Trajectory- and Topology-Aware Multimodal Modeling for Pediatric Sleep Event Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sleep disorders in children are common yet often underdiagnosed, and manual scoring of overnight polysomnography (PSG) is slow while labels for key events are sparse. We study 30-second pediatric PSG epochs represented by fixed embeddings from a multimodal masked-autoencoder. We investigate and augment these embeddings with (i) PHATE-derived per-epoch coordinates and whole-night movement descriptors, (ii) persistent-homology summaries computed on the high-dimensional embedding cloud, and (iii) routine EHR context. An AHI-stratified screen shows clinically coherent shifts in movement/topology. In predictive benchmarks, a late-fusion MLP that integrates all branches improves rare-event detection over a linear probe, leading in 3/4 binary tasks (Desaturation AUPRC = 0.370, EEG arousal = 0.484, Hypopnea = 0.290), while Apnea favors the EHR-only late-fusion variant (AUPRC = 0.147). Results suggest that clinical context and latent geometry/topology provide complementary signals beyond the generative embeddings, yielding interpretable links to disease burden and better performance under extreme imbalance.

1 Introduction

Pediatric sleep disorders affect cognition, behavior, and cardiometabolic health, yet real-world diagnosis is constrained by manual PSG scoring and highly imbalanced event labels [1–3]. We start from per-epoch PedSleepMAE [4] embeddings—fixed, multimodal representations learned generatively from raw PSG channels via masked-autoencoder [5]—and ask whether their (a) latent trajectory information, (b) topological shape, and (c) augmentation with EHR can (i) reflect disease burden across AHI strata and (ii) improve detection of apnea, hypopnea, desaturation, EEG arousal, and five-stage sleep under session-wise splits.

To motivate, we mapped per-epoch PedSleepMAE embeddings to 2-D PHATE [6] (Fig. 1). PedSleepMAE was trained by treating every 30 seconds of PSG as an independent sample, i.e. it was reconstructing signals without knowing what time of the night it is or who it is from. Yet, Fig. 1 reveals that the embeddings captured time-dependent information despite not knowing it explicitly in training. PHATE maps each night to a smooth, time-ordered path whose geometry matches expert stages: lighter stages at the entrance, N3 near the center, and REM along peripheral arcs. Across the session we observe consistent curvature, drift, and occasional bifurcations that align with canonical sleep progressions. This motivates our novel research question of investigating the *session-wide diagnostic information* contained in the *sequences of multimodal generative embeddings*.

Manifold learning is widely used to visualize high-dimensional trajectories [7]; PHATE's diffusion geometry preserves local neighborhoods while maintaining global progression and denoises noisy biological measurements, making it suitable for sleep dynamics [8]. Prior PSG work more often models raw or time–frequency inputs with sequence architectures (e.g., SleepTransformer) [9].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

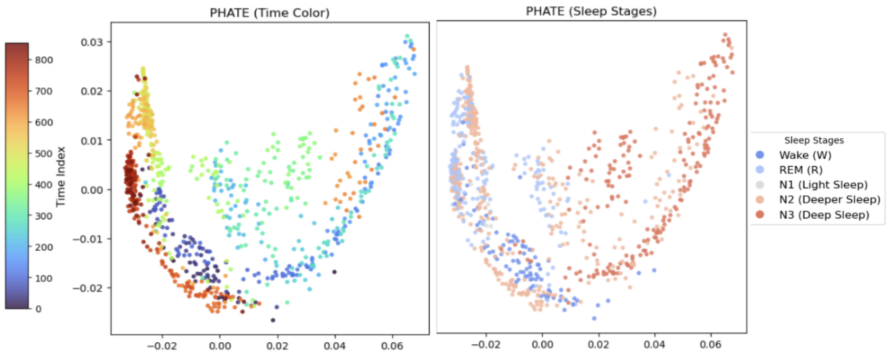

Figure 1: Parallel PHATE views for one study: left colored by epoch index (time), right by sleep stages. The 2-D diffusion map reveals a smooth, time-ordered trajectory whose regions align with expert staging.

