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

Guido Gagliardi[1*], Javier García Ciudad[2], Letizia Micca[3], Birgitte Rahbek Kornum[2], Moran Gilat[3], Antonio Luca Alfeo[4,7], Mario G.C.A Cimino[5], Maarten De Vos[1,6*]

[1]Department of Electrical Engineering, KU Leuven, Leuven, Belgium.
[2]Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark.
[3]Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium.
[4]Dept. of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy.
[5]Department of Information Engineering, University of Pisa, Pisa, Italy.
[6]Department of Development and Regeneration, KU Leuven, Leuven, Belgium.
[7]SMARTEST 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 including Parkinson's and Alzheimer's. Automated sleep staging networks achieve competitive performance on multi-channel polysomnography, but they operate as black boxes, limiting clinical trust and hiding whatever sleep biomarkers the models may have learned.

We introduce the Proto-Sequence-To-Sequence framework, a prototype-based wrapper that adds interpretability to any sequence-to-sequence sleep staging backbone by constraining the model to rely on a discrete codebook learned post-hoc. The codebook entries, i.e. the model prototypes, can be reconstructed in the original EEG, EOG and EMG input space, validated against the sleep staging clinical practice, and clinically probed.

Applied to two architectures: ProtoSleepNet (PSN, extending SeqSleepNet) and ProtoSleepTransformer (PST, extending SleepTransformer), the framework has been validated across 11 independent datasets encompassing 12,317 subjects, forming 16,412 overnight PSG recordings, consisting of more than 120,000 total hours of data. The framework proved to maintain competitive staging performance while conferring robustness under channel occlusion conditions and intrinsic explainability.

Prototype analysis reveals physiologically meaningful sub-stage patterns, such as delta-dominated deep sleep, phasic–tonic REM differentiation, and an N2 $\boldsymbol{\delta}$–$\boldsymbol{\sigma}$ reciprocity continuum, that converge across both backbone architectures. In our disease detection pipeline, per-subject prototype-based statistics, separate subjects with Alzheimer's and Parkinson's disease from controls in within-cohort leave-one-subject-out validation, with disease-specific stage signatures consistent with known neuropathological progression.

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

Introduction

Sleep disorders, affecting 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, with growing evidence identifying sleep disturbances as both markers 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 therefore support earlier detection and intervention, yet such screening demands both accurate automated sleep staging and interpretable representations of what staging models learn about sleep physiology.

Sleep assessment involves recording polysomnographic (PSG) data and scoring each 30-second epoch into one of five stages (Wake, N1, N2, N3, and REM) according to established criteria [4]. Although manual scoring is costly and labor-intensive, in-home PSG devices [5] and wearable sensors [2] have made large-scale data collection feasible, and automatic sleep staging algorithms now achieve competitive accuracy [6]. Among these, the sequence-to-sequence deep learning framework [7] has emerged as a particularly effective paradigm, pairing an epoch encoder that embeds individual

30-second windows with a sequence encoder that processes L successive epochs jointly, exploiting the longitudinal dependence between sleep states within a sleep cycle [8–10]. This architecture spans two dominant design families: recurrent networks [8] and self-attention transformers [10], whose architectural diversity offers a natural opportunity for cross-validation: if an interpretable representation yields similar physiological findings when built on fundamentally different encoders, the findings are more likely to reflect properties of the data than artifacts of the architecture.

Despite their effectiveness, these models share three interrelated limitations. Their opaque decision-making prevents clinicians from verifying whether the model relies on physiologically expected features (delta power for N3, sigma-band activity for N2, EOG dynamics for REM) or on spurious correlations. Beyond stage labels, sub-stage features captured during training, i.e., patterns that may reflect sleep-spindle dynamics, arousal structure, or pathological disruptions associated with neurodegenerative disease, remain locked within the latent space, inaccessible to clinical interpretation or downstream analysis. Finally, naïve multi-channel fusion may introduce a multimodal competition problem [11]: EEG signals, which carry the strongest staging gradient, tend to dominate training at the expense of EOG and EMG, making the model brittle when the dominant modality is noisy or absent. This scenario could be common in home-based [5] and wearable [2] recordings where sensor quality cannot be guaranteed.

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, however, produce instance-level explanations whose precision degrades when aggregated into global summaries [16]. Prototype-based learning offers a different route: the model routes its decisions through a discrete set of interpretable representations, i.e. the prototypes, that function as a condensed vocabulary of the input space, providing simultaneous local and global interpretability [17–20].

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 neighbourhoods. This assumption generally holds at the final classification layer but often fails in intermediate representations, where small perturbations can produce drastically different outcomes [21, 22]. Conversely, constraining the architecture to route all predictions through learned prototypes requires end-to-end retraining and typically reduces performance due to quantization of the decision space. Finally, prototypes are discrete latent-space vectors, and as such, it is not possible for human experts to directly inspect and understand their physiological meaning. Bridging the gap between each vocabulary entry and a clinically meaningful physiological description requires its reconstruction in the input space, decomposition into interpretable features, and validation against established scoring criteria.

Here, we propose the Proto-Sequence-To-Sequence (PSS) framework (Fig. 1), an architecture-agnostic prototype-based wrapper of sequence-to-sequence sleep staging that tackles these challenges by shaping the embedding space for post-hoc quantization via dual residual connections, so that a codebook of discrete prototypes can be constructed after training without degrading staging quality. The framework wraps any backbone encoder: we demonstrate it on SeqSleepNet [8] (ProtoSleepNet, PSN) and SleepTransformer [10] (ProtoSleepTransformer, PST), and validate it across 11 independent PSG dataset composed by different cohorts and subgroups yielding 20 different dataset-sources encompassing over 12,000 subjects. We trained the PSS variants on the same backbone training setup: PSN was trained using the 5 cohorts of the MASS dataset (200 subjects) and PST was trained using the visit-1 of the SHHS dataset (5,793 subjects). Because prototypes are latent vectors with no direct physiological meaning, we develop a reconstruction and characterization pipeline that maps each codebook entry back to the input space, decomposes it into physiologically defined spectral and temporal features, and distills the result into compact matching summaries accessible to sleep clinicians. The framework provides explanations at both the local level, i.e. why a specific epoch is assigned to a given prototype, and the global level, i.e. what each prototype encodes across the full population of assigned epochs. The global characterizations are validated against AASM scoring criteria independently on both backbones, so that convergent findings across architectures provide evidence that the discovered patterns reflect properties of the data rather than of a particular encoder. As a further test of physiological validity, we probe whether statistics extracted from the prototype-gram, i.e., the subject flow of prototypes during the night, computed from frozen staging-only models that have never seen disease labels, can separate clinical groups known to differ in sleep architecture.

