MOSAIC: A DIFFERENTIABLE PHASE-AMPLITUDE COUPLING TRANSFORMER FOR MEDICAL TIME SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural oscillations interact through phase-amplitude coupling (PAC): the amplitude of a fast rhythm is modulated by the phase of a slow one. PAC is an established clinical biomarker, yet transformers for medical time series largely ignore it, treating signals as undifferentiated tokens. Classical PAC measures are typically computed offline on hand-chosen bands, so the bands that determine which coupling is detectable lie outside the classification objective. We present **MOSAIC** (Multi-Oscillatory Signal Attention with Integrated Coupling), which makes PAC estimation end-to-end differentiable. A spectrally applied learnable Gaussian filterbank lets gradients reach the band parameters, an FFT-based Hilbert transform gives instantaneous amplitude and phase, and a soft phase-binned amplitude histogram keeps the shape of the modulation rather than a single scalar. The resulting *PAC tokens* feed a cross-band transformer that learns which coupling pairs matter. On three subject-independent EEG and ECG benchmarks, jointly learning the bands and the coupling representation raises macro-F1 over fixed-band PAC features by $1.2\times$ to $1.6\times$, and MOSAIC attains the highest reported accuracy, recall, and macro-F1 on all three against eleven published transformer baselines.

1 INTRODUCTION

The brain is a multi-frequency oscillatory system. Neurons communicate through synchronized rhythms from delta (1–4 Hz) to gamma (30–150 Hz), each reflecting distinct computational states (Buzsaki & Draguhn, 2004), and these rhythms interact: the amplitude of a fast oscillation is modulated by the phase of a slow one. This *phase-amplitude coupling* (PAC) is among the most clinically informative signatures in electroencephalography (EEG), with disease-specific disruptions reported in dementia, movement disorders, and sleep pathology (Canolty et al., 2006; Tort et al., 2010; Jensen & Colgin, 2007).

Modern architectures for medical time series do not explicitly represent this structure. Patch-based transformers tokenize fixed-length windows without regard to oscillatory content (Nie et al., 2023; Wang et al., 2024), and convolutional models such as EEGNet (Lawhern et al., 2018) learn channel-specific filters but no explicit cross-frequency relationship. The natural remedy, feeding classical PAC measures to a classifier, runs into a specific technical obstacle. PAC requires narrowband filtering, a Hilbert transform, and a dependence statistic between slow-band phase and fast-band amplitude, and the standard pipeline performs all three offline on hand-fixed bands, so the classifier cannot adapt the bands or the coupling features. Filtering, the Hilbert transform, and MVL admit differentiable formulations, but hard phase binning has zero phase gradient almost everywhere. Joint learning therefore requires trainable band parameters and a soft alternative to hard binning.

The fixed-band assumption is not a harmless convenience. Aru et al. (2015) show that filter bandwidth determines what is even admitted as an oscillatory component: if the band around the fast oscillation is too narrow to contain the sidebands induced by the slow rhythm, genuine coupling becomes undetectable, and a fixed bandwidth biases the analysis toward slow modulators regardless of the physiology. Setting these parameters correctly requires knowing in advance which coupling the disease expresses, which is precisely what a classifier ought to discover. Yet many studies scan center frequencies at a fixed bandwidth, often without explicit justification.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

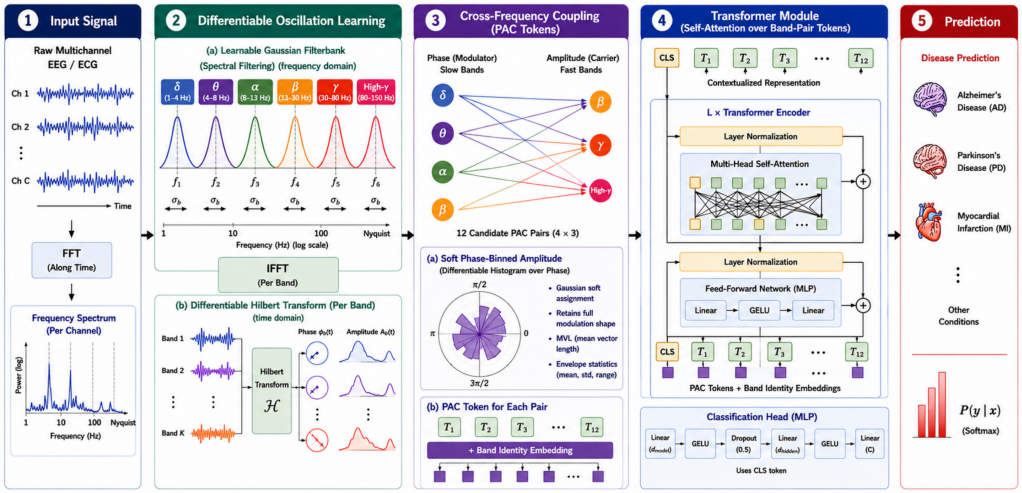

Figure 1: **The MOSAIC pipeline.** A learnable Gaussian filterbank ($B = 6$ bands, trainable f_b, σ_b) and a differentiable Hilbert transform give each band's phase and amplitude; four slow and three fast bands form 12 PAC tokens, which a cross-band transformer with a CLS token classifies.

Evaluation compounds the difficulty. Medical time series benchmarks hold out entire subjects, so every test subject is an unseen domain, and performance drops sharply relative to sample-level splits (Wang et al., 2024). A representation that encodes a physiological mechanism, rather than waveform statistics that vary with electrode placement and anatomy, is a plausible route to closing part of that gap.

