SD-Former: Sparse Directed Attention as a Structural Prior for Robust Physiological Time Series

Ankur Sharma
Exea Labs

Abstract

Physiological time-series models frequently suffer performance degradation under subject, device, and dataset shifts, largely because dense sequence architectures treat all input channels as symmetrically interacting and fail to exploit underlying structured dependencies. To address this limitation, we introduce SD-Former, a graph-conditioned biosignal architecture that integrates spectral tokenization, sparse dependency inference, and transformer representations modulated by a learned directed graph. We frame this dependency structure as a causally inspired inductive bias, combining ideas from continuous DAG learning with structured attention masking in an end-to-end sequence model. Across three public benchmarks evaluated over five independent seeds, SD-Former achieves the best results on Sleep-EDF (macro F1 0.7569 ± 0.0037, AUROC 0.9402 ± 0.0012) and the best macro F1 and calibration on PTB-XL (F1 0.5521 ± 0.0109, ECE 0.0282 ± 0.0051), while its AUROC on PTB-XL is comparable to the strongest baseline. In cross-dataset transfer from CHB-MIT to Sleep-EDF, SD-Former achieves the highest AUROC (0.696 ± 0.034) among evaluated baselines. An ablation study shows that the utility of the composite objective depends on dataset-specific statistical properties, with the clearest benefits observed on CHB-MIT. These findings suggest that learned sparse directed graphs used as dynamic attention masks can serve as an effective structural prior for physiological modeling under distribution shift.

1 Introduction

Physiological machine learning systems are routinely deployed in environments characterized by distribution shifts. Electroencephalography (EEG) signals exhibit substantial variability across subjects, electrode montages, and underlying neurological states; similarly, electrocardiography (ECG) recordings vary across hardware devices, rhythm disturbances, and institutional preprocessing pipelines. Deep sequence models frequently absorb these variations as spurious correlations rather than isolating invariant features that remain robust across recording conditions [Roy et al., 2019, Strodthoff et al., 2021]. Consequently, such architectures may achieve high performance on identically distributed held-out splits yet degrade substantially when subjected to covariate shifts inherent in novel clinical acquisition processes.

This work investigates whether structured graph conditioning can serve as an effective structural prior for mitigating such generalization failures. Our central hypothesis is that physiological channels are non-exchangeable; therefore, an architecture that infers sparse, directed, and state-dependent inter-channel dependencies may exhibit improved out-of-distribution generalization compared to approaches relying on dense attention or static graph topologies. We formulate this directed dependency learning as a causally inspired inductive bias, deliberately distinguishing it from formal causal identification procedures, which require assumptions generally unsatisfied in observational physiological datasets.

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

We operationalize this hypothesis via SD-Former, a graph-conditioned transformer architecture designed for physiological time series. The model processes raw waveforms by projecting them into a frequency-resolved spectral token space, infers a sparse directed adjacency matrix characterizing token interactions via continuous optimization, and uses this graph to enforce structured sparsity on the self-attention mechanism. The key idea is to combine continuous DAG inference—previously studied in the structure learning literature—with transformer sequence modeling, using the learned graph as an attention mask. While graph-constrained attention and differentiable structure learning have each been explored independently, their integration as a dynamic attention bottleneck for physiological signals has, to our knowledge, not been previously studied. The architecture is optimized via a two-phase training protocol that separates structural representation learning from classifier refinement.

Our empirical evaluation delineates the specific contexts in which this structural prior confers measurable advantages. SD-Former achieves the best results on Sleep-EDF and competitive results on PTB-XL, while its performance on CHB-MIT exhibits higher variance. The clearest benefits appear in cross-dataset transfer, particularly from CHB-MIT to Sleep-EDF, where SD-Former outperforms the evaluated baselines in AUROC. An ablation analysis further shows that the utility of the composite objective function depends on dataset-specific characteristics rather than being universally beneficial.