In parallel, topological data analysis (TDA) offers stable vectorizations that capture multiscale loop/cluster structure for learning [10–12]; pediatric sleep EEGs have related such structure to respiratory burden and desaturation [13]. Finally, combining signal representations with structured EHR via late fusion is a common and effective pattern in clinical prediction [14, 15].

Ablations are ordered for interpretability and deployment: linear probe on embeddings (M0), add routine EHR via late-fusion MLP (M1), add PHATE point+time branches (M2), then add topological descriptors (M3). This isolates the incremental value of context, trajectory, and topology beyond the generative embeddings.

2 Methods

2.1 Data

We use the Nationwide Children's Hospital Sleep DataBank (NCHSDB), which contains pediatric overnight polysomnography (PSG) with technologist labels for sleep stages and respiratory events [2]. The analysis set includes 2,522 complete studies, each identified by a (person ID, session ID) pair. Signals include EEG, ECG, EMG, respiratory effort, airflow, and oxygen saturation. Recordings are divided into consecutive 30-second epochs in temporal order. Each epoch is represented by a 7,680-dimensional PedSleepMAE embedding (120×64) learned generatively from raw PSG channels [4]. Labels for sleep stage, apnea, hypopnea, desaturation, and EEG arousal align one-to-one with the embeddings. We use *session-wise*, stratified splits (70/10/20% train/val/test) per label.

Structured EHR from NCHSDB [2] is linked to each session. Routine EHR provides low-overhead clinical context that can reduce confounding and improve generalization when fused with signal features. We include a demographic and comorbidity set to our analysis; see Appendix A for the list.

2.2 Feature Sets

Our features mirror the ablation order: per-epoch PedSleepMAE embeddings as baselines, EHR (Sec. 2.1), then (i) PHATE-based trajectory features and (ii) topological descriptors. See Appendix A for formal definitions.

PHATE trajectory features. PHATE is fit on training sessions and applied out of sample to validation/test. We use (a) *trajectory-local* per-epoch coordinates/derivatives and (b) *trajectory-global* session summaries of movement/fragmentation: mean and max inter-epoch step, mean turning angle, directional entropy of turns, tortuosity (path-length vs. end-to-end), and a change-point count on the step-length series using RUPTURES with PELT (Pruned Exact Linear Time) [16, 17]. Session-level quantities are broadcast to all epochs of that session.

Topological features. To quantify shape directly in representation space, we compute persistent homology on the original 7,680-D PedSleepMAE point cloud via a Vietoris–Rips filtration and summarize with a compact six-statistic panel: `H0_sum_pers`, `H0_n_bars`, `H1_n_bars`, `H1_max_pers`,

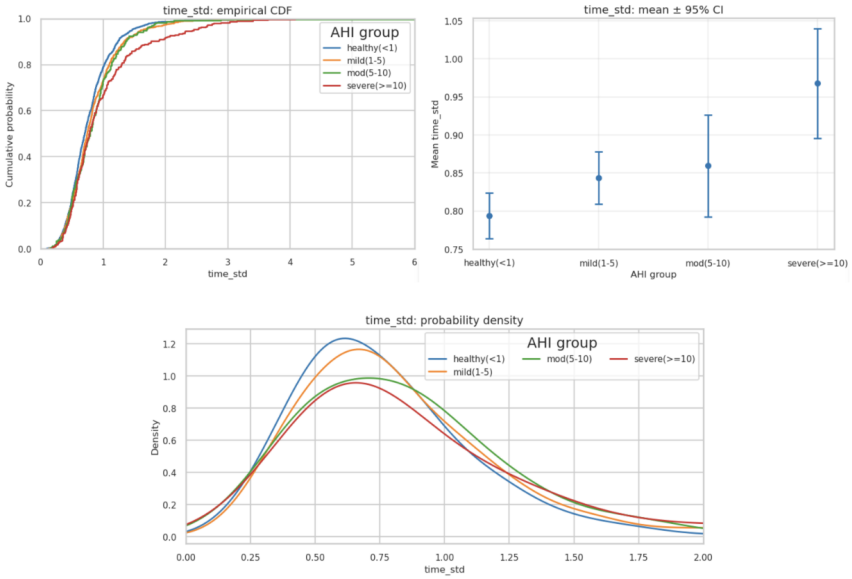

Figure 2: AHI associations for a representative movement metric (`time_std`). Top-left: ECDF; top-right: mean±95% CI; bottom: KDE density. Groups: healthy (<1), mild (1–5), moderate (5–10), severe (≥10).