MOSAIC. We make the entire PAC pipeline differentiable and train it jointly with the classifier that consumes it. We call the model **M**ulti-**O**scillatory **S**ignal **A**ttention with **I**ntegrated **C**oupling (MOSAIC); *integrated* indicates that coupling is estimated inside the network and optimized by the task loss, rather than computed offline for a classifier that cannot influence it. Three design decisions carry the work: Gaussian filters with free center frequencies and bandwidths, applied in the frequency domain so that gradients reach the band definitions; a *soft phase-binned amplitude histogram* that keeps the shape of the modulation across the phase cycle rather than only its magnitude; and one token per candidate band pair, so that a transformer learns which cross-frequency interactions matter instead of being told.

Contributions.

- A *differentiable PAC front end*: a learnable Gaussian filterbank and an FFT-based Hilbert transform that expose instantaneous amplitude and phase as differentiable functions of the band parameters, with an orthogonality penalty that keeps the filters spectrally distinct (§3).
- *Soft phase-binned amplitude* (PBA) coupling features, a differentiable relaxation of phase binning that retains the full modulation profile and approaches the classical estimator as the smoothing scale shrinks (§3).
- *PAC tokens and a cross-band transformer* with learned band-identity embeddings, letting attention select discriminative coupling pairs and exposing that selection for inspection (§3).
- An evaluation on three subject-independent EEG and ECG benchmarks showing that the learned coupling representation improves on fixed-band PAC features by $1.2\times$ to $1.6\times$ macro-F1, that MOSAIC leads twelve methods on accuracy, recall, and macro-F1 on all three, and that coupling and waveform modeling are complementary (§4).

2 Related Work

Transformers for medical time series. Transformers (Vaswani et al., 2017) are now widely used for time series, and a central design axis is what constitutes a token. PatchTST (Nie et al., 2023) embeds a patch of timestamps from one channel, iTransformer (Liu et al., 2024) embeds an entire channel, and Crossformer (Zhang & Yan, 2023) attends across both time and dimension. Medformer (Wang et al., 2024) targets medical classification directly, combining cross-channel patching, multi-granularity patch lengths, and two-stage self-attention within and across granularities; notably, its multi-granularity patching is motivated as a way to *simulate* frequency bands without band filters. MOSAIC instead represents the bands, and the coupling between them, explicitly. A parallel line of work explains such models post hoc, through shapelets (Huang et al., 2026) or joint time-frequency views (Tran et al., 2026), and a related effort develops transferability metrics for time series representations (Zhang et al., 2026). MOSAIC builds the frequency structure into the representation, so attention operates directly over the clinical measure.

EEG models and subject shift. EEGNet (Lawhern et al., 2018) uses depthwise separable convolutions tuned to EEG and remains a strong, very compact architecture. Because these benchmarks hold out entire subjects, each unseen subject is effectively a new domain, and the setting overlaps with domain generalization. ManyDG (Yang et al., 2023) treats every patient as a domain, EEG-DG (Zhong et al., 2024) targets multi-source generalization for motor imagery, and VREx (Krueger et al., 2021) and domain-adversarial training (Ganin et al., 2016) supply general-purpose objectives. These methods address the shift through a training objective on raw waveforms; we ask what the representation itself should encode, so the two directions are complementary.

Phase-amplitude coupling. Canolty et al. (2006) established that gamma amplitude is phase-locked to theta phase in human neocortex, Tort et al. (2010) introduced the mean vector length and modulation index that remain standard estimators, and Jensen & Colgin (2007) systematized the functional roles of cross-frequency coupling. Aru et al. (2015) provide a key methodological critique, documenting how bandwidth selection, non-stationarity, and sharp transients produce both false negatives and spurious coupling, and recommending control for band power when comparing coupling across conditions. Closest in spirit to our work, Soulat et al. (2022) avoid bandpass filtering by estimating oscillatory components with a state space model, which sidesteps band selection and yields credible intervals on the coupling parameters. That method is generative and unsupervised, fit per recording, and not trained against a downstream objective. MOSAIC differs in objective: it is discriminative, learns its parameters from a classification loss across a training population, and runs as a GPU layer inside the classifier. To our knowledge, it is the first PAC estimator trained jointly with the model that consumes it. Learning filter parameters inside a network has precedent in audio, where SincNet (Ravanelli & Bengio, 2018) learns the cutoffs of parametrized band-pass filters from raw waveforms. We adopt that principle but learn the band parameters *through* a coupling statistic, so the gradient shaping them carries cross-frequency rather than generic spectral information.

3 MOSAIC

Figure 1 shows the pipeline and Algorithm 1 states the PAC pathway and its update step.

3.1 Setup and Background

Let $X \in \mathbb{R}^{C\times T}$ be a sample with C channels and T timestamps, and $y \in \{1, \dots, Q\}$ its label among Q classes. We follow the subject-independent protocol of Wang et al. (2024), so that $\mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{test}} = \emptyset$. Each channel passes through the pipeline independently and the per-channel logits are averaged, so coupling is defined within a channel.