Our core contributions are:

1. SD-Former, a graph-conditioned biosignal architecture that combines continuous DAG inference with transformer self-attention for multivariate physiological time series.
2. The use of dynamically learned sparse directed graphs as attention masks, providing a structured bottleneck that avoids reliance on static topologies.
3. An informal capacity argument observing that for a *fixed* sparse attention mask with bounded in-degree k, standard Rademacher complexity techniques yield a $\sqrt{k/N}$ reduction in the attention capacity term relative to dense attention. We note that this argument does not account for the additional capacity introduced by learning the graph itself.
4. An artifact-backed empirical evaluation with 5-seed supervised experiments, 3-seed ablations, and hyperparameter analyses across Sleep-EDF, CHB-MIT, and PTB-XL.
5. A failure-mode analysis showing that while sparse graph-conditioning can improve transfer robustness, its benefits depend on the dependency structure of the target domain.

2 Related Work

Biosignal representation models. Recent biosignal models scale representation learning for EEG and related modalities using self-supervision and transformer backbones [Kostas et al., 2021, Jiang et al., 2024]. These systems focus primarily on sequence modeling and large-scale pretraining. They do not explicitly constrain inter-channel interactions with a learned sparse graph.

Directed and structure-aware time-series learning. Continuous optimization methods such as NOTEARS and its time-series variants provide a differentiable route to directed graph learning [Zheng et al., 2018, Pamfil et al., 2020]. Our work borrows that structural perspective but uses it as a neural inductive bias inside an end-to-end sequence model. We do not claim causal identification, which would require interventions, identifiability assumptions, and confounding control beyond the available EEG and ECG benchmarks.

Time-series transformers and graph biosignal models. PatchTST and related transformer models are strong baselines for time-series prediction and classification [Nie et al., 2023, Wu et al., 2023]. Their default attention patterns are dense or channel-independent, which makes them flexible but agnostic to structured dependencies among recording channels. Fixed graph neural networks have also been used for EEG tasks [Song et al., 2020, Covert et al., 2019]; our comparison focuses on whether adding learned graph constraints changes robustness under shift.

3 Problem Formulation

Let $X \in \mathbb{R}^{C \times T}$ denote a multi-channel physiological time series with C channels and T time points. We divide the signal into windows $X_w \in \mathbb{R}^{C \times T_w}$ and associate each window with a task label $y_w \in \{1, \ldots, K\}$. The learning problem is to map each window to its label while maintaining robustness to shifts in subject composition, sensor layout, and dataset origin.

Our modeling assumption is that each window is governed by a sparse dependency structure over channels or channel-derived tokens. We use this assumption only as a structural prior: there exists a directed graph $G_w = (V, E_w)$ that is useful for organizing interactions within the window, and this graph may vary across windows and datasets. We do not assume the learned graph is a scientifically correct mechanistic graph.

We evaluate two settings. In-distribution evaluation trains and tests on subject-stratified splits of the same dataset. Cross-dataset transfer trains the backbone on a source dataset, freezes it, and fits a linear probe on a target dataset. This transfer protocol directly tests whether the learned representation retains information that survives a dataset shift.

4 Method

Spectral tokenizer. Raw windows are transformed into short-time frequency representations and summarized into five physiologically motivated bands: delta, theta, alpha, beta, and gamma. For each channel-band pair, a small MLP projects the corresponding spectral summary into a token embedding. This produces a token sequence of length $N = C \times 5$ with token dimension $d_{\text{token}} = 128$.

For high-channel datasets, full channel-band tokenization yields many pairwise interactions. We therefore optionally pool the five band tokens per channel into a single channel token for graph inference, while retaining the unpooled band tokens for reconstruction. Band pooling is disabled for Sleep-EDF and enabled for CHB-MIT and PTB-XL.

Sparse dependency inferencer. The token sequence is passed through a lightweight encoder that produces context-aware embeddings $H \in \mathbb{R}^{N \times d}$. For each ordered token pair (i, j), an edge-scoring MLP predicts a directed dependency score,

$$s_{ij} = \text{MLP}_{edge}([h_i \| h_j]), \qquad \hat{a}_{ij} = \mathbb{I}[\sigma(s_{ij}) > 0.5].$$

Gradients pass through the threshold with a straight-through estimator. We regularize the graph with a NOTEARS-style acyclicity penalty and an ℓ_1 sparsity penalty.

Graph-conditioned transformer. The backbone is a 6-layer pre-norm transformer with 4 attention heads, hidden width 128, and feed-forward width 512. The learned adjacency matrix acts as an attention mask: if $\hat{a}_{ij} = 0$, token j cannot influence token i through attention. Self-attention on the diagonal remains allowed. This masking mechanism provides a structured dependency bottleneck that can reduce effective interaction complexity and potentially improve transfer robustness.

