A Large-Scale EEG Foundation Model Enables Label-Efficient, Multi-Device Sleep Analytics

Jason Manley, Pierre Bartet, Franz Fürbass, Antoine Guillot, Alexander M. Chan
Beacon Biosignals, Boston, MA

BEACONBIOSIGNALS

P-01 Board 2, #341

We developed a sleep-focused EEG foundation model, pretrained on >128,000 overnight recordings spanning PSG, 10-20 EEG, and dry-electrode (Waveband) data, that matches or outperforms task-specific architectures with 5-10× fewer labels across four sleep analytics tasks.

Introduction

Self-supervised learning has advanced pretraining in natural language processing, computer vision and, more recently, EEG. However, most prior EEG work targets BCI or high-density research recordings, leaving sleep analytics, which spans diverse devices and is chronically label-limited, underserved. We pretrain across clinical and at-home recordings to build a generic encoder that transfers across devices and tasks.

Methods

We pretrained a 4.7M-parameter hybrid RNN-Transformer encoder on >128,000 overnight sleep studies (~1.8M hours, 14 datasets) using a contrastive objective (Figure 1). The same encoder was then fine-tuned on four downstream tasks — sleep staging on PSG, sleep staging on dry-EEG headband data (Waveband), artifact detection on Waveband, and arousal detection on PSG — using between 10 and 300 recordings, or the full labelled set for tasks with fewer available recordings. We compared against (i) task-specific architectures and (ii) the same generic encoder trained from scratch, holding the number of fine-tuning recordings fixed.

