

Prototype-based interpretable sleep staging with physiologically plausible sub-stage pattern discovery

Guido Gagliardi^{1*}, Javier García Ciudad², Letizia Micca³, Birgitte Rahbek Kornum²,
Moran Gilat³, Antonio Luca Alfeo^{4,7}, Mario G.C.A Cimino⁵, Maarten De Vos^{1,6*}

¹Department of Electrical Engineering, KU Leuven, Leuven, Belgium.

²Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark.

³Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium.

⁴Dept. of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy.

⁵Department of Information Engineering, University of Pisa, Pisa, Italy.

⁶Department of Development and Regeneration, KU Leuven, Leuven, Belgium.

⁷SMARTTEST Research Center, eCampus University, Novedrate, Italy.

*Corresponding author(s). E-mail(s): guido.gagliardi@kuleuven.be; maarten.devos@kuleuven.be;

Abstract

Sleep disturbances reduce quality of life and are recognized risk factors for neurodegenerative diseases such as Parkinson’s and Alzheimer’s. Automated sleep staging networks achieve competitive performance on multi-channel polysomnography, but their decisions are opaque. This limits clinical trust and leaves the sleep features they learn inaccessible.

We introduce the Proto-Sequence-to-Sequence (PSS) framework, a prototype-based wrapper that, after training, builds a codebook of discrete entries in the latent space of sequence-to-sequence sleep staging models. Staging quality approaches that of the continuous model as the codebook grows. The codebook entries, i.e. the prototypes, can be reconstructed in the original EEG, EOG and EMG input space, compared against an AASM-inspired checklist, and used to describe what the model encodes. Applied to two architectures, ProtoSleepNet (PSN, extending SeqSleepNet) and ProtoSleepTransformer (PST, extending SleepTransformer), the framework was evaluated across 11 independent datasets encompassing 12,317 subjects (over 120,000 recording hours). Without quantization, both models stage as well as their backbones. A new accuracy-weighted modality dropout makes them more robust to channel occlusion.

The learned prototypes show sub-stage patterns. Against an AASM-inspired checklist, 10 of 24 prototypes are coherent, 10 partially coherent and 4 non-canonical. Both backbones show delta-dominated deep sleep and an EOG-driven versus EEG-driven REM split. PSN also shows an N2 delta/sigma relevance gradient, which PST lacks. In an exploratory analysis of two clinical cohorts with nested leave-one-subject-out evaluation, twenty per-subject prototype statistics separated Alzheimer’s and Parkinson’s patients from controls at 86.8% and 80.5% area under the ROC curve. A comparator built on 34 descriptors of the expert hypnogram, on the same subjects and folds, reached 76.2% and 67.7%. The Alzheimer cohort analysis is not adjusted for age. Patients visit fewer distinct prototypes than controls, and the gap widens with the codebook size. An interactive online demo presents the prototypes.

Keywords: Prototype Learning, Explainable AI, Sleep Staging, Physiological Time Series.

Introduction

Sleep disorders affect 10–20% of the population [1], substantially reduce quality of life and are closely linked to overall health status. Changes in sleep architecture may reflect neurodegenerative processes. Growing evidence identifies sleep disturbances as markers of, and risk factors for, diseases such as Parkinson’s disease (PD) and Alzheimer’s dementia (AD) [2, 3]. Large-scale screening for sleep abnormalities could support earlier detection and intervention, but it requires accurate automated sleep staging and interpretable representations of what staging models learn about sleep physiology.

Sleep assessment records polysomnographic (PSG) data and scores each 30-second epoch into one of five stages (Wake, N1, N2, N3, and REM) according to established criteria [4]. Manual scoring is costly and labor-intensive. In-home PSG devices [5] and wearable sensors [2] make large-scale data collection feasible, and automatic sleep staging algorithms now achieve competitive accuracy [6]. The sequence-to-sequence deep learning framework [7] is particularly effective.

An epoch encoder embeds each 30-second window, and a sequence encoder processes L successive epochs jointly, which captures the dependence between sleep states within a sleep cycle [8–10]. The two main families of this framework are recurrent networks [8] and self-attention transformers [10].

These models share three limitations. First, their decisions are opaque: clinicians cannot verify whether a model relies on physiologically expected features (delta power for N3, sigma-band activity for N2, EOG dynamics for REM) or on spurious correlations. Second, models also learn sub-stage features during training, for example patterns that may reflect sleep-spindle dynamics, arousal structure or disruptions associated with neurodegenerative disease. These features stay in the latent space, where clinicians cannot inspect them and downstream analyses cannot use them. Third, naive multi-channel fusion may cause a multimodal competition problem [11]: EEG gives the strongest staging signal, so it tends to dominate training over EOG and EMG. The model is then weak when the EEG is noisy or absent. Such recordings are common in home-based [5] and wearable [2] settings, where sensor quality varies.

Explainable AI (XAI) methods address interpretability through post-hoc approaches (e.g., LIME [12], SHAP [13]) that probe input–output relationships and mechanistic approaches (e.g., LRP [14], CRP [15]) that trace contributions of specific model components. Both produce instance-level explanations, which lose precision when aggregated into global summaries [16]. Prototype-based models take a different route. They base their decisions on a discrete set of interpretable representations, the prototypes, which form a condensed vocabulary of the input space and give local and global explanations at the same time [17–20].

Explainable sleep-staging methods vary by how prototypes are obtained and presented. Rule-based approaches fix prototypes before training by aggregating embeddings of epochs that match expert rules, then classify epochs by similarity; explanations are limited to the chosen rules [21]. Prototype networks instead learn latent prototypes jointly with the classifier, with the number of prototypes set as a hyperparameter that requires retraining to change [22]. These networks display each prototype as its nearest training segment in latent space, and a loss term keeps prototypes close to real patches. Experts then label the prototypes by inspection. Other work explains standard classifiers with input relevance heatmaps that experts compare to scoring rules, yielding local explanations without a global vocabulary [23].

It remains unclear whether prototypes should be integrated into the model’s decision-making or constructed post-hoc from a pre-trained encoder [20]. Post-hoc approaches offer broader applicability but rely on the assumption that the decision function is locally continuous around prototype neighborhoods. The assumption usually holds at the final classification layer. In intermediate representations, though, it often fails, because small perturbations can change the output drastically [24, 25]. Models that pass every prediction through learned prototypes need end-to-end retraining, and quantizing the decision space can lower their performance.

Prototypes are latent vectors, so experts cannot read their physiological meaning directly. Reconstructing a prototype from its nearest real input has two limits: first, the reconstruction has low fidelity if the latent space is sparse around the prototype; second, a faithful reconstruction still has to be decomposed to show which features of that input place it close to the prototype in the space. Hence, describing a prototype in clinical terms requires three steps: reconstruct it in the input space, decompose it into interpretable features, and compare these features with established scoring criteria.

We propose the Proto-Sequence-to-Sequence (PSS) framework¹ (Figure 1), a prototype-based wrapper for sequence-to-sequence sleep staging. Dual residual connections prepare the embedding space for post-hoc quantization. After training, a codebook of discrete prototypes is built from the continuous encoder, a post-hoc prototype approximation of the backbone, with a measured loss of staging quality. Users choose the codebook size M without retraining, according to how much staging performance they want to retain and how many prototypes they want to inspect. At inference, predictions are routed through the prototypes, or through the continuous model when no quantization is needed. Each prototype is then reconstructed in the input space and described in a short summary.

We demonstrate the framework on SeqSleepNet [8] (ProtoSleepNet, PSN) and SleepTransformer [10] (ProtoSleepTransformer, PST), and evaluate it on 11 independent PSG datasets, which yield 20 dataset-sources and over 12,000 subjects. To make the prototypes readable, we develop a reconstruction and characterization pipeline. It maps each codebook entry back to the input space, decomposes it into physiologically defined spectral and temporal features, and summarizes the result in a short text that sleep clinicians can read. The explanations work at two levels: locally, why a specific epoch is assigned to a given prototype, and globally, what each prototype encodes across the full population of assigned epochs. The global characterizations are compared against an AASM-inspired checklist independently on both backbones. Patterns found in both architectures are not specific to one architecture or training corpus. Finally, we test the prototype-gram, the sequence of prototype assignments across a subject’s night in a neurodegenerative disease detection pipeline. The models that produce the prototype-gram are frozen and trained for staging only, with no disease labels. We ask whether its statistics separate clinical groups known to differ in sleep architecture better than those of the expert’s hypnogram of the same recordings.

We show that both PSS variants preserve competitive staging performance while conferring robustness to missing channels. Against an AASM-inspired checklist, 10 of the 24 prototypes are Coherent, 10 Partially coherent and 4 Non-canonical. Delta-dominated N3 and an EOG-driven versus EEG-driven REM split appear in both backbones. PSN shows an N2 delta/sigma relevance gradient, which PST lacks. In exploratory clinical probing, twenty per-subject prototype-gram statistics reach AUCs of 86.8% (Alzheimer’s) and 80.5% (Parkinson’s), against 76.2% and 67.7% for a comparator

¹An interactive explainability demo is available at <https://protosleepnet-demo.pages.dev>, and the source code at <https://github.com/guidogagl/protosleepnet>.

built on 34 hypnogram descriptors of the same recordings and folds. The Alzheimer analysis is not adjusted for age. In both cohorts, patients visit fewer distinct prototypes than controls, and the gap widens with the codebook size, reaching 28% (Alzheimer’s) and 21% (Parkinson’s) at $M = 100$. The clinical findings are hypothesis-generating and require replication on independent cohorts.

Results

We evaluate the Proto-Sequence-to-Sequence (PSS) framework along four axes: sleep staging performance, robustness under channel occlusion, prototype-based explanations, and exploratory detection of neurodegenerative conditions such as Alzheimer’s and Parkinson’s disease. Four model variants are compared throughout: SeqSleepNet (SSN) and its PSS extension ProtoSleepNet (PSN), and SleepTransformer (ST) and its PSS extension ProtoSleepTransformer (PST). Both PSS variants were trained with the same settings as the original backbones: PSN on MASS (200 subjects, 5 cohorts) and PST on SHHS visit-1 (5,793 subjects).