Given narrowband signals $x_p(t)$ and $x_a(t)$, PAC measures how strongly the amplitude envelope $A_a(t)$ of the faster band is modulated by the instantaneous phase $\phi_p(t)$ of the slower one. Both quantities derive from the analytic signal $z_b(t) = x_b(t) + j\,\mathcal{H}[x_b](t)$, where $\mathcal{H}$ is the Hilbert transform, giving $A_b(t) = |z_b(t)|$ and $\phi_b(t) = \angle z_b(t)$. The classical summary is the mean vector length (Tort et al., 2010),

$$\mathrm{MVL}_{pa} = \left| \tfrac{1}{T} \textstyle\sum_{t=1}^{T} A_a(t)\, e^{\,j\phi_p(t)} \right|, \qquad 0 \le \mathrm{MVL}_{pa} \le \bar{A}_a, \tag{1}$$

where $\bar{A}_a = T^{-1}\sum_{t=1}^{T} A_a(t)$ is the mean amplitude. The upper bound follows from the triangle inequality and is attained when all nonzero amplitude-weighted vectors share one phase; MVL is therefore sensitive to amplitude scale. Two further estimators serve as baselines. The modulation index compares the phase-binned amplitude distribution $P_k = \mathrm{PBA}_{pa}[k]/\sum_{k'} \mathrm{PBA}_{pa}[k']$ against the uniform distribution,

$$\mathrm{MI}_{pa} = \frac{\log K + \sum_{k=1}^{K} P_k \log P_k}{\log K} \in [0, 1], \tag{2}$$

and the phase-locking value measures the consistency of the phase difference between the slow band and the phase $\psi_a(t)$ of the amplitude envelope $A_a(t)$,

$$\mathrm{PLV}_{pa} = \left| \tfrac{1}{T} \sum_{t=1}^{T} e^{\,j(\phi_p(t) - \psi_a(t))} \right| \in [0, 1]. \tag{3}$$

These estimators are commonly computed as fixed features before classifier training; the stages below bring band learning and soft phase binning into the classifier.

3.2 Learnable Gaussian Filterbank

For each of B bands we maintain a Gaussian bandpass filter with center frequency $0 < f_b < f_s/2$ and bandwidth $\sigma_b > 0$, where f_s is the sampling rate,

$$G_b(f) = \exp\!\Big(-\tfrac{(f-f_b)^2}{2\sigma_b^2}\Big) + \exp\!\Big(-\tfrac{(f+f_b)^2}{2\sigma_b^2}\Big), \tag{4}$$

where the second term enforces Hermitian symmetry so that the impulse response is real. Filtering is applied spectrally,

$$x_b(t) = \mathcal{F}^{-1}\big(G_b(f) \cdot \mathcal{F}[x](f)\big), \tag{5}$$

which is differentiable in f_b and σ_b through the FFT. A Gaussian response suits this purpose: it is smooth in both parameters, has no ripple or discontinuity to differentiate through, and has a closed form at every frequency bin. A softplus keeps σ_b positive. To keep the centers ordered and below Nyquist, we use $B + 1$ positive gaps $q_i = \mathrm{softplus}(u_i)$ and set

$$f_b = \frac{f_s}{2} \frac{\sum_{i=1}^{b} q_i}{\sum_{i=1}^{B+1} q_i}, \qquad b = 1, \ldots, B.$$

This gives $0 < f_1 < \cdots < f_B < f_s/2$ and preserves the ordering of the phase- and amplitude-providing bands. The gaps are initialized to the clinical centers of the delta, theta, alpha, beta, gamma, and high-gamma bands, $\{f_b\} = \{2, 6, 10.5, 21.5, 55, 115\}$ Hz, with bandwidths $\{1.5, 2, 2.5, 8.5, 25, 35\}$ Hz, so training begins from established clinical practice and departs from it only where the data warrant.

Nothing so far prevents two filters from converging on the same sub-band, which would make their tokens redundant and the attention pattern hard to interpret. We therefore add an orthogonality penalty on the frequency responses,

$$\mathcal{L}_{\text{orth}} = \big\|\mathbf{G}^{\top}\mathbf{G} - \mathbf{I}_B\big\|_F^2, \tag{6}$$

where $\mathbf{G} \in \mathbb{R}^{F \times B}$ collects the B responses evaluated at F frequency bins, normalized per column, so that each token continues to describe a distinct oscillatory component as the bands adapt.

Differentiable Hilbert transform. For even T (all our inputs: $T = 256$ for EEG, $T = 300$ for ECG), the analytic signal of $x_b \in \mathbb{R}^T$ follows from the one-sided spectrum:

$$\hat{X}_b = \mathcal{F}[x_b], \tag{7}$$

$$\hat{H}_b[k] = \begin{cases} \hat{X}_b[k] & k = 0 \text{ or } k = T/2, \\ 2\,\hat{X}_b[k] & 1 \le k < T/2, \\ 0 & T/2 < k < T, \end{cases} \tag{8}$$

$$z_b(t) = \mathcal{F}^{-1}[\hat{H}_b] \in \mathbb{C}^T. \tag{9}$$

Instantaneous amplitude and phase are then $A_b(t) = |z_b(t)|$ and $\phi_b(t) = \mathrm{atan2}(\mathrm{Im}\, z_b(t), \mathrm{Re}\, z_b(t))$, both differentiable in x_b and hence in $\{f_b, \sigma_b\}$. Because Eqs. (5) and (7) both operate in the frequency domain, the filterbank and the Hilbert transform share one forward FFT; the stage costs $O(T \log T)$ and runs batched on GPU.

3.3 Coupling Features and PAC Tokens

We designate $B_\phi = 4$ phase-providing bands (delta, theta, alpha, beta) and $B_A = 3$ amplitude-providing bands (beta, gamma, high-gamma), giving $B_\phi \times B_A = 12$ candidate pairs; beta appears in both sets so that intra-band modulation remains representable. For each pair we compute mvl_{pa} as in Eq. (1).

