Hypnos: Learning Generalisable Representations of Sleep Physiology via Next-Token Prediction

Jonathan F. Carter*
Institute of Biomedical Engineering
University of Oxford

Lionel Tarassenko
Institute of Biomedical Engineering
University of Oxford

Abstract

Foundation models offer a promising route to compress multi-modal physiological signals into compact representations of human health, with broad applications across sleep medicine, cardiology, neurology and other healthcare domains. Existing models have typically been trained with masked-reconstruction or contrastive objectives. However, masked reconstruction may be poorly suited to the stochastic nature of these signals, while contrastive approaches rely on positive-pair definitions despite the semantic invariances of physiological signals being poorly understood. In this work, we show that next-token prediction is a simple and scalable alternative. We develop Hypnos, a multi-modal sleep foundation model trained using eight different sensing modalities (e.g. EEG, ECG, respiratory signals) drawn from over 20,000 overnight polysomnography recordings. We tokenize each modality into streams of discrete tokens using residual vector quantization, then train a large auto-regressive RQ-Transformer to jointly predict the next token across all modalities in parallel. After training, Hypnos can be applied to continuous streams of sensor data from any subset of supported modalities, generating embeddings for downstream tasks. Hypnos significantly outperforms existing foundation models on most benchmarks evaluated. In sleep stage classification, we match the performance of strong supervised baselines on held-out test sets whilst using $100\times$ less labelled data. Hypnos even generalises beyond sleep recordings, surpassing a dedicated ECG foundation model at detecting atrial fibrillation. Our results demonstrate that auto-regressive token modelling is a strong self-supervised approach for representation learning from multi-modal physiological signals.

1 Introduction

Physiological recordings such as polysomnography (PSG) capture hours of continuous, multi-modal sensor data from the brain and body. A single overnight study with eight channels recorded at 128 Hz yields over 30 million data points. How can we compress minutes, hours, or even days of continuous, multi-modal sensor data into better measures of health for tasks such as identifying neurodegenerative or cardiovascular disease? A key motivation for physiological foundation models is to use large quantities of unlabelled sensor data to address this challenge. Prior work has shown that self-supervised learning (SSL) techniques can be used to learn effective representations from a broad range of physiological sensors [5, 1, 68]. These models have predominantly been trained using either contrastive learning [58, 1] or masked input-space reconstruction [31, 40, 19] approaches. However, each has known shortcomings on stochastic, continuous signals (see Section 2). Next-token prediction is a simple alternative: it underpins modern Large Language Models (LLMs) [46, 47, 8], has been demonstrated to scale to context lengths exceeding 1 M tokens [20], and has been successfully applied in the analogous, continuous-signal domain of audio [7, 17].

*jonathan.carter@eng.ox.ac.uk

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

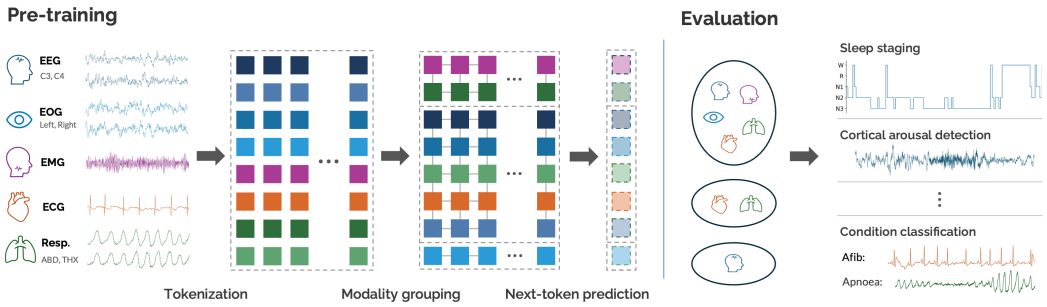


Figure 1: **Overview.** Hypnos is a large auto-regressive RQ-transformer trained via multi-modal next-token prediction on tokenized streams of physiological sensor data. During pre-training, cross-modal attention is restricted to randomly sampled sub-groups, improving test-time generalisation to subsets of supported modalities. After pre-training, Hypnos can be used to generate high-quality embeddings for a diverse range of sensor configurations and downstream tasks over different timescales. We evaluate on tasks including sleep stage classification of overnight recordings, the detection of cortical arousal events, and atrial fibrillation (Afib) detection.

In this work, we show that next-token prediction is a simple and scalable self-supervised learning objective for multi-modal physiological sensor data (Figure 1). We introduce Hypnos, a sleep foundation model trained on eight sensing modalities including electroencephalogram (EEG), electrocardiogram (ECG), and respiratory signals. Each modality is tokenized into a stream of discrete tokens using residual vector quantization (RVQ), and an auto-regressive Transformer is trained to jointly predict the next token across modalities in parallel. Using over 20,000 overnight PSG recordings drawn from nine public datasets, we find that both next-token perplexity and downstream probing performance continue to improve with scale. Our key contributions are as follows:

- **Physiological next-token prediction:** We demonstrate that next-token prediction is an effective method for joint self-supervised learning from diverse physiological signals, unifying generative modelling and representation learning into a single architecture.
- **State-of-the-art performance:** Hypnos significantly outperforms prior sleep foundation models on most benchmark tasks and datasets evaluated. Embeddings from Hypnos even transfer beyond sleep, outperforming a dedicated ECG foundation model on external single-lead ECG benchmarks.
- **Deployment-focused design:** Hypnos can run in a streaming fashion, supports diverse sensor combinations and arbitrary length recordings, and generates embeddings at a convenient rate of 1 Hz. This can enable real-time applications such as closed-loop neuromodulation.

Pre-trained model weights and inference code are available at: https://github.com/joncarter1/hypnos.

2 Background and Motivation

Vector Quantization Existing physiological foundation models typically operate on short windows of sensor data, e.g. 5 minutes or less [58, 53, 36], or on derived quantities such as step counts and heart rate statistics [40]. SleepMaMi [41] is a notable exception, training a Mamba-based macro-encoder over full-night sequences of epoch-level embeddings from a frozen micro-encoder. High sampling rates are essential to capture fine details from modalities such as EEG or ECG, e.g. peak-peak timings. However, these signals are also highly compressible. For example, it has long been known that continuous ECG data can be modelled using coupled differential equations with a small number of latent variables [37]. This motivates an initial compression of the data (e.g. tokenization) prior to sequence modelling. Vector quantization [60] learns a discrete codebook that maps continuous inputs to tokens, and has been used to build foundation models across domains including brain [25, 65] and audio [14, 7] signals. NeuroLM [26] trained an autoregressive Transformer over single-codebook VQ tokens of EEG signals. We extend this line of work to multi-modal physiological signals, using residual vector quantization and an RQ-Transformer [29] to model multiple high-rate streams in

parallel. Our technical approach, discussed in Section 3, is most similar to Moshi [17], a state-of-the-art speech-text foundation model.

Stochasticity Next-token prediction is well-suited to handle the stochastic nature of physiological signals. Rather than trying to exactly reconstruct physiological signals in input space, e.g. [31, 19, 62], which may be sensitive to exact peak timings and waveform morphology, next-token prediction can assign probability mass to the plausible subset of next tokens (see Figure 2). Alternatively, the signal could be modelled using a continuous distribution, e.g. via diffusion [54, 22], as recently used in the analogous domain of audio [51]. However, we adopt discrete next-token prediction in this work to inherit architectures, training recipes, and simple tractable likelihoods from audio and language modelling.

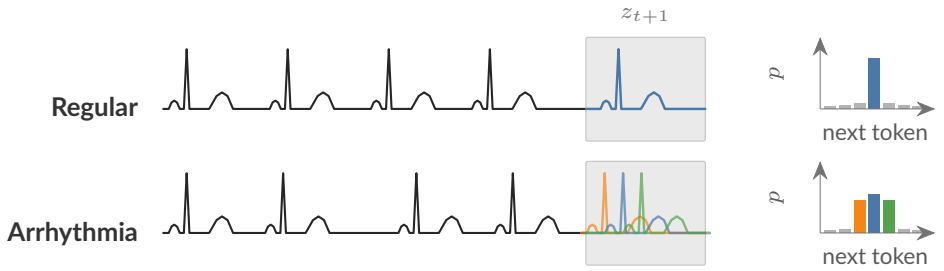


Figure 2: For stochastic signals such as an ECG (left), the distribution over the future may be well-characterised by a peaked distribution p over some discretisation (i.e. tokenization) z_{t+1}. For example, during arrhythmia, a subset of tokens with similar QRS morphologies may be equally likely.

Discretisation Tokenization is itself a design bias. However, sensor data already has a finite sampling rate and precision, and its effective resolution is limited by the noise floor. Our tokenizers achieve low reconstruction error (Figure 9 in the Appendix), and downstream performance is stable across a range of token lengths (Figure 11).

Multimodality Contrastive methods are characterised by the use of positive pairs. For example, prior work has constructed positive pairs using physiological signals from different modalities over the same time range [70, 58], segments from adjacent time ranges [28], and segments from the same subject at different time ranges [44]. Positive pair definitions and augmentations are inductive biases that implicitly shape the learnt latent space. For example, the motivation for the leave-one-out method used by SleepFM is to 'encourage each embedding to align semantically with all other modalities' [58]. A potential limitation of SleepFM's leave-one-out construction is that it encourages the model to extract information *shared* between modalities, and may discard information that is not [15]. SleepMaMi addresses this directly, pairing private per-modality encoders with a shared encoder and using additional non-contrastive losses [41]. However, this approach widens the design space.

Different sensing modalities are intentionally recorded because they give complementary 'views' of physiology, with independent variations providing useful information for downstream tasks. For example, brain activity measured by the EEG often looks similar between Wake and Rapid Eye Movement (REM) sleep, but is easily distinguished using activity measured by the electrooculogram (EOG) [23]. This is unlike contrastive learning in vision [13], where augmentations such as crops or colour jitters are chosen precisely because they do not change the underlying semantic content. For physiological signals, the space of invariant augmentations is not well understood. For example, should ECG embeddings from one subject during exercise be closer to that of another subject during exercise, or the same subject when stationary? Human-defined augmentations may limit the effectiveness of both contrastive learning and the related approach of predicting transformations applied to the input [68, 24].

Sequential Learning Finally, existing physiological foundation models have not been designed for sequential updates and streaming inference. This may be beneficial across a range of applications, ranging from closed loop neuromodulation to remote patient monitoring, by enabling real-time

tracking of physiological state. The autoregressive nature of next-token prediction is well-suited for this downstream use case.

3 Method

3.1 Datasets

To train and evaluate our models, we use overnight polysomnography (PSG) recordings drawn from datasets available from the National Sleep Research Resource (NSRR, [71]): SHHS [45], CCSHS [49] CFS [48], CHAT [34], MESA [12], MrOS [55], NCHSDB [30], and WSC [67]; using the same training, validation and test set splits for each dataset as Shuai *et al.* [53]. We use the Dreem Open Datasets (DOD-H and DOD-O) for further external evaluation [21]. Collectively, these datasets contain over 20,000 overnight polysomnography recordings spanning a broad range of patient demographics, sensors and recording configurations.

Modalities We use up to $M = 8$ modalities per recording, drawn from five physiological modality groups commonly available across the NSRR cohorts: central electroencephalogram (EEG), electrooculogram (EOG), chin electromyogram (EMG), electrocardiogram (ECG), and abdominal (ABD) / thoracic (THX) respiratory effort. These were selected based on their prevalence across cohorts and their relevance to downstream sleep analysis tasks. Table 1 summarises the supported input channels to Hypnos and the sampling rates used for each modality.

Table 1: **Supported modalities.**

Modality	Rate (Hz)	Tokenizer group
EEG (C3–M2)	128	EEG
EEG (C4–M1)	128	EEG
EOG (E1–M2)	128	EOG
EOG (E2–M1)	128	EOG
Chin EMG	128	EMG
ECG	128	ECG
Abdominal effort	32	Resp
Thoracic effort	32	Resp

Pre-processing Each channel was minimally pre-processed by resampling to a consistent rate across recordings before filtering and normalisation. Full pre-processing details are given in Section A.1.

3.2 Tokenizing Physiological Signals

For each modality, we train tokenizers to transform the continuous raw signals into discrete tokens. We do this using an encoder-decoder architecture with a residual vector quantization (RVQ) layer, previously used by several foundation models, including for audio [17] and brain data [65]. Residual vector quantization [27, 35] extends vector quantization by representing each input unit (e.g. a signal segment) as a sum of K tokens drawn from K separate codebooks. The first token quantizes the input vector; each subsequent token quantizes the residual error left by the preceding partial reconstruction.

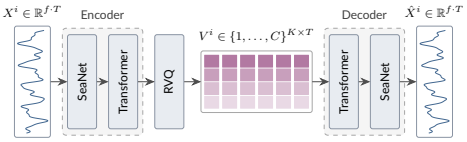


Figure 3: **Tokenizer training.** Each signal X^i is encoded with RVQ into discrete residual tokens V^i. Encoder and decoder are trained jointly to reconstruct X^i.

We tokenize each 1D channel $X^i \in \mathbb{R}^{f \cdot T}$ into a stream of discrete residual tokens $V^i \in \{1, \ldots, C\}^{K \times T}$, where f is the sampling rate in Hz, T is the sequence length in seconds, K is the number of residual levels, and C is the per-codebook vocabulary size. This differs from the design of BrainTokenizer [65], which compresses EEG or MEG data with varying numbers of channels into a fixed number of 'virtual' channels. This allows us to delegate the handling of missing modalities to downstream sequence modelling described in Section 3.3.

Architecture Our encoders and decoders consist of stacks of convolutional (SeaNet [57]) and Transformer layers [61], as illustrated in Figure 3. We configure the stride in the convolutional layers such that the encoders produce embeddings at 1 Hz. All convolutions are causal [59], and each Transformer layer uses a causal sliding window. This design means that after training the tokenizer can be applied to arbitrary length input sequences, supporting streaming inference and enabling us to apply the tokenizers to an entire night of data in a single forward pass.

