OSF: On Pre-training and Scaling of Sleep Foundation Models

Zitao Shuai[1], **Zongzhe Xu**[1], **David Yang**[2], **Wei Wang**[1], **Yuzhe Yang**[1†]

[1]University of California, Los Angeles, [2]Emory University

Polysomnography (PSG) provides the gold standard for sleep assessment but suffers from substantial heterogeneity across recording devices and cohorts. There have been growing efforts to build general-purpose foundation models (FMs) for sleep physiology, but lack an in-depth understanding of the *pre-training* process and *scaling* patterns that lead to more generalizable sleep FMs. To fill this gap, we curate a massive corpus of 166,500 hours of sleep recordings from nine public sources and establish `SleepBench`, a comprehensive, fully open-source benchmark. Leveraging `SleepBench`, we systematically evaluate four families of self-supervised pre-training objectives and uncover three critical findings: (1) existing FMs fail to generalize to missing channels at inference; (2) channel-invariant feature learning is essential for pre-training; and (3) scaling sample size, model capacity, and multi-source data mixture consistently improves downstream performance. With an enhanced pre-training and scaling recipe, we introduce OSF, a family of sleep FMs that achieves state-of-the-art performance across nine datasets on diverse sleep and disease prediction tasks. Further analysis of OSF also reveals intriguing properties in sample efficiency, hierarchical aggregation, and cross-dataset scaling. Codes are available at: https://github.com/yang-ai-lab/OSF-Open-Sleep-FM.

Code: https://github.com/yang-ai-lab/OSF-Open-Sleep-FM
Website: https://yang-ai-lab.github.io/osf

1. Introduction

Sleep is a fundamental physiological process for human health [36, 39, 41]. Assessing sleep quality [28] and detecting sleep disorder events [34] often rely on polysomnography (PSG). PSG typically includes multiple complementary signals that capture brain activity, respiratory effort, muscle movement, and cardiac activity. In practice, these recordings differ across patient cohorts and devices [16, 19], and the available channel set is often inconsistent due to differences in acquisition protocols and occasional channel dropouts. For example, home studies often lack brain signals [3, 46], and sensors can become dislodged during the night. Together, these challenges limit the transferability of sleep analysis models [45].

Recent foundation models (FMs) [36] have shown rapid progress in modeling sleep recordings. They learn generalizable representations via self-supervised

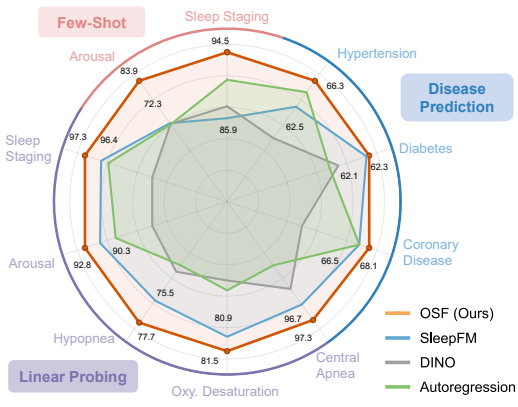


Figure 1 | **Performance comparison across downstream tasks.** OSF consistently achieves state-of-the-art on downstream tasks.

†*Correspondence to: yuzhey@ucla.edu.*

learning (SSL) on large-scale 30-second PSG epochs. These representations support efficient adaptation to diverse epoch-level tasks (e.g., sleep staging, arousal detection) and can be aggregated over time to predict patient-level outcomes. Prior work [36] has primarily focused on contrastive approaches, where sleep signals are grouped into multiple modalities and their embeddings are aligned in a shared latent space. While contrastive learning has achieved strong results, the broader design space for pre-training sleep FMs remains largely under-explored.

In this work, we move beyond reporting gains from a single pre-training recipe and ask what actually drives transfer in sleep FMs. Existing sleep FMs differ in their pre-training objectives, but it is unclear which of these choices reliably improve downstream performance across cohorts, devices, and channel sets. A central question arises:

*Which **pre-training** and **scaling** design choices truly improve the generalization of sleep FMs, especially under cohort shift and missing-channel inference?*

To answer this question with controlled evaluation, we conduct a systematic study that quantifies how different pre-training design choices affect downstream performance. We also construct `SleepBench`, the largest fully open, multi-source sleep benchmark, aggregated from nine publicly available datasets with diverse demographics, comprising 166,500 hours of recordings from over 21,000 sleep studies, and spanning a broad set of PSG channels.

Leveraging `SleepBench`, we identify three main findings. *First*, existing sleep FMs are not reliable under inference-time input incompleteness, showing large performance drops under realistic missing-channel settings. *Second*, through a controlled, step-by-step study of pre-training design choices, we find that learning channel-invariant features is critical for invariance-driven methods to learn stronger representations. *Third*, we observe consistent scaling trends in sleep: increasing training data and model capacity, together with multi-source data mixing, improves downstream performance across tasks. Guided by these insights, we propose **O**pen **S**leep **F**oundation Model (**OSF**), a family of sleep foundation models that achieves state-of-the-art across diverse sleep-related tasks and datasets (Fig. 1).

Our contribution. In summary, we present a systematic study of how to build generalizable sleep FMs. We curate a large corpus of 166,500 hours of sleep recordings from nine public sources and establish `SleepBench`, a comprehensive, fully open-source benchmark. Leveraging `SleepBench`, we evaluate four families of self-supervised pre-training objectives and identify three key findings: ❶ existing sleep FMs fail to generalize to missing channels at inference; ❷ channel-invariant feature learning is essential for strong pre-training; and ❸ scaling sample size, model capacity, and multi-source data mixture consistently improves downstream performance. Guided by these results, we introduce OSF, a family of sleep FMs trained with an improved pre-training and scaling recipe, achieving state-of-the-art performance across nine datasets on diverse sleep and disease prediction tasks. Further analysis of OSF reveals strong sample efficiency, effective hierarchical aggregation, and consistent scaling across datasets.

2. Fully Open Sleep Benchmarking at Scale

Our benchmark aggregates nine publicly available datasets hosted by the National Sleep Research Resource (NSRR) [46]: SHHS [29], NCHSDB [18], CFS [32], CCSHS [33], WSC [42], MROS [4], MESA [7], CHAT [22], and SOF [35]. In total, it contains **166,500** hours of PSG recordings spanning more than **21,000** nights. We further partition the data into an in-domain cohort for pre-training and an external cohort for out-of-domain (OOD) evaluation. Each cohort is split into training, validation, and test sets. All samples in the external cohort are held out from pre-training and used only for downstream evaluation. For the in-domain cohort, we use the training split for pre-training and also

for downstream fine-tuning in the in-domain setting. Detailed statistics across cohorts, datasets, and splits are provided in Appendix B.1.

We focus on a shared subset of channels that is available across most cohorts. Specifically, we utilize a standardized 12-channel montage grouped by physiological modality: ❶ *Brain (EEG/EOG):* `C3-A2`, `C4-A1`, `E1-A2`, `E2-A1`, ❷ *Respiration:* `Abdominal`, `Thorax`, `Nasal Pressure`, `Snore`, ❸ *Cardiac:* `ECG`, and ❹ *Somatic:* `EMG-Chin`, `EMG-LLeg`, `EMG-RLeg`. Detailed settings and preprocessing steps are provided in Appendix B.3.

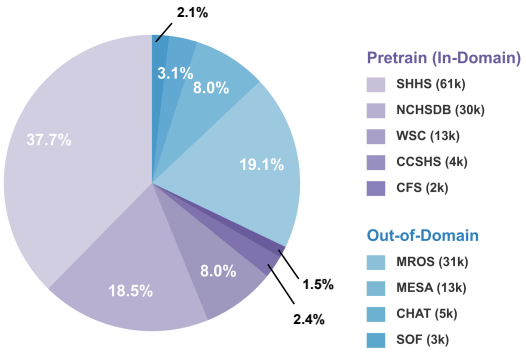


Figure 2 | **Distribution of our established `SleepBench`.**

We process each night's recording with a standardized pipeline to reduce cross-dataset heterogeneity. We first perform manual quality control to trim prolonged wake periods at the beginning and end of the night, removing extreme artifacts from sensor setup and removal. Next, we apply per-night z-score normalization to each channel. For respiratory channels, we additionally apply area-dependent z-score normalization to reduce amplitude drift and improve cross-night consistency.

We follow prior work and segment each night into non-overlapping 30-second epochs, yielding roughly **20 million** epochs for self-supervised pre-training. For segmented samples, we resample all channels to 64 Hz and zero-pad missing channels. Epoch-level statistics are provided in Table 17. To the best of our knowledge, this corpus constitutes the *largest* fully public benchmark for pre-training and evaluating sleep foundation models across diverse channel types. We compare our `SleepBench` other published benchmarks in Appendix B.1. With large-scale sleep data spanning diverse cohorts, `SleepBench` enables controlled benchmarking of different pre-training methods on sleep data, and supports systematic studies of how performance scales with pre-training sample size and how multi-source data blending affects generalization.

We acknowledge ongoing community efforts to harmonize public sleep datasets. Rather than treating `SleepBench` as a standalone contribution, we view it as a methodological infrastructure that enables controlled multi-source pre-training and evaluation under a unified setting. We will open-source the codebase and processing pipeline, and maintain `SleepBench` as a living resource that can be expanded with additional PSG datasets and community contributions.

3. On Pre-training Sleep FM

In this section, we study practical design choices for self-supervised pre-training of sleep foundation models. Following [36], we conduct pre-training and downstream evaluation on 30-second epochs segmented from full-night recordings.

3.1. Generalization Breaks Under Missing-Channel Inference

We first assess whether existing sleep FMs can cope with real-world deployment challenges. Real-world sleep recordings exhibit heterogeneous channel availability across cohorts, devices, and study protocols, such that channels present in the pre-training corpus may be absent at test time. For instance, home-based recordings [10] sometimes do not collect EEG and EOG channels, whereas EEG-centric sleep architecture studies [11] may not include breathing signals. This input mismatch motivates a key research question: *Do current sleep FMs generalize under missing-channel settings?*

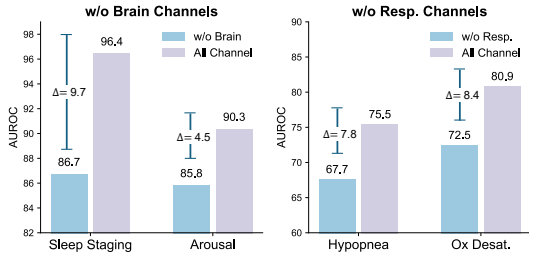


Figure 3 | **Inference with full versus missing channels.** Existing sleep FM fails to generalize to missing channel samples.

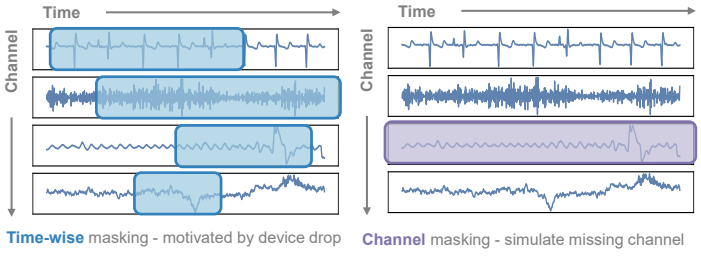


Figure 4 | **Illustration of considered augmentations.** We consider time-wise masking and channel masking strategies.

To this end, we simulate realistic missing-channel cases by zero-masking a subset of channels at inference time. We group the 12 channels into three groups: (1) *brain activity* channels, (EEG and EOG), which provide key cues for sleep staging; (2) *respiratory* channels, including breathing effort and airflow waveforms that characterize respiratory physiology; and (3) all remaining channels (ECG and EMG). We focus on two realistic settings: *in-home studies* that do not provide brain signals, and *micro-event studies* where breathing channels are not collected.

We evaluate current sleep FMs under full-channel input and the two missing-channel settings described above. As shown in Fig. 3, removing brain activity channels leads to a substantial drop in sleep staging performance. Similarly, when respiratory channels are removed, hypopnea detection performance degrades significantly. These results match *clinical intuition*: hypopnea is driven by respiration, and sleep staging depends primarily on brain signals.

Finding 1: Existing sleep FMs fail to generalize under missing-channel inference, motivating pre-training designs that explicitly handle channel incompleteness.

