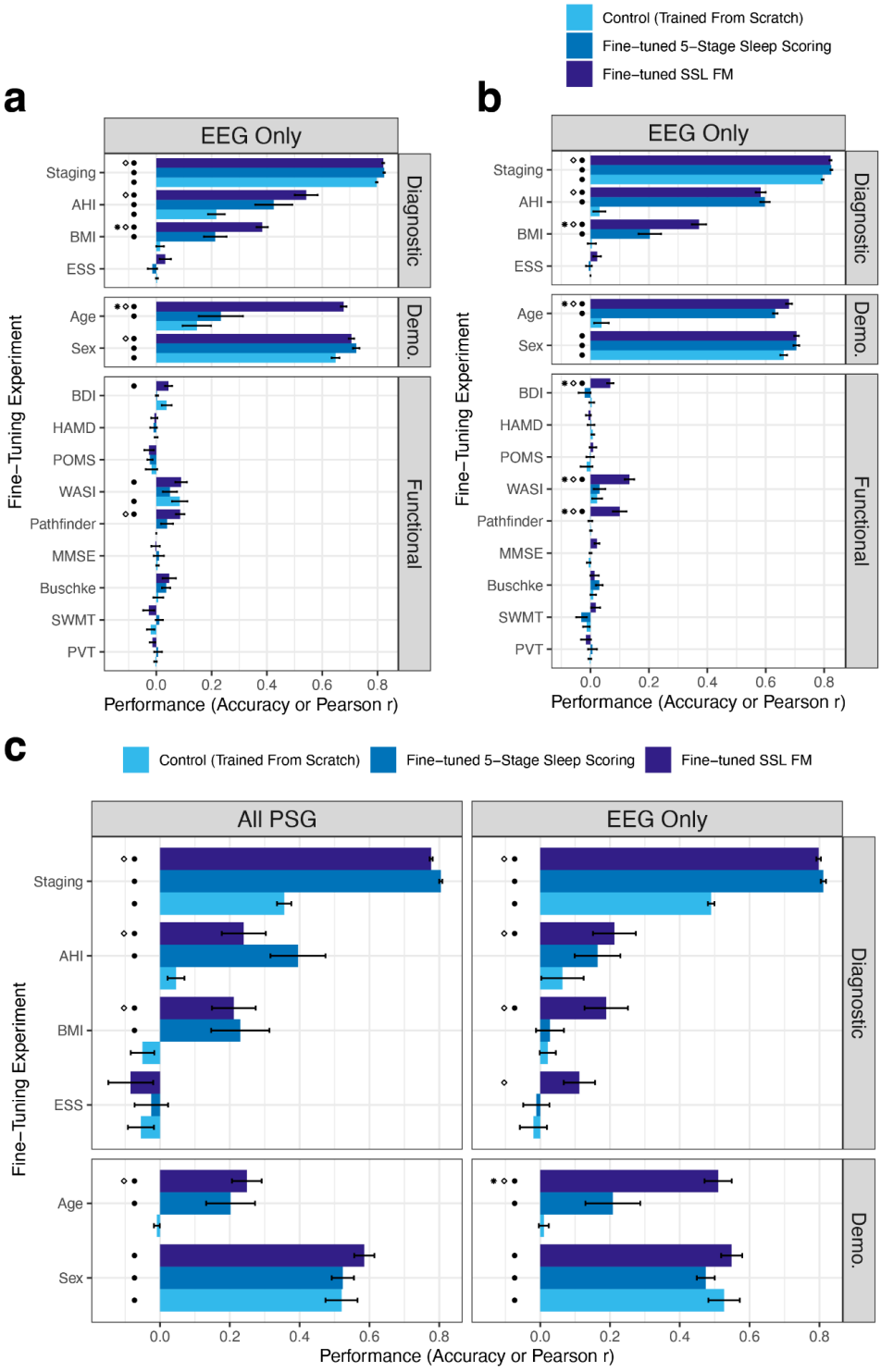
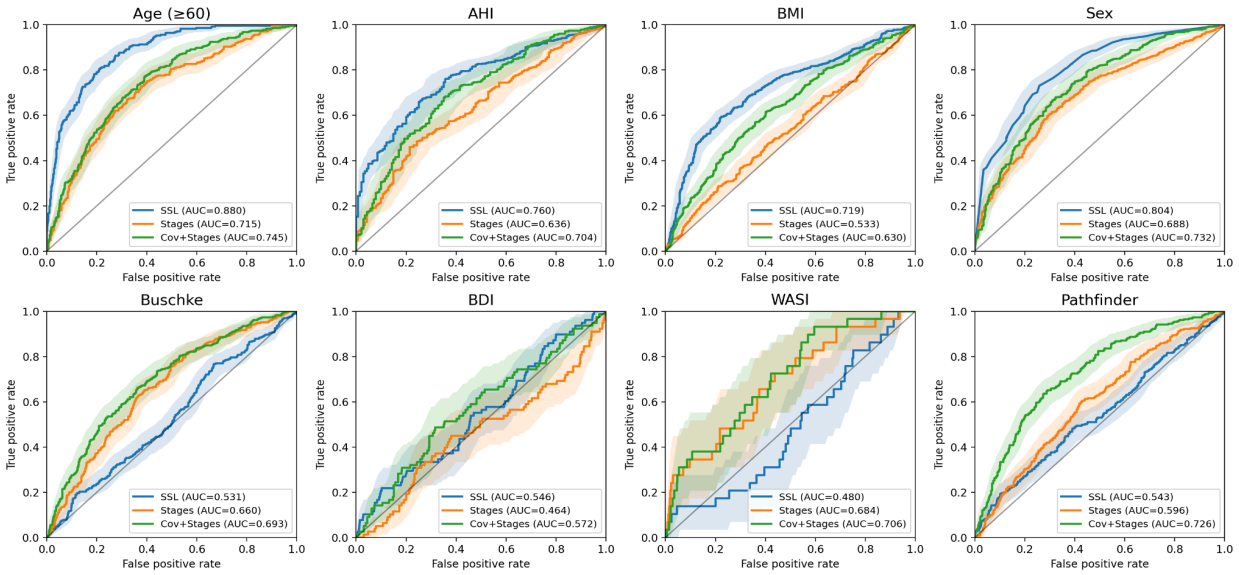
Supplementary Figures