Training objective. Training uses two phases. In Phase 1, the model optimizes

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{recon}\mathcal{L}_{recon} + \lambda_{dep}\mathcal{L}_{dep} + \lambda_{dag}\mathcal{L}_{dag} + \lambda_{sparse}\mathcal{L}_{sparse}.$$

In Phase 2, training switches to classification-only refinement. Optimization uses AdamW, cosine decay, warmup, gradient clipping, and early stopping.

5 Capacity Motivation

We provide an informal capacity argument motivating why sparse attention masking may reduce overfitting. This is not a novel generalization bound; it applies standard Rademacher complexity reasoning to the restricted interaction structure of a *fixed* sparse mask and serves to motivate the architectural design.

Consider a *fixed* binary attention mask with maximum in-degree $k \ll N$. Assume the spectral token embeddings satisfy $\|z_i\| \leq B_x$ and each transformer layer is α-Lipschitz (achievable via spectral normalization or bounded weight norms).

Remark 1 (Capacity reduction under fixed sparse masking). Under these assumptions, the per-layer attention capacity term in a Rademacher complexity bound scales with the number of active attention edges per token. For a fixed mask with in-degree k, the softmax in each attention head operates over k entries rather than N, and the resulting attention output for token i is a convex combination of at most k value vectors. Following the layer-wise contraction argument of Golowich et al. [2018], composing L such layers yields an overall capacity term of order

$$O\left(\frac{\rho B_x \alpha^L \sqrt{k d_h}}{\sqrt{n}}\right),$$

where d_h is the attention head projection dimension, n is the number of training samples, and ρ subsumes constants (weight matrix norms, number of heads, FFN parameters) that are independent of N and k. For dense attention the corresponding term replaces k with N, giving a factor of $\sqrt{k/N}$ reduction.

Important caveats. This argument has two significant limitations. First, it assumes a *fixed* sparse mask. In SD-Former, the mask is learned from data by the edge-scoring MLP, so the effective hypothesis class is the union over all k-sparse directed graphs of the per-graph function classes. The capacity cost of this graph selection is not accounted for above and could, in principle, offset the sparsity benefit. Bounding this additional term rigorously would require a covering number or PAC-Bayes argument over the graph space, which we do not attempt. Second, the assumptions—bounded degree, Lipschitz layers, correct sparsity structure matching the data—are strong and unlikely to hold exactly in practice. We therefore treat this remark as architectural motivation rather than as a formal guarantee.

6 Experimental Setup

We evaluate on Sleep-EDF [Goldberger et al., 2000], CHB-MIT [Shoeb, 2009], and PTB-XL [Wagner et al., 2020]. The main supervised study uses five seeds (42, 123, 7, 0, 256). The ablation study uses three seeds (42, 123, 7). Hyperparameter sweeps use one seed (42). We report mean and standard deviation over the relevant seed budget and make no significance claims from one-seed studies.

Table 1: Dataset summary.

Dataset	Task	Signal	Ch.	Classes	Window
Sleep-EDF	Sleep staging	EEG	2	5	30 s
CHB-MIT	Seizure detection	EEG	23	2	2 s
PTB-XL	Arrhythmia classification	ECG	12	5	10 s

Baselines. We compare against four architectures: (1) **PatchTST** [Nie et al., 2023], a channel-independent patching transformer (patch length 16, 3 layers, 4 heads); (2) a **Vanilla Transformer** with the same depth and width as SD-Former's backbone (6 layers, 4 heads, width 128) but without graph masking or auxiliary losses; (3) a **Static GNN** that uses a fixed adjacency matrix derived from electrode distance (2-layer graph convolutional network, width 128); and (4) a **Raw Waveform** 1D-CNN baseline (5 convolutional blocks, width 128). All baselines use the same training schedule, optimizer, and early stopping criteria as SD-Former for a controlled comparison.