A scalar magnitude, however, discards the shape of the modulation: two signals with equal coupling strength but different preferred phases are indistinguishable. The classical alternative averages amplitude within phase bins, but hard binning has zero gradient in ϕ_p almost everywhere and so blocks learning of the phase-band parameters. We use a soft assignment instead. Partition $[-\pi, \pi)$ into K equal bins of width $2\pi/K$ with centers $c_k = -\pi + (k - \frac{1}{2})2\pi/K$, and use a periodic kernel with smoothing scale $\delta = \pi/K$, distinct from the bin width,

$$w_{tk} = \frac{\exp\big((\cos(\phi_p(t) - c_k) - 1)/\delta^2\big)}{\sum_{k'=1}^{K} \exp\big((\cos(\phi_p(t) - c_{k'}) - 1)/\delta^2\big)}, \tag{10}$$

$$\widetilde{\mathrm{PBA}}_{pa}[k] = \frac{\sum_t A_a(t)\, w_{tk}}{\sum_t w_{tk}}, \quad k = 1, \ldots, K. \tag{11}$$

The cosine kernel is smooth and periodic, treating phases near $-\pi$ and $+\pi$ as neighbors, and is locally Gaussian for small angular differences. The representation is differentiable in A_a and ϕ_p, approaches hard nearest-bin assignment as $\delta \to 0$ (except for ties at bin boundaries), and records *where* in the phase cycle amplitude concentrates, not merely how much.

Writing $s_a = [\, \bar{A}_a \,\|\, \hat{\sigma}_{A_a} \,\|\, \mathrm{range}(A_a) \,]$ for the mean, standard deviation, and range of the amplitude envelope, each pair yields a token

$$\mathbf{t}_{pa} = W_{\mathrm{tok}} \big[\, \widetilde{\mathrm{PBA}}_{pa} \,\|\, \mathrm{mvl}_{pa} \,\|\, s_a \,\big] + b_{\mathrm{tok}}, \tag{12}$$

with $W_{\mathrm{tok}} \in \mathbb{R}^{d \times (K+4)}$. Such a token is otherwise anonymous, so we add learned band-identity embeddings $e_p^\phi, e_a^A \in \mathbb{R}^d$ naming the two bands,

$$\tilde{\mathbf{t}}_{pa} = \mathbf{t}_{pa} + e_p^\phi + e_a^A, \tag{13}$$

which play the role of positional encodings: without them, attention over an unordered set of pairs cannot tell which frequencies a coupling value refers to.

3.4 Cross-Band Transformer

The $N = B_\phi B_A$ PAC tokens are processed by a standard multi-head transformer encoder (Vaswani et al., 2017) with a prepended CLS token $\mathbf{c} \in \mathbb{R}^d$,

$$[\mathbf{h}_{\mathrm{CLS}}, \mathbf{h}_1, \ldots, \mathbf{h}_N] = \mathrm{Transformer}\big([\mathbf{c}, \tilde{\mathbf{t}}_{11}, \ldots, \tilde{\mathbf{t}}_{B_\phi B_A}]\big), \tag{14}$$

$$\hat{y} = W_{\mathrm{cls}}\, \mathrm{LayerNorm}(\mathbf{h}_{\mathrm{CLS}}) + b_{\mathrm{cls}}. \tag{15}$$

Self-attention over 12 tokens costs little, but it lets the model condition one coupling pair on another rather than scoring each in isolation, which matters because pairs sharing a band are not statistically independent. Writing $\alpha_{\mathrm{CLS},pa}^{(\ell,h)}$ for the weight the CLS query places on token (p, a) in layer ℓ and head h, we define the aggregate reliance on that pair, which we read off for interpretation, as

$$\bar{\alpha}_{pa} = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \alpha_{\mathrm{CLS},pa}^{(\ell,h)}, \tag{16}$$

3.5 Waveform Stream and Channel Fusion

Twelve PAC tokens per channel is a deliberately compact description that necessarily discards waveform detail no coupling statistic retains. We therefore define an optional parallel waveform branch, used where coupling alone is not sufficient and ablated in §4. Given patch lengths $\{L_g\}_{g=1}^{G}$, channel c is divided into non-overlapping patches at each granularity and every patch is linearly embedded,

$$u_{g,i}^c = W_g\, x^c\big[(i-1)L_g + 1 : iL_g\big] + b_g, \quad i = 1, \ldots, \lfloor T/L_g \rfloor, \tag{17}$$

Algorithm 1 MOSAIC forward and update (PAC pathway)

Input: batch $x \in \mathbb{R}^{M\times C\times T}$ (M: batch size), sampling rate f_s, labels y; weights λ_o, λ_m
Output: logits $\hat{y}$ and updated parameters

Initialize ordered centers through positive gaps and bandwidths through softplus at the clinical defaults
$x \leftarrow \mathrm{reshape}(x, (MC, T))$ channels are independent
for each mini-batch **do**
  $x_b \leftarrow \mathcal{F}^{-1}\big(G_b(f; f_b, \sigma_b)\cdot \mathcal{F}[x]\big), b = 1..B$ (Eq. 5)
  $A_b, \phi_b \leftarrow |z_b|, \angle z_b$ (Eqs. 7–9)
  $\mathrm{mvl}_{pa}, \widetilde{\mathrm{PBA}}_{pa}$ for all 12 pairs (Eqs. 1, 11)
  $\tilde{\mathbf{t}}_{pa} \leftarrow W_{\mathrm{tok}}[\,\widetilde{\mathrm{PBA}}_{pa} \| \mathrm{mvl}_{pa} \| s_a\,] + e^{\phi}_{p} + e^{A}_{a}$ (Eqs. 12, 13)
  $\mathbf{h}_{\mathrm{CLS}}, \bar{\alpha} \leftarrow \mathrm{Transformer}([\mathbf{c}, \tilde{\mathbf{t}}_{11}, \ldots])$ (Eq. 14)
  per-channel logits $\leftarrow W_{\mathrm{cls}}\,\mathrm{LayerNorm}(\mathbf{h}_{\mathrm{CLS}})$ (Eq. 15)
  $\hat{y} \leftarrow$ mean of the per-channel logits over the C channels
  Backpropagate $\mathcal{L}$ into every parameter *including* $\{f_b, \sigma_b\}$ (Eq. 19)
end for
return $\hat{y}$

with $W_g \in \mathbb{R}^{d\times L_g}$. The tokens from all granularities pass through a transformer encoder with its own CLS token, yielding $h^c_{\mathrm{wave}} \in \mathbb{R}^d$; multiple patch lengths let the branch describe both fast morphology and slow envelope structure.