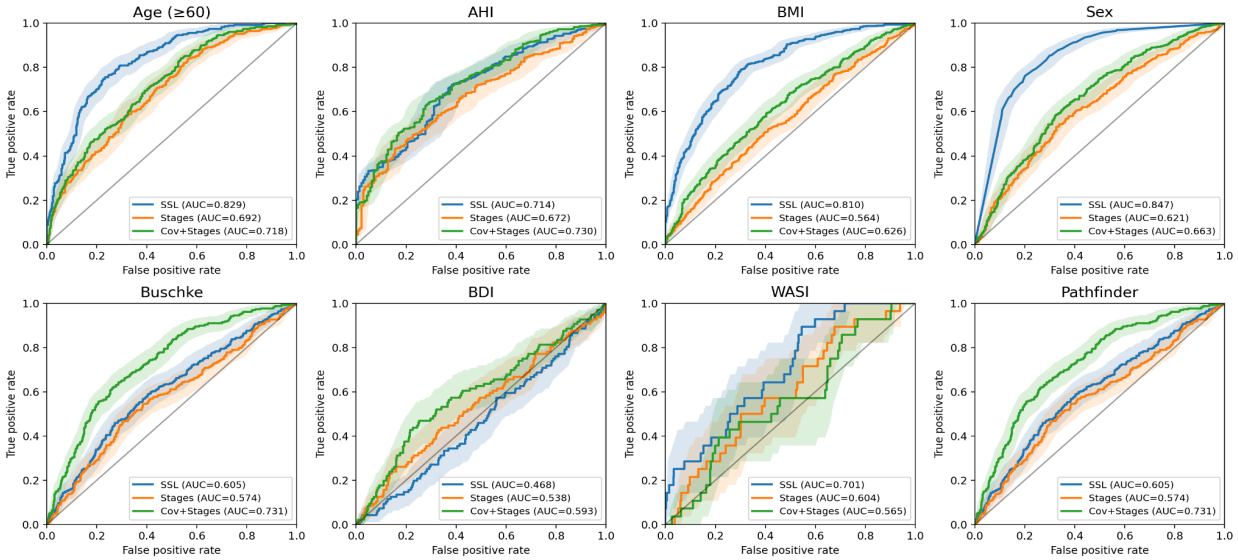


Supplementary Figure 1 | Downstream patterns are stable across transformer depth and replicate on held-out HomePAP data.
Architecture-sensitivity and external-replication analyses comparing models trained from scratch, models pretrained for five-stage sleep scoring, and SSL-pretrained foundation models. **(a)**, APPLES fine-tuning results using a compact

2-layer transformer architecture (128 CNN channels, 2 transformer layers, 384 embedding dimensionality, 6 transformer heads per layer, 768 transformer feed forward dimensionality) for the from-scratch condition only (bars for SSL and 5-Stage models are the same as in the main text and are depicted here for comparison with the smaller from-scratch model). **(b)**, APPLES fine-tuning results using a 6-layer transformer architecture (256 CNN channels, 6 transformer layers, 384 embedding dimensionality, 8 transformer heads per layer, 1536 transformer feed forward dimensionality) for all three conditions (SSL, 5-Stage, and from-scratch). **(c)**, HomePAP fine-tuning results using the primary 12-layer transformer, shown for all-PSG and EEG-only inputs on outcomes shared with APPLES. Performance is accuracy for classification tasks and Pearson r for continuous outcomes; bars show mean ± SD across folds. Across transformer depths, pretrained models showed qualitatively similar downstream patterns, indicating that the main conclusions are not specific to the particular (12-layer) architecture presented throughout the main text. The HomePAP analysis further supports generalization to a second unseen cohort, with SSL and five-stage-pretrained models retaining strong performance on shared diagnostic and demographic outcomes. Smaller from-scratch models performed relatively better than the 12-layer from-scratch model, consistent with lower-capacity architectures being better matched to the limited labeled fine-tuning data.



Supplementary Figure 2 | EEG-derived screening scores provide incremental predictive value beyond demographic covariates and staging-derived sleep-report features. Receiver operating characteristic (ROC) curves for participant-level screening of Age ≥ 60, AHI ≥ 15, BMI ≥ 30, Sex (male), Buschke ≤ Q25, BDI ≥ 14, WASI ≤ 85, and Pathfinder ≤ Q25. Curves compare out-of-fold logistic models using the EEG-derived SSL outcome score (blue), staging-derived sleep-report features alone (orange), and covariates + staging-derived sleep-report features (green; stage proportions and standard sleep-architecture summaries including efficiency/latencies, transition metrics, and bout-duration statistics; see Methods). Shaded bands show 95% confidence intervals; insets report AUROC; the diagonal indicates chance performance.



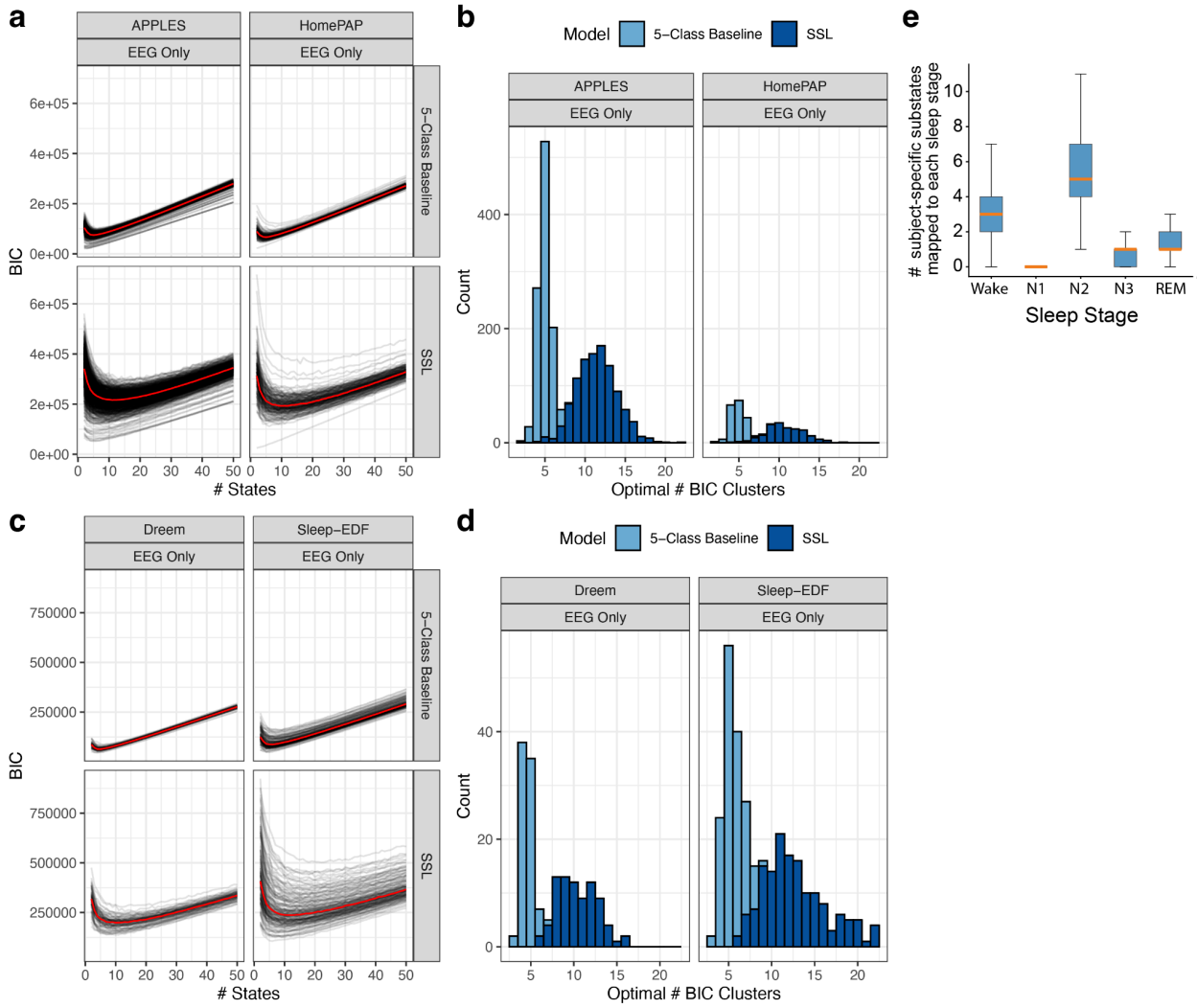
Supplementary Figure 3 | PSG-derived screening scores retain incremental predictive value beyond demographic covariates and staging-derived sleep-report features. ROC curves for the same outcomes and modeling framework as Supplementary Figure 2, but using SSL outcome scores derived from the multimodal PSG foundation model (PSG channels; see Methods). Curves compare the PSG-derived SSL score (blue) against staging-derived sleep-report features alone (orange) and covariates + staging-derived sleep-report features (green). Shaded bands show 95% confidence intervals and insets report AUROC.