Betti–1 L^2 norm, and the H_1/H_0 lifetime ratio. These capture cluster spread/fragmentation (H_0), loop prevalence/strength (H_1/Betti–1 energy), and loop-vs-cluster balance, producing stable, fixed-length vectors for learning [10–12].

2.3 AHI-stratified feature analysis (pre-specification)

We used an AHI-stratified screen to decide which session-level descriptors advance to modeling. Sessions were grouped by pediatric AHI thresholds into healthy (<1), mild (1–5), moderate (5–10), and severe (≥10), following commonly used pediatric criteria [3, 18]. For each session-level candidate we ran a Kruskal–Wallis omnibus test [19], Dunn post-hoc comparisons [20] with Holm correction [21], reported Cliff's δ as an effect size [22], and visualized box/ECDF/KDE with adjusted q values. Because AHI is defined per session, the screen applied only to trajectory-global PHATE features and to TDA summaries. Trajectory-local features are per-epoch and do not align to a session label; they were not screened. EHR features were pre-specified and likewise not AHI-screened to avoid label leakage and to preserve a stable confounder block across all tasks.

2.4 Diagnostic models

We compare four epoch-level classifiers on identical *session-wise*, stratified splits and a shared training recipe. M0 (Linear Probe) applies a single linear layer to each 7,680-D PedSleepMAE embedding to set a lower bound on representation quality [4]. M1 (Emb+EHR, late-fusion MLP) replaces the linear head with a two-branch MLP: embeddings and EHR are encoded separately and concatenated. M2 (Emb+EHR+Trajectory, late-fusion MLP) keeps capacity matched to M1 and adds two PHATE trajectory branches: (i) per-epoch point features and (ii) per-session global summaries, to expose local state and whole-night structure to the classifier [6, 8, 16]. M3 (Emb+EHR+Trajectory+TDA, late-fusion MLP) further adds a session-level topological branch built from persistent-homology statistics, allowing loop/cluster structure to inform decisions [10–12]. All architectural/optimization details are specified in Appendix B.

3 Results

3.1 Clinical association with AHI

Trajectory movement and topology co-vary with AHI. Permutation omnibus tests are significant for all six TDA descriptors and several PHATE movement metrics, showing monotone shifts

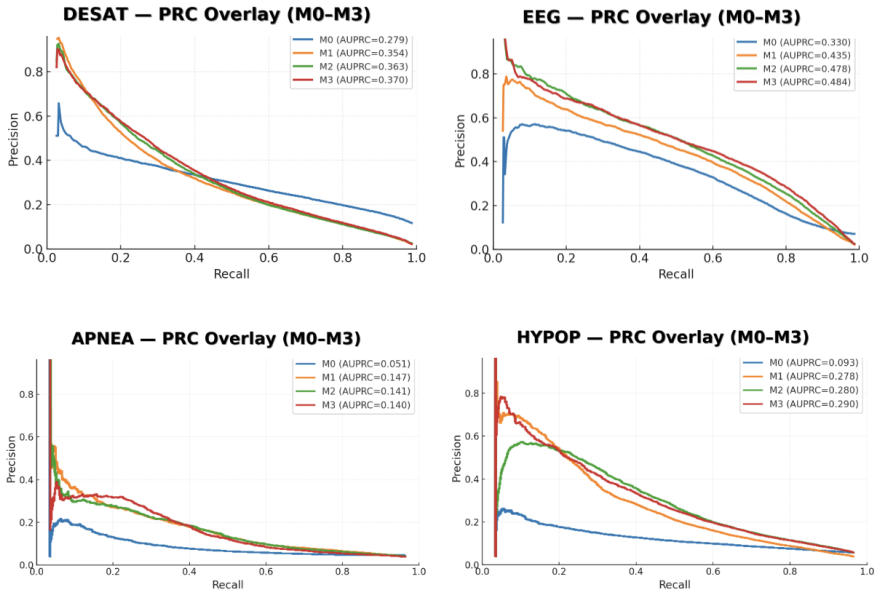

Figure 3: PR curves for the four binary tasks (test set), M0–M3 overlaid; legend reports AUPRC for each model.