We show that both PSS variants preserve competitive staging performance while conferring robustness to missing channels. Prototype characterization reveals physiologically coherent sub-stage patterns that converge across both backbone architectures, i.e., delta-dominated N3, phasic–tonic REM differentiation, and an N2 delta–sigma reciprocity continuum, and that are largely consistent with AASM scoring criteria. In exploratory clinical probing, prototype-grams discriminate Alzheimer's and Parkinson's disease from controls, with disease-specific stage signatures consistent with known neuropathological progression. These 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

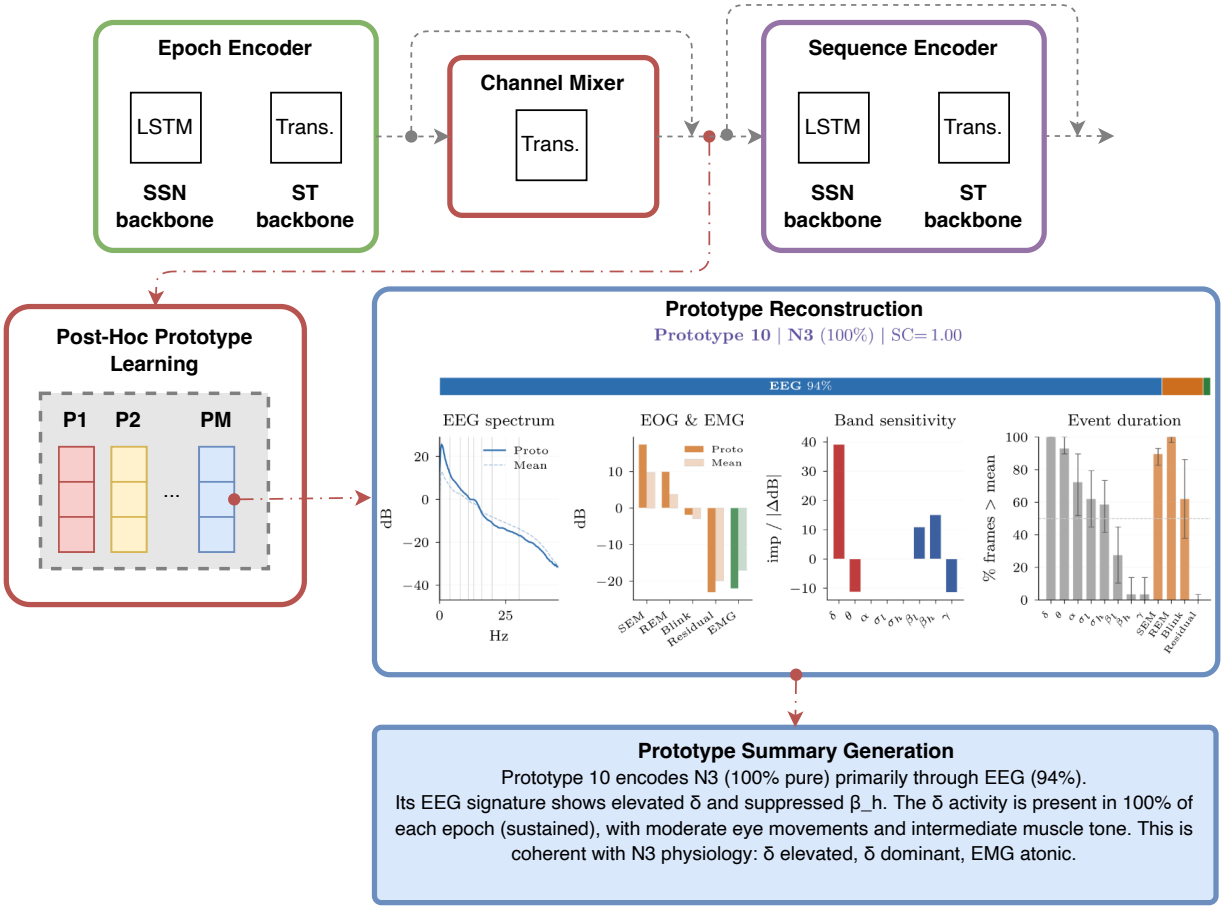

Fig. 1 The Proto-Sequence-to-Sequence framework for interpretable sequence-to-sequence sleep staging. The framework extends any sequence-to-sequence architecture with three components: (i) channel dropout and a channel mixer with dual residual connections that process each PSG channel independently before fusion, (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 (PSD) of each channel together with its importance. Illustrated for ProtoSleepNet (PSN); the ProtoSleepTransformer (PST) variant replaces the recurrent encoder with a self-attention encoder.

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. 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, i.e. on the complete held-out dataset-sources, 15 for PSN and 19 for PST, we performed linear probing on frozen sequence-encoder outputs using 5-fold subject-stratified cross-validation [23] for each domain. Per-subject differences were assessed using paired Wilcoxon signed-rank tests with Bonferroni correction.

Figure 2a shows per-dataset κ as spider charts (filled: in-domain; open: out-of-domain). In-domain, both PSS variants preserve staging quality: PSN $\kappa = 0.801 \pm 0.070$ vs. SSN 0.788 ± 0.068 on MASS, and PST $\kappa = 0.814 \pm 0.100$ vs. ST 0.812 ± 0.103 on SHHS. Out-of-domain, PSN significantly outperforms SSN on 13 of 15 datasets (Wilcoxon $p < 0.05$, Bonferroni-corrected), with the largest gain on MESA; PST significantly outperforms ST on 5 of 19 datasets, with near-parity on the remaining 14 (largest gain: MESA, +0.051). The out-of-domain gains reflect channel dropout trading in-domain specialization for cross-dataset robustness (Supplementary Section 1). Staging performance is stable across six independent training seeds (coefficient of variation $< 1\%$ for κ; Supplementary Section 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 five scenarios: 25% and 50% random channel replacement, complete EEG removal, and complete EOG+EMG removal. Progressive ablation

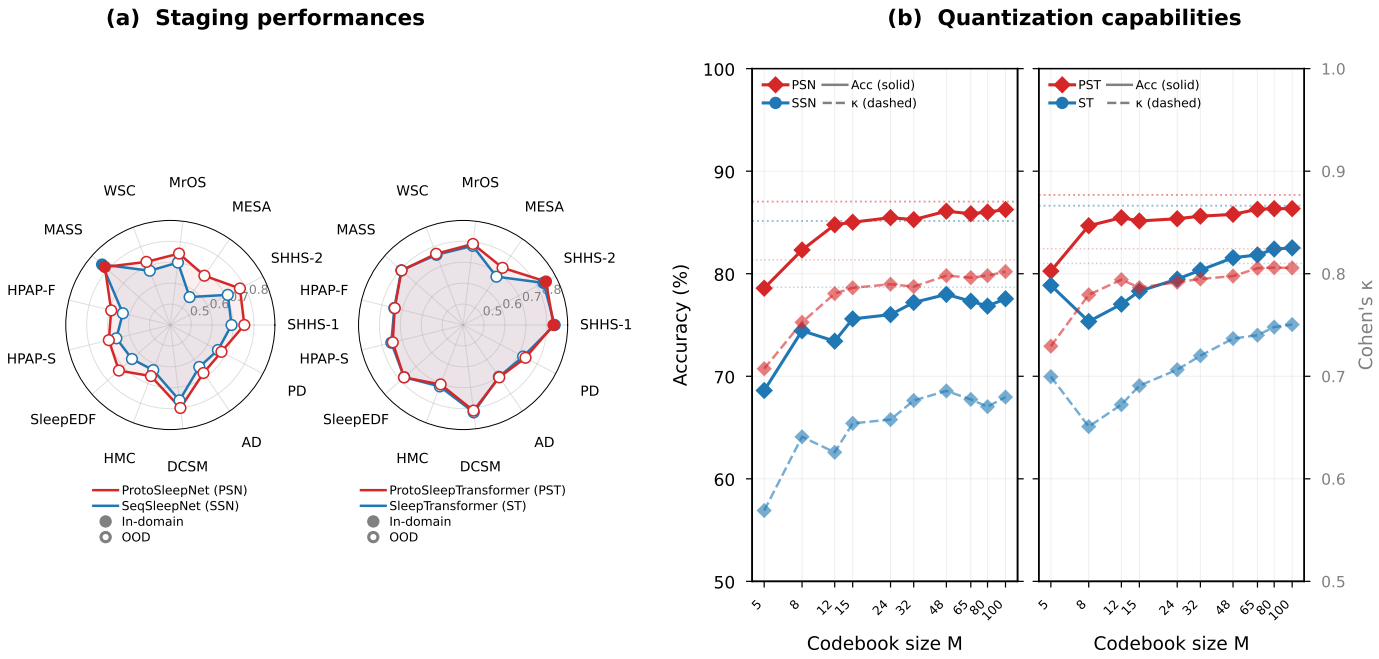

Fig. 2 Sleep staging performance and post-hoc quantization capabilities. (a) Cohen's κ across 13 evaluation dataset-sources for ProtoSleepNet (PSN) vs. SeqSleepNet (left) and ProtoSleepTransformer (PST) vs. SleepTransformer (right). MASS results are aggregated across 5 cohorts, WSC results are aggregated across 4 different visits, yielding to a total of 20 dataset-sources. Filled markers indicate in-domain datasets; open markers indicate out-of-domain datasets. PSN consistently outperforms SSN on out-of-domain cohorts. **(b)** Staging accuracy (solid) and Cohen's κ (dashed) as a function of post-hoc quantized codebook size M (5–100). Horizontal dotted lines show non-quantized baseline performance of the respective backbone model and the PSS variant. Performance saturates at $M \approx 48$–65 for both backbones, confirming that post-hoc discretization preserves staging quality.