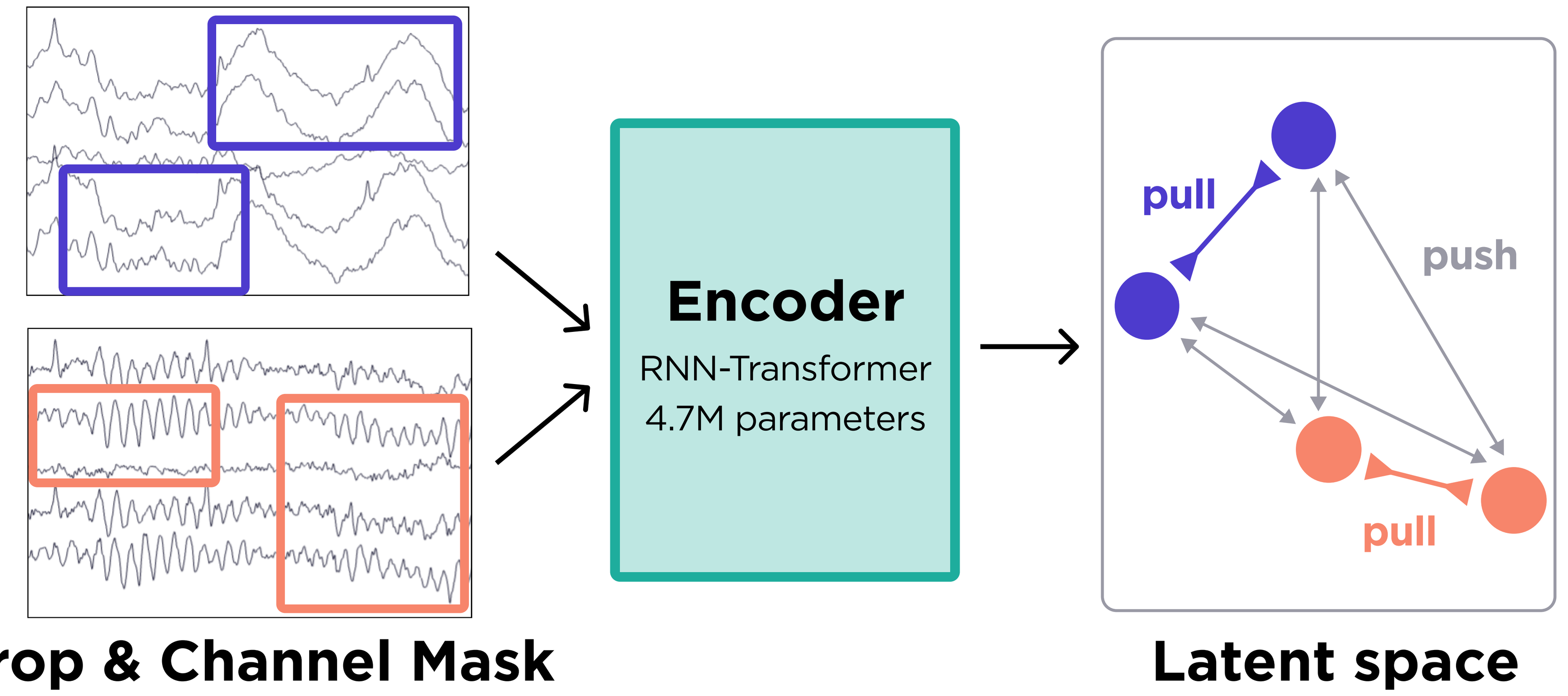

Encoder
RNN-Transformer
4.7M parameters
pull
push
pull
Crop & Channel Mask
Latent space

Figure 1: Schematic of self-supervised contrastive learning approach. Two augmented views of each EEG window (via cropping and channel masking) are encoded and pulled together in latent space, while views from different windows are pushed apart.

Table 1: Downstream task performance scales with model size and pretraining dataset size. SSL-pretrained encoders outperform task-specific baselines across tasks, with gains growing as model capacity and pretraining hours increase. Bold: configuration used in Fig. 2.

Architecture	Model capacity (parameters)	Pretraining dataset size (hours)	Waveband Sleep Staging	PSG Sleep Staging	Waveband Signal Quality	PSG Arousal Detection
Task-Specific	0.04-0.2 million	—	0.571	0.542	0.948	0.653
Generic	2.2 million	0.4 million	0.678	0.673	0.964	0.686
Generic	4.7 million	0.4 million	0.695	0.688	0.966	0.689
Generic	**4.7 million**	**1.8 million**	**0.731**	**0.716**	**0.968**	**0.711**