Supplementary Table 1 | EEG-only screening analyses: functional outcomes.

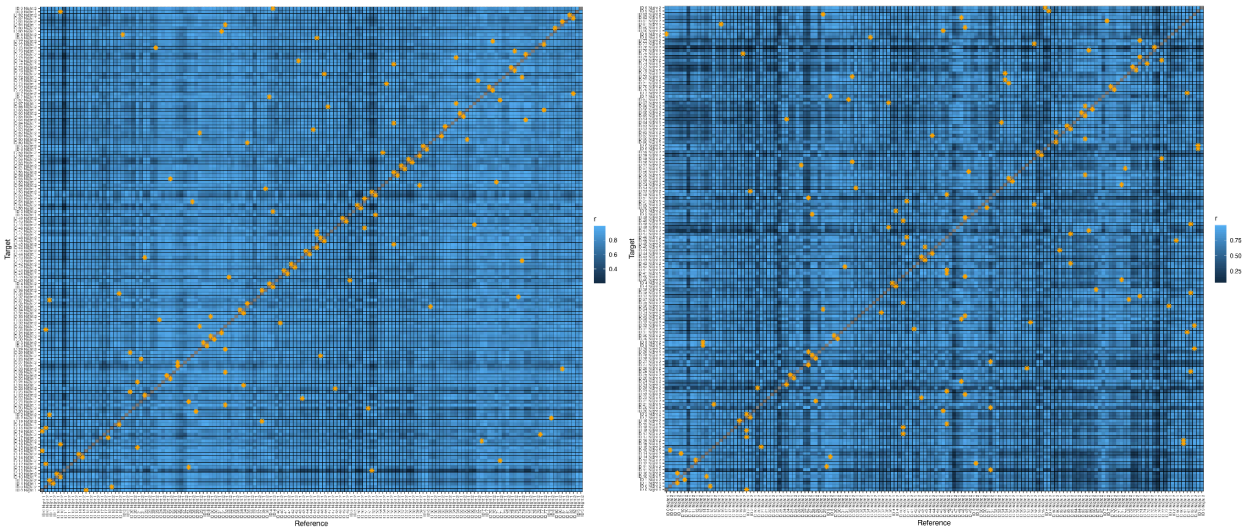
endpoint	auc	brier_oof	prevalence	n	pathfinder_q25
WASI ≤ 85	0.4933598937583001	0.02816693641094637	0.029013539651837523	1034	
Pathfinder ≤ Q25	0.5416216216216216	0.1875010269214683	0.2504835589941973	1034	20.015
HAMD ≥ 8	0.4935800698428536	0.14983564989850257	0.18375241779497098	1034	
Age ≥ 60	0.8765460105012927	0.13296947610688	0.27176015473887816	1034	
Sex (male)	0.8018524775479388	0.1680022082075603	0.660541586073501	1034	

Supplementary Table 2 | PSG screening analyses: functional outcomes.

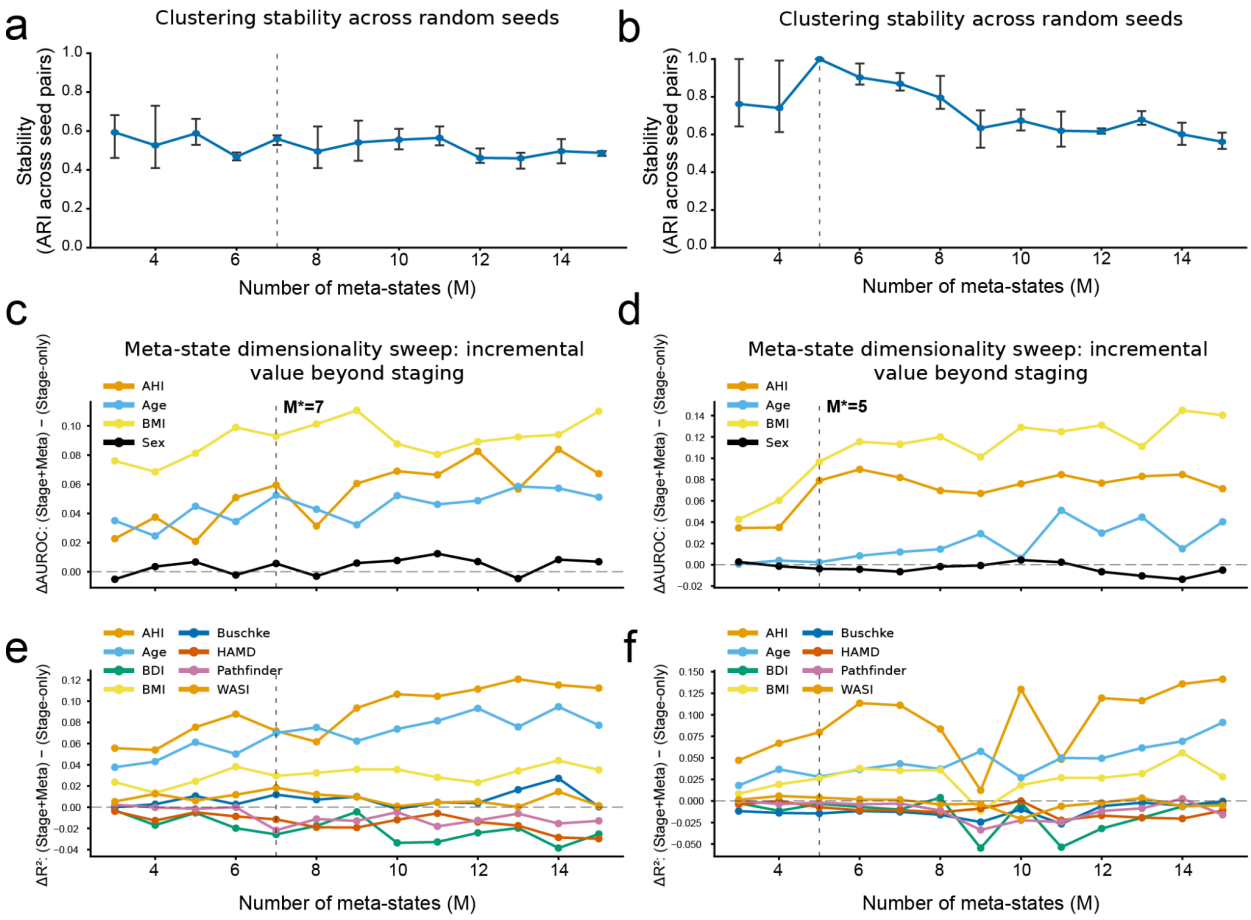
endpoint	auc	brier_oof	prevalence	n	pathfinder_q25
WASI ≤ 85	0.7132138114209827	0.02772339790082954	0.029013539651837523	1034	
Pathfinder ≤ Q25	0.607582513388965	0.18411963741701864	0.2504835589941973	1034	20.015
HAMD ≥ 8	0.5916188575704664	0.14895371432156862	0.18375241779497098	1034	
Age ≥ 60	0.8283308048943017	0.14929806760654873	0.27176015473887816	1034	
Sex (male)	0.8474406944392303	0.14242519871334838	0.660541586073501	1034	



