STEP: SCIENTIFIC TIME-SERIES ENCODER PRETRAINING VIA CROSS-DOMAIN DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific time series are central to scientific AI but are typically sparse, highly heterogeneous, and limited in scale, making unified representation learning particularly challenging. Meanwhile, foundation models pretrained on relevant time series domains such as audio, general time series, and brain signals contain rich knowledge, but their applicability to scientific signals remains underexplored. In this paper, we investigate the transferability and complementarity of foundation models from relevant time series domains, and study how to effectively leverage them to build a unified encoder for scientific time series. We first systematically evaluate relevant foundation models, showing the effectiveness of knowledge transfer to scientific tasks and their complementary strengths. Based on this observation, we propose **STEP**, a **S**cientific **T**ime Series **E**ncoder **P**retraining framework via cross domain distillation. STEP introduces adaptive patching to handle extreme-length sequences and a statistics compensation scheme to accommodate diverse numerical scales. It further leverages cross-domain distillation to integrate knowledge from multiple foundation models into a unified encoder. By combining complementary representations across different domains, STEP learns general-purpose and transferable features tailored for scientific signals. Experiments on seven scientific time series tasks demonstrate that STEP provides both an effective structure and an effective pretraining paradigm, taking a STEP toward scientific time series representation learning[1].

1 INTRODUCTION

Artificial intelligence for science (AI4Sci) has become one of the most active and impactful frontiers of modern AI research (Jumper et al., 2021; Wang et al., 2023; Bai et al., 2025), aiming to accelerate scientific discovery by learning from complex observational and experimental data. Among various scientific data modalities, time series play a central role, arising in domains such as astronomy, geophysics, neuroscience, and bioacoustics. These signals encode rich temporal dynamics and physical processes, making them indispensable for understanding, modelling, and predicting scientific phenomena. Despite their importance, how to build unified and general-purpose representations for scientific time series remains largely underexplored.

Recent studies reveal that time series should be treated as a dedicated data modality (Wu et al., 2026): serializing long temporal signals as text is inefficient and difficult to scale, while rasterizing them as images sacrifices numerical precision and struggles with high-dimensional signals. Consequently, existing general-purpose multimodal LLMs, which typically rely on vision encoders and language models, are insufficient for capturing the richness of scientific time series. This motivates the design of specialized time series encoders.

Conventional time series encoders face additional challenges when applied to scientific data, primarily due to the extreme heterogeneity of scientific time series. Conventional time series encoders typically build on general time series such as sales, electricity load, or weather data (Shi et al., 2025; Woo et al., 2024; Zhou et al., 2021; Nie et al., 2023). These signals are typically sampled at coarse resolutions (*e.g.*, minute-level or daily) and exhibit noticeable trends and seasonality. However, scientific temporal signals from different scientific domains differ drastically in their semantic content, sequence length,

[1]Code and model will be made publicly available upon acceptance.

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.

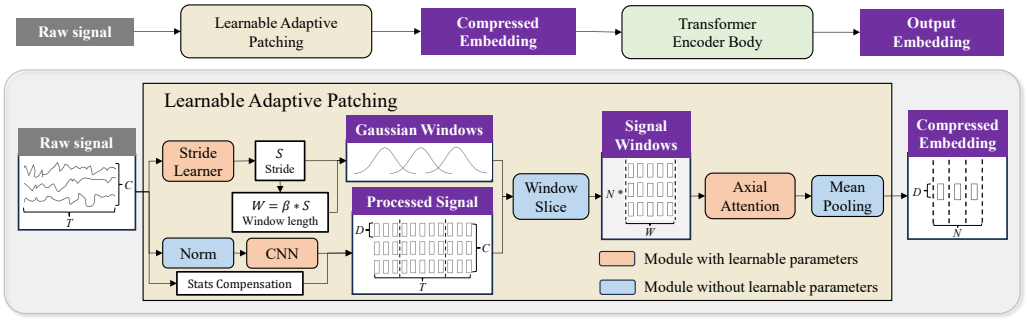


Figure 1: Architecture of the STEP Encoder. It includes an adaptive module that dynamically compresses the input signal and a statistics compensation scheme that recovers task-relevant statistical