Results: Label Efficiency

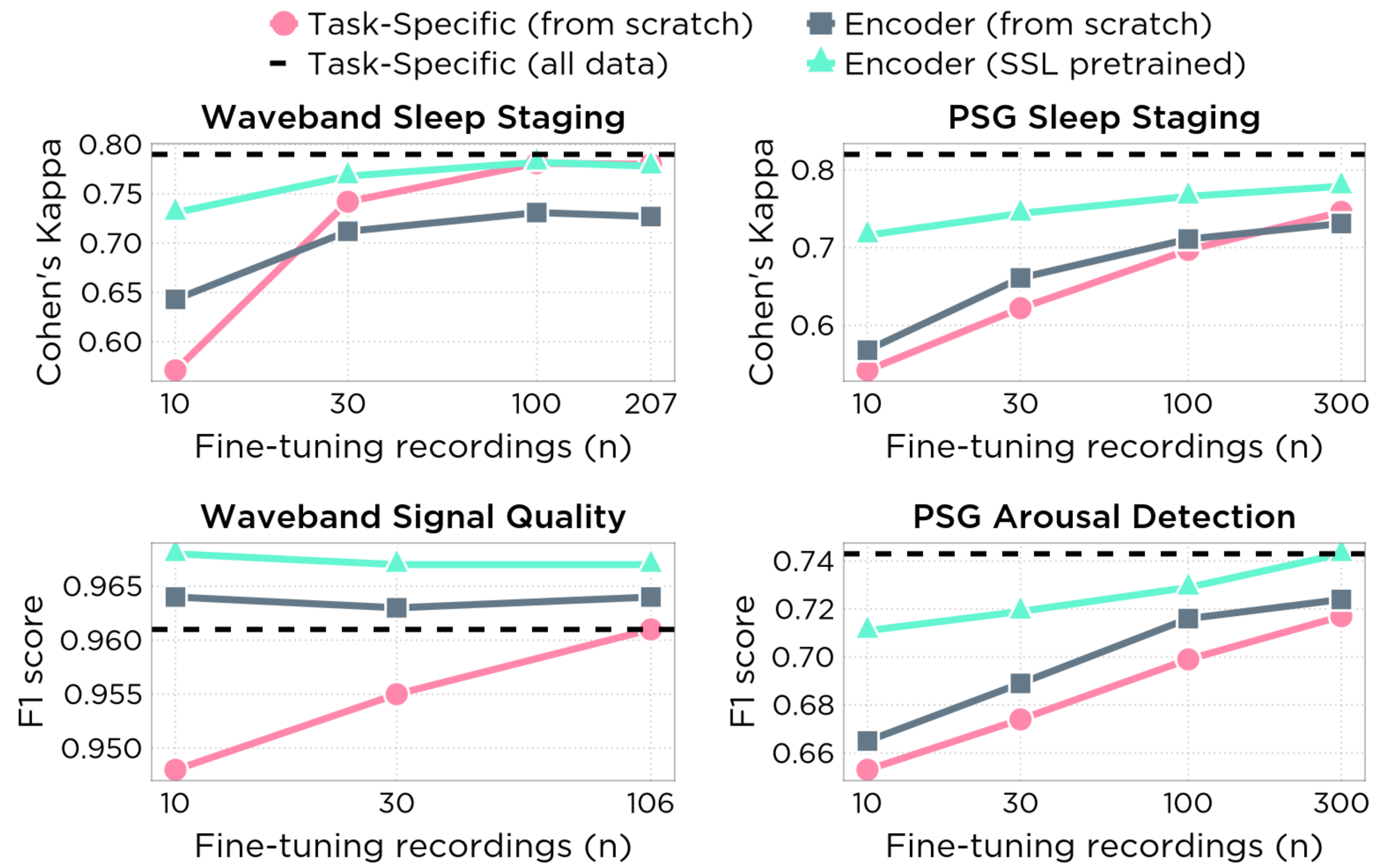

Task-Specific (from scratch)
Task-Specific (all data)
Encoder (from scratch)
Encoder (SSL pretrained)
Waveband Sleep Staging
Cohen's Kappa
0.80
0.75
0.70
0.65
0.60
10
30
100
207
Fine-tuning recordings (n)
PSG Sleep Staging
Cohen's Kappa
0.8
0.7
0.6
10
30
100
300
Fine-tuning recordings (n)
Waveband Signal Quality
F1 score
0.965
0.960
0.955
0.950
10
30
106
Fine-tuning recordings (n)
PSG Arousal Detection
F1 score
0.74
0.72
0.70
0.68
0.66
10
30
100
300
Fine-tuning recordings (n)

Figure 2: SSL pretraining substantially improves label efficiency across tasks and devices. The pretrained encoder (teal) matches or approaches task-specific supervised baselines (pink) with 5–10× fewer labelled recordings, and consistently outperforms an identical encoder trained from scratch (gray). Shown for the configuration from Table 1 and 120 s crops from Figure 3. Black dashed lines inidcate the task-specific backbone when trained on all available data.

