Learning Structured Sleep Transitions with Sequential EEG Foundation Models

Koki Yoshida, Zheng Chen, Rikuto Kotoge, Yasuko Matsubara, Yasushi Sakurai,
SANKEN, The University of Osaka, Japan
{yoshida88, chenz, rikuto88, yasuko,yasushi}@sanken.osaka-u.ac.jp

Abstract

In this paper, we investigate a novel sleep EEG foundation model designed to capture long-term temporal dependencies in EEG sequences and reduce implausible predictions in sleep stages. Unlike existing models that treat short EEG epochs as independent samples, our method aggregates a sequence of pre-tokenized EEG epochs and learns structured dynamics spanning multiple stages. We adopt a masked language modeling framework, leveraging masked token prediction to enable robust temporal representation learning. Empirical results on the SHHS dataset show that our model outperforms four state-of-the-art EEG foundation models across standard classification metrics. Moreover, we introduce a novel metricIrregular Transition Rate to assess the biological plausibility of stage transitions. Our method significantly reduces ITR to 15.2%, compared to 29.6% (BIOT) and 33.7% (EEGPT), confirming its superior ability to model coherent sleep dynamics.

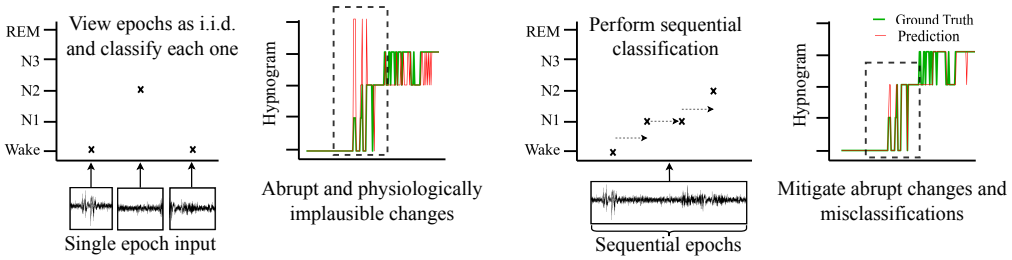
1 Introduction

Sleep is a vital biological process marked by a series of physiological *transitions*. Effective sleep requires understanding of the macrostructure sleep [Phan et al., 2022, Chen et al., 2023], such as assessing the proportion of time spent in deep sleep. Achieving such insights relies on accurate annotation of overnight patient recordings, typically electroencephalograms (EEG). The American Academy of Sleep Medicine (AASM) categorizes sleep into five stages: Wake, N1, N2, N3, and REM [Berry et al., 2013]. Clinically, experts must inspect patient recordings and assign a sleep stage to each 30-second EEG epoch based on distinct features. This manual process is labor-intensive and time-consuming, often requiring collaboration among multiple clinicians.

Deep learning models have demonstrated impressive success in automating sleep stage annotation [Phan et al., 2019, Jia et al., 2020, Chen et al., 2022, Pradeepkumar et al., 2024]. Such achievements stem from learning stage-specific representations for noisy EEGs. However, most existing methods rely on supervised learning with large amounts of costly, labeled data. Several self-supervised learning (SSL) methods have been proposed, aiming to learn meaningful representations from unlabeled EEGs [Eldele et al., 2021b, Kotoge et al., 2024]. However, real-world EEG signals are acquired under diverse scenarios and varying settings. These SSL methods are case-specific with carefully designed architectures and fixed channel settings, limiting their scalability across data formats or settings.

Related Works in EEG Foundation Models (FMs). Recently, the transformative success of LLMs has sparked growing interest in building EEG FMs. Namely, BIOT [Yang et al., 2024] is the first benchmark for EEG FMs, establishing a SSL pretraining and evaluation protocol across diverse data formats, including varying channel counts and sequence lengths. EEG2Rep [Mohammadi Foumani et al., 2024] enhances EEG representation learning by leveraging latent-space reconstruction and semantic-aware masking to robustly capture informative patterns in noisy, continuous EEG signals. LaBraM [Jiang et al., 2024] introduces a vector-quantisation-based tokenizer that discretizes EEG signals into neural code tokens, enabling masked token prediction in a vision-inspired pretraining

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.