identifies channel dropout as the primary robustness mechanism: without it, both baselines collapse catastrophically (SSN: $\kappa = 0.788 \rightarrow 0.076$ at 50% random occlusion; ST: $0.812 \rightarrow 0.556$ without EEG), whereas adding dropout alone, a zero-parameter regularization, restores near-clean performance. The channel mixer then recovers clean-condition accuracy lost to the dropout regularization. Modality importance is backbone-dependent: PSN exhibits a pronounced EEG dependency ($\kappa = 0.357$ without EEG), whereas PST distributes capacity more evenly ($\kappa = 0.756$ in the same condition). Full ablation results are reported in Supplementary Section 2.

PSN and PST allow Post-Hoc Quantization without staging loss

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. If this quantization degrades staging quality, the prototypes would misrepresent the model's actual representations.

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 gap never closes for the unmodified backbones: at $M = 100$, PSN retains $\kappa = 0.788$ (non-quantized: 0.801), whereas SSN drops to 0.675 (from 0.788); PST retains 0.794 (from 0.814), whereas ST drops to 0.736 (from 0.812).

Ablation pinpoints the dual residual connection as the enabler: without residuals, K-Means quantization yields $\kappa = 0.568$ at $M = 5$; adding residuals recovers $\kappa = 0.715$ (+0.147), shaping the embedding space into a smoother manifold amenable to centroid-based clustering (Supplementary Section 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 while avoiding the training instabilities of gradient-based codebook learning (κ ranging from 0.534 to 0.709 for randomly initialized VQ; Supplementary Section 3). A randomization test further confirms that K-Means codebooks capture genuine embedding structure: at all tested sizes, K-Means significantly outperforms random exemplar codebooks for both backbones ($p \leq 0.04$ for PSN, $p < 0.05$ for PST at $M \geq 8$; Supplementary Section 8). A seed stability analysis across six independent training runs per backbone shows that the resulting prototypes are reproducible: staging κ varies by at most 1% (CV), and the input-space distributional shift between matched prototype reconstructions remains within 2× of the intra-seed sampling baseline for both backbones at all but one codebook size (Supplementary Section 9).

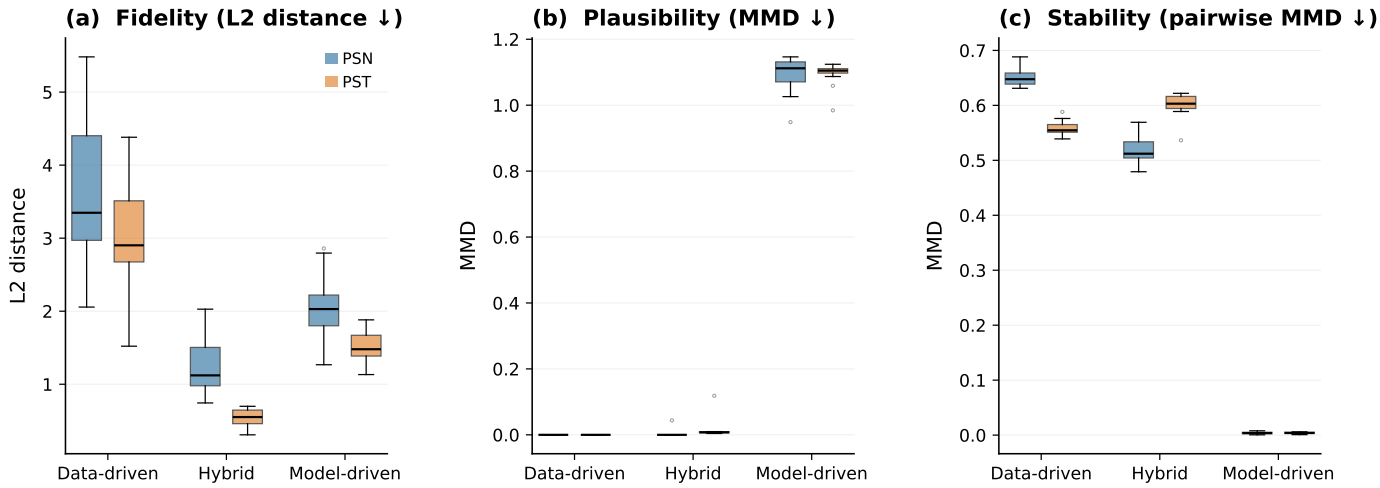

Fig. 3 Quantitative comparison of three prototype reconstruction methods across both backbones (PSN and PST). **(a)** Fidelity: L2 distance between reconstruction embeddings and prototype embeddings (lower is better). Hybrid achieves the best fidelity. **(b)** Plausibility: MMD between reconstructions and empirical instances proximate to each prototype (lower is better). Data-driven plausibility is zero by construction; model-driven reconstructions diverge substantially from the empirical distribution. **(c)** Stability: pairwise MMD between reconstructions derived from independent datasets (lower is better). Model-driven reconstructions are dataset-independent by design; data-driven and hybrid exhibit cross-dataset variability. Each box summarizes the distribution across $M = 12$ prototypes.

Reconstructing quantized entries in the PSG input space

Before inspecting individual prototypes, the practitioner must choose the inspection level M. More prototypes yield a more faithful representation of the network's decision mechanism but require analyzing more instances; fewer prototypes give a coarser but more manageable vocabulary. We set $M = 12$ for all subsequent analyses, a trade-off between granularity and interpretability confirmed by the codebook size sweep in the previous subsection.

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). This is non-trivial because prototypes capture only the features relevant for prototype assignment and unconstrained 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 achieve dataset independence, i.e. greater stability, but collapse to non-physiological spectral attractors: the sigma-to-delta power ratio is inflated 17-fold for PSN and 9-fold for PST relative to empirical recordings. Hybrid reconstruction achieves the best fidelity–plausibility trade-off, with near-zero plausibility MMD across external cohorts (PSN: 0.004 ± 0.012; PST: 0.017 ± 0.031) while maintaining substantially better fidelity than the data-driven approach (PSN: 1.24 ± 0.35; PST: 0.55 ± 0.12).

Based on these results, we adopt hybrid reconstruction as the default for all subsequent prototype characterization analyses, as it provides reconstructions that are both faithful to the model's internal representation and plausible as real PSG signals (Supplementary Section 5).

Prototype characterization and coherence with AASM criteria

Each of the $M = 12$ hybrid-reconstructed prototypes per backbone is characterized via combinatorial ablation of physiologically defined spectral bands (8-band EEG decomposition including separate σ_l [11–13 Hz] and σ_h [13–16 Hz] sub-bands; 4-band EOG; EMG tone), producing per-prototype sensitivity profiles that quantify which signal features drive prototype assignment. Each profile is summarized as a compact prototype summary in a structured four-sentence format covering stage purity, discriminative spectral bands, temporal structure, and an AASM coherence verdict (Supplementary Section 6, Tables 1–2).