Supplementary Figure 4 | Cross-dataset BIC sweeps show that SSL embeddings support finer within-record latent compression than five-stage-supervised embeddings. Per-record k-means/BIC sweeps were applied to final-layer EEG embeddings from the matched five-stage-supervised baseline and the SSL model. **a,c**, BIC curves across candidate cluster counts for APPLES and HomePAP (**a**) and Dreem and Sleep-EDF (**c**). Thin gray lines show individual recordings; red lines show the dataset median. **b,d**, Distributions of BIC-selected cluster counts for each model and dataset. Across all four held-out datasets, the five-stage-supervised representation converged near the expected five-cluster solution, whereas SSL embeddings selected higher cluster counts, consistent with finer within-record latent structure under the same compression procedure. Median selected counts were APPLES: SSL 11 vs five-stage 5; HomePAP: 10 vs 5; Dreem: 10 vs 5; and Sleep-EDF: 12 vs 5. Paired tests were significant in each dataset (APPLES: $p < 2.2\times10^{-16}$, HomePAP: $p < 2.2\times10^{-16}$, Sleep-EDF: $p < 2.2\times10^{-16}$; Dreem: $p = 4.7\times10^{-15}$), and these results were consistent across randomly initialized k-means seeds (across 10 additional sweeps, optimal BIC scores different at most by 1 for any given run: s.d. across runs for each dataset ranged from 0.42 to 0.48 for the SSL model and from 0 to 0.31 for the five-stage baseline model). These BIC results are interpreted comparatively, not ontologically: they indicate that SSL embeddings support higher-resolution compression than matched five-stage embeddings, with downstream utility evaluated separately in the Rosetta/meta-state analyses. **(e)** “Rosetta stone” boxplots of the number of subject-specific microstates assigned to each AASM stage (dominant stage per microstate), showing that N2 and Wake tend to subdivide more than N3/REM across recordings APPLES.



Supplementary Figure 5 | Cross-night fingerprinting indicates a stable subject-specific component in SSL sleep embeddings. To examine whether the learned representation retains stable subject-specific structure across nights, we performed an exploratory repeated-night analysis. Because APPLES did not include multiple nights, we used another held-out corpus, the Sleep-EDF cohort, which includes 74 participants with two recordings each . Using a correlation-based fingerprinting procedure, each query night was matched to the most similar recording in the dataset on the basis of its night-level embedding summary (without fine-tuning the models). SSL embeddings correctly identified the alternate night from the same participant in 60 of 148 queries (40.5%; 95% CI 32.6–48.9%), exceeding both a conservative chance baseline (1/74; exact binomial test, $p < 2.2\times10^{-16}$) and the five-stage-pretrained baseline (22.9%; Wilcoxon signed-rank test on night-pair correlation scores: $V = 1311.5$, $p = 8.0 \times 10^{-4}$). Although well below ceiling, these findings suggest that the SSL embedding retains a stable, trait-like component across nights within an individual while remaining sensitive to meaningful night-to-night variation.

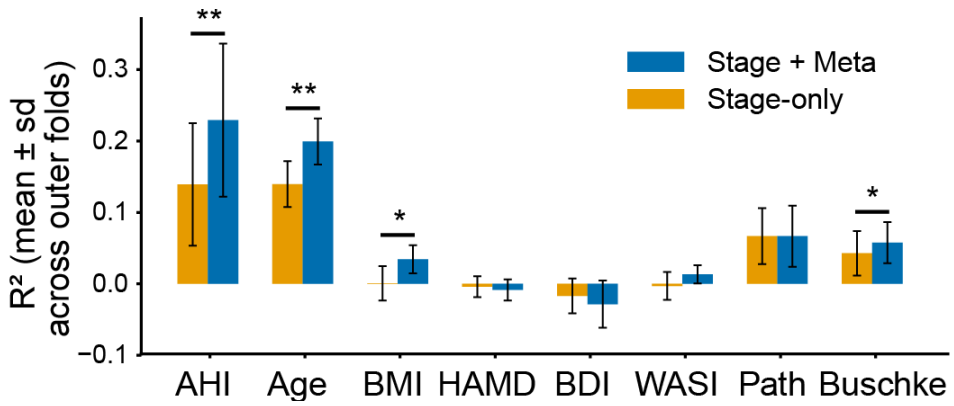


Supplementary Figure 6 | Meta-state dimensionality sweep and clustering stability.
Left column (a,c,e): Analysis applied to EEG-only embeddings from the self-supervised (SSL) foundation model. **Right column (b,d,f):** Same analysis applied to EEG-only embeddings from the five-stage supervised sleep-staging model.