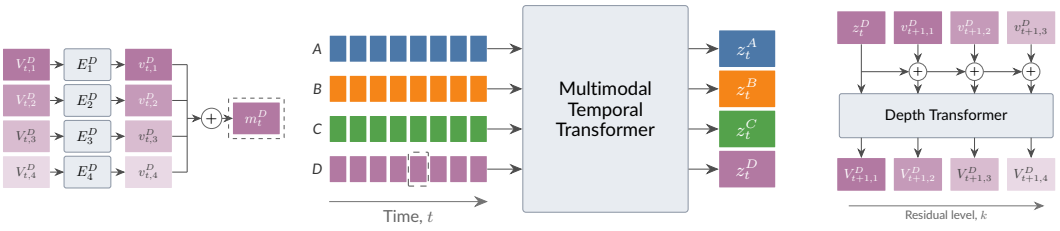


Figure 4: **Hypnos training.** (a) For each time t, the K discrete residual tokens (illustrated with $K = 4$) from modality i are combined to form an embedding m_t^i. (b) A Transformer backbone mixes information to produce embeddings z_t^i for each modality, e.g. $i \in \{A, B, C, D\}$. (c) For all i, t, the Depth Transformer auto-regressively predicts the next residual token $V_{t+1,k}^i$ conditioned on z_t^i.

Because they have similar signal characteristics, we share tokenizers between EEG channels, between EOG channels, and between respiratory channels, resulting in five tokenizers for eight modalities. The number of quantizers K was chosen for each modality to maintain high reconstruction accuracy. Full architecture and training hyper-parameters are given in Section A.2. An investigation into the effect of quantization depth is performed in Section B.1.

Optimisation Each tokenizer is trained on 64-second windows sampled from the training split of the pre-training datasets. We use a multi-term reconstruction loss combined with an RVQ commitment penalty closely following BrainTokenizer [65]. Full loss specifications and hyper-parameters are given in Section A.2.

3.3 Hypnos

After tokenization, we train a RQ-Transformer [29] to minimise the conditional log-likelihood over next residual tokens with all modalities, timesteps and residual levels weighted equally:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{k=1}^{K} \log p_\theta\left(V_{t,k}^i \mid V_{<t,1:K},\, V_{t,1:k-1}^i\right). \tag{1}$$

Here $V_{<t,1:K}$ denotes the tokens at all residual levels and modalities before time t, restricted to the same attention group as modality i during training (Section 3.4). A key advantage of the RQ-Transformer design is that, rather than flattening the $K \cdot T$ tokens per modality into a single sequence, the architecture decouples temporal modelling from residual-depth modelling, meaning self-attention scales with T rather than $K \cdot T$. This is illustrated in Figure 4.

The first stage of the model is a learnt embedding layer which aggregates discrete residual tokens $V_{t,k}^i$ into a single embedding m_t^i for each modality and timestep following prior work [29, 17]. This is followed by a stack of Transformer layers [61] which aggregate information over time and across modalities from the sequences of embeddings m_t^i. To reduce computational complexity, each layer alternates between temporal and modality attention, rather than all-to-all attention. Finally, a depth transformer which auto-regressively predicts residual tokens for the next timestep conditioned on the modality embedding z_t^i. To allow the shared depth transformer to disambiguate modalities, a learnt per-modality embedding e^i is also added to the conditioning. Predictions are made in parallel for all modalities i and timesteps t. The backbone and depth transformer parameters are shared across all modalities.

After training, we use the output embeddings z_t^i from the temporal transformer for downstream tasks. The RQ-Transformer design means that the outputs of the temporal transformer are a natural choice to use as embeddings for downstream tasks. The training process encourages these embeddings to simultaneously encode coarse to fine details required by the depth transformer to predict each output residual token.

Design and Optimisation We use modern Transformer components throughout, with sliding-window attention [6] used to enable length-generalisation during inference. Models are trained with AdamW [32] and a cosine learning rate schedule, following prior language and audio work [8, 17].

By default, we use batch size $B = 512$ and context length $T = 512$, i.e. ≈ 8.5 minutes of data per sequence at 1 Hz. We evaluated three model scales (Table 2) informed by Vision Transformer configurations [18]. Full architecture and training hyper-parameters are given in Section A.3.

Table 2: **Hypnos model configurations**. Parameter counts include token embedding tables, both Transformer stacks, and per-codebook output heads.

	Temporal Transformer			Depth Transformer			
Model	Layers	Hidden D	Heads	Layers	Hidden D	Heads	Params
Hypnos-Tiny	12	192	3	4	128	2	37M
Hypnos-Small	12	384	6	4	192	3	81M
Hypnos-Base	12	768	12	4	384	6	222M

3.4 Modality Masking

A desired property of our model is robustness to missing modalities. Sleep studies use a wide range of sensor configurations, from full polysomnography (PSG) in clinical settings to single-channel EEG or cardio-respiratory only recordings in the home using wearable devices. To improve generalisation to subsets of supported modalities during inference, we randomly divide modalities into one or more groups of varying sizes during training. We then restrict attention in each Transformer layer so that modalities within a group can only attend to each other, as illustrated in Figure 5.

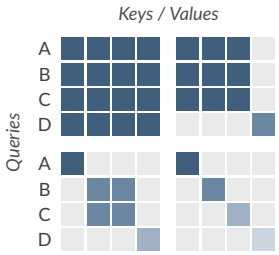


Figure 5: **Example cross-modal attention matrices ($M = 4$).** During training, attention is restricted to random sub-groups.

We split modalities into a variable number of groups G by sampling from a Chinese Restaurant Process [2] with concentration parameter α. Prior work has improved robustness to missing modalities by randomly masking out sensing modalities during training, e.g. [11, 66, 53]. We use the Chinese Restaurant Process as a convenient way to interpolate between using a single group ($\alpha \to 0$, G=1) and fully independent ($\alpha \to \infty$, G=M), i.e. no cross-modal fusion during training. We use $\alpha = 1$ by default, which exposes the model to a wide range of group sizes during training, and which gives $\mathbb{E}[G] \approx 2.72$ for M=8. We expected this to give better test-time performance than random dropout, since no tokens go unused during training, but found that both perform comparably (Section 4.6).

3.5 Embedding Aggregation

We take a simple approach to produce embeddings over different timescales and modalities. To produce a summary vector z_t for each timestep t, we simply take the average over the embeddings from available modalities z_t^i. To produce a single embedding for a time range $[t_1, t_2]$, we simply average embeddings over that time range. For example, to produce an embedding for a 30-second interval for sleep stage classification, we average embeddings z_t over that interval. We leave the design of more effective embedding aggregation strategies to future work.

3.6 Evaluation set-up

In our main experiments, we compare Hypnos with OSF [53], SleepFM [58], and sleep2vec [69]. We re-train SleepFM using the open-source code and re-implement sleep2vec from the paper description. Both re-trained models use the same data splits, modalities and pre-processing as Hypnos. We directly use the public OSF model weights. Additional compute-matched comparisons with re-trained OSF, and a comparison with SleepMaMi [41], are given in Sections B.9 and B.10. Further implementation details are given in Section A.4.

To simplify comparison across foundation models, which produce embeddings at different timescales, we formulate sleep analysis tasks as classification over 30-second windows of data, following [53].

For example, a 30-second window is marked as 'positive' for apnoea if it has any overlap with an apnoea event. Recordings are chunked to match the context length of each model. Using Hypnos, we can generate embeddings for an arbitrary length recording in a single forward pass.

For each model, we use all supported modalities as inputs. All probes are trained using only the training and validation splits of the pre-training datasets, such that MrOS, DOD-H and DOD-O serve as fully external test sets to evaluate generalisation. For our supervised baselines, we use the SLEEPYLAND framework [50] to train and evaluate SleepTransformer [43] and U-Sleep [42] on the same dataset splits. Additional details of our evaluation set-up are described in Section A.4.

4 Experiments

4.1 Comparison with existing foundation models

In Table 3, we compare Hypnos with existing foundation models across common sleep analysis tasks using four clinically motivated modality configurations: full PSG ($M = 8$), single-channel EEG ($M = 1$), EOG ($M = 2$), and cardio-respiratory signals (ECG, ABD and THX; $M = 3$). Additional information on the evaluation tasks is provided in Section A.4. We report performance using mean AUROC and AUPRC values over recordings. To assess statistical significance, we perform a Wilcoxon signed-rank test over recordings with Benjamini-Hochberg FDR correction ($q < 0.05$). The same procedure is used for all comparisons throughout this section. Across sensor configurations, Hypnos consistently achieves better performance across downstream tasks.

Table 3: **Foundation model comparison across tasks and sensor configurations.** MLP probe results on held-out MrOS, under full-modality (M=8) and restricted-modality configurations. AUROC / AUPRC, mean over subjects; staging AUROC/AUPRC are macro-averaged over the five sleep stages. **Best** per metric in bold; * indicates Hypnos is significantly better.

		Staging		Arousal		Apnoea		Desat.	
Setting	Method	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Full (M=8)	SleepFM	0.952	0.718	0.883	0.698	0.777	0.396	0.731	0.714
	sleep2vec	0.960	0.741	0.886	0.704	0.754	0.382	0.736	0.712
	OSF	0.960	0.747	0.922	0.777	0.775	0.392	0.737	0.715
	Hypnos	**0.973***	**0.796***	**0.934***	**0.808***	**0.795***	**0.418***	**0.767***	**0.755***
EEG-C3 (M=1)	SleepFM	0.925	0.650	0.849	0.635	0.686	0.305	0.621	0.609
	sleep2vec	0.946	0.707	0.834	0.610	0.693	0.321	0.636	0.624
	OSF	0.937	0.674	0.906	0.749	0.659	0.291	**0.643**	0.620
	Hypnos	**0.969***	**0.779***	**0.922***	**0.790***	**0.723***	**0.351***	0.635	**0.628***
EOG (M=2)	SleepFM	0.916	0.638	0.824	0.593	0.681	0.303	0.654	0.627
	sleep2vec	0.933	0.671	0.804	0.545	0.667	0.304	0.653	0.631
	OSF	0.940	0.691	0.907	0.747	0.688	0.309	0.651	0.630
	Hypnos	**0.962***	**0.752***	**0.912***	**0.766***	**0.717***	**0.335***	**0.666***	**0.645***
ECG+ABD+THX (M=3)	SleepFM	0.849	0.512	0.798	0.544	0.753	0.373	0.737	0.714
	sleep2vec	0.909	0.618	0.830	0.607	0.748	0.378	0.737	0.718
	OSF	0.870	0.545	0.849	0.632	0.762	0.379	0.742	0.722
	Hypnos	**0.931***	**0.659***	**0.878***	**0.684***	**0.797***	**0.417***	**0.795***	**0.777***

4.2 Comparison with supervised sleep stage classification models

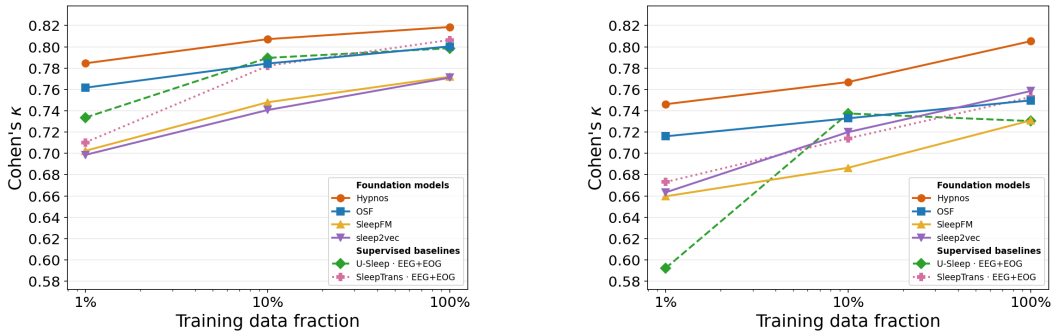
In Table 4, we compare the performance of Hypnos against strong supervised baselines on the task of sleep stage classification, evaluating against AASM (5-class) expert-annotated sleep stages: Wake, N1, N2, N3 and REM. We report Cohen's κ (the most common metric in the automated sleep staging literature [43]), macro-averaged AUROC and macro-averaged AUPRC, for one representative in-domain cohort (SHHS) and one held-out cohort (MrOS). Additional per-dataset results across eight cohorts (SHHS, CCSHS, CFS, NCHSDB, MrOS, DOD-H, DOD-O, MESA), including MLP probe results, are reported in the 100% column of Tables 12 and 13. Using only a linear probe, Hypnos outperforms strong supervised baselines on every metric on both SHHS and MrOS, and (as detailed in Section B.6) outperforms both baseline models on the majority of datasets and metrics.

Table 4: **Sleep stage classification performance.** Cohen's κ, macro-averaged AUROC and macro-averaged AUPRC (per-subject mean) on one in-domain (SHHS) and one held-out (MrOS) cohort. **Best** per (cohort, metric) in bold; * indicates Hypnos is significantly better than every other foundation model (paired Wilcoxon, one-sided, per-subject; Benjamini–Hochberg FDR-corrected $q < 0.05$).