PSN and PST match their backbone models in staging performance

Any prototype extension must first match the backbone’s staging accuracy. We evaluated PSN against SSN and PST against ST on 20 dataset-sources, differentiating between in-domain and out-of-domain settings. Figure 2a shows 13 markers for the 20 sources: the five MASS cohorts are pooled into one point and the four WSC visits into another. In-domain, we tested the full end-to-end models on held-out test-split recordings (5 MASS cohorts for PSN/SSN; SHHS visit-1 for PST/ST) using sliding-window inference with stride 1 [8]. Out-of-domain evaluation used the complete held-out dataset-sources, 15 for PSN and 19 for PST. On each, we trained a linear probe on frozen sequence-encoder outputs with 5-fold subject-stratified cross-validation [26]. Per-subject differences were assessed using paired Wilcoxon signed-rank tests with Bonferroni correction.

Figure 2a shows per-dataset κ (filled markers: in-domain, end-to-end models; open markers: out-of-domain, linear probes on frozen encoders). The baselines are the official single-channel reference implementations, SeqSleepNet-1ch and SleepTransformer-1ch. PSN exceeds SeqSleepNet-1ch on 13 of 15 dataset-sources (Wilcoxon $p < 0.05$, Bonferroni-corrected), with the largest gain on MESA. PST exceeds SleepTransformer-1ch on 5 of 19, with near-parity on the remaining 14 (largest gain: MESA, +0.051). The corresponding three-channel variants (SeqSleepNet-3ch, SleepTransformer-3ch) were also evaluated in-domain: PSN $\kappa = 0.801 \pm 0.070$ vs. 0.788 ± 0.068 on MASS, and PST $\kappa = 0.814 \pm 0.100$ vs. 0.812 ± 0.103 on SHHS (Supplementary Note 2).

The out-of-domain protocol therefore measures how linearly separable the stages are in the frozen representation of each cohort. Modality dropout may produce broader representations that transfer better out of domain (Supplementary Note 1). Staging performance is stable across six independent training seeds (coefficient of variation of about 1% for κ ; Supplementary Note 9). For reference, published inter-rater agreement among human scorers on the same data ranges from $\kappa = 0.68$ to 0.82 [6].

We also evaluated robustness under channel occlusion conditions on the in-domain datasets, testing four scenarios plus the clean condition: 25% and 50% random channel replacement, complete EEG removal, and complete EOG+EMG removal. Progressive ablations show that accuracy-weighted modality dropout is the main source of robustness. Without modality dropout, both baselines degrade (SSN: $\kappa = 0.788 \rightarrow 0.076$ at 50% random occlusion; ST: $0.812 \rightarrow 0.556$ without EEG). Modality dropout restores near-clean performance without additional parameters. The channel mixer then recovers the clean-condition accuracy that modality dropout reduces. Modality importance depends on the backbone: PSN relies strongly on EEG ($\kappa = 0.357$ without EEG), while PST relies less on it ($\kappa = 0.756$ in the same condition). Full ablation results are reported in Supplementary Note 2.

PSN and PST allow post-hoc quantization

All downstream prototype analyses depend on discretizing the embedding space into M codebook entries. Because the codebook is built post-hoc via K-Means on frozen embeddings, M is decoupled from training and can be chosen afterward as an explanation granularity parameter. The prototypes represent the model faithfully only if quantization preserves staging quality, so we measure that quality first.

Figure 2b shows staging accuracy (solid) and κ (dashed) as a function of codebook size M . For both PSS variants, performance rises steeply from $M = 5$ and saturates at $M \approx 48$ –65, approaching the non-quantized baselines (dotted lines). The unmodified backbones lose more under quantization. At $M = 100$, PSN keeps $\kappa = 0.788$ (non-quantized: 0.801), while the single-channel SSN drops to 0.675 (from 0.769). PST keeps 0.794 (from 0.814), while the single-channel ST drops to 0.736 (from 0.799).

Ablation shows that the dual residual connection improves quantization quality. Without residuals, K-Means quantization gives $\kappa = 0.568$ at $M = 5$. With residuals it gives $\kappa = 0.715$ (+0.147). The residuals may make the embedding space smoother and better suited to centroid-based clustering (Supplementary Note 3).

An alternative to post-hoc K-Means is supervised Vector Quantization (VQ), which refines the codebook via gradient descent on the staging loss using straight-through estimation. We compared K-Means against VQ with random initialization and VQ with K-Means initialization: K-Means matches supervised VQ in staging fidelity and is stable, whereas

randomly initialized VQ ranges from $\kappa = 0.534$ to 0.709 (Supplementary Note 3). A randomization test shows that K-Means codebooks capture embedding structure: K-Means outperforms random exemplar codebooks for PSN at all tested sizes ($p \leq 0.04$) and for PST for $M \geq 8$ ($p < 0.05$; Supplementary Note 8). A seed stability analysis across six independent training runs per backbone shows that the prototypes are reproducible. Staging κ varies by about 1% (CV). At all but one codebook size, the input-space shift between matched prototype reconstructions is at most $2\times$ the intra-seed sampling baseline for both backbones (Supplementary Note 9).

Reconstructing quantized entries in the PSG input space

The codebook size M sets the inspection level. We use $M = 12$ for all the interpretability analyses that follow. Larger codebooks follow the network more faithfully and take longer to analyze, and smaller codebooks give a coarser vocabulary. Twelve entries can be inspected one by one, at a measured staging cost (previous subsection). The characterization, reconstruction and local-explanation results therefore refer to the twelve-entry codebook. Supplementary Note 9.4 reports which of them also hold at neighbouring codebook sizes.

The $M = 12$ codebook entries are latent vectors with no direct physiological meaning. To understand what each prototype encodes, we map it back to the PSG input space (EEG, EOG, EMG time-frequency representation). The mapping is under-determined: a prototype constrains only the dimensions that drive assignment, so the remaining dimensions can take arbitrary values during optimization.

We propose three reconstruction strategies: *data-driven* (nearest real epochs to the prototype), *model-driven* (gradient optimization from random noise), and *hybrid* (gradient optimization seeded with nearby real epochs). Quality is assessed by three complementary metrics: fidelity (L2 distance to the prototype embedding), plausibility (MMD against the empirical data distribution), and stability (pairwise MMD between reconstructions from independent datasets). For each of the 12 prototypes per backbone, 256 reconstructions are generated and evaluated across 16 independent datasets for PSN and 20 for PST.

The three methods trade off differently across metrics (Figure 3). Data-driven reconstruction guarantees plausibility by construction (panel b) but exhibits the worst fidelity (panel a: PSN 3.61 ± 0.93 , PST 2.95 ± 0.80) and highest dataset dependence (panel c). Fully model-driven reconstructions are the most stable across datasets. Their spectra, however, depart from the empirical ones: the sigma-to-delta power ratio is 17 times the empirical value for PSN and 9 times for PST. Hybrid reconstruction gives the best balance of fidelity and plausibility. Its plausibility MMD across external cohorts is near zero (PSN: 0.004 ± 0.012 ; PST: 0.017 ± 0.031), and its fidelity is much better than that of data-driven reconstruction (PSN: 1.24 ± 0.35 ; PST: 0.55 ± 0.12).

We therefore use hybrid reconstruction in all subsequent prototype characterization analyses. Its reconstructions are faithful to the model’s internal representation and close in distribution (MMD) to real PSG signals (Supplementary Note 5).

Prototype characterization and coherence with an AASM-inspired checklist

We characterize each of the $M = 12$ prototypes per backbone with two ablations of its hybrid reconstruction. The first is a combinatorial sweep over an 8-band EEG decomposition, with separate σ_l [11–13 Hz] and σ_h [13–16 Hz] sub-bands, holding EOG and EMG at their reconstructed values. The second is three whole-channel ablations that partition importance across EEG, EOG and EMG. Together they give each prototype a band sensitivity profile and a channel-importance partition, reported below. Each profile is then written up as a compact prototype summary of four sentences: stage purity, discriminative spectral bands, temporal structure, and a coherence verdict against the AASM-inspired checklist (Supplementary Note 6).

Band relevance is the marginal contribution of a band to how closely the reconstruction matches the prototype. It is expressed as a percentage of the total EEG contribution, not of spectral power. A band has negative relevance when restoring it moves the reconstruction away from the prototype, and raw values range from -3.4% to 155.2% . The tables quote raw values. Figure 4 rescales the absolute relevances of each prototype to sum to 100, so a raw δ relevance of 88.5% appears as 68%. Values quoted from Figure 4 in the N3, REM, N2 and Wake paragraphs below use this rescaled scale.

Figure 4 shows the codebook summaries for both backbones, organized by dominant sleep stage. Table 1 presents the corresponding prototype summaries.

In the N3 rows of Figure 4, PSN P10 ($\delta = 68\%$) and PST P4/P8 ($\delta = 88\%/82\%$) are dominated by δ . Every other band has relevance of at most $|10\%|$. Both backbones also show a mild θ suppression (-10% to -11% in PSN). Restoring θ , a band shared with N1 and drowsy wakefulness, moves the reconstruction away from the N3 prototype. EEG carries 94–95% of the channel importance. This is the clearest cross-backbone convergence in the figure: PSN and PST have similar N3 signatures, dominated by slow-wave (δ) activity.