3.2. Channel-Invariant Pre-training Improves Robustness and Transfer

The failures of existing sleep FMs under missing-channel inference suggest that *pre-training should encourage the model to learn representations that are invariant to the available channel subset.*

To test this hypothesis, we perform a controlled, step-by-step study using SimCLR [6]while varying its augmentation strategy. We focus on SimCLR because it learns by aligning two augmented views of the same input, and its objective is largely shaped by the augmentation design. Specifically, following prior work on self-supervised learning for time series [20], we use time-wise block masking as the default augmentation. We compare it against an enhanced strategy that additionally masks a randomly selected subset of input channels.

Table 1 | **Comparisons of different masking strategies.** Channel masking consistently improves downstream performance, and combining time and channel masking yields the best results across pre-training methods.

Model	Mask Strategy		Sleep Staging		Hypopnea	
	Time	Channel	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SimCLR	✓	✗	81.4	56.8	58.4	53.1
	✗	✓	94.8	84.3	73.3	59.2
	✓	✓	**96.7**	**89.0**	**75.6**	**60.9**
DINO	✓	✗	93.6	81.4	72.7	59.5
	✗	✓	96.4	87.9	77.1	61.4
	✓	✓	**97.3**	**90.4**	**77.7**	**62.2**

The augmentation strategies are illustrated in Fig. 4. For block masking, given an input $\mathbf{x} \in \mathbb{R}^{C \times T}$ with C channels and sequence length T, we apply masking independently to each channel. For each

channel, we sample a masking ratio $r \in 0.3, 0.6$ and a start index $p \sim \mathcal{U}\{0, \ldots, \lfloor 1 - rT \rfloor\}$, and set the contiguous segment $p, p\ \lfloor rT \rfloor$ to zero. For channel masking, we randomly drop 50% of the channels by setting all time steps in the selected channels to zero.

As shown in Table 1, SimCLR pre-trained with channel masking consistently outperforms the default time-only masking by a large margin across multiple downstream tasks. To isolate the contribution of channel masking, we further train a SimCLR variant that uses only channel masking, while keeping all other hyperparameters and masking ratios the same. Table 1 confirms that this channel-only variant still significantly outperforms the default setting, suggesting that SimCLR learns more robust features when the two views differ in channel availability.

To test whether this effect extends beyond contrastive learning, we conduct an analogous controlled study with DINO [26], a distillation-based method. As shown in Table 1, we observe a similar pattern: although DINO performs strongly with time masking alone, adding channel masking yields a clear improvement. Based on these results, we summarize our second finding as follows:

Finding 2: Explicitly encouraging channel-invariant feature learning during pre-training improves robustness and downstream transfer, particularly for contrastive and distillation-based methods.

3.3. Scaling Laws Emerge in Sleep FM Pre-training

Recent sleep FMs are pre-trained on millions of 30-second epochs, yet it remains unclear whether current designs continue to improve as we scale up the amount of pre-training data. We therefore ask *whether sleep FM performance improves consistently as the pre-training sample size increases.*

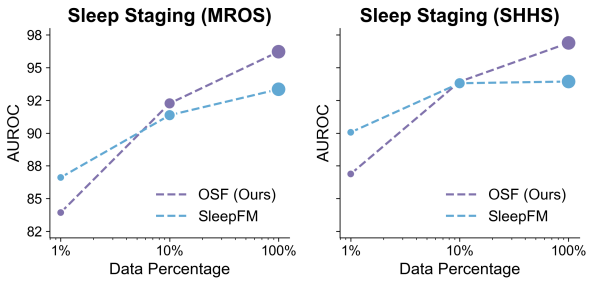


Figure 5 | **Comparison of scaling behavior.** Current sleep FM is less scalable. In contrast, OSF better utilizes training data.

We start by pre-training the prior sleep FM [36] on our pre-training cohort, matching the original parameter budget. As shown in Fig. 5, sleep staging performance on SHHS improves when increasing the pre-training data from 1% to 10% of the corpus, but shows little or no gain when further scaling from 10% to 100%. We observe a similar saturation trend on MROS. These results suggest that the baseline pre-training design may not fully benefit from large-scale sleep data. We then repeat the same data-scaling study using the pre-training recipe identified in Sec. 3.2, while keeping the model size fixed. As shown in Fig. 5, performance improves consistently as we scale up the pre-training sample size, indicating that data scaling can hold for sleep FM pre-training with appropriate SSL designs.

We further analyze scaling and multi-source data mixture in more detail. We find that (1) consistent gains can be observed when scaling both sample size and model size, and (2) increasing data diversity via multi-source blending improves generalization relative to training on a single source. Results are presented in Sec. 5. Taken together, these results motivate our final pre-training recipe: we scale model capacity and pre-train on large-scale, multi-source data, while explicitly encouraging the model to learn channel-invariant representations.

Finding 3: Baseline sleep FMs can saturate as pre-training data grows; with channel-invariant SSL, performance scales more consistently with larger data size, larger models, and multi-source data mixtures.

Table 2 | **Sleep staging and sleep event detection.** OSF achieves the best overall performance among all compared methods. We report macro-AUC and macro-AUPRC.

Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.		Central Apnea	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
Supervised:										
ViT [9]	96.9	89.9	88.4	83.0	81.3	66.2	81.8	81.6	97.3	70.5
Linear Probing:										
SleepFM [36]	96.4	87.7	90.3	84.9	75.5	60.6	80.9	80.3	96.7	70.3
SimCLR [6]	81.4	56.8	73.6	67.8	58.4	53.1	66.8	65.0	79.4	53.6
DINO [26]	93.6	81.4	82.2	75.7	72.7	59.5	78.6	78.2	96.1	66.8
VQ-VAE [37]	95.8	86.7	86.2	80.3	75.0	60.8	79.8	79.5	96.7	68.8
MAE [12]	96.5	88.7	89.7	84.3	77.6	61.9	81.2	80.8	**97.3**	**71.7**
Autoregression [30]	96.0	87.6	87.8	81.9	72.1	58.9	79.0	78.5	95.2	65.7
OSF	**97.3**	**90.4**	**92.8**	**88.3**	**77.7**	**62.2**	**81.5**	**81.0**	**97.3**	70.7
Full Fine-tuning:										
SleepFM [36]	97.0	89.5	93.7	89.6	85.0	69.0	82.8	82.4	97.7	71.0
SimCLR [6]	95.2	85.5	84.0	77.4	65.9	56.2	80.5	80.1	96.8	69.3
DINO [26]	97.0	90.2	92.3	88.0	84.5	68.9	82.6	82.3	97.9	75.2
VQ-VAE [37]	97.4	90.8	93.5	89.5	81.8	65.8	82.4	82.1	98.0	75.2
MAE [12]	97.3	90.6	93.3	89.3	**85.1**	**69.3**	83.3	83.0	**98.1**	**76.0**
Autoregression [30]	97.1	90.0	93.0	88.9	84.6	68.9	82.8	82.4	97.9	75.6
OSF	**97.9**	**92.0**	**94.6**	**91.0**	85.0	69.2	**83.5**	**83.1**	**98.1**	75.5

4. OSF: Open Sleep Foundation Models

4.1. Experimental Setup

Pre-training Setup. To rigorously benchmark self-supervised pre-training methods on sleep data, in addition to SleepFM [36], we adapt four representative method families: ❶ *Contrastive Learning*: SimCLR [6]; ❷ *Reconstruction-based Method*: MAE [12] and VQ-VAE [37]; ❸ *Autoregressive Modeling*: Autoregression [30]; and ❹ *Self-Distillation*: DINO [26]. In our implementation, we use DINO as the base method and incorporate our pre-training recipes, yielding OSF. Unless otherwise specified, all methods reported in the main tables use a Transformer [9] encoder with 85M parameters. All models are pre-trained on the same split of the pre-training cohorts in `SleepBench` for up to 30 epochs, with early stopping upon convergence. We use AdamW [21] with linear warmup (10% of total steps) followed by cosine annealing, and we tune the learning rate and batch size for each method.

Evaluation Setup. For main experiments, we evaluate pre-trained models across five *epoch-level* tasks and three *patient-level* disease prediction tasks. The epoch-level tasks include (1) 4-class sleep staging, and (2) four binary sleep-event detection tasks: Arousal, Hypopnea, Oxygen Desaturation (Ox. Desat.), and Central Apnea. We also evaluate three patient-level disease classifications: Coronary Disease, Diabetes, and Hypertension. We report AUROC and AUPRC due to label imbalance. For epoch-level tasks, we consider three transfer protocols: (1) linear probing, (2) full fine-tuning, and (3) few-shot adaptation. For disease classification, we segment each recording into 30-second epochs, extract an embedding for each epoch using the pre-trained encoder, and then aggregate epoch embeddings for patient-level prediction. We use MROS as the primary OOD evaluation cohort and SHHS as the primary in-domain cohort. Unless otherwise specified, we report results on MROS in the main text. All pre-training and evaluation configurations, as well as implementation details, are provided in Appendix A.

Table 3 | **Linear probing sleep staging across diverse cohorts.** OSF achieves the best performance across all datasets.

Method	CCSHS		CFS		NCHSDB		WSC		CHAT		MESA		SOF	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
	In-Domain Datasets								*Out-of-Domain Datasets*					
SleepFM	97.8	94.9	97.4	93.4	95.2	89.0	97.7	89.2	97.8	95.1	94.6	82.8	96.2	89.7
SimCLR	84.5	65.4	83.7	64.5	80.3	56.5	84.2	56.3	87.9	70.2	73.8	46.6	79.5	56.4
DINO	95.6	88.9	95.9	89.4	91.2	78.8	96.2	85.3	96.5	90.7	89.4	71.5	94.3	85.0
VQ-VAE	97.4	93.6	97.1	92.5	93.8	85.2	97.2	88.1	97.6	93.8	93.0	79.0	95.0	86.3
MAE	97.7	94.8	97.4	93.5	95.0	88.3	97.7	89.1	97.9	94.9	92.3	78.4	95.8	88.8
AR	97.6	94.4	97.3	93.2	94.7	87.4	97.6	89.3	97.8	94.5	90.4	73.7	95.7	88.6
OSF	**98.6**	**96.7**	**97.9**	**94.9**	**95.8**	**90.1**	**98.1**	**90.6**	**98.4**	**96.2**	**95.9**	**85.7**	**97.0**	**91.8**

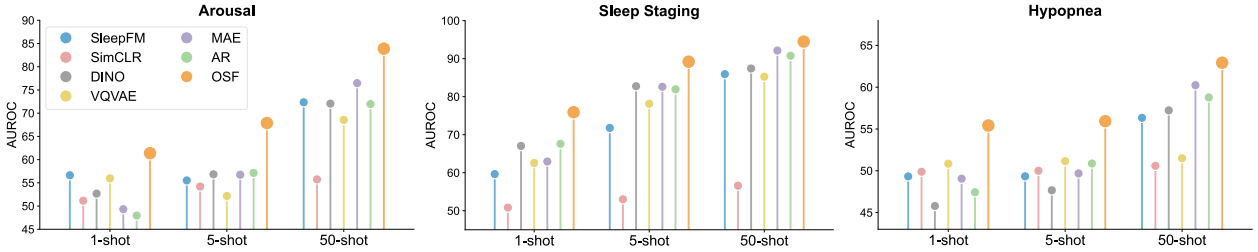


Figure 6 | **Few-shot adaptation to downstream tasks.** OSF achieves the best performance across all tasks at every shot count.

4.2. Main Results

OSF achieves state-of-the-art on sleep analysis tasks. We evaluate linear-probing performance for sleep staging across all cohorts in `SleepBench`. As shown in Table 3, OSF achieves the best performance across all cohorts. We further evaluate models on four sleep event detection tasks. On MROS, Table 2 shows that OSF achieves state-of-the-art performance on both sleep staging and event detection under linear probing and fine-tuning. We observe similar trends on SHHS (Appendix C.1).

OSF adapts better in few-shot settings. To measure adaptation under limited supervision, we evaluate all models in a few-shot setting on the out-of-domain MROS cohort. We consider $K \in \{1, 5, 50\}$ labeled examples per class, train a linear classifier on the K-shot training subset, and evaluate on the same held-out test set. As shown in Fig. 6, OSF consistently outperforms all other methods across three sleep analysis tasks for all shot budgets. Detailed results are provided in Appendix C.1.