(a) Epoch-level EEG FM (b) Epoch-sequence EEG FM

Figure 1: Compared with conventional epoch-level models, our epoch-sequence model leverages cross-epoch context. Our model generates time-dependent outputs from sequential epochs. In sleep staging, it reduces predictions of irregular, abrupt stage transitions.

framework. EEG-GPT [Wang et al., 2024] further leverages both contrastive learning and make prediction to build an EEG representation space capable of generalizing across datasets.

- Limitations. Existing FMs typically model EEG data at the single-epoch level (such as 5-second or 30-second segments), which is problematic for the sleep stage annotation task. This modeling treats each EEG epoch as an independent and identically distributed (i.i.d.) sample, overlooking the strong temporal dependencies inherent in sleep architecture. In clinical practice, sleep experts often manually account for temporal dependencies across consecutive epochs to determine ambiguous or transitional stages [Chen et al., 2023]. However, the i.i.d. assumption in single-epoch FMs often lead to biologically implausible stage transitions (e.g., abrupt jumps from Wake to REM), which reduces both model interpretability and clinical trustworthiness.

In this work, we investigate a new EEG foundation model that operates beyond the single-epoch modeling and instead learns from EEG sequences. **Empirically,** our central hypothesis is that capturing transitions between sleep stages, rather than treating each short EEG segment as independent, enables the model to better characterize physiological dynamics. We assume that an individuals brain states exhibit temporal continuity and typically do not change rapidly. **Methodologically**, our model aggregates multiple pre-tokenized EEG epochs into continuous representations, enabling the learning of inter-epoch dependencies and global sleep architecture. We pretrain the model using a combination of sequentially masked token prediction and cross-entropy objectives, encouraging it to learn both contextual and discriminative patterns across time. **Experimentally**, we evaluate our model on the SHHS dataset and we outperform recent four EEG FMs across classification metrics. Moreover, we introduce a novel evaluation metric of Irregular Transition Rate (ITR), which quantifies biologically implausible stage changes within predicted sleep sequences. Our model consistently yields lower ITR scores, demonstrating it reduce unrealistic transitions in sleep stage predictions.

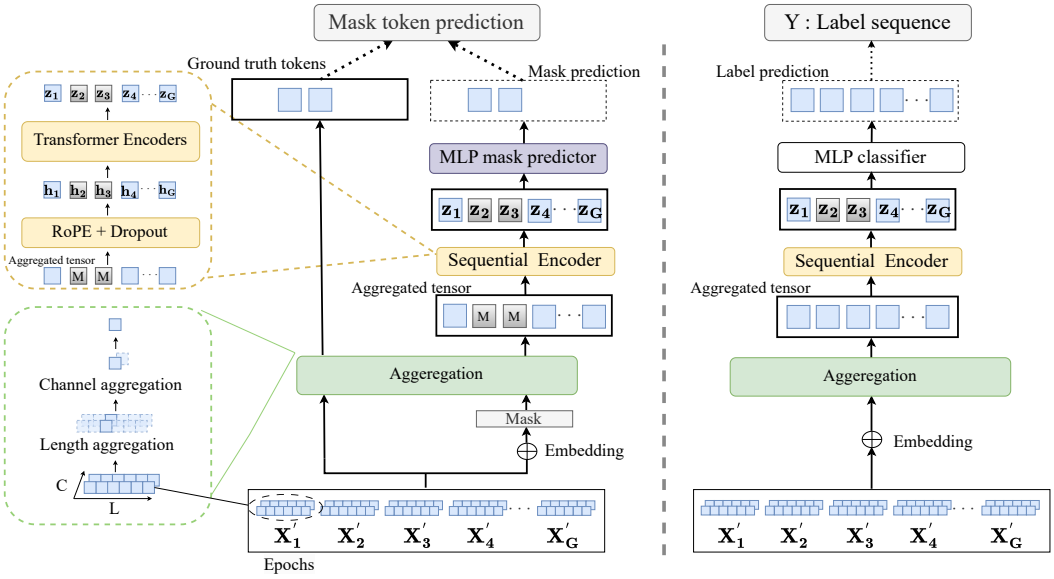
Remark (Potential for Broader Applicability). Our model holds promise for broader EEG-based applications, such as seizure detection or braincomputer interface tasks. We assume that individuals tend to remain within a particular brain or bodily state for extended durations, exhibiting stable and structured temporal patterns. Modeling these long-term dependencies is crucial for capturing gradual transitions, persistent abnormalities, or recurring neural signatures across time. Extending the proposed method to other domains is a key future work.