from healthy→mild→moderate→severe. Severe nights exhibit reduced topological richness (fewer components/loops, lower Betti energy), larger average/variable steps on the manifold, and higher `H1_max_pers`—fewer but more persistent loops. Healthy nights show the opposite pattern. We therefore retain the six PHATE-global and six TDA summaries for prediction. Pairwise "3-vs-all" and "0-vs-all" contrasts are strongly significant (Appendix C). See Fig. 2 for an example.

3.2 Predictive performance (Models 0–3)

We report AUPRC as the primary metric for imbalanced tasks [23]. Fig. 3 shows clear separation from the linear probe (M0) once contextual branches are added. The full late-fusion model (M3) is best on three of four labels— Desaturation 0.370, EEG arousal 0.484, Hypopnea 0.290—while Apnea favors the EHR-only late-fusion model (M1; 0.147 vs. 0.141–0.140 for M2–M3). The M0→M1 jump reflects both capacity (linear→MLP) and genuine value from EHR, which is particularly helpful for the rarest outcomes. Adding PHATE trajectory features (M2) yields further gains on Desaturation, EEG, and Hypopnea—consistent with Sec. 3.1 where directional entropy and step statistics tracked AHI—and is competitive on secondary metrics (e.g., top accuracy for Desaturation 0.8877 and Hypopnea 0.9730; top ROC-AUC for Hypopnea 0.8963; Appendix D). Adding topology (M3) provides small, label-dependent lifts, most notable for Hypopnea, suggesting complementary loop-geometry signals. For Apnea, EHR and local waveform cues dominate at very low recall; trajectory/TDA add little and can trade precision for recall.

4 Conclusion

We investigated diagnostic information contained in the *sequences* of per-epoch PedSleepMAE embeddings, and presented a late-fusion pipeline that augments the embeddings with PHATE-based temporal descriptors, topological summaries of the latent trajectory, and EHR context. On 2.5k+ pediatric sleep studies, these time/shape features showed AHI-stratified shifts and improved rare-event detection beyond a linear probe. The full model (M3) led in three of four tasks—Desaturation AUPRC 0.370, EEG arousal 0.484, Hypopnea 0.290—while Apnea favored the EHR-only variant (M1; AUPRC 0.147). These results highlight that latent geometry, topology, and clinical context provide meaningful signals that capture disease burden and reduce reliance on manual scoring. Future work will refine an embeddings-only MLP for clearer baselines and explore end-to-end integration of manifold/topological structure to ensure robustness under severe imbalance.

References

[1] T. F. Anders and L. A. Eiben, "Pediatric sleep disorders: A review of the past 10 years," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 36, no. 1, pp. 9–20, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0890856709636947

[2] H. Lee, B. Li, S. DeForte, M. L. Splaingard, Y. Huang, Y. Chi, and S. L. Linwood, "A large collection of real-world pediatric sleep studies," *Scientific Data*, vol. 9, no. 1, p. 421, jul 2022. [Online]. Available: https://doi.org/10.1038/s41597-022-01545-6

[3] A. A. of Sleep Medicine *et al.*, "Aasm manual for the scoring of sleep and associated events american academy of sleep medicine," *Darien, IL.[Google Scholar]*, 2007.

[4] S. R. Pandey, A. Saeed, and H. Lee, "Pedsleepmae: Generative model for multimodal pediatric sleep signals," 2024. [Online]. Available: https://arxiv.org/abs/2411.00718

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[6] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, "Visualizing structure and transitions in high-dimensional biological data," *Nature Biotechnology*, vol. 37, no. 12, pp. 1482–1492, dec 2019. [Online]. Available: https://doi.org/10.1038/s41587-019-0336-3

[7] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, jan 2019. [Online]. Available: https://doi.org/10.1038/nbt.4314

[8] M. Kuchroo, J. Huang, P. Wong, J.-C. Grenier, D. Shung, A. Tong, C. Lucas, J. Klein, D. Burkhardt, S. Gigante *et al.*, "Multiscale phate exploration of sars-cov-2 data reveals multimodal signatures of disease," *BioRxiv*, pp. 2020–11, 2020.