	SHHS (in-domain)			MrOS (held-out)		
Method	κ	AUROC	AUPRC	κ	AUROC	AUPRC
Supervised						
U-Sleep [42]	0.799	0.969	0.825	0.730	0.949	0.748
SleepTransformer [43]	0.806	0.974	0.832	0.753	0.958	0.758
Self-supervised + linear probe						
SleepFM [58]	0.716	0.941	0.731	0.658	0.919	0.659
sleep2vec [69]	0.763	0.964	0.792	0.751	0.958	0.737
OSF [53]	0.791	0.968	0.809	0.740	0.957	0.738
Hypnos	**0.811***	**0.976***	**0.836***	**0.799***	**0.973***	**0.794***

4.3 Few-shot learning

Next, we evaluate model performance in the few-shot learning regime, scaling the proportion of labelled recordings used for training and validation from 1% up to 100%. Figure 6 reports sleep staging performance for an in-domain (SHHS) and a held-out (MrOS) dataset, alongside supervised baselines re-trained on the same data fractions. Hypnos outperforms existing foundation models and supervised baselines across all proportions of labelled data; on MrOS, Hypnos trained on as little as 1% of recordings matches U-Sleep trained on the full dataset.



(a) SHHS (in-domain)

(b) MrOS (held-out)

Figure 6: **Few-shot sleep stage classification.** We train MLP probes on each foundation model and re-train supervised baselines using varying fractions of in-domain data. Using as little as 1% of the probe-training data, Hypnos matches U-Sleep trained on the full dataset on held-out MrOS.

4.4 Transfer to external ECG benchmarks

To assess whether Hypnos's representations transfer beyond PSG recordings, we also evaluated performance on three external single-lead ECG benchmarks: PhysioNet/CinC 2017 (atrial fibrillation detection), Apnea-ECG (overnight per-minute apnoea detection) and CPSC 2021 (paroxysmal AF detection). We compare against xECG [33], a foundation model pre-trained on 12-lead clinical ECG. In Table 5, we see that Hypnos matches xECG on atrial fibrillation (AF) detection using CinC 2017, and beats it by 4 and 5 percentage points on Apnea-ECG and CPSC 2021 respectively, indicating the model effectively generalises beyond PSG recordings.

Table 5: **External ECG benchmarks.** Frozen-encoder linear-probe AUROC; full details in Section B.7.

Dataset	Hypnos	xECG
CinC 2017	**0.984**	**0.985**
Apnea-ECG	**0.925**	0.884
CPSC 2021	**0.985**	0.934

4.5 Generative modelling

Although the focus of our work is representation learning, the next-token prediction objective makes Hypnos fully generative: for any subset of supported modalities, tokens can be auto-regressively sampled and decoded back to waveforms via the tokenizers. Figure 7 shows that, conditioned on real multi-modal context, Hypnos produces coherent continuations that preserve waveform morphology and cross-modal structure, indicating that next-token prediction captures the joint distribution of the underlying signals. Additional generated ECG examples during atrial fibrillation and regular rhythm are shown in Figure 15 in the Appendix.

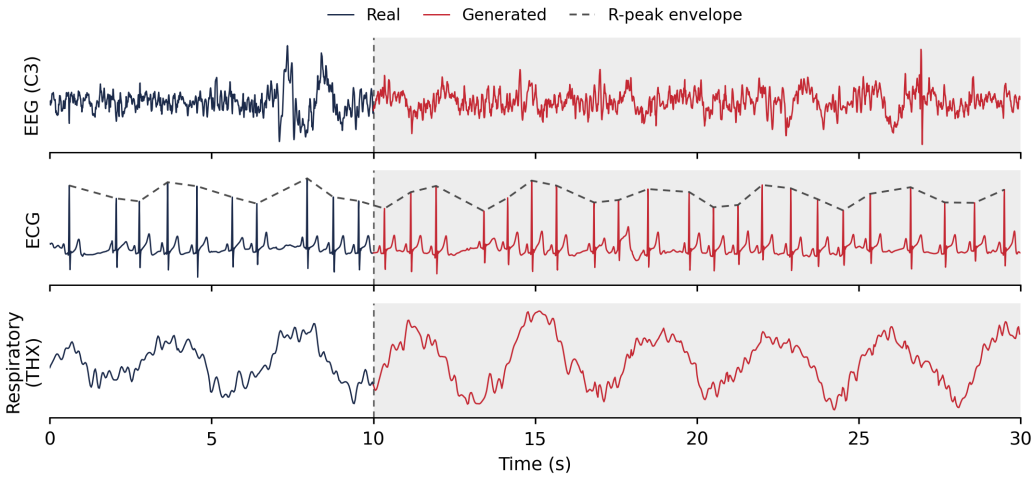


Figure 7: **Autoregressive generation of physiological signals.** Hypnos can be used to jointly generate physiological signals for any subset of supported modalities. Here we see that conditioned on 10 s of real context (blue), Hypnos generates plausible signals with cross-modal consistency. For example, we can observe respiration-induced amplitude modulation of R-peaks in the ECG.

4.6 Modality-masking

We ablate our group masking approach (*Default*, $\alpha = 1$) against three alternatives: *No masking* ($\alpha \to 0$), where modalities are processed with full cross-modal attention during training; *Independent* ($\alpha \to \infty$), where each modality is processed with no cross-modal attention; and *Random* where modalities are randomly masked out with probability $p = 0.5$. Table 6 reports linear-probe performance on full (M=8) and restricted-modality (*Restr.*) inputs, with full results in Table 11. The *Independent* variant lags across every configuration, highlighting the benefit of cross-modal fusion. Training-time masking improves restricted-modality robustness over *No masking* whilst maintaining full-modality performance. Surprisingly, both masking strategies perform comparably, despite a roughly $20\times$ difference in exposure to single-channel sequences during training, indicating that missing-modality robustness is insensitive to the specific masking strategy.

4.7 Model scaling

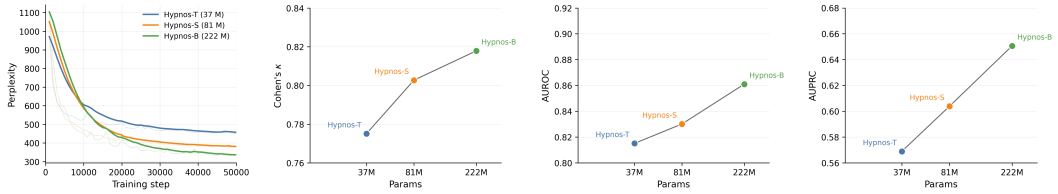
We investigated the effect of model scaling from Tiny to Base configurations. Further scaling experiments, including to larger model sizes and context lengths of over an hour ($T = 4096$) using unimodal model variants, can be found in Section B.2. In Figure 8, we observe monotonic improvements in both next-token perplexity and performance on downstream tasks as we increase model size.

5 Limitations and Future Work

Latent-space objectives Our results suggest that a key ingredient may be prediction in a compressed latent space rather than in input space: masked-token prediction over the same tokens performs

Table 6: **Modality-masking ablation.** Per-subject mean of the primary metric per task. *Restr.* is the mean across three restricted-modality configurations (EEG-C3, EOG, ECG+ABD+THX); *Full* uses all eight modalities. Full results are reported in Table 11.

	Staging (κ)		Arousal (AUROC)		Apnoea (AUROC)		Desat. (AUROC)	
Variant	Full	Restr.	Full	Restr.	Full	Restr.	Full	Restr.
No training-time masking								
Independent ($\alpha \to \infty$)	0.773	0.656	0.866	0.838	0.847	0.798	0.817	0.776
No masking ($\alpha \to 0$)	0.792	0.659	0.900	0.864	0.861	0.802	0.830	0.769
Training-time modality masking								
Random ($p = 0.5$)	0.795	0.672	0.902	0.873	0.860	0.813	0.831	0.796
Default ($\alpha = 1$)	0.797	0.676	0.901	0.874	0.859	0.814	0.831	0.798



(a) Validation loss (b) Sleep staging (c) Apnoea detection (d) Arousal detection

Figure 8: **Scaling model size improves the performance of Hypnos.** Next-token perplexity and downstream metrics all improve with model scale. Sleep stage classification, apnoea detection and arousal detection performance are reported using a linear probe on the SHHS validation set.

similarly to next-token prediction, while input-space MAE is weaker (Section B.8). We expect other latent-space approaches, such as JEPA [4] and latent diffusion [51], to also work well, and comparing them at scale is an important direction for future work.

Long-context learning Many clinically meaningful properties of physiological signals are only visible across hours or days of sensor data, including circadian phase, multi-night sleep regularity, and the clustering of rare events such as nocturnal seizures or periodic limb movements. Recent work has begun to reach these timescales by first compressing short segments into embeddings, then modelling sequences of those embeddings over a full night [41], a full day [63], or multiple days [9]. Combining this kind of hierarchical compression with next-token prediction could extend Hypnos to days of continuous multi-modal data while retaining its generative and streaming capabilities.

Further discussion of limitations and future directions, including clinical outcomes and biomarker discovery, is provided in Section C.

6 Conclusions

We have presented Hypnos, a multi-modal sleep foundation model trained with next-token prediction over residual vector quantised (RVQ) tokens drawn from eight physiological sensing modalities. Hypnos can be applied to real-time continuous streams of physiological sensor data from modalities such as EEG and ECG signals, generating high-quality embeddings for a range of downstream tasks. Our model outperforms existing foundation models and strong supervised baselines on most tasks and datasets evaluated, across diverse physiological sensing tasks such as sleep staging, atrial fibrillation and apnoea detection. Hypnos is one step toward a broader goal: foundation models that compress hours to days of multi-modal physiological signals into better measures of human health. That this simple recipe scales with model size, transfers to downstream tasks, generalises to held-out datasets, and is robust to sensor configurations suggests it can extend beyond sleep.

Acknowledgements

Funding from ARIA and DSIT and Pillar VC under the Encode: AI for Science Fellowship. JC thanks Will Bolton, Botos Csaba and Bryan Li for their feedback on drafts of the paper. We kindly thank the National Sleep Research Resource (NSRR) for providing access to the datasets used. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). Additional acknowledgements for the datasets used in this work are listed in Section E.

References

[1] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale Training of Foundation Models for Wearable Biosignals. In *The Twelfth International Conference on Learning Representations*, March 2024. doi: 10.48550/arXiv.2312.05409. 1

[2] David J. Aldous. Exchangeability and related topics. In David J. Aldous, Illdar A. Ibragimov, Jean Jacod, and P. L. Hennequin, editors, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg, 1985. Springer. ISBN 978-3-540-39316-0. doi: 10.1007/BFb0099421. 6

[3] Thomas Andrillon, Yuval Nir, Richard J. Staba, Fabio Ferrarelli, Chiara Cirelli, Giulio Tononi, and Itzhak Fried. Sleep Spindles in Humans: Insights from Intracranial EEG and Unit Recordings. *The Journal of Neuroscience*, 31(49):17821–17834, December 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2604-11.2011. 20

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 10, 18

[5] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. 1

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. 5

[7] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, and Marco Tagliasacchi. Audiolm: A language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533, 2023. 1, 2

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. 1, 5, 17

[9] Peng Cao, Zhijian Yang, Tennison Liu, Jonathan Wang, Jiang Wu, Magdalena Proszewska, Arvind Pillai, Mingwu Gao, Amir Farjadian, Lawrence Cai, Emily Blanchard, Daniel McDuff, Pramod Rudrapatna, Matthew Thompson, Anupam Pathak, Mark Malhotra, Shwetak Patel, Dina Katabi, Paolo Di Achille, and Ming-Zher Poh. WavesFM: Hierarchical Representation Learning for Longitudinal Wearable Sensor Waveforms, May 2026. 10

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. 18

[11] Jonathan F. Carter and Lionel Tarassenko. Wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals. In *Proceedings of the 4th Machine Learning for Health Symposium*, pages 186–202. PMLR, February 2025. 6

[12] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, June 2015. ISSN 1550-9109. doi: 10.5665/sleep.4732. 4

[13] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.01549. 3

[14] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and Controllable Music Generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, December 2023. 2

[15] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability Results for Multimodal Contrastive Learning, March 2023. 3

[16] Shaun Davidson, Rachel Sharman, Simon D. Kyle, and Lionel Tarassenko. Is it time to revisit the scoring of slow wave (N3) sleep? *Sleep*, 48(10), October 2025. ISSN 0161-8105. doi: 10.1093/sleep/zsaf063. 29

[17] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue, October 2024. 1, 3, 4, 5, 16, 21

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020. 6

[19] Benjamin Fox, Joy Jiang, Sajila Wickramaratne, Patricia Kovatch, Mayte Suarez-Farinas, Neomi A Shah, Ankit Parekh, and Girish N Nadkarni. A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. *Sleep*, 48(8):zsaf061, August 2025. ISSN 0161-8105. doi: 10.1093/sleep/zsaf061. 1, 3

[20] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://arxiv.org/abs/2403.05530v5, March 2024. 1

[21] Antoine Guillot, Fabien Sauvet, Emmanuel H. During, and Valentin Thorey. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(9):1955–1965, September 2020. ISSN 1558-0210. doi: 10.1109/TNSRE.2020.3011181. 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*. arXiv, December 2020. doi: 10.48550/arXiv.2006.11239. 3

[23] C. Iber. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification. 2007. 3

[24] Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The Brain's Bitter Lesson: Scaling Speech Decoding With Self-Supervised Learning. In *Proceedings of the 42nd International Conference on Machine Learning*, June 2025. 3

[25] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations (ICLR 2024)*, May 2024. doi: 10.48550/arXiv.2405.18765. 2

[26] Weibang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals. In *The Thirteenth International Conference on Learning Representations*, October 2024. 2

[27] Biing-Hwang Juang and A. Gray. Multiple stage vector quantization for speech coding. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 597–600, May 1982. doi: 10.1109/ICASSP.1982.1171604. 4

[28] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients, May 2021. 3

[29] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive Image Generation using Residual Quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*. arXiv, March 2022. doi: 10.48550/arXiv.2203.01941. 2, 5

[30] Harlin Lee, Boyue Li, Shelly DeForte, Mark L. Splaingard, Yungui Huang, Yuejie Chi, and Simon L. Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):421, July 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01545-6. 4

