evaluating FMs. Model specifications are described in Table 1 and further discussed per domain below.

Supervised Epilepsy Models Seizure-TF is an encoder-decoder model used for seizure detection, sleep stage classification and pathological detection. The encoder comprises convolutional and transformer layers. The checkpoint that we used was trained on the datasets TUSZ [71] and Siena [21], using windows 60 seconds. At inference, a sliding window of 10 seconds is passed through the model (encoder-only) with no overlap (stride of 10 seconds).

DeepSOZ is a Transformer+LSTM model used for seizure detection and seizure-onset-zone (SOZ) localization. The checkpoint that we used was trained on the datasets TUSZ [71] and Siena [21],

using windows of 1 second. At inference, a sliding window of 10 seconds is passed through the model with no overlap (stride of 10 seconds). We extract the embedding from the Transformer global token, averaged over all seconds in the window.

Supervised Sleep Models SleepTransformer is a sequence-to-sequence transformer-based sleep staging model using single-channel EEG, trained on SHHS [62]. Each 30-second epoch is converted to a short-time Fourier transform (STFT) representation. At inference, a sliding window of sequence length 21, i.e., containing 21 consecutive epochs, is passed through the model with a stride of 1 epoch. We extract the embedding from the CLS token output of the outer transformer at the center epoch (index 10 of 21).

SleePyCo is a sleep staging model using single-channel EEG, trained using supervised contrastive learning [46]. We evaluate the published checkpoint trained on SHHS. During inference, a sliding sequence of 10 consecutive epochs is fed with a stride of 1 epoch. Extracted embeddings are the attention-pooled transformer output, averaged across the three feature-pyramid scales, attributed to the last epoch of the sequence.

CoRe-Sleep is a transformer-based bimodal sleep staging model trained on EEG and EOG, trained on SHHS [42]. At inference, only the EEG channel is used, and each epoch is converted using STFT. Embeddings are extracted using non-overlapping sequences of length 21 epochs, as the summed cross-attention fused representation of the outer Transformer outputs, retained for all epochs in the window.

Thus, in contrast to the evaluated FM embeddings, the Supervised-Pre models explicitly encode inter-epoch temporal context across sequences of 30-second epochs.

Supervised Brain Ageing Model We additionally include the sleep-EEG brain-age model of Sun et al. as a task-specific supervised reference [79]. In contrast to the frozen foundation-model probes, this is a feature-based model trained to predict chronological age from sleep EEG features and is therefore directly optimized for brain-age estimation. We use it as an upper reference for the brain-age prediction and BAG analyses, allowing us to compare general-purpose EEG and time-series representations against a model designed specifically for sleep-EEG ageing.

Supervised BCI/MI Model For the supervised EEGNet baseline, we used publicly available pretrained EEGNetv4 checkpoints released on Hugging Face by Guetschel et al. [25], available at `https://huggingface.co/PierreGtch/EEGNetv4`. These pretrained models were only available for the Motor Imagery (MI) paradigm, which is why EEGNet is exclusively included in the MI comparison. Following the same protocol used for EEG foundation models and time-series foundation models, we froze the pretrained encoder, extracted latent embeddings, and evaluated them using linear probing under a leave-one-subject-out (LOSO) setting.

Table 1: Overview of models included in NEUROATLAS and their input specifications.

Model	Channels	Sampling rate(s)	Scaling / Normalization	Referencing	Channel ordering	Strict vs Flexible
BIOT	Variable; token-based; pretrained on 16–18 bipolar channels, e.g., C3-A2/C4-A1.	200 Hz; resampled; FFT-based segment embedding.	Per-channel / per-sample normalization using the 95% quantile of absolute amplitude.	Dataset-dependent; commonly bipolar 10–20 derivations.	Index-based; order-sensitive; channel identity encoded by learned channel embeddings.	Flexible; supports variable channels and sequence lengths; missing segments handled at token level.
CBraMod pretrained; rand	19 EEG; fixed pretraining subset from standard 10–20 channels.	200 Hz; 30 s samples; 1 s non-overlapping patches.	Bandpass 0.3–75 Hz; notch 60 Hz; bad-sample removal; amplitude scaled by 100 μV to [-1,1].	Reference-agnostic design.	Spatial–temporal encoding with asymmetric conditional positional encoding.	Semi-flexible; fixed pretraining channels with adaptive positional encoding for downstream variation.
EEGPT	58 pretraining channels; name-based embedding space.	256 Hz; patch-based input.	Amplitude normalization to a consistent scale, depending on dataset convention.	Average reference / global CAR.	Name-based channel identity; ordering not strictly enforced.	Flexible; supports subsets of predefined channels.
LaBraM	Variable multi-channel EEG; 10–20 system; patch-based segmentation.	200 Hz; non-overlapping patch window.	Bandpass 0.1–75 Hz; notch 50 Hz; amplitude scaled by 0.1 mV to [-1,1].	Not explicitly fixed; heterogeneous referencing across datasets.	Spatially aware channel-identity embeddings.	Flexible; supports heterogeneous EEG montages but requires correct channel-to-embedding mapping.
NeuroGPT	22 EEG channels; fixed extended 10–20 montage.	250 Hz; resampled from heterogeneous sources.	Bandpass 0.5–100 Hz; notch 60 Hz; detrending; z-score per recording.	Average reference / global CAR.	Fixed aligned 10–20 ordering.	Semi-flexible; supports multi-dataset training through strict preprocessing alignment.
NeuroLM	Variable multi-channel EEG; typically 4–62 channels; 10–20 system.	200 Hz; patch-based tokenizer.	Bandpass 0.1–75 Hz; notch 50/60 Hz; amplitude normalization; z-score in frequency branch.	Not explicitly constrained.	Spatially aware; channel identity tied to electrode position.	Flexible; supports heterogeneous EEG configurations.
NeuroRVQ	Variable multi-dataset EEG; up to 64 channels.	Unified to 200 Hz.	Fourier-based representation with log-amplitude scaling.	Dataset-dependent.	Learned spatial embedding.	Flexible; relatively channel-agnostic representation interface.

Continued on next page

Model	Channels	Sampling rate(s)	Scaling / Normalization	Referencing	Channel ordering	Strict vs Flexible
REVE	Variable multi-channel EEG; coordinate-based arbitrary layouts.	200 Hz; overlapping temporal patches.	Per-record z-score; clipping at $\pm 15\sigma$; bandpass 0.5–99.5 Hz.	Recording-dependent; no fixed re-referencing.	Spatial representation based on 3D electrode coordinates.	Flexible; supports arbitrary layouts and sequence lengths.
SleepFM	Multimodal PSG; EEG, EOG, EMG, ECG, and respiratory channels.	256 Hz; resampled.	Minimal preprocessing; 30 s clips; raw signals mostly preserved.	Dataset-dependent PSG derivations.	Modality-wise temporal encoding with temporal fusion.	Flexible; handles heterogeneous modalities and cross-modal representation tasks.
ST-EEGFormer S/B/L	142 unique EEG channels during pretraining; downstream channels are dataset-dependent.	128 Hz for pretraining; downstream data resampled to model-native rate.	Bandpass 0.1–64 Hz; notch filtering; per-channel zero-mean/unit-variance standardization.	Not explicitly fixed.	Spatial–temporal tokenization with learned spatial and temporal positional embeddings.	Flexible; designed for heterogeneous EEG datasets and channel configurations.
MIRepNet	MI-specific channel template over FC, C, CP, and T regions; implementation uses 45 template channels.	250 Hz; resampled.	Bandpass 8–30 Hz; Euclidean Alignment / whitening.	Not explicitly specified.	Template-based spatial alignment using inverse-distance interpolation.	Flexible; maps heterogeneous MI layouts to a unified motor-imagery template.
CoRe-Sleep	Multimodal EEG + EOG; e.g., C4-A1 and L-R.	100 Hz; resampled from EEG and EOG recordings.	Time-frequency representation; EEG 0.3–40 Hz; EOG 0.3–23 Hz; STFT over 30 s segments.	Dataset-dependent derivations.	Multi-stream temporal encoding with cross-attention fusion.	Flexible; robust to missing or noisy modalities.
SleePyCo	Single-channel EEG; e.g., Fpz-Cz, C4-A1, C3-A2.	100 Hz; 30 s epochs.	Minimal preprocessing except downsampling; raw signal input; internal BatchNorm.	Dataset-dependent sleep EEG derivations.	Single-channel temporal ordering.	Flexible across single-channel sleep datasets; not designed for arbitrary multi-channel montages.
DeepSOZ-HEM	19 EEG channels; fixed 10–20 configuration.	256 Hz input; resampled to 200 Hz.	Bandpass 1.6–30 Hz; clipping at 2σ.	Dataset-dependent / not specified.	Fixed channel positional encoding.	Strict; requires fixed channel configuration.
Seizure-TF	RNS-derived A+B pattern features from two RNS channels.	Daily time series; no raw EEG sampling rate.	Per-patient z-score normalization of A+B patterns.	Not applicable; structured RNS biomarkers rather than raw scalp EEG.	Temporal patch ordering with learnable positional encoding.	Flexible for structured RNS features; not designed for raw scalp EEG layouts.
SleepTransformer	C4-A1 for SHHS, Fpz-Cz for SleepEDF; code also supports EEG/EOG/EMG.	100 Hz.	STFT time-frequency input; log-amplitude spectrum; zero-mean/unit-variance normalization.	Dataset-defined derivations.	Temporal ordering with positional encoding.	Flexible in modality input; mainly evaluated on single-channel sleep EEG.
EEGNet	Multi-channel EEG; channel count depends on dataset.	Typically resampled according to task setting.	No fixed preprocessing; BatchNorm used internally.	Dataset-dependent.	Spatial structure learned implicitly through convolution.	Flexible; can adapt to different channel counts and paradigms.
Brain Age (Sun et al.	6-channel EEG montage, e.g., F3, F4, C3, C4, O1, O2; also 2-channel variant.	200 Hz.	Sleep-EEG feature normalization; commonly includes robust per-channel scaling and z-scored features.	Contralateral mastoid reference, e.g., F3-M2.	Fixed PSG montage.	Strict; channel-specific model.
Chronos T/S/B/L	Univariate time series; no channel structure.	Dataset-dependent; original sampling generally preserved.	Mean scaling followed by value quantization into discrete tokens.	Not applicable.	Temporal sequence ordering.	Flexible for arbitrary-length univariate sequences; not topology-aware.
Moirai S/B/L	Variable multivariate series; arbitrary numbers of variables.	Variable; supports multi-frequency time series.	Instance normalization across heterogeneous datasets.	Not applicable.	Flattened temporal-variate ordering with any-variate attention.	Flexible; universal forecasting model across domains, frequencies, and variable dimensions.
MOMENT S/B/L	Univariate input; multivariate handled through channel-wise processing.	Variable; fixed input length after resampling or padding.	Reversible Instance Normalization.	Not explicitly constrained.	Patch-based temporal tokenization.	Flexible; supports forecasting, classification, imputation, anomaly detection, and representation extraction.

B Dataset Description

B.1 Epilepsy

The seven continuously-labelled cohorts span neonatal (Helsinki [76]), pediatric (CHB-MIT [27]), adult seizure monitoring (Siena [21], SeizeIT1 [11], AUB-Med [55]), and large-scale clinical archives like TUSZ [71], Epilepsiae [41] and SeizeIT2 [9]. Three additional cohorts (TUAB [20], NMT [39], Bonn [3]) provide recording-level abnormality labels rather than continuous seizure annotations and contribute $\sim$2,500 abnormal recordings to the recording-level evaluation only; of these, only Bonn's 200 ictal clips carry explicit seizure annotation, while TUAB and NMT label broader pathology including epileptiform activity, slowing, and artifacts. We evaluate on splits that are patient-level throughout.

B.2 Sleep

The 15 cohorts evaluated for sleep span lab and clinical polysomnography (PSG) recordings (Cleveland Family Study (CFS) [67, 101], DCSM Sleep Staging Dataset [60], Dreem Open Datasets (DOD) [83, 26], Haaglanden Medisch Centrum (HMC) sleep staging database [2], HomePAP [70, 101], ISRUC-Sleep [38], Montréal Archive of Sleep Studies (MASS) [58], George B. Moody PhysioNet Challenge 2026 (PN2026), a subset of the Human Sleep Project [88], Sleep-EDF Database Expanded [37], Stanford Technology Analytics and Genomics in Sleep (STAGES) [101], St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCDDB), [53], Wisconsin Sleep Cohort (WSC) [101, 98]), at-home PSG recordings (Multi-Ethnic Study of Atherosclerosis

(MESA) [15, 101], Osteoporotic Fractures in Men Sleep Study (MrOS) [12, 101], Sleep Heart Health Study (SHHS) [64, 101], Sleep-EDF). The datasets span multiple clinical populations, including children (CFS), and people with cognitive impairment (CI) (PhysioNet 2026) and obstructive sleep apnea (OSA) (DOD, UCDDB). Furthermore, MASS is annotated for microarousals, PhysioNet 2026 contains labels for microarousals, respiratory events and limb movements, and UCDDB is annotated for respiratory events.

B.3 Brain Aging and Neurodegenerative Diseases

The brain age domain is evaluated on 10 cohorts of PSG recordings which include age labels, spanning $\sim$15,000 patients and $\sim$193k hours (CFS, HomePAP, ISRUC, MESA, MrOS, PhysioNet 2026, SHHS, SleepEDF Cassette subset (SC), STAGES, and WSC). Furthermore, PhysioNet 2026 includes patients with cognitive impairment as well as healthy controls.

B.4 BCI

The benchmark spans a diverse set of BCI paradigms, including **motor imagery (MI)** (BNCI2014_001 [81], BNCI2014_004 [81], BNCI2015_001 [23], Weibo2014 [97], Dreyer2023 [63], Shin2017A [74], Liu2024 [51]), **ERP** (BI2013a [85], BI2014a [43], EPFLP300 [29], BNCI2014_008 [68], ErpCore2021_N170 [33]), and **SSVEP** (Nakanishi2015 [54], Kim2025BetaRange [40]), complemented by cognitive and affective datasets (DREAMER valence/arousal [36], EEGMat [103], ArithmeticTask [69]). In total, the benchmark comprises 509 participants and exhibits substantial heterogeneity in acquisition, including variation in channel configurations, sampling rates, and recording designs (event-related trials vs. continuous sessions). Paradigm structure varies accordingly, with MI datasets providing multi-second trials with moderate sample sizes per subject, ERP (P300 and N170) datasets yielding high-density event-locked epochs, and SSVEP datasets consisting of short ($\approx$2-6 s) frequency-tagged trials. Combined with continuous recordings, this amounts to approximately 159 hours of EEG.

C Methodology

C.1 Epilepsy

To evaluate seizure detection, we split EEG recordings in windows of 10 seconds without overlap, and extract embeddings per window for EEG-FMs, TS-FMs, and Supervised-Pre models as frozen-backbone encoders. We assess embeddings using class-balanced logistic regression trained using LBFGS. The L2 regularization strength C is selected by grid search over $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$ on a held-out validation fold, with selection criterion AUPRC for binary seizure detection and macro-F1 for multi-class seizure-type classification. The probe is refit at the selected C on the full training fold before test evaluation. Folds are defined at the patient-level.

C.2 Sleep

To evaluate sleep, EEG-FMs, TS-FMs, as well as Supervised-Pre models are consistently evaluated as embedding extractors, by extracting embeddings per 30-second window. For epoch-level tasks, a linear probe is applied to individual epoch-level embeddings; for recording-level diagnosis, embeddings are average-pooled across each recording before classification. Event annotations were converted to binary epoch labels, which however does not assess precise event onset, offset, or duration. We fit a linear probe using LBFGS. For event-detection and diagnosis tasks, a balanced cross-entropy loss is used to account for class imbalance, while unweighted loss is applied for the sleep staging task. Positive epochs are defined using task-specific duration thresholds derived from AASM criteria [84]. Prior to training, embeddings are standardized to ensure zero mean and unit variance. Default L2 regularization is employed with C = 1.0. Folds are defined at the patient level.