In the REM rows of Figure 4, PSN P4 relies mainly on EOG (EOG = 78%), while PSN P1 relies mainly on EEG (EEG = 68%). The same inversion appears in PST: P7 (EOG = 62%) versus P1 (EEG = 62%). This channel-importance contrast separates an EOG-driven from an EEG-driven REM prototype. Both backbones show it independently, without phasic/tonic labels in the training data. This split is consistent with the phasic and tonic REM substates of Simor et al. [27]. Testing the correspondence would require phasic/tonic annotations. In the heatmap, both REM variants show moderate δ relevance (33–44% in PSN, 36–41% in PST) and σ_h relevance (19–20%). In both backbones the ablation

Table 1 Representative prototype summaries for six sub-stage patterns across both backbones (12 prototypes per backbone, hybrid reconstruction). For each pattern, one prototype per backbone illustrates convergent and divergent classification strategies. Each summary has four sentences: (S1) stage identity and channel importance, (S2) primary and secondary spectral features by sensitivity, (S3) temporal structure and EMG/EOG dynamics, (S4) checklist coherence verdict. The muscle-tone and eye-movement words in S3 use the 25th and 75th percentile of the codebook, whereas the checklist uses the 40th and 60th (Supplementary Note 6). The letter after the prototype index gives the verdict: C is coherent, P partially coherent and X non-canonical. Spectral bands appear in the table by their usual symbols (δ , θ , α , σ with low and high sub-bands, β with low and high sub-bands, γ). Summaries for all 24 prototypes are in Supplementary Note 6.

Pattern	BB	P	Prototype Summary
N3	PSN	P10 (C)	Encodes N3 (100% pure) primarily through EEG (94%), with EOG (5%) and EMG (1%). EEG signature shows elevated δ (+14.7 dB, sensitivity=21.6) and suppressed β_h (-0.1 dB, sensitivity=13.0). δ activity is present in 100% of each epoch, with moderate eye movements (SEM 90% active, likely EEG crosstalk) and intermediate muscle tone. Coherent with AASM N3 criteria: δ elevated, δ dominant, EMG atonic — all confirmed.
	PST	P4 (C)	Encodes N3 (100% pure) primarily through EEG (94%), with EOG (5%) and EMG (2%). EEG signature shows elevated δ (+10.3 dB, sensitivity=16.9) and absence of β_h (-0.8 dB, sensitivity=1.8). δ activity is present in 97% of each epoch, with moderate eye movements and atonic muscle tone. Coherent with AASM N3 criteria: δ elevated, δ dominant, EMG atonic — all confirmed.
REM (phasic)	PSN	P4 (C)	Encodes REM (100% pure) primarily through EOG (78%), with EEG (14%) and EMG (9%). EEG signature shows elevated δ (+1.0 dB, sensitivity=11.5) and elevated σ_h (+1.2 dB, sensitivity=6.8). δ activity is present in 24% of each epoch, with active eye movements (REM 41% active) and atonic muscle tone. Coherent with REM physiology: EOG dominant, EMG atonic, θ elevated — all confirmed.
	PST	P7 (C)	Encodes REM (100% pure) primarily through EOG (62%), with EEG (30%) and EMG (7%). EEG signature shows suppressed δ (-0.3 dB, sensitivity=7.7) and suppressed γ (-0.1 dB, sensitivity=4.2). δ activity is present in 14% of each epoch, with moderate eye movements (REM 31% active) and atonic muscle tone. Coherent with REM physiology: EOG dominant, EMG atonic, θ elevated — all confirmed.
REM (tonic)	PSN	P1 (P)	Encodes REM (100% pure) primarily through EEG (68%), with EOG (21%) and EMG (11%). EEG signature shows elevated δ (+0.6 dB, sensitivity=13.8) and elevated σ_h (+1.2 dB, sensitivity=6.3). δ activity is present in 21% of each epoch, with quiescent eye movements (REM 10% active) and atonic muscle tone. Partially coherent: confirmed EMG atonic, θ elevated; missing EOG dominant. Model uses δ rather than θ .
	PST	P1 (P)	Encodes REM (100% pure) primarily through EEG (62%), with EOG (37%) and EMG (1%). EEG signature shows suppressed δ (-1.9 dB, sensitivity=18.2) and elevated γ (+1.2 dB, sensitivity=8.5). δ activity is present in 7% of each epoch, with quiescent eye movements (REM 3% active) and intermediate muscle tone. Partially coherent: confirmed EOG dominant, θ elevated; missing EMG atonic.
N2 (σ -rich)	PSN	P0 (C)	Encodes N2 (100% pure) primarily through EEG (92%), with EOG (4%) and EMG (3%). EEG signature shows elevated β_h (+1.0 dB, sensitivity=11.7) and elevated σ_h (+7.6 dB, sensitivity=5.4). β_h activity is present in 10% of each epoch, with quiescent eye movements (SEM 24% active) and intermediate muscle tone. Coherent with N2: σ elevated, σ relevant, EOG quiescent — all confirmed. Model primarily uses β_h , atypical for N2.
	PST	P2 (C)	Encodes N2 (100% pure) primarily through EEG (81%), with EOG (14%) and EMG (5%). EEG signature shows elevated δ (+5.7 dB, sensitivity=9.1) and elevated σ_h (+2.0 dB, sensitivity=2.9). δ activity is present in 69% of each epoch, with moderate eye movements (SEM 38% active) and intermediate muscle tone. Coherent with N2: σ elevated, σ relevant, EOG quiescent — all confirmed.
N2 (δ -rich)	PSN	P5 (X)	Encodes N2 (100% pure) primarily through EEG (96%), with EOG (2%) and EMG (2%). EEG signature shows elevated δ (+9.8 dB, sensitivity=18.2) and elevated β_h (+0.3 dB, sensitivity=17.5). δ activity is present in 86% of each epoch, with moderate eye movements (SEM 66% active, likely EEG crosstalk) and intermediate muscle tone. Non-canonical for N2: missing σ relevant, EOG quiescent. Deep-N2 classified through δ rather than σ -band relevance.
	PST	P0 (X)	Encodes N2 (100% pure) primarily through EEG (73%), with EOG (23%) and EMG (4%). EEG signature shows elevated γ (+3.3 dB, sensitivity=4.3) and elevated β_h (+3.2 dB, sensitivity=4.3). γ activity is present in 17% of each epoch, with moderate eye movements (SEM 45% active) and intermediate muscle tone. Non-canonical for N2: missing σ relevant, EOG quiescent. Model uses broadband high-frequency power, atypical for N2.
Wake	PSN	P7 (C)	Encodes W (100% pure) primarily through EEG (92%), with EOG (6%) and EMG (2%). EEG signature shows elevated θ (+4.0 dB, sensitivity=13.9) and elevated δ (+7.8 dB, sensitivity=11.2). θ activity is present in 21% of each epoch, with moderate eye movements (Blink 79% active) and tonic muscle tone. Coherent with W: α elevated+relevant, EMG tonic, EOG active — all confirmed. Model primarily uses θ , atypical for W.
	PST	P3 (P)	Encodes W (100% pure) primarily through EEG (50%), with EOG (46%) and EMG (4%). EEG signature shows elevated γ (+21.8 dB, sensitivity=2.7) and elevated δ (+21.6 dB, sensitivity=1.9). γ activity is present in 100% of each epoch, with active eye movements (Blink 100% active) and tonic muscle tone. Partially coherent: confirmed EMG tonic, EOG active; missing α elevated+relevant.

relevance of θ is zero or negative (-8% to -16% in PSN, below threshold in PST). The checklist criterion “ θ elevated” refers to spectral power relative to the training mean. θ also marks N1 and drowsy wakefulness, which may explain why the models separate REM prototypes by other features. The EOG-driven prototype is characterized by eye-movement (EOG) dynamics, and the EEG-driven one by δ level combined with high-frequency activation.

N2 is the most heterogeneous group in both panels of Figure 4. Six PSN prototypes and four PST prototypes lie along a continuous spectral gradient. In the PSN panel, the gradient is visible along the σ_h column. P0 ($\delta = 6\%$, $\sigma_h = 32\%$) is mainly sigma-driven. P5 ($\delta = 48\%$, $\sigma_h = 8\%$) and P11 ($\delta = 43\%$, σ_h absent) are delta-dominated. Intermediate prototypes (P8: $\sigma_h = 25\%$; P9: $\sigma_h = 23\%$; P6: $\sigma_h = 22\%$) have both σ and δ relevance. The gradient reflects the model’s ablation relevance, that is, how much each band contributes to matching an epoch to a prototype.

In the 19 MASS recordings, spindle density detected by YASA rises with σ relevance across nine PSN prototypes (Spearman $\rho = 0.75$, $p = 0.043$). σ power predicts spindle density equally well ($\rho = 0.77$), so this check links relevance to spindles at the level of band power (Supplementary Note 6). De Gennaro & Ferrara [28] report an inverse sigma-delta relationship in *power*, which is a different quantity.

The PST panel differs. Its σ_h column is mostly white ($< 11\%$ in all PST N2 prototypes). PST classifies N2 primarily through broadband power levels and high-frequency suppression (P5: $\beta_h = 28\%$; P0: $\beta_h = 18\%$, $\gamma = 19\%$). PSN weights the sigma sub-bands more heavily than PST. This is the largest systematic difference between the two codebooks. The two systems differ in architecture, training corpus, channel derivations and scoring conventions, so the difference reflects the system as a whole. The finding applies to these two codebooks. Where a sigma-resolved N2 vocabulary is needed, the PSN codebook provides one.

In the Wake rows of Figure 4, the α column is nearly empty. The highest alpha relevance across all six Wake prototypes is 18% (PSN P7), far below what AASM’s α -centric definition would predict. The β_l , β_h , and γ columns show warm colours throughout the Wake group. PSN P3 ($\beta_l = 18\%$, $\beta_h = 17\%$, $\gamma = 11\%$) and PST P3 ($\gamma = 34\%$) show elevated beta and gamma relevance. The channel bars also differ between the backbones. PST Wake prototypes allocate 24–46% of channel importance to EOG (teal), as a secondary feature. PSN allocates 69–92% to EEG. The α criterion fails in five of the six Wake prototypes: all four PST prototypes and PSN P3. Two explanations are consistent with this: diversity in the pooled training populations (eyes-open and eyes-closed recordings, individual α variability, drowsy states), or artefact contamination of the EEG. Additional data are needed to distinguish them.