[31] Simon A. Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Minji Han, Rachel Choi, Md Sazzad Hissain Khan, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, Viswam Nathan, Md Mahbubur Rahman, Li Zhu, Subramaniam Venkatraman, and Sharanya Arcot Desai. HiMAE: Hierarchical Masked Autoencoders Discover Resolution-Specific Structure in Wearable Time Series. In *The Fourteenth International Conference on Learning Representations*, October 2025. doi: 10.48550/arXiv.2510.25785. 1, 3

[32] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 5

[33] Riccardo Lunelli, Angus Nicolson, Samuel Martin Pröll, Sebastian Johannes Reinstadler, Axel Bauer, and Clemens Dlaska. BenchECG and xECG: A benchmark and baseline for ECG foundation models, September 2025. 8, 27

[34] Carole L. Marcus, Reneé H. Moore, Carol L. Rosen, Bruno Giordani, Susan L. Garetz, H. Gerry Taylor, Ron B. Mitchell, Raouf Amin, Eliot S. Katz, Raanan Arens, Shalini Paruthi, Hiren Muzumdar, David Gozal, Nina Hattiangadi Thomas, Janice Ware, Dean Beebe, Karen Snyder, Lisa Elden, Robert C. Sprecher, Paul Willging, Dwight Jones, John P. Bent, Timothy Hoban, Ronald D. Chervin, Susan S. Ellenberg, Susan Redline, and Childhood Adenotonsillectomy Trial (CHAT). A randomized trial of adenotonsillectomy for childhood sleep apnea. *The New England Journal of Medicine*, 368(25):2366–2376, June 2013. ISSN 1533-4406. doi: 10.1056/NEJMoa1215881. 4

[35] Julieta Martinez, Holger H. Hoos, and James J. Little. Stacked Quantizers for Compositional Vector Compression, November 2014. 4

[36] Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An Open Electrocardiogram Foundation Model, May 2025. 2

[37] P.E. McSharry, G.D. Clifford, L. Tarassenko, and L.A. Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, March 2003. ISSN 1558-2531. doi: 10.1109/TBME.2003.808805. 2

[38] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for Polyphonic Sound Event Detection. *Applied Sciences*, 6(6):162, June 2016. ISSN 2076-3417. doi: 10.3390/app6060162. 29

[39] George B. Moody and Roger G. Mark. A new method for detecting atrial fibrillation using R-R intervals. In *Computers in Cardiology*, volume 10, pages 227–230, 1983. 29

[40] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling Wearable Foundation Models. In *The Thirteenth International Conference on Learning Representations*, October 2024. 1, 2

[41] Keondo Park, Younghoon Na, Yourim Choi, Hyunwoo Ryu, Hyun-Woo Shin, and Hyung-Sin Kim. SleepMaMi: A Universal Sleep Foundation Model for Integrating Macro- and Micro-structures, June 2026. 2, 3, 6, 10, 28

[42] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, 4(1):72, April 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00440-5. 7, 8, 25, 26

[43] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, August 2022. ISSN 1558-2531. doi: 10.1109/TBME.2022.3147187. 7, 8, 25, 26

[44] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open Foundation Models for Optical Physiological Signals, February 2025. 3

[45] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep*, 20(12):1077–1085, December 1997. ISSN 0161-8105. 4

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[48] S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci. The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1):682–687, March 1995. ISSN 1073-449X. doi: 10.1164/ajrccm/151.3_Pt_1.682. 4

[49] Carol L. Rosen, Emma K. Larkin, H. Lester Kirchner, Judith L. Emancipator, Sarah F. Bivins, Susan A. Surovec, Richard J. Martin, and Susan Redline. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: Association with race and prematurity. *The Journal of Pediatrics*, 142(4): 383–389, April 2003. ISSN 0022-3476. doi: 10.1067/mpd.2003.28. 4

[50] Alvise Dei Rossi, Matteo Metaldi, Michal Bechny, Irina Filchenko, Julia van der Meer, Markus H. Schmidt, Claudio L. A. Bassetti, Athina Tzovara, Francesca D. Faraci, and Luigi Fiorillo. SLEEPYLAND: Trust begins with fair evaluation of automatic sleep staging models. *npj Digital Medicine*, 9(1):55, December 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02237-2. 7

[51] Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. Continuous Audio Language Models. In *The Fourteenth International Conference on Learning Representations*, January 2026. doi: 10.48550/arXiv.2509.06926. 3, 10

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs, June 2016. 21

[53] Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On Pre-training and Scaling of Sleep Foundation Models. In *Proceedings of the 43rd International Conference on Machine Learning*, February 2026. doi: 10.48550/arXiv.2603.00190. 2, 4, 6, 8, 25, 26

[54] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, November 2015. 3

[55] Yeonsu Song, Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Katie L. Stone, and Osteoporotic Fractures in Men (MrOS) Study Group. Relationships between sleep stages and changes in cognitive function in older men: The MrOS Sleep Study. *Sleep*, 38(3):411–421, March 2015. ISSN 1550-9109. doi: 10.5665/sleep.4500. 4

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. 17

[57] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. SEANet: A Multi-modal Speech Enhancement Network, October 2020. 4

[58] Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2):752–762, February 2026. ISSN 1546-170X. doi: 10.1038/s41591-025-04133-4. 1, 2, 3, 6, 8, 17, 25, 26, 29

[59] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, September 2016. 4

[60] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. 4, 5

[62] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024. 3

[63] Yong Wang, Dylan McGagh, Katya Broomberg, Zizheng Zhang, Jonathan Carter, Junayed Naushad, Laura Brocklebank, Yang Sun, George Nicholson, Dianjianyi Sun, Canqing Yu, Jun Lv, Maxim Barnard, Hubert Lam, Andrew Steptoe, David W. Eyre, Liming Li, Zhengming Chen, Naomi Wray, Spiros Denaxas, Gary S. Collins, Huaidong Du, Aiden Doherty, and Hang Yuan. Learning Human Health and Diseases from 24-hour Wrist Movement, August 2026. 10

[64] Simon C. Warby, Sabrina L. Wendt, Peter Welinder, Emil G. S. Munk, Oscar Carrillo, Helge B. D. Sorensen, Poul Jennum, Paul E. Peppard, Pietro Perona, and Emmanuel Mignot. Sleep-spindle detection: Crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, 11(4):385–392, April 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2855. 29

[65] Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A Brain Foundation Model for Unified EEG and MEG Signals. In *Advances in Neural Information Processing Systems*, volume 38, October 2025. doi: 10.48550/arXiv.2505.18185. 2, 4, 5, 17, 19

[66] Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. LSM-2: Learning from Incomplete Wearable Sensor Data, June 2025. 6

[67] Terry Young, Mari Palta, Jerome Dempsey, Paul E. Peppard, F. Javier Nieto, and K. Mae Hla. 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):246–249, August 2009. ISSN 1098-1861. 4

[68] Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 7(1):1–10, April 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01062-3. 1, 3

[69] Weixuan Yuan, Zengrui Jin, Yichen Wang, Donglin Xie, Ziyi Ye, Chao Zhang, and Xuesong Chen. Sleep2vec: Unified Cross-Modal Alignment for Heterogeneous Nocturnal Biosignals. https://arxiv.org/abs/2602.13857v1, February 2026. 6, 8, 17, 25, 26

[70] Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. Brant-X: A Unified Physiological Signal Alignment Framework. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4155–4166, Barcelona Spain, August 2024. ACM. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671953. 3

[71] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064. 4

A Additional Implementation Details

A.1 Preprocessing

Referencing and filtering EEG and EOG channels were re-referenced against the contralateral mastoid (C3–M2, C4–M1 for EEG; E1–M2, E2–M1 for EOG). Chin EMG was derived bipolarly from the chin electrode pair. ECG and respiratory effort (ABD, THX) were used directly. All signals were resampled using a polyphase filter with an anti-aliasing low-pass. Brain, muscle, and cardiac channels were resampled to 128 Hz. Respiratory effort signals (ABD, THX) were resampled to 32 Hz, reflecting their lower frequency content. All channels were notch-filtered to suppress mains interference. Per-modality bandpass filters were then applied:

- EEG, EOG, EMG: 0.5–45 Hz bandpass.
- ECG: 0.05–45 Hz bandpass.
- ABD, THX: 0.05 Hz high-pass.

Normalisation and amplitude compression For each channel of each recording we performed an online rolling z-score normalisation: the mean and variance were tracked online by a 60-second exponential moving average. To prevent transient artefacts from inflating the running variance, the per-sample squared deviation was clipped at 6σ of the current variance estimate before accumulation. After normalisation, values lying beyond $\pm 8\sigma$ were log-compressed, so that large artefacts remain order-preserving without dominating the reconstruction loss.

A.2 Tokenizer training details

Architecture Each tokenizer consists of a SeaNet and Transformer components plus an RVQ bottleneck, as illustrated in Figure 3. SeaNet stride ratios are chosen so that the product of strides corresponds to one second of input samples, producing tokens at 1 Hz at every modality's sampling rate. The encoder right-pads by one hop length and drops the first warm-up token, so each output token is right-aligned to its segment boundary. All other architecture and optimisation hyperparameters are listed in Table 7; weight decay is applied to Transformer parameters only, following the Mimi convention [17].

Table 7: **Tokenizer hyperparameters.**

Component	Setting
SeaNet `n_filters`	64
SeaNet stride ratios	$[2, 4, 4, 4]$ at 128 Hz; $[4, 4, 2]$ at 32 Hz
SeaNet residual blocks	1 per stage, dilation base 2
Encoder embedding dim	512
Codebook entry dim	256
Codebook size C	2048
Residual levels K	8 (EEG, EOG, EMG); 4 (ECG, respiratory)
EMA decay	0.99
Quantization dropout	0.5
Encoder/decoder Transformer	4 layers, 8 heads, FFN dim 2048, sliding window 32
LayerScale init	10^{-2}
Loss weights	$\lambda_\phi = 0.5$, $\lambda_{\text{RVQ}} = 0.25$
Optimiser	AdamW, lr 2×10^{-4}, wd 10^{-2} (Tx params)
Gradient clipping	1.0 (norm)
Schedule	Cosine, 500-step linear warm-up
Training steps	50,000
Batch size $\times$ window	1024×64 s
Precision	bf16 mixed

Loss formulation Following BrainTokenizer [65], the tokenizer is trained with a multi-term reconstruction loss combined with an RVQ commitment penalty. The reconstruction loss has (i) a time-domain term between the original waveform X and the reconstruction $\hat{X}$:

$$\mathcal{L}_{\text{time}} \;=\; \|X - \hat{X}\|_1, \tag{2}$$

where $\|\cdot\|_1$ denotes the ℓ_1 distance; and (ii) a frequency-domain term between the amplitude spectra A, $\hat{A}$ and phase spectra Φ, $\hat{\Phi}$ of the Hamming-windowed signals:

$$\mathcal{L}_{\text{freq}} \;=\; \|A - \hat{A}\|_1 + \lambda_\phi \cdot \|\Phi - \hat{\Phi}\|_1. \tag{3}$$

Letting z_k and z_{q_k} denote the residual entering the k-th codebook and its nearest entry, the RVQ commitment loss is:

$$\mathcal{L}_{\text{rvq}} \;=\; \sum_{k=1}^{K} \big\| z_k - \text{sg}[z_{q_k}] \big\|_2^2, \tag{4}$$

where $\text{sg}[\cdot]$ denotes a stop-gradient. This term pulls the encoder output z_k toward its assigned codebook entry. Each tokenizer is trained to minimise:

$$\mathcal{L}_{\text{tok}} \;=\; \mathcal{L}_{\text{time}} + \mathcal{L}_{\text{freq}} + \lambda_{RVQ} \cdot \mathcal{L}_{\text{rvq}}, \tag{5}$$

with scalar weights $\lambda_\phi = 0.5$ and $\lambda_{RVQ} = 0.25$, identical to BrainTokenizer [65].

A.3 Hypnos hyperparameters and training

The temporal Transformer applies rotary position embeddings (RoPE, [56]) to attention queries and keys, while the depth Transformer uses a learnt position embedding indexed by codebook level. Linear and embedding parameters are initialised from $\mathcal{N}(0, 0.02^2)$ following GPT-2 [8]. We use a residual dropout of 0.1. Per-scale learning rates are set inversely to hidden width: 1.2×10^{-3} for Hypnos-Tiny, 6×10^{-4} for Hypnos-Small, and 3×10^{-4} for Hypnos-Base. All other training hyperparameters are listed in Table 8.

A.4 Comparing sleep foundation models

Baseline implementation SleepFM [58] and sleep2vec [69] were re-implemented and re-trained using the same modalities, pre-processing steps, and dataset splits as Hypnos, which match the dataset splits used by OSF. We used the 'Large' configuration of sleep2vec, which has around 240 million

Table 8: **Hypnos training hyperparameters.**

Component	Setting
Norm / activation	LayerNorm + SwiGLU; RMSNorm on QK
Position encoding	RoPE (temporal); learnt level-index embedding (depth)
Sliding-window pattern	Local window 64; every 4th layer global ($T/2$)
CRP concentration	$\alpha = 1.0$
Dropout	0.1
LayerScale init	10^{-2}
Weight init	$\mathcal{N}(0, 0.02^2)$ for Linear / Embedding
Optimiser	AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$, wd 0.1
Gradient clipping	1.0 (norm)
Per-scale lr	Tiny 1.2×10^{-3}; Small 6×10^{-4}; Base 3×10^{-4}
Schedule	Cosine, 500-step linear warm-up
Training steps	50,000
Batch size	512
Context length	512
Precision	bf16 mixed, activation checkpointing

parameters, including a slightly larger Transformer backbone than Hypnos-Base. For completeness, in Table 9 we additionally report performance using the open-source SleepFM checkpoint, evaluated using all supported modalities and the original pre-processing steps. We adopt our re-trained variant as the stronger baseline throughout.