[9] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[10] P. Bubenik *et al.*, "Statistical topological data analysis using persistence landscapes." *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 77–102, 2015.

[11] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, "Persistence images: A stable vector representation of persistent homology," *Journal of Machine Learning Research*, vol. 18, no. 8, pp. 1–35, 2017.

[12] N. Atienza, R. González-Díaz, and M. Soriano-Trigueros, "On the stability of persistent entropy and new summary functions for tda," *arXiv preprint arXiv:1803.08304*, 2018.

[13] A. Sathyanarayana, S. Manjunath, and J. A. Perea, "Topological data analysis based characteristics of electroencephalogram signals in children with sleep apnea," *Journal of sleep research*, p. e70017, 2025.

[14] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 06 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy068

[15] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, "Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection," *Scientific reports*, vol. 10, no. 1, p. 22147, 2020.

[16] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020.

[17] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[18] C. L. Marcus, L. J. Brooks, S. D. Ward, K. A. Draper, D. Gozal, A. C. Halbower, J. Jones, C. Lehmann, M. S. Schechter, S. Sheldon *et al.*, "Diagnosis and management of childhood obstructive sleep apnea syndrome," *Pediatrics*, vol. 130, no. 3, pp. e714–e755, 2012.

[19] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[20] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[21] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[22] N. Cliff, *Ordinal methods for behavioral data analysis*. Psychology Press, 2014.

[23] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[24] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, and K. Hess, "giotto-tda: A topological data analysis toolkit for machine learning and data exploration," 2020.

A Feature definitions

A.1 Branch inventory

Table 1: Hand-crafted branches (input dims per scope) used in late-fusion. Session-level vectors are broadcast to epochs.

Branch	Scope	Dim	Contents
EHR–Demographics	per-session	11	Age (z), gender (3), race (6), ethnicity (1).
EHR–Comorbidities	per-session	12	Asthma, obesity, diabetes, hypertension, depression/mood, anxiety, ADHD, seizure disorder/epilepsy, GERD, cerebral palsy, autism, developmental delay. (Sleep apnea excluded to avoid label leakage.)
Trajectory-local (PHATE)	per-epoch	6	`delta_dist`, `cum_dist`, `turn`, `curv`, `dist_start`, `segment_id` (PELT).
Trajectory-global (PHATE)	per-session	6	mean/max step, mean turn, `dir_entropy` (20-bin), `tortuosity`, `#segments`.
TDA (embedding cloud)	per-session	6	`H0_sum_pers`, `H0_n_bars`, `H1_n_bars`, `H1_max_pers`, `Betti-L`^{2}, `ratio_sum_H1_H0`.

A.2 PHATE trajectory quantities

Let $p_t \in \mathbb{R}^2$ be the PHATE coordinates at epoch t.

$$
\begin{aligned}
\texttt{delta\_dist}_t &= \|p_t - p_{t-1}\|_2 \\
\texttt{cum\_dist}_t &= \sum_{i=2}^{t} \texttt{delta\_dist}_i \\
\theta_t &= \operatorname{atan2}(p_t^y - p_{t-1}^y,\, p_t^x - p_{t-1}^x) \\
\texttt{turn}_t &= \operatorname{wrap}(\theta_t - \theta_{t-1}) \\
\texttt{curv}_t &= \frac{|\texttt{turn}_t|}{\texttt{delta\_dist}_t + \varepsilon} \\
\texttt{dist\_start}_t &= \|p_t - p_1\|_2 \\
\texttt{dir\_entropy} &= -\sum_{b=1}^{20} \hat{p}_b \log(\hat{p}_b) \quad \text{(20-bin histogram of } \texttt{turn}_t\text{)} \\
\texttt{tortuosity} &= \frac{\sum_t \texttt{delta\_dist}_t}{\|p_T - p_1\|_2 + \varepsilon} \\
\texttt{n\_segments} &= \#\{\text{PELT change points on } \texttt{delta\_dist}_t\}
\end{aligned}
$$

A.3 Topological descriptors

Let $\mathcal{X} = \{x_i\}$ be the 7,680-D embedding cloud for a session; we compute Vietoris–Rips persistence with H_0 and H_1 barcodes having lifetimes $\{\ell_j^{(0)}\}$ and $\{\ell_k^{(1)}\}$ using the `giotto-tda` library [24].