6.1 Implementation Notes

All experiments were run on a single AMD MI300X with ROCm 6.2. Mixed precision uses bfloat16. Training uses AdamW with learning rate 1×10^{-3}, weight decay 1×10^{-2}, batch size 64, cosine learning rate decay with 5-epoch linear warmup, and gradient clipping at norm 1.0. Phase 1 trains for up to 50 epochs; Phase 2 trains for up to 30 epochs. Default loss weights are $\lambda_{\text{recon}} = 0.1$, $\lambda_{\text{dep}} = 0.1$, $\lambda_{\text{dag}} = 0.5$, $\lambda_{\text{sparse}} = 0.01$. CHB-MIT checkpoint selection uses validation AUROC; Sleep-EDF and PTB-XL use validation macro F1. All tables and figures were regenerated from saved JSON results and logs. The ablation study reports one spectral/reconstruction condition to avoid counting duplicated auxiliary-loss settings as independent evidence.

7 Results

7.1 Main Supervised Results

Table 2 presents the supervised evaluation across all benchmarks. On Sleep-EDF, SD-Former achieves the best results across all evaluated metrics (macro F1, AUROC, and ECE). On PTB-XL, the architecture obtains the highest macro F1 and lowest ECE, while its AUROC is comparable to the raw waveform baseline. On CHB-MIT, the Vanilla Transformer marginally outperforms SD-Former in both F1 and AUROC, and calibration differences are within the cross-seed variance. This mixed pattern illustrates that the benefit of explicit structural priors varies across tasks and datasets.

Table 2: Main supervised results from the 5-seed rerun. Bold indicates best per dataset-metric column.

Model	Sleep-EDF F1↑	Sleep-EDF AUROC↑	Sleep-EDF ECE↓	CHB-MIT F1↑	CHB-MIT AUROC↑	CHB-MIT ECE↓	PTB-XL F1↑	PTB-XL AUROC↑	PTB-XL ECE↓
SD-Former	**0.7569 ± 0.0037**	**0.9402 ± 0.0012**	**0.0149 ± 0.0032**	0.6251 ± 0.0762	0.7671 ± 0.1466	0.1011 ± 0.0313	**0.5521 ± 0.0109**	0.8498 ± 0.0048	**0.0282 ± 0.0051**
PatchTST	0.7312 ± 0.0037	0.9303 ± 0.0009	0.0338 ± 0.0061	0.5480 ± 0.1324	0.6620 ± 0.1999	0.1292 ± 0.0678	0.5245 ± 0.0052	0.8368 ± 0.0025	0.0415 ± 0.0049
Vanilla Transformer	0.5293 ± 0.0017	0.8148 ± 0.0046	0.2843 ± 0.0420	**0.6815 ± 0.0516**	**0.7777 ± 0.0911**	**0.1008 ± 0.0515**	0.2797 ± 0.0082	0.6199 ± 0.0094	0.3279 ± 0.0539
Static GNN	0.4983 ± 0.0014	0.8033 ± 0.0010	0.1637 ± 0.0075	0.4430 ± 0.0310	0.6214 ± 0.0752	0.1306 ± 0.0530	0.1749 ± 0.0115	0.5138 ± 0.0122	0.0744 ± 0.0341
Raw Waveform	0.7037 ± 0.0075	0.9228 ± 0.0038	0.0522 ± 0.0062	0.5578 ± 0.1429	0.7318 ± 0.1079	0.1214 ± 0.0883	0.5169 ± 0.0075	**0.8527 ± 0.0030**	0.0449 ± 0.0081

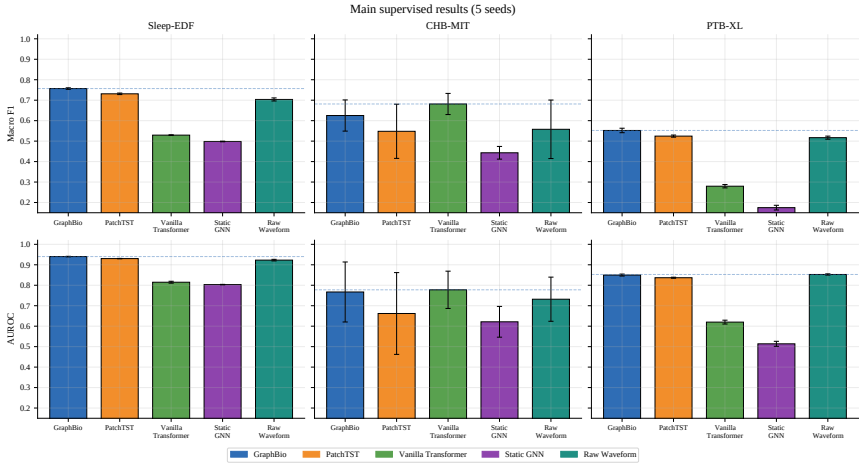

Figure 1: Main supervised results for the latest 5-seed rerun. SD-Former is strongest on Sleep-EDF and remains strong on PTB-XL, while CHB-MIT is the least stable dataset in the suite.