Hypnogram Features

Table 2: Hypnogram feature definitions, adapted from [28]. All durations are in hours. The sleep period spans from the first non-Wake epoch to the last non-Wake epoch. Epoch length is 30 s.

Feature	**Unit**	**Definition**
Duration & efficiency		
TST	h	Total sleep time (all non-Wake epochs)
TimeW	h	Total time in Wake
TimeN1	h	Total time in N1
TimeN2	h	Total time in N2
TimeN3	h	Total time in N3
TimeR	h	Total time in REM
Sleep-stage proportions		
RelN1	–	TimeN1/TST
RelN2	–	TimeN2/TST
RelN3	–	TimeN3/TST
RelR	–	TimeR/TST
Transitions		
Awakenings	count	Transitions into Wake from any other stage
SlStCh	count	Total epoch-to-epoch stage changes
ChToR	count	Transitions into REM from any other stage
ChToN3	count	Transitions into N3 from any other stage
Bout analysis		
TimeBetweenW	h	Mean inter-bout interval between Wake bouts
TimeBetweenR	h	Mean inter-bout interval between REM bouts
TimeBetweenN3	h	Mean inter-bout interval between N3 bouts
WBoutDur	h	Mean duration of contiguous Wake bouts
RBoutDur	h	Mean duration of contiguous REM bouts
N3BoutDur	h	Mean duration of contiguous N3 bouts
Latencies		
WASO	h	Wake after sleep onset (Wake time within sleep period)
REMLatency	h	Time from sleep onset to first REM epoch
normalized transition rates		
RelAwakenings	h^{-1}	Awakenings/TST
RelSlStCh	h^{-1}	SlStCh/TST
RelChToR	h^{-1}	ChToR/TST
RelChToN3	h^{-1}	ChToN3/TST
RelWakeTime	–	WASO/TST
Temporal distribution within sleep period		
RelOccN1	–	Mean normalized position of N1 epochs (0–1)
RelOccN2	–	Mean normalized position of N2 epochs (0–1)
RelOccN3	–	Mean normalized position of N3 epochs (0–1)
RelOccR	–	Mean normalized position of REM epochs (0–1)

C.3 Brain Age Prediction

Brain-Age Regression We mapped 30-second sleep epochs to embedding vectors and trained linear ridge-regression probes on the resulting cached features, without updating the encoder. The primary analysis used subject-level mean pooling, where epoch embeddings from each recording were averaged into a single feature vector before regression.

Ridge regression was used as the downstream probe because it provides a linear result while controlling for overfitting through the regularization parameter α. For each dataset, we used a fixed five-fold subject-level cross-validation protocol shared across all models. Folds were stratified by age bins years. Within each outer training fold, α was selected by nested cross-validation over:

$$\alpha \in \{0.001, \ldots, 100\},$$

using validation MAE. The selected model was then refit on the full outer training set and evaluated on the held-out test fold. For multi-visit cohorts, folds were to ensure that recordings from the same individual never appeared in both training and test sets.

Brain-age Gap and Clinical Evaluation For each held-out subject, we computed the brain-age gap (BAG) as:

$$\mathrm{BAG}(x) = \hat{y}(x) - y(x),$$

where $\hat{y}(x)$ is the predicted age from sleep EEG recording x and $y(x)$ is the chronological age. Since the age-prediction model was trained on the remaining healthy subjects in the training split, BAG reflects deviation from a normative healthy ageing trajectory. Positive values indicate an older-appearing brain.

We report two complementary BAG-based statistics. The mean BAG difference:

$$\Delta\mathrm{BAG} = \overline{\mathrm{BAG}}_{\mathrm{CI}} - \overline{\mathrm{BAG}}_{\mathrm{Healthy}},$$

which quantifies group-level shifts in apparent brain age. Its 95% confidence interval was computed using the independent-samples standard error.

We additionally report AUROC to assess rank-based Healthy–Cognitive Impairement separation:

$$\mathrm{AUROC} = P\left(\mathrm{BAG}_{\mathrm{CI}} > \mathrm{BAG}_{\mathrm{Healthy}}\right),$$

Values above 0.5 indicate separation in the expected accelerated-ageing direction, whereas values below 0.5 indicate separation opposite to this direction. For AUROC, 95% confidence intervals were estimated using 1,000 matched-pair bootstrap resamples. Each resample selected age-matched case–control pairs with replacement and included both members of each selected pair, preserving the matching structure.

C.4 BCI

Heterogeneity of BCI tasks in EEG signal characteristics. A key challenge in evaluating EEG foundation models is the large variability across BCI paradigms. Our benchmark includes four paradigms that differ in neural activity, timing, number of classes, and recording hardware. All foundation models are kept entirely frozen and evaluated via a logistic regression probe on the extracted embeddings. Raw EEG is converted to microvolts (μV) and preprocessed (notch and bandpass filtering, optional common average re-referencing, resampling); each model applies its own internal amplitude scaling on the μV input. We use leave-one-subject-out (LOSO) cross-validation: in each fold one subject is held out for testing and the remaining subjects form the training set. Embeddings are z-score standardized per fold (training statistics applied to the test set). The probe is an ℓ_2-regularized logistic regression (`sklearn.linear_model.LogisticRegression`, C=1.0, L-BFGS solver, up to 1 000 iterations), with balanced class weights for binary tasks.

Motor imagery (MI) relies on event-related modulations of sensorimotor rhythms in the α (8-13 Hz) and β (13-30 Hz) bands [99]. Trials typically span 3-6 s of imagery within 6-10 s epochs, and discriminative information is distributed over time, requiring longer integration windows. Performance is particularly sensitive to inter-subject variability, as individual differences in brain anatomy, motor strategy, and neurophysiological patterns can lead to pronounced distribution shifts between participants [1]. Within MI alone, the benchmark already exhibits marked heterogeneity: the number of classes ranges from C=2 (e.g. BNCI2014-004, Dreyer2023, Liu2024, Shin2017A) through

C=4 (BNCI2014-001) to C=7 (Weibo2014); channel counts vary from 3 (BNCI2014-004) to 60 (Weibo2014); sampling rates span 200-512Hz; and cohort sizes range from 9 to 87 subjects.

Event-related potential (ERP) paradigms (e.g. P300, N170) involve transient neural responses within 170-600 ms post-stimulus. Trials are considerably shorter (0.5-1 s), but decoding typically aggregates repetitions over longer sequences, with low signal-to-noise ratios demanding precise temporal alignment [34]. All five ERP datasets in our benchmark are binary (C=2), yet they vary widely in hardware and temporal resolution: channel counts range from 8 (BNCI2014-008) to 32 (EPFLP300), and sampling rates from 256 Hz to 2048 Hz, reflecting the diversity of clinical and research-grade setups. These differences are particularly challenging for a frozen encoder, which must produce informative embeddings without adjusting to the specific temporal structure or hardware characteristics of each dataset.

Steady-state visual evoked potentials (SSVEP) exploit frequency-locked cortical responses to periodic visual stimulation, achieving high information transfer rates but remaining sensitive to attention, fatigue, and harmonic interference [102]. This paradigm introduces the most extreme class counts in our benchmark: C=12 (Nakanishi2015) and C=40 (Kim2025, Liu2020B), bringing the theoretical chance level down to as low as 2.5%. Channel counts also span an order of magnitude, from 8 to 64. Distinguishing among 40 frequency targets from frozen embeddings, without any spectral fine-tuning, represents one of the most demanding evaluations in the benchmark.

Cognitive and affective state decoding introduces yet another source of variability. Trial windows range from short segments (e.g. 2 s for workload [75]) to longer epochs lasting tens of seconds (e.g. the DREAMER affective dataset [36]). Class counts vary from C=2 (DREAMER, EEGMat) to C=3 (ArithmeticTask), and the hardware ranges from consumer-grade devices (Emotiv EPOC at 128 Hz, 14 channels) to research-grade systems (500 Hz, 23 channels). Given that subject variability in cognitive and affective tasks can be attributed to differences in brain anatomy, personality, and cognitive abilities among individuals, the absence of any subject-specific calibration or model adaptation renders these tasks particularly challenging.

Taken together, these paradigms span class counts from 2 to 40, channel counts from 3 to 64, sampling rates from 128 Hz to 2048 Hz, and cohort sizes from 8 to 100 subjects, all evaluated on entirely frozen representations with no fine-tuning whatsoever. This heterogeneity motivates both the use of *normalized balanced accuracy* (Eq. 2), which places all tasks on a common scale where 50% equals chance regardless of C, and the per-dataset *binomial significance threshold* (Eq. 3), which accounts for the finite number of evaluation samples. Together, these two normalizations ensure that cross-paradigm comparisons are neither biased by differing chance levels nor inflated by small sample sizes, providing a rigorous and unified framework for assessing the quality of pretrained EEG representations.

C.4.1 Metrics

Normalized balanced accuracy We evaluate all models using *balanced accuracy*, defined as the arithmetic mean of per-class recall values:

$$\mathrm{BA} \;=\; \frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FN}_c}\,, \tag{1}$$

where C denotes the number of classes and TP_c, FN_c are the true positives and false negatives for class c. Balanced accuracy is preferred over standard accuracy because it is insensitive to class-frequency imbalance and reduces to the theoretical chance level $1/C$ for a random classifier regardless of the class distribution. However, because the benchmark spans paradigms with widely different numbers of classes - ranging from C=2 (e.g. binary motor imagery) to C=40 (e.g. 40-target SSVEP) - raw balanced accuracies are not directly comparable: a score of 60% is far above chance for a 40-class problem but only marginally so for a binary one. To enable fair comparison both within and across paradigms, we apply an affine normalization that maps chance-level performance to 50% and perfect classification to 100%:

$$\widetilde{\mathrm{BA}} \;=\; \frac{\mathrm{BA}-\frac{1}{C}}{1-\frac{1}{C}} \;\times\; 0.5 \;+\; 0.5\,. \tag{2}$$

Under this transformation every dataset shares a common chance baseline of 50% and a common ceiling of 100%, so that scores can be meaningfully averaged and compared across tasks with different numbers of classes.

Statistical significance of decoding accuracy The theoretical chance level $100/C\,\%$ is strictly valid only in the limit of infinitely many test samples. As demonstrated by [17], classifiers applied to small datasets which is a common scenario in brain-computer interface (BCI) research, can yield empirical accuracies well above the theoretical chance level purely by statistical fluctuation. For example, with only 40 observations a two-class decoder can reach 75% accuracy on random data at a rate exceeding p=0.001. To guard against such inflated chance-level estimates, we derive a sample-size–dependent significance threshold using the binomial cumulative distribution. Under the assumption that classification errors are binomially distributed, the probability of correctly classifying at least z out of n samples by chance with C classes is

$$P(z) \;=\; \sum_{i=z}^{n} \binom{n}{i} \left(\frac{1}{C}\right)^{i} \left(\frac{C-1}{C}\right)^{n-i}. \tag{3}$$

For each dataset we compute the minimum balanced accuracy BA^* such that $P\big(\lceil \mathrm{BA}^* \cdot n \rceil\big) < \alpha$ with α=0.05, using the inverse survival function of the binomial distribution. This threshold is then mapped to the normalized scale via Eq. (2), yielding a per-dataset significance boundary $\widetilde{\mathrm{BA}}^*$ that is directly comparable across all paradigms. In all reported figures, results that do not exceed the corresponding normalized significance threshold are marked accordingly, ensuring that only decodings that are statistically better than chance, accounting for both the number of classes and the finite sample size, are interpreted as meaningful.

C.5 Statistical tests

To identify models that reliably rank among the best per domain and across the domains, we apply a binomial win test. For each of the datasets, we rank all models by their mean performance on the main domain-task. Datasets with missing model performances are excluded. We then count, for each model, the number of datasets on which it finishes in the top 3. Under the null hypothesis that rankings are exchangeable across models, the probability of any given model finishing in the top 3 on a single dataset is $p_0 = 3/k$, where k is the number of models compared. We test whether the observed top-3 count exceeds this chance level using a one-sided exact binomial test at $\alpha = 0.05$.

D Results

D.1 Epilepsy

D.1.1 Per-type score distributions across model families

Scalar metrics summarise a detector's predictions but do not reveal how its confidence distributes across seizure subtypes and background, structure that matters when assessing deployment behaviour on rare or clinically critical seizure families. To expose this structure, we plot the per-window seizure probability emitted by classifier, stratified by the ILAE-2017 [5] main type of the underlying event. Figure 6 contrasts four representative models on TUSZ, SeizeIt2 and Epilepsiae: the best EEG foundation model per dataset (NeuroLM-VQ for TUSZ and SeizeIt2, REVE for Epilepsiae), the best time-series foundation model (Moirai-base), the supervised seizure specialist SeizureTransformer, and a CBraMod random-init baseline that isolates the contribution of pretraining from the architecture and the linear head.

Three patterns stand out. First, the random-init CBraMod baseline emits high seizure probabilities across every type group, including background. The contrast of random initialization with the EEG foundation model, whose distribution concentrates near zero on background, identifies pretraining as the dominant contributor. Second, EEG-FM dominance is dataset-dependent: on TUSZ and SeizeIt2 the green violins on focal, generalised and focal-to-bilateral seizures sit tightly near one, but on Epilepsiae the same architecture exhibits long downward tails and a visibly bimodal distribution on focal events. It is therefore evident that on different datasets the performance on each kind of seizures differs. This analysis suggests that the long pre-surgical Epilepsiae traces, with their subtler

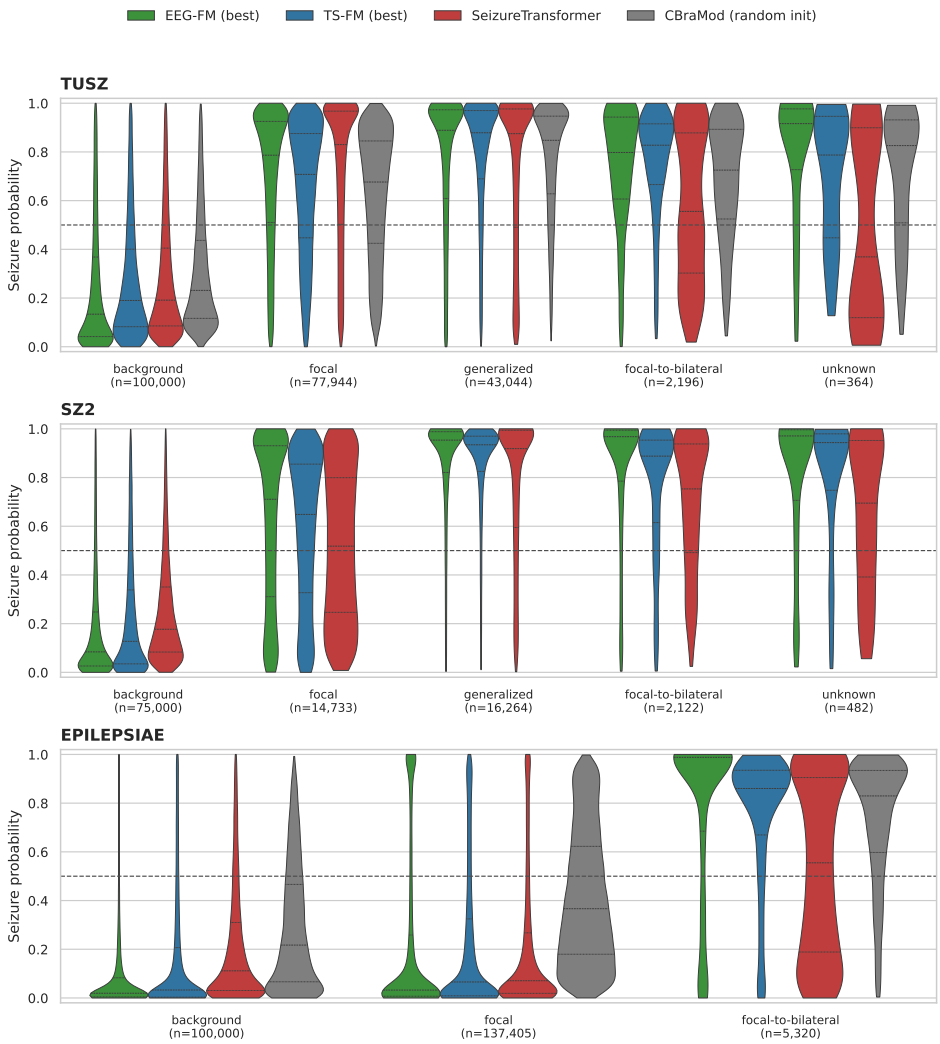

Figure 6: Per-window seizure-probability distributions, stratified by ILAE-2017 main type, on TUSZ (top), SeizeIt2 (middle) and Epilepsiae (bottom). Within each panel, four models are compared: the best EEG foundation model per cohort (*green*; NeuroLM-VQ for TUSZ and SeizeIt2, REVE for Epilepsiae), the best time-series foundation model (*blue*; Moirai-base), the supervised SeizureTransformer (*red*), and a CBraMod random-init baseline (*grey*). Each violin shows the distribution of $\hat{P}(\text{seizure} \mid \text{window})$ over all 10 s test windows of the indicated type pooled across the five subject-disjoint folds; dashed horizontal lines mark $\hat{P}{=}0.5$. Sample counts per group are noted beneath the x-axis (background windows are subsampled to $25{,}000$ per cell for legibility; all seizure-type windows are retained). Generalised and unknown groups are omitted from the Epilepsiae panel because its annotation schema (1981 ILAE) categorises every event as focal-onset. The CBraMod random-init baseline is absent from the SeizeIt2 panel because its embeddings were cached at a different window stride and cannot be byte-exactly aligned with the canonical pipeline.

focal-onset events, remain difficult for every model in the panel. Third, the supervised seizure specialist does not consistently outperform the foundation-models: on TUSZ it exhibits a pronounced false-alarm shoulder on background that the EEG-FM does not.