In code we extracted a wide panel of persistence-derived statistics, including lifetime sums, maxima, entropy-based measures, midlife and birth/death summaries, Betti curve energies, and persistence image ratios. For stability and interpretability in the late-fusion model, we retained six robust statistics as the TDA branch:

$$\texttt{H0\_sum\_pers} = \sum_j \ell_j^{(0)} \qquad \texttt{H0\_n\_bars} = \#\{\ell_j^{(0)} > 0\}$$

$$\texttt{H1\_n\_bars} = \#\{\ell_k^{(1)} > 0\} \qquad \texttt{H1\_max\_pers} = \max_k \ell_k^{(1)}$$

$$\texttt{Betti-}L^2 = \left\|\beta_1(r)\right\|_2 \quad \text{(Betti-1 curve } L^2 \text{ norm)}$$

$$\texttt{ratio\_sum\_H1\_H0} = \frac{\sum_k \ell_k^{(1)}}{\sum_j \ell_j^{(0)} + \varepsilon}$$

B Model specifics

Training protocol. All models were trained under the same protocol. We used the AdamW optimizer (learning rate 10^{-3}, weight decay 10^{-5}), batch size 256, and automatic mixed precision. A ReduceLROnPlateau scheduler (factor 0.5, patience 3) controlled learning rate decay, and training stopped early if validation performance did not improve for 8 epochs. Splits were stratified at the session level (70/10/20 for train/validation/test, seed $= 42$). Binary decision thresholds were chosen by maximizing F1-score on the validation set, while multiclass tasks (sleep staging) reported macro–F1.

Normalization. All branches were normalized using train-only mean and standard deviation. For numerical stability, features were clipped to the range $[-8, 8]$. Time-series inputs (embeddings and PHATE-point features) were standardized over all epochs, while session-level vectors (EHR, PHATE-time, and TDA) were standardized across sessions and then broadcast to all epochs. Non-finite values were replaced with zeros before standardization.

Loss functions and imbalance handling. To account for severe class imbalance, we applied class-weighted losses with weights $w_k \propto 1/\text{freq}_k$. Binary tasks used focal cross-entropy with focusing parameter $\gamma = 1.5$, while multiclass sleep staging used weighted cross-entropy.

Branch encoders. Each modality was encoded separately by a shallow MLP block. An encoder consisted of a linear layer mapping the raw input dimension to 128 units, followed by layer normalization and a ReLU activation:

$$z^{(k)} = \mathrm{ReLU}\Big(\mathrm{LN}\big(W^{(k)}x^{(k)} + b^{(k)}\big)\Big), \qquad z^{(k)} \in \mathbb{R}^{128}.$$

The five input modalities used in the full model (M3) were: per-epoch embeddings (7680-D), session-level EHR features (23-D), per-epoch PHATE-point features (6-D), session-level PHATE-time features (6-D), and session-level TDA features (6-D). Each branch produced its own 128-dimensional latent representation.

Fusion and classifier head. The encoded features were concatenated into a single latent vector. In M3, this produced a fused representation of size 640 (five times 128). This vector was then passed through a two-layer classifier head: a linear transformation to 256 units, ReLU activation, and dropout with probability 0.30, followed by a final linear layer mapping to logits. In summary, the head contained two linear layers with one hidden nonlinearity, while each branch contributed an additional encoder block upstream.

Model variants. The linear probe baseline (M0) consisted only of a direct linear mapping from embeddings (7680-D) to logits, without an encoder. M1 combined embeddings and EHR, producing a 256-dimensional fused vector before classification. M2 added both PHATE-time and PHATE-point, producing a 512-dimensional fused vector. M3 incorporated all five branches, producing a 640-dimensional vector before the classifier. Removing branches yielded the simpler models without changing the classifier head.