7.2 Cross-Dataset Transfer

The clearest evidence for the utility of the structural prior appears in cross-dataset transfer (Table 3). In the CHB-MIT $\rightarrow$ Sleep-EDF direction, SD-Former achieves the highest AUROC among evaluated models, exceeding the Raw Waveform baseline by 2.1 absolute points. However, the Raw Waveform baseline retains a small advantage in macro F1, so the improvement is metric-dependent rather than uniform.

The reverse direction (Sleep-EDF $\rightarrow$ CHB-MIT) is difficult for all architectures: macro F1 scores approach chance level and AUROC estimates have high variance. SD-Former's AUROC of 0.568 ± 0.183 nominally leads in this direction, but the standard deviation is large enough that the difference from the Vanilla Transformer (0.561 ± 0.062) is not meaningful; we bold it in the table only for visual consistency but do not claim a real advantage. The cross-dataset generalization is therefore asymmetric, and we characterize SD-Former's transfer benefit as specific to the CHB-MIT $\rightarrow$ Sleep-EDF direction.

Table 3: Cross-dataset transfer via linear probing. Bold indicates best mean per column; see text for caveats on high-variance entries.

Model	CHB-MIT → Sleep-EDF F1	CHB-MIT → Sleep-EDF AUROC	Sleep-EDF → CHB-MIT F1	Sleep-EDF → CHB-MIT AUROC
SD-Former	0.157 ± 0.050	**0.696 ± 0.034**	0.440 ± 0.033	**0.568 ± 0.183**
PatchTST	0.086 ± 0.020	0.532 ± 0.041	0.440 ± 0.033	0.430 ± 0.054
Vanilla Transformer	0.164 ± 0.009	0.525 ± 0.014	**0.441 ± 0.031**	0.561 ± 0.062
Static GNN	0.108 ± 0.006	0.517 ± 0.006	0.440 ± 0.033	0.475 ± 0.040
Raw Waveform	**0.172 ± 0.017**	0.675 ± 0.079	0.440 ± 0.033	0.523 ± 0.046

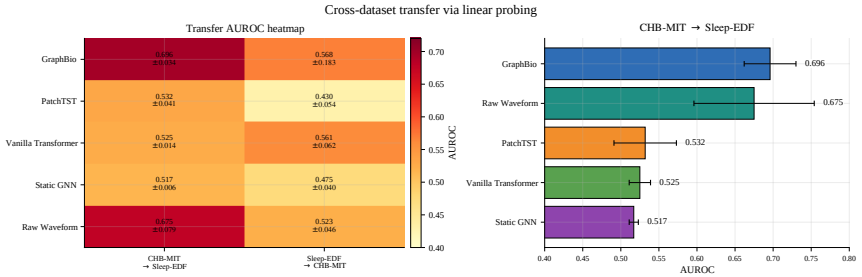

Figure 2: Transfer AUROC heatmap and CHB-MIT → Sleep-EDF comparison. The clearest positive result is CHB-MIT → Sleep-EDF, where SD-Former attains the best AUROC.

7.3 Ablation Study

The ablation study (Table 4) shows that the benefit of the full SD-Former objective depends on the dataset. On CHB-MIT, removing any individual component degrades macro F1, providing the clearest support for the composite objective.

On Sleep-EDF, only removing the DAG penalty produces a clear performance drop. On PTB-XL, the classification-only configuration slightly outperforms the full multiphase objective in our three-seed evaluation, which challenges the necessity of the auxiliary losses for this dataset. To avoid inflating evidence through correlated conditions, we report a single combined spectral/reconstruction ablation rather than separate entries for overlapping auxiliary-loss settings.

Table 4: Ablation study from the 3-seed experimental budget. The spectral/reconstruction row represents the single retained overlapping auxiliary-loss ablation.