Figure 4 shows the resulting 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 the most visually striking: PSN P10 ($\delta = 68\%$) and PST P4/P8 ($\delta = 88\%/82\%$) appear nearly monochromatic on the δ column, with all remaining

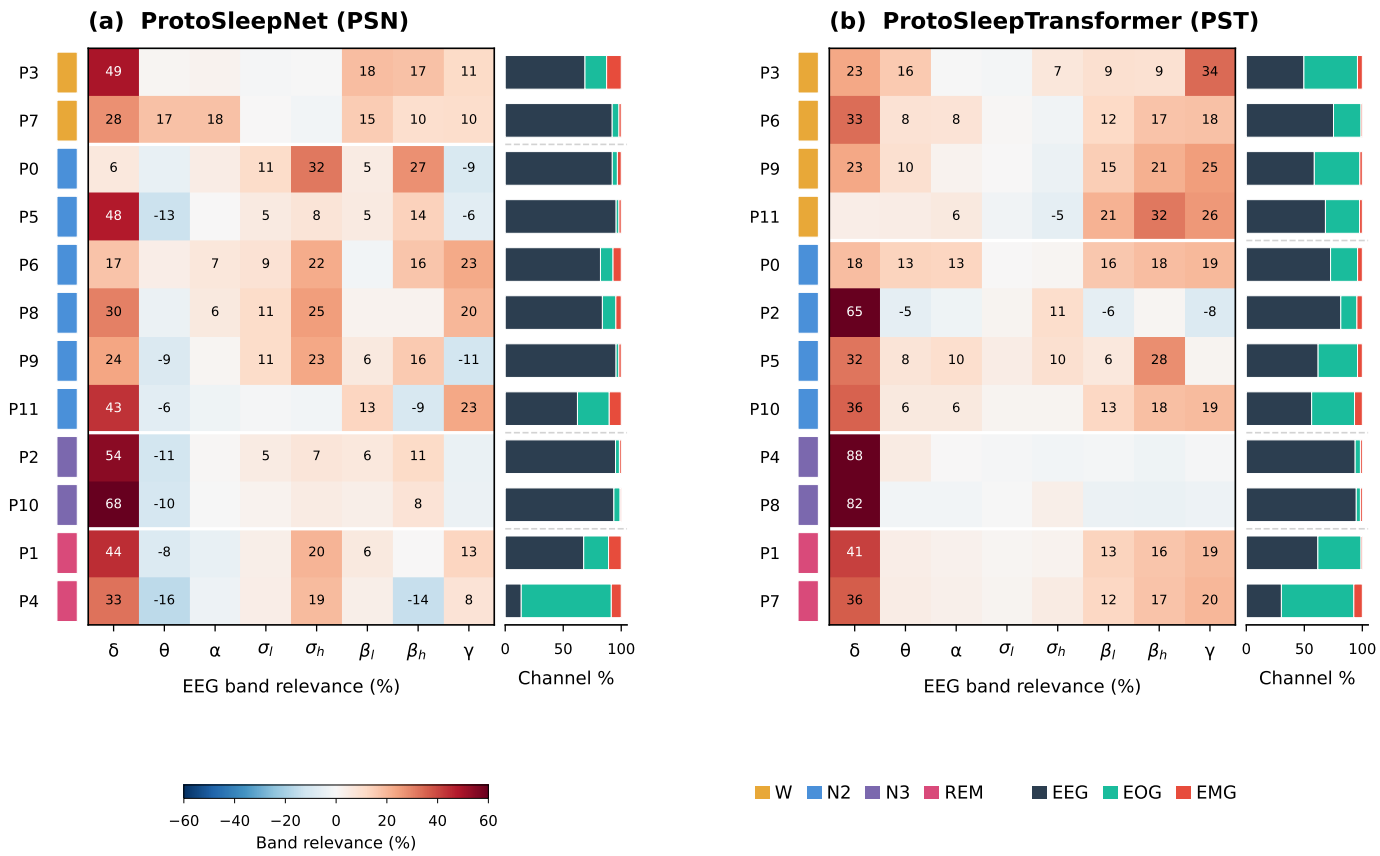

Fig. 4 Prototype codebook summary for both backbones ($M = 12$ prototypes each). Each row is a prototype, grouped by dominant sleep stage (colour strip, left). The central heatmap shows normalized EEG band relevance (%, diverging colour scale); values $\geq |5|$% are annotated. The right stacked bars show per-prototype channel importance (EEG, EOG, EMG). Dashed lines separate stage groups. **(a)** ProtoSleepNet (PSN, trained on MASS): 2 Wake, 6 N2, 2 N3, 2 REM prototypes. **(b)** ProtoSleepTransformer (PST, trained on SHHS): 4 Wake, 4 N2, 2 N3, 2 REM prototypes. Both backbones converge on delta-dominated N3, a phasic–tonic REM bifurcation (EOG- vs. EEG-driven), and an N2 delta–sigma reciprocity continuum. Neither backbone allocates a dedicated N1 prototype.

bands at or below |10%| relevance. Both backbones also show a mild θ suppression (-10% to -11% in PSN), indicating that the model actively penalizes theta, a frequency band shared with N1 and drowsy wakefulness, to sharpen the N3 boundary. The channel bars are nearly entirely EEG (94–95%), confirming that N3 is encoded as a purely cortical phenomenon. This is the strongest cross-backbone convergence in the entire figure: PSN and PST produce visually identical N3 signatures, reflecting the unambiguous spectral dominance of slow-wave activity.

In the REM rows of Figure 4, the channel bars reveal a sharp split: PSN P4 is nearly entirely teal (EOG = 78%), while PSN P1 is nearly entirely dark (EEG = 68%). The same inversion appears in PST: P7 (EOG = 62%) versus P1 (EEG = 62%). This channel-importance contrast reflects a phasic–tonic bifurcation within REM sleep that both backbones discovered independently without explicit phasic/tonic labels in the training data, recapitulating the framework of Simor et al. [24]. In the heatmap, both REM variants show moderate δ relevance (33–44% in PSN, 36–41% in PST) and σ_h relevance (19–20%), but θ is consistently negative or absent (-8% to -16% in PSN, below threshold in PST). This theta paradox, both backbones ignore AASM's defining REM marker, reflects the model's discovery that θ is insufficiently discriminative because it also characterizes N1 and drowsy wakefulness; instead, the model relies on EOG dynamics (phasic) or δ level combined with high-frequency activation (tonic).

The N2 group is the most heterogeneous in both panels of Figure 4, with six PSN prototypes and four PST prototypes tiling a continuous spectral gradient. In the PSN panel, the gradient is clearly visible along the σ_h column: P0 ($\delta = 6\%$, $\sigma_h = 32\%$) is a nearly pure sigma-driven prototype, while 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: $\sigma_h = 22\%$) populate the balanced zone where sigma-band and delta-band power coexist. This gradient reflects the reciprocal relationship between spindle density and slow-wave power documented by De Gennaro & Ferrara [25], where σ/δ ratio indexes NREM depth. Crucially, the PST panel tells a different story: the σ_h column is mostly white ($< 11\%$ in all PST N2 prototypes), and instead PST classifies N2 primarily through broadband power levels and high-frequency suppression (P5: $\beta_h = 28\%$; P0: $\beta_h = 18\%$, $\gamma = 19\%$). This difference, i.e., PSN extracts transient spindle oscillations more effectively than PST, is the largest systematic difference between the two codebooks and has clinical implications: when spindle detection is diagnostically important, recurrent architectures like PSN may provide more physiologically grounded N2 representations.