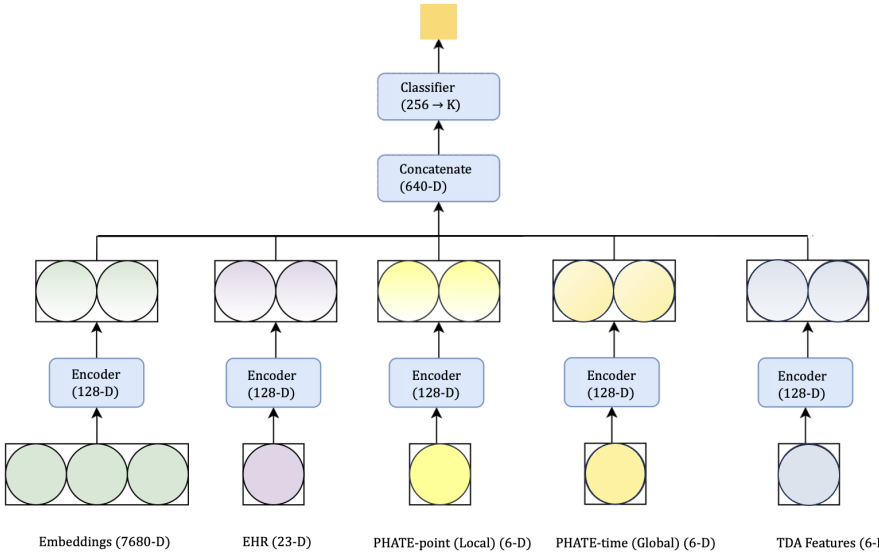

Figure 4: Late-fusion MLP (M3). Each branch input is encoded with Linear→LayerNorm→ReLU (128-D). Latents are concatenated (5×128=640-D) and passed through a classifier head: Linear 640→256, ReLU, Dropout(0.30), and Linear 256→K. Per-epoch branches are Embeddings and PHATE-point; session-level branches are EHR, PHATE-time, and TDA (broadcast across epochs).

C AHI Tests

Table 2: Permutation Kruskal–Wallis omnibus tests across AHI groups for candidate session-level descriptors. Larger H (with small q after Holm correction) indicates stronger distributional differences across AHI strata.

Feature	H	q
H0_n_bars	54.0954	0.00237
H1_n_bars	23.2657	0.00237
H1_max_pers	13.1744	0.00237
betti_L1	29.7612	0.00237
betti_L2	29.0318	0.00237
ratio_sum_H1_H0	21.9597	0.00237
time_std	19.2169	0.00237
time_l2	17.0812	0.00237
time_mean	16.1793	0.00633
PI_H1_long_short_ratio	13.4909	0.00949
H0_sum_pers	12.2354	0.00949
PI_H1_long	13.0261	0.00949
BettiH1_peak_loc_norm	8.7578	0.04526

Table 3: Permutation Mann–Whitney contrasts for AHI extremes. Effect is median(group)−median(others); δ is Cliff's delta. "Direction" summarizes whether the feature tends to be higher or lower in the target group.