Configuration	Sleep-EDF F1	Sleep-EDF Δ	CHB-MIT F1	CHB-MIT Δ	PTB-XL F1	PTB-XL Δ
Full model	0.7538	–	0.6887	–	0.5523	–
No dependency loss	0.7563	+0.0025	0.6532	-0.0355	0.5542	+0.0019
No DAG penalty	0.7489	-0.0049	0.6535	-0.0352	0.5595	+0.0072
No spectral/recon	0.7597	+0.0059	0.6444	-0.0443	0.5643	+0.0120
Classification only	0.7550	+0.0012	0.6560	-0.0327	0.5684	+0.0161

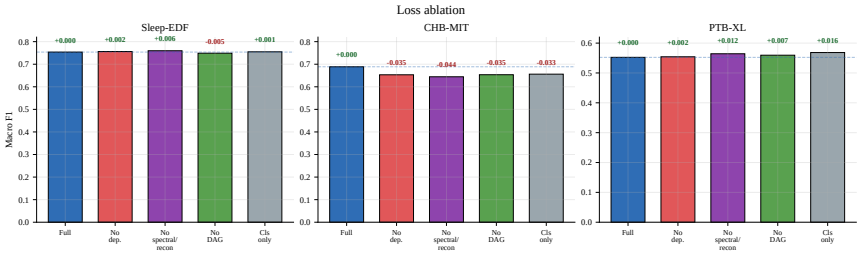

Figure 3: Ablation F1 by dataset. The effect of the full objective is clearest on CHB-MIT and mixed on Sleep-EDF and PTB-XL.

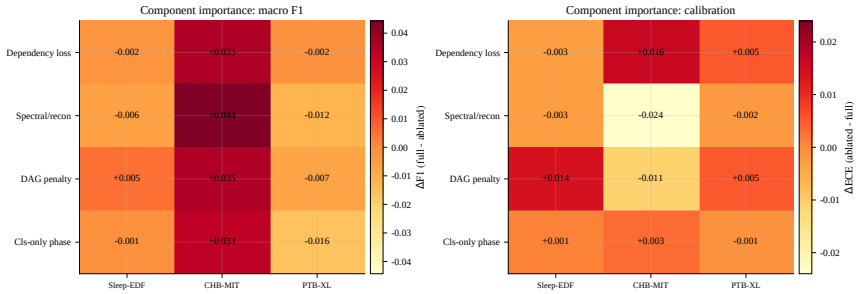

Figure 4: Descriptive component-importance heatmaps from the ablation study. In these heatmaps, the sign convention is inverted relative to Table 4: positive ΔF1 here indicates the full model outperforms the ablation (i.e., removing the component hurts), and positive ΔECE indicates the ablation has higher (worse) calibration error than the full model.

7.4 Hyperparameter Sweeps and Calibration

The single-seed hyperparameter sweeps are intended to identify directional trends rather than to establish stability. On Sleep-EDF, the results support the default $\lambda_{\text{dep}} = 0.1$, which jointly maximizes macro F1 and AUROC. Both CHB-MIT and PTB-XL show sensitivity to the latent token dimensionality, suggesting room for dataset-specific tuning. We caution against overinterpreting single-seed results.

Across the five-seed evaluation, SD-Former achieves the lowest ECE on Sleep-EDF and PTB-XL. On CHB-MIT, its calibration is comparable to the Vanilla Transformer. These results are consistent with structural constraints providing calibration benefits on some datasets, though they do not establish the approach as universally superior for calibration.

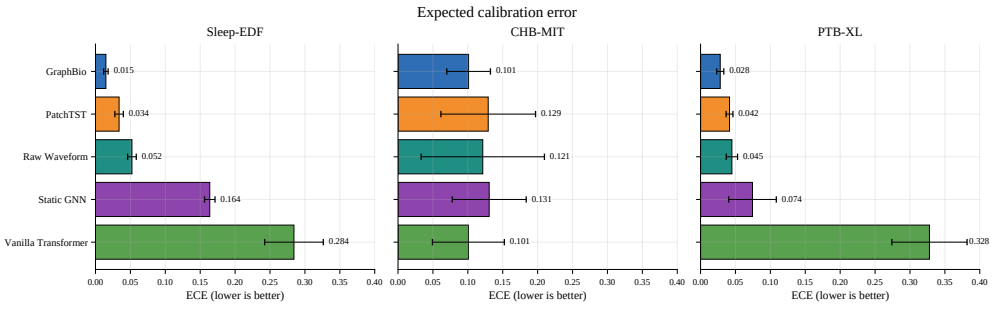

Figure 5: Expected calibration error. Lower is better.