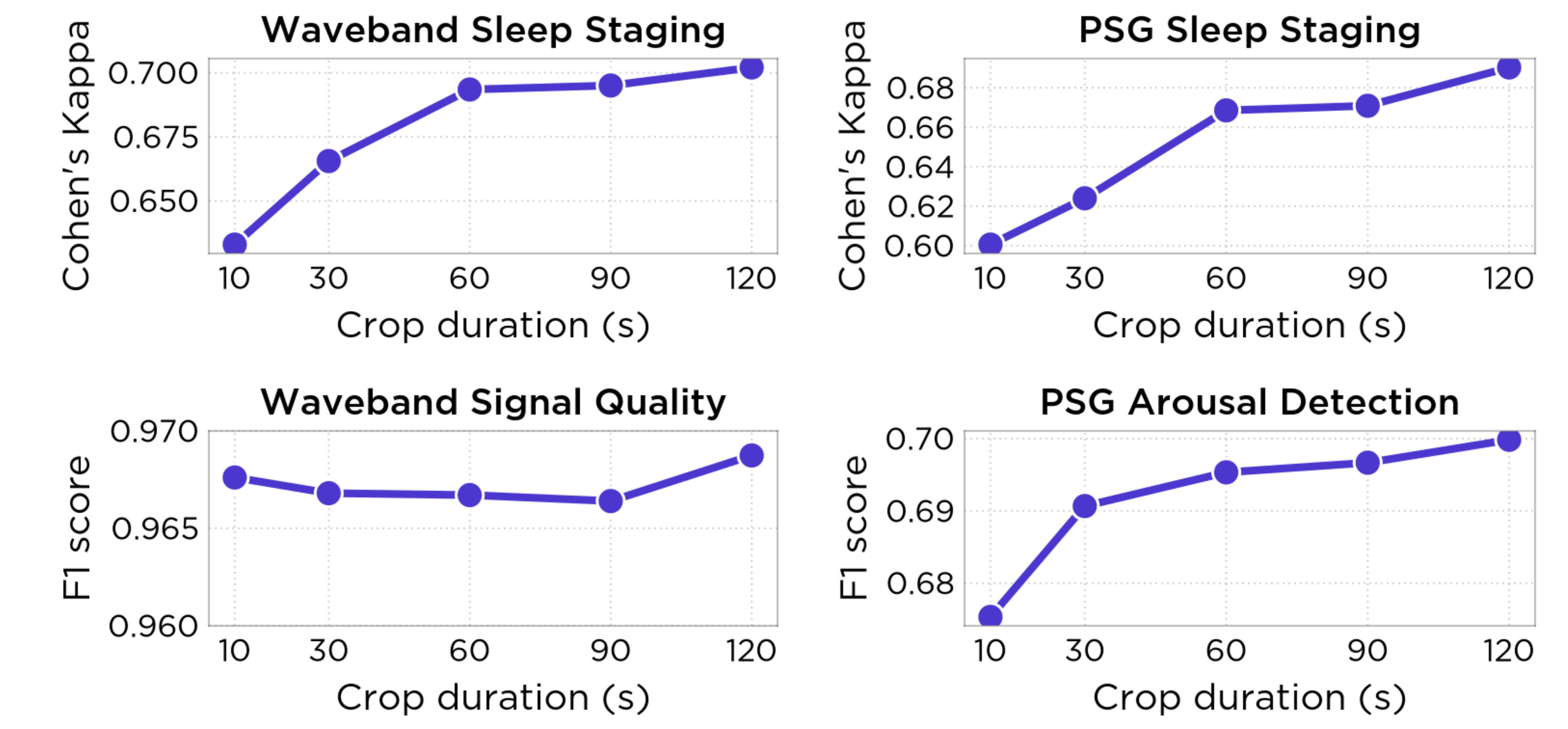

Waveband Sleep Staging
Cohen's Kappa
0.700
0.675
0.650
10
30
60
90
120
Crop duration (s)
PSG Sleep Staging
Cohen's Kappa
0.68
0.66
0.64
0.62
0.60
10
30
60
90
120
Crop duration (s)
Waveband Signal Quality
F1 score
0.970
0.965
0.960
10
30
60
90
120
Crop duration (s)
PSG Arousal Detection
F1 score
0.70
0.69
0.68
10
30
60
90
120
Crop duration (s)

Figure 3: Downstream task performance scales with context length. Increasing the crop duration from 10 s to 120 s improves performance across all four tasks (Waveband sleep staging, PSG sleep staging, Waveband signal quality, PSG arousal detection), with the largest gains on sleep staging. Fine-tuning performed on n=10 labelled recordings.