N1 makes up 1–5% of training sleep, and neither backbone has a dedicated N1 prototype. PSN P11 (N2 with negative σ relevance and δ -dominant spectral profile) covers the N1/N2 transition. PSN P3 (Wake with 80% purity, 20% N1 content) covers the W/N1 transition. Three properties of N1 may explain why it has no dedicated prototype, and this design does not rank them. N1 is defined by the *absence* of graphoelements (no spindles, no K-complexes, no α). It has the lowest inter-rater reliability among stages ($\kappa \approx 0.25$ [6]). It also provides a weak training signal due to low prevalence. The number of prototypes per stage follows both spectral distinctiveness and stage prevalence, and the two effects are

confounded here. N3 receives 2 prototypes in both codebooks. N2 (heterogeneous) receives 4–6, and Wake receives 2–4. N1 is spectrally ambiguous and transient, and its epochs are assigned to other prototypes.

To compare these findings with the AASM scoring criteria, we assessed each prototype against a stage-specific checklist inspired by them. Of the 24 prototypes of the two twelve-entry codebooks, 10 are rated Coherent, 10 Partially coherent, and 4 Non-canonical. A Coherent verdict means that all checklist conditions hold for the model’s attributions (Methods). Blinded expert scoring is a next step. The assessment covers the two twelve-entry codebooks and was not run at the other codebook sizes used in the clinical analysis.

As a negative control, we assigned each prototype a stage other than its own, keeping the number of prototypes per stage in each codebook, and applied the criteria of that stage (10,000 random assignments). Against their own stage, 10 of the 24 prototypes are coherent. Against another stage, 2.0 are coherent on average and at most 4 ($p < 0.001$; PSN 6 against 1.3, PST 4 against 0.6). The checklist therefore separates the stages (Supplementary Note 6).

The thresholds of the coherence checklist are chosen by the investigators, so the verdict counts depend on them. To measure this dependence, we shift together the six thresholds that determine the verdicts: the α , σ , δ and EOG relevance criteria and the two percentile bounds. Each threshold moves up or down by the same fraction of its own value. The N1 bound plays no role, because no prototype is assigned N1. We call this fraction the tolerance. For the two percentile thresholds, it applies to their 10-point distance from the median. For example, a tolerance of 25% lets a relevance threshold of 30% take any value from 22.5% to 37.5%.

The margin of a prototype is the smallest tolerance at which its verdict changes (Table 2, column Threshold margin). It can exceed 100%, because a threshold can rise without limit. The margin is partly arithmetic: a criterion “ x exceeds T ” changes answer as soon as T crosses x , so the margin equals the distance of a prototype from its nearest threshold. All verdicts are stable below a tolerance of 10%. At 25%, 21 of the 24 verdicts are stable, including all ten Coherent verdicts. The three verdicts that change at 25% are all non-Coherent: PST P6 in Wake (margin 10%), PSN P5 in N2 (13%) and PST P1 in REM (23%).

Table 2 Summary of the coherence assessment against the AASM-inspired checklist across both backbones. C denotes Coherent (all stage-specific criteria met) and X Non-canonical (fewer than 50% of criteria met). The threshold margin is the range, across the prototypes of a stage, of the tolerance at which a verdict changes (all six thresholds shifted together by the same fraction of their own values).

Stage	PSN	PST	Combined	Threshold margin	Key observation
N3	1C, 1P	2C	3C, 1P	150–243%	Delta dominance meets the checklist
REM	1C, 1P	1C, 1P	2C, 2P	23–160%	Phasic coherent; tonic partial (phasic/tonic labels are an interpretation; PSN misses 1)
N2	3C, 2P, 1X	1C, 2P, 1X	4C, 4P, 2X	13–145%	PSN uses sigma more than PST
W	1C, 1X	3P, 1X	1C, 3P, 2X	10–104%	Alpha criterion failed by 5 of 6
N1	—	—	—	—	No dedicated prototype; N1 epochs are assigned to boundary prototypes
Total	6C, 4P, 2X	4C, 6P, 2X	10C, 10P, 4X	10–243%	

Non-canonical verdicts (PSN P3, P5; PST P0, P6) come from two patterns: Wake prototypes whose α relevance is below the criterion (PSN P3, PST P6), and N2 prototypes that fail the σ relevance and the quiescent-EOG criteria (PSN P5, PST P0). PSN has 6 Coherent verdicts and PST has 4, out of 12 prototypes each, and the difference was not tested statistically. PSN uses σ more than PST in N2. The two backbones differ in architecture and in training corpus together, so the difference reflects the system as a whole. The complete set of 24 summaries and the full coherence table are reported in Supplementary Note 6. Local instance-to-prototype explanations are detailed in Supplementary Note 4.

Exploratory clinical probing with prototype-grams

We asked whether subject-level summaries of prototype-based sleep representations encode disease-related structure beyond the conventional hypnogram. We extracted summary statistics of the prototype-gram to classify clinical groups known to differ in sleep structure, and compared them with features from the expert hypnogram [2]. This study is an interpretability probe of what the prototypes encode. No disease information entered the staging model during training.

We used two KU Leuven cohorts: ASD (32 controls, 37 Alzheimer’s disease, $N = 69$) and KPD (40 controls, 46 Parkinson’s disease, $N = 86$). The KPD groups are age- and sex-matched (Mann–Whitney $p = 0.883$ for age, Fisher $p = 1.000$ for sex). The frozen backbones PSN (MASS [29]) and PST (SHHS [5]) assign each subject’s epochs to the nearest codebook entries, which gives a per-subject prototype-gram. From it we compute 20 quantities at each of 10 codebook sizes ($M = 5$ –100): 19 coverage quantities (set C) and the untraversed mass of the training transition graph (T). C-T is their union. The prior and transitions come only from the backbone’s training population (Supplementary Note 7). The comparator is the 34 hypnogram descriptors of Heremans et al. [2], computed from the expert scoring and passed through the same classifier and folds.

A logistic regression under nested leave-one-subject-out cross-validation selects among 96 configurations for each held-out subject; AUCs are compared by a one-sided DeLong test (Methods, Supplementary Note 7).

C-T reaches 86.8% AUC on AD ($p = 0.038$) and 80.5% on PD ($p = 0.034$), against 76.2% and 67.7% for the comparator (Figures 5a and 5b). These two comparisons are the confirmatory tests; the one-sided DeLong probabilities are 0.038 and 0.034 before correction, Benjamini–Hochberg at $m = 2$ gives 0.038 for both, and a Bonferroni correction

would give 0.076 and 0.068. The results for T and C alone are descriptive (Table 3). For the coverage-based sets, the inner loop selected PST in 83–100% of Parkinson’s folds and PSN in 81–86% of Alzheimer’s folds.

Patients visit fewer distinct prototypes within a fixed epoch budget from $M = 12$ onward (Figures 5e and 5h); the groups do not differ at $M = 5$ and 8. At $M = 100$ the deficit reaches 28% below controls on Alzheimer’s and 21% on Parkinson’s. At $M = 12$, controls traverse 83% of the training-edge mass against 63% in Alzheimer’s patients, and 85% against 73% in Parkinson’s (Figures 5c, 5d, 5f and 5g; in-sample AUC 81.8% and 74.8%, Hedges’ g +1.15 and +0.95).

Split by the dominant stage of each entry, REM entries account for 38.0% of the unreached prior mass in Alzheimer’s patients (bootstrap interval 26.3–52.3) against 20.4% in controls; the patients’ interval lies above the 18.1% that those entries hold in the prior. The inferential comparison is the test on absolute unreached REM mass, and the share difference is descriptive. On Parkinson the REM share difference is larger, and the test on absolute REM mass is not significant. The transitions most often left untraversed link two N2 entries. This type also makes up the largest share of the transition prior, so it is enriched by only about a quarter. The Parkinson REM-share difference holds at every codebook size, and the Alzheimer one at $M = 12$ (Supplementary Note 7). Self-supervised models of unlabelled recordings have independently reported reproducible substructure in conventional N2 [30]. The codebook identifies the entries where the deficit falls. The margin over the comparator reflects two differences, the scorer (expert or model) and the level of description (five stages or individual prototypes).

Table 3 Nested leave-one-subject-out performance against the hypnogram comparator. Areas under the ROC curve (AUC) are those of the seed-averaged predictor, with DeLong 95% confidence intervals; accuracies carry Wilson intervals. Probabilities are one-sided DeLong tests against the comparator. The combined feature set (C-T) is the confirmatory analysis: its two comparisons, one per cohort, are adjusted by the Benjamini–Hochberg procedure at $m = 2$ (0.038 for both). The coverage-only (C) and transition-only (T) feature sets show the discriminative value of each family on its own. Their probabilities are descriptive and carry no correction for the feature-set search.

Cohort	Feature set	AUC (%)	Accuracy (%)	Macro-F1 (%)
Alzheimer ($N = 69$)	Expert hypnogram (34)	76.2 (64.3–88.1)	76.8 (65.6–85.2)	76.8
	T (1)	82.8 (72.7–92.9), $p = 0.188$	72.5 (61.0–81.6)	72.4
	C (19)	87.2 (79.2–95.3), $p = 0.031$	76.8 (65.6–85.2)	76.7
	C-T (20)	86.8 (78.0–95.7), $p = 0.038$	82.6 (72.0–89.8)	82.6
Parkinson ($N = 86$)	Expert hypnogram (34)	67.7 (56.3–79.1)	61.6 (51.1–71.2)	61.6
	T (1)	76.2 (66.3–86.2), $p = 0.139$	70.9 (60.6–79.5)	70.9
	C (19)	75.4 (65.0–85.7), $p = 0.151$	67.4 (57.0–76.4)	67.4
	C-T (20)	80.5 (71.1–90.0), $p = 0.034$	73.3 (63.1–81.5)	73.3

Discussion

The PSS framework adds interpretability to sequence-to-sequence sleep staging by building, after training, a discrete codebook of sub-stage patterns that can be reconstructed, characterized and used to describe a subject’s night. The central question is what the prototype analysis shows about what a staging model encodes. PSN and PST stage competitively across 20 dataset-sources, and quantizing them after training costs staging performance in proportion to the quantization granularity.