2 Method

In this study, we propose a framework for modeling long-term EEG sequences (Figure 2). Our framework consists of two modules: (i) *Epoch-level EEG encoder* learns diverse EEG waveform patterns within each 30-second EEG epoch and tokenizes them. (ii) *Sequence-level EEG encoder* captures long-term dependencies by modeling the sequence of tokenized epochs, enabling the representation of smooth transitions across sleep stages.

2.1 Epoch-level EEG Tokenizer

Given the raw EEG signal $\mathbf{X} \in \mathbb{R}^{C \times T}$, where C denotes the number of EEG channels (electrodes) and T is the total number of time steps, the epoch-level EEG encoder transforms the input $\mathbf{X}$ into



(a) Pretraining (b) Fine-tuning

Figure 2: System overview: (a) Given an epoch sequence, we mask a subset, embed and aggregate each epoch to form a compact EEG sequence. We pretrain the encoder by predicting the original tokens at masked positions after passing the sequence through it, which encourages learning of temporal dependencies. (b) Initialize the encoder with pre-trained weights and attach an MLP classifier. We fine-tune the model to predict per-epoch labels from a sequence of tokenized epochs.

a discrete token representation by learning salient waveform patterns across channels and time. Formally, the encoder outputs a token $x' \in \mathcal{V}$, where $\mathcal{V}$ denotes the predefined token vocabulary.

Our model can be combined with any epoch-level EEG foundation model for tokenizing EEG signals. In this paper, we use the recently proposed TFM-Tokenizer [Pradeepkumar et al., 2025] as the example epoch-level encoder. For each EEG epoch, both the raw signal $\mathbf{X}$ and its short-time Fourier transform (STFT)-based time-frequency representation are provided as inputs to the TFM-tokenizer. From these two representations, the tokenizer extracts local time-frequency motifs and quantizes each motif to the discrete tokens $\mathbf{X}' \in \mathbb{R}^{C \times L}$, where L denotes the token sequence length.

2.2 Sequence-level EEG Encoder

The Sequence-level EEG Encoder consists of: (i) *Pretraining* with self-supervised learning to capture sequentially temporal dependencies; (ii) *Finetuning* on labeled data for downstream sleep stage tasks.

Pretraining for Capturing Temporal Dependencies. We train a Transformer-based Sequential model with masked token prediction, following the Masked Language Modeling (MLM) objective [Warner et al., 2024] to capture the temporal transitions across EEG epochs. Given a sequence of epoch-level EEG tokens $\mathbf{S} = (\mathbf{X'}_1, \ldots, \mathbf{X'}_G)$, where G is an epoch length, we apply token-level masking for MLM and non-masked tokens are mapped to embedding vectors. We then aggregate the embedded tensor by max pooling over the channel C and token length L dimensions, and apply rotary positional embedding (RoPE) [Su et al., 2024] to inject temporal order across epochs, producing a sequence of epoch-level features $(\mathbf{h_1}, \ldots, \mathbf{h_G})$. This epoch-level feature sequence is passed to a sequential encoder to produce contextual representations $(\mathbf{z}_1, \ldots, \mathbf{z}_G)$. We select only $\mathbf{z}_i$ corresponding to epochs with masked tokens and predict the masked token $\mathbf{x'_i}$ via an MLP head and softmax: $\mathcal{L}_{\text{pretrain}} = -\frac{1}{|M|} \sum_{i \in M} \log p_{i,\mathbf{x'}}$, where M is the number of masked tokens and p_{i,x'_i} is the predicted probability.