Results: Learned Representations

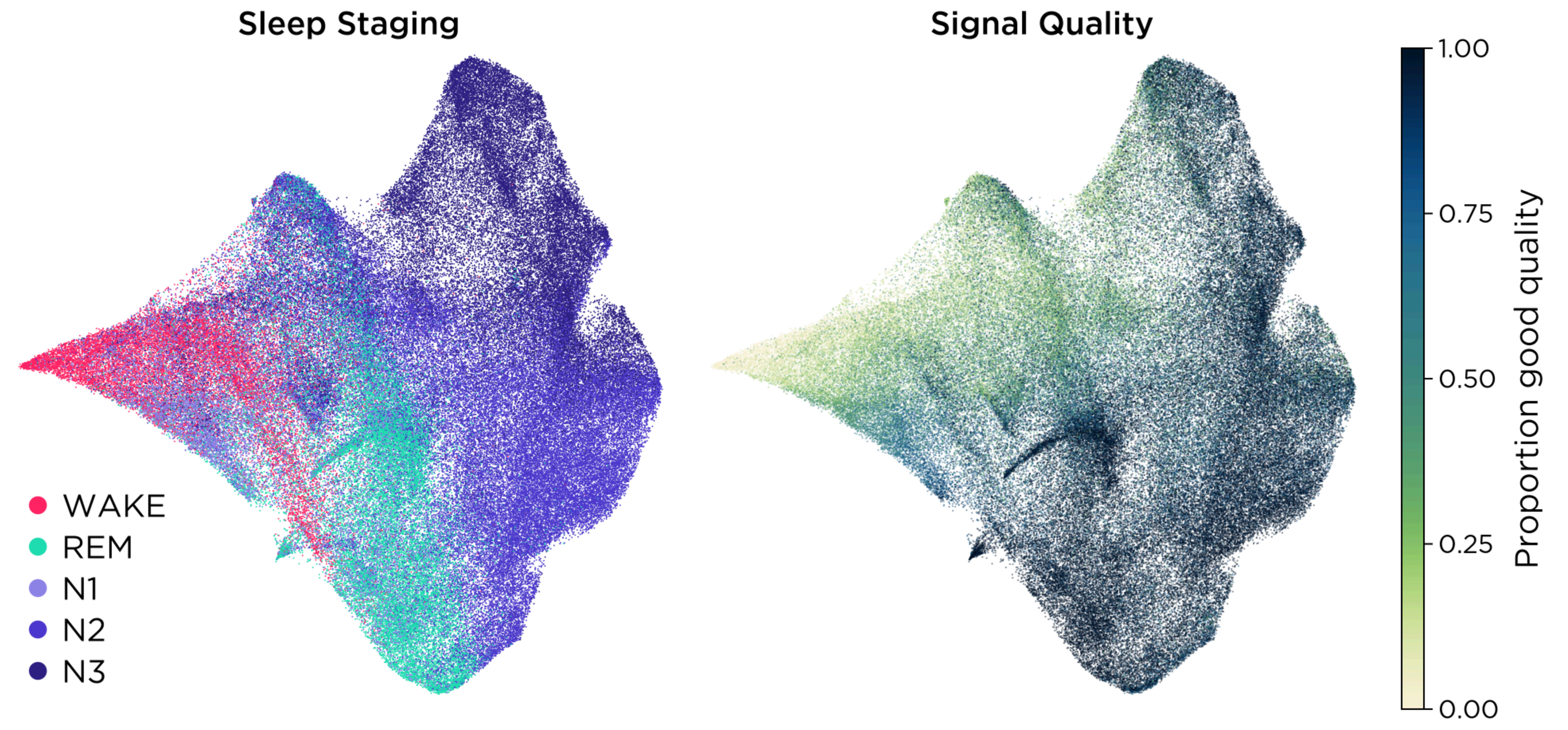

Sleep Staging
Signal Quality
WAKE
REM
N1
N2
N3
1.00
0.75
0.50
0.25
0.00
Proportion good quality

Figure 4: The pretrained encoder learns sleep- and signal-relevant structure without labels. UMAP projection of embeddings from held-out recordings reveals clear separation by sleep stage (left; WAKE, REM, N1, N2, N3) and by signal quality (right), despite neither label being used during pretraining. This suggests that the encoder captures physiologically meaningful representations transferable across downstream tasks.

Conclusions

- → **Label efficiency**. SSL pretraining matches or approaches state-of-the-art performance with 5–10× fewer labelled recordings across four sleep analytics tasks.
- → **Cross-device transfer**. A single encoder, pretrained jointly on PSG, 10-20 EEG, and dry-electrode recordings, transfers across all three device classes — supporting deployment from clinical to consumer-grade hardware.
- → **Predictable scaling**. Downstream performance scales with model size, pretraining corpus size, and input context length, suggesting further gains are achievable with additional scale.
- → **Meaningful representations emerge without labels**. Pretrained embeddings separate by sleep stage and signal quality despite neither being used during pretraining, indicating physiologically relevant structure.

Contact
Jason Manley, PhD
Algorithm Engineer, Beacon Biosignals
jason.manley@beacon.bio