Prototype-based models built for intrinsic explainability, such as ProtoPNet [31] and PIP-Net [32], learn their prototypes jointly with the classifier. In PSS instead, the codebook is fitted after training in the encoder’s latent space, and the encoder is never constrained by it. Post-hoc attribution methods such as integrated gradients and GradCAM give instance-level saliency maps without a global vocabulary, and attention weights in Transformer models such as Sleep-Transformer do not faithfully explain the decision [33]. PSS offers both levels: a global codebook of what the model distinguishes, and a local assignment of each epoch to one entry.

Both PSN and PST converge to delta-dominated N3 and an EOG-driven versus EEG-driven REM split, and PSN alone resolves the N2 delta/sigma relevance gradient. The two models differ in architecture and in training corpus at the same time, so their agreement and disagreement may reflect either factor.

Assessment of all 24 prototype codebook entries against an AASM-inspired checklist classifies 10 as coherent, 10 as partially coherent, and 4 as non-canonical. The checklist separates the stages: prototypes meet the criteria of their own stage far more often than those of another (10 against 2.0 coherent verdicts, $p < 0.001$). Both backbones share features that dominate a stage (δ in N3) or follow eye movements (EOG in REM). Features that appear in only one backbone rest on bands linked to brief events: σ in N2 and α in Wake. The four non-canonical prototypes (PSN P3, P5; PST P0, P6) may reflect alternative classification strategies, artefacts or errors, and expert review can tell these apart.

In two clinical cohorts, we asked whether the prototype-gram adds to conventional staging, using frozen PSS models. The twenty quantities of the C-T feature set reach an area under the ROC curve of 86.8% on Alzheimer’s disease and 80.5% on Parkinson’s, against 76.2% and 67.7% for 34 hypnogram descriptors computed from the expert scoring of the same recordings. Two properties of the prototype-gram are associated with the separation. Patients reach fewer distinct

prototypes than controls, and the deficit grows with the codebook size from $M = 12$. Patients also traverse less of the training transition mass. The unreached prototypes and transitions lie within stages, mainly in N2 entries and, on Alzheimer’s at $M = 12$, in REM entries.

The Parkinson AUC is lower than the Alzheimer AUC (80.5 against 86.8%), and its margin over the comparator is 12.8 points against 10.6, with a weaker comparator (67.7 against 76.2%). The KPD cohort has mild-to-moderate disease (Hoehn and Yahr II–III) [34], a stage at which slow-wave differences linked to later motor progression [35] may be small. The cohort has no reported cognitive impairment, which is widely associated with EEG alterations in PD [36]. RBD was not formally assessed in the KPD dataset, and REM sleep without atonia may have moved REM epochs to non-REM prototypes, which could affect REM-related features in the Parkinson result.

Prototypes are learned from supervised staging labels, so they follow the scorer’s labeling conventions. They may over-represent patterns that distinguish scoring categories and may under-represent physiological phenomena that cross stage boundaries. We assessed prototype coherence and sub-stage patterns against an AASM-inspired checklist. Blinded expert review, or comparison with expert-annotated micro-events such as sleep spindles, K-complexes and arousals, is the next step for these assessments.

The clinical results are hypothesis-generating. The samples are small (ASD: $N = 69$; KPD: $N = 86$), and an external cohort is the next test. The twenty quantities of C-T were chosen during development on these same cohorts and then fixed before the reported evaluation, so the margins may be optimistic for new subjects. The confirmatory claim concerns the combined C-T set. C alone is nominally significant on Alzheimer’s ($p = 0.031$, uncorrected), and T alone is not (Table 3). The reported AUCs are those of a predictor averaging five inner-loop seeds; the individual seeds span 78.3 to 88.6 on Alzheimer’s and 74.8 to 80.8 on Parkinson’s.

Age is a possible contributor to the Alzheimer margin. The ASD patients are older than the controls (74.6 ± 6.5 against 70.6 ± 7.1 years in the cohort’s published description), and subject-level age is unavailable, so age was not adjusted within the training folds. The comparator study did not adjust for age either. The Parkinson classification compares age- and sex-matched groups, so age and sex show no detectable difference between patients and controls. The analysis describes which prototypes a night occupies and which training transitions it omits; the physiological reasons remain to be tested.

Boundary wake trimming removes contiguous Wake epochs before sleep onset and after final awakening. This may have removed clinically relevant disease signal such as sleep onset latency or early morning awakening, so Wake entries may contribute less to the deficit than they would in untrimmed recordings.

Looking forward, unsupervised prototype learning, independent of staging labels, could reveal physiologically relevant patterns that supervised training suppresses. The prototype-gram could also be modelled as a sequence, beyond the coverage and transition summaries used here. Blinded expert evaluation of prototype coherence is the most direct next step.

Methods

Eleven polysomnographic datasets spanning 12,317 unique subjects were used (Supplementary Note 1): nine benchmark cohorts (SHHS [5], MrOS [37], MESA [38], WSC [39], HomePAP [40], DCSM [41], MASS [29], HMC [42], SleepEDF [43]) and two clinical cohorts for Alzheimer’s disease (ASD [2], $N=69$) and Parkinson’s disease (KPD [44], $N=86$). All recordings were harmonized to three channels (EEG, EOG, EMG), resampled to 100 Hz, bandpass-filtered (EEG/EOG: 0.3–40 Hz; EMG: highpass 10 Hz), and converted to log-power spectrograms via STFT. No per-subject normalization was applied. Dataset demographics, channel mappings, and full preprocessing details are in Supplementary Note 1.

Sequence-to-Sequence Sleep Staging

Let $\mathbf{S} \in \mathbb{R}^{L \times C \times T \times F}$ denote an input sequence of L consecutive 30-second PSG epochs, each with C recording channels, $T = 29$ time frames, and $F = 129$ frequency bins. A sequence-to-sequence sleep staging model learns a mapping $f : \mathbf{S} \mapsto \hat{\mathbf{Y}} \in \mathbb{R}^{L \times K}$, where $K = 5$ sleep stages. The architecture decomposes into two stages: an epoch encoder f_{EE} that maps each epoch independently to a D -dimensional embedding, $\mathbf{H} = f_{\text{EE}}(\mathbf{S}) \in \mathbb{R}^{L \times D}$, and a sequence encoder f_{SE} that models temporal dependencies across epochs, $\mathbf{Z} = f_{\text{SE}}(\mathbf{H}) \in \mathbb{R}^{L \times D}$. A linear classifier maps each sequence-level embedding to stage logits: $\hat{\mathbf{Y}} = \mathbf{Z}\mathbf{W} + \mathbf{b}$.

We evaluate the PSS framework on two backbone architectures:

- **SeqSleepNet** [8]: a learnable filterbank followed by a 4-layer bidirectional LSTM with attention pooling as epoch encoder, and a 4-layer bidirectional GRU as sequence encoder. Trained on MASS [29] ($L=20$, $D=128$).
- **SleepTransformer** [10]: a 4-layer Transformer encoder with sinusoidal positional encoding and attention pooling as epoch encoder, and a 4-layer Transformer encoder as sequence encoder. Trained on SHHS [5] ($L=21$, $D=128$).

In both cases, each recording channel is encoded independently by a shared epoch encoder before channel aggregation, unlike the original architectures which concatenated channels along the frequency dimension.

Proto-Sequence-to-Sequence Framework

The Proto-Sequence-to-Sequence (PSS) framework extends the standard sequence-to-sequence architecture by adding modular components that improve robustness and interpretability while allowing quantization. It is shown here on two

sequence encoders (SeqSleepNet and SleepTransformer) and is in principle compatible with others. When applied to SeqSleepNet, the result is ProtoSleepNet (PSN); when applied to SleepTransformer, the result is ProtoSleepTransformer (PST).

Each of the C channels is encoded independently by the shared epoch encoder, yielding per-channel embeddings $\mathbf{h}_c = f_{\text{EE}}(\mathbf{S}_{:,c,:}) \in \mathbb{R}^{L \times D}$ for $c = 1, \dots, C$. A learned modality embedding $\mathbf{e}_c \in \mathbb{R}^D$ is added to identify each channel type: $\mathbf{h}_c \leftarrow \mathbf{h}_c + \mathbf{e}_c$.

During training, accuracy-weighted modality dropout stochastically replaces channels with a zero-input embedding $\mathbf{h}_0 = f_{\text{EE}}(\mathbf{0})$. A per-channel auxiliary classifier produces accuracy estimates a_c ; the drop probability of channel c is proportional to its accuracy, $p_c = a_c / \sum_j a_j$, so that higher-accuracy channels are occluded more frequently, forcing the model to learn from secondary modalities. At each training step, with probability $p_{\text{apply}} = 0.5$, a number n of channels, drawn at random between 1 and $C-1$ with all values equally likely, is sampled without replacement according to $\{p_c\}$ and replaced with $\mathbf{h}_0$.

The channel embeddings $[\mathbf{h}_1, \dots, \mathbf{h}_C] \in \mathbb{R}^{C \times D}$ are processed by a single-layer Transformer encoder f_{CM} operating over the channel dimension ($d_{\text{ff}} = 256$, 4 heads). An additive residual connection preserves the original per-channel representations:

$$\mathbf{h}'_c = \mathbf{h}_c + f_{\text{CM}}([\mathbf{h}_1, \dots, \mathbf{h}_C])_c \quad (1)$$

Attention pooling aggregates across channels: $w_c = \text{softmax}(\mathbf{v}^\top \mathbf{h}'_c)$, yielding $\bar{\mathbf{h}} = \sum_c w_c \mathbf{h}'_c \in \mathbb{R}^{L \times D}$.