Table 9: **Off-the-shelf vs. re-trained SleepFM.** Performance on held-out MrOS with an MLP probe.

	Staging		Arousal		Apnoea		Desat.	
SleepFM	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Open weights	0.928	0.657	0.771	0.465	0.729	0.349	0.694	0.679
Re-trained	**0.952**	**0.718**	**0.883**	**0.698**	**0.777**	**0.396**	**0.731**	**0.714**

Evaluation set-up We believe an important property of a 'foundation model' is effective 'out-of-the-box' embeddings, enabling strong downstream performance with limited task-specific supervision, and downstream multimodal fusion with other information streams, e.g. electronic healthcare records. In computer vision, representation quality is commonly evaluated with linear or MLP probes, e.g. [4, 10]. However, existing sleep foundation models have varied drastically in their evaluation set-up.

We aimed to standardise the use of embeddings to perform a fair comparison across different foundation models. We primarily report linear and MLP probing performance, applying simple temporal and modality pooling for each model where necessary to evaluate at 30-second resolution on downstream tasks. We believe this most closely matches evaluations in domains such as computer vision, e.g. mean-pooled embeddings from image patches. For OSF, which uses 30-second windows of input data, we use the dedicated CLS token for downstream probing experiments. For SleepFM and sleep2vec, which generate embeddings for each modality at different temporal resolutions, we first mean-pooled in the temporal dimension to obtain embeddings for each 30-second window. To combine embeddings from each modality, we evaluated both mean-pooling and concatenation and found the latter to work best across baselines.

Sleep analysis tasks We primarily compare performance on the following tasks:

- **Staging**: Sleep stage classification into one of five classes: Wake, N1, N2, N3 (deep) or rapid-eye-movement (REM) sleep.
- **Arousal**: Detection of cortical arousal events, characterised by an abrupt shift to wakefulness-like EEG activity. A high rate of cortical arousals during sleep correlates with excessive daytime sleepiness, impaired vigilance/cognition, and reduced quality of life.

- **Apnoea**: Detection of apnoea events, i.e. a temporary cessation in breathing. All apnoeas (central, obstructive, mixed) are aggregated into a single class for the purpose of evaluation.
- **Desat.**: Detection of blood oxygen desaturation events (transient drops in SpO_2).

Annotations for each task come from the NSRR-harmonised scoring provided alongside each overnight recording.

Probe training Each foundation model was evaluated with an identical probing pipeline. Before fitting any probe, we standardised each embedding dimension to zero mean and unit variance, with the scaling statistics computed on the training split only and then applied to the validation and test splits. The *linear probe* is a single linear layer (multinomial logistic regression), trained to minimise cross-entropy with AdamW (learning rate 10^{-3}, weight decay 10^{-4}) for 30 epochs. The *MLP probe* has two hidden layers of width 512 and 256 with ReLU activations and dropout ($p = 0.1$), and is trained to minimise cross-entropy with AdamW (learning rate 10^{-3}, weight decay 10^{-4}) for up to 10 epochs, with early stopping on a held-out validation split (patience 2). A separate probe is fit per task, and a fixed random seed is used throughout. We used the same probe configurations and hyper-parameters for all models reported. Probe hyper-parameters were chosen as sensible defaults and were not optimised.

A.5 Compute usage

Each tokenizer described in Section 3.2 was trained using a single NVIDIA H100 GPU using bf16 mixed-precision, requiring 60 GB of GPU RAM and around 5 hours of training time. After training, tokenization was performed using a single NVIDIA L40S GPU, with each entire overnight recording of each channel (10+ hours) tokenized in a single forward pass in around 250 ms. All Hypnos models were trained using H100 GPUs with bf16 mixed-precision training and activation checkpointing in each Transformer layer. Training Hypnos-Base to 50k steps required 1.5 days distributed across 8 GPUs, using around 45 GB of RAM on each GPU. For downstream probing evaluations, embeddings were generated using a single NVIDIA L40S GPU, with the embedding of each overnight recording taking around 3 seconds.

To reduce compute usage, many of our experiments including initial model design and ablation studies were performed using a unimodal variant of Hypnos-Small using EEG data, which took around 1 day to train on a single H100 GPU. Across all experiments our total compute usage was approximately 8000 H100 GPU-hours.

B Additional Experiments

B.1 Tokenizer design

Reconstruction quality Tokenizer model capacities were chosen to achieve near lossless compression before downstream sequence modelling. Figure 9 shows reconstructions from each of our five tokenizers on held-out MrOS recordings. Over eight random 64-second windows from each of 32 records, mean SNR is 9.1 dB for EEG (C3; 9.2 dB for C4), 10.2 dB for EOG (E1 and E2), 9.4 dB for chin EMG (28 records with this channel), 15.5 dB for ECG, and 27.2 dB and 29.0 dB for thoracic and abdominal respiration, with mean Pearson correlations from 0.91 to 0.99. Reconstruction is weakest for EEG, EOG and EMG, where small residuals remain visible.

Residual depth BrainOmni uses a codebook with $K = 4$ quantization layers, which leads to visibly smoothed reconstructions of EEG data (see Fig. 4 of [65]). Instead, we used $K = 8$ quantizers for neural signals to increase reconstruction accuracy of higher frequency details, e.g. gamma activity. Here we investigate the effect of residual depth on downstream performance. We vary the quantization depth at both the input and output to Hypnos, K_{in} and K_{out}, which determine the residual tokens available for sequence modelling (Figure 4a) and the residual tokens to be predicted (Figure 4c) respectively. Figure 10 shows the performance of unimodal Hypnos variants across downstream tasks using EEG data as we vary K_{in} and K_{out}. We generally observe improved performance as we increase K_{in}, but a decrease in performance as we increase K_{out}. This indicates that high-frequency information is useful for sequence modelling, but trying to predict high-frequency information does not improve the quality of the learnt representation on the tasks evaluated. This suggests that K_{out}

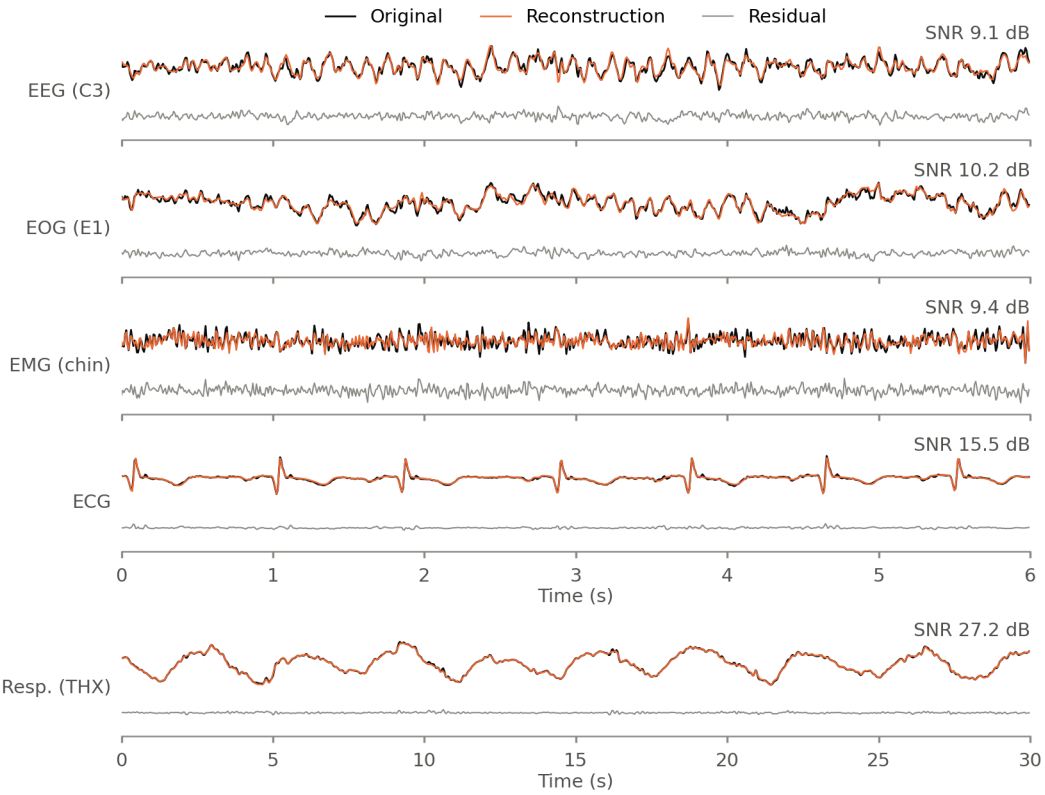


Figure 9: **Tokenizer reconstructions on held-out MrOS.** One time-aligned example per tokenizer, with the residual (original minus reconstruction) on the same scale below each trace. Respiration is shown over 30 s to span several breaths. The SNR on each row is the mean over held-out records.

could be decreased during training. Reducing K_{out} from 8 to 2 would reduce Depth Transformer FLOPs by 75%, and overall FLOPs by around 25% for Hypnos-Base. However, this would come at the expense of generative capabilities.

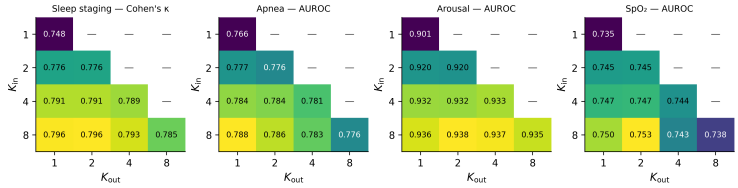
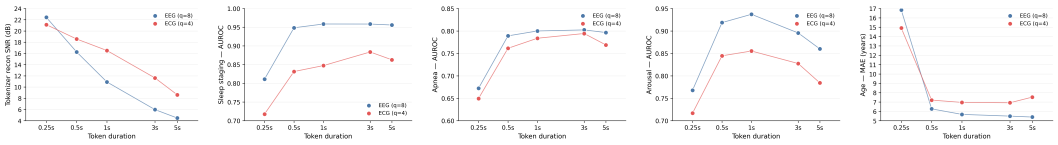


Figure 10: **Effect of input and output residual depth on downstream performance.** Linear probe metrics for unimodal EEG models on the SHHS validation set varying K_{in} and K_{out}. Performance slightly improves when increasing the number of input residual tokens but worsens when increasing the number of output residual tokens.

Tokenization length In our main experiments, we designed our tokenizers to produce tokens at a rate of 1 Hz. This also determines the rate at which unique output embeddings are produced by the model. 1 Hz is a natural choice for real-world applications, aligning with standard units. Additionally, relevant physiological events such as heartbeats and sleep spindles [3] commonly occur on this timescale. Here we investigate the sensitivity of our approach to the tokenization length of different signals. For EEG and ECG signals, we trained 5 tokenizers with tokenization lengths, $\tau \in (0.25, 0.5, 1.0, 3.0, 5.0)$ seconds. This was achieved by modifying the convolutional stride lengths in the convolutional components of the tokenizers. For 3-second and 5-second tokenization lengths, we also increased the input sequence length by 3x and 5x respectively, so that the input sequence length

to Transformer components in the tokenizers remained constant. We then re-tokenized all datasets and investigated the effect of tokenization length on downstream performance. Figure 11 shows both the reconstruction SNR and downstream performances as we vary the tokenization length. As expected, increasing tokenization length with fixed model capacity leads to a decrease in signal-to-noise ratio as the compression rate increases. Above $\tau = 0.25$ s, performance is reasonably stable across a range of token durations for both ECG and EEG data, with $\tau = 1$ s working well across tasks and inputs. Performance is noticeably poorer with $\tau = 0.25$ across tasks. For very small tokenization lengths τ, the auto-regression task reduces to trivial extrapolation over short timescales, empirically reducing the quality of the learnt representations.



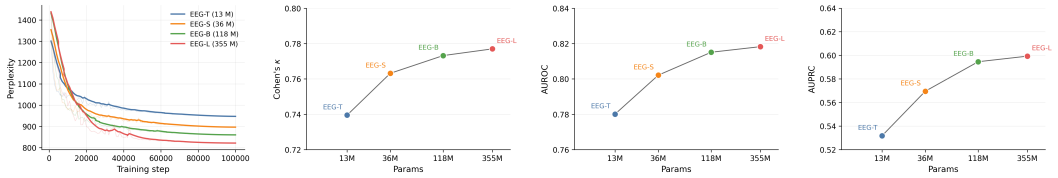
(a) Recon. SNR (b) Sleep staging (c) Apnoea detection (d) Arousal detection (e) Age regression

Figure 11: **Effect of token duration on downstream performance.** (left) Reconstruction quality decreases as token duration is varied from 0.25 s to 5 s, i.e. the compression rate increases. However, performance is worst at high token rates (0.25 s) and saturates or regresses beyond 1 s. We adopt a 1 s token duration in all other experiments.

Adversarial losses Défossez *et al.* [17] recently observed that removing reconstruction losses and solely relying on adversarial losses led to better performance in downstream audio modelling tasks. In early experiments, we tried incorporating adversarial losses but found that this did not have a significant effect on downstream performance. Additionally, it significantly increased computational requirements: each training run required substantially higher activation memory for the discriminator; and, more training runs were required to find stable hyper-parameters given the well-known instability issues in adversarial training [52].

B.2 Scaling unimodal EEG models

Compute requirements precluded scaling our multimodal model beyond Base size in our main experiments. To extend the scaling analysis to larger models, we trained unimodal EEG variants of Hypnos from Tiny up to Large. Figure 12 shows next-token perplexity and downstream probing performance on the SHHS validation set as we scale model size. Trends mirror those observed in the multimodal setting: validation loss and downstream metrics continue to improve with scale through to Large.