Let $h^c = [\,\mathbf{h}^c_{\mathrm{CLS}} \,\|\, h^c_{\mathrm{wave}}\,] \in \mathbb{R}^{2d}$ be the per-channel representation. With this branch enabled, the logit averaging of Algorithm 1 can be replaced by a learned gate over channels, so that an uninformative electrode is not weighted as heavily as an informative one,

$$g_c = \frac{\exp(w^\top h^c)}{\sum_{c'=1}^{C}\exp(w^\top h^{c'})}, \qquad h = \sum_{c=1}^{C} g_c\, h^c, \tag{18}$$

where $w \in \mathbb{R}^{2d}$ is the only added parameter and the fused h replaces $\mathbf{h}_{\mathrm{CLS}}$ in Eq. (15). Both the waveform branch and the gate are optional; Table 4 measures the contribution of each.

Training objective. The total loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_o\,\mathcal{L}_{\mathrm{orth}} + \lambda_m\,\mathcal{L}_{\mathrm{smooth}}, \tag{19}$$

with $\mathcal{L}_{\mathrm{cls}}$ the cross-entropy and

$$\mathcal{L}_{\mathrm{smooth}} = \sum_{p,a}\sum_{k=1}^{K}\big(\widetilde{\mathrm{PBA}}_{pa}[k+1] - \widetilde{\mathrm{PBA}}_{pa}[k]\big)^2 \tag{20}$$

where $\widetilde{\mathrm{PBA}}_{pa}[K+1] = \widetilde{\mathrm{PBA}}_{pa}[1]$. This circular penalty encodes the prior that amplitude varies smoothly with phase and suppresses histogram noise when a band carries little energy. Gradients from $\mathcal{L}_{\mathrm{cls}}$ propagate through the transformer, the tokens, the soft binning, and the Hilbert transform into $\{f_b, \sigma_b\}$, so the band definitions and the classifier are optimized together.

Per-channel PAC features are cheap once the filterbank is applied, so the front end adds little over the transformer it feeds. MOSAIC has roughly 0.8M parameters, comparable to the patch and inverted transformer baselines (about 0.6M) and below multi-granularity models (about 1.3M).

4 Experiments

4.1 Setup

Datasets. We evaluate on three public medical time series benchmarks (Table 1): the EEG datasets APAVA (Escudero et al., 2006) and TDBrain (Van Dijk et al., 2022), and the ECG dataset PTB (Bousseljot et al., 1995; Goldberger et al., 2000). All experiments are subject-independent: no

Dataset	Mod.	Subj.	Samples	Cls	Ch	Task
APAVA	EEG	23	5,967	2	16	Alzheimer's
TDBrain	EEG	50	6,240	2	33	Parkinson's
PTB	ECG	198	64,356	2	15	Myocard. infarct.

Table 1: Benchmark statistics (Wang et al., 2024). Segments are 1 s at 256 Hz (EEG) and 1.2 s at 250 Hz (ECG). TDBrain is the eyes-closed subset of 25 Parkinson's and 25 control subjects.