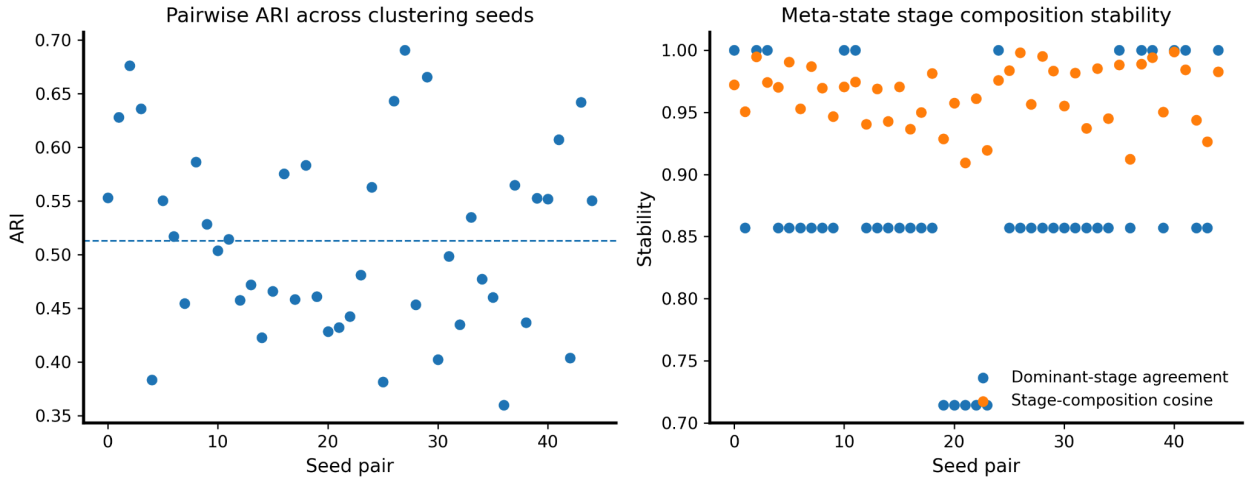
a–b | Clustering stability across random initializations.
For each candidate number of meta-states ($M \in \{3,\ldots,15\}$), clustering stability was quantified as the Adjusted Rand Index (ARI) between meta-state assignments obtained from different random seeds. Points indicate the mean ARI across seed pairs and error bars show the min–max range. The dashed vertical line marks the selected dimensionality M* determined by the predefined stability–performance tradeoff rule (see Methods): SSL embeddings support a compact solution at M*=7, whereas embeddings derived from the five-stage supervised model support M*=5.

c–d | Incremental value for diagnostic and demographic outcomes.
For each M, meta-state features were computed within training folds of a participant-level 5-fold cross-validation procedure (three random seeds: 0,1,2). The plots show the fold-averaged incremental predictive value beyond staging, defined as ΔAUROC = (Stage+Meta) − (Stage-only), for diagnostic and demographic outcomes (AHI, age, BMI, sex).

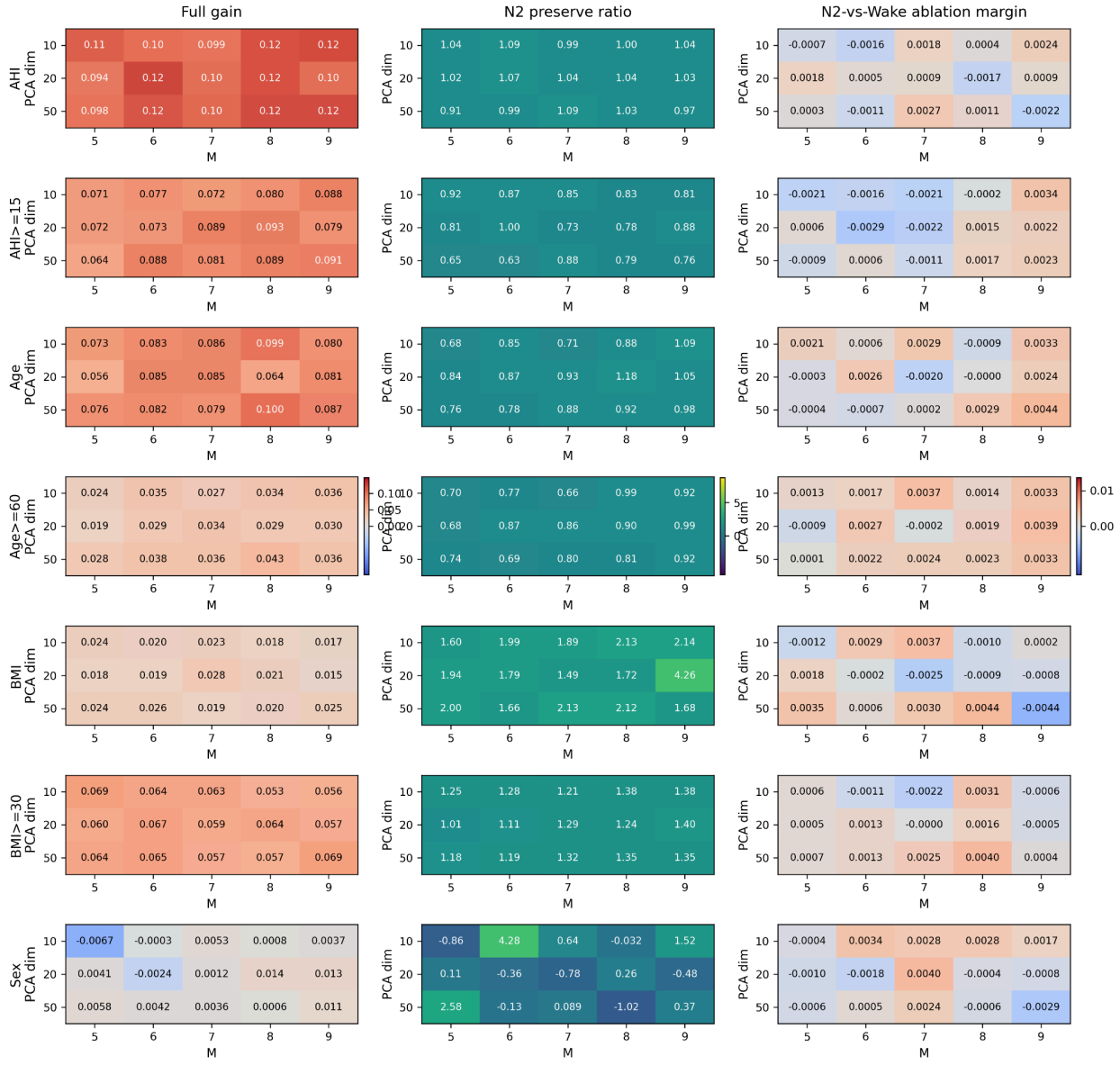
e–f | Incremental value for functional outcomes.
The same procedure was applied to functional mood and cognitive outcomes. Panels plot the fold-averaged ΔR^2 = (Stage+Meta) − (Stage-only) for BDI, HAM-D, POMS, WASI-IQ, Pathfinder, Buschke, SWMT, and PVT. Across outcomes, meta-state features consistently improve prediction relative to staging-only summaries, with the largest gains observed for diagnostic and demographic outcomes.