(a) Validation loss (b) Sleep staging (c) Apnoea detection (d) Arousal detection

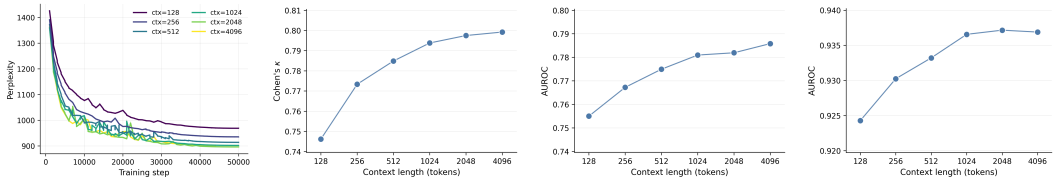
Figure 12: **Scaling unimodal EEG models from Tiny to Large.** Next-token perplexity and downstream metrics continue to improve with model scale. Sleep stage classification, apnoea detection and arousal detection performance are reported using a linear probe on the SHHS validation set.

B.3 Scaling context length

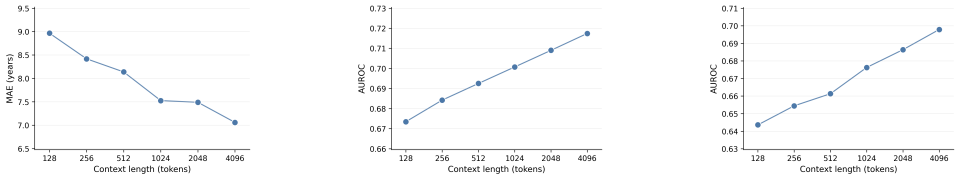
To quantify the effect of context length on Hypnos, we trained unimodal EEG and ECG variants of Hypnos-Small with context lengths $T \in \{128, 256, 512, 1024, 2048, 4096\}$ tokens, corresponding to roughly 2 minutes up to over an hour of data at 1 Hz. Models were trained for 50k steps, whilst all

other hyper-parameters were held fixed across runs. Figures 13 and 14 show validation perplexity and downstream probing performance on the SHHS validation set as we vary T.

For both modalities, validation perplexity improves monotonically with context length, and most downstream metrics continue to improve out to 4096 tokens. Summary tasks benefit most: age regression, CVD risk and obstructive sleep apnoea (OSA) classification all improve steadily across the full range. Meanwhile, sleep staging and arousal detection saturate earlier, at around 1024–2048 tokens. These trends are consistent across EEG and ECG, suggesting that the benefit of longer context is not specific to a single modality.

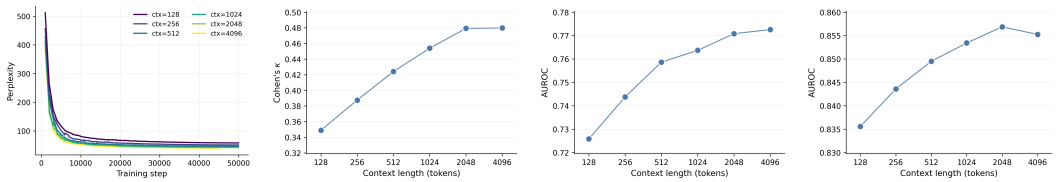


(a) Validation perplexity (b) Sleep staging (c) Apnoea detection (d) Arousal detection

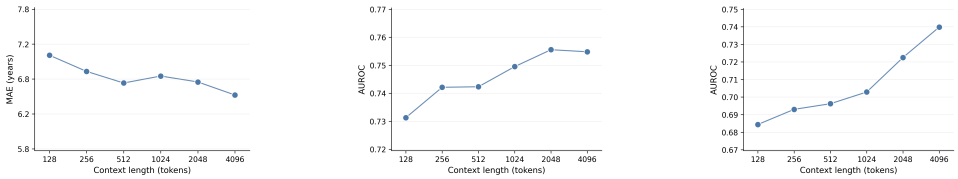


(e) Age regression (f) CVD risk (g) Moderate OSA

Figure 13: **Effect of context length on single-channel EEG models.** Validation perplexity decreases and downstream probing performance improves as the training context length is increased from 128 to 4096 tokens. Sleep staging and arousal detection saturate at around 1024–2048 tokens, while age regression, CVD risk and moderate OSA detection continue to improve at the longest context lengths.



(a) Validation perplexity (b) Sleep staging (c) Apnoea detection (d) Arousal detection



(e) Age regression (f) CVD risk (g) Moderate OSA

Figure 14: **Effect of context length on ECG-only models.** As with EEG, perplexity and downstream metrics improve with longer context. The largest relative gains are again on summary tasks such as moderate OSA detection and CVD risk.

B.4 Training stability

To reduce compute usage, our main pre-training experiments were performed with a single random seed (42). We observed stable and predictable training dynamics across our training runs including our model scaling experiments. To further assess variance due to neural network training dynamics, we re-trained Hypnos-Small on single-channel EEG using three different random seeds. In Table 10 we see that the variance in validation loss and probing metrics is low.

Table 10: **Training stability of Hypnos trained on single-channel EEG across three random seeds.** We report mean ± std and coefficient of variation (CV) across seeds. Downstream task performance is computed on validation data using an online linear probe.

Metric	Mean ± std	CV (%)
Validation perplexity (in-domain)	757.27 ± 0.50	0.066
Validation perplexity (MrOS held-out)	876.73 ± 0.68	0.078
Sleep staging κ (SHHS)	0.7930 ± 0.0004	0.050
Sleep staging κ (MrOS held-out)	0.7646 ± 0.0013	0.170
Apnoea AUROC (SHHS)	0.7825 ± 0.0014	0.179
Apnoea AUROC (MrOS held-out)	0.7812 ± 0.0095	1.216

B.5 Modality-masking ablation

Table 11 reports the full results of the modality-masking ablation summarised in Section 4.6, including the linear-probe head and the secondary metric per task. The overall ordering is consistent across both probes: *Independent* is the weakest variant, *No masking* closes most of the gap on *Full*-modality evaluation but degrades on restricted-modality inputs, and *Default* ($\alpha = 1$ group masking) and *Random* are comparable. All runs were trained on the same data splits using Hypnos-Small.

Table 11: **Modality-masking ablation — full results.** Linear and MLP probes on SHHS (in-domain). Per-subject mean. **Best** per (probe, subset, metric) column in bold (ties bolded both). Variants: *Independent* ($\alpha \to \infty$, each modality stream trained independently), *No masking* ($\alpha \to 0$, always-full attention), *Random* ($p = 0.5$ per-modality random masking, following prior work), *Default* (ours, Chinese Restaurant Process with $\alpha = 1$; Section 3.4).

		Staging			Arousal		Apnoea		Desat.	
Subset	Variant	κ	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Linear probe										
EEG-C3	Independent ($\alpha \to \infty$)	0.747	0.958	0.774	0.862	0.622	0.771	0.573	0.766	0.455
(M=1)	No masking ($\alpha \to 0$)	0.755	0.960	0.784	0.884	0.670	0.772	0.572	0.774	0.474
	Random ($p = 0.5$)	0.773	0.965	0.797	0.896	0.691	0.786	0.597	0.793	**0.508**
	Default ($\alpha = 1$)	**0.774**	**0.966**	**0.800**	**0.898**	**0.697**	**0.788**	**0.599**	**0.795**	0.507
EOG	Independent ($\alpha \to \infty$)	0.750	0.958	0.779	0.859	0.601	0.773	0.588	0.754	0.441
(M=2)	No masking ($\alpha \to 0$)	0.760	**0.962**	**0.791**	0.877	0.643	0.776	0.596	0.745	0.428
	Random ($p = 0.5$)	0.758	0.955	0.770	**0.891**	0.670	0.791	0.621	**0.795**	**0.513**
	Default ($\alpha = 1$)	**0.764**	0.957	0.778	0.888	**0.671**	**0.793**	**0.622**	0.794	0.510
Cardio-resp.	Independent ($\alpha \to \infty$)	0.472	0.869	0.581	0.793	0.502	0.851	0.722	**0.810**	**0.533**
(M=3)	No masking ($\alpha \to 0$)	0.463	0.869	0.575	0.831	0.565	0.858	0.731	0.789	0.490
	Random ($p = 0.5$)	0.486	0.876	0.588	0.833	0.569	**0.862**	**0.740**	0.801	0.512
	Default ($\alpha = 1$)	**0.490**	**0.878**	**0.591**	**0.835**	**0.573**	0.861	0.737	0.806	0.522
Full	Independent ($\alpha \to \infty$)	0.773	0.966	0.801	0.866	0.657	0.847	0.708	0.817	0.546
(M=8)	No masking ($\alpha \to 0$)	0.792	0.970	0.816	0.900	0.710	**0.861**	0.731	0.830	0.582
	Random ($p = 0.5$)	0.795	**0.971**	0.819	**0.902**	0.713	0.860	**0.733**	**0.831**	**0.584**
	Default ($\alpha = 1$)	**0.797**	**0.971**	**0.820**	0.901	**0.716**	0.859	0.730	**0.831**	0.582
MLP probe										
EEG-C3	Independent ($\alpha \to \infty$)	0.784	0.969	0.811	0.917	0.752	0.788	0.599	0.788	0.499
(M=1)	No masking ($\alpha \to 0$)	0.786	0.970	0.813	0.924	0.771	0.790	0.606	0.794	0.510
	Random ($p = 0.5$)	**0.796**	**0.972**	**0.823**	0.932	0.790	0.798	0.617	0.809	0.537
	Default ($\alpha = 1$)	**0.796**	**0.972**	0.822	**0.933**	**0.792**	**0.799**	**0.624**	**0.810**	**0.538**
EOG	Independent ($\alpha \to \infty$)	0.781	0.969	0.808	0.903	0.712	0.792	0.616	0.778	0.487
(M=2)	No masking ($\alpha \to 0$)	0.787	0.970	0.813	0.916	0.743	0.794	0.624	0.768	0.466
	Random ($p = 0.5$)	**0.795**	**0.972**	**0.820**	**0.923**	**0.760**	**0.804**	**0.640**	**0.812**	**0.544**
	Default ($\alpha = 1$)	0.794	**0.972**	**0.820**	0.922	0.759	0.803	0.634	0.810	0.542
Cardio-resp.	Independent ($\alpha \to \infty$)	0.539	0.895	0.626	0.853	0.613	0.869	0.752	0.832	**0.583**
(M=3)	No masking ($\alpha \to 0$)	0.543	0.897	0.628	0.873	0.650	0.879	0.771	0.815	0.539
	Random ($p = 0.5$)	0.559	0.903	0.639	0.876	**0.659**	**0.881**	**0.772**	0.827	0.564
	Default ($\alpha = 1$)	**0.562**	**0.905**	**0.643**	**0.877**	**0.659**	**0.881**	**0.772**	**0.835**	0.579
Full	Independent ($\alpha \to \infty$)	0.800	0.974	0.826	0.931	0.784	0.864	0.737	0.842	0.598
(M=8)	No masking ($\alpha \to 0$)	0.812	0.976	0.835	0.943	0.813	**0.880**	0.766	0.856	0.630
	Random ($p = 0.5$)	**0.814**	**0.977**	**0.839**	**0.944**	0.815	0.879	**0.768**	**0.860**	**0.636**
	Default ($\alpha = 1$)	0.813	**0.977**	0.837	**0.944**	**0.816**	**0.880**	0.767	0.858	0.631

B.6 Few-shot scaling across all datasets

Figure 6 reports few-shot sleep-staging curves for one in-domain (SHHS) and one held-out (MrOS) dataset. For completeness, Tables 12 and 13 reports the full per-dataset breakdown of sleep staging results across eight cohorts, three labelled-data fractions (1%, 10%, 100%), and three metrics (Cohen's κ, macro-AUROC, macro-AUPRC) for linear and MLP probes. Hypnos consistently achieves the best macro-AUPRC in every cohort at every data fraction, and is best or near-best on κ and AUROC throughout. Using 1% of labelled data, Hypnos outperforms both U-Sleep and SleepTransformer using 100% of the labelled data on 3 out of 4 held-out test sets evaluated.

Table 12: **Few-shot sleep stage classification across all datasets — linear probe.** Per-subject mean of Cohen's κ, macro-AUROC and macro-AUPRC, macro-averaged over the five AASM stages, as the fraction of labelled recordings used to train the probe is scaled from 1% to 100%. Foundation models are evaluated with a frozen encoder and a **linear probe**; U-Sleep and SleepTransformer are supervised (EEG+EOG), retrained at each fraction. **Best** per (fraction, metric) within each dataset block in bold.