7.5 Computational Cost

Table 5 summarizes parameter counts. SD-Former's overhead relative to the Vanilla Transformer comes from the edge-scoring MLP ($\sim$0.13M parameters) and the reconstruction decoder ($\sim$0.26M). Total model size varies with the number of input channels: approximately 1.8M parameters on Sleep-EDF (2 channels), 2.4M on PTB-XL (12 channels), and 3.1M on CHB-MIT (23 channels). All configurations fit on a single AMD MI300X with room to spare.

Table 5: Approximate parameter counts (millions) on the Sleep-EDF configuration.

Model	Params (M)
SD-Former	1.8
Vanilla Transformer	1.4
PatchTST	0.9
Static GNN	0.5
Raw Waveform	1.2

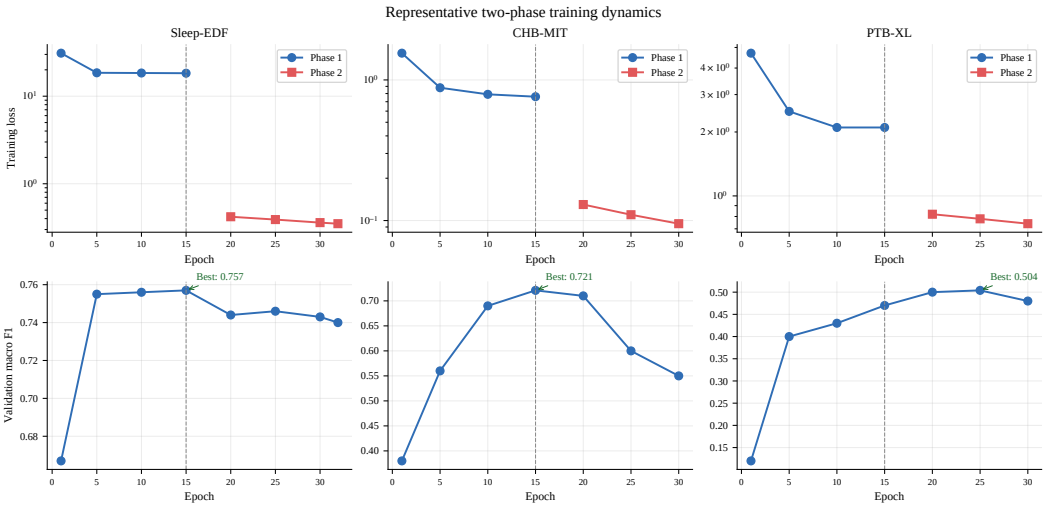

Figure 6: Representative two-phase training dynamics reconstructed from the original paper figure. The switch separates joint representation shaping from task-only refinement.

8 Limitations and Broader Impact

We acknowledge several limitations.

1. **Inconsistent empirical gains.** Sleep-EDF and the CHB-MIT $\rightarrow$ Sleep-EDF transfer support our hypothesis, but supervised performance on CHB-MIT is unstable, and PTB-XL ablations suggest the full auxiliary objective is not always necessary. The method does not uniformly outperform all baselines on all datasets.
2. **Ablation granularity.** The ablation study aggregates spectral and reconstruction penalties into a single condition to avoid double-counting correlated auxiliary losses, which limits fine-grained attribution.
3. **Single-seed hyperparameter sweeps.** These provide only directional guidance, not stability guarantees.
4. **No graph validation.** We do not visualize the learned graphs, analyze their stability across seeds, or compare them against random graphs with matched sparsity. Consequently, we cannot distinguish whether the learned graph captures meaningful structure or simply acts as a form of dropout-like regularization. This is an important direction for future work.