In the Wake rows of Figure 4, the α column is nearly white: the highest alpha relevance across all six Wake prototypes is 18% (PSN P7), far below what AASM's α-centric definition would predict. Instead, the β_l, β_h, and γ columns show

Table 1 Representative prototype summaries for six sub-stage patterns across both backbones (M=12, hybrid reconstruction). One prototype per pattern is selected from each backbone to illustrate convergent and divergent classification strategies. Each summary follows a four-sentence format: (S1) stage identity and channel importance, (S2) primary and secondary spectral features by sensitivity, (S3) temporal structure and EMG/EOG dynamics, (S4) AASM coherence verdict. C = coherent, P = partially coherent, X = non-canonical. Full prototype summaries for all 24 prototypes are in Supplementary Section 6, Tables 1–2.

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 (sustained), 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 (sustained), 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 (transient), 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 (transient), 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 (transient), 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 (transient), 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 (transient), 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 (intermittent), 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 (sustained), 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 spindle activity.
	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 (transient), 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 (transient), 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 (sustained), with active eye movements (Blink 100% active) and tonic muscle tone. Partially coherent: confirmed EMG tonic, EOG active; missing α elevated+relevant.

warm colours throughout the Wake group: PSN P3 (β_l = 18%, β_h = 17%, γ = 11%) and PST P3 (γ = 34%) classify wakefulness through elevated beta and gamma power. The channel bars reveal a further divergence: PST Wake prototypes allocate 24–46% to EOG (teal), using eye-movement patterns as a secondary classification feature, whereas PSN relies almost exclusively on EEG (69–92%). This systematic alpha bypass is consistent across both architectures and training populations, suggesting that α is insufficiently reliable for wake classification when diverse populations (eyes-open vs. eyes-closed, individual α variability, drowsy states) are pooled in training.

Neither backbone allocates a dedicated N1 prototype, despite N1 occupying 1–5% of training sleep. Instead, N1 epochs are absorbed into boundary prototypes: PSN P11 (N2 with negative σ relevance and δ-dominant spectral profile) and PSN P3 (Wake with 80% purity, 20% N1 spillover) straddle the N1/N2 and W/N1 transitions respectively. This absence reflects a physiologically sensible allocation strategy: N1 is defined by the *absence* of graphoelements (no spindles, no K-complexes, no α) rather than by distinctive features, exhibits the lowest inter-rater reliability among stages ($\kappa \approx 0.25$ [6]), and provides a weak training signal due to low prevalence. The codebook allocates capacity proportionally to spectral distinctiveness: N3 (unambiguous δ) receives 2 prototypes universally, N2 (heterogeneous) receives 4–6, Wake receives 2–4, and N1, which is spectrally ambiguous and transient, is absorbed rather than represented.

To quantify the alignment of these findings with clinical practice, each prototype was assessed against stage-specific AASM criteria (Table 2). Of the 24 prototypes, 10 are rated Coherent (42%), 10 Partially coherent (42%), and 4 Non-canonical (17%). Partially coherent verdicts do not indicate model failure; they identify prototypes whose discriminative strategy diverges from canonical AASM criteria while remaining physiologically valid (tonic REM lacking EOG bursts by definition, deep N2 prioritizing δ over σ, Wake using high-frequency rather than α). Non-canonical verdicts (PSN P3, P5; PST P0, P6) arise from two systematic causes: population-driven α bypass in Wake, and deep-N2 prototypes whose σ/δ ratio approaches N3 territory. Importantly, PSN achieves higher coherence than PST (6C vs 4C), reflecting the recurrent architecture's greater sigma utilization in N2. Table 1 presents representative prototype summaries for each sub-stage pattern; the complete set of 24 summaries and the full coherence table are reported in Supplementary Section 6. Local instance-to-prototype explanations are detailed in Supplementary Section 4.

Table 2 Summary of AASM coherence assessment across both backbones. C = Coherent (all stage-specific criteria met), P = Partially coherent (≥50% criteria met), X = Non-canonical (<50% criteria met).

Stage	PSN	PST	Combined	Key observation
N3	1C, 1P	2C	3C, 1P	δ dominance aligns with AASM
REM	1C, 1P	1C, 1P	2C, 2P	Phasic coherent; tonic partial (EOG absent by definition)
N2	3C, 2P, 1X	1C, 2P, 1X	4C, 4P, 2X	Coherence tracks σ relevance
W	1C, 1X	3P, 1X	1C, 3P, 2X	α bypassed in both backbones
N1	—	—	—	No dedicated prototype; absorbed by boundaries
Total	6C, 4P, 2X	4C, 6P, 2X	10C, 10P, 4X	

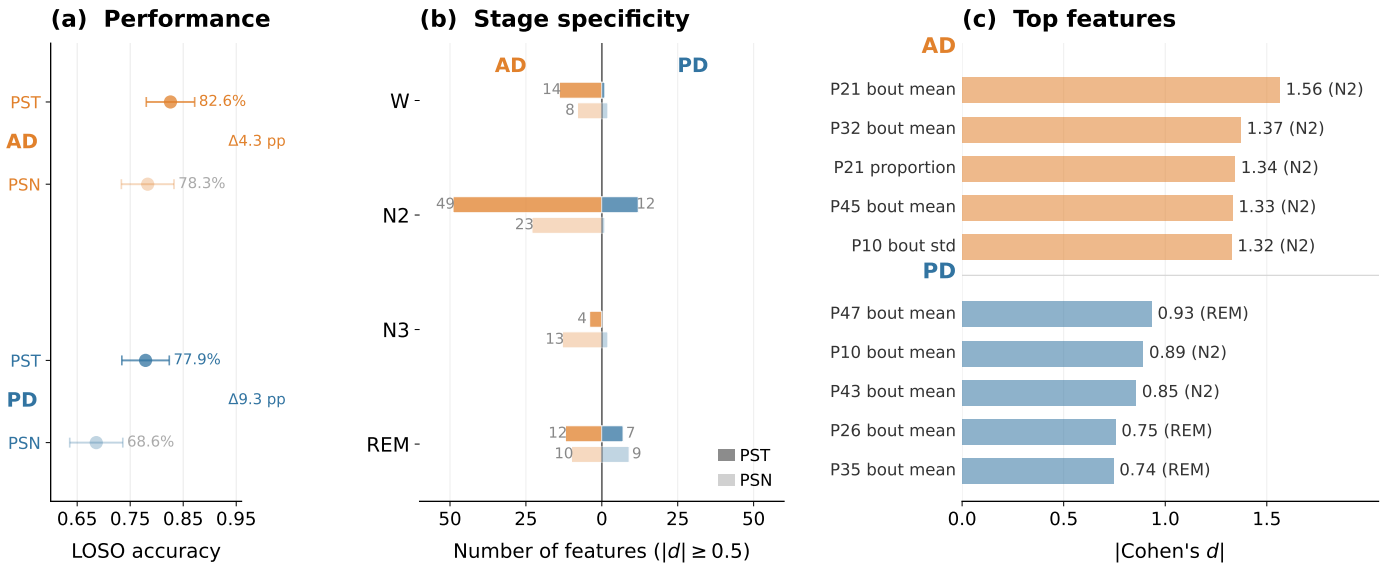

Fig. 5 Exploratory clinical probing with prototype-grams. (a) LOSO accuracy for AD and PD diagnosis, comparing PST and PSN architectures (best feature set per architecture; see text). Error bars: ±1 binomial standard deviation. Parkinson's cohort in blue, Alzheimer's in orange. **(b)** Stage-specificity butterfly: number of discriminative features ($|d| \geq 0.5$) by sleep stage, mirrored AD (left) vs. PD (right). Solid bars: PST; faded bars: PSN. AD is N2-concentrated; PD draws on both REM and N2. PST yields a richer N2 codebook than PSN (49 vs. 23 features). **(c)** Top five discriminative features per disease, ranked by |Cohen's d|, annotated with feature type and dominant sleep stage. Temporal consolidation (bout mean) dominates both signatures. All results use within-cohort LOSO cross-validation on frozen PSSmodels at $M = 65$.