		1%			10%			100%		
Dataset	Method	κ	AUROC	AUPRC	κ	AUROC	AUPRC	κ	AUROC	AUPRC
In-domain										
SHHS	U-Sleep [42]	0.734	0.916	0.741	0.790	0.958	0.815	0.799	0.969	0.825
	SleepTransformer [43]	0.710	0.947	0.755	0.782	0.967	0.805	0.806	0.974	0.832
	SleepFM [58]	0.699	0.942	0.733	0.668	0.911	0.671	0.716	0.941	0.731
	sleep2vec [69]	0.681	0.932	0.716	0.743	0.958	0.774	0.763	0.964	0.792
	OSF [53]	0.754	0.955	0.773	0.780	0.964	0.796	0.791	0.968	0.809
	Hypnos	**0.762**	**0.964**	**0.798**	**0.799**	**0.973**	**0.825**	**0.811**	**0.976**	**0.836**
CCSHS	U-Sleep	0.785	0.931	0.796	0.843	0.973	0.874	0.851	0.981	0.888
	SleepTransformer	0.776	0.965	0.819	0.852	0.980	0.867	0.873	0.987	0.896
	SleepFM	0.776	0.960	0.801	0.742	0.934	0.738	0.784	0.958	0.797
	sleep2vec	0.771	0.959	0.798	0.819	0.975	0.844	0.833	0.980	0.862
	OSF	0.827	0.973	0.842	0.842	0.979	0.859	0.849	0.983	0.872
	Hypnos	**0.843**	**0.980**	**0.871**	**0.875**	**0.988**	**0.893**	**0.881**	**0.990**	**0.901**
CFS	U-Sleep	0.743	0.918	0.756	0.808	0.961	0.825	0.803	0.971	0.845
	SleepTransformer	0.760	0.958	0.798	0.779	0.962	0.816	0.804	0.973	0.847
	SleepFM	0.705	0.937	0.735	0.662	0.906	0.677	0.725	0.933	0.738
	sleep2vec	0.738	0.940	0.757	0.792	0.962	0.808	0.804	0.966	0.820
	OSF	0.757	0.954	0.784	0.784	0.962	0.802	0.805	0.968	0.816
	Hypnos	**0.785**	**0.967**	**0.829**	**0.824**	**0.979**	**0.858**	**0.839**	**0.983**	**0.873**
NCHSDB	U-Sleep	0.626	0.850	0.671	0.718	0.939	0.774	0.733	0.951	0.794
	SleepTransformer	0.618	0.925	0.731	0.740	0.948	0.783	0.773	0.963	0.816
	SleepFM	0.595	0.909	0.692	0.570	0.880	0.634	0.639	0.918	0.707
	sleep2vec	0.632	0.915	0.708	0.694	0.941	0.761	0.713	0.948	0.774
	OSF	0.609	0.909	0.715	0.674	0.932	0.749	0.704	0.942	0.764
	Hypnos	**0.723**	**0.948**	**0.784**	**0.767**	**0.964**	**0.817**	**0.777**	**0.967**	**0.826**
Held-out										
MROS	U-Sleep	0.592	0.872	0.621	0.737	0.934	0.730	0.730	0.949	0.748
	SleepTransformer	0.673	0.938	0.699	0.714	0.943	0.724	0.753	0.958	0.758
	SleepFM	0.622	0.918	0.638	0.601	0.889	0.604	0.658	0.919	0.659
	sleep2vec	0.656	0.925	0.664	0.734	0.951	0.719	0.751	0.958	0.737
	OSF	0.698	0.940	0.702	0.707	0.948	0.723	0.740	0.957	0.738
	Hypnos	**0.701**	**0.944**	**0.718**	**0.766**	**0.968**	**0.778**	**0.799**	**0.973**	**0.794**
DOD-H	U-Sleep	0.622	0.886	0.729	0.721	0.952	0.818	0.753	0.956	0.830
	SleepTransformer	0.628	0.952	0.787	0.757	0.967	0.846	0.785	0.972	0.861
	SleepFM	0.478	0.911	0.717	0.402	0.861	0.617	0.635	0.925	0.740
	sleep2vec	0.626	0.924	0.742	0.701	0.955	0.810	0.724	0.963	0.837
	OSF	0.690	0.951	0.808	0.639	0.953	0.821	0.679	0.957	0.833
	Hypnos	**0.767**	**0.970**	**0.866**	**0.814**	**0.977**	**0.889**	**0.805**	**0.979**	**0.898**
DOD-O	U-Sleep	0.602	0.873	0.701	0.736	0.960	0.813	**0.766**	**0.968**	0.829
	SleepTransformer	0.295	0.894	0.610	0.554	0.937	0.730	0.720	0.956	0.795
	SleepFM	0.386	0.891	0.666	0.409	0.856	0.593	0.612	0.917	0.703
	sleep2vec	0.607	0.912	0.702	0.679	0.943	0.770	0.693	0.950	0.785
	OSF	0.645	0.945	0.768	0.656	0.954	0.784	0.717	0.960	0.801
	Hypnos	**0.746**	**0.967**	**0.828**	**0.747**	**0.972**	**0.847**	0.721	**0.968**	**0.847**
MESA	U-Sleep	0.576	0.863	0.604	0.640	0.904	0.683	0.655	0.931	0.712
	SleepTransformer	0.475	0.882	0.633	0.663	0.940	0.726	0.661	0.942	0.740
	SleepFM	0.544	0.890	0.619	0.526	0.879	0.585	0.562	0.904	0.640
	sleep2vec	0.628	0.914	0.664	0.684	0.939	0.716	0.695	0.946	0.731
	OSF	0.623	0.917	0.667	0.637	0.926	0.685	0.652	0.936	0.700
	Hypnos	**0.693**	**0.948**	**0.744**	**0.747**	**0.959**	**0.773**	**0.750**	**0.964**	**0.782**

Table 13: **Few-shot sleep stage classification across all datasets — MLP probe.** Per-subject mean of Cohen's κ, macro-AUROC and macro-AUPRC, macro-averaged over the five AASM stages, as the fraction of labelled recordings used to train the probe is scaled from 1% to 100%. Foundation models are evaluated with a frozen encoder and a **MLP probe**; U-Sleep and SleepTransformer are supervised (EEG+EOG), retrained at each fraction. **Best** per (fraction, metric) within each dataset block in bold.

		1%			10%			100%		
Dataset	Method	κ	AUROC	AUPRC	κ	AUROC	AUPRC	κ	AUROC	AUPRC
In-domain										
SHHS	U-Sleep [42]	0.734	0.916	0.741	0.790	0.958	0.815	0.799	0.969	0.825
	SleepTransformer [43]	0.710	0.947	0.755	0.782	0.967	0.805	0.806	0.974	0.832
	SleepFM [58]	0.702	0.943	0.735	0.748	0.957	0.772	0.772	0.964	0.791
	sleep2vec [69]	0.699	0.940	0.725	0.741	0.957	0.767	0.771	0.966	0.799
	OSF [53]	0.762	0.959	0.777	0.784	0.968	0.806	0.800	0.971	0.818
	Hypnos	**0.785**	**0.971**	**0.821**	**0.807**	**0.975**	**0.832**	**0.819**	**0.978**	**0.844**
CCSHS	U-Sleep	0.785	0.931	0.796	0.843	0.973	0.874	0.851	0.981	0.888
	SleepTransformer	0.776	0.965	0.819	0.852	0.980	0.867	0.873	0.987	0.896
	SleepFM	0.778	0.963	0.808	0.831	0.977	0.850	0.849	0.983	0.869
	sleep2vec	0.773	0.961	0.793	0.813	0.975	0.838	0.839	0.982	0.868
	OSF	0.829	0.976	0.843	0.844	0.982	0.872	0.863	0.985	0.884
	Hypnos	**0.856**	**0.984**	**0.880**	**0.871**	**0.988**	**0.894**	**0.888**	**0.991**	**0.914**
CFS	U-Sleep	0.743	0.918	0.756	0.808	0.961	0.825	0.803	0.971	0.845
	SleepTransformer	0.760	0.958	0.798	0.779	0.962	0.816	0.804	0.973	0.847
	SleepFM	0.704	0.938	0.741	0.740	0.953	0.783	0.783	0.965	0.812
	sleep2vec	0.748	0.944	0.755	0.778	0.960	0.794	0.810	0.970	0.828
	OSF	0.774	0.959	0.790	0.796	0.968	0.816	0.813	0.972	0.829
	Hypnos	**0.793**	**0.974**	**0.843**	**0.821**	**0.979**	**0.865**	**0.838**	**0.983**	**0.880**
NCHSDB	U-Sleep	0.626	0.850	0.671	0.718	0.939	0.774	0.733	0.951	0.794
	SleepTransformer	0.618	0.925	0.731	0.740	0.948	0.783	0.773	0.963	0.816
	SleepFM	0.610	0.919	0.708	0.680	0.940	0.755	0.721	0.950	0.778
	sleep2vec	0.628	0.921	0.710	0.687	0.940	0.753	0.733	0.954	0.787
	OSF	0.630	0.914	0.723	0.689	0.940	0.762	0.719	0.947	0.775
	Hypnos	**0.749**	**0.960**	**0.807**	**0.770**	**0.967**	**0.822**	**0.782**	**0.969**	**0.833**
Held-out										
MROS	U-Sleep	0.592	0.872	0.621	0.737	0.934	0.730	0.730	0.949	0.748
	SleepTransformer	0.673	0.938	0.699	0.714	0.943	0.724	0.753	0.958	0.758
	SleepFM	0.660	0.934	0.675	0.686	0.943	0.696	0.731	0.952	0.718
	sleep2vec	0.663	0.930	0.668	0.720	0.947	0.701	0.758	0.960	0.741
	OSF	0.716	0.946	0.709	0.733	0.956	0.735	0.750	0.960	0.747
	Hypnos	**0.746**	**0.961**	**0.763**	**0.767**	**0.968**	**0.776**	**0.805**	**0.973**	**0.796**
DOD-H	U-Sleep	0.622	0.886	0.729	0.721	0.952	0.818	0.753	0.956	0.830
	SleepTransformer	0.628	0.952	0.787	0.757	0.967	0.846	0.785	0.972	0.861
	SleepFM	0.531	0.911	0.720	0.684	0.949	0.792	0.647	0.943	0.788
	sleep2vec	0.639	0.934	0.746	0.679	0.950	0.797	0.745	0.963	0.836
	OSF	0.698	0.954	0.815	0.723	0.959	0.829	0.700	0.955	0.820
	Hypnos	**0.788**	**0.975**	**0.882**	**0.786**	**0.976**	**0.889**	**0.820**	**0.982**	**0.906**
DOD-O	U-Sleep	0.602	0.873	0.701	**0.736**	0.960	0.813	**0.766**	**0.968**	0.829
	SleepTransformer	0.295	0.894	0.610	0.554	0.937	0.730	0.720	0.956	0.795
	SleepFM	0.549	0.911	0.700	0.541	0.919	0.708	0.581	0.926	0.731
	sleep2vec	0.615	0.919	0.707	0.660	0.940	0.758	0.707	0.952	0.788
	OSF	0.722	0.953	0.785	0.694	0.954	0.792	0.703	0.953	0.796
	Hypnos	**0.742**	**0.966**	**0.833**	0.689	**0.962**	**0.835**	0.733	0.966	**0.843**
MESA	U-Sleep	0.576	0.863	0.604	0.640	0.904	0.683	0.655	0.931	0.712
	SleepTransformer	0.475	0.882	0.633	0.663	0.940	0.726	0.661	0.942	0.740
	SleepFM	0.572	0.909	0.647	0.586	0.926	0.671	0.629	0.933	0.697
	sleep2vec	0.626	0.919	0.667	0.669	0.935	0.699	**0.705**	0.947	0.732
	OSF	0.639	0.925	0.683	0.626	0.931	0.690	0.666	0.937	0.703
	Hypnos	**0.690**	**0.956**	**0.757**	**0.702**	**0.958**	**0.763**	0.692	**0.956**	**0.754**

B.7 Transfer to external single-lead ECG benchmarks

We evaluate Hypnos's frozen ECG-only embeddings against xECG [33], a foundation model pretrained on 12-lead clinical ECG, on three external single-lead ECG benchmarks (Table 14). Neither model's pretraining corpus contains any of these datasets. Both models use a frozen encoder followed by a `StandardScaler` + L2-regularised logistic-regression probe, with C tuned on validation AUROC over $\{10^{-4}, \ldots, 10^{2}\}$.

Apnea-ECG protocol. We follow BenchECG's `sleep_apnea` 5-way ensemble linear-probe recipe: each sample is a 300 s window centred on a target minute; the backbone produces per-minute features for five consecutive positions ($m-2, \ldots, m+2$); a linear head is trained on each position supervised by the central-minute label; the five predictions are averaged at test time.

Table 14: **External single-lead ECG benchmarks.** Frozen-encoder linear-probe AUROC across three publicly available external single-lead ECG datasets covering handheld (CinC 2017), overnight single-lead (Apnea-ECG), and dynamic Holter (CPSC 2021) settings. CinC 2017 reports 5-fold CV mean; other datasets report a single subject-disjoint split.

Dataset	Task / eval unit	Hypnos	xECG
CinC 2017	AF-vs-rest (per-record)	**0.984**	**0.985**
Apnea-ECG	Apnoea (per-minute)	**0.925**	0.884
CPSC 2021	Paroxysmal AF (30 s window)	**0.985**	0.934

B.8 Learning objective

The comparisons in Section 4 differ in objective, architecture, context length and aggregation, so they establish overall model performance rather than the effect of the learning objective. Here we compare next-token prediction (NTP) directly with two alternatives.

Masked-token prediction We trained a BERT-style bi-directional masked-token prediction (MTP) model, tuning the masking ratio over $\{25, 50, 75\}\%$ and selecting 50%. Table 15 shows that the two objectives achieve very similar downstream performance. The advantage of next-token prediction is practical: a single architecture supports both representation learning and auto-regressive generation with no train-test shift, and permits sequential state updates for streaming inference.

Input-space masked autoencoding We compared NTP with an input-space masked autoencoder (MAE) using unimodal EEG models with matched Transformer widths, context length and pretraining compute, tuning the learning rate per configuration. Table 16 shows that MAE is weaker at every scale.

Table 15: **Next-token vs. masked-token prediction.** Hypnos-Small trained on all eight modalities with a context length of 512, using the same tokenizers and depth head. MTP uses a bidirectional temporal transformer with 50% of timesteps masked, shared across modalities. Linear probes trained and evaluated on SHHS.

	Staging		Arousal		Apnoea		Desat.	
Objective	κ	AUROC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
MTP	0.813	0.971	0.926	0.754	0.869	0.785	0.829	0.590
NTP	0.817	0.974	0.920	0.737	0.864	0.777	0.837	0.612

B.9 Model size and compute

Tables 17 and 18 compare Hypnos with our re-trained OSF and sleep2vec respectively, using matched Transformer backbones, identical dataset splits and identical pre-processing. Encoder parameter counts exclude contrastive projection heads for sleep2vec, student parameters for OSF and discrete output codebooks for Hypnos. At each scale, Hypnos and OSF are trained with compute matched to within 10%. Hypnos achieves higher performance on all four tasks at every scale. Re-training

Table 16: **Input-space MAE vs. NTP across model scale.** Sleep staging κ (MLP probe) on held-out MrOS using single-channel EEG (C3).