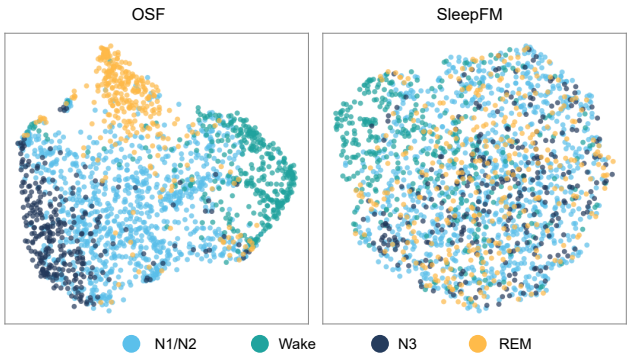


Figure 7 | **UMAP visualization of learned representations.** OSF produces more stage-separable clusters, indicating stronger alignment between the embedding geometry and sleep stage structure.

OSF outperforms physiological-signal-specific pre-training methods. We further compare OSF with five recent domain-specific baselines: TS2Vec [44], ST-MEM [24], LSM-2 [38], Ped-SleepMAE [27], and SleepGPT [14]. We pre-train PedSleepMAE on `SleepBench`, evaluate SleepGPT from its released checkpoint, and adapt the remaining methods to our benchmark. Table 4 verifies that OSF achieves the strongest performance across all downstream tasks on MROS in this comparison. More results are in the Appendix C.1.

OSF learns clinically useful features for disease prediction. We evaluate patient-level disease prediction on MROS using the same embedding aggregation for all pre-trained models. As shown in Table 5, representations learned by OSF transfer better to unseen data and consistently outperform SleepFM [36].

Table 4 | **Comparison with domain-specific pre-training methods.** OSF outperforms baselines tailored for physiological signals.

Method	Sleep Staging		Hypopnea	
	AUC↑	AUPRC ↑	AUC ↑	AUPRC ↑
TS2Vec [44]	86.9	67.3	61.9	54.5
ST-MEM [24]	96.1	87.5	75.4	60.0
LSM-2 [38]	95.8	86.7	74.6	60.1
PedSleepMAE [27]	84.0	59.0	59.4	53.4
SleepGPT [14]	95.4	85.7	65.1	55.7
OSF	**97.3**	**90.4**	**77.7**	**62.2**

The embedding space of OSF shows a more structured pattern. We further study what properties of the learned representation space may contribute to the strong performance of OSF. We visualize embeddings from OSF and SleepFM [36] using UMAP [23] on the SHHS pre-training cohort. Embeddings from OSF form clearer clusters aligned with sleep stage labels, while embeddings from SleepFM [36] appear less structured. This suggests that OSF learns more separable representations, which may help downstream adaptation.

Overall, OSF provides a practical recipe for building sleep FMs and achieves strong performance, improved sample efficiency, and effective hierarchical aggregation.

5. On Scaling Sleep FM

In this section, we study how multi-source data mixtures affect downstream performance, and we examine the scaling behavior of sleep FM pre-training with respect to sample size and model capacity.

Table 5 | **Disease prediction results.** OSF achieves the best overall performance.

Model	Coronary Dis.		Diabetes		Hypertension	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
End2End ViT	61.4	60.1	55.8	52.5	57.7	57.4
SleepFM	<u>66.5</u>	<u>65.5</u>	<u>62.1</u>	<u>54.9</u>	62.5	61.4
SimCLR	60.4	60.0	60.4	54.8	58.8	57.9
DINO	58.0	57.3	60.1	54.3	58.2	56.1
VQ-VAE	65.5	64.4	61.8	54.6	62.1	59.7
MAE	65.4	65.6	57.6	53.5	64.4	<u>63.9</u>
AR	66.5	65.3	59.5	54.4	<u>64.6</u>	63.2
OSF	**68.1**	**67.4**	**62.3**	**56.3**	**66.3**	**66.1**

Does multi-source data mixture help pre-training? Prior work often pre-trains sleep FMs on multi-source corpora. However, cohorts differ in acquisition protocols and population characteristics, and mixing datasets introduces distribution shift. To examine whether multi-source data mixture helps pre-training, we train two variants of OSF: one pre-trained on *SHHS only* and the other pre-trained on the *full* multi-cohort corpus. As shown in Fig. 8, multi-source pre-training yields consistent improvements across all tasks on the out-of-domain MROS dataset. Moreover, when we vary the pre-training sample size and train both single-source and multi-source variants under identical settings, multi-source pre-training consistently outperforms single-source pre-training. Additional results are provided in Appendix D.2.

Takeaway 1: Multi-source pre-training improves generalization by increasing data diversity.

Does scaling model size help pre-training on sleep data? To answer this question, we pre-train OSF with ViT encoders of increasing capacity, ranging from 1M and 5M to 85M parameters, and evaluate them via linear probing on hypopnea detection. As shown in Fig. 9, larger models consistently achieve better performance. We observe a similar trend for sleep staging (Appendix D). These results suggest that, given the richness and diversity of epoch-level sleep signals, increasing model capacity does not

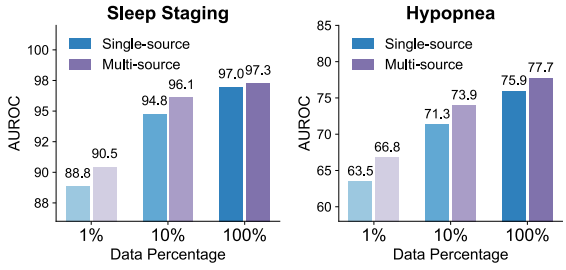


Figure 8 | **Single-source vs. multi-source pre-training.** Multi-source pre-training yields consistently better downstream performance.

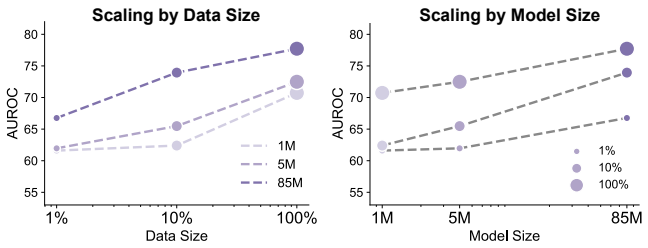


Figure 9 | **Scaling behavior.** Linear probing results on hypopnea detection show that OSF improves with both model capacity and pre-training sample size.

lead to obvious overfitting in our setting and translates into stronger transferable representations. Additional details are provided in Appendix D.1.

Does scaling sample size help pre-training on sleep data? Beyond the results in Sec. 3, we further test data scaling for OSF on harder settings and larger model sizes. Specifically, we pre-train OSF with ViT-1M, 5M, and 85M encoders using 1%, 10%, and 100% of the multi-source corpus, and evaluate via linear probing on MROS. As shown in Fig. 9, hypopnea detection performance improves consistently as the pre-training sample size increases. This indicates that pre-training performance scales positively with data size in our setting. Results on additional tasks and cohorts are provided in Appendix D.1.

Table 6 | **Robustness of pre-training with fewer channels.** We pre-train both models using only ECG and respiratory signals and evaluate downstream performance. OSF makes better use of this limited channel set.

Method	Hypopnea		Ox. Desat.	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SleepFM	71.3	58.6	79.7	78.9
OSF	**77.6**	**62.4**	**80.8**	**80.1**

Takeaway 2: Scaling laws emerge in sleep data; jointly scaling model and data size yields the strongest gains.

6. Discussion and Analysis

6.1. On Pre-training and Inference with Incomplete Channels

Evaluation under missing-channel inference. We evaluate pre-trained models under realistic missing-channel settings. Motivated by real-world deployment scenes, we consider four common study configurations: ❶ ***head-band device only*** [2]; ❷ ***sleep disorder studies*** [10]; ❸ ***sleep micro-event studies*** [11]; and ❹ ***in-home studies*** [31] that provide breathing signals only. We report linear probing results on sleep analysis tasks on MROS. As shown in Table 7, OSF consistently outperforms SleepFM across these missing-channel settings. Specifically, (1) OSF *makes better use of the available channels.* With brain-activity channels only, it achieves stronger sleep staging and arousal detection, suggesting stronger brain-related representations. Similarly, with respiratory channels only, it achieves stronger performance on hypopnea and oxygen desaturation. (2) OSF *is more robust when key modalities are missing.* When respiratory signals are removed, both methods degrade on hypopnea and oxygen desaturation, but OSF remains consistently better. Conversely, when brain-related channels are unavailable, sleep staging becomes much harder for both models; nevertheless, OSF better uses the remaining channels and yields stronger performance.

Table 7 | **Linear probing results under realistic missing-channel settings.** OSF is more robust to missing channels.

Realistic Settings Ref.	Brain	Resp.	ECG	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
					AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
Head-Band Data [2]	✓	✗	✗	SleepFM	96.6	88.7	90.8	85.5	67.5	56.7	69.9	68.9
				OSF	**97.2** (+0.6)	**90.0** (+1.3)	**92.2** (+1.4)	**87.6** (+2.1)	**68.8** (+1.3)	**57.1** (+0.4)	**70.7** (+0.8)	**69.9** (+1.0)
Sleep Disorder Study [10]	✗	✓	✓	SleepFM	86.7	64.9	85.8	79.8	75.7	60.8	80.9	80.1
				OSF	**86.8** (+0.1)	**65.3** (+0.4)	**86.4** (+0.6)	**80.5** (+0.7)	**77.6** (+1.9)	**62.2** (+1.4)	**81.1** (+0.2)	**80.4** (+0.3)
Sleep Micro-Event. Study [11]	✓	✗	✓	SleepFM	96.2	87.2	89.5	83.7	67.7	56.7	**72.5**	**71.8**
				OSF	**97.1** (+0.9)	**89.9** (+2.7)	**92.3** (+2.8)	**87.5** (+3.8)	**68.7** (+1.0)	**57.1** (+0.4)	71.6 (-0.9)	70.7 (-1.1)
In-Home Study [31]	✗	✓	✗	SleepFM	86.7	65.2	85.7	79.7	77.4	62.0	81.1	80.5
				OSF	**87.4** (+0.7)	**65.9** (+0.7)	**86.6** (+0.9)	**80.8** (+1.1)	**81.0** (+3.6)	**65.1** (+3.1)	**81.9** (+0.8)	**81.3** (+0.8)

Evaluation under more types of channel corruption. To further evaluate robustness beyond missing-channel settings, we consider two additional corruption scenarios applied to non-brain-related channels: *noise interference* and *temporal channel detachment*. We show linear probing results on MROS dataset in Table 8, OSF achieves the best performance in most settings and tasks, suggesting that its robustness extends beyond channel removal to other practically relevant forms of channel corruption. The advantage is particularly clear under temporal detachment, which is consistent with our use of channel masking and time-wise masking during pre-training.

Table 8 | **Inference with different types of channel corruption.** OSF demonstrates stronger robustness across practical channel corruption scenarios.

Setting	Method	Sleep Staging		Hypopnea	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑
Random Noise	SleepFM	96.1	87.1	**73.1**	**59.7**
	OSF	**97.1**	**90.0**	69.4	57.5
Temporal Detach	SleepFM	96.1	87.0	69.1	57.5
	OSF	**97.3**	**90.4**	**76.2**	**61.2**

Pre-training with fewer channels. Brain-related channels provide strong cues for many sleep analysis tasks, but they are inconvenient for patients to collect. We therefore study a practical pre-training setup that uses only ECG and respiratory channels [10]. We evaluate on hypopnea and oxygen desaturation, since these tasks are primarily driven by respiratory and cardiac physiology. As shown in Table 6, OSF makes better use of the available channels and consistently outperforms SleepFM by a large margin on both hypopnea and oxygen desaturation.

Overall, OSF better addresses practical constraints in inference and pre-training with incomplete channels. However, a meaningful performance gap remains when task-critical channels are absent. We encourage future work to further improve robustness under channel missingness.