Exploratory clinical probing with prototype-grams

If prototypes are physiologically meaningful, the per-subject prototype-gram should encode not only staging but also disease-related disruptions of sleep architecture. We tested this through exploratory clinical probing: summary statistics from the prototype-gram were used to classify clinical groups known to differ in sleep structure. This is an interpretability validation, not a biomarker discovery study; no disease information entered the staging model during training.

Two KU Leuven cohorts recorded under identical conditions were evaluated: the Alzheimer Sleep Dataset (ASD; 32 healthy controls vs. 37 individuals with Alzheimer's disease, $N = 69$) and the KU Leuven Parkinson dataset (KPD; 40 healthy older adults vs. 46 individuals with Parkinson's disease, $N = 86$). The KPD cohort is age- and sex-matched; the ASD cohort exhibits a significant age difference between groups ($p = 0.018$), which we acknowledge as a confound (Supplementary Section 7.2). Both PSS backbones, PSN (trained on MASS, 200 subjects) and PST (trained on SHHS, 5,793 subjects), with zero subject overlap with either clinical cohort, were applied frozen. For each subject, three features per prototype were extracted from the prototype-gram at codebook resolution $M = 65$: proportion (occupancy), mean bout duration (consolidation), and bout-duration standard deviation (fragmentation), yielding a 195-dimensional feature vector. A logistic regression classifier with ℓ_1/ℓ_2 regularisation and class balancing was evaluated in within-cohort leave-one-subject-out cross-validation for two binary tasks: AD diagnosis and PD diagnosis. The best-performing hyperparameter configuration per task was selected post-hoc from a grid of 16 configurations (2 penalties × 4 regularization strengths × 2 standardization options); accuracy varies across the grid (Supplementary Section 7). This exploratory screening is an analytical tool for hypothesis generation, not a model-selection procedure for clinical deployment.

Figure 5a shows that PST outperforms PSN on both tasks: AD diagnosis reaches 82.6% accuracy (PST) vs. 78.3% (PSN), both with all-stage features; PD diagnosis reaches 77.9% (PST, all-stage) vs. 68.6% (PSN, all-stage). For PSN, restricting to REM-only features improves PD accuracy to 73.3%, indicating that PSN's non-REM prototypes introduce noise rather than signal for PD classification. The butterfly plot (Figure 5b) reveals sharply distinct stage-specificity profiles: AD discrimination is concentrated in N2-dominant features (49 features with $|d| \geq 0.5$ for PST), whereas PD draws on a mixture of REM and N2 features. PST consistently produces a richer discriminative landscape than PSN, particularly in Stage N2 where it yields twice as many discriminative features (49 vs. 23). PSN partially compensates through stronger N3-associated features (13 vs. 4), but this stage carries less disease-specific signal. Figure 5c ranks the top five discriminative features per disease: all five AD features are N2-associated, led by bout-mean consolidation ($d = 1.56$), while the leading PD feature is an REM bout-mean ($d = 0.93$) followed by N2 and additional REM features, confirming the dual-stage signature. Temporal consolidation (bout mean) dominates both disease signatures.

The stage-specificity profiles align with known neurophysiology. The N2-concentrated AD signature is consistent with the thalamocortical disconnection and sleep-spindle disruption documented in AD [26]: because N2 prototypes tile a delta–sigma-reciprocity continuum (see Explanations), disease-driven spindle loss translates directly into altered N2 prototype occupancy and bout structure. The mixed REM–N2 profile of PD reflects joint brainstem pathology (disrupting REM regulation) and thalamocortical involvement (disrupting N2 consolidation) [27, 28]. The architectural comparison reveals that PST's richer N2 codebook, visible in the butterfly plot (Figure 5b), makes the disease signal more accessible

to a linear classifier; this advantage likely stems from the SleepTransformer's multi-head attention, which decomposes within-epoch spectral structure more finely than SeqSleepNet's recurrent architecture, combined with the larger and more diverse SHHS training corpus. Disease-discriminative information is present in both representations but the quality of the probing depends on the granularity of the learned codebook.

Full pipeline details, per-configuration results, univariate feature rankings by Cohen's d, and demographic comparisons for both cohorts are provided in Supplementary Section 7.

Discussion

The PSSframework adds interpretability to sequence-to-sequence sleep staging by learning a discrete codebook of physiological archetypes that can be reconstructed, characterized, and clinically probed. The central question is not whether prototypes improve staging accuracy, but what the prototype analysis reveals about how sleep staging models encode physiological knowledge. Both PSN and PST achieve competitive staging across 20 datasets, confirming that the prototype mechanism preserves representation quality while enabling dissection of the model's learned sleep vocabulary.

Other intrinsically explainable prototype-based models include ProtoPNet [29], which learns prototypical image patches and explains decisions by comparing input regions with these prototypes, and PIP-Net [30], which discovers prototypes self-supervisedly with spatial feature localization. While effective for image classification, these methods face fundamental limitations when applied to sleep staging. First, they treat prototypes as translation-invariant image objects, whereas time-frequency features are frequency-specific: a 12 Hz spindle-band pattern is qualitatively different from 2 Hz delta activity. Second, sleep staging often depends on a small number of distinctive markers, and learning many prototypes for such well-defined features introduces redundancy. The PSSframework overcomes these limitations by learning abstract prototypes in the epoch encoder's latent space. Critically, both PSN and PST independently discover the same core physiological motifs: delta-dominated N3, phasic/tonic REM differentiation, and an N2 delta–sigma reciprocity continuum, providing cross-backbone validation that these patterns reflect genuine signal properties rather than architectural artifacts.

Systematic coherence assessment of all 24 prototype codebook entries against AASM criteria reveals that 42% are fully coherent with standard scoring guidelines, 42% are partially coherent, and 17% are non-canonical. Partially coherent prototypes are arguably more informative: both backbones bypass canonical alpha activity for Wake classification, instead relying on elevated beta and gamma power. Similarly, N2 prototypes tile a delta–sigma reciprocity continuum revealing finer-grained sub-stage structure than the binary spindle/K-complex criteria suggest. Neither backbone allocates dedicated N1 prototypes, mirroring the clinical reality of N1 as the least reliable stage in inter-rater agreement. The four non-canonical prototypes (PSN P3, P5; PST P0, P6) reflect alternative classification strategies rather than physiological errors. The two Wake prototypes rely on beta/gamma power rather than alpha, consistent with eyes-open wakefulness where alpha is physiologically absent. The two N2 prototypes detect deep-N2 via delta dominance (PSN P5) or broadband high-frequency patterns (PST P0), capturing the spectral heterogeneity of N2 near the N3 boundary. These cases suggest that the model has learned valid sub-stage distinctions that extend beyond the criteria evaluated by our coherence checklist.

To assess whether the learned prototypes encode disease-relevant information, we performed exploratory clinical probing on two independent cohorts using frozen PSSmodels. PST achieves higher discrimination than PSN for both Alzheimer's disease (82.6% vs. 78.3% accuracy) and Parkinson's disease (77.9% vs. 68.6% with all-stage features; PSN reaches 73.3% when restricted to REM-only features). The gap is driven by PST's richer N2 codebook (49 vs. 23 discriminative features with $|d| \geq 0.5$). The stage-specificity of discriminative features differs between diseases in a pathophysiologically coherent way: AD produces an N2-concentrated signature consistent with spindle disruption, while PD yields a mixed REM–N2 signature reflecting joint brainstem and thalamocortical involvement. These results are exploratory and hypothesis-generating; the clinical cohorts are small and lack external validation (see Limitations).