The pooled epoch embeddings $\bar{\mathbf{h}}$ are passed through the sequence encoder with a second additive residual:

$$\mathbf{z} = \bar{\mathbf{h}} + f_{\text{SE}}(\bar{\mathbf{h}}) \in \mathbb{R}^{L \times D} \quad (2)$$

The dual residual design ensures that the epoch-level representations $\bar{\mathbf{h}}$ remain directly accessible to the classifier. At inference, each epoch embedding can be replaced by its nearest codebook entry $\bar{\mathbf{h}}_q = \arg \min_{\mathbf{P}_j \in \mathbf{C}} \|\bar{\mathbf{h}} - \mathbf{P}_j\|_2$, discretizing the continuous representation into one of M prototypes. The quantized embedding then follows the same forward path: $\mathbf{z}_q = \bar{\mathbf{h}}_q + f_{\text{SE}}(\bar{\mathbf{h}}_q)$. This post-hoc quantization requires no retraining; the residual skip connection preserves the prototype vector in the classifier input, which is why models with residual connections tolerate quantization far better than those without (Supplementary Note 3).

Three cross-entropy losses supervise the network at different depths, each with equal weight:

$$\mathcal{L}_{\text{channel}} = \sum_{c=1}^C \text{CE}(\text{clf}_c(\mathbf{h}_c), \mathbf{y}) \quad (3)$$

$$\mathcal{L}_{\text{epoch}} = \text{CE}(\text{clf}(\bar{\mathbf{h}}), \mathbf{y}) \quad (4)$$

$$\mathcal{L}_{\text{main}} = \text{CE}(\text{clf}(\mathbf{z}), \mathbf{y}) \quad (5)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{epoch}} + \mathcal{L}_{\text{channel}} \quad (6)$$

$\mathcal{L}_{\text{channel}}$ provides per-channel accuracy estimates used by the modality dropout mechanism; $\mathcal{L}_{\text{epoch}}$ supervises the pre-sequence representations used for prototype learning; $\mathcal{L}_{\text{main}}$ is the primary staging objective.

Prototype Selection and Reconstruction

After training, the encoder is frozen and all training-set epoch embeddings are extracted. A codebook $\mathbf{C} = [\mathbf{P}_1, \dots, \mathbf{P}_M] \in \mathbb{R}^{M \times D}$ is constructed via K-Means clustering. Each test epoch is assigned to its nearest prototype: $j^* = \arg \min_j \|\bar{\mathbf{h}}_i - \mathbf{P}_j\|_2^2$. The codebook size M is selected post-hoc as an explanation granularity parameter, independently of the training procedure; a randomization test confirms that K-Means codebooks significantly outperform random exemplar codebooks for PSN at all tested sizes and for PST for $M \geq 8$ (Supplementary Note 8), and a seed stability analysis shows that the resulting prototypes are reproducible across independent training runs (Supplementary Note 9).

To interpret each prototype in the PSG input space, three reconstruction strategies are employed: (1) data-driven, which retrieves the nearest real epochs to $\mathbf{P}_k$ in latent space; (2) model-driven, which optimizes a spectrogram $\bar{\mathbf{S}}$ from random noise to minimize $\|f_{\text{EE}}(\bar{\mathbf{S}}) - \mathbf{P}_k\|_2^2$; and (3) hybrid, which seeds the optimization from the data-driven reconstruction. Reconstruction quality is assessed via fidelity (L_2 distance to prototype embedding), plausibility (MMD to the empirical data distribution), and stability (pairwise MMD across reconstructions from independent datasets). Full evaluation across 16–20 datasets per backbone is in Supplementary Note 5.

Local Instance-to-Prototype Explanations

For a given epoch $\bar{\mathbf{S}}$ assigned to prototype $\bar{\mathbf{P}}$, we define a scalar proximity objective $f_k(\bar{\mathbf{S}}) = -\|f_{\text{EE}}(\bar{\mathbf{S}}) - \bar{\mathbf{P}}\|_2^2$. Local attribution is computed by accumulating gradients along the straight-line interpolation path from a population-average baseline $\mathbf{S}_0$ to $\bar{\mathbf{S}}$:

$$\mathbf{A}(\bar{\mathbf{S}}) = (\bar{\mathbf{S}} - \mathbf{S}_0) \odot \int_0^1 \frac{\partial f_k}{\partial \mathbf{S}}(\mathbf{S}_0 + \alpha(\bar{\mathbf{S}} - \mathbf{S}_0)) d\alpha \quad (7)$$

approximated via 128 Riemann steps. The resulting attribution map $\mathbf{A} \in \mathbb{R}^{C \times T \times F}$ identifies which time-frequency regions of the input epoch drive its assignment to $\tilde{\mathbf{P}}$, with signed values indicating whether each feature increases or decreases prototype affinity. Example attributions are in Supplementary Note 4.

Prototype-Summary Generation

Each prototype is characterized through combinatorial spectral ablation on its hybrid reconstruction. The EEG spectrogram is decomposed into 8 bands ($\delta, \theta, \alpha, \sigma_l, \sigma_h, \beta_l, \beta_h, \gamma$); EOG and EMG enter as whole channels. The procedure is *add-back*. From a baseline in which the EEG spectrogram is replaced by the training-set mean and the peripheral channels are held at their reconstructed values, every one of the $2^8 = 256$ band subsets is restored and passed through the frozen encoder. The marginal importance of band i is the mean, over all subsets S that exclude it, of the reduction in latent distance obtained by adding it,

$$\text{imp}_i = \mathbb{E}_{S \not\ni i} \left[\|f_{\text{EE}}(\tilde{\mathbf{S}}_S) - \mathbf{P}_k\|_2^2 - \|f_{\text{EE}}(\tilde{\mathbf{S}}_{S \cup \{i\}}) - \mathbf{P}_k\|_2^2 \right], \quad (8)$$

a Shapley-like decomposition in which all subsets receive equal weight, averaged over the 256 reconstructed epochs of the prototype. A positive imp_i therefore means that restoring the band moves the reconstruction towards the prototype. Channel importance is computed separately, by replacing each of the three channels in turn with its training mean. Sensitivity normalizes importance by spectral deviation: $\text{sens}_i = \text{imp}_i / \max(|\Delta \text{dB}_i|, 3.0)$, where the floor of 3.0 keeps the ratio finite for features close to baseline. Band relevance is reported as a percentage of the total EEG contribution, that is, of the distance reduction obtained by restoring all eight bands at once; it is consequently not bounded to $[0, 100]$ and can be negative for bands whose restoration moves the reconstruction away from the prototype.

Each prototype is distilled into a compact four-sentence prototype summary. Sentence S1 identifies the predicted sleep stage, classification purity (fraction of assigned epochs belonging to that stage), and the channel dominance hierarchy (e.g., “EEG-driven, 72% importance”). Sentence S2 reports the two most discriminative EEG bands ranked by sensitivity, revealing which spectral features the model is most responsive to per unit of spectral deviation. Sentence S3 describes the temporal structure of the top band, as the fraction of the epoch’s frames in which it exceeds the training mean and the variability of its envelope, and the peripheral-channel context (EOG activity pattern, EMG tone level). Sentence S4 evaluates coherence against a stage-specific checklist derived from the AASM criteria [4] and assigns a verdict.

The coherence assessment applies the following investigator-defined checklist, inspired by the AASM criteria and adapted from them (Supplementary Note 6). Wake requires α elevated relative to the training mean *and* relevant ($>10\%$ band relevance; one criterion), tonic EMG, and active EOG. N1 requires θ elevated and α either suppressed or irrelevant ($<5\%$). N2 requires σ elevated in at least one of its two sub-bands, σ relevant in at least one of them ($>15\%$, applied to each sub-band separately), and quiescent EOG. N3 requires δ elevated, δ dominant ($>30\%$), and atonic EMG. REM requires EOG dominant ($>30\%$ channel importance), atonic EMG, and θ elevated. The four qualitative peripheral conditions are operationalized as percentile ranks *within the codebook*: tonic and active mean above the 60th percentile of EMG tone and EOG power across the M prototypes of that backbone, atonic and quiescent below the 40th. Each prototype is classified as *coherent* (all criteria satisfied), *partially coherent* ($\geq 50\%$), or *non-canonical* ($<50\%$). To test whether the checklist separates the stages, we assigned each prototype a stage other than its own 10,000 times, keeping the number of prototypes per stage in each codebook, and compared the number of coherent verdicts with the resulting distribution, with $p = (1 + n)/(10,000 + 1)$, where n is the number of assignments with at least as many coherent verdicts as observed. The thresholds were fixed before the verdicts were inspected. How far they can move before a verdict changes is reported in Table 2 and Supplementary Note 6. Because the peripheral conditions are relative to the codebook and the relevance thresholds are investigator-chosen, the verdicts are a structured summary of the model’s own band and channel attributions, not an independent physiological validation. Full prototype summaries, coherence results, and per-prototype spectral cards for both backbones are in Supplementary Note 6.

Evaluation Protocol

All staging metrics are computed per subject and reported as mean ± 1 standard deviation across subjects. Per-subject metric distributions are non-Gaussian (confirmed by Shapiro–Wilk tests); we therefore use Wilcoxon signed-rank tests with Bonferroni correction as the significance test for all paired model comparisons, with m the number of paired comparisons of the same analysis (staging, occlusion or residual) over all datasets and metrics together. The two confirmatory clinical comparisons use one-sided DeLong tests with Benjamini–Hochberg correction ($m = 2$). Accuracy is the fraction of correctly classified epochs. Cohen’s κ measures agreement corrected for chance. Macro-F1 is the unweighted mean of per-class F1 scores. Per-class F1 is reported for each stage (W, N1, N2, N3, REM).

Five-fold subject-stratified cross-validation with linear probing (a single linear classifier trained on frozen sequence-level embeddings, isolating representation quality from classifier complexity) and sliding-window voting is used for out-of-domain staging evaluation (Supplementary Note 1). Ablation experiments use fixed train/test splits: 34 held-out subjects from MASS (PSN), 1,636 from SHHS (PST), with four occlusion scenarios plus the clean condition (Supplementary Notes 2–3). For clinical probing, PSS models trained on benchmark datasets were applied frozen to the KPD and ASD cohorts, with none of their subjects in the training sets. Twenty per-subject features (nineteen describing which prototypes

a night visits and one describing how much of the training transition graph it never traverses) were classified by logistic regression under nested leave-one-subject-out cross-validation: an outer leave-one-subject-out loop, an inner stratified five-fold loop in which backbone, dimensionality reduction, penalty, regularization strength and scaling are all selected, and five inner-loop seeds whose probabilities are averaged. Areas under the ROC curve are compared against a classifier built on 34 hypnogram descriptors from the expert scoring over the same subjects and folds, using the one-sided DeLong test for correlated areas. Full pipeline in Supplementary Note 7.