6.2. Ablation on Design Variations

How to aggregate epoch-level embeddings for patient-level tasks? Patient-level disease prediction requires aggregating epoch-level representations into a single patient-level feature. Given a full-night recording, we segment it into N non-overlapping 30-second epochs, feed each epoch into the pre-trained encoder, and obtain a sequence of embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$ where $\mathbf{z}_i \in \mathbb{R}^D$. Prior work [36] typically uses simple pooling or a learnable temporal aggregator (e.g., an LSTM [13]) to summarize this sequence. However, downstream datasets often contain only a limited number of patient-level recordings, which can increase the risk of overfitting. This motivates

Table 9 | **Comparisons of different embedding aggregation methods.** Top-k selection consistently outperforms mean pooling for patient-level prediction.

Aggregation	Coronary Dis.		Diabetes		Hypertension	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
Avg Pool	65.4	64.6	60.0	55.4	**64.2**	**64.1**
LSTM	63.5	63.0	59.9	55.9	61.8	60.9
MIL [15]	66.1	65.5	58.6	53.7	62.3	61.4
Top-k Selection	**66.8**	**66.3**	**62.8**	**56.3**	64.1	63.5

us to test whether a learnable aggregation module is necessary.

We hypothesize that disease-related signals concentrate in specific sleep periods, and thus simple average pooling may be suboptimal for aggregating epoch embeddings. To test this idea, we evaluate multiple-instance learning (MIL) [15] and an LSTM-based aggregator. We also test a selection module trained with straight-through estimation to select and average the top-k disease-relevant epoch embeddings. We apply each method to embeddings extracted from multiple pre-trained models and average performance across models. As shown in Table 9, top-k selection outperforms mean pooling overall.

Takeaway 3: Patient-level prediction benefits from better joint modeling of epoch-level embedding.

Vision-style augmentations transfer to sleep data, but are not necessary for strong performance. To test whether standard vision augmentations (e.g., cropping) can be adapted to sleep data, we implement temporal cropping by randomly selecting a contiguous segment of each input signal, where the crop ratio r is sampled uniformly from $0.25, 0.75$. We pre-train models under the same configuration using (1) crop-only augmentation and (2) cropping combined with our channel-masking augmentation. As shown in Table 10, cropping alone yields reasonable performance but remains worse than channel masking alone. Moreover, adding cropping on top of our augmentation does not yield consistent gains. These results suggest that additional augmentations can help in some cases, but are not required to obtain strong performance in our setting.

Table 10 | **Comparison of augmentation strategies.** We report linear-probing performance (AUC$^{\uparrow}$). Standard vision augmentations offer limited additional benefit beyond channel masking.

Crop	Time	Channel	Sleep Staging	Arousal	Hypopnea	Ox. Desat.
✓	✗	✗	95.4	85.7	77.2	80.6
✗	✗	✓	96.4	90.5	77.1	80.3
✓	✓	✓	96.6	89.3	77.3	81.0
✗	✓	✓	**97.3**	**92.8**	**77.7**	**81.5**

Ablation on channel masking ratio. We further ablate the channel masking ratio by pre-training two additional variants with masking ratios of 0.1 and 0.9, while keeping all other settings unchanged. As shown in Table 11, the default ratio of 0.5 achieves the best overall linear probing performance across tasks, suggesting that moderate masking better balances pretext-task difficulty and input preservation for downstream transfer. More results are in Appendix E.

Table 11 | **Ablation on channel masking ratios.** A moderate masking ratio works the best.

Masking Ratio	Sleep Staging		Hypopnea	
	AUC$^{\uparrow}$	AUPRC$^{\uparrow}$	AUC$^{\uparrow}$	AUPRC$^{\uparrow}$
0.1	95.9	87.2	77.4	61.5
0.5	**97.3**	**90.4**	**77.7**	**62.2**
0.9	95.3	85.9	75.0	60.3

7. Related Work

Physiological Foundation Models. Foundation models (FMs) have emerged as a strong approach for analyzing physiological time-series data [14, 39, 43]. Building on this line of work, [36] propose SleepFM for sleep physiology, showing strong transfer performance on downstream tasks such as sleep staging and disease prediction. SleepFM partitions channels into modality groups and applies contrastive objectives to align their embeddings. Despite these advances, the sleep domain still lacks a clear understanding of which pre-training design choices drive strong representations, and it remains unclear whether scaling trends reported for other physiological FMs [25, 47] reliably hold for sleep data. Our work fills this gap by establishing a fully open multi-source benchmark `SleepBench` and conducting a controlled study of pre-training objectives and scaling behavior for sleep foundation models.

Self-Supervised Learning. Self-supervised learning is widely used to build FMs across domains. Invariance-based methods, including contrastive learning [6, 40] and self-distillation [26], have achieved strong results, while reconstruction-based methods such as MAE [12] and VQ-VAE [37] are also effective. In language and time-series modeling, autoregressive pre-training has been especially successful [1, 8, 30]. However, existing sleep-focused studies cover only a subset of this design space [27, 36] and lack a unified framework for evaluating and comparing methods. As a result, the field still lacks a systematic analysis of how pre-training design choices affect downstream performance for sleep FMs. In contrast, we benchmark representative objectives from these SSL families under a unified protocol and derive practical design guidelines for scalable sleep FM pre-training.

8. Conclusion

In this paper, we present a systematic study of key design choices for building sleep FMs through controlled experiments on a fully open, multi-source corpus. Through open benchmarking, we identify three main findings and establish best practices for pre-training sleep FMs. Guided by these insights, we scale both pre-training data and model capacity, explicitly encourage channel-invariant feature learning, and develop the OSF family of sleep FMs, which consistently outperforms existing models. Extensive evaluations across nine datasets and eight downstream tasks demonstrate the effectiveness, robustness, and intriguing behaviors of OSF.

Acknowledgments

We gratefully acknowledge the support by Amazon Science Hub and UCLA DataX. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[2] Pierrick J Arnal, Valentin Thorey, Eden Debellemaniere, Michael E Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, et al. The dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, 43(11):zsaa097, 2020.

[3] Indu Ayappa, Robert G Norman, Vijay Seelall, and David M Rapoport. Validation of a self-applied unattended monitor for sleep disordered breathing. *Journal of Clinical Sleep Medicine*, 4(1):26–37, 2008.

[4] Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011.

[5] Jonathan F Carter and Lionel Tarassenko. wav2sleep: A unified multi-modal approach to sleep stage classification from physiological signals. In *Machine Learning for Health (ML4H)*, pages 186–202. PMLR, 2025.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[7] 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, 2015.

[8] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. *arXiv preprint arXiv:2505.14766*, 2025.

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Kevin Gleason, Donghoon Shin, Michael Rueschman, Tanya Weinstock, Rui Wang, James H Ware, Murray A Mittleman, and Susan Redline. Challenges in recruitment to a randomized controlled study of cardiovascular disease reduction in sleep apnea: an analysis of alternative strategies. *Sleep*, 37(12):2035–2038, 2014.

[11] Naihua N Gong, Aditya Mahat, Samya Ahmad, Daniel Glaze, Mirjana Maletic-Savatic, Matthew McGinley, Anne Marie Morse, Alcibiades J Rodriguez, Audrey Thurm, Susan Redline, et al. Leveraging clinical sleep data across multiple pediatric cohorts for insights into neurodevelopment: the retrospective analysis of sleep in pediatric (rasp) cohorts study. *Sleep*, page zsaf157, 2025.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Weixuan Huang, Yan Wang, Hanrong Cheng, Wei Xu, Tingyue Li, Xiuwen Wu, Hui Xu, Pan Liao, Zaixu Cui, Qihong Zou, et al. A unified time-frequency foundation model for sleep decoding. *Nature Communications*, 2026.

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[16] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and Li-wei H Lehman. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

[17] Magnus Ruud Kjaer, Rahul Thapa, Gauri Ganjoo, Hyatt Moore IV, Poul Joergen Jennum, Brandon M Westover, James Zou, Emmanuel Mignot, Bryan He, and Andreas Brink-Kjaer. Stanford sleep bench: Evaluating polysomnography pre-training methods for sleep foundation models. *arXiv preprint arXiv:2512.09591*, 2025.

[18] 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, 2022.

[19] Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. Hearts: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.

[20] Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23(9):4221, 2023.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] 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, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368(25):2366–2376, 2013.

[23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[24] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

[25] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[27] Saurav Raj Pandey, Aaqib Saeed, and Harlin Lee. Pedsleepmae: Generative model for multimodal pediatric sleep signals. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.

[28] 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, 2021.

[29] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

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

[31] Susan Redline, Daniela Sotres-Alvarez, Jose Loredo, Martica Hall, Sanjay R Patel, Alberto Ramos, Neomi Shah, Andrew Ries, Raanan Arens, Janice Barnhart, et al. Sleep-disordered breathing in hispanic/latino individuals of diverse backgrounds. the hispanic community health study/study of latinos. *American journal of respiratory and critical care medicine*, 189(3):335–344, 2014.

[32] Susan Redline, Peter V Tishler, Tor D Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, 151(3):682–687, 1995.

[33] 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, 2003.

[34] Scott A Sands, Philip I Terrill, Bradley A Edwards, Luigi Taranto Montemurro, Ali Azarbarzin, Melania Marques, Camila M De Melo, Stephen H Loring, James P Butler, David P White, et al. Quantifying the arousal threshold using polysomnography in obstructive sleep apnea. *Sleep*, 41(1):zsx183, 2018.

[35] Adam P Spira, Terri Blackwell, Katie L Stone, Susan Redline, Jane A Cauley, Sonia Ancoli-Israel, and Kristine Yaffe. Sleep-disordered breathing and cognition in older women. *Journal of the American Geriatrics Society*, 56(1):45–50, 2008.

[36] Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore IV, Umaer Hanif, Gauri Ganjoo, M Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, et al. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, pages 1–11, 2026.

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[38] Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

[39] Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S Aysola, Rajesh Kumar, and Yuzhe Yang. Sleeplm: Natural-language intelligence for human sleep. *arXiv preprint arXiv:2602.23605*, 2026.

[40] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. In *The Eleventh International Conference on Learning Representations*, 2023.

[41] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature Medicine*, 28(10):2207–2215, 2022.

[42] 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, 2009.

[43] 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):91, 2024.

[44] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8980–8987, 2022.

[45] Bing Zhai, Greg J Elder, and Alan Godfrey. Challenges and opportunities of deep learning for wearable-based objective sleep assessment. *npj Digital Medicine*, 7(1):85, 2024.

[46] 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*, 25(10):1351–1358, 2018.

[47] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025.

A. Implementation Details

A.1. Detailed Pre-training Setups

In this section, we provide implementation details of our pre-training configurations for both the benchmarked baselines and our method. For our explorations, we aim to directly adapt well-established self-supervised pre-training methods, and benchmark their performance on the sleep data. In addition to the existing pre-training method SleepFM [36], we have considered four representative families of SSL objectives:

(1) **Contrastive learning**: SimCLR [6];

(2) **Reconstruction-based methods**: MAE [12] and VQ-VAE [37];

(3) **Autoregressive modeling**: Autoregression [30];

(4) **Self-distillation**: DINO [26].

For OSF model family, we use DINO as a specification and incorporate it with our identified pre-training recipes, and form OSF used in the main experiments. We have also incorporated OSF with SimCLR, whose performance in the main evaluation tasks is shown in Table 22. We consider them because these invariance-based methods are a natural testbed for applying our pre-training recipes. Specifically, contrastive learning and self-distillation-based methods are invariance-based methods, which aim to align two (or more) augmented views of the same input in the latent space. As a result, their downstream performance usually depends on the approach we use to augment the raw input data.

Backbone and training protocol. Unless otherwise specified, all methods reported in the main tables use the same ViT-85M Transformer backbone [9]. We replace the patchify layer in the original ViT with a simple convolution layer to project the input signal into tokens, and the architecture details are in Table 12. We pre-train all models on the same split of the pre-training cohorts in our `SleepBench`. Each run is trained for up to 30 epochs, with early stopping when we observe training converges, based on the loss curves.

Table 12 | **Encoder backbone configurations used in our experiments.**

Model	Architecture				Params (M)
	Width	Depth	Heads	MLP dim	
ViT-1M	128	6	4	512	1.7M
ViT-5M	192	12	3	768	5.5M
ViT-85M	768	12	12	3072	85.5M

Optimization and scheduling. We use AdamW [21] for all methods. For the scheduler of learning rate, we follow a standard configuration of self-supervised pre-training, and use a linear warmup for the first 10% of total training steps, and then apply cosine annealing decay. We have tuned the learning rate and batch size for each method independently to ensure a fair comparison under its best-performing regime. We first search batch size, for the maximized usage of GPU memory; and then optimize learning rate choices based on the performance.