D.1.2 Window-level AUROC and event-level sensitivity yield different model rankings

We report two scalar metrics on every seizure-detection cohort, and we argue they are not interchangeable. The first is the mean five-fold window-level AUROC of the linear probe, the metric

conventionally reported in the EEG machine-learning literature. The second is the event-based AUC of the median sensitivity-vs-false-alarm-rate curve, integrated against $\log_{10}(\text{FA/h})$ over $[0.1, 100]$ FA/h ("Event-Sens@FA AUC", bounded in $[0, 1]$). The two quantities measure substantively different things. AUROC summarises whether each ten-second window is, in isolation, distinguishable from the background distribution. Event-Sens@FA AUC aggregates the same per-window scores into clinically meaningful seizure events and quantifies how many of those events a deployed system would recover at clinically tolerable alarm budgets. Because seizure-detection systems are evaluated and deployed at the event level under explicit false-alarm constraints, Event-Sens@FA AUC is the metric of clinical relevance. We therefore designate it as our primary metric and report AUROC alongside it to expose the consequences of the prevailing convention.

Figure 7 establishes that the choice of metric is consequential. For each of the seven seizure cohorts in our benchmark we rank every probed model on AUROC and on Event-Sens@FA AUC separately, then draw each model as a single line whose top endpoint is its AUROC rank and whose bottom endpoint is its Event-Sens@FA rank. Two lines crossing indicates that the corresponding pair of models swaps relative order between the two metrics. Spearman's ρ printed at the right of each row summarises the agreement (1.0 denotes identical rankings).

Across the seven cohorts, ρ ranges from 0.61 (Helsinki, Siena) to 0.81 (Epilepsiae), with mean $\bar{\rho} \approx 0.72$. The two metrics never produce identical orderings, and on Helsinki and Siena their agreement is weak enough that they carry largely independent information about which model performs best. The disagreement is most pronounced on TUSZ, where the AUROC-best model, the EEG foundation model NeuroLM, is displaced at the top of the Event-Sens@FA ranking by the fully-supervised seizure detector Seizure-TF. A representation that achieves the highest per-window discriminability does not, in general, produce the cleanest event-level alarm stream. Model selection driven by AUROC, the dominant convention in the broader EEG ML literature, therefore identifies a different model than selection driven by the clinically relevant event-level metric. This motivates our decision to foreground Event-Sens@FA AUC throughout the main results.

D.1.3 Non-window EEG benchmarks: datasets and foundation-model rankings

Three datasets in our suite, Bonn, TUAB, and NMT, carry trial- or recording-level binary labels rather than continuous seizure annotations. They therefore do not admit window-level event metrics (sensitivity at fixed false-alarm rate, etc.) and are evaluated separately from the seizure-detection cohorts in the main text.

The Bonn dataset [3] comprises five sets (A–E) of 100 single-channel surface and intracranial EEG segments, each 23.6 s at 173.6 Hz, originally released as a small seizure/non-seizure benchmark. We use the canonical *S vs. rest* binary split, evaluating set E (ictal) against the four non-ictal sets A–D; after windowing at 10 s, the cohort contains 1,000 segments (800 non-S, 200 S; 1:4 imbalance), evaluated under stratified 5-fold cross-validation at the segment level. Bonn provides no per-trial subject identifier, so subject-grouped cross-validation is not applicable. The Temple University Hospital Abnormal EEG Corpus (TUAB) [57] contains 2,993 clinical EEG sessions from 2,329 subjects (train: 2,717; official eval: 276), labeled normal or abnormal by board-certified neurologists, with the cohort approximately balanced (1,521 normal, 1,472 abnormal); we use the official train/eval split, stratifying the train subjects into 5 folds and reporting on the held-out eval set. The NMT EEG Scalp Abnormal dataset [39] contains 2,417 recordings collected at NMT Karachi, labeled normal or abnormal (train: 2,232; eval: 185), with severe class imbalance (2,002 normal vs. 415 abnormal, approximately 83/17 %) that makes threshold-dependent metrics more sensitive than on TUAB.

For every (model, dataset) pair, we extract the embedding (10 s windows, 10 s stride; window-level features pooled to recording level on TUAB and NMT) and fit a 5-fold sklearn logistic-regression probe with $C \in \{10^{-3}, \ldots, 10^{2}\}$ tuned on the validation fold. Figure 8 pairs the threshold-free AUROC (●) with the threshold-dependent balanced accuracy (○) for every model on each dataset; the connecting segment indexes the calibration gap. Models are colored

Bonn saturates: eleven models reach AUROC ≥ 0.99 (BIOT, REVE, STEEGFormer-L, all four Chronos sizes, NeuroLM, SleepFM, MOMENT-S, and CBraMod), confirming that the S-versus-rest split is recoverable from almost any representation. We therefore treat Bonn as a sanity check rather than a discriminating benchmark; the only meaningful cross-model signal lies in the lower tail, where NeuroGPT (0.881), EEGPT (0.924), and the CBraMod random-init ablation (0.927) trail the saturation ceiling. TUAB orders the EEG foundation models more cleanly: REVE leads at 0.898

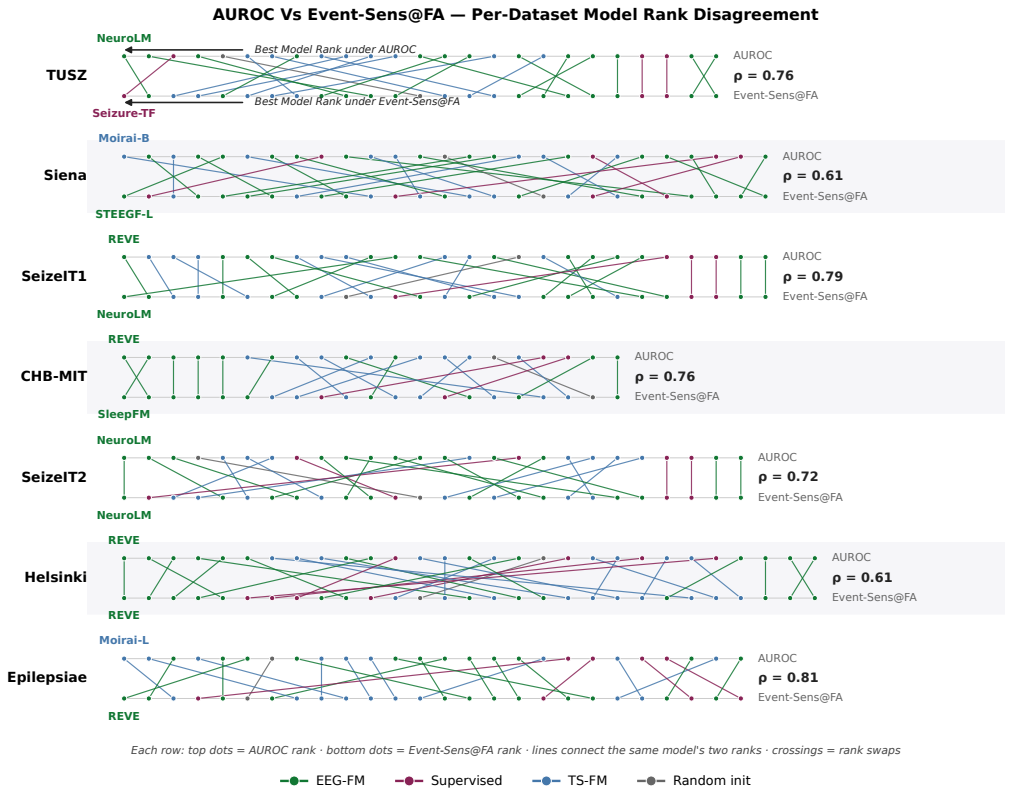

Figure 7: AUROC and event-level Sens@FA yield systematically different model rankings on every seizure-detection cohort. Each row corresponds to one cohort. Within a row, the upper rail of dots is the per-model rank by five-fold-mean window-level AUROC and the lower rail is the per-model rank by the event-based sensitivity-vs-FA/h AUC integrated over $[0.1, 100]$ FA/h; rank 1 (best) is leftmost on both rails. Each line connects the same model's two ranks, coloured by family (green: EEG foundation models; red: supervised pretraining; blue: time-series foundation models; grey: random initialisation). Every line crossing is a pair of models whose relative order flips between the two metrics. Spearman's ρ between the two rankings (right of each row) ranges from 0.61 (Helsinki, Siena) to 0.81 (Epilepsiae) and is below 1 on every cohort, with mean $\bar{\rho} \approx 0.72$ across the seven cohorts. On TUSZ (top row, annotated), NeuroLM is rank-1 by AUROC but is overtaken by the supervised Seizure-TF model on Event-Sens@FA, illustrating that the conventional window-level metric selects a different model than the clinically relevant event-level one. 181 (cohort, model) pairs across the seven seizure-detection cohorts are shown.

AUROC, followed by STEEGFormer-L (0.880), NeuroLM (0.868), STEEGFormer-B (0.858), and EEGPT (0.853); the supervised SeizureTransformer is sixth (0.852), competitive but not specialized for this task, while DeepSOZ-HEM places mid-low (0.790), suggesting that a TUSZ-fitted seizure head transfers only partially to abnormality detection. The three sleep-trained EEG models (SleepFM 0.771, CoRe-Sleep 0.737, Sleep-TF 0.731) sit at the bottom of the panel, indicating limited transfer from sleep-stage pretraining to clinical abnormality. On the imbalanced NMT abnormality task, REVE again leads (0.845), but the next four positions are essentially tied between time-series and EEG foundation models (Moirai-B 0.804, Chronos-B 0.800, Moirai-S 0.800, Chronos-L 0.799), making NMT the dataset on which TS-FMs come closest to parity with EEG-FMs; DeepSOZ-HEM (0.719) and the SHHS-pretrained sleep transformers (CoRe-Sleep 0.686, Sleep-TF 0.682) again sit at the bottom.

Two cross-cutting patterns emerge. First, calibration gaps are systematic rather than anomalous: across every (model, dataset) pair, the AUROC-to-balanced-accuracy gap (segment length in Figure 8) lies between 0.05 and 0.11, the expected offset between threshold-free discrimination and the probe's default 0.5 decision boundary, with no model showing a calibration anomaly that would disqualify a strong AUROC on balanced-accuracy grounds. Second, pretraining provides a small but consistent

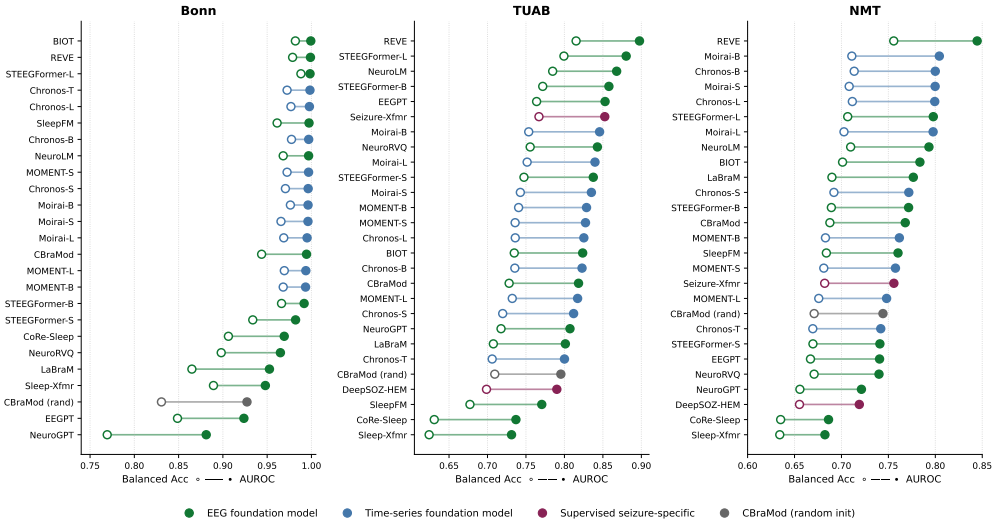

Figure 8: Foundation-model rankings on the three non-window EEG benchmarks (Bonn, TUAB, NMT). For each model and dataset, we plot AUROC (●) and balanced accuracy (○); the connector indicates the threshold-calibration gap (typically 0.05–0.11). Models are sorted within each panel by AUROC and colored by family: dark-green for general EEG foundation models, blue for time-series foundation models, burgundy for supervised seizure-specific baselines (SeizureTransformer, DeepSOZ-HEM), and grey for the CBraMod random-init ablation. All values are means over 5 stratified folds. REVE leads on all three datasets, Bonn is saturated, and NMT is the dataset on which TS-FMs come closest to parity with EEG-FMs.

boost: the CBraMod random-init points fall short of the pretrained CBraMod points by 0.067 AUROC on Bonn (0.994 → 0.927), 0.023 on TUAB (0.818 → 0.795), and 0.024 on NMT (0.768 → 0.744). The Bonn drop is largest in absolute terms but both endpoints remain near the saturation ceiling; on the more discriminating TUAB and NMT, pretraining buys roughly 2–3 percentage points of AUROC.

D.1.4 Statistical testing across epilepsy dataset

Ranking first on a single dataset does not establish broad competence: a model can lead one benchmark through alignment with a particular recording style or through an artefact of a single split. We therefore evaluate cross-dataset consistency directly. For each model we count the dataset on which it ranks in the top three by Event-Sens@FA AUC, and apply an exact one-sided binomial test against the chance rate $p_0 = 3/k$, where k is the size of the comparator pool. Results are reported for two pools on the 7 epileptic datasets. The first comprises all 23 models and tests competitiveness against the full benchmark, which includes supervised seizure detectors, time-series foundation models, and a random-initialised baseline. The second restricts to the 11 EEG foundation models and tests competitiveness within the family.