Training Budget and Model Selection

PSS models and their backbone baselines were trained under identical budgets with code-default hyperparameters. Every run used a maximum of 50 epochs, learning rate 10^{-4} , no weight decay, batch size 32 and early-stopping patience of 10 epochs, with sequence length 20 and validation every 100 optimizer steps within the SeqSleepNet family (PSN and the one-channel, three-channel, modality-dropout and channel-mixer baselines), and sequence length 21 and validation every 1,000 steps within the SleepTransformer family. The epoch limit was set to the same value for all runs, and all other values are code defaults.

Early stopping counts epochs without improvement in validation loss. Model selection keeps the checkpoint with the lowest validation loss, identically for PSS and for the baselines.

Ethics approval and consent

All research involving human participants was conducted in accordance with the Declaration of Helsinki. This study uses publicly available datasets that received prior ethical approval and informed consent from their respective institutions, as documented in the corresponding primary publications. The KU Leuven Parkinson’s dataset (KPD) was collected during the screening phase of a project that had previously received ethical approval from the Ethics Committee Research UZ/KU Leuven (reference number S61792); participants provided written informed consent for the re-use of their data in the respective informed consent form, and the retrospective analyses reported here were subsequently approved by the Ethics Committee Research UZ/KU Leuven (reference number S70708).

Data Availability

Of the nine benchmark datasets, five are available on the NSRR [45] (SHHS, MESA, MrOS, WSC, HomePAP), two on PhysioNet (SleepEDF, HMC), and two under request (MASS, DCSM). The ASD [2] and KPD [44] datasets are under restricted access at KU Leuven and can be made available upon reasonable request to authors MG and LM respectively, pending ethical approval and a data transfer agreement.

Code Availability

The PyTorch implementations of the PSS framework, including ProtoSleepNet and ProtoSleepTransformer, are available as part of the PhysioEx library (v2.0.0) [46]. PhysioEx provides the data loaders, preprocessing pipelines, dataset splits, random seeds, and evaluation protocols used in this work. The library is available on GitHub (<https://github.com/guidogagl/physioex>) and on PyPI (`pip install physioex`). The experimental setups for model training, prototype reconstruction, prototype explanations, and clinical probing are available via <https://github.com/guidogagl/protosleepnet>. The results reported in this paper correspond to PhysioEx release v2.0.0. An interactive explainability demo and the full documentation are available at <https://protosleepnet-demo.pages.dev> and <https://guidogagl.github.io/protosleepnet>, respectively.

Acknowledgements

The funding sources below supported this work and the acquisition and maintenance of the datasets used. The dataset funders had no role in study design, analysis, interpretation, or the decision to publish. The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, “Outcomes of Sleep Disorders in Older Men,” under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The Wisconsin Sleep Cohort (WSC) Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

This research received funding from the Flemish Government (AI Research Program), FWO PhD SB Grants (1SH4Z24N), HORIZON-HLTH-2022-STAYHLTH: “Artificial intelligence-based Parkinson’s disease risk assessment and prognosis (AI-PROGNOSIS),” under Grant Agreement 101080581 and FWO project G046925N: “Task- and device-agnostic Artificial Intelligence (AI) for EEG analysis.”

Work partially supported by the Italian Ministry of Education and Research (MUR) in the framework of the FoReLab project (Departments of Excellence), and by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

Author Contributions

GG conceived the study and conducted the investigation under the supervision of MDV, MGCAC and ALA. LM and MG acquired the KU Leuven Parkinson’s dataset, and LM reviewed the physiological plausibility of the disease detection pipeline. JGC participated in the implementation of the proposed architecture under the supervision of BRK. GG wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Competing Interests

The authors declare no competing financial or non-financial interests.

References

- [1] Phan, H., Mikkelsen, K.: Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological Measurement* **43**(4), 04–01 (2022) <https://doi.org/10.1088/1361-6579/ac6049>
- [2] Heremans, E.R., Devulder, A., Borzée, P., Vandenberghe, R., De Winter, F.-L., Vandenbulcke, M., Van Den Bossche, M., Buyse, B., Testelmans, D., Van Paesschen, W., *et al.*: Wearable sleep recording augmented by artificial intelligence for alzheimer’s disease screening. *npj Aging* **11**(1), 34 (2025) <https://doi.org/10.1038/s41514-025-00219-y>
- [3] Simmonds, E., Levine, K.S., Han, J., Iwaki, H., Koretsky, M.J., Kuznetsov, N., Faghri, F., Solsberg, C.W., Schuh, A., Jones, L., *et al.*: Sleep disturbances as risk factors for neurodegeneration later in life. *npj Dementia* **1**(1), 6 (2025) <https://doi.org/10.1038/s44400-025-00008-0>
- [4] Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Lloyd, R.M., Marcus, C.L., Vaughn, B.V.: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.1. American Academy of Sleep Medicine, Darien, IL (2014)
- [5] Quan, S.F., Howard, B.V., Iber, C., Kiley, J.P., Nieto, F.J., O’Connor, G.T., Rapoport, D.M., Redline, S., Robbins, J., Samet, J.M., *et al.*: The sleep heart health study: design, rationale, and methods. *Sleep* **20**(12), 1077–1085 (1997) <https://doi.org/10.1093/sleep/20.12.1077>
- [6] Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., *et al.*: Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *Journal of sleep research* **18**(1), 74–84 (2009) <https://doi.org/10.1111/j.1365-2869.2008.00700.x>
- [7] Phan, H., Andreotti, F., Cooray, N., Chén, O.Y., De Vos, M.: Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering* **66**(5), 1285–1296 (2018) <https://doi.org/10.1109/tbme.2018.2872652>
- [8] Phan, H., Andreotti, F., Cooray, N., Chén, O.Y., De Vos, M.: Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**(3), 400–410 (2019) <https://doi.org/10.1109/tnsre.2019.2896659>
- [9] Phan, H., Lorenzen, K.P., Heremans, E., Chén, O.Y., Tran, M.C., Koch, P., Mertins, A., Baumert, M., Mikkelsen, K.B., De Vos, M.: L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics* (2023) <https://doi.org/10.1109/jbhi.2023.3303197>

- [10] Phan, H., Mikkelsen, K., Chén, O.Y., Koch, P., Mertins, A., De Vos, M.: Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering* **69**(8), 2456–2467 (2022) <https://doi.org/10.1109/tbme.2022.3147187>
- [11] Kontras, K., Chatzichristos, C., Phan, H., Suykens, J., De Vos, M.: Core-sleep: A multimodal fusion framework for time series robust to imperfect modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024) <https://doi.org/10.1109/tnsre.2024.3354388>
- [12] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13–17, 2016, pp. 1135–1144 (2016). <https://doi.org/10.1145/2939672.2939778>
- [13] Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41**(3), 647–665 (2014) <https://doi.org/10.1007/s10115-013-0679-x>
- [14] Samek, W., Arras, L., Osman, A., Montavon, G., Müller, K.-R.: Explaining the decisions of convolutional and recurrent neural networks. *Mathematical Aspects of Deep Learning* **229** (2022) <https://doi.org/10.1017/9781009025096.006>
- [15] Achibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence* **5**(9), 1006–1019 (2023) <https://doi.org/10.1038/s42256-023-00711-8>
- [16] Gagliardi, G., Alfeo, A.L., Catrambone, V., Cimino, M.G., De Vos, M., Valenza, G.: Model-driven validation of visual explanations for multimodal emotion recognition. *Machine Learning* **114**(12), 273 (2025) <https://doi.org/10.1007/s10994-025-06921-y>
- [17] Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. *The Annals of Applied Statistics* **5**(4), 2403–2424 (2011) <https://doi.org/10.1214/11-AOAS495>
- [18] Chong, P., Cheung, N.-M., Elovici, Y., Binder, A.: Toward scalable and unified example-based explanation and outlier detection. *IEEE Transactions on Image Processing* **31**, 525–540 (2021) <https://doi.org/10.1109/tip.2021.3127847>
- [19] Arik, S.O., Pfister, T.: Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research* **21**(210), 1–35 (2020)
- [20] Dreyer, M., Achibat, R., Samek, W., Lapuschkin, S.: Understanding the (extra-) ordinary: Validating deep model decisions with prototypical concept-based explanations. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3491–3501 (2024). <https://doi.org/10.1109/cvprw63382.2024.00353>. IEEE
- [21] Al-Hussaini, I., Xiao, C., Westover, M.B., Sun, J.: SLEEPER: interpretable sleep staging via prototypes from expert rules. In: *Proceedings of the 4th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research*, vol. 106, pp. 721–739. PMLR, ??? (2019)
- [22] Pei, Y., Xu, J., Yu, F., Zhang, L., Luo, W.: WaveSleepNet: An interpretable network for expert-like sleep staging. *IEEE Journal of Biomedical and Health Informatics* **29**, 1371–1382 (2025) <https://doi.org/10.1109/JBHI.2024.3498871>
- [23] Lee, H., Choi, Y.R., Lee, H.K., Jeong, J., Hong, J., Shin, H.-W., Kim, H.-S.: Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals. *npj Digital Medicine* **8**, 55 (2025) <https://doi.org/10.1038/s41746-024-01378-0>
- [24] Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., Shea-Brown, E.: Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications* **12**(1), 1417 (2021) <https://doi.org/10.1038/s41467-021-21696-1>
- [25] Somepalli, G., Fowl, L., Bansal, A., Yeh-Chiang, P., Dar, Y., Baraniuk, R., Goldblum, M., Goldstein, T.: Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In: *Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition*, pp. 13689–13698 (2022). <https://doi.org/10.1109/cvpr52688.2022.01333>
- [26] Kontras, K., et al.: NeuroAtlas: benchmarking foundation models for clinical EEG and brain–computer interfaces. *arXiv preprint arXiv:2605.14698* (2026) [arXiv:2605.14698](https://arxiv.org/abs/2605.14698) [cs.LG]