Supplementary Figure 7 | Meta-state summaries add continuous-outcome signal beyond stage-only sleep summaries. Mean out-of-fold R^2 for continuous outcome prediction using conventional stage-only sleep-architecture summaries versus stage summaries augmented with meta-state occupancy/diversity features. Error bars show SD across outer folds. Adding meta-state features improved prediction most clearly for AHI and age: mean R^2 increased from 0.139 to 0.211 for AHI (ΔR^2 = +0.072) and from 0.140 to 0.210 for age (ΔR^2 = +0.070). Smaller gains were observed for BMI and Buschke, whereas mood and other cognitive outcomes showed limited or inconsistent benefit at this level of aggregation. Asterisks denote paired fold-wise improvement over the stage-only model (*p < 0.05, **p < 0.01).

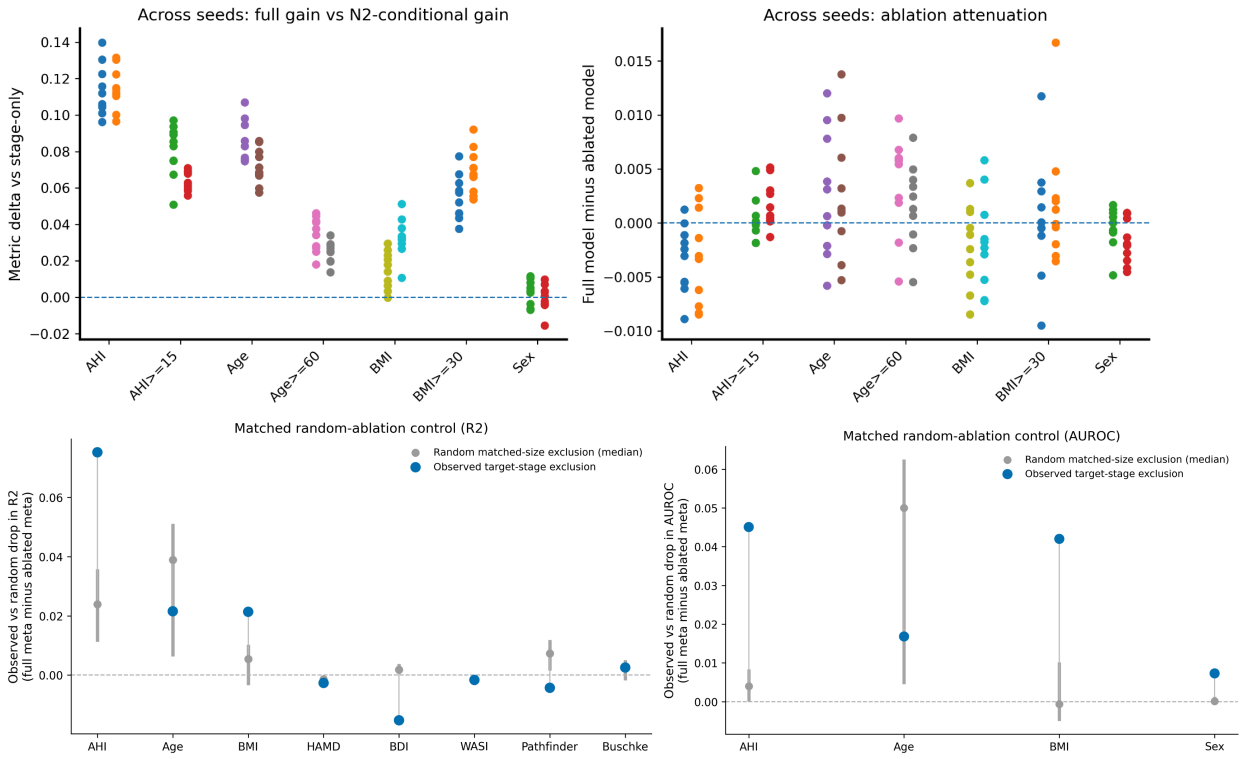


Supplementary Figure 8 | Seed sensitivity and stage-composition stability of the EEG meta-state solution. Because subject-specific microstates are used as an intermediate representation for mapping each night into a shared cross-subject meta-state space, we tested whether the selected EEG meta-state solution was stable to clustering initialization. The analysis was repeated across 10 random clustering seeds at M = 7. Left, pairwise adjusted Rand index (ARI) across all seed pairs showed moderate label-level recoverability of exact meta-state assignments (mean ARI = 0.513; range, 0.359–0.691), as expected for unsupervised clustering in which cluster labels and boundaries can vary across initializations. Right, after matching meta-states across runs by their stage-composition profiles, the physiologically interpretable organization of the solution was substantially more stable: dominant-stage agreement averaged 0.879 and stage-composition cosine similarity averaged 0.964 across seed pairs. Thus, although exact microstate labels are not identical across random initializations, the stage-anchored structure of the shared meta-state space is highly reproducible.

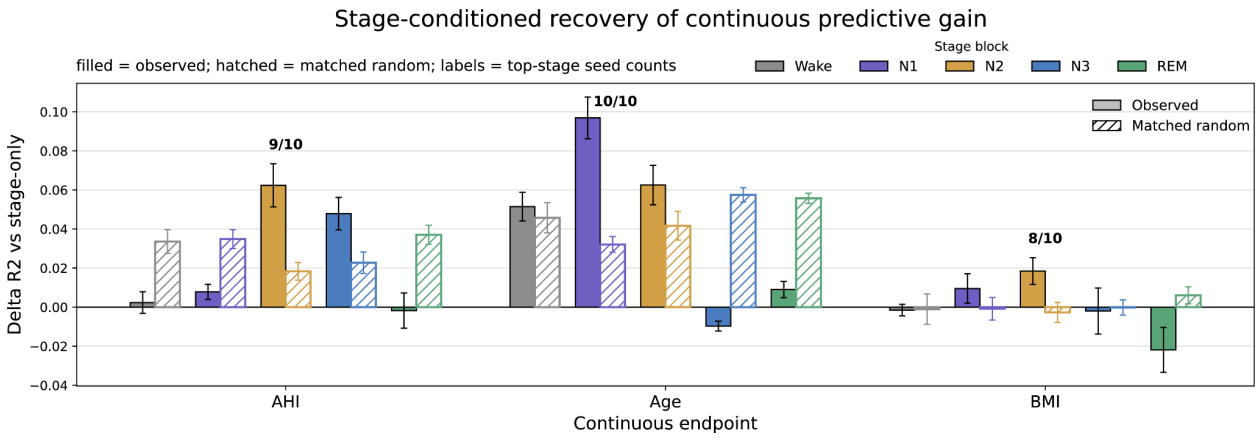


Supplementary Figure 9 | Meta-state results are locally robust to the number of meta-states and PCA dimensionality. Local hyperparameter sensitivity analysis for the EEG meta-state framework across a grid of meta-state counts (M = 5–9) and PCA dimensionalities used before cross-subject meta-state clustering (10, 20, or 50 PCs). Rows show diagnostic and demographic outcomes; columns summarize three robustness quantities. **Full gain** denotes the improvement of stage + meta-state features over stage-only summaries, measured as ΔR^2 for continuous outcomes (AHI, age, BMI) and ΔAUROC for binary outcomes (AHI ≥ 15, age ≥ 60, BMI ≥ 30, sex). **N2 preserve ratio** denotes the fraction of the full stage + meta-state gain recovered when meta-state occupancy is computed conditionally within N2, with values near or above 1 indicating that within-N2 structure preserves most or all of the full meta-state benefit. **N2-vs-Wake ablation margin** compares the loss of performance after removing N2-dominant versus Wake-dominant meta-state features; positive values indicate greater attenuation after N2-dominant ablation. Across the local grid, full meta-state gains remained positive for the principal diagnostic and demographic outcomes, and within-N2 conditional occupancy generally preserved a substantial fraction of these gains. In contrast, N2-vs-Wake ablation margins were small and variable across settings, supporting a conservative