	APAVA				TDBrain				PTB			
Method	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Autoformer	$68.64_{1.82}$	$68.48_{2.10}$	$68.77_{2.27}$	$68.06_{1.94}$	$87.33_{3.79}$	$88.06_{3.56}$	$87.33_{3.79}$	$87.26_{3.84}$	$73.35_{2.10}$	$72.11_{2.89}$	$63.24_{3.17}$	$63.69_{3.84}$
Crossformer	$73.77_{1.95}$	$79.29_{4.36}$	$68.86_{1.70}$	$68.93_{1.85}$	$81.56_{2.19}$	$81.97_{2.25}$	$81.56_{2.19}$	$81.50_{2.20}$	$80.17_{3.79}$	$85.04_{1.83}$	$71.25_{6.29}$	$72.75_{7.19}$
FEDformer	$74.94_{2.15}$	$74.59_{1.50}$	$73.56_{3.55}$	$73.51_{3.39}$	$78.13_{1.98}$	$78.52_{1.91}$	$78.13_{1.98}$	$78.04_{2.01}$	$76.05_{2.54}$	$77.58_{3.61}$	$66.10_{3.55}$	$67.14_{4.37}$
Informer	$73.11_{4.40}$	$75.17_{6.06}$	$69.17_{4.56}$	$69.47_{5.06}$	$89.02_{2.50}$	$89.43_{2.14}$	$89.02_{2.50}$	$88.98_{2.54}$	$78.69_{1.68}$	$82.87_{1.02}$	$69.19_{2.90}$	$70.84_{3.47}$
iTransformer	$74.55_{1.66}$	$74.77_{2.10}$	$71.76_{1.72}$	$72.30_{1.79}$	$74.67_{1.06}$	$74.71_{1.06}$	$74.67_{1.06}$	$74.65_{1.06}$	$83.89_{0.71}$	**$88.25_{1.18}$**	$76.39_{1.01}$	$79.06_{1.06}$
MTST	$71.14_{1.59}$	$79.30_{0.97}$	$65.27_{2.28}$	$64.01_{3.16}$	$76.96_{3.76}$	$77.24_{3.59}$	$76.96_{3.76}$	$76.88_{3.83}$	$76.59_{1.90}$	$79.88_{1.90}$	$66.31_{2.95}$	$67.38_{3.71}$
Nonformer	$71.89_{3.81}$	$71.80_{4.58}$	$69.44_{3.56}$	$69.74_{3.84}$	$87.88_{2.48}$	$88.86_{1.84}$	$87.88_{2.48}$	$87.78_{2.56}$	$78.66_{0.49}$	$82.77_{0.86}$	$69.12_{0.87}$	$70.90_{1.00}$
PatchTST	$67.03_{1.65}$	$78.76_{1.28}$	$59.91_{2.02}$	$55.97_{3.10}$	$79.25_{3.79}$	$79.60_{4.09}$	$79.25_{3.79}$	$79.20_{3.77}$	$74.74_{1.62}$	$76.94_{1.51}$	$63.89_{2.71}$	$64.36_{3.38}$
Reformer	$78.70_{2.00}$	**$82.50_{3.95}$**	$75.00_{1.61}$	$75.93_{1.82}$	$87.92_{2.01}$	$88.64_{1.40}$	$87.92_{2.01}$	$87.85_{2.08}$	$77.96_{2.13}$	$81.72_{1.61}$	$68.20_{3.35}$	$69.65_{3.88}$
Transformer	$76.30_{4.72}$	$77.64_{5.95}$	$73.09_{5.01}$	$73.75_{5.38}$	$87.17_{1.67}$	$87.99_{1.68}$	$87.17_{1.67}$	$87.10_{1.68}$	$77.37_{1.02}$	$81.84_{0.66}$	$67.14_{1.80}$	$68.47_{2.19}$
Medformer	$78.74_{0.64}$	$81.11_{0.84}$	$75.40_{0.66}$	$76.31_{0.71}$	$89.62_{0.81}$	$89.68_{0.78}$	$89.62_{0.81}$	$89.62_{0.81}$	$83.50_{2.01}$	$85.19_{0.94}$	$77.11_{3.39}$	$79.18_{3.31}$
MOSAIC	**$80.98_{1.81}$**	$80.85_{1.90}$	**$79.61_{2.15}$**	**$79.97_{2.03}$**	**$91.60_{2.41}$**	**$91.91_{2.38}$**	**$91.60_{2.41}$**	**$91.59_{2.42}$**	**$85.11_{1.53}$**	$84.36_{1.97}$	**$81.65_{2.73}$**	**$82.56_{2.00}$**

Table 2: Subject-independent results (%, five-seed mean with standard deviation as subscript). Baselines are published values (Wang et al., 2024); MOSAIC is from our experiments. **Bold blue**: best per column; MOSAIC takes 10 of 12.

subject contributes samples to more than one split. We adopt the established split definitions (Wang et al., 2024): the published subject identifiers for APAVA and TDBrain, and a 60/20/20 subject-level ratio for PTB. The split is fixed across the five training seeds, so reported standard deviations reflect training stochasticity rather than which subjects are held out.

Baselines. We compare against eleven transformer architectures previously evaluated on these benchmarks under the same protocol, quoting their published values (Wang et al., 2024). They span decomposition and frequency-domain forecasters (Wu et al., 2021; Zhou et al., 2022), efficient-attention models (Zhou et al., 2021; Kitaev et al., 2020), multi-resolution and non-stationary variants (Zhang et al., 2024; Liu et al., 2022), the patch and inverted tokenizations discussed above, and Medformer. We also compare against fixed-feature PAC classifiers that feed fixed-band mean vector length, modulation index, or phase-locking value features to a two-layer MLP. These are the three estimators most commonly compared in the coupling literature (Hülsemann et al., 2019), and they separate the contribution of a differentiable pipeline from that of using coupling at all.

Metrics. We report accuracy and macro-averaged precision, recall, and F1. Because several benchmarks are class-imbalanced, accuracy alone rewards predicting the majority class, so macro-F1 is our primary metric: it weights every class equally and penalizes buying precision at the cost of recall. Precision and recall are reported alongside it so the trade-off stays visible.

Implementation. MOSAIC uses $B = 6$ bands, $K = 18$ phase bins, token dimension $d = 128$, and a 4-layer, 4-head transformer with FFN width 512. We train with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 64) for up to 100 epochs with early stopping on validation F1 (patience 10), and set $\lambda_o = 0.1$ and $\lambda_m = 0.05$. These values were chosen by light manual tuning on validation F1 rather than an exhaustive search. Each configuration is run with seeds 0 through 4, which control initialization and batch ordering; reported values are means over these five runs. All experiments run on a single NVIDIA RTX PRO 5000 Blackwell GPU (48 GB) under Linux, using Python 3.10, PyTorch 2.1, and CUDA 12.1.