Table 3 reports the outcome. Two models reach significance in both pools at $\alpha = 0.05$. NeuroLM ranks top-3 on 6 of 7 datasets ($p < 10^{-3}$ in the full pool, $p = 0.002$ within EEG-FMs), and REVE on 5 of 7 in the full pool and 6 of 7 within EEG-FMs (both $p \leq 0.002$). No other model exceeds the chance rate at $\alpha = 0.05$ in either pool. The supervised seizure specialist Seizure-TF is the closest non-significant case at 3 of 7 wins ($p = 0.052$), indicating that task-specific supervision is strong on a subset of datasets but does not generalise reliably across recording styles. No time-series foundation model reaches significance, and the random-initialised CBraMod baseline ranks top-3 on no datasets, consistent with the score-distribution analysis in Section D.1.1. Among the eleven EEG foundation models, nine fail to exceed the chance rate even within the EEG-FM-only pool. The claim that "EEG foundation models work for seizure detection" is therefore supported, on this benchmark, by two specific models (NeuroLM and REVE) rather than by the family as a whole.

Table 3: **Binomial top-3 win test across the seven seizure-detection datasets.** For each model, the exact one-sided binomial test asks whether the number of datasets on which the model ranks in the top 3 (by Event-Sens vs FA/h AUC, integrated over $[0.1, 100]$ FA/h) exceeds the chance rate $p_0 = 3/k$. Two pools are tested independently: all $k = 23$ probed models, and the $k = 11$ EEG foundation models in isolation. Bold rows are significant at $\alpha = 0.05$.

		All models ($k = 23$)		EEG-FM ($k = 11$)	
Model	Family	Wins	p	Wins	p
NeuroLM	**EEG-FM**	**6/7**	$<10^{-3}$	**6/7**	**0.002**
REVE	**EEG-FM**	**5/7**	$<10^{-3}$	**6/7**	**0.002**
Seizure-TF	Supervised-Pre	3/7	0.052	–	–
Moirai-L	TS-FM	2/7	0.229	–	–
Chronos-B	TS-FM	1/7	0.624	–	–
Moirai-B	TS-FM	1/7	0.624	–	–
Moirai-S	TS-FM	1/7	0.624	–	–
NeuroRVQ	EEG-FM	1/7	0.624	1/7	0.892
SleepFM	EEG-FM	1/7	0.624	2/7	0.610
BIOT	EEG-FM	0/7	1.000	1/7	0.892
CBraMod	EEG-FM	0/7	1.000	1/7	0.892
CBraMod (rand)	Random-Init	0/7	1.000	–	–
Chronos-S	TS-FM	0/7	1.000	–	–
Chronos-T	TS-FM	0/7	1.000	–	–
CoRe-Sleep	Supervised-Pre	0/7	1.000	–	–
EEGPT	EEG-FM	0/7	1.000	1/7	0.892
LaBraM	EEG-FM	0/7	1.000	1/7	0.892
MOMENT-B	TS-FM	0/7	1.000	–	–
MOMENT-S	TS-FM	0/7	1.000	–	–
NeuroGPT	EEG-FM	0/7	1.000	0/7	1.000
Sleep-TF	Supervised-Pre	0/7	1.000	–	–
STEEGFormer-B	EEG-FM	0/7	1.000	2/7	0.610
STEEGFormer-S	EEG-FM	0/7	1.000	0/7	1.000

D.2 AUROC, balanced accuracy, and event-level metrics per model

To complement the per-dataset view in the main text, this section collapses the seven seizure-detection datasets into a single per-model summary across the three metrics. The pooled view makes it easier to see which models are competitive on average, how the random-init baseline is positioned once dataset-specific noise is averaged out, and how rankings shift between window-level (AUROC, balanced accuracy) and event-level (Event-Sens@FA AUC) metrics.

Table 4 reports across-dataset means and standard deviations for each model on the three metrics. Three observations stand out. First, only NeuroLM (0.553) and REVE (0.549) clearly exceed the random-init Event-Sens@FA AUC of 0.440; SleepFM (0.487) and STEEGFormer-B (0.451) sit just above, while the remaining seven EEG-FMs sit at or below random init, with NeuroGPT (0.277) far below. Pretrained CBraMod records the same 0.440 as CBraMod-rand, indicating that for some models pretraining does not deliver an event-level gain. Second, metric choice changes the ranking: REVE leads on AUROC (0.839) followed by Moirai-L (0.830) and NeuroLM (0.821), whereas the event-level ranking is NeuroLM, REVE, then Seizure-TF, with Moirai-L falling to fourth (consistent with Appendix D.1.2). Third, families do not separate cleanly: Seizure-TF reaches 0.519 on Event-Sens@FA partly through pretraining overlap with TUSZ and Siena, and Moirai-L outperforms 9 of 11 EEG-FMs on Event-Sens@FA despite no EEG-specific pretraining, so EEG-specific pretraining alone does not guarantee strong event-level performance.

D.3 Sleep

D.3.1 Sleep staging

Sleep staging performance per model and dataset is presented in Figure 9. Despite substantial variability across datasets, NeuroLM, and REVE consistently rank among the top-3-performing EEG-FMs. While both models evaluate sleep staging as a downstream task, only REVE incorporates

Table 4: **Per-model summary across the seven seizure-detection cohorts.** Each cell is the mean $\pm$ across-cohort standard deviation (7 datasets: TUSZ, Siena, SeizeIT1, CHB-MIT, SeizeIT2, Helsinki, Epilepsiae). Window-level AUROC and balanced accuracy come from the linear probe (5-fold subject-disjoint mean per cohort). The event-level metric is the area under the mean event sensitivity vs. $\log_{10}$(FA/h) curve, integrated over the clinically relevant $[0.1, 100]$ FA/h range under any-overlap event scoring. Models are grouped by family and sorted within family by Event-Sens vs FA/h AUC. *All 23 models in the canonical* 7×23 *panel are reported.*

Model	Family	AUROC	Balanced Acc.	Event-Sens vs FA/h AUC
NeuroLM	EEG-FM	0.821 ± 0.079	0.640 ± 0.050	0.553 ± 0.093
REVE	EEG-FM	0.839 ± 0.086	0.664 ± 0.069	0.549 ± 0.098
SleepFM	EEG-FM	0.789 ± 0.099	0.603 ± 0.052	0.487 ± 0.117
STEEGFormer-B	EEG-FM	0.776 ± 0.077	0.598 ± 0.061	0.451 ± 0.118
CBraMod	EEG-FM	0.787 ± 0.110	0.611 ± 0.070	0.440 ± 0.118
EEGPT	EEG-FM	0.752 ± 0.097	0.607 ± 0.081	0.431 ± 0.099
LaBraM	EEG-FM	0.764 ± 0.079	0.582 ± 0.062	0.418 ± 0.095
NeuroRVQ	EEG-FM	0.756 ± 0.111	0.596 ± 0.066	0.410 ± 0.095
BIOT	EEG-FM	0.756 ± 0.082	0.591 ± 0.054	0.408 ± 0.074
STEEGFormer-S	EEG-FM	0.746 ± 0.073	0.577 ± 0.051	0.405 ± 0.099
NeuroGPT	EEG-FM	0.618 ± 0.113	0.555 ± 0.050	0.277 ± 0.048
Seizure-TF	Supervised-Pre	0.763 ± 0.111	0.613 ± 0.074	0.519 ± 0.101
Sleep-TF	Supervised-Pre	0.667 ± 0.066	0.554 ± 0.050	0.357 ± 0.086
CoRe-Sleep	Supervised-Pre	0.668 ± 0.068	0.556 ± 0.050	0.354 ± 0.080
Moirai-L	TS-FM	0.830 ± 0.067	0.616 ± 0.053	0.474 ± 0.119
Moirai-B	TS-FM	0.807 ± 0.093	0.625 ± 0.053	0.453 ± 0.123
Chronos-S	TS-FM	0.755 ± 0.078	0.590 ± 0.062	0.443 ± 0.110
MOMENT-S	TS-FM	0.754 ± 0.069	0.590 ± 0.053	0.443 ± 0.068
Moirai-S	TS-FM	0.785 ± 0.138	0.601 ± 0.076	0.436 ± 0.160
Chronos-B	TS-FM	0.760 ± 0.083	0.583 ± 0.056	0.431 ± 0.094
MOMENT-B	TS-FM	0.756 ± 0.101	0.586 ± 0.065	0.428 ± 0.101
Chronos-T	TS-FM	0.743 ± 0.097	0.575 ± 0.058	0.422 ± 0.087
CBraMod (rand)	Random-Init	0.773 ± 0.086	0.590 ± 0.064	0.440 ± 0.102

sleep-specific data during pretraining; however, this dataset represents only 34 out of the 24,274 patients included across studies [59, 30]. Together with TS-FMs achieving performance comparable to EEG-FMs without relying on EEG-specific pretraining, this further supports the claim that such pretraining may be less influential than other design factors. Within TS-FMs, increasing the number of parameters does not consistently lead to improved performance within a given model family. Furthermore, no clear relationship is observed between performance and either dataset size or embedding dimensionality (Figure 10). On average, all models, except for the TS-FM Chronos, perform above the random initialized baseline.

Per-stage performance (Figure 11) indicates that wake is the most reliably classified stage across models, whereas N1 and REM are most frequently misclassified. While reduced performance on N1 is consistent with its low inter-rater reliability among human scorers, REM sleep is generally considered less ambiguous [47]. The discrepancy in REM classification performance may be attributed to the reliance on EEG-only inputs, whereas the AASM scoring criteria incorporate additional modalities, such as EOG and EMG, for REM identification [84]. Moreover, supervised models derive notable benefits from incorporating inter-epoch temporal context, particularly for N1 and REM stages, as AASM guidelines rely on information from adjacent epochs for accurate staging [84].

Sleep staging performance was further evaluated across dataset subsets by training a separate linear probe for each subset (Figure 12). Subsets recorded at different sites can show large variability, as acquisition protocols can differ in device, channels ... Contrary to expectations, clinical populations, i.e., including patients with disorders or experimental interventions, were not consistently more difficult to stage than healthy populations.

In datasets with annotations from multiple experts, sleep staging agreement was consistently higher when evaluated against consensus scoring (Figure 13). This aligns with expectations, as consensus labels tend to exclude ambiguous epochs, which are also considered hard to stage automatically.

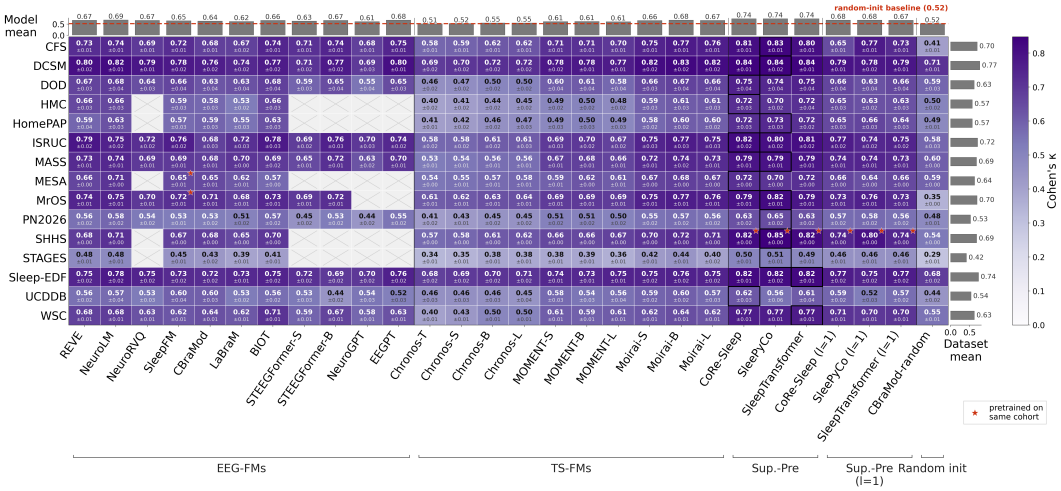

Figure 9: **Sleep staging performance is strongly dataset-dependent.** Cohen's κ sleep staging agreement with ground-truth labels for all datasets and models per model category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models, with and without temporal context (sequence length l=1), and a random-initialized reference. Cohen's κ (above) and standard deviation over folds (below) are shown for each. Black borders mark the cohort best-performing model. Red stars flag pretraining/evaluation overlap. Grey bars show mean performance over dataset and models respectively, with the red dashed line indicating mean random-initialized baseline. Missing results are indicated with a cross.

Table 5: **Binomial top-3 win test across nine sleep-staging cohorts.** For each model, the exact one-sided binomial test asks whether the number of cohorts on which the model ranks in the top 3 (by mean Cohen's κ) exceeds the chance rate $p_0 = 3/k$. Three pools are tested independently: all $k = 28$ probed models, $k = 25$ models excluding the Supervised-Pre models using sequence modeling, and the $k = 11$ EEG-FMs in isolation. Bold cells indicate significance at $\alpha = 0.05$.

		All models (k=28)		No Sup.-Pre (k=25)		EEG-FM only (k=11)	
Model	Family	Wins	p	Wins	p	Wins	p
$N = 9$ cohorts; $p_0 = 3/28 = 0.107$ vs. $3/25 = 0.120$ vs. $3/11 = 0.273$. Bold = significant at $\alpha = 0.05$.							
CoRe-Sleep	Sup.-Pre	**9/9**	$<10^{-3}$	–	–	–	–
SleepTransformer	Sup.-Pre	**9/9**	$<10^{-3}$	–	–	–	–
SleePyCo	Sup.-Pre	**8/9**	$<10^{-3}$	–	–	–	–
CBraMod	EEG-FM	1/9	0.639	1/9	0.684	1/9	0.943
Moirai-B	TS-FM	0/9	1.000	**5/9**	**0.002**	–	–
CoRe-Sleep (l=1)	Sup.-Pre (ℓ=1)	0/9	1.000	**5/9**	**0.002**	–	–
NeuroLM	EEG-FM	0/9	1.000	3/9	0.083	**7/9**	**0.002**
BIOT	EEG-FM	0/9	1.000	3/9	0.083	5/9	0.069
SleePyCo (l=1)	Sup.-Pre (ℓ=1)	0/9	1.000	3/9	0.083	–	–
Moirai-L	TS-FM	0/9	1.000	2/9	0.295	–	–
SleepTransformer (l=1)	Sup.-Pre (ℓ=1)	0/9	1.000	2/9	0.295	–	–
REVE	EEG-FM	0/9	1.000	1/9	0.684	**6/9**	**0.016**
SleepFM	EEG-FM	0/9	1.000	1/9	0.684	2/9	0.751
Moirai-S	TS-FM	0/9	1.000	1/9	0.684	–	–
EEGPT	EEG-FM	0/9	1.000	0/9	1.000	3/9	0.463
STEEGFormer-B	EEG-FM	0/9	1.000	0/9	1.000	2/9	0.751
NeuroRVQ	EEG-FM	0/9	1.000	0/9	1.000	1/9	0.943
LaBraM	EEG-FM	0/9	1.000	0/9	1.000	0/9	1.000
STEEGFormer-S	EEG-FM	0/9	1.000	0/9	1.000	0/9	1.000
NeuroGPT	EEG-FM	0/9	1.000	0/9	1.000	0/9	1.000
Chronos-T	TS-FM	0/9	1.000	0/9	1.000	–	–
Chronos-S	TS-FM	0/9	1.000	0/9	1.000	–	–
Chronos-B	TS-FM	0/9	1.000	0/9	1.000	–	–
Chronos-L	TS-FM	0/9	1.000	0/9	1.000	–	–
MOMENT-S	TS-FM	0/9	1.000	0/9	1.000	–	–
MOMENT-B	TS-FM	0/9	1.000	0/9	1.000	–	–
MOMENT-L	TS-FM	0/9	1.000	0/9	1.000	–	–
CBraMod-random	Random-Init	0/9	1.000	0/9	1.000	–	–

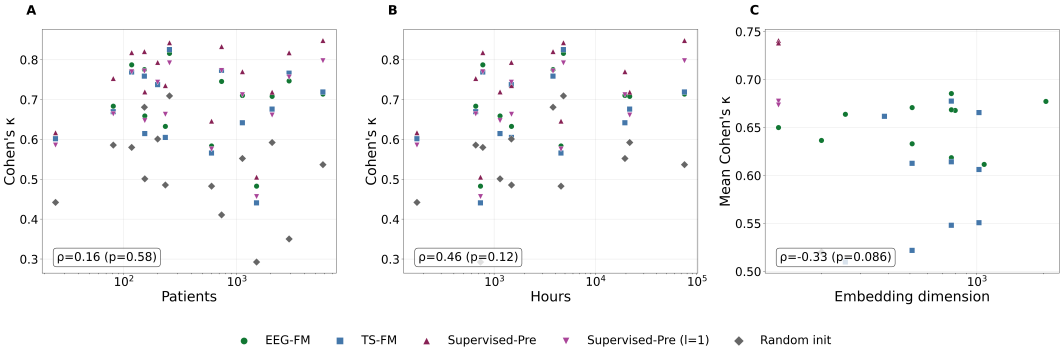

Figure 10: **Performance is not related to dataset size or embedding dimension. (a), (b)** Cohen's κ sleep staging agreement for best performing model per model category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models, with and without temporal context (sequence length l=1), and a random-initialized reference, is shown. **(a)** Dataset size in number of patients. **(b)** Dataset size in number of hours. **(c)** Mean Cohen's κ over datasets and embedding dimension.