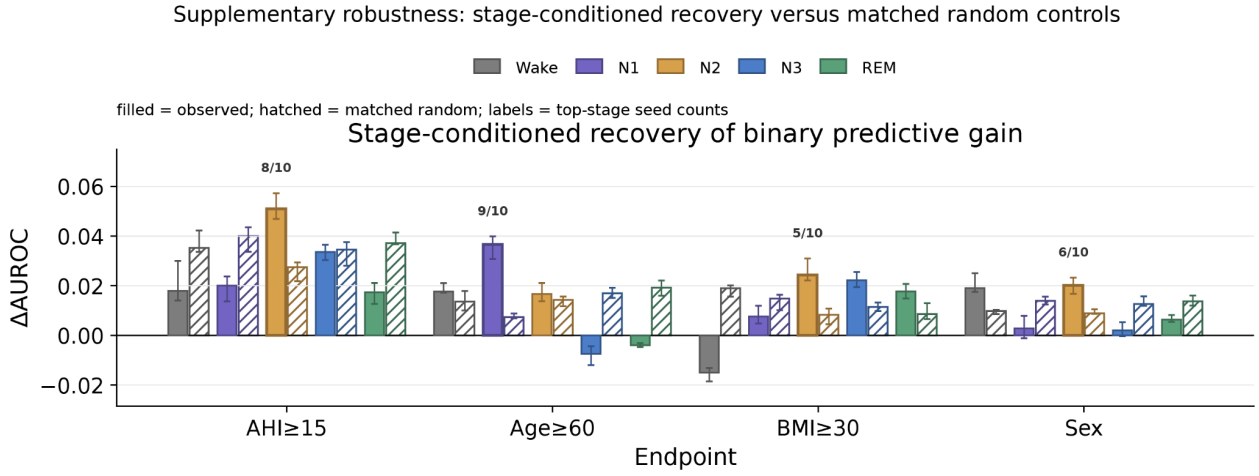
interpretation that within-N2 heterogeneity contributes importantly to the meta-state signal, while not depending on a single hyperparameter setting or a uniquely defined N2-dominant cluster subset. Cell values are printed within each heatmap, and color limits are standardized within each robustness metric to aid visual comparison.



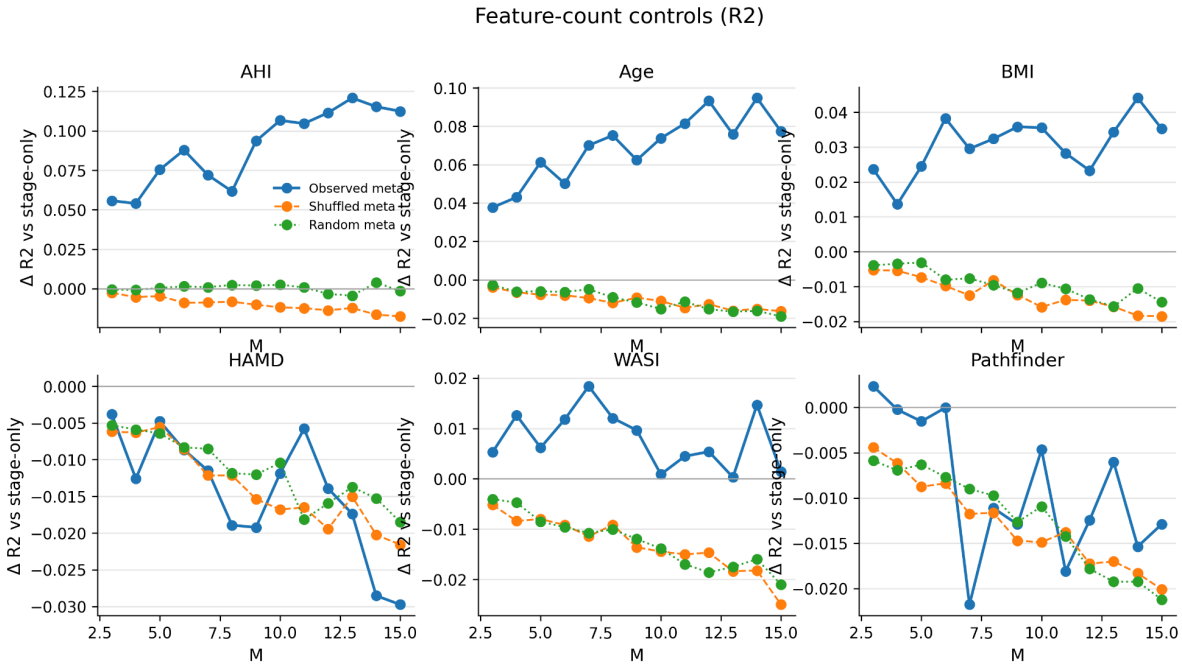
Supplementary Figure 10 | Seed-robust stage+meta gains and stage-localization controls. Stage+meta analyses were repeated across clustering seeds to assess robustness of downstream gains and their localization to N2-linked structure. **Top left**, incremental gain over stage-only features across seeds for the full stage+meta model and for N2-conditioned meta-state occupancy features. Gains remained positive across all 10 seeds for AHI, AHI ≥ 15, Age, Age ≥ 60, and BMI ≥ 30, and in 9/10 seeds for continuous BMI; sex showed small and less consistent effects. N2-conditioned features recovered a substantial fraction of the full stage+meta gain, particularly for AHI- and age-related outcomes. The weaker N2-conditioned recovery for Age ≥ 60 relative to continuous Age suggests that age-related information is more apparent across the continuous age gradient than at a single threshold. **Top right**, dominant-stage ablation analyses across seeds, plotted as the full model minus the ablated model. Ablation effects were smaller and more variable than conditional-occupancy gains, indicating that no single dominant-stage subset uniquely explains the full stage+meta signal. **Bottom row**, matched random-ablation controls for continuous outcomes (R^2, left) and binary outcomes (AUROC, right). Blue points show the observed target-stage exclusion; gray points/intervals summarize matched random exclusions of the same feature count. Together, these analyses support an important contribution of within-N2 heterogeneity to the incremental stage+meta signal, especially for apnea- and BMI-related outcomes, while indicating that the SSL-derived advantage is not exclusively driven by N2-linked features.