4.2 Results

Table 2 reports all four metrics, and Figure 2(a) places MOSAIC within the spread of published results. The MOSAIC row is the full backbone of Table 4, which combines the PAC pathway with the waveform stream and gated channel fusion (§3.5). MOSAIC obtains the best accuracy, recall,

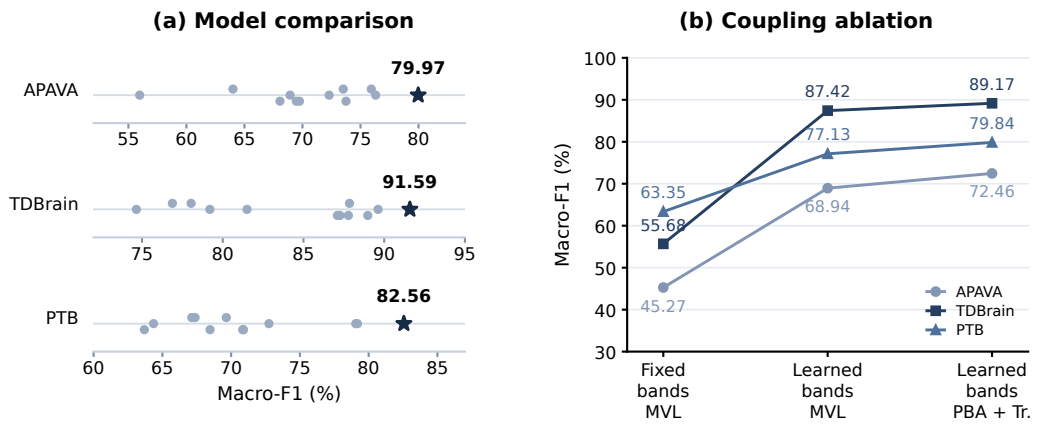

Figure 2: **(a)** Macro-F1 of the twelve methods in Table 2 (circles: baselines; stars: MOSAIC, annotated); horizontal scales differ by dataset. **(b)** Coupling ablation of Table 3 (Tr.: cross-band Transformer). Five-seed means; uncertainty intervals not shown.

Coupling representation	APAVA	TDBrain	PTB
PLV, fixed bands	43.82	55.37	40.16
MI, fixed bands	44.63	51.24	64.71
MVL, fixed bands	45.27	55.68	63.35
MVL, learned bands	68.94	87.42	77.13
PBA, learned bands	**72.46**	**89.17**	**79.84**

Table 3: Macro-F1 (%, five-seed mean) by coupling representation, without waveform features. Rows 1–3: fixed-band estimators with a two-layer MLP; row 4 learns the bands; row 5 adds PBA and the cross-band transformer (the PAC pathway of Algorithm 1). **Bold blue**: best per dataset.

and macro-F1 on every benchmark and the best precision on TDBrain. It trails only in precision on APAVA (Reformer) and PTB (iTransformer), where that precision comes with markedly lower recall (75.00 and 76.39 versus MOSAIC's 79.61 and 81.65), a less balanced operating point that macro-F1 reflects.

The margin over the strongest prior architecture is 3.7 macro-F1 on APAVA, 2.0 on TDBrain, and 3.4 on PTB. The margin is largest on APAVA, the smallest dataset at 23 subjects, although the ablation in Table 4 indicates that on APAVA the waveform stream contributes more than the PAC stream, whereas on TDBrain and PTB the PAC stream contributes more.

4.3 Does Learning the Coupling Help?

To isolate the effect of a learnable representation, Table 3 varies how coupling is represented, from classical fixed-band statistics through learned bands to the full PBA representation, using the PAC pathway alone. Figure 2(b) shows both transitions. Learning the bands raises MVL macro-F1 from 45.27 to 68.94 on APAVA, 55.68 to 87.42 on TDBrain, and 63.35 to 77.13 on PTB, gains of 23.67, 31.74, and 13.78 points. The PBA-and-Transformer row adds a further 3.52, 1.75, and 2.71 points, although this comparison does not separate the effects of the representation and the classifier.

MVL is the strongest fixed-band estimator on APAVA and TDBrain, and MI on PTB. Relative to the best fixed-band result on each dataset, the final PBA row improves macro-F1 by 27.19, 33.49, and 15.13 points, approximately $1.60\times$, $1.60\times$, and $1.23\times$ the fixed-band scores.

4.4 Component Ablation

Table 4 reports component removals and two controls, which replace PAC with amplitude-only features or disrupt phase–amplitude alignment, for a backbone carrying both the PAC pathway and the multi-granularity waveform pathway of §3.5. The full backbone has the highest reported mean

Variant	APAVA	TDBrain	PTB
Full backbone	**79.97**	**91.59**	**82.56**
− PAC stream	75.23 (−4.74)	85.67 (−5.92)	71.47 (−11.09)
Replace PAC with amplitude-only features	76.14 (−3.83)	87.03 (−4.56)	75.42 (−7.14)
Disrupt phase–amplitude alignment	74.86 (−5.11)	86.12 (−5.47)	73.95 (−8.61)
− waveform stream	71.48 (−8.49)	86.63 (−4.96)	80.08 (−2.48)
− gated pooling	77.78 (−2.19)	87.32 (−4.27)	81.10 (−1.46)

Table 4: Component ablation and PAC controls (macro-F1 %, five-seed mean; Δ vs. full backbone in parentheses). **Bold blue**: best per dataset; no variant improves the reported mean.

on every dataset. Because this ablation and Table 3 use different configurations, the PAC-only scores in Table 3 and the "− waveform stream" row here are not directly comparable.

Removing the PAC stream reduces macro-F1 by 4.74, 5.92, and 11.09 points on APAVA, TDBrain, and PTB, and removing the waveform stream by 8.49, 4.96, and 2.48; the waveform stream thus matters more on APAVA and the PAC stream on TDBrain and PTB. Removing gated pooling costs 2.19, 4.27, and 1.46 points.