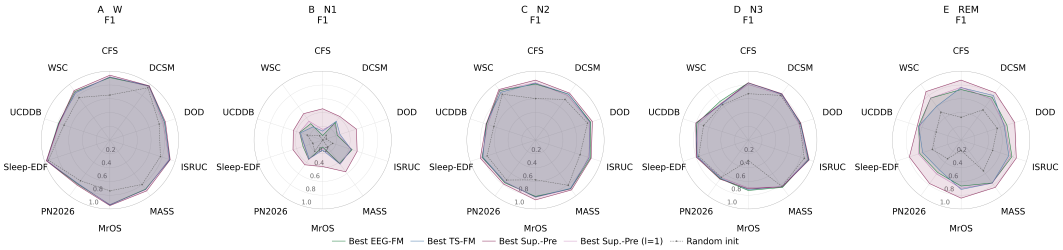

Figure 11: **Sleep staging performance is lowest for N1 and REM.** F1-score per sleep stage for best performing model per category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models, with and without temporal context (sequence length l=1), and a random-initialized reference.

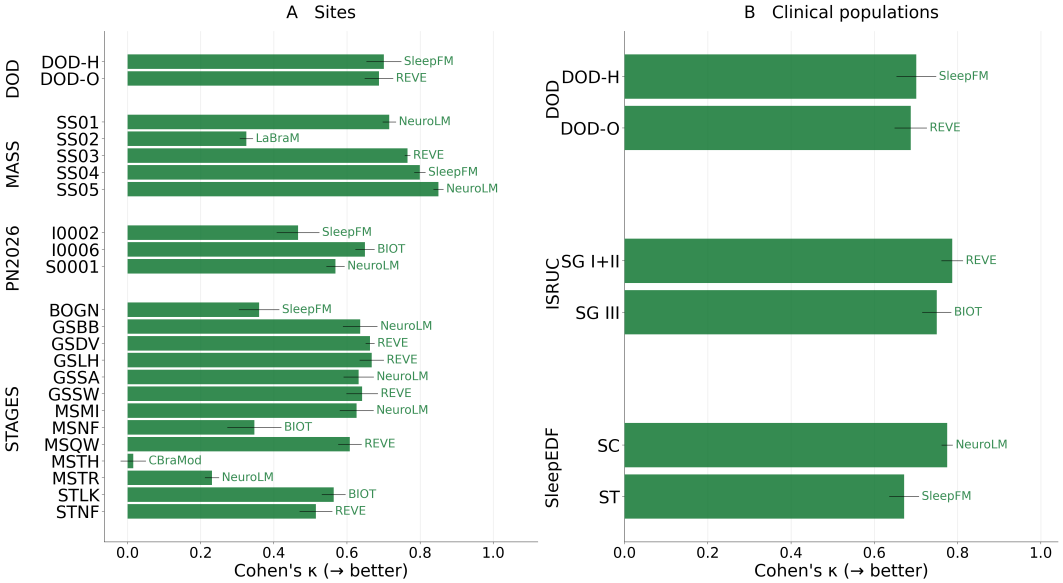

Figure 12: Cohen's κ sleep staging agreement for best performing EEG-FM per subset. **(a) Subsets recorded at different sites differ significantly due to differing acquisition protocols. (b) Clinical cohorts are not systematically associated with lower performance.** DOD-O contains patients with OSA, DOD-H contains healthy controls. ISRUC Subgroup I and II contain patients with sleep disorders, Subgroup III contains healthy people. Parkinson's PD contains patients with Parkinson's disease, HOA contains healthy older adults. Sleep-EDF Cassette (SC) contains healthy patients, Telemetry (ST) contains medicated people.

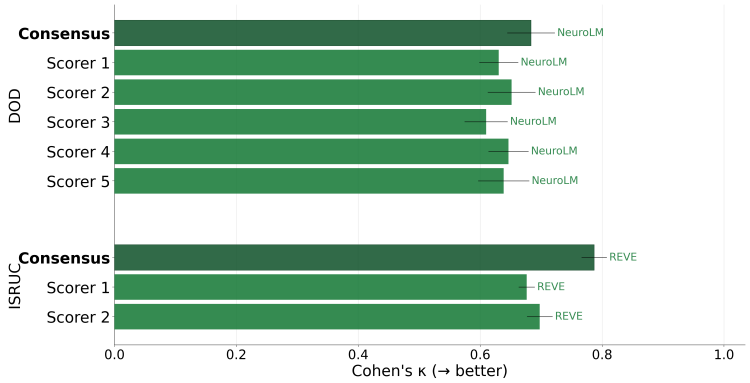

Figure 13: **Sleep staging agreement is highest with consensus scoring, as ambiguous epochs are excluded.** Cohen's κ sleep staging agreement with ground-truth labels scored by different experts, and consensus scoring for best performing EEG-FM.

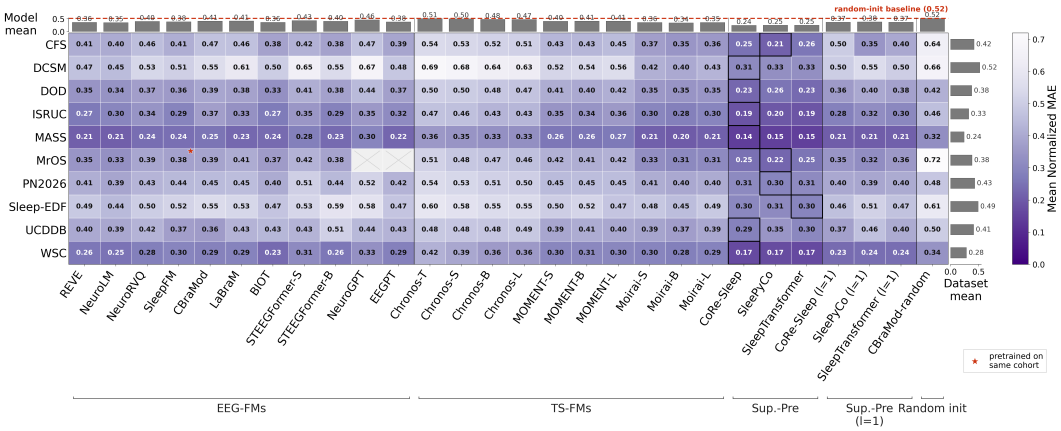

Figure 14: **Sleep staging and hypnogram feature performance are not consistently aligned across datasets.** Normalized Mean Absolute Error (MAE) between ground truth and predicted hypnogram features for all datasets and models per model category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models, with and without temporal context (sequence length l=1), and a random-initialized reference. Black borders mark the cohort best-performing model. Red stars flag pretraining/evaluation overlap. Grey bars show mean performance over dataset and models respectively, with the red dashed line indicating mean random-initialized baseline. Missing results are indicated with a cross.

D.3.2 Hypnogram features

Hypnogram-derived features used for downstream clinical diagnosis and analysis are presented in Figure 14. Similar to sleep staging, substantial variability is observed across datasets. however, datasets with strong staging performance do not necessarily achieve the best performance on hypnogram features. In addition, supervised models benefit from incorporating inter-epoch temporal context through their sequence-based architecture.

D.3.3 Event Detection

Similar to sleep staging, supervised models outperform FMs in microarousal detection, suggesting that FMs do not adequately encode these events, although all model categories exceed the chance baseline set by prevalence. Performance is more limited for respiratory events and limb movements. As these are primarily defined using other modalities (e.g., respiratory signals and EMG), their representation in EEG embeddings is less essential. Although event detection focuses on identifying discrete events rather than performing longitudinal staging, the inter-epoch temporal context captured by supervised models still provides a modest benefit.

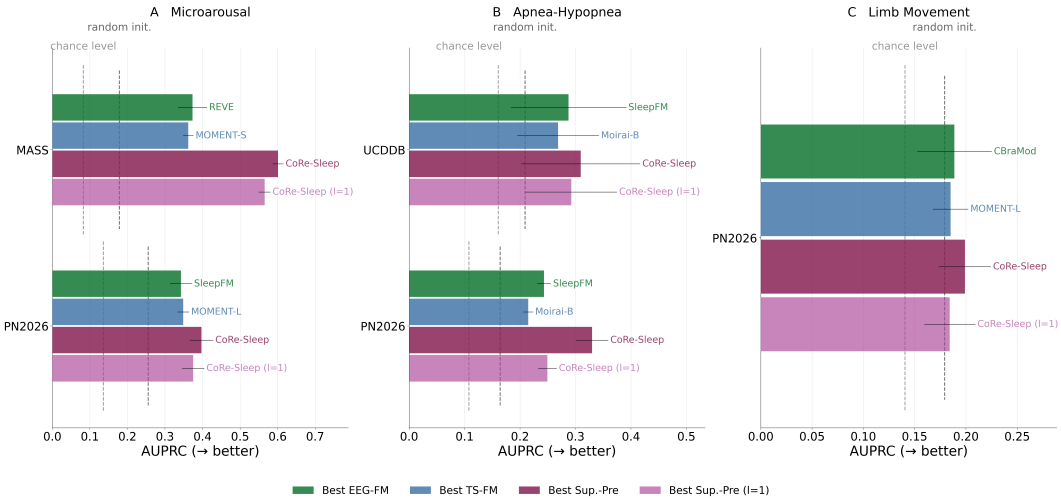

Figure 15: (**FMs fail to adequately capture microarousals in their embeddings.** Area under the precision-recall curve (AUPRC) for sleep event detection for best performing model per category. Prevalence indicates chance baseline. **(a)** Microarousals with a minimal duration of 3 seconds. **(b)** Apnea and hypopnea events with a minimal duration of 10 seconds. **(c)** Periodic limb movement with a minimal duration of 0.5 seconds.

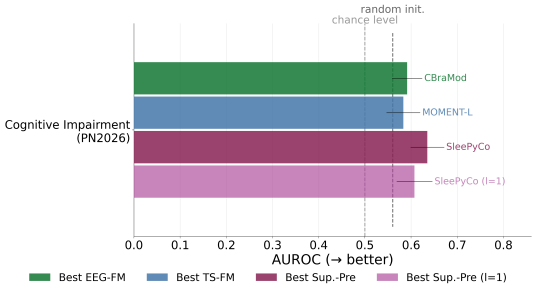

Figure 16: **Models insufficiently encode cognitive impairment for reliable detection.** Area under the receiver-operator curve (AUROC) for diagnosis of cognitive impairment in the PhysioNet 2026 dataset for best performing model per model category, i.e., EEG-FMs, TS-FMs, and Supervised-Pre models, with and without temporal context (sequence length l=1), and a random-initialized reference, is shown.

D.3.4 Cognitive Impairment Diagnosis

Recent studies have highlighted associations between sleep and cognitive impairment, and demonstrated that cognitive impairment can be detected in EEG recordings [48, 13]. However, the models evaluated on cognitive impairment detection in the PhysioNet 2026 dataset achieved only modest performance (AUROC around 0.60), suggesting that the extracted representations do not sufficiently encode features relevant for reliable detection of cognitive impairment.

D.4 Brain Age

Cohort- and model-family-level brain-age results. Detailed cohort- and model-family-level results are reported in Supplemental Tables 6, 7, 8, and 9. Across most cohorts and metrics, the supervised feature-based age-prediction model remains the strongest reference. One exception occurs in cohorts with a restricted age range, such as MrOS and WSC, where its advantage is reduced, suggesting that cohort-specific age distributions partly modulate apparent performance differences. Within each model family, performance also varies across datasets and metrics.

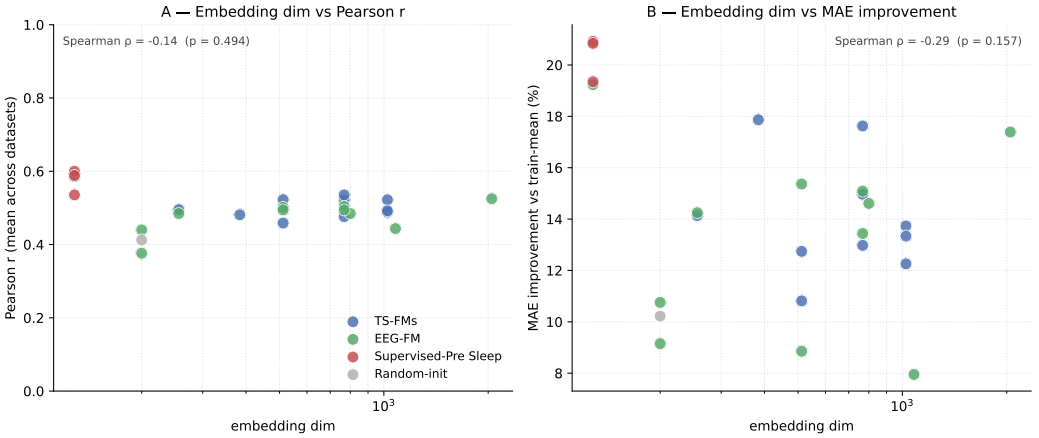

Figure 17: **Embedding dimensionality does not correlate with brain-age performance.** Each marker is one frozen-encoder checkpoint, with metrics fold-averaged across the seven main-pool brain-age cohorts (PN2026, WSC, ISRUC, SleepEDF SC, CFS, MrOS). Marker color denotes encoder family. **(A)** Embedding dim versus mean Pearson r between predicted and chronological age. **(B)** Embedding dim versus mean MAE improvement over a predict-cohort-mean baseline (%). The relationship is essentially flat in (A) (Spearman $\rho = -0.14$, $p = 0.49$) and weakly negative in (B) ($\rho = -0.29$, $p = 0.16$).

Effect of linear-probe regularization. The effect of probe regularization was examined on a representative subset of cohorts and models (Table 10). We compared the Ridge regression probe with an unregularized ordinary least squares (OLS) alternative. Regularization consistently improved stability, particularly in smaller cohorts. This analysis also indicates that the reported linear-probing results are partly shaped by the choice of regularization.

Relationship between embedding dimensionality and brain-age performance. We next examined whether brain-age performance scales with embedding dimensionality across the evaluated frozen-encoder checkpoints. Across 25 models, embedding dimension showed no significant association with the Pearson correlation metric and only a weak, non-significant negative association with MAE improvement over the train-mean baseline ($\rho = -0.29$, $p = 0.16$; Figure 17. Thus, larger embedding spaces did not translate into systematically better age prediction.

This lack of dimensional scaling is most apparent for supervised sleep-stage models, which use comparatively compact 128-dimensional representations yet achieve some of the strongest average MAE improvements. In contrast, higher-dimensional EEG and time-series foundation models are competitive but not consistently superior. These results suggest that the information encoded by the representation, rather than the dimensionality of the embedding space, is the primary determinant of brain-age regression performance.