Augmentations for OSF. For OSF, as discussed in Sec. 3, we adopt a two-stage masking augmentation. We first apply channel masking by randomly dropping 50% of input channels. We then apply block-wise temporal masking, where the temporal masking ratio is independently sampled for each training example from a uniform distribution over $0.3, 0.6$.

The complete pre-training configuration, including detailed hyper-parameters and method-specific setup, is summarized in Table 13.

Table 13 | **Pretraining configurations across different methods.**

Configuration	SleepFM	SimCLR	DINO	MAE	VQ-VAE	AR	OSF (SimCLR)	OSF (DINO)
Max Epoch					30			
Batch Size	2560	3200	1024	9600	3200	3200	3200	1024
Base Learning Rate	1.00E-04	1.00E-04	5.00E-05	3.00E-04	1.00E-04	1.00E-04	1.00E-04	5.00E-05
Optimizer					AdamW			
Opt. momentum β_1, β_2					$0.9, 0.95$			
Weight Decay	0.2	0.2	0.2	0.2	0.2	0.05	0.2	0.2
Gradient Clipping	–	–	–	–	3	–	–	3
Learning Rate Scheduler					LinearWarmup & CosineDecay			
Token window length					64			

Table 14 | **Training hyperparameters for sleep analysis tasks and disease prediction tasks.**

Hyperparameter	Sleep Analysis Tasks			Disease Prediction				
	Linear Probing	Finetuning	Supervised	Mean Pool	MIL	LSTM	TopK	End2End
Base learning rate	0.1	1e-4	1e-3	1e-2	5e-3	5e-3	5e-3	1e-3
Batch size	800 per GPU			256 per GPU	128 per GPU			16 per GPU
Max step	500			–				
Max epoch	–		5	50				5
Learning rate scheduler	LinearWarmup & CosineDecay							
Optimizer	AdamW							

A.2. Detailed Downstream Evaluation Setups

Evaluation setup. We evaluate pre-trained models on three types of downstream tasks under three standard transfer learning protocols. For epoch-level evaluation, we consider a 4-class sleep staging task, and four sleep-event detection tasks. For sleep staging, we follow [5] and merge N1 and N2 into a single light sleep category, and predict four classes: awake, light sleep, deep sleep, and REM. For sleep-event detection, we perform 2-class classification for the four following event detection tasks: arousal, hypopnea, oxygen desaturation, and central apnea. We also consider three patient-level disease classification: coronary disease, diabetes, and hypertension. We report both AUC and AUPRC since all tasks are class-imbalanced.

For epoch-level evaluation, we consider three transfer learning settings: linear probing, full fine-tuning, and few-shot adaptation. In linear probing, we freeze the encoder and train a linear classifier on the full training split of each target dataset, then evaluate on the held-out test set. In full fine-tuning, we adopt a similar setting on data usage, and in addition to training a linear classifier, we also jointly optimize the encoder and classifier. In few-shot adaptation, we sample K examples per class from the training split. In particular, we consider $K \in 1, 5, 50$ in our main experiments. We keep the encoder frozen, train a linear classifier, and evaluate on the same test split across methods.

For patient-level prediction, we segment each recording into non-overlapping 30-second epochs and extract an embedding for each epoch using the pre-trained encoder. For each pre-trained checkpoint, we pre-extract epoch embeddings for all recordings before training and evaluating on the downstream disease prediction tasks. During embedding extraction, we pad the epoch sequence length to 1200, corresponding to 10 hours. During downstream training, we consider two approaches of aggregating the epoch embeddings into a single patient-level representation: (i) we learn a lightweight aggregation module on top of epoch embeddings of each patient, or (ii) we use simple mean pooling over all epochs. We then feed the aggregated patient-level representation to a classifier for disease prediction.

Detailed configurations of evaluation tasks can be seen in Table 14.

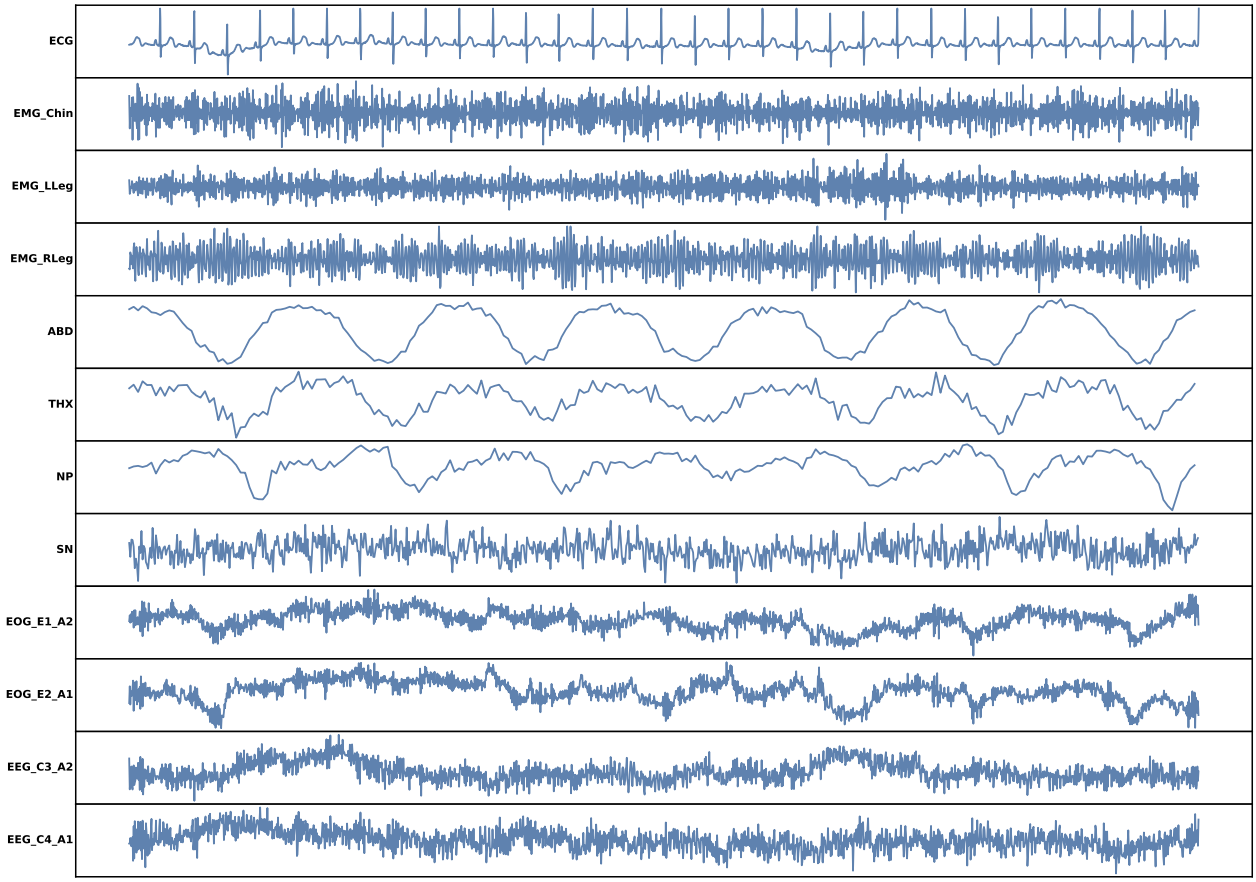


Figure 10 | **Visualization of the pre-processed 12 channels of the sleep data used in our paper.** We pick a 30-sec epoch for demonstration.

B. Details of SleepBench

B.1. Dataset Statistics

To enable open benchmarking and our controlled explorations, we curated a large-scale benchmark SleepBench, comprising nine diverse sleep datasets: SHHS [29], NCHSDB [18], WSC [42], CCSHS [33], CFS [32], MROS [4], MESA [7], CHAT [22], and SOF [35]. This consolidated corpus represents a significant variety of demographic profiles and sleep disorder prevalences, totaling 20M epochs of sleep recording data. Table 16 and Table 17 report the statistics of patients and 30-second epochs, across different datasets and data splits.

Table 15 | **Comparison between SleepBench and existing open-source sleep datasets.**

Dataset	Num. Recordings	Num. Hours
SSC [17]	17,467	163,000
SLEEPBENCH	21,482	166,500

B.2. Downstream Task Labels

Sleep Event Distribution. We report the prevalence of sleep events of datasets used in main experiments, in the Table 18. The datasets exhibit substantial heterogeneity in event distributions, which enables the evaluation of the generalization capability of pre-trained models. During dataset splitting,

Table 16 | **Distribution of sleep recordings across different datasets and splits.** Models are pretrained on pretrain datasets and evaluated on both pretrain and out-of-domain (OOD) datasets. The test split of pretrain data and all OOD data are unseen during pre-training.

Split	Pretrain Datasets					O.O.D. Datasets				Total
	SHHS	NCHSDB	WSC	CCSHS	CFS	MROS	MESA	CHAT	SOF	
Train	6,663	3,094	1,397	403	269	3,101	1,367	531	361	17,186
Valid	811	392	182	59	34	385	174	66	45	2,148
Test	802	403	194	53	26	381	180	67	42	2,148
Total	8,276	3,889	1,773	515	329	3,867	1,721	664	448	21,482

Table 17 | **Distribution of 30-second epochs across different datasets and splits.**

Split	Internal Pretrain Datasets					External Eval. Datasets				Total
	SHHS	NCHSDB	WSC	CCSHS	CFS	MROS	MESA	CHAT	SOF	
Train	5,919k	2,892k	1,270k	432k	255k	3,015k	1,312k	562k	341k	15,997k
Valid	724k	367k	164k	64k	32k	370k	164k	69k	42k	1,996k
Test	711k	376k	175k	57k	24k	372k	172k	70k	38k	1,995k
Total	7,354k	3,634k	1,609k	553k	311k	3,757k	1,648k	701k	421k	19,989k

Table 18 | **Distributions of sleep event detection tasks across datasets used for sleep event evaluation.** For each sleep disorder event, we report its prevalence in each split of each used dataset.

Dataset	Split	Epochs	Arousal	Hypopnea	Oxygen Desat.	Central Apnea
SHHS	Train	5,919,217	17.85%	29.13%	23.46%	0.71%
SHHS	Valid	724,056	17.73%	29.00%	23.71%	0.74%
SHHS	Test	710,821	17.60%	29.00%	23.32%	0.71%
MROS	Train	3,014,915	19.46%	12.27%	52.28%	1.32%
MROS	Valid	370,378	19.55%	12.62%	52.65%	1.66%
MROS	Test	371,873	19.77%	12.26%	51.93%	1.42%

we perform stratified sampling at the patient level to prevent information leakage and to keep event distributions consistent across splits. Specifically, we first randomly assign patients to the training, validation, and test splits. We then assign each epoch to the split of its corresponding patient via patient ID. As shown in Table 18, the data was divided into train, validate, and test sets with an approximate global ratio of 80:10:10 strictly on the subject level.

Disease Classification: For more clinically valuable downstream tasks, we perform patient-level disease classification on the MROS dataset. We focused on three diagnosis tasks: Coronary Disease, Diabetes, and Hypertension. These labels are formed from disease-related variables provided MROS dataset. We set the disease label as positive, if the patient has self-reported related medication usage of diagnosis history. As detailed in Table 19, the MROS dataset was split into training, validation, and testing sets. The three labels cover both class-balance and class-imbalance scenes, therefore more comprehensively evaluating the pre-trained model's ability on disease prediction tasks.

Table 19 | **Distribution of subject-level disease labels in the MROS dataset.** We filtered out patients with NaN values on the selected labels. Here we report the number and percentage of both negative and positive samples across splits.

Disease Category	Split	Negative		Positive		Total Samples
		Count	Ratio	Count	Ratio	
Coronary Disease	Train	1,626	71.3%	654	28.7%	2,280
	Valid	188	67.1%	92	32.9%	280
	Test	183	64.2%	102	35.8%	285
Diabetes	Train	1,987	86.9%	300	13.1%	2,287
	Valid	247	88.2%	33	11.8%	280
	Test	248	87.0%	37	13.0%	285
Hypertension	Train	1,152	50.4%	1,134	49.6%	2,286
	Valid	138	49.3%	142	50.7%	280
	Test	142	49.8%	143	50.2%	285

B.3. Dataset Pre-processing Details

As mentioned in Sec. 2, we utilize a standardized 12-channel montage covering 4 physiological modality groups: ❶ *Brain (EEG/EOG):* `C3-A2`, `C4-A1`, `E1-A2`, `E2-A1`, ❷ *Respiration:* `Abdominal`, `Thorax`, `Nasal Pressure`, `Snore`, ❸ *Cardiac:* `ECG`, and ❹ *Somatic:* `EMG-Chin`, `EMG-LLeg`, `EMG-RLeg`. The detailed channel descriptions are in Table 20.