Replacing PAC with amplitude-only features lowers macro-F1 by 3.83, 4.56, and 7.14 points, and disrupting phase–amplitude alignment by 5.11, 5.47, and 8.61. These controls suggest that phase-dependent information contributes beyond amplitude alone, although the reported mean differences do not establish statistical significance.

5 Conclusion

We presented MOSAIC, which makes phase-amplitude coupling estimation end-to-end differentiable and trains it jointly with its classifier. A learnable Gaussian filterbank, an FFT-based Hilbert transform, and a soft phase-binned amplitude histogram replace the fixed-band offline pipeline that has kept classical PAC measures from serving as trainable features, and a cross-band transformer over PAC tokens learns which coupling pairs matter. On three subject-independent benchmarks, the learned representation improves on fixed-band PAC features by $1.2\times$ to $1.6\times$ macro-F1, and MOSAIC attains the highest accuracy, recall, and macro-F1 among the twelve methods in the reported comparison.

The results also delimit what coupling contributes. In our ablations, removing the PAC stream costs most on PTB and TDBrain, whereas on APAVA, the smallest dataset, the waveform stream contributes more; coupling therefore complements rather than replaces waveform modeling. Three extensions follow. An adaptive smoothing scale would generalize the fixed one used in soft binning. The envelope statistics in each token leave band power and coupling entangled, and separating them may sharpen the learned attention. Finally, pretraining the front end on large unlabeled recordings could transfer band definitions to the small-subject settings where they are hardest to estimate.

Ethics Statement

This study is a secondary analysis of three open, publicly available EEG and ECG datasets (APAVA, TDBrain, and PTB), used under their providers' terms of use. We did not record, collect, or annotate any new data from human participants for this research, and we made no attempt to re-identify any individual. MOSAIC is a research prototype evaluated on retrospective benchmarks; it has not been clinically validated and should not be used for diagnosis without further prospective evaluation.

Reproducibility Statement

We have provided a complete description of the model, its parameterization, initialization, and training objective, together with pseudocode, dataset statistics, subject-independent split definitions, evaluation metrics, hyperparameters, random seeds, and the software and hardware environment. All three datasets are open and publicly available. Upon acceptance, we will publicly release our full code repository, including preprocessing scripts, split definitions, configuration files, and the commands needed to reproduce every reported table and figure.

AI Use Statement

Generative AI tools were used to assist with manuscript preparation, including editing and formatting. The authors reviewed all outputs and take full responsibility for the final manuscript and its claims.

References

Juhan Aru, Jaan Aru, Viola Priesemann, Michael Wibral, Luiz Lana, Gordon Pipa, Wolf Singer, and Raul Vicente. Untangling cross-frequency coupling in neuroscience. *Current opinion in neurobiology*, 31:51–61, 2015.

Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 1995.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304 (5679):1926–1929, 2004.

Ryan T Canolty, Erik Edwards, Sarang S Dalal, Maryam Soltani, Srikantan S Nagarajan, Heidi E Kirsch, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *science*, 313(5793):1626–1628, 2006.

J Escudero, Daniel Abásolo, Roberto Hornero, Pedro Espino, and Miguel López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091–1106, 2006.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Bosong Huang, Ming Jin, Yuxuan Liang, Johan Barthelemy, Debo Cheng, Qingsong Wen, Chenghao Liu, and Shirui Pan. Shapex: Shapelet-driven post hoc explanations for time series classification models. *Advances in Neural Information Processing Systems*, 38:26639–26677, 2026.

Mareike J Hülsemann, Ewald Naumann, and Björn Rasch. Quantification of phase-amplitude coupling in neuronal oscillations: comparison of phase-locking value, mean vector length, modulation index, and generalized-linear-modeling-cross-frequency-coupling. *Frontiers in neuroscience*, 13:573, 2019.

Ole Jensen and Laura L Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in cognitive sciences*, 11(7):267–269, 2007.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations (ICLR)*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2023. Accepted at International Conference on Learning Representations (ICLR 2023).

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pp. 1021–1028. IEEE, 2018.

Hugo Soulat, Emily P Stephen, Amanda M Beck, and Patrick L Purdon. State space methods for phase amplitude coupling analysis. *Scientific Reports*, 12(1):15940, 2022.

Adriano BL Tort, Robert Komorowski, Howard Eichenbaum, and Nancy Kopell. Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of neurophysiology*, 104(2):1195–1210, 2010.

Viet-Hung Tran, Ngoc Phu Doan, Zichi Zhang, Tuan Pham, Phi Hung Nguyen, Xuan Nguyen, Hans Vandierendonck, Ira Assent, and Thai Son Mai. Mix: a multi-view time-frequency interactive explanation framework for time series classification. *Advances in Neural Information Processing Systems*, 38:49181–49226, 2026.

Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific data*, 9(1):333, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Manydg: Many-domain generalization for healthcare applications. *arXiv preprint arXiv:2301.08834*, 2023. Published as a conference paper at the International Conference on Learning Representations (ICLR 2023).

Weiyang Zhang, Xinyang Chen, Xiucheng Li, Kehai Chen, Weili Guan, and Liqiang Nie. Unified transferability metrics for time series foundation models. *Advances in Neural Information Processing Systems*, 38:41155–41178, 2026.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *International conference on artificial intelligence and statistics*, pp. 4222–4230. PMLR, 2024.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Xiao-Cong Zhong, Qisong Wang, Dan Liu, Zhihuang Chen, Jing-Xiao Liao, Jinwei Sun, Yudong Zhang, and Feng-Lei Fan. Eeg-dg: A multi-source domain generalization framework for motor imagery eeg classification. *IEEE Journal of Biomedical and Health Informatics*, 29(4):2484–2495, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.