A methodological note on codebook resolution: the interpretability showcase uses $M = 12$ prototypes, which are visually inspectable and map directly onto AASM stage categories, while clinical probing uses $M = 65$, which captures finer-grained variation relevant for disease discrimination. The discriminative prototypes at $M = 65$ are drawn from stage-specific clusters that map onto the same physiological motifs identified at $M = 12$ (delta-dominated N3, sigma-rich N2, phasic REM), but the higher resolution resolves within-stage heterogeneity that a 12-entry codebook absorbs. This creates a trade-off between interpretability and clinical utility: coarser codebooks are more inspectable but less sensitive to subtle disease signatures.

Examining the Parkinson's disease findings in greater detail, PD patients presented limited prototype-level alterations, consistent with the cross-sectional design and the mild-to-moderate disease severity (Hoehn and Yahr II–III) of the KPD cohort [31]. The restricted severity range likely prevented detection of slow-wave sleep differences, where higher slow-wave density has been associated with slower motor symptom progression after 4–5 years [32]. Additionally, participants did not present signs of cognitive impairment, a symptom widely associated with EEG alterations in PD [33].

Despite these constraints, differences with healthy controls emerged in N2 sub-stage consolidation: PD patients exhibited reduced bout duration across multiple N2-associated prototypes (Cohen's $d = 0.5$–0.9), with the affected prototypes tiling the sigma–delta reciprocity continuum identified in the prototype characterization analysis. This pattern is consistent with disrupted sleep spindle generation reported in PD [28] and its association with later development of

dementia [34], although the probing framework measures prototype-gram statistics rather than spectral power directly; validation against expert-annotated spindle events was not performed.

Finally, PD patients showed reduced and fragmented REM sleep in the REM-associated prototypes, in agreement with the established literature [27]. People with PD may also exhibit REM sleep without atonia or REM sleep behaviour disorder (RBD). Although RBD was not formally assessed in the KPD dataset, REM sleep without atonia may have contributed to the assignment of REM epochs to non-REM prototypes. This warrants caution in the interpretation of prototype features that may contain information about REM sleep in the PD population.

The prototype-based approach occupies a distinct niche in the interpretability landscape for sleep staging. Post-hoc attribution methods (integrated gradients, GradCAM) produce local, instance-level saliency maps that identify which input features drive a single prediction, but provide no global vocabulary of what the model has learned. Attention visualization in Transformer-based models (e.g., SleepTransformer) reveals which epochs or frequency bins receive high weight, but attention weights are not faithful explanations of the decision mechanism [35]. Rule-based expert systems offer global interpretability by design but are hand-crafted and cannot learn from data. The PSS framework combines global codebook-level interpretability (what prototypes exist and what they encode) with local instance-to-prototype attribution (why a specific epoch maps to a specific prototype), providing a bridge between the granularity of post-hoc methods and the global structure of rule-based systems.

Looking forward, unsupervised prototype learning, independent of staging labels, could reveal physiologically relevant patterns that supervised training suppresses. The prototype-gram representation would benefit from sequential descriptors that capture dynamic sleep disruption beyond the current summary statistics. Finally, blinded expert validation of prototype coherence remains the most direct way to establish clinical credibility.

Limitations

Because prototypes are learned from supervised staging labels, they are inherently tied to the scorer's labeling conventions and may over-represent patterns that distinguish scoring categories while under-representing physiologically relevant phenomena that do not map cleanly to stage boundaries. Prototype coherence and sub-stage patterns were assessed by the study authors against AASM criteria and have not been independently validated, either through blinded expert review or through direct comparison with expert-annotated micro-events such as sleep spindles, K-complexes, or arousals.

The clinical probing results should be interpreted with caution. All analyses are within-cohort leave-one-subject-out on small samples (ASD: $N = 69$; KPD: $N = 86$) and are intended as hypothesis-generating. The ASD cohort exhibits a significant age difference between healthy controls and Alzheimer's disease patients ($p = 0.018$); age was not included as a covariate in the logistic regression, and the N2-concentrated AD signature may partly reflect age-related sleep architecture changes rather than disease-specific effects. The KPD cohort is age- and sex-matched, providing a partial control; age-corrected AD probing is planned as future work.

Finally, boundary wake trimming removes contiguous Wake epochs before sleep onset and after final awakening; we cannot rule out that this removed clinically relevant disease signal such as sleep onset latency or early morning awakening. The discriminative features are dominated by N2 and REM statistics, but this may partly reflect the removal of Wake-related signal rather than its absence.

Methods

Eleven polysomnographic datasets spanning 12,317 unique subjects were used (Supplementary Section 1): nine benchmark cohorts (SHHS [5], MrOS [36], MESA [37], WSC [38], HomePAP [39], DCSM [40], MASS [41], 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 Section 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: $\hat{\mathbf{Y}} = \mathbf{ZW} + \mathbf{b}$.

We evaluate the PSSframework 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 [41] (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 remaining compatible with any epoch/sequence encoder pair and allowing quantization. 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 $\mathbf{h}_c = f_{\mathrm{EE}}(\mathbf{S}_{:,:,c,:,:}) \in \mathbb{R}^{L\times D}$ for $c = 1, \ldots, C$. 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, channels are stochastically replaced with a zero-input embedding $\mathbf{h}_0 = f_{\mathrm{EE}}(\mathbf{0})$. A per-channel auxiliary classifier produces accuracy estimates a_c; the drop probability for channel c is $p_c = (1-a_c)/\sum_j (1-a_j)$, so that higher-accuracy channels are occluded more frequently, forcing the model to learn from secondary modalities. At each training step, with probability $p_{\mathrm{apply}} = 0.5$, a random number $n \sim \mathrm{Uniform}(1, C-1)$ of channels is sampled according to $\{p_c\}$ and replaced with $\mathbf{h}_0$.

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

$$\mathbf{h}'_c = \mathbf{h}_c + f_{\mathrm{CM}}([\mathbf{h}_1, \ldots, \mathbf{h}_C])_c \tag{1}$$

Attention pooling aggregates across channels: $w_c = \mathrm{softmax}(\mathbf{v}^\top \mathbf{h}'_c)$, 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:

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

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: $\mathbf{z}_q = \bar{\mathbf{h}}_q + f_{\mathrm{SE}}(\bar{\mathbf{h}}_q)$. 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 Section 3).

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

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

$$\mathcal{L}_{\mathrm{epoch}} = \mathrm{CE}\big(\mathrm{clf}(\bar{\mathbf{h}}),\ \mathbf{y}\big) \tag{4}$$

$$\mathcal{L}_{\mathrm{main}} = \mathrm{CE}\big(\mathrm{clf}(\mathbf{z}),\ \mathbf{y}\big) \tag{5}$$

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{main}} + \mathcal{L}_{\mathrm{epoch}} + \mathcal{L}_{\mathrm{channel}} \tag{6}$$

$\mathcal{L}_{\mathrm{channel}}$ provides per-channel accuracy estimates used by the dropout mechanism; $\mathcal{L}_{\mathrm{epoch}}$ supervises the pre-sequence representations used for prototype learning; $\mathcal{L}_{\mathrm{main}}$ 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 $\mathbf{C} = [\mathbf{P}_1, \ldots, \mathbf{P}_M] \in \mathbb{R}^{M\times D}$ 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 at all tested sizes (Supplementary Section 8), and a seed stability analysis shows that the resulting prototypes are reproducible across independent training runs (Supplementary Section 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 $\tilde{\mathbf{S}}$ from random noise to minimize $\|f_{\mathrm{EE}}(\tilde{\mathbf{S}}) - \mathbf{P}_k\|_2^2$; 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 Section 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 $f_k(\bar{\mathbf{S}}) = -\|f_{\mathrm{EE}}(\bar{\mathbf{S}}) - \bar{\mathbf{P}}\|_2^2$. Local attribution is computed by accumulating gradients along the straight-line interpolation path from a population-average

baseline $\mathbf{S}_0$ to $\bar{\mathbf{S}}$:

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

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 $\bar{\mathbf{P}}$, with signed values indicating whether each feature increases or decreases prototype affinity. Example attributions are in Supplementary Section 4.