For the pre-processing of the full-night recordings, we follow [5], and first resample each channel to a unified sampling frequency, as detailed in Table 20. Next, we conduct manual quality control by trimming prolonged wake periods at the beginning and end of the night, which also removes extreme artifacts introduced during sensor setup and removal. We then apply per-night z-score normalization to each channel, followed by clipping the normalized signals to the range $-6, 6$ to suppress outliers. The clipping thresholds are chosen based on the empirical distribution of z-scored signals. For respiratory channels, we additionally apply area-dependent z-score normalization to reduce amplitude drift and improve cross-night consistency.

We follow prior work and segment each night into non-overlapping 30-second epochs, yielding roughly **20 million** epochs for self-supervised pre-training. We resample all channels in the segmented epochs to 64 Hz and zero-pad missing channels. The pre-processed data in `SleepBench` is demonstrated in Fig. 10.

C. Additional Results and Analysis

C.1. Additional Results of Sleep Analysis Tasks

OSF achieves state-of-the-art on sleep analysis tasks on in-domain evaluation set. We also conduct evaluation on sleep analysis tasks on in domain dataset SHHS. As shown in Table 21. OSF achieves the best performance under both linear probing and full fine-tuning. The same trend also holds on the out-of-domain cohorts reported in the main text, further supporting the superiority of our method.

OSF achieves state-of-the-art on few-shot adaptation tasks. We conduct few-shot learning evaluations on the MROS dataset, on three sleep event detection tasks and the sleep staging task, across

Table 20 | **Channel configuration used for main experiments.** During post-processing, they are resampled to 64 Hz.

Group	Channel	Full name	Electrode Detail	Preprocessing Frequency
Cardiac	**ECG**	Electrocardiography	ECG left - ECG right	128 Hz
Somatic	**EMG**	Electromyography	EMG_Chin	64 Hz
			EMG from left leg	64 Hz
			EMG from right leg	64 Hz
Resp. Effort	**ABD**	Abdominal respiratory effort		8 Hz
	THX	Thoracic respiratory effort		8 Hz
	NP	Nasal pressure		8 Hz
	SN	Snore		32 Hz
Brain Activity	**EOG**	Electrooculography	EOG E1 - A2	64 Hz
			EOG E2 - A1	64 Hz
	EEG	Electroencephalography	EEG C3 - A2	64 Hz
			EEG C4 - A1	64 Hz

1-shot, 5-shot, and 50-shot settings. OSF achieves the overall best performance under 1-shot and 5-shot evaluation (Tables 23 and 24), and performs best on all tasks under 50-shot evaluation (Table 25).

OSF outperforms physiological-signal-specific pre-training methods. To assess whether methods tailored to physiological signals transfer effectively to sleep foundation modeling, we further compare OSF with five recent domain-specific baselines: TS2Vec [44], ST-MEM [24], LSM-2 [38], PedSleepMAE [27], and SleepGPT [14]. We pre-train PedSleepMAE on `SleepBench`, evaluate SleepGPT using its released checkpoint, and adapt the remaining methods to our benchmark following their original settings when applicable. As shown in Table 26, OSF consistently achieves the strongest performance on all of the downstream tasks on MROS dataset.

Table 27 | **Disease prediction results on SHHS.** OSF outperforms the baseline method on most metrics.

Method	Diabetes		Hypertension	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SleepFM	79.5	56.5	76.7	**76.1**
OSF	**79.8**	**58.1**	**76.8**	74.7

OSF improves in-domain disease prediction. We further evaluate in-domain disease prediction on SHHS using two additional outcomes, *hypertension* and *diabetes*. As shown in Table 27, OSF consistently outperforms the compared baselines, further supporting the effectiveness of our pre-training strategy for clinically relevant downstream prediction tasks.

Confidence intervals. We additionally report confidence intervals for the main results to enable a more reliable comparison across methods. As shown in Table 28, the confidence intervals show that the performance gains of OSF over competing baselines are consistent across tasks, further supporting the robustness of our improvements.

Additional evaluation metrics for sleep staging. Since several downstream tasks are label-imbalanced, we use AUROC and AUPRC as the primary threshold-independent metrics. For

Table 29 | **Additional sleep staging evaluation metrics.** OSF remains the best.

Method	Acc. (micro)	F1 (weighted)	Cohen Kappa
SleepFM	85.2	85.0	<u>79.9</u>
SimCLR	63.2	59.5	25.2
DINO	79.4	78.8	63.9
VQ-VAE	83.6	83.4	74.6
MAE	<u>85.4</u>	<u>85.2</u>	79.6
AR	84.0	83.7	75.7
OSF	**87.0**	**86.9**	**82.8**

Table 21 | **Sleep staging and sleep event detection results on SHHS**. OSF achieves overall best performance among all compared methods.

Eval	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.		Central Apnea	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
Supervised ViT		97.0	92.3	88.7	82.3	79.1	74.8	76.0	70.1	95.2	61.7
LP	SleepFM	96.6	91.1	92.2	86.3	82.0	77.0	80.6	75.0	95.9	64.7
	SimCLR	86.1	67.4	71.9	64.2	64.9	60.7	62.9	58.5	74.6	51.3
	DINO	95.1	87.4	87.0	79.9	78.3	73.7	76.5	70.5	94.8	60.6
	VQ-VAE	96.7	91.3	89.4	83.0	80.2	75.4	78.0	72.2	95.7	63.6
	MAE	96.9	91.9	91.4	85.7	**83.0**	**78.4**	80.5	75.0	96.7	66.0
	AR	96.9	91.8	90.6	88.4	79.0	73.3	78.6	72.9	94.4	61.2
	OSF	**97.5**	**93.2**	**94.0**	**89.3**	82.8	78.1	**81.3**	**75.6**	**96.9**	**66.5**
FT	SleepFM	97.0	92.0	95.1	91.0	86.7	83.3	82.5	77.1	97.1	68.6
	SimCLR	95.4	88.3	85.8	78.4	80.3	76.3	75.5	69.6	95.7	64.1
	DINO	97.1	92.4	93.8	89.3	86.2	83.0	81.8	76.1	97.4	69.0
	VQ-VAE	97.5	93.4	94.7	90.7	85.5	82.2	81.7	76.2	97.2	68.4
	MAE	97.4	93.0	94.6	90.6	87.1	84.2	83.2	77.9	**97.7**	68.5
	AR	97.4	93.1	94.2	90.0	86.4	83.3	82.4	77.1	97.2	68.2
	OSF	**97.9**	**94.3**	**95.7**	**92.1**	**87.6**	**84.6**	**83.7**	**78.6**	97.6	**69.8**

Table 22 | **Sleep staging and sleep event detection results on SHHS.** SimCLR can also be upgraded by our pre-training recipts and achieve better performance.

Eval	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.		Central Apnea	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
	SimCLR	86.1	67.4	71.9	64.2	64.9	60.7	62.9	58.5	74.6	51.3
	OSF (SimCLR)	97.3	92.9	92.9	87.4	82.2	77.2	80.4	74.7	96.3	64.8

completeness, we also report Accuracy and weighted F1 for sleep staging results shown in the main table, as shown in Table 29. The results are consistent with the main findings, with OSF remaining the strongest overall method.

Demographic analysis. We further evaluate subgroup generalization within cohorts using age subgroups in CFS dataset and gender subgroups in MESA dataset. For each pretrained model, we fine-tune and evaluate the downstream classifier separately within each subgroup. As shown in Tables 30, OSF outperforms SleepFM in most subgroup-task settings and shows lower performance variation across demographic groups. These results suggest that the benefits of OSF extend beyond cohort-level averages to subgroup-level evaluation within heterogeneous sleep cohorts.

C.2. Additional Analysis Studies

Channel masking consistently improves downstream performance across cohorts and tasks. Here, we report additional experiments on more downstream tasks and data cohorts. In Table 32, we report results across 3 sleep disorder event detection tasks and sleep staging tasks, on both the out-of-domain dataset MROS and the in-domain dataset SHHS. Under these broader evaluation settings, we observe that channel masking remains a strong augmentation strategy, compared to crop or time-masking strategies. Other findings in Sec. 3.2 and Sec. 6.2 also hold. These results suggest that encouraging channel-invariant representations is important for improving downstream performance.

Table 23 | **1-shot few-shot results on MROS.** OSF shows better adaptability to unseen out-of-domain data.

Few-shot	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
$k=1$	SleepFM	59.6	32.3	56.6	53.8	49.3	49.8	55.2	54.1
	SimCLR	50.8	25.1	51.2	50.6	49.9	50.0	50.9	50.8
	DINO	67.0	41.6	52.7	51.2	45.8	48.7	**60.2**	**58.9**
	MAE	63.0	36.6	49.3	50.0	49.1	49.9	54.4	53.4
	VQ-VAE	62.6	33.7	56.0	53.3	50.8	50.3	55.8	54.7
	AR	67.6	43.4	48.0	49.4	47.4	49.4	49.3	49.4
	OSF	**75.9**	**48.9**	**61.4**	**57.0**	**55.4**	**51.8**	56.6	55.7

Table 24 | **5-shot few-shot results on MROS.** OSF shows better adaptability to unseen out-of-domain data.

Few-shot	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
$k=5$	SleepFM	71.8	45.3	55.5	53.0	49.3	49.8	59.8	58.6
	SimCLR	53.0	27.1	54.2	52.2	50.0	50.0	50.7	50.5
	DINO	82.8	55.8	56.8	53.3	47.7	49.2	57.8	56.1
	MAE	82.6	59.2	56.7	53.7	49.7	49.9	**62.9**	**61.2**
	VQ-VAE	78.1	51.8	52.2	51.1	51.1	50.4	58.9	57.6
	AR	81.9	60.1	57.1	54.0	50.9	50.4	55.1	53.9
	OSF	**89.2**	**70.1**	**67.9**	**61.3**	**55.9**	**52.0**	56.4	55.7

Our identified pre-training recipes also apply to other pre-training methods. To verify whether our pre-training recipes generalize to other self-supervised learning methods, we adopt our identified pre-training recipes to the original SimCLR objective, yielding an upgraded variant (OSF (SimCLR)). We evaluate both SimCLR and OSF (SimCLR) models on the in-domain SHHS cohort. As shown in Table 22, the upgraded SimCLR (OSF (SimCLR)) consistently outperforms the original SimCLR by a large margin across all tasks.

Table 33 | **Heart-rate prediction results on SHHS and MROS**. Reconstruction-based SSL objectives help the regression task.

Method	MROS		SHHS	
	MAE↓	RMSE↓	MAE↓	RMSE↓
SleepFM	3.38	5.00	3.52	5.06
SimCLR	8.11	10.09	8.14	10.06
DINO	7.39	9.26	7.40	9.23
VQ-VAE	4.03	5.79	3.50	5.22
MAE	2.33	3.99	2.04	3.63
AR	**2.17**	**3.91**	**1.89**	**3.56**
OSF	3.22	4.94	3.42	4.94
Mean	7.97	10.30	7.97	10.15

Ablation on data normalization strategy. We further compare our per-night per-channel z-score normalization with min-max normalization, which is commonly used in prior physiological self-supervised learning studies. Using the same model architecture and training protocol, we pre-train a variant with min-max normalized inputs and evaluate it on SHHS dataset and MROS dataset for sleep staging and hypopnea detection. As shown in Table 34, per-night z-score normalization achieves better performance across both datasets and tasks, suggesting that it better handles cross-subject and cross-cohort variation in sleep signal scale.

Additional UMAP visualizations on imbalanced downstream tasks. UMAP is computed from 2,000 randomly sampled epochs from the unseen SHHS test split. We visualize embeddings for hypopnea

Table 25 | **50-shot few-shot results on MROS.** OSF shows better adaptability to unseen out-of-domain data on all tasks.