Fine-tuning for Downstream Task. For downstream sleep stage classification, we append a two-layer MLP classifier to the output of the sequential encoder. Given a sequence of epoch-level EEG tokens $\mathbf{S}$ for each epoch, the sequential encoder and classifier predict sleep stage labels Y. The model is fine-tuned using the cross-entropy loss: $\mathcal{L}_{\text{fine-tune}} = -\sum_{i \in G} \log p_{i,y_i}$.

Table 2: Comparison of sleep staging performance. Task-specific models (regular) and pre-trained models (*italic*); **best** and second-best scores are indicated.

Method	Overall Metrics (%)				Per-class F1 (%)				
	Acc(↑)	κ(↑)	MF1(↑)	ITR(↓)	W(↑)	N1(↑)	N2(↑)	N3(↑)	REM(↑)
AttnSleep	83.7	0.775	75.9	43.7	91.2	37.7	**86.6**	**85.9**	78.0
BIOT	84.4	0.780	73.2	29.6	**92.8**	28.5	85.5	76.5	82.8
LaBram	81.9	0.746	69.2	35.4	90.8	18.4	83.3	76.7	76.9
EEGPT	83.0	0.767	72.0	33.7	92.0	26.0	85.0	75.0	81.0
TFM tokenizer	77.0	0.671	61.6	40.6	86.7	7.9	79.1	72.4	62.6
Our Method (90 epochs)	**84.6**	**0.784**	**76.6**	**15.2**	92.1	**44.2**	85.1	75.3	**86.3**

3 Experiments

3.1 Experiment Setup

Dataset and Baselines. We used SHHS [Stephansen et al., 2018, Zhang et al., 2018], which contains large-scale sleep recording data from 5,793 individuals. The data was divided into seven parts at the subject level, with five parts used for training, one for validation, and one for testing to balance between model generalization and reliable evaluation. Both the training and validation datasets were used in both the pretraining and fine-tuning. We collected five baselines. AttnSleep [Eldele et al., 2021a] classifies sleep stages using CNN and self-attention to capture long-term dependencies. TFM-Tokenizer [Pradeepkumar et al., 2025] transforms EEG into discrete tokens and applies them to downstream tasks. BIOT [Yang et al., 2024] is a EEG FM with the contrastive loss. LaBraM [Jiang et al., 2024] is a Vector-Quantization AutoEncoder (VQVAE)-based model with MLM. EEGPT [Wang et al., 2024] is also a large-scale general EEG FM that learns EEG tokens via MLM and contrastive learning. Regarding foundation models, we used pre-trained models.

Metrics. We evaluate all model performance using Accuracy (Acc), macro-averaged F1-score (MF1) [Yang and Liu, 1999], and Cohen's kappa(κ) [McHugh, 2012] for overall performance and class-specific F1 score. To objectively assess the physiological plausibility of predicted transitions, we propose the *Irregular Transition Rate (ITR)*. ITR is defined as the proportion of irregular sleep stage transitions ζ in the predicted sequence that correspond to transitions considered impossible according to standard clinical rules of sleep stages, as defined in Table 1 [Zhu et al., 2025]. ITR is as follows:

Table 1: Irregular Sleep Stage Transitions : (ζ_{sleep})

Irregular Transitions
Wake → N3 / REM
N1 → N3 / REM
N2 → Wake
N3 → N1
REM → N1 / N3

$$\text{ITR}(\%) = 100 \times \frac{\zeta_{\text{sleep}}}{\text{Number of all transitions (excluding } y_i = y_{i+1})} \tag{1}$$

where y_i is the predicted sleep stage at epoch i.