Comparison of subject-level and epoch-level aggregation. Finally, we compared two aggregation strategies for using epoch-level embeddings. In the subject-level pipeline, epoch embeddings were averaged within each recording before fitting Ridge regression from the resulting subject-level vector to age. In the epoch-level pipeline, Ridge regression was fitted to individual 30-second epoch embeddings, and predicted ages were averaged within each recording. This comparison was performed on a subset of models and datasets. The subject-level pipeline generally outperformed the epoch-level predict-then-average strategy, supporting mean-pooled subject embeddings as the main approach for the reported brain-age analyses.

D.5 BCI

Statistical analysis. To identify which foundation models reliably rank among the best across BCI tasks, we apply a one-sided exact binomial test to each model's top-3 finish count. For each of the $N = 17$ datasets, we rank the $k = 7$ models by normalized balanced accuracy and record whether a model places in the top 3. Under the null hypothesis, each model has a $p_0 = 3/7 \approx 0.43$ probability

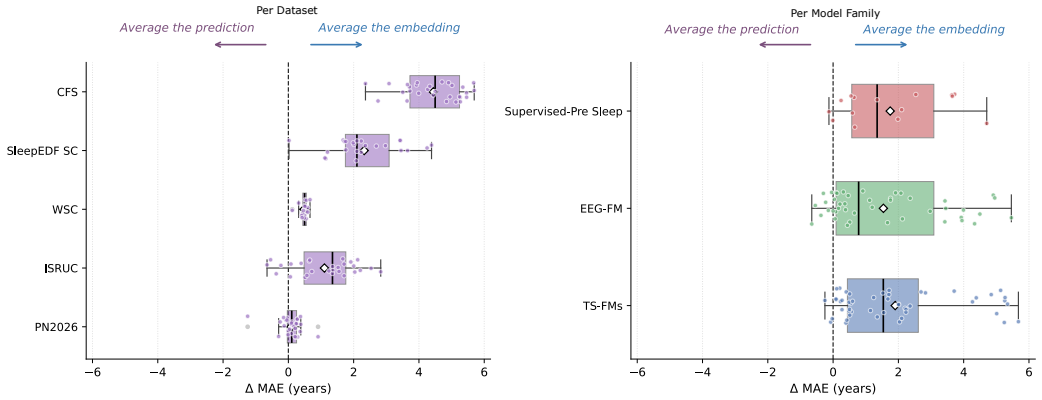

Figure 18: **Aggregating per subject tends to outperform aggregating per epoch.** Each marker is one (model, cohort) pair, fold-averaged across the 5 folds; The horizontal axis shows $\Delta\,\mathrm{MAE} = \mathrm{MAE}_{\mathrm{epoch}} - \mathrm{MAE}_{\mathrm{subject}}$; positive values indicate that subject-level Ridge (mean-pool a recording's epoch embeddings into one subject vector before fitting to age beats epoch-level Ridge fit per 30s epoch, then average predictions per subject). **(left)** Per-cohort distributions, ordered by $n_{\mathrm{outer_train}}$ ascending;

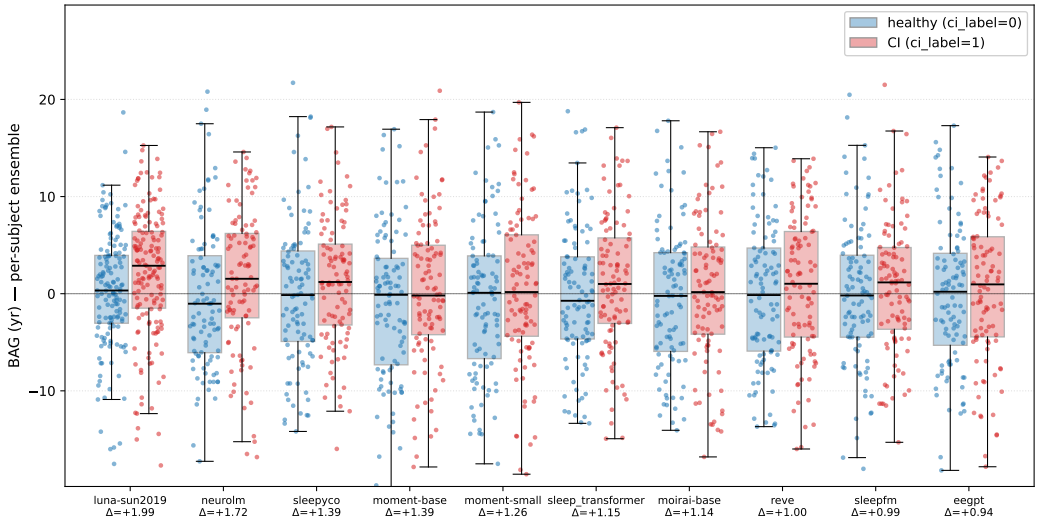

Figure 19: **Brain-age-gap (BAG) distributions on the PN2026 Healthy/Cognitive Impairment holdout.** BAG was computed per subject as predicted age minus chronological age using the 5-fold ensemble prediction. We show the 10 models with the largest positive mean BAG difference, ordered by $\Delta\mathrm{BAG} = \overline{\mathrm{BAG}}_{\mathrm{CI}} - \overline{\mathrm{BAG}}_{\mathrm{H}}$. For each model, box plots show the BAG distributions of 100 age-matched healthy subjects and 100 cognitively impaired (CI) subjects; dots indicate individual subjects. The horizontal line marks BAG $= 0$. Values below model names report ΔBAG in years, where positive values indicate older-appearing EEG in the CI group.

of finishing in the top 3 on any given dataset by chance. We then test whether the observed number of top-3 finishes significantly exceeds this rate using `scipy.stats.binomtest` (one-sided, $\alpha = 0.05$). Two models pass the significance threshold: **REVE** (14/18 datasets, $p = 0.003$) and **NeuroLM** (12/18, $p = 0.036$), indicating that both consistently outperform the other foundation models across paradigms. EEGPT narrowly misses significance (11/18, $p = 0.093$), while the remaining models, ST-EEG-S, LaBraM, CBraMod, and Moment-B do not finish in the top 3 more often than expected by chance.

Model	PN2026	WSC	ISRUC	SleepEDF SC	CFS	MrOS
EEG-FM						
BIOT	6.94 ± 0.60 (-0.72)	5.94 ± 0.21 (-0.50)	11.27 ± 1.25 (-2.51)	<u>11.98</u> ± 1.69 (-6.35)	9.38 ± 0.58 (-6.99)	3.99 ± 0.04 (-0.63)
CBraMod	7.57 ± 1.13 (-0.08)	5.88 ± 0.25 (-0.56)	11.78 ± 2.19 (-2.00)	15.55 ± 2.03 (-2.79)	9.53 ± 0.33 (-6.84)	4.22 ± 0.02 (-0.41)
LaBraM	7.47 ± 0.43 (-0.19)	5.81 ± 0.20 (-0.63)	10.64 ± 0.90 (-3.14)	15.26 ± 1.96 (-3.07)	9.05 ± 0.28 (-7.32)	4.21 ± 0.08 (-0.42)
NeuroLM-VQ	7.22 ± 0.57 (-0.44)	5.53 ± 0.18 (-0.91)	10.95 ± 1.22 (-2.83)	13.13 ± 1.88 (-5.21)	7.78 ± 0.43 (-8.59)	3.86 ± 0.10 (-0.76)
REVE	<u>6.72</u> ± 0.27 (-0.93)	5.83 ± 0.24 (-0.61)	11.24 ± 2.00 (-2.54)	14.33 ± 3.05 (-4.00)	8.81 ± 0.41 (-7.56)	3.85 ± 0.06 (-0.78)
SleepFM	7.28 ± 0.34 (-0.37)	5.78 ± 0.33 (-0.67)	9.40 ± 1.25 (-4.39)	12.73 ± 1.09 (-5.60)	8.81 ± 0.91 (-7.56)	3.95 ± 0.08 (-0.68)
EEGPT	7.54 ± 0.43 (-0.11)	**5.17** ± 0.20 (-1.27)	<u>8.99</u> ± 1.87 (-4.80)	14.31 ± 2.79 (-4.02)	<u>7.67</u> ± 0.58 (-8.70)	3.89 ± 0.11 (-0.74)
NeuroRVQ-EEG	7.06 ± 1.02 (-0.60)	5.34 ± 0.17 (-1.10)	11.94 ± 1.48 (-1.84)	12.34 ± 2.61 (-6.00)	8.12 ± 0.62 (-8.25)	<u>3.81</u> ± 0.10 (-0.81)
Neuro-GPT	7.23 ± 0.51 (-0.42)	6.06 ± 0.18 (-0.38)	12.60 ± 1.89 (-1.19)	14.56 ± 1.66 (-3.78)	9.65 ± 0.47 (-6.72)	4.20 ± 0.08 (-0.42)
STEEGFormer-S	7.48 ± 0.84 (-0.18)	5.67 ± 0.23 (-0.77)	11.79 ± 1.16 (-1.99)	15.23 ± 1.85 (-3.11)	8.36 ± 0.69 (-8.01)	3.93 ± 0.04 (-0.70)
STEEGFormer-B	7.23 ± 1.03 (-0.43)	5.56 ± 0.23 (-0.88)	10.82 ± 1.11 (-2.97)	14.24 ± 1.27 (-4.09)	8.26 ± 0.34 (-8.11)	3.86 ± 0.03 (-0.76)
Random-init						
CBraMod (random init)	7.07 ± 0.59 (-0.58)	5.83 ± 0.22 (-0.61)	11.54 ± 2.05 (-2.24)	12.30 ± 2.04 (-6.04)	11.23 ± 0.57 (-5.14)	4.60 ± 0.03 (-0.03)
TS-FMs						
Chronos-T5 (tiny)	7.68 ± 0.94 (+0.03)	5.88 ± 0.23 (-0.57)	10.91 ± 1.58 (-2.88)	<u>11.69</u> ± 0.90 (-6.64)	9.44 ± 0.64 (-6.93)	4.08 ± 0.06 (-0.55)
Chronos-T5 (small)	7.46 ± 0.73 (-0.19)	5.84 ± 0.25 (-0.60)	11.34 ± 1.46 (-2.44)	13.33 ± 1.43 (-5.01)	8.74 ± 0.37 (-7.63)	4.16 ± 0.04 (-0.47)
Chronos-T5 (base)	7.66 ± 0.80 (+0.00)	5.75 ± 0.23 (-0.69)	10.77 ± 1.27 (-3.01)	13.28 ± 2.81 (-5.06)	8.66 ± 0.72 (-7.71)	4.13 ± 0.03 (-0.50)
Chronos-T5 (large)	7.63 ± 0.78 (-0.03)	5.73 ± 0.24 (-0.72)	12.15 ± 1.97 (-1.63)	12.71 ± 1.85 (-5.63)	8.77 ± 0.52 (-7.59)	4.14 ± 0.01 (-0.49)
Moirai (small)	7.13 ± 0.69 (-0.53)	5.78 ± 0.30 (-0.67)	<u>9.13</u> ± 1.57 (-4.65)	14.04 ± 2.60 (-4.29)	8.89 ± 0.40 (-7.48)	4.07 ± 0.16 (-0.56)
Moirai (base)	7.10 ± 0.69 (-0.56)	5.75 ± 0.32 (-0.69)	10.22 ± 1.49 (-3.57)	14.03 ± 2.18 (-4.30)	8.02 ± 0.18 (-8.35)	4.17 ± 0.27 (-0.46)
Moirai (large)	<u>7.08</u> ± 0.69 (-0.58)	5.88 ± 0.32 (-0.57)	10.03 ± 1.57 (-3.75)	14.07 ± 3.18 (-4.26)	9.15 ± 0.46 (-7.22)	4.13 ± 0.05 (-0.50)
MOMENT (small)	7.11 ± 0.67 (-0.55)	5.79 ± 0.18 (-0.65)	10.27 ± 1.20 (-3.51)	14.30 ± 2.07 (-4.03)	8.31 ± 0.31 (-8.06)	3.99 ± 0.05 (-0.64)
MOMENT (base)	7.26 ± 0.59 (-0.40)	<u>5.70</u> ± 0.13 (-0.74)	10.20 ± 0.98 (-3.58)	12.62 ± 2.54 (-5.72)	<u>7.99</u> ± 0.57 (-8.38)	**3.71** ± 0.08 (-0.92)
MOMENT (large)	7.33 ± 0.63 (-0.32)	5.90 ± 0.19 (-0.55)	11.27 ± 1.93 (-2.52)	13.48 ± 2.26 (-4.86)	8.28 ± 0.27 (-8.09)	3.99 ± 0.04 (-0.64)
Supervised-Pre Sleep						
CoreSleep	7.15 ± 0.65 (-0.51)	5.43 ± 0.19 (-1.02)	<u>8.80</u> ± 1.42 (-4.98)	<u>10.39</u> ± 1.07 (-7.94)	8.17 ± 0.32 (-8.20)	<u>3.89</u> ± 0.09 (-0.74)
SleepTransformer	7.15 ± 0.46 (-0.51)	<u>5.43</u> ± 0.14 (-1.02)	10.30 ± 1.32 (-3.49)	11.75 ± 2.67 (-6.58)	<u>7.87</u> ± 0.30 (-8.50)	3.93 ± 0.06 (-0.70)
SleepyCo	<u>6.88</u> ± 0.67 (-0.78)	5.51 ± 0.20 (-0.94)	10.43 ± 1.58 (-3.35)	13.02 ± 2.51 (-5.31)	8.01 ± 0.48 (-8.35)	3.93 ± 0.05 (-0.70)
Supervised-Pre Age						
Sun et al. (2019)	**4.86** ± 0.48 (-2.79)	5.65 ± 0.20 (-0.79)	**5.54** ± 0.70 (-8.25)	**6.09** ± 0.89 (-12.25)	**6.22** ± 0.40 (-10.15)	4.98 ± 0.19 (+0.35)
Predict train-mean						
Predict train-mean	7.66 ± 0.82	6.44 ± 0.19	13.78 ± 1.01	18.33 ± 0.40	16.37 ± 0.59	4.63 ± 0.02

Table 6: Overall brain-age prediction performance. Cells report cv-test MAE in years, shown as mean ± standard deviation across five outer folds. Gray parentheses show the difference relative to the predict-train-mean baseline. The best non-baseline MAE per dataset is bolded, and the second-best is underlined.

Confound filtering on other BCI paragigms does not lead to significant changes in performance. The comparison between no filtering and confound filtering shows that adopting a more task-specific filtering strategy does not substantially alter the performance of frozen EEG foundation models relative to the filtering configurations originally suggested by each model. Here, confound filtering refers to paradigm-specific preprocessing designed to retain task-relevant neural activity while reducing potential recording-related artifacts: ERP datasets were filtered using a 0.5–40,Hz band-pass filter, SSVEP datasets were only high-pass filtered without low-pass filtering in order to preserve stimulus harmonics, and Cognitive datasets were filtered within the 4–40,Hz range. As seen in 20 Across ERP, SSVEP, and Cognitive paradigms, the observed differences between filtering strategies are generally small and fail to reach statistical significance. Moreover, with the exception of the SSVEP paradigm, most results remain below the aggregated statistical significance threshold, indicating that current frozen EEG foundation models still struggle to consistently extract robust task-relevant representations across paradigms. These findings suggest that model architecture and pretraining strategy currently have a stronger impact on downstream decoding performance than the specific filtering configuration applied during evaluation.

Best performing model families per paradigm Figure 21 provides a per-dataset comparison of the best-performing model families under task-specific confound filtering using radar plots of normalized balanced accuracy. Each spoke corresponds to an individual dataset, while the radial axis maps chance-level performance to 0.5 and perfect classification to 1.0, allowing direct comparison across paradigms with different numbers of classes and intrinsic task difficulty. For each family, we report the best-performing model across all members on each dataset, highlighting the strongest achievable performance of EEG foundation models (EEG-FM), time-series foundation models (TS-FM), and supervised baselines under a unified frozen-evaluation framework.