Few-shot	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
$k=50$	SleepFM	85.9	64.0	72.3	65.8	56.3	52.3	61.7	60.3
	SimCLR	56.6	29.1	55.8	53.1	50.6	50.2	49.5	49.6
	DINO	87.4	67.0	72.1	64.5	57.2	52.7	63.5	61.3
	MAE	92.1	79.6	76.5	68.9	60.2	53.6	66.8	65.1
	VQ-VAE	85.3	63.3	68.6	62.3	51.5	50.5	65.4	63.7
	AR	90.8	75.4	72.0	64.8	58.8	53.2	63.0	61.4
	OSF	**94.5**	**82.5**	**83.9**	**77.3**	**62.9**	**54.9**	**68.4**	**66.7**

Table 26 | **Comparison with domain-specific pretraining methods.** OSF outperforms baselines tailored for physiological signals.

Method	SHHS				MROS			
	Sleep Staging		Hypopnea		Sleep Staging		Hypopnea	
	AUC ↑	AUPRC ↑	AUC ↑	AUPRC ↑	AUC ↑	AUPRC ↑	AUC ↑	AUPRC ↑
TS2Vec [44]	90.2	76.7	68.6	64.3	86.9	67.3	61.9	54.5
ST-MEM [24]	97.0	92.2	81.3	76.3	96.1	87.5	75.4	60.0
LSM-2 [38]	95.9	89.6	80.3	75.4	95.8	86.7	74.6	60.1
PedSleepMAE [27]	89.2	73.4	64.8	61.1	84.0	59.0	59.4	53.4
SleepGPT [14]	96.2	90.0	71.3	66.4	95.4	85.7	65.1	55.7
OSF	**97.5**	**93.2**	**82.8**	**78.1**	**97.3**	**90.4**	**77.7**	**62.2**

and arousal events, which are substantially more imbalanced than sleep staging. As shown in Fig. 11, although rare-event tasks show less clean separation, OSF generally exhibits clearer task-related structure than the baseline.

Reconstruction objectives transfer better on regression tasks on sleep data. As shown in Table 2, reconstruction-based methods such as MAE, VQ-VAE, and AR can be adapted to sleep signals with minimal modification and achieve reasonable performance; MAE even outperforms SleepFM [36] in many cases. In contrast, vanilla SimCLR and DINO lag behind both reconstruction-based methods and SleepFM. This indicates that invariance-based objectives require domain-appropriate augmentations to work well on sleep data. We also designed a heart-rate prediction task to test each pre-trained model's capability in regression, and found that MAE and AR achieve good performance, as shown in Table 33.

D. Additional Results on Scaling Analysis

D.1. Scaling across Pre-training Sample Scale and Model Capacity

Scaling pre-training data leads to larger gains in few-shot adaptation. Besides training with the full training split of the downstream task, we also tested the sample efficiency brought by scaling the pre-training sample size. Particularly, we run 1-shot, 5-shot, and 50-shot adaptation experiments to test whether increasing the pre-training sample size improves few-shot transfer. We conduct these experiments on MROS. As shown in Fig. 12, in both the 1-shot and 50-shot settings, models pre-trained

Table 28 | **Confidence intervals for sleep staging and sleep event detection.** OSF achieves the best overall performance among all compared methods with statistical significance.

Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SleepFM	96.4 ± 0.04	87.7 ± 0.15	90.3 ± 0.12	84.9 ± 0.17	75.5 ± 0.20	60.6 ± 0.17	80.9 ± 0.14	80.3 ± 0.15
SimCLR	81.4 ± 0.11	56.8 ± 0.21	73.6 ± 0.21	67.8 ± 0.21	58.4 ± 0.25	53.1 ± 0.10	66.8 ± 0.18	65.0 ± 0.17
DINO	93.6 ± 0.06	81.4 ± 0.18	82.2 ± 0.17	75.7 ± 0.21	72.7 ± 0.22	59.5 ± 0.16	78.7 ± 0.14	78.2 ± 0.15
VQ-VAE	95.8 ± 0.05	86.7 ± 0.16	86.2 ± 0.15	80.3 ± 0.20	75.1 ± 0.21	60.9 ± 0.17	79.9 ± 0.14	79.5 ± 0.15
MAE	96.5 ± 0.04	88.7 ± 0.14	89.7 ± 0.12	84.3 ± 0.17	77.6 ± 0.19	61.9 ± 0.17	81.2 ± 0.13	80.8 ± 0.14
Autoregression	96.0 ± 0.04	87.6 ± 0.15	87.8 ± 0.13	81.9 ± 0.18	72.1 ± 0.22	58.9 ± 0.15	79.0 ± 0.14	78.5 ± 0.15
OSF	**97.3 ± 0.04**	**90.4 ± 0.14**	**92.8 ± 0.10**	**88.3 ± 0.15**	**77.7 ± 0.19**	**62.2 ± 0.18**	**81.5 ± 0.13**	**81.0 ± 0.15**

Table 30 | **Robustness analysis on demographic subgroups.** OSF achieves better performance and deliver fairer outcome.

Subgroup of CFS

Subgroup	Setting	Sleep Staging	Hypopnea
Young	SleepFM	97.9	**92.2**
	OSF	**98.1**	91.0
Old	SleepFM	96.8	83.6
	OSF	**97.6**	**85.3**
Disparity	SleepFM	0.78	6.08
	OSF	**0.35**	**4.03**

Subgroup of MESA

Subgroup	Setting	Sleep Staging	Hypopnea
Male	SleepFM	94.6	74.5
	OSF	**95.5**	**77.4**
Female	SleepFM	94.2	76.8
	OSF	**96.1**	**77.9**
Disparity	SleepFM	**0.28**	1.63
	OSF	0.42	**0.35**

with more data consistently outperform those pre-trained with less data.

Additional results on the effect of sample size scaling. We pre-train OSF and SleepFM with a ViT-85M backbone using different fractions of the pre-training corpus. As shown in Tables 36 and 35, we conduct more comprehensive evaluations on sleep staging task and three sleep event detection tasks. The downstream performance of OSF improves consistently as we increase the percentage of pre-training data. We observe a similar scaling trend for SleepFM when using the same 85M backbone, as shown in Tables 38 and 37. The models pre-trained with larger parameter size demonstrate good scaling behaviors, which motivates us to jointly scale both model size and sample size.

Single-source pre-training also exhibits scaling with sample size. To further verify the generalization of scalability on sample size, we pre-train models on with different percentage of the SHHS cohort. As shown in Tables 36, Tables 35, Tables 38, and Tables 37, models pre-trained on the single-source SHHS cohort also show clear scaling behavior. When we increase the fraction of SHHS used for pre-training, downstream performance improves consistently.

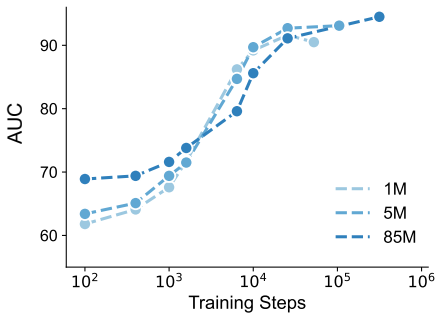


Figure 13 | **Scaling on pretraining steps.** OSF scales on training computation.

Comprehensive results on model-size scaling. We pre-train OSF and SleepFM on the full pre-training corpus, with different parameters of the backbone: 1M, 5M, 85M. We then evaluate these checkpoints on both the in-domain SHHS cohort and the out-of-domain MROS cohort, on three sleep event detection tasks and the sleep staging task. As shown in Table 40, the scaling trend holds across all tasks and both cohorts, therefore, increasing model size leads to improved downstream performance.

Table 31 | **Controlled study of masking strategies for self-supervised pretraining on sleep data.** We report linear-probing results on sleep analysis tasks (**AUC**) on SHHS and MROS datasets. Channel masking as an augmentation consistently improves representation quality.

Method	Mask Strategy		SHHS				MROS			
	Time	Channel	Sleep Staging	Arousal	Hypopnea	Ox. Desat.	Sleep Staging	Arousal	Hypopnea	Ox. Desat.
SimCLR	✓	✗	86.1	71.9	64.9	62.9	81.4	73.6	58.4	66.8
	✗	✓	96.5	90.1	79.0	75.5	94.8	86.5	73.3	79.1
	✓	✓	**97.3**	**92.9**	**82.2**	**80.4**	**96.7**	**90.8**	**75.6**	**81.1**
DINO	✓	✗	95.1	87.0	78.3	76.5	93.6	82.2	72.7	78.6
	✗	✓	97.4	93.9	82.8	80.1	96.4	90.5	77.1	80.3
	✓	✓	**97.5**	**94.0**	**82.8**	**81.3**	**97.3**	**92.8**	**77.7**	**81.5**

Table 32 | **Ablation studies between augmentation from the vision domain (crop) and our masking-based strategies.** We show linear probing results (**AUC**$^{\uparrow}$) here. Vision domain's augmentation works for sleep data, but not necessary to be added.

Method	Crop	Masking		SHHS				MROS			
		Time	Channel	Sleep Staging	Arousal	Hypopnea	Ox. Desat.	Sleep Staging	Arousal	Hypopnea	Ox. Desat.
SimCLR	✓	✗	✗	95.2	86.7	79.2	77.1	94.7	83.8	75.0	80.2
	✗	✗	✓	96.5	90.1	79.0	75.5	94.8	86.5	73.3	79.1
	✓	✓	✓	97.1	91.9	**83.4**	80.4	96.6	89.7	**79.3**	**81.5**
	✗	✓	✓	**97.3**	**92.9**	82.2	**80.4**	**96.7**	**90.8**	75.6	81.1
DINO	✓	✗	✗	95.6	88.1	81.0	78.5	95.4	85.7	77.2	80.6
	✗	✗	✓	97.4	93.9	82.8	80.1	96.4	90.5	77.1	80.3
	✓	✓	✓	97.0	91.4	82.1	79.8	96.6	89.3	77.3	81.0
	✗	✓	✓	**97.5**	**94.0**	**82.8**	**81.3**	**97.3**	**92.8**	**77.7**	**81.5**

D.2. Additional Results of Multi-Source Data Mixture

Here we report detailed results comparing multi-source data mixture pre-training with single-source pre-training. We pre-train two groups of models on SHHS only, as well as on data sampled from the full pre-training corpus. For each group, we consider using 1%, 10%, 100% of the given pre-training cohorts. As shown in Tables 36, Tables 35, Tables 38, and Tables 37, across four sleep analysis tasks and three data scales, multi-source pre-training consistently outperforms single-source pre-training, over all data percentages.

D.3. Jointly Scale Model and Sample Size is Needed

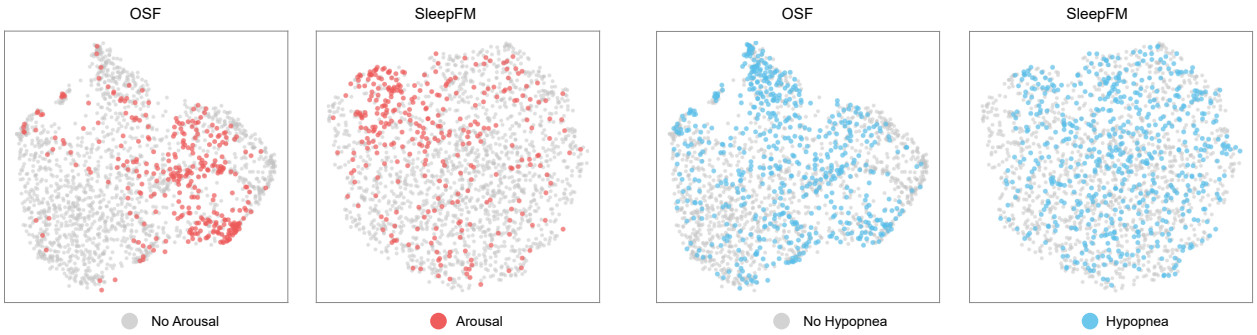
We further examine how model capacity and pre-training sample size jointly influence downstream performance. We pre-train both OSF and the baseline SleepFM [36]with different model sizes, on different sample size, and evaluate them on four sleep analysis tasks. As shown in Fig. 15, we find that increasing model capacity can improve the utilization of additional pre-training data. In particular, while SleepFM with its original model capacity shows only marginal gains as the pre-training sample size grows, scaling its each backbone encoder to 85M parameters yields substantially larger improvements. Motivated by this observation, we adopt a ViT-85M backbone when benchmarking different pre-training strategies and pre-train on the full multi-source data.