Model size	Width	MAE	NTP
Tiny	192	0.649	0.766
Small	384	0.657	0.783
Base	768	0.725	0.795

also improves on the released OSF weights on some metrics (e.g. MrOS κ of 0.785 for OSF-Base vs 0.750).

Our sleep2vec re-implementation uses far less compute (around 10^{18} FLOPs for the Medium configuration). Scaling its compute beyond the original recipe, via a $4\times$ larger batch size with a tuned learning rate and up to 80k training steps, gave only a minor improvement (MrOS MLP κ from 0.757 to 0.765).

Table 17: **Hypnos vs. OSF at matched backbones and compute.** MLP probe on held-out MrOS. *Enc* is the encoder parameter count. All OSF results in this table use our re-trained models, rather than the released weights used in Section 4. Results are computed on the subset of recordings common to OSF, sleep2vec and Hypnos, so may differ slightly from Table 3, which uses the subset common to all models in that table.

Model	Backbone	Enc (M)	FLOPs	Staging κ	Apnoea AUROC	Arousal AUROC	Desat. AUROC
OSF-Tiny	12L/192d	5.4	7.5×10^{18}	0.709	0.761	0.904	0.704
Hypnos-Tiny	12L/192d	19.7	8.3×10^{18}	0.771	0.780	0.912	0.736
OSF-Small	12L/384d	21.4	2.5×10^{19}	0.766	0.768	0.909	0.728
Hypnos-Small	12L/384d	53.6	2.6×10^{19}	0.793	0.786	0.918	0.756
OSF-Base	12L/768d	85.3	9.3×10^{19}	0.785	0.791	0.918	0.738
Hypnos-Base	12L/768d	163.8	1.0×10^{20}	0.815	0.808	0.934	0.767

Table 18: **Hypnos vs. sleep2vec at matched backbones.** MLP probe on held-out MrOS, on the same recordings as Table 17. *Enc* is the encoder parameter count.

Model	Backbone	Enc (M)	Staging κ	Apnoea AUROC	Arousal AUROC	Desat. AUROC
sleep2vec-Tiny	12L/192d	19.0	0.679	0.723	0.829	0.700
Hypnos-Tiny	12L/192d	19.7	0.771	0.780	0.912	0.736
sleep2vec-Small	12L/384d	49.3	0.732	0.748	0.859	0.710
Hypnos-Small	12L/384d	53.6	0.793	0.786	0.918	0.756
sleep2vec-Medium	12L/768d	144.1	0.757	0.765	0.873	0.733
Hypnos-Base	12L/768d	163.8	0.815	0.808	0.934	0.767
sleep2vec-Large	16L/1024d	282.8	0.766	0.777	0.883	0.750

B.10 Comparison with SleepMaMi

SleepMaMi [41] trains a micro-encoder with private and shared components using contrastive, input-space reconstruction and KoLeo losses, followed by a macro-encoder over the full night using demographic-guided contrastive learning. Both models therefore separate local learning from long-range learning, whereas Hypnos requires no patient metadata and uses causal components throughout.

We compare the two models under linear probing on CHAT and MESA, generating SleepMaMi embeddings with its public weights, pre-processing and inference code, on the matched subset of recordings with the signals it requires. We excluded datasets such as SHHS because they appear in both pre-training corpora under different splits, and DOD-H and DOD-O because they do not record SpO_2, which SleepMaMi requires.

In Table 19, Hypnos performs better on sleep staging and arousal detection. SleepMaMi performs better on desaturation detection, where it receives SpO_2 directly while Hypnos must infer desaturation from indirect signals such as ECG and respiratory effort, so this comparison is not modality-matched. Apnoea results are mixed: Hypnos is better on CHAT AUROC, SleepMaMi is better on MESA AUPRC, and the remaining differences are not significant.

We also compare per-second detection of arousal and apnoea events in Table 20, reporting per-second macro-F1 and event-level F1, which rewards detecting whole events rather than individual seconds [38, 64]. Hypnos leads on event-level F1 for both tasks on both cohorts, at an IoU of 0.3 and at all thresholds from 0.1 to 0.5. Per-second F1 shows the same pattern, except for apnoea on CHAT, where the difference is not significant. SleepMaMi reports no evaluation under missing modalities, so we do not compare robustness.

Table 19: **Hypnos vs. SleepMaMi.** Linear probe on a matched subset of recordings. * indicates a significant difference (Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction, $q < 0.05$).

		Staging	Arousal		Apnoea		Desat.	
Cohort	Method	κ	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
CHAT	SleepMaMi	0.799	0.946	0.742	0.893	0.459	0.909^*	0.699^*
	Hypnos	0.852^*	0.965^*	0.793^*	0.925^*	0.460	0.863	0.502
MESA	SleepMaMi	0.754	0.897	0.717	0.789	0.455^*	0.915^*	0.885^*
	Hypnos	0.803^*	0.927^*	0.766^*	0.787	0.436	0.833	0.776

Table 20: **Event detection: Hypnos vs. SleepMaMi.** Per-second macro-F1 and event-level F1 (IoU 0.3). * indicates a significant difference, using the same test as Table 19.

		Arousal		Apnoea	
Cohort	Method	Per-second	Event	Per-second	Event
CHAT	SleepMaMi	0.577	0.603	0.327	0.333
	Hypnos	0.695^*	0.742^*	0.292	0.394^*
MESA	SleepMaMi	0.495	0.530	0.365	0.405
	Hypnos	0.632^*	0.681^*	0.382^*	0.453^*

B.11 Rhythm-conditioned ECG generation

As an example of Hypnos' ability to model stochastic structure, we auto-regressively generated single-lead ECG conditioned on either atrial fibrillation (AF) or regular rhythm using the MIT-BIH Atrial Fibrillation Database (AFDB) [39]. From a 60 s input segment, we generate 30 s signal continuations. Figure 15 shows that conditioning on inputs during atrial fibrillation leads to more irregular beat-to-beat intervals in the generated signals.

C Extended Limitations and Future Work

Clinical Outcomes In this work, we evaluated the quality of Hypnos' embeddings for existing diagnostic tasks such as sleep stage classification. Due to data availability, we were unable to evaluate Hypnos on broader clinical outcomes data like SleepFM [58]. Evaluating Hypnos on a similar broad range of clinical outcomes is an important direction for future work. Given the large performance improvements observed across our evaluations, we expect Hypnos would also enable a significant improvement in these tasks.

Biomarker Discovery There are well-known issues with existing sleep stage definitions. For example, small changes to the brittle definition of deep (N3) sleep can have a significant effect on downstream analyses, such as how sleep architecture purportedly varies with age for women [16]. A promising direction is using Hypnos for data-driven biomarker discovery and identifying better

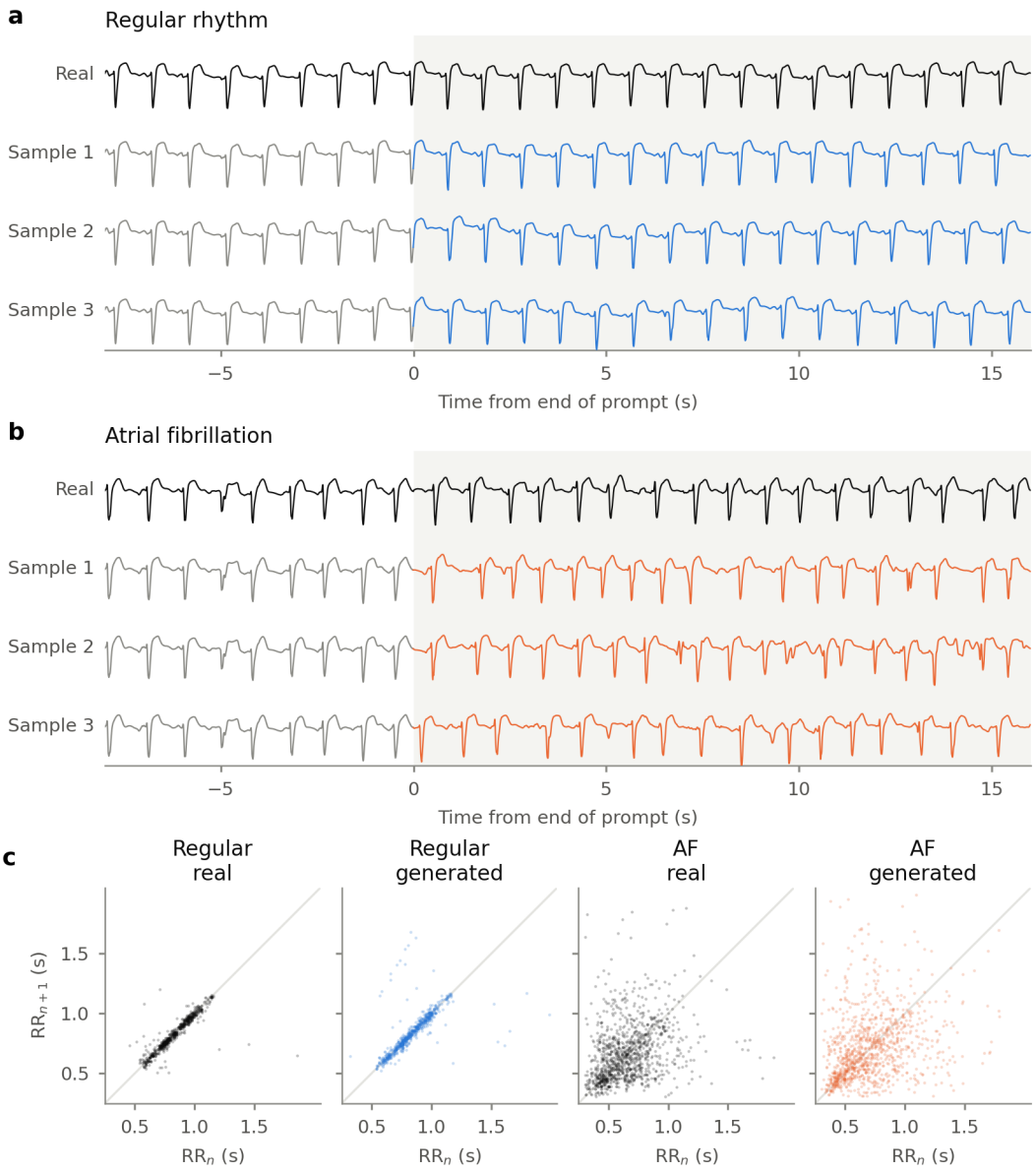


Figure 15: **Rhythm-conditioned ECG generation on MIT-BIH AFDB.** (a, b) Real continuation (black) and three generated samples (colour) from a prompt (grey; final 8 s of 60 s shown, first 16 s of 30 s generated) for regular rhythm and AF in the same patient. (c) Poincaré plots of successive RR intervals for real and generated electrocardiograms.

continuous measures of health. For example, identifying latent directions or sparse features predictive of incident neurodegenerative or cardiovascular disease. The generative capability of Hypnos could be used for visual interpretation of such features in input space.

D Broader Impact

Sleep disorders such as obstructive sleep apnoea and insomnia are common but routinely under-diagnosed, in part because polysomnography is expensive to record and labour-intensive to score. Foundation models for physiological signals offer a path to reducing the cost of sleep analysis and broadening access to sleep medicine, through improved generalisation with fewer labelled examples

and via fast adaptation to novel sensor configurations, e.g. from consumer wearables. The recordings used in this work are drawn from nine public datasets, predominantly curated by the National Sleep Research Resource (NSRR). These cohorts are each biased by exclusion criteria; for example, MrOS is a study that solely recruited older males. For real-world deployment, Hypnos should be evaluated on populations of interest, and we encourage further work auditing performance across demographic subgroups.

E Additional Acknowledgements

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

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

The Wisconsin Sleep Cohort 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 Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380).

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39).

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.

NCH Sleep DataBank was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 4 shows that Hypnos outperforms prior models on most sleep analysis tasks and datasets evaluated, that performance improves with model scale, and that the model is robust to missing modalities.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalise to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated section before conclusions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Paper does not contain theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All implementation details are discussed in the main paper or sign-posted to the Appendix. This includes pre-processing details, dataset splits, model hyper-parameters, and model training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All datasets are publicly available. We release pre-trained model weights, an inference library and a worked example of embedding generation and linear probing at `https://github.com/joncarter1/hypnos`. Code for pre-training and for reproducing all experiments is not currently released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits follow prior work. All hyper-parameters are provided and were either carefully informed by prior work or tuned using validation data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our primary results compare with prior work using Wilcoxon signed-rank tests computed over a large number of test subjects (paired across methods), as described in Section 4. We did not repeat our main pre-training experiments with different random seeds due to computational resources. However, in Section B.4 we investigate variance due to random seeds using smaller scale unimodal models, reporting mean $\pm$ standard deviation across three seeds, and find this to be very low relative to our performance improvements over prior work.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section A.5 we report GPU types, memory usage and training time for all major model components. We additionally provide an estimate of total compute usage.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and believe the paper conforms to them.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Discussed in Section D.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not believe release of code or models requires safeguards. Data used is already publicly available.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Datasets used have been appropriately cited as requested by the National Sleep Research Resource which provided access. External ECG datasets, and the code and weights of baseline models, are credited by citation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The model weights and inference library are released under the MIT license, documented in a README and model card, with a worked example notebook.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Paper uses retrospective, publicly available datasets. Specific details regarding data collection can be found in the original works which are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: All datasets used in this work are publicly available and have been de-identified or fully anonymised, thus not requiring IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an original or core component of the methods described in this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.