Training Setup. Our experiments were performed using the AdamW optimizer with a learning rate of 1×10^{-4}. The optimizer hyperparameters were set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay $= 0.01$. We used a batch size of 128. Pre-training was conducted for 50 epochs, while fine-tuning was performed for 20 epochs with early stopping on three NVIDIA RTX A6000 GPUs.

3.2 Results

Main Results. Table 2 presents results for our model that uses a 90-epoch sequence as the atomic training unit, compared with baseline methods on the SHHS dataset. It shows our model achieves superior overall performance to both task-specific and pre-trained models in the sleep staging task. In particular, higher κ and lower ITR indicate that our model produces more reliable predictions. Consistent with this, our ITR is less than half the baselines, suggesting that sequence-based EEG training enables the model to learn physiologically plausible stage-transition patterns without explicit penalties for forbidden transitions.

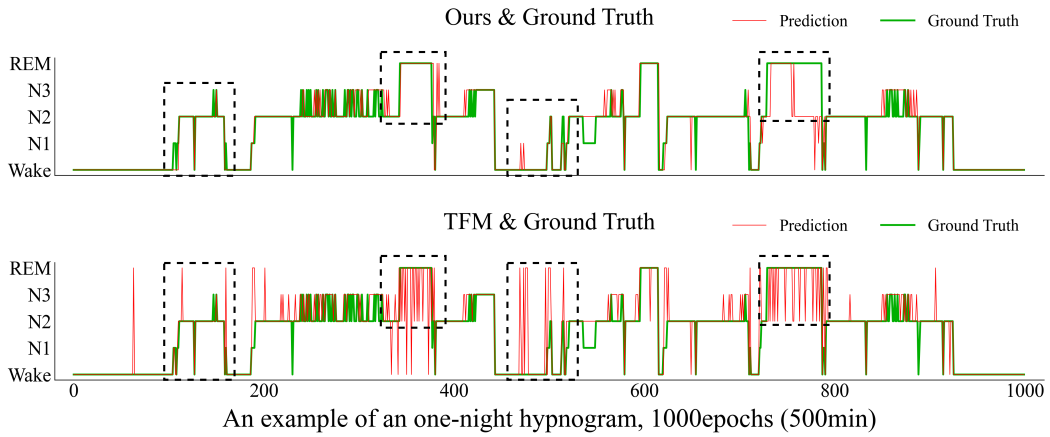


Figure 4: Hypnograms: Our model predicts physiologically plausible sleep-stage progressions through sequence learning, whereas TFM often produces physiologically impossible transitions.

EEG Epoch Length Evaluation. We evaluate the number of epochs in a sequence. As shown in Fig. 3, the best performance occurs with 90 epochs (45 min), approximately half of a typical 90-min sleep cycle [Hartmann, 1968]. This result suggests that too long sequences do not contribute to performance, possibly due to the inclusion of redundant or less informative temporal patterns.

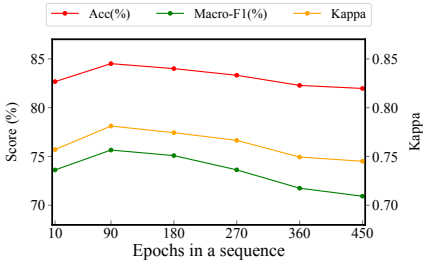


Figure 3: Effect of the number of epochs on our model performance.