Prototype-Summary Generation

Each prototype is characterized through combinatorial feature-level ablation on its hybrid reconstruction. Thirteen features are defined: 8 EEG spectral bands (δ, θ, α, σ_l, σ_h, β_l, β_h, γ), 4 EOG bands (slow eye movements, rapid eye movements, blinks, residual), and EMG tone. For each feature i, importance is measured as the change in prototype affinity when the feature is replaced by its population-distribution center: $\text{imp}_i = \|f_{\text{EE}}(\tilde{\mathbf{S}}) - \mathbf{P}_k\|_2^2 - \|f_{\text{EE}}(\tilde{\mathbf{S}}_{\setminus i}) - \mathbf{P}_k\|_2^2$. Sensitivity normalizes importance by spectral deviation: $\text{sens}_i = \text{imp}_i \,/\, \max(|\Delta\text{dB}_i|,\ 3.0)$, where the soft floor prevents pathological values for near-baseline features.

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 (sustained vs. transient activation) and the peripheral-channel context (EOG activity pattern, EMG tone level). Sentence S4 evaluates physiological coherence against stage-specific AASM criteria [45] and assigns a verdict.

The coherence assessment applies a stage-specific checklist: Wake requires elevated α and tonic EMG; N2 requires σ-band relevance (>15% combined ablation importance); N3 requires δ dominance (>30% importance) and EMG atonia; REM requires EOG dominance (>30% channel importance) and EMG atonia. Each prototype is classified as *coherent* (all criteria satisfied), *partially coherent* ($\geq$50%), or *non-canonical* (<50%). Full prototype summaries, coherence results, and per-prototype spectral cards for both backbones are in Supplementary Section 6.

Evaluation Protocol

All staging metrics are computed per subject and reported as mean $\pm$ 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 for K pairwise comparisons per dataset as the significance test for all paired model comparisons, rather than bootstrap confidence intervals. 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 Section 1). Ablation experiments use fixed train/test splits: 34 held-out subjects from MASS (PSN), 1,636 from SHHS (PST), with five occlusion scenarios (Supplementary Sections 2–3). For clinical probing, PSSmodels trained on benchmark datasets were applied frozen to KPD and ASD cohorts (no subject overlap). Prototype-gram features (proportion, bout mean, bout std per prototype; 195-dimensional vector at $M = 65$) were classified via logistic regression with ℓ_1/ℓ_2 penalty and within-cohort LOSO CV for two tasks: AD diagnosis and PD diagnosis. Full pipeline in Supplementary Section 7.

Ethics Approval and Consent to Participate

This study makes use of publicly available datasets that have received prior ethical approval and informed consent from the respective institutions. Data from the KU Leuven PD dataset were collected as part of the screening of a project that had previously received ethical approval (S61792). The participants indicated their approval for re-use of the data in the respective ICF and the retrospective analyses performed in this manuscript have been subsequently approved by the ethics committee UZ/KU Leuven (S70708).

Data Availability

Of the eleven datasets, five are available on the NSRR [46] (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 PSSframework, including ProtoSleepNet and ProtoSleepTransformer, are available as part of the PhysioEx library (v2.0.0) [47]. 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.

Authors Contribution

GG conceived and conducted the investigation of the study with and under the supervision of MDV, MGCAC and ALA. LM and MG acquired the KU Leuven Parkinson's dataset and LM validated the physiological results 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 participated in reviewing the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgments

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 HHSN2682015000003I, 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).

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)

[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)

[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)

[4] Berry, R.B., Abreu, A.R., Krishnan, V., Quan, S.F., Strollo Jr, P.J., Malhotra, R.K.: A transition to the american academy of sleep medicine–recommended hypopnea definition in adults: initiatives of the hypopnea scoring rule task force. Journal of clinical sleep medicine **18**(5), 1419–1425 (2022)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[13] Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems **41**(3), 647–665 (2014)

[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)

[15] Achtibat, 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)

[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)

[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)

[19] Arik, S.O., Pfister, T.: Protoattend: Attention-based prototypical learning. Journal of Machine Learning Research **21**(210), 1–35 (2020)

[20] Dreyer, M., Achtibat, 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). IEEE

[21] 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)

[22] 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. 13699–13708 (2022)

[23] Kontras, K., et al.: NeuroAtlas: benchmarking foundation models for clinical EEG and brain–computer interfaces. arXiv preprint arXiv:2605.14698 (2026) arXiv:2605.14698 [cs.LG]

[24] 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

[25] 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

[26] Christensen, J.A., Nikolic, M., Warby, S.C., Koch, H., Zoetmulder, M., Frandsen, R., Moghadam, K.K., Sorensen, H.B., Mignot, E., Jennum, P.J.: Sleep spindle alterations in patients with parkinson's disease. Frontiers in human neuroscience **9**, 233 (2015)

[27] Zhang, Y., Ren, R., Sanford, L.D., Yang, L., Zhou, J., Tan, L., Li, T., Zhang, J., Wing, Y.-K., Shi, J., *et al.*: Sleep in parkinson's disease: a systematic review and meta-analysis of polysomnographic findings. Sleep medicine reviews **51**, 101281 (2020)

[28] Villamar-Flores, C.I., Rodríguez-Violante, M., Abundes-Corona, A., Alatriste-Booth, V., Valencia-Flores, M., Rodríguez-Agudelo, Y., Cervantes-Arriaga, A., Solís-Vivanco, R.: Association between alterations in sleep spindles and cognitive decline in persons with parkinson's disease. Neuroscience letters **842**, 138006 (2024)

[29] Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence **3**(12), 1061–1070 (2021)

[30] 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)

[31] Hoehn, M.M., Yahr, M.D.: Parkinsonism: onset, progression, and mortality. Neurology **17**(5), 427–427 (1967)

[32] 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)

[33] 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)

[34] Latreille, V., Carrier, J., Lafortune, M., Postuma, R.B., Bertrand, J.-A., Panisset, M., Chouinard, S., Gagnon, J.-F.: Sleep spindles in parkinson's disease may predict the development of dementia. Neurobiology of aging **36**(2), 1083–1090 (2015)

[35] Jain, S., Wallace, B.C.: Attention is not Explanation. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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 . https://aclanthology.org/N19-1357/

[36] 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)

[37] 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)

[38] 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)

[39] Rosen, C.L., Auckley, D., Bender, R., Bolden, N., Chervin, R.D., Johnson, N.L., Malhotra, A., Mancuso, J.J., Martin, R.J., Mehra, R., et al.: Home positive airway pressure. National Sleep Research Resource (2022). Version 0.2.0

[40] 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)

[41] 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)

[42] Alvarez-Estevez, D., Rijsman, R.: Haaglanden Medisch Centrum sleep staging database. PhysioNet (2022) https://doi.org/10.13026/t79q-fr32 . Version 1.1

[43] Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A.C., Oberye, 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., Vandenberghe, 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] Berry, R.B.: The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. version 2.1. Darien Illinois: American Academy of Sleep Medicine (2014)

[46] 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)

[47] 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)