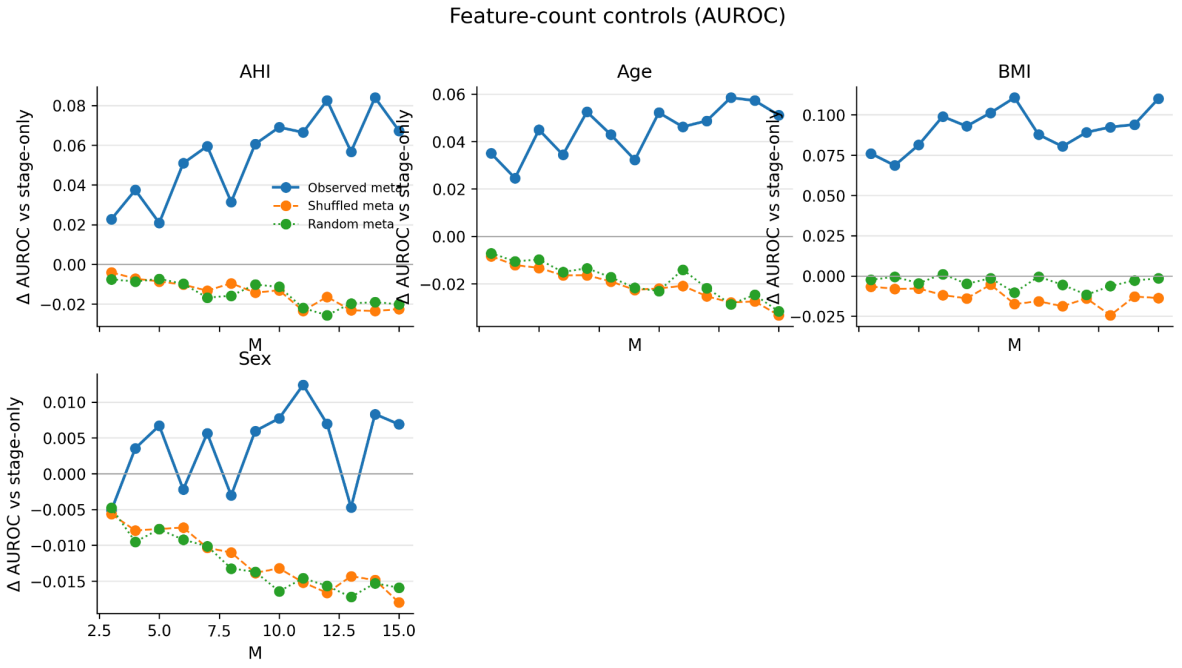


Supplementary Figure 11 | Stage-conditioned recovery of continuous predictive gain versus matched random controls. For each continuous outcome, meta-state occupancy features were recomputed conditionally within each canonical sleep stage and added to the stage-only baseline. Filled bars show observed ΔR^2 relative to the stage-only model; hatched bars show matched-size random feature-block controls evaluated under the same cross-validation folds and clustering seeds. Error bars show variability across seeds. Labels above each outcome indicate how often the corresponding stage-conditioned block was the top-performing observed stage across seeds. These analyses provide the continuous-outcome counterpart to the binary ΔAUROC control (Supplementary Figure 12), showing whether within-stage meta-state structure recovers predictive information beyond total stage duration and beyond matched random feature additions.



Supplementary Figure 12 | Stage-conditioned recovery of binary predictive gain versus matched random controls. For each binary outcome, meta-state occupancy features were recomputed conditionally within each canonical sleep stage and added to the stage-only baseline. Filled bars show observed ΔAUROC relative to the stage-only model; hatched bars show matched-size random feature-block controls evaluated under the same cross-validation folds and clustering seeds. Error bars show variability across seeds. Labels above each outcome indicate how often the corresponding stage-conditioned block was the top-performing observed stage across seeds. N2-conditioned features recovered substantial gain for AHI≥15 and BMI≥30, while age and sex showed more distributed or weaker stage specificity. These controls indicate that within-stage meta-state structure carries outcome-specific predictive information beyond total stage duration and beyond matched random feature additions, while also supporting a more cautious interpretation that the SSL advantage is not exclusively N2-driven for all outcomes.





Supplementary Figure 13 | Meta-state predictive gains are not explained by adding more features. For each candidate number of meta-states *M*, we compared the observed stage+meta model against two matched feature-count controls in the same 5-fold cross-validation framework. The observed meta-state features improved held-out AUROC/R^2 beyond stage-only summaries for AHI, age, and BMI across a broad range of *M*. In contrast, adding an equally sized shuffled meta-state feature block or a randomized microstate-to-meta-state occupancy block

produced near-zero or negative changes in AUROC/R^2. These controls indicate that the predictive gain depends on the learned subject-specific meta-state organization rather than simply on adding more predictors.

Dataset Summary

Dataset	Age (Mean/Mdn)	Age Range (Min./Max.)	Sex (F/M)	Participants	Overnights	Total Examples
MESA	69.42/68	54/90	991/861	1,852	1,852	87,366
MrOS	74.67/74	67/90	0/903	903	1,808	96,713
SHHS	62.47/62	39/90	1,229/1,071	2,300	4,698	191,880
WSC	58.16/58	37/78	438/481	919	1,660	53,989
SOF	83/82	76/90	393/0	398	398	17,581
CFS	41.44/43.31	6.77/88.53	362/295	657	657	28,826
NCHSDB	21.38/19.93	18/58.35	88/84	172	188	5,915
Train Total	**63.15/65**	**6.77/90**	**3,506/3,695**	**7,201**	**11,261**	**482,270**
Held-Out Train	**63.37/65.5**	**10.76/90**	**381/398**	**779**	**779**	**32,405**
DreemO + DreemH	35.3, 45.6/-	-/-	54/26	80	80	742
Sleep-EDF	59.21/57	25/101	41/36	77	152	1,849
HomePAP	46.05/45.98	20.25/79.77	92/99	191	191	1,668
APPLES	51.59/52	18/83	353/691	1,044	1,044	9,683

Supplementary Table 3 | Overview of polysomnography datasets and cohort demographics. Summary of the publicly available datasets utilized for model pretraining, internal validation, and downstream clinical evaluation (e.g., APPLES, HomePAP, Dreem, Sleep-EDF). For each dataset partition, the total number of PSG recordings, unique participants, and baseline demographic characteristics (such as age, sex distribution, and relevant clinical metrics) are provided. Total examples are the number of 50.5min input sequences derived from each corresponding dataset (note that Training data were augmented by 4x oversampling with a 12.5-epoch hop size).