Case Study. Furthermore, by visualizing the model's sleep stage predictions in Fig. 4, we evaluate its validity. This figure presents a one-night hypnogram composed of 1000 epochs (500 minutes), comparing sleep stage classification results from our model and the TFM tokenizer baseline. Both models utilize TFM as the base encoder for single-epoch tokenization. Compared to TFM, our model exhibits substantially improved temporal consistency and better alignment with the ground truth, reducing abrupt and biologically implausible stage transitions. Notably, in the first and third highlighted region, our model corrects misclassifications where TFM predicted a direct transition from Wake to REM, an implausible pattern rarely observed in real sleep dynamics. In the second and forth region, we refine transitions where TFM incorrectly predicts N2 to REM. These corrections illustrate that our models superior performance stems from its ability to mitigate irregular transitions. While we adopt the same per-epoch tokenizer as TFM, our model further captures long-range temporal dependencies across multiple epochs to produce smoother and more physiologically plausible stage sequences.

4 Conclusion

In this study, we proposed a framework that capture long-term dependencies across EEG tokenized EEG epochs, exceeding the short context window of conventional methods. By further aggregating tokenized EEG epochs into multi-epoch sequences, our method enabled learning with approximately half the EEG length of a sleep cycle. The model learns both sequence-level context and discriminative patterns through pretraining that combines time-series masked token prediction with a cross-entropy objective function. Experiments on sleep stages demonstrated that our model improves overall prediction performance and significantly, reducing biologically implausible stage transitions. These findings show that modeling long-term dependencies lays a foundation for broader clinical/BCI applications and opening the possibility of extension to other domains.

5 Acknowledgment

This work was supported by Graduate Degree Program for Interdisciplinary Research Fields and Program for Leading Graduate Schools of The University of Osaka, Japan, JSPS KAKENHI Grant-in-Aid for Scientific Research Number JP24K20778, JST BOOST JPMJBS2402, JST CREST

JPMJCR23M3, JST START JPMJST2553,JST CREST JPMJCR20C6, JST K Program JPMJKP25Y6, JST COI-NEXT JPMJPF2009, JPMJPF2115.

References

Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Carole L. Marcus, and Bradley V. Vaughn. The AASM Manual for the Scoring of Sleep and Associated Events. *American Academy of Sleep Medicine*, pages 1689–1699, 2013.

Zheng Chen, Lingwei Zhu, Ziwei Yang, and Renyuan Zhang. Multi-tier platform for cognizing massive electroencephalogram. In *IJCAI-22*, pages 2464–2470, 2022.

Zheng Chen, Ziwei Yang, Lingwei Zhu, Wei Chen, Toshiyo Tamura, Naoaki Ono, Md Altaf-Ul-Amin, Shigehiko Kanaya, and Ming Huang. Automated sleep staging via parallel frequency-cut attention. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1974–1985, 2023.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021a.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2352–2359, 2021b.

Ernest Hartmann. The 90-minute sleep-dream cycle. *Archives of General Psychiatry*, 18(3):280–286, 1968.

Ziyu Jia, Xiyang Cai, Gaoxing Zheng, Jing Wang, and Youfang Lin. Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Transactions on Artificial Intelligence*, pages 248–257, 2020.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. Splitsee: A splittable self-supervised framework for single-channel eeg representation learning. In *2024 IEEE International Conference on Data Mining (ICDM)*, pages 741–746, 2024.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, pages 1285–1296, 2019.

Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L Kappel, Anjula C De Silva, and Chamira US Edussooriya. Towards interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Single-channel eeg tokenization through time-frequency modeling, 2025. URL `https://arxiv.org/abs/2502.16060`.

Jens B Stephansen, Alexander N Olesen, Mads Olsen, Aditya Ambati, Eileen B Leary, Hyatt E Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications*, 9(1):5229, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1999.

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.

Lingwei Zhu, Zheng Chen, Yukie Nagai, and Jimeng Sun. Towards physiologically sensible predictions via the rule-based reinforcement learning layer, 2025. URL `https://arxiv.org/abs/2501.19055`.