Feature	Effect (Δ)	δ	q	Direction in severe
Severe (group 3) vs all others				
H0_n_bars	-16.5	-0.227	0.00271	lower
H1_n_bars	-20.5	-0.168	0.00271	lower
betti_L2	-23.66	-0.166	0.00271	lower
betti_L1	-170.0	-0.166	0.00271	lower
ratio_sum_H1_H0	-0.00414	-0.153	0.00271	lower
PI_H1_long_short_ratio	-0.00920	-0.109	0.00271	lower
time_mean	$+0.104$	$+0.137$	0.00271	higher
PI_H1_long	-0.3119	-0.112	0.00844	lower
ratio_count_H1_H0	-0.02274	-0.095	0.00844	lower
H1_max_pers	$+0.1968$	$+0.108$	0.01035	higher
time_l2	$+0.0885$	$+0.103$	0.01035	higher
time_std	$+0.0503$	$+0.109$	0.01582	higher
H0_sum_pers	$+33.51$	$+0.087$	0.02482	higher
Healthy (group 0) vs all others				
H0_n_bars	$+8.0$	$+0.107$	0.00475	higher
betti_L1	$+84.0$	$+0.089$	0.00475	higher
betti_L2	$+10.64$	$+0.089$	0.00475	higher
time_l2	-0.1407	-0.085	0.00475	lower
time_std	-0.0540	-0.092	0.00542	lower
H0_sum_pers	-22.25	-0.0697	0.00542	lower
ratio_sum_H1_H0	$+0.00250$	$+0.0763$	0.00542	higher
H1_max_pers	-0.0669	-0.0645	0.01661	lower
PI_H1_long	$+0.24295$	$+0.0611$	0.01898	higher
H1_n_bars	$+5.0$	$+0.0594$	0.01898	higher
time_mean	-0.0180	-0.0581	0.02416	lower
PI_H1_long_short_ratio	$+0.00419$	$+0.0542$	0.03638	higher

D Full Test-Set Metrics

Label prevalences

Table 4: Class distributions by split. Sleep staging shows % for {W, REM, N1, N2, N3}; binary tasks list positive prevalence (%). Use these to contextualize AUPRC baselines.

Task	Train	Val	Test
Sleep staging	[17.77, 3.55, 39.01, 22.48, 17.19]	[18.51, 3.70, 39.30, 22.09, 16.40]	[17.60, 3.72, 38.84, 22.97, 16.87]
Desaturation (+)	8.738	8.843	9.419
EEG arousal (+)	4.676	4.741	4.746
Apnea (+)	0.846	0.622	0.808
Hypopnea (+)	1.969	1.797	2.054

Sleep staging (5-class)

Table 5: Sleep staging metrics (test set). Macro-F1 is the primary multi-class metric; higher is better.

Model	Accuracy	F1 (macro)
M0 Linear probe	0.6860	0.6594
M1 MLP+EHR (late)	0.6916	0.6657
M2 +Trajectory	0.7047	0.6779
M3 +TDA	0.7076	0.6800

Binary tasks

Table 6: Desaturation (test set). AUPRC is the primary metric under imbalance; ROC-AUC and F1 are provided for completeness.

Model	Accuracy	F1	ROC-AUC	AUPRC
M0 Linear probe	0.8260	0.3520	0.7797	0.2793
M1 MLP+EHR (late)	0.8663	0.3760	0.7887	0.3537
M2 +Trajectory	0.8877	0.3799	0.7907	0.3626
M3 +TDA	0.8760	0.3923	0.7974	0.3700

Table 7: EEG arousal (test set). Late-fusion models progressively improve AUPRC and ROC-AUC over the linear probe.

Model	Accuracy	F1	ROC-AUC	AUPRC
M0 Linear probe	0.9384	0.4186	0.8507	0.3300
M1 MLP+EHR (late)	0.9472	0.4795	0.9054	0.4349
M2 +Trajectory	0.9477	0.5024	0.9191	0.4776
M3 +TDA	0.9486	0.5184	0.9285	0.4836

Table 8: Apnea (test set). The EHR-only late-fusion variant (M1) attains the highest AUPRC, consistent with the main text.

Model	Accuracy	F1	ROC-AUC	AUPRC
M0 Linear probe	0.9842	0.1255	0.7164	0.0506
M1 MLP+EHR (late)	0.9868	0.2361	0.8998	0.1472
M2 +Trajectory	0.9886	0.2282	0.8974	0.1409
M3 +TDA	0.9896	0.2479	0.8725	0.1397

Table 9: Hypopnea (test set). Adding trajectory (M2) and topology (M3) yields gains in AUPRC over M1, with small trade-offs on secondary metrics.

Model	Accuracy	F1	ROC-AUC	AUPRC
M0 Linear probe	0.9425	0.1681	0.7843	0.0931
M1 MLP+EHR (late)	0.9686	0.3269	0.8813	0.2782
M2 +Trajectory	0.9730	0.3676	0.8963	0.2798
M3 +TDA	0.9729	0.3524	0.8934	0.2901