5. **Graph interpretability.** The learned graphs are structural priors for information routing, not validated causal or mechanistic physiological networks—they should not be interpreted as such.
6. **Limited transfer evaluation.** We evaluate only two transfer directions between two EEG datasets. A more comprehensive evaluation would include additional source-target pairs, domain adaptation baselines (e.g., CORAL, DAN), and invariance-based methods (e.g., IRM).
7. **Narrow baseline set.** We compare against four baselines. Comparison to additional modern architectures (e.g., iTransformer, TimesFM) and to recent EEG-specific models would strengthen the empirical picture.

From a broader impact perspective, improved calibration may benefit clinical decision support by providing better-calibrated confidence estimates. A corresponding risk is that visualized graph structures could be misinterpreted as physiological mechanisms, potentially leading to misguided clinical conclusions. We emphasize that the inferred graphs should be treated as mathematical regularizers, not as clinical ground truth.

9 Conclusion

We presented SD-Former, a structure-aware biosignal architecture that uses learned sparse directed graphs as attention masks in a transformer backbone. Across three physiological benchmarks, the architecture achieves the best results on Sleep-EDF, competitive calibration on PTB-XL, and the highest transfer AUROC in the CHB-MIT $\rightarrow$ Sleep-EDF setting. The ablation analysis shows that the full composite objective is beneficial on some datasets but not others, indicating that the value of the structural prior is context-dependent. These results do not demonstrate causal discovery, but they provide empirical evidence that graph-conditioned attention can improve generalization of physiological time-series models under distribution shift in specific settings.

A Sketch for Remark 1

Consider a single attention head at one layer. Under dense attention, each token attends to all N tokens: the softmax operates over N logits, and the output is a convex combination of N value vectors. The Rademacher complexity of the resulting function class scales with $\sqrt{Nd_h}$ via standard arguments [Golowich et al., 2018]. Under a *fixed* sparse mask with in-degree k, the softmax operates over k entries, and the corresponding Rademacher term scales with $\sqrt{kd_h}$.

Composing L such layers with per-layer Lipschitz constant α and applying the layer-wise contraction lemma yields

$$O\left(\frac{\rho B_x \alpha^L \sqrt{kd_h}}{\sqrt{n}}\right),$$

where ρ collects terms (weight norms, number of heads, FFN capacity) that do not depend on N or k.

This argument applies to a fixed mask. When the mask is learned, the effective function class is the union of per-graph classes over all k-sparse directed graphs, and the additional capacity cost of graph selection is not bounded by the above. Providing such a bound is an open direction.

B Extended Experimental Details

The main supervised study uses five seeds, the ablation study uses three seeds, and hyperparameter sweeps use one seed. CHB-MIT checkpoint selection uses validation AUROC; Sleep-EDF and PTB-XL use validation macro F1. All reported error bars are standard deviations over the corresponding seed budget.

References

I. Covert, B. Krishnan, I. Najm, and J. Zhan. Temporal graph convolutional networks for automatic seizure detection. *Machine Learning for Healthcare*, 2019.

Table 6: SD-Former main 5-seed recap.

Dataset	F1	AUROC	ECE
Sleep-EDF	0.7569 ± 0.0037	0.9402 ± 0.0012	0.0149 ± 0.0032
CHB-MIT	0.6251 ± 0.0762	0.7671 ± 0.1466	0.1011 ± 0.0313
PTB-XL	0.5521 ± 0.0109	0.8498 ± 0.0048	0.0282 ± 0.0051

A. Goldberger et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, 2000.

Y. Jiang, X. Li, and others. THD-BAR: Time-frequency hierarchical decomposition for brain activity representation. *NeurIPS*, 2024.

D. Kostas, S. Aroca-Ouellette, and F. Bhatt. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Y. Nie, N. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words. *ICLR*, 2023.

R. Pamfil, N. Srber, S. Acerbi, and others. DYNOTEARS: Structure learning from time-series data. *AISTATS*, 2020.

Y. Roy et al. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019.

A. Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, MIT, 2009.

T. Song, W. Zheng, P. Song, and Z. Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020.

N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE JBHI*, 25(5):1519–1528, 2021.

P. Wagner et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

H. Wu et al. TimesNet: Temporal 2D-variation modeling for general time series analysis. *ICLR*, 2023.

X. Zheng, B. Aragam, P. Ravikumar, and E. Xing. DAGs with NO TEARS. *NeurIPS*, 2018.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *COLT*, 2018.