D.4. Scaling with Pre-training Computation

We further study whether OSF benefits from increased pre-training computation. Specifically, we evaluate few-shot sleep staging performance across different numbers of pre-training steps using

Table 34 | **Ablation on normalization strategies.** Per-night z-score normalization leads to better performance.

Normalization	SHHS				MROS			
	Sleep Staging		Hypopnea		Sleep Staging		Hypopnea	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
Min-Max	96.3	90.4	80.6	76.0	95.8	86.4	77.1	62.1
Per-Night Z-Score	**97.5**	**93.2**	**82.8**	**78.1**	**97.3**	**90.4**	**77.7**	**62.2**



(a) Arousal **(b)** Hypopnea

Figure 11 | **UMAP visualization of learned representations.** OSF produces more separable embeddings, indicating stronger alignment between the embedding geometry and sleep event labels, including arousal and hypopnea events.

models with 1M, 5M, and 85M parameters. As shown in Fig. 13, all models improve rapidly during the first 10^4 steps. After that, smaller models tend to saturate, while the 85M model continues to improve with additional pre-training. These results suggest that OSF can benefit from increased pre-training computation when paired with sufficient model capacity.

E. Additional Results on Handling Missing Channels

OSF better handles missing channel inference problems on in-domain data corpus as well. To validate the robustness of the conclusion in Sec. 6.1, We report additional results for inference on the SHHS cohort on the four realistic missing channel settings. As shown in Table 41, across four sleep analysis tasks, OSF consistently outperforms SleepFM [36], indicating that OSF learns more robust representations under inference-time channel missingness.

Additional results on pre-training with fewer channels. Here we report detailed results for pre-training only with breathing channels and ECG channel. We compare OSF against SleepFM under the same channel subset. As shown in Table 42, across both the in-domain SHHS cohort and the out-of-domain MROS cohort, on most sleep analysis tasks, OSF makes better use of the available channels and achieves higher downstream performance.

F. Additional Ablation Results

Additional results of ablation studies on channel masking ratio. We further ablate the channel masking ratio by pre-training two additional variants with masking ratios of 0.1 and 0.9, while

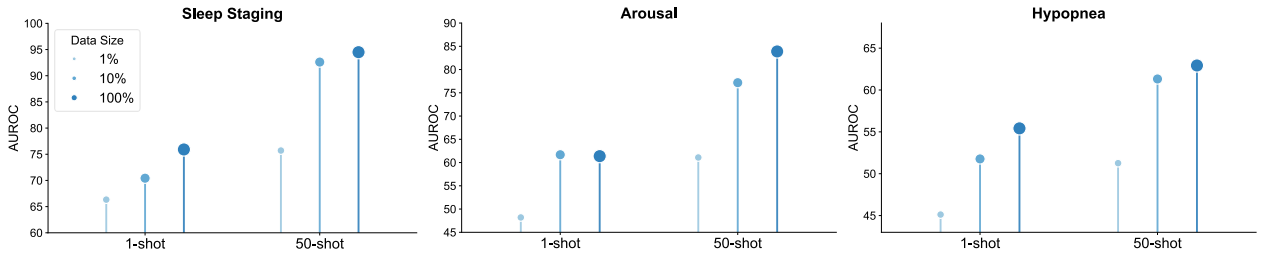
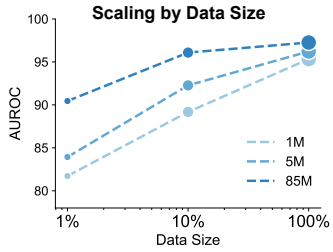


Figure 12 | **Scaling performance on few-shot adaptation.** OSF scales on the pre-training sample size.

Table 35 | **Detailed data scaling results (linear probing) for OSF evaluated on SHHS.** All methods are trained with ViT-85M. Large model capacities or better pretraining strategies enable scalability.

Pretrain Set	Data Pct	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SINGLE-SRC	1%	91.8	79.8	74.9	67.2	70.0	65.2	68.5	62.7
SINGLE-SRC	10%	96.3	90.4	90.2	84.1	79.4	74.8	77.6	71.6
SINGLE-SRC	100%	97.5	93.3	94.1	89.5	83.1	78.4	81.4	75.9
MULTI-SRC	1%	92.8	81.9	78.0	69.9	72.0	67.1	70.3	64.5
MULTI-SRC	10%	96.8	91.6	91.9	86.3	80.2	75.4	78.8	73.1
MULTI-SRC	100%	97.5	93.2	94.0	89.3	82.8	78.1	81.3	75.6



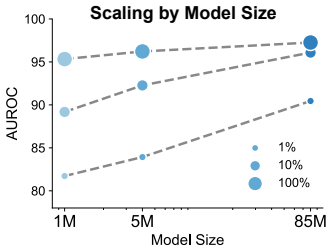


Figure 14 | **Scaling performance on linear probing sleep staging.** OSF scales on both the model capacity and the pre-training sample size.

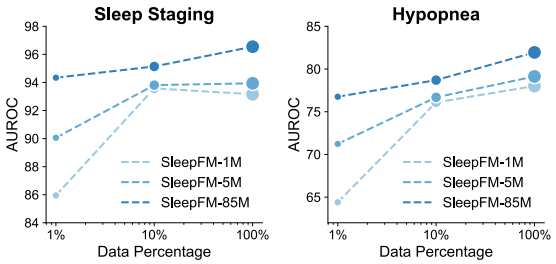


Figure 15 | **Improved scaling performance of SleepFM.** Increasing model capacity helps weaker pre-training methods utilize more training samples effectively.

keeping all other settings unchanged. As shown in Table 39, the default ratio of 0.5 achieves the best overall linear probing performance across SHHS and MROS. This suggests that moderate channel masking provides a better balance between task difficulty and preserved input information: weak masking may make the pretext task less informative, whereas overly aggressive masking can remove too much training signal and hurt downstream transfer.

Table 36 | **Detailed data scaling results (linear probing) for OSF evaluated on MROS.** All methods are trained with ViT-85M. Large model capacities or better pretraining strategies enable scalability.

Pretrain Set	Data Pct	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SINGLE-SRC	1%	88.9	71.3	75.4	68.7	63.5	55.3	74.5	73.5
SINGLE-SRC	10%	94.8	84.5	84.7	78.3	71.3	58.7	78.1	77.6
SINGLE-SRC	100%	97.0	89.5	92.1	87.4	75.9	60.9	80.8	80.4
MULTI-SRC	1%	90.5	74.8	77.6	70.8	66.8	56.6	76.4	75.7
MULTI-SRC	10%	96.1	87.8	88.4	82.8	73.9	60.1	79.8	79.5
MULTI-SRC	100%	97.3	90.4	92.8	88.3	77.7	62.2	81.5	81.0

Table 37 | **Detailed data scaling results (linear probing) for SleepFM [36] evaluated on SHHS.** All methods are trained with ViT-85M. Large model capacities or better pretraining strategies enable scalability.

Pretrain Set	Data Pct	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SINGLE-SRC	1%	93.7	84.0	84.2	76.0	76.3	71.2	74.4	68.7
SINGLE-SRC	10%	95.1	87.5	88.1	81.1	79.1	74.2	77.6	71.7
SINGLE-SRC	100%	96.2	90.1	91.8	85.6	81.5	76.3	80.0	74.3
MULTI-SRC	1%	94.3	85.7	84.0	75.5	76.8	72.0	75.3	69.8
MULTI-SRC	10%	95.1	87.7	88.7	81.9	78.7	73.8	77.3	71.3
MULTI-SRC	100%	96.7	91.1	92.2	86.3	82.0	77.0	80.6	75.0

Table 38 | **Detailed data scaling results (linear probing) for SleepFM evaluated on MROS.** All methods are trained with ViT-85M. Large model capacities or better pretraining strategies enable scalability.

Pretrain Set	Data Pct	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SINGLE-SRC	1%	91.2	77.2	80.7	73.9	69.1	57.9	78.3	77.9
SINGLE-SRC	10%	92.9	80.5	82.6	76.5	70.4	58.3	78.4	77.9
SINGLE-SRC	100%	96.0	86.8	89.2	83.4	74.2	60.0	80.0	79.6
MULTI-SRC	1%	92.3	79.1	81.2	74.6	68.3	57.4	77.8	77.3
MULTI-SRC	10%	94.4	82.9	85.0	78.6	70.9	58.5	78.3	78.0
MULTI-SRC	100%	96.4	87.7	90.3	84.9	75.5	60.6	80.9	80.3

Table 39 | **Ablation on channel masking ratios.** A moderate masking ratio works the best.

Masking Ratio	SHHS				MROS			
	Sleep Staging		Hypopnea		Sleep Staging		Hypopnea	
	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
0.1	96.9	91.9	81.7	77.0	95.9	87.2	77.4	61.5
0.5 (main setting)	**97.5**	**93.2**	**82.8**	**78.1**	**97.3**	**90.4**	**77.7**	**62.2**
0.9	95.4	88.7	79.5	74.5	95.3	85.9	75.0	60.3

Table 40 | **Model size scaling results (linear probing) on SHHS and MROS.** Large models show better performances.

Dataset	Model	Enc Size	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
			AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SHHS	OSF	ViT-1M	96.2	90.2	89.2	82.7	78.1	72.9	76.7	70.5
		ViT-5M	96.9	91.8	91.8	86.2	80.2	75.1	78.5	72.5
		ViT-85M	97.5	93.2	94.0	89.3	82.8	78.1	81.3	75.6
	SleepFM	ViT-1M	93.2	82.6	85.9	78.0	78.0	72.7	76.1	69.7
		ViT-5M	93.9	84.4	89.4	82.6	79.1	73.1	77.9	71.8
		ViT-85M	96.6	91.1	92.2	86.3	82.0	77.0	80.6	75.0
MROS	OSF	ViT-1M	95.3	86.1	86.0	80.1	70.7	58.4	78.7	78.3
		ViT-5M	96.2	88.1	88.9	83.6	72.5	59.3	79.5	79.1
		ViT-85M	97.3	90.4	92.8	88.3	77.7	62.2	81.5	81.0
	SleepFM	ViT-1M	92.0	75.8	81.6	75.2	68.3	57.2	76.1	69.7
		ViT-5M	93.4	78.8	85.0	78.7	69.1	57.4	78.3	77.5
		ViT-85M	96.4	87.7	90.3	84.9	75.5	60.6	80.9	80.3

Table 41 | **Linear probing results on SHHS under realistic missing-channel settings.** OSF (DINO) is more robust to missing channels compared to existing sleep FM.

Brain	Resp.	ECG	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
				AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
✓	✗	✗	SleepFM	97.0	92.0	92.2	86.2	75.0	69.7	73.8	67.5
			OSF	97.5	93.4	93.7	88.9	75.9	70.5	75.3	69.0
✗	✓	✓	SleepFM	86.7	69.1	87.3	80.3	82.3	77.5	80.1	74.5
			OSF	85.2	66.6	87.4	80.7	82.8	78.4	80.1	74.4
✓	✗	✓	SleepFM	96.7	91.3	92.3	86.3	76.1	70.6	76.4	70.6
			OSF	97.5	93.3	94.0	89.2	76.4	70.9	76.7	70.8
✗	✓	✗	SleepFM	86.6	68.7	87.0	79.8	82.8	78.2	80.1	74.3
			OSF	85.5	67.0	86.9	80.2	84.1	80.3	80.3	74.7

Table 42 | **Robustness of pre-training with fewer channels.** We pre-train the two models and evaluate them using only ECG and respiratory signals, OSF better utilizes these channels.

Dataset	Method	Sleep Staging		Arousal		Hypopnea		Ox. Desat.	
		AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑	AUC↑	AUPRC↑
SHHS	SleepFM	82.1	61.1	84.4	77.1	79.5	74.5	78.1	72.6
	OSF	81.5	60.4	84.8	77.8	81.5	77.0	78.6	73.3
MROS	SleepFM	83.2	59.0	82.7	76.8	71.3	58.6	79.7	78.9
	OSF	83.2	59.4	83.3	77.6	77.6	62.4	80.8	80.1