- [27] Simor, P., Wijk, G., Nobili, L., Peigneux, P.: The microstructure of REM sleep: why phasic and tonic? *Sleep Medicine Reviews* **52**, 101305 (2020) <https://doi.org/10.1016/j.smrv.2020.101305>
- [28] De Gennaro, L., Ferrara, M.: Sleep spindles: an overview. *Sleep Medicine Reviews* **7**(5), 423–440 (2003) <https://doi.org/10.1053/smrv.2002.0252>
- [29] O’reilly, C., Gosselin, N., Carrier, J., Nielsen, T.: Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research* **23**(6), 628–635 (2014) <https://doi.org/10.1111/jsr.12169>
- [30] Coon, W., Ogg, M.: Sleep EEG foundation models reveal within-stage microstructure that improves health screening beyond traditional stages. *npj Digital Medicine* (2026) <https://doi.org/10.1038/s41746-026-02970-2>
- [31] Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This looks like that: Deep learning for interpretable image recognition. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (2019)
- [32] Nauta, M., Schlötterer, J., Van Keulen, M., Seifert, C.: Pip-net: Patch-based intuitive prototypes for interpretable image classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753 (2023). <https://doi.org/10.1109/cvpr52729.2023.00269>
- [33] Jain, S., Wallace, B.C.: Attention is not Explanation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, Minneapolis, Minnesota (2019). <https://doi.org/10.18653/v1/N19-1357>
- [34] Hoehn, M.M., Yahr, M.D.: Parkinsonism: onset, progression, and mortality. *Neurology* **17**(5), 427–442 (1967) <https://doi.org/10.1212/01.wnl.0000405146.06300.91>
- [35] Schreiner, S.J., Imbach, L.L., Werth, E., Poryazova, R., Baumann-Vogel, H., Valko, P.O., Murer, T., Noain, D., Baumann, C.R.: Slow-wave sleep and motor progression in parkinson disease. *Annals of neurology* **85**(5), 765–770 (2019) <https://doi.org/10.1002/ana.25459>
- [36] Bugalho, P., Ladeira, F., Barbosa, R., Marto, J.P., Borbinha, C., Conceição, L., Salavisa, M., Saraiva, M., Meira, B., Fernandes, M.: Polysomnographic predictors of sleep, motor and cognitive dysfunction progression in parkinson’s disease: a longitudinal study. *Sleep Medicine* **77**, 205–208 (2021) <https://doi.org/10.1016/j.sleep.2020.06.020>
- [37] Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K.E., Stefanick, M.L., Laffan, A., Stone, K.L., Men Study Group, O.F.: Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society* **59**(12), 2217–2225 (2011) <https://doi.org/10.1111/j.1532-5415.2011.03731.x>
- [38] Chen, X., Wang, R., Zee, P., Lutsey, P.L., Javaheri, S., Alcántara, C., Jackson, C.L., Williams, M.A., Redline, S.: Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep* **38**(6), 877–888 (2015) <https://doi.org/10.5665/sleep.4732>
- [39] Young, T., Palta, M., Dempsey, J., Peppard, P.E., Nieto, F.J., Hla, K.M.: Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study. *WMJ: official publication of the State Medical Society of Wisconsin* **108**(5), 246 (2009)
- [40] Rosen, C.L., Auckley, D., Benca, R., Foldvary-Schaefer, N., Iber, C., Kapur, V., Rueschman, M., Zee, P., Redline, S.: 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 (2012) <https://doi.org/10.5665/sleep.1870>
- [41] Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., Igel, C.: U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine* **4**(1), 72 (2021) <https://doi.org/10.1038/s41746-021-00440-5>
- [42] Alvarez-Estevéz, D., Rijsman, R.: Haaglanden Medisch Centrum sleep staging database. *PhysioNet* (2022) <https://doi.org/10.13026/t79q-fr32> . Version 1.1
- [43] Kemp, B., Zwirnerman, A.H., Tuk, B., Kamphuisen, H.A.C., Obery, J.J.L.: Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering* **47**(9), 1185–1194 (2000) <https://doi.org/10.1109/10.867928>

- [44] Micca, L., Albouy, G., King, B.R., d’Cruz, N., Nieuwboer, A., Vandenbergh, W., Borzée, P., Buyse, B., Testelmans, D., Nicolas, J., et al.: The effect of a post-learning nap on motor memory consolidation in people with parkinson’s disease: A randomised controlled trial. *Journal of Sleep Research*, 70203 (2025)
- [45] Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., Redline, S.: The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* **25**(10), 1351–1358 (2018) <https://doi.org/10.1093/jamia/ocy064>
- [46] Gagliardi, G., Alfeo, A.L., Cimino, M.G.C.A., Valenza, G., De Vos, M.: Physioex: a new python library for explainable sleep staging through deep learning. *Physiological Measurement* (2025) <https://doi.org/10.1088/1361-6579/adaf73>

Figure Legends

Figure 1. The Proto-Sequence-to-Sequence framework for interpretable sequence-to-sequence sleep staging. The framework extends sequence-to-sequence architectures such as the two tested here with three components: (i) a channel mixer with dual residual connections, drawn as the two dashed arcs that bypass the mixer and the sequence encoder, fusing the channels after a shared epoch encoder has encoded each PSG channel independently, (ii) a sequence encoder that captures inter-epoch temporal context, and (iii) post-hoc K-Means prototype assignment on the resulting embedding space. This design yields a discrete, characterizable codebook of sleep sub-stage patterns while preserving staging accuracy. Prototype instances can be reconstructed in the input space and summarized as compact prototype summaries describing the power spectral density of each channel together with its importance. Each encoder uses the module of its backbone: recurrent for ProtoSleepNet (PSN) and self-attention for ProtoSleepTransformer (PST). Modality dropout acts during training only and is described in the Methods.

Figure 2. Sleep staging performance and post-hoc quantization capabilities. (a) Cohen’s kappa for ProtoSleepNet (PSN) versus the single-channel SeqSleepNet-1ch (left) and ProtoSleepTransformer (PST) versus SleepTransformer-1ch (right). Filled markers are in-domain datasets, evaluated with the end-to-end models; open markers are out-of-domain datasets, evaluated by linear probing on frozen encoders. The 13 markers cover 20 dataset-sources: the five MASS cohorts form one marker and the four WSC visits another. Per-source values are in Supplementary Note 1. PSN outperforms SeqSleepNet-1ch on 13 of 15 out-of-domain sources. (b) Staging accuracy (solid lines) and Cohen’s kappa (dashed lines) as a function of the post-hoc quantized codebook size, denoted M , ranging from 5 to 100. As in panel (a), SeqSleepNet-1ch (SSN) and SleepTransformer-1ch (ST) are the single-channel references. Horizontal dotted lines show the non-quantized performance of each model. Performance rises with codebook size and flattens between approximately 48 and 65 for both backbones. At $M = 100$ quantization costs Cohen’s κ 0.013 for PSN and 0.020 for PST; the epoch-level cost is in Supplementary Note 3.5.

Figure 3. Quantitative comparison of three prototype reconstruction methods across both backbones (PSN and PST). (a) Fidelity: the L2 distance between reconstruction embeddings and prototype embeddings (lower is better). The hybrid method achieves the best fidelity. (b) Plausibility: the maximum mean discrepancy between reconstructions and empirical instances close to each prototype (lower is better). Data-driven plausibility is zero by construction; model-driven reconstructions diverge substantially from the empirical distribution. (c) Stability: the pairwise maximum mean discrepancy between reconstructions derived from independent datasets (lower is better). Model-driven reconstructions are dataset-independent by design, whereas data-driven and hybrid reconstructions show cross-dataset variability. Each box summarizes the distribution across the 12 prototypes.

Figure 4. Prototype codebook summary for both backbones (12 prototypes each). Each row is a prototype, grouped by dominant sleep stage (colour strip on the left). The central heatmap shows EEG band relevance as a percentage on a diverging colour scale, with cool colours indicating negative relevance and warm colours indicating positive relevance; values with an absolute magnitude of at least 5 percent are annotated. Relevance is renormalized per prototype so that the absolute values of the eight bands sum to 100, which is what makes the rows comparable; the raw quantities differ from the raw values. The stacked bars on the right show per-prototype channel importance (EEG, EOG and EMG). Dashed lines separate stage groups. (a) ProtoSleepNet (PSN, trained on MASS): two Wake, six N2, two N3 and two REM prototypes. (b) ProtoSleepTransformer (PST, trained on SHHS): four Wake, four N2, two N3 and two REM prototypes. Both backbones converge on delta-dominated N3 and on an eye-movement-driven versus EEG-driven REM split. PSN resolves the N2 delta/sigma relevance gradient; the N2 entries of PST have σ_h below 11%. Neither backbone has a dedicated N1 prototype.

Figure 5. Exploratory clinical probing with prototype-grams. (a, b) Nested leave-one-subject-out area under the ROC curve for Alzheimer’s disease and Parkinson’s disease, for three feature sets: C, the 19 coverage features; T, the single transition feature; and C-T, their union. The dashed line marks the expert scoring-based comparator. Bars are 95 percent confidence intervals. An asterisk marks the two confirmatory comparisons, C-T against the comparator,

corrected for multiplicity (Benjamini–Hochberg, $m = 2$); the probabilities on the T and C rows are descriptive. **(c, f)** Training transition graph at a codebook size of 12, one panel per group. Each node is a prototype, sized by its use in training and drawn as a pie of its mean staging posterior; arrow width is the transition mass and darkness the fraction of the group traversing it, so untraversed edges are faint. **(d, g)** The same transition-mass per subject, with in-sample (not cross-validated) area under the curve and Hedges’ g at a codebook size of 12. **(e, h)** Distinct prototypes a night visits within a fixed epoch budget, against codebook size; thin lines are subjects, thick lines group medians. The Alzheimer cohort is orange, the Parkinson cohort blue and controls grey throughout.