Model	PN2026	WSC	ISRUC	SleepEDF SC	CFS	MrOS
EEG-FM						
BIOT	0.407 ± 0.126	0.393 ± 0.037	0.472 ± 0.112	<u>0.806</u> ± 0.087	0.782 ± 0.029	0.473 ± 0.012
CBraMod	0.294 ± 0.110	0.407 ± 0.032	0.477 ± 0.207	0.566 ± 0.110	0.779 ± 0.015	0.395 ± 0.011
LaBraM	0.257 ± 0.172	0.409 ± 0.061	0.552 ± 0.091	0.587 ± 0.112	0.803 ± 0.021	0.380 ± 0.056
NeuroLM-VQ	0.400 ± 0.108	0.498 ± 0.045	0.556 ± 0.079	0.731 ± 0.109	<u>0.855</u> ± 0.023	0.488 ± 0.062
REVE	<u>0.458</u> ± 0.090	0.420 ± 0.060	0.493 ± 0.131	0.681 ± 0.123	0.806 ± 0.019	<u>0.527</u> ± 0.015
SleepFM	0.356 ± 0.135	0.431 ± 0.059	0.696 ± 0.048	0.726 ± 0.067	0.815 ± 0.036	0.489 ± 0.027
EEGPT	0.252 ± 0.161	<u>0.646</u> ± 0.019	<u>0.712</u> ± 0.074	0.609 ± 0.142	0.855 ± 0.033	0.418 ± 0.077
NeuroRVQ-EEG	0.421 ± 0.114	0.624 ± 0.022	0.409 ± 0.125	0.767 ± 0.079	0.840 ± 0.038	0.501 ± 0.108
Neuro-GPT	0.287 ± 0.093	0.491 ± 0.018	0.364 ± 0.185	0.637 ± 0.089	0.773 ± 0.012	0.381 ± 0.034
STEEGFormer-S	0.264 ± 0.096	0.562 ± 0.032	0.459 ± 0.121	0.605 ± 0.093	0.833 ± 0.023	0.468 ± 0.035
STEEGFormer-B	0.339 ± 0.064	0.541 ± 0.108	0.543 ± 0.092	0.687 ± 0.069	0.843 ± 0.011	0.454 ± 0.066
Random-init						
CBraMod (random init)	0.389 ± 0.102	0.431 ± 0.031	0.419 ± 0.235	0.756 ± 0.096	0.686 ± 0.048	0.126 ± 0.025
TS-FMs						
Chronos-T5 (tiny)	0.313 ± 0.082	0.423 ± 0.022	0.617 ± 0.069	<u>0.783</u> ± 0.055	0.784 ± 0.035	0.445 ± 0.022
Chronos-T5 (small)	0.291 ± 0.105	0.417 ± 0.042	0.517 ± 0.134	0.713 ± 0.082	0.820 ± 0.013	0.410 ± 0.037
Chronos-T5 (base)	0.255 ± 0.105	0.449 ± 0.029	0.583 ± 0.102	0.725 ± 0.066	0.824 ± 0.031	0.437 ± 0.020
Chronos-T5 (large)	0.249 ± 0.101	0.454 ± 0.022	0.480 ± 0.152	0.731 ± 0.051	0.819 ± 0.017	0.441 ± 0.015
Moirai (small)	0.350 ± 0.175	0.442 ± 0.041	<u>0.694</u> ± 0.122	0.642 ± 0.152	0.809 ± 0.017	0.377 ± 0.144
Moirai (base)	0.377 ± 0.130	0.454 ± 0.043	0.578 ± 0.236	0.643 ± 0.128	0.848 ± 0.017	0.359 ± 0.155
Moirai (large)	0.355 ± 0.188	0.417 ± 0.040	0.611 ± 0.148	0.633 ± 0.185	0.793 ± 0.038	0.392 ± 0.100
MOMENT (small)	<u>0.410</u> ± 0.080	0.453 ± 0.020	0.626 ± 0.056	0.642 ± 0.165	0.844 ± 0.014	0.455 ± 0.041
MOMENT (base)	0.347 ± 0.106	<u>0.465</u> ± 0.038	0.645 ± 0.050	0.725 ± 0.129	<u>0.848</u> ± 0.032	<u>0.567</u> ± 0.009
MOMENT (large)	0.347 ± 0.107	0.414 ± 0.025	0.528 ± 0.114	0.735 ± 0.082	0.841 ± 0.014	0.470 ± 0.034
Supervised-Pre Sleep						
CoreSleep	0.353 ± 0.163	0.529 ± 0.022	<u>0.710</u> ± 0.069	<u>0.825</u> ± 0.046	0.843 ± 0.014	<u>0.516</u> ± 0.021
SleepTransformer	0.379 ± 0.169	<u>0.530</u> ± 0.028	0.604 ± 0.090	0.759 ± 0.168	<u>0.856</u> ± 0.015	0.500 ± 0.015
SleepyCo	<u>0.432</u> ± 0.132	0.509 ± 0.022	0.554 ± 0.157	0.809 ± 0.033	0.850 ± 0.013	0.502 ± 0.022
Supervised-Pre Age						
Sun et al. (2019)	**0.786** ± 0.070	**0.770** ± 0.010	**0.922** ± 0.010	**0.969** ± 0.014	**0.933** ± 0.009	**0.586** ± 0.026

Table 7: Overall brain-age prediction Pearson correlation between predicted and chronological age. Cells report cv-test Pearson r, shown as mean ± standard deviation across five outer folds. The best model per dataset is bolded, and the best model within each model family is underlined.

For the MI paradigm, we additionally compare against pretrained supervised models obtained by [25], which is why EEGNet [45] is only included in this section. To ensure fair comparison with EEG-FMs and TS-FMs, embeddings were extracted from the pretrained supervised models using the same frozen-encoder protocol employed for the foundation models, followed by linear probing using a leave-one-subject-out (LOSO) evaluation strategy. The same evaluation pipeline was applied under task-specific confound filtering, where the linear classification layer was adapted to account for the reduced temporal dimensionality introduced by the three-second filtered segments. For MIRepNet [50], we similarly froze the pretrained encoder and extracted embeddings before linear probing. MIRepNet was originally trained as a paradigm-specific motor imagery foundation model using a hybrid pretraining strategy combining supervised MI classification with self-supervised masked token reconstruction, together with MI-specific preprocessing including 8–30 Hz filtering, temporal alignment, and channel-template normalization designed to capture sensorimotor rhythm dynamics associated with ERD/ERS phenomena in motor imagery tasks. Although MIRepNet follows a supervised pretraining philosophy conceptually related to EEGNet-style approaches, it additionally incorporates self-supervised representation learning and large-scale MI-specific pretraining across multiple datasets. Nevertheless, despite these specialized training strategies, neither EEGNet nor MIRepNet outperformed the best EEG foundation models in the frozen linear-probing setting, and their normalized performance generally remained close to chance level, suggesting that supervised specialization alone does not necessarily yield more transferable EEG representations under strict cross-subject frozen evaluation protocols.

Model	SHHS	STAGES	MESA	HPAP
EEG-FM				
BIOT	7.53 ± 0.21 (-1.78)	–	7.40 ± 0.14 (-0.36)	9.08 ± 1.03 (-0.69)
CBraMod	7.53 ± 0.21 (-1.78)	10.69 ± 0.28 (-1.39)	7.15 ± 0.13 (-0.61)	9.24 ± 0.90 (-0.53)
LaBraM	7.00 ± 0.11 (-2.32)	10.42 ± 0.29 (-1.66)	6.99 ± 0.04 (-0.77)	9.18 ± 1.10 (-0.59)
NeuroLM-VQ	<u>6.11</u> ± 0.11 (-3.21)	<u>8.45</u> ± 0.53 (-3.63)	<u>6.15</u> ± 0.17 (-1.61)	9.28 ± 2.37 (-0.49)
REVE	6.70 ± 0.16 (-2.61)	9.77 ± 0.55 (-2.32)	6.81 ± 0.08 (-0.95)	8.91 ± 1.20 (-0.86)
SleepFM	6.53 ± 0.10 (-2.79)	10.06 ± 0.23 (-2.03)	6.42 ± 0.08 (-1.34)	<u>8.64</u> ± 1.06 (-1.12)
Random-init				
CBraMod (random init)	7.37 ± 0.10 (-1.94)	11.54 ± 0.24 (-0.54)	6.98 ± 0.19 (-0.78)	8.98 ± 1.24 (-0.79)
TS-FMs				
Chronos-T5 (tiny)	6.84 ± 0.08 (-2.47)	10.78 ± 0.33 (-1.31)	6.71 ± 0.11 (-1.04)	9.28 ± 0.94 (-0.49)
Chronos-T5 (small)	6.66 ± 0.10 (-2.65)	10.49 ± 0.20 (-1.59)	6.55 ± 0.12 (-1.21)	9.13 ± 0.88 (-0.63)
Chronos-T5 (base)	6.22 ± 0.11 (-3.10)	10.47 ± 0.30 (-1.61)	6.53 ± 0.05 (-1.23)	8.92 ± 0.61 (-0.84)
Chronos-T5 (large)	6.21 ± 0.17 (-3.11)	10.06 ± 0.26 (-2.02)	6.43 ± 0.17 (-1.33)	8.90 ± 0.65 (-0.86)
Moirai (small)	6.68 ± 0.22 (-2.64)	10.01 ± 0.13 (-2.07)	6.70 ± 0.10 (-1.06)	8.90 ± 0.71 (-0.87)
Moirai (base)	6.31 ± 0.23 (-3.00)	9.46 ± 0.26 (-2.62)	6.28 ± 0.08 (-1.48)	8.81 ± 0.63 (-0.96)
Moirai (large)	6.52 ± 0.15 (-2.79)	9.94 ± 0.17 (-2.14)	6.55 ± 0.10 (-1.21)	8.76 ± 0.78 (-1.01)
MOMENT (small)	6.18 ± 0.13 (-3.14)	9.05 ± 0.28 (-3.04)	6.16 ± 0.13 (-1.60)	<u>8.42</u> ± 0.72 (-1.35)
MOMENT (base)	<u>6.00</u> ± 0.14 (-3.32)	<u>9.05</u> ± 0.13 (-3.04)	<u>6.01</u> ± 0.10 (-1.75)	8.95 ± 0.86 (-0.82)
MOMENT (large)	6.11 ± 0.17 (-3.21)	9.51 ± 0.39 (-2.57)	6.26 ± 0.06 (-1.50)	9.14 ± 0.97 (-0.62)
Supervised-Pre Sleep				
CoreSleep	6.05 ± 0.10 (-3.27)	8.05 ± 0.24 (-4.04)	6.29 ± 0.13 (-1.47)	<u>7.85</u> ± 1.17 (-1.92)
SleepTransformer	<u>6.00</u> ± 0.08 (-3.31)	<u>7.83</u> ± 0.24 (-4.25)	<u>6.17</u> ± 0.18 (-1.59)	8.33 ± 1.25 (-1.44)
SleepyCo	6.07 ± 0.09 (-3.24)	8.28 ± 0.53 (-3.80)	6.35 ± 0.21 (-1.41)	7.92 ± 1.17 (-1.85)
Supervised-Pre Age				
Sun et al. (2019)	**4.72** ± 0.11 (-4.59)	**7.06** ± 0.13 (-5.02)	**5.96** ± 0.17 (-1.80)	**6.66** ± 0.71 (-3.11)
Predict train-mean				
Predict train-mean	9.31 ± 0.08	12.08 ± 0.31	7.76 ± 0.08	9.77 ± 1.00

Table 8: Brain-age prediction performance on the remaining cohorts: SHHS, STAGES, MESA, and HPAP. Cells report cv-test MAE in years, shown as mean ± standard deviation across five outer folds. Gray parentheses show the difference relative to the predict-train-mean baseline. The best non-baseline MAE per dataset is bolded, and the second-best is underlined.

Model	SHHS	STAGES	MESA	HPAP
EEG-FM				
BIOT	0.457 ± 0.127	—	0.252 ± 0.027	0.397 ± 0.156
CBraMod	0.546 ± 0.022	0.458 ± 0.060	0.338 ± 0.029	0.332 ± 0.131
LaBraM	0.615 ± 0.013	0.493 ± 0.057	0.393 ± 0.023	0.387 ± 0.133
NeuroLM-VQ	<u>0.713</u> ± 0.008	<u>0.656</u> ± 0.056	<u>0.563</u> ± 0.022	0.404 ± 0.158
REVE	0.655 ± 0.014	0.565 ± 0.067	0.428 ± 0.018	0.420 ± 0.125
SleepFM	0.675 ± 0.007	0.542 ± 0.042	0.509 ± 0.013	<u>0.424</u> ± 0.086
Random-init				
CBraMod (random init)	0.565 ± 0.018	0.339 ± 0.032	0.395 ± 0.036	0.420 ± 0.152
TS-FMs				
Chronos-T5 (tiny)	0.628 ± 0.008	0.456 ± 0.044	0.460 ± 0.021	0.342 ± 0.075
Chronos-T5 (small)	0.654 ± 0.014	0.483 ± 0.035	0.485 ± 0.026	0.379 ± 0.077
Chronos-T5 (base)	0.693 ± 0.030	0.492 ± 0.053	0.494 ± 0.021	0.390 ± 0.102
Chronos-T5 (large)	0.686 ± 0.041	0.541 ± 0.020	0.515 ± 0.019	0.386 ± 0.097
Moirai (small)	0.619 ± 0.061	0.541 ± 0.040	0.463 ± 0.035	0.446 ± 0.092
Moirai (base)	0.632 ± 0.147	0.597 ± 0.041	0.541 ± 0.027	0.449 ± 0.099
Moirai (large)	0.647 ± 0.041	0.538 ± 0.056	0.484 ± 0.029	0.455 ± 0.086
MOMENT (small)	0.699 ± 0.025	0.634 ± 0.022	0.564 ± 0.023	<u>0.471</u> ± 0.060
MOMENT (base)	<u>0.724</u> ± 0.020	<u>0.637</u> ± 0.022	<u>0.589</u> ± 0.012	0.388 ± 0.064
MOMENT (large)	0.689 ± 0.054	0.584 ± 0.034	0.540 ± 0.026	0.400 ± 0.065
Supervised-Pre Sleep				
CoreSleep	0.732 ± 0.008	0.720 ± 0.023	0.551 ± 0.029	<u>0.584</u> ± 0.153
SleepTransformer	<u>0.737</u> ± 0.007	<u>0.745</u> ± 0.026	<u>0.563</u> ± 0.029	0.491 ± 0.139
SleepyCo	0.727 ± 0.004	0.700 ± 0.054	0.524 ± 0.037	0.560 ± 0.155
Supervised-Pre Age				
Sun et al. (2019)	**0.861** ± 0.008	**0.870** ± 0.017	**0.729** ± 0.017	**0.810** ± 0.062

Table 9: Brain-age prediction Pearson correlation on the remaining cohorts: SHHS, STAGES, MESA, and HPAP. Cells report Pearson r between predicted and chronological age on cv-test folds, shown as mean ± standard deviation across five outer folds. The best model per dataset is bolded, and the best model within each model family is underlined.

Table 10: **Ridge regularization stabilizes linear probing, especially in smaller cohorts.** Baseline MAE denotes the dataset-specific dummy error obtained by predicting the training-set mean age. Δ denotes OLS MAE minus Ridge MAE; positive values indicate that Ridge performs better. Dataset-level values are aggregated across evaluated checkpoints within each cohort, whereas model family-level values are aggregated across all model–dataset evaluations within each model family.

A. Dataset-level summary

Dataset	n_{train}	Baseline MAE	Ridge MAE	OLS MAE	Δ
WSC	2055	6.44	5.71 ± 0.18	5.89 ± 0.36	$+0.17$
CFS	584	16.37	8.55 ± 0.62	12.04 ± 3.51	$+3.49$
PhysioNet 2026	169	7.66	7.23 ± 0.27	18.79 ± 13.75	$+11.56$
ISRUC	99	13.78	10.51 ± 0.86	19.69 ± 5.13	$+9.18$
SleepEDF-SC	62	18.33	13.19 ± 1.54	18.10 ± 4.31	$+4.92$

B. Model-family summary

Family	Ridge MAE	OLS MAE	Δ
EEG-FM	9.32 ± 2.97	15.81 ± 10.31	$+6.49$
Supervised-Pre Sleep	8.24 ± 2.11	13.27 ± 7.50	$+5.02$
TS-FMs	9.00 ± 2.74	14.17 ± 4.86	$+5.17$

Table 11: **Binomial top-3 win test across the 7 main-pool brain-age cohorts (MAE).** For each model, the exact one-sided binomial test asks whether the number of cohorts on which the model ranks in the top 3 (by cv-test mean absolute error, lower-is-better) exceeds the chance rate $p_0 = 3/k$. Two pools are tested independently: all $k = 26$ probed models, and the $k = 11$ EEG foundation models in isolation. Bold rows are significant at $\alpha = 0.05$.

		All models ($k = 26$)		EEG-FM only ($k = 11$)	
Model	Family	Wins	p	Wins	p
$N = 7$ cohorts; $p_0 = 3/26 = 0.115$ vs. $3/11 = 0.273$. Bold = significant at $\alpha = 0.05$.					
Sun et al. (2019)	**Supervised-Pre Age**	**5/7**	$< 10^{-3}$	–	–
EEGPT	**EEG-FM**	**3/7**	**0.038**	**3/7**	**0.292**
NeuroRVQ-EEG	**EEG-FM**	**2/7**	0.189	**5/7**	**0.019**
REVE	EEG-FM	2/7	0.189	3/7	0.292
CoreSleep	Supervised-Pre Sleep	2/7	0.189	–	–
SleepTransformer	Supervised-Pre Sleep	2/7	0.189	–	–
SleepFM	EEG-FM	1/7	0.576	3/7	0.292
NeuroLM-VQ	EEG-FM	1/7	0.576	2/7	0.610
Chronos-T5 (tiny)	TS-FMs	1/7	0.576	–	–
MOMENT (base)	TS-FMs	1/7	0.576	–	–
SleepyCo	Supervised-Pre Sleep	1/7	0.576	–	–
BIOT	EEG-FM	0/7	1.000	2/7	0.610
STEEGFormer-B	EEG-FM	0/7	1.000	2/7	0.610
LaBraM	EEG-FM	0/7	1.000	1/7	0.892
CBraMod	EEG-FM	0/7	1.000	0/7	1.000
Neuro-GPT	EEG-FM	0/7	1.000	0/7	1.000
STEEGFormer-S	EEG-FM	0/7	1.000	0/7	1.000
CBraMod (random init)	Random-init	0/7	1.000	–	–
Chronos-T5 (base)	TS-FMs	0/7	1.000	–	–
Chronos-T5 (large)	TS-FMs	0/7	1.000	–	–
Chronos-T5 (small)	TS-FMs	0/7	1.000	–	–
MOMENT (large)	TS-FMs	0/7	1.000	–	–
MOMENT (small)	TS-FMs	0/7	1.000	–	–
Moirai (base)	TS-FMs	0/7	1.000	–	–
Moirai (large)	TS-FMs	0/7	1.000	–	–
Moirai (small)	TS-FMs	0/7	1.000	–	–

Table 12: **Binomial top-3 win test across the 7 main-pool brain-age cohorts (Pearson r).** For each model, the exact one-sided binomial test asks whether the number of cohorts on which the model ranks in the top 3 (by cv-test Pearson r between predicted and chronological age, higher-is-better) exceeds the chance rate $p_0 = 3/k$. Two pools are tested independently: all $k = 26$ probed models, and the $k = 11$ EEG foundation models in isolation. Bold rows are significant at $\alpha = 0.05$.

		All models ($k = 26$)		EEG-FM only ($k = 11$)	
Model	Family	Wins	p	Wins	p
$N = 7$ cohorts; $p_0 = 3/26 = 0.115$ vs. $3/11 = 0.273$. Bold = significant at $\alpha = 0.05$.					
Sun et al. (2019)	**Supervised-Pre Age**	**7/7**	$< 10^{-3}$	–	–
EEGPT	EEG-FM	2/7	0.189	4/7	0.093
REVE	EEG-FM	2/7	0.189	2/7	0.610
CoreSleep	Supervised-Pre Sleep	2/7	0.189	–	–
SleepTransformer	Supervised-Pre Sleep	2/7	0.189	–	–
SleepyCo	Supervised-Pre Sleep	2/7	0.189	–	–
NeuroRVQ-EEG	EEG-FM	1/7	0.576	4/7	0.093
NeuroLM-VQ	EEG-FM	1/7	0.576	3/7	0.292
SleepFM	EEG-FM	1/7	0.576	3/7	0.292
MOMENT (base)	TS-FMs	1/7	0.576	–	–
BIOT	EEG-FM	0/7	1.000	2/7	0.610
STEEGFormer-S	EEG-FM	0/7	1.000	2/7	0.610
STEEGFormer-B	EEG-FM	0/7	1.000	1/7	0.892
CBraMod	EEG-FM	0/7	1.000	0/7	1.000
LaBraM	EEG-FM	0/7	1.000	0/7	1.000
Neuro-GPT	EEG-FM	0/7	1.000	0/7	1.000
CBraMod (random init)	Random-init	0/7	1.000	–	–
Chronos-T5 (base)	TS-FMs	0/7	1.000	–	–
Chronos-T5 (large)	TS-FMs	0/7	1.000	–	–
Chronos-T5 (small)	TS-FMs	0/7	1.000	–	–
Chronos-T5 (tiny)	TS-FMs	0/7	1.000	–	–
MOMENT (large)	TS-FMs	0/7	1.000	–	–
MOMENT (small)	TS-FMs	0/7	1.000	–	–
Moirai (base)	TS-FMs	0/7	1.000	–	–
Moirai (large)	TS-FMs	0/7	1.000	–	–
Moirai (small)	TS-FMs	0/7	1.000	–	–

Table 13: **Binomial top-3 win test across 17 BCI datasets (DREAMER counted twice for two distinct tasks).** For each model, the exact one-sided binomial test asks whether the number of datasets on which the model ranks in the top 3 (by normalized balanced accuracy) exceeds the chance rate $\bar{p}_0 = 3/k$. The pool contains $k = 7$ models (6 EEG-FMs + 1 TS-FM) evaluated under confound filtering with mean-pool aggregation. Bold rows are significant at $\alpha = 0.05$.

Model	Family	Wins	p
$N = 18$ datasets; $k = 7$ models; $\bar{p}_0 = 3/7 = 0.429$. Bold = significant at $\alpha = 0.05$.			
REVE	**EEG-FM**	**14/18**	**0.003**
NeuroLM	**EEG-FM**	**12/18**	**0.036**
EEGPT	EEG-FM	11/18	0.093
ST-EEG-S	EEG-FM	10/18	0.197
LaBraM	EEG-FM	4/18	0.981
CBraMod	EEG-FM	2/18	0.999
Moment-B	TS-FM	1/18	1.000

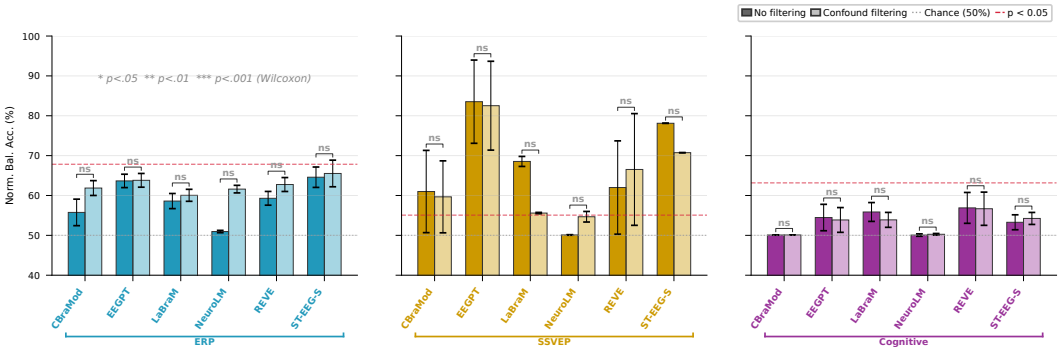

Figure 20: **Per-paradigm comparison of frozen EEG foundation models under No filtering vs. Confound filtering for ERP, SSVEP, and Cognitive BCI datasets.** Bars show the mean normalized balanced accuracy (%) across datasets within each paradigm; error bars denote the standard error of the mean. Asterisks indicate significant paired differences between tracks (Wilcoxon signed-rank test; $*p < .05$, $**p < .01$, $***p < .001$; "ns" = not significant). The dashed red line marks the aggregated significance threshold ($p < 0.05$, binomial test); the dotted gray line marks chance level (50%). All models use per-patch embeddings except NeuroLM (mean-pool) and REVE (mean-pool fallback where per-patch is unavailable). Overall, adopting a more task-specific confound filtering strategy does not lead to substantial performance changes compared to using the filtering configurations originally suggested by the models. Across paradigms, No filter and Confound filter results remain largely comparable, with most pairwise differences failing to reach statistical significance. **ERP:** Most models achieve above-chance performance, with EEGPT and ST-EEG-S reaching the highest accuracy ($\sim$65%). Confound filtering produces only marginal variations in performance, and none of the observed differences are statistically significant, suggesting that ERP decoding primarily relies on robust neural representations rather than recording-related confounds. **SSVEP:** EEGPT achieves the strongest performance ($\sim$84% in No filtering), followed by ST-EEG-S ($\sim$78%), whereas NeuroLM and REVE remain near chance level. Although filtering effects remain limited overall, SSVEP is the only paradigm showing performance consistently above the aggregated significance threshold, albeit with substantial inter-dataset variability reflected by the large error bars. **Cognitive:** All models remain near or below the aggregated significance threshold, with REVE reaching the highest accuracy ($\sim$59%). Differences between filtering strategies are again not statistically significant, indicating that current frozen EEG foundation models struggle to capture the subtler cognitive-state distinctions represented in these datasets.

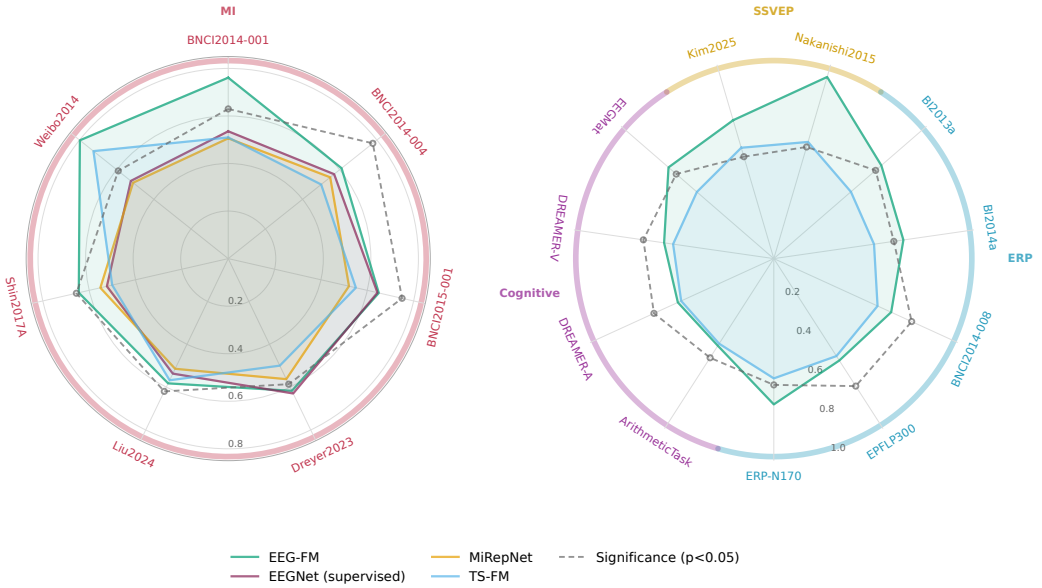

Figure 21: Normalized balanced accuracy of foundation model families under task-specific confound filtering across four BCI paradigms. Each spoke represents a dataset, and the radial axis maps chance-level performance to 0.5 and perfect classification to 1.0, enabling comparisons across paradigms with different class counts and difficulty levels. **Left:** Motor Imagery (MI) datasets comparing EEG foundation models (EEG-FM, per-patch pooling), time-series foundation models (TS-FM, mean pooling), pretrained supervised EEGNet models, and MIRepNet against the aggregated binomial significance threshold ($p < 0.05$). **Right:** SSVEP, ERP, and Cognitive datasets comparing EEG-FM and TS-FM families. Colored arcs indicate paradigm groupings. For each family, the best-performing model across all members is shown per dataset.

Table 14: LOSO EEGNet supervised classification results (balanced accuracy). No filter, uses the original filtering; the Confound Filter takes into account the appropriate per-paradigm filtering.

Dataset	**No filter (BAcc ± std)**	**Confound Filter (BAcc ± std)**
ArithmeticTask	0.341 ± 0.029	0.333 ± 0.000
BI2013a	0.770 ± 0.075	0.760 ± 0.073
BI2014a	0.614 ± 0.082	0.608 ± 0.080
BI2015a	0.780 ± 0.081	0.774 ± 0.084
BNCI2014_001	0.536 ± 0.124	0.324 ± 0.082
BNCI2014_004	0.732 ± 0.081	0.639 ± 0.093
BNCI2014_008	0.696 ± 0.043	0.699 ± 0.033
BNCI2014_009	0.778 ± 0.075	0.787 ± 0.071
BNCI2015_001	0.630 ± 0.104	0.559 ± 0.075
Cho2017	0.721 ± 0.124	0.680 ± 0.142
Dong2023	0.499 ± 0.263	0.128 ± 0.210
DREAMER_arousal	0.499 ± 0.028	0.501 ± 0.017
DREAMER_valence	0.511 ± 0.028	0.506 ± 0.029
Dreyer2023A	0.880 ± 0.096	0.656 ± 0.128
Dreyer2023	0.878 ± 0.098	0.673 ± 0.138
EEGMat	0.524 ± 0.074	0.499 ± 0.033
EPFLP300	0.638 ± 0.064	0.594 ± 0.071
ErpCore2021_ERN	0.544 ± 0.070	0.531 ± 0.053
ErpCore2021_LRP	0.553 ± 0.036	0.503 ± 0.019
ErpCore2021_MMN	0.585 ± 0.041	0.582 ± 0.038
ErpCore2021_N170	0.694 ± 0.110	0.650 ± 0.096
ErpCore2021_N2pc	0.643 ± 0.065	0.640 ± 0.064
ErpCore2021_N400	0.681 ± 0.079	0.686 ± 0.077
ErpCore2021_P3	0.701 ± 0.081	0.665 ± 0.077
GuttmannFlury2025_SSVEP	0.348 ± 0.105	0.505 ± 0.208
Han2024Fatigue	0.106 ± 0.038	0.053 ± 0.016
HefmiIch2025	0.494 ± 0.040	0.496 ± 0.016
Hinss2021	—	0.667 ± 0.131
Kim2025BetaRange	0.530 ± 0.279	0.063 ± 0.026
Lee2019_ERP	0.858 ± 0.086	0.850 ± 0.090
Lee2019_MI	0.654 ± 0.137	0.668 ± 0.133
Lee2019_SSVEP	0.932 ± 0.132	0.786 ± 0.260
Liu2020BETA	0.577 ± 0.286	0.545 ± 0.328
Liu2022EldBETA	0.681 ± 0.267	0.684 ± 0.256
Liu2024	0.508 ± 0.068	0.508 ± 0.058
Nakanishi2015	0.917 ± 0.108	0.088 ± 0.015
PhysionetMI	0.568 ± 0.135	0.298 ± 0.101
Schirrmeister2017	0.634 ± 0.119	0.424 ± 0.105
Shin2017A	0.636 ± 0.104	0.563 ± 0.080
Shin2017B	0.702 ± 0.108	0.570 ± 0.094
Wang2016	0.864 ± 0.149	0.583 ± 0.354
Weibo2014	0.198 ± 0.048	0.153 ± 0.035
Yang2025	0.611 ± 0.108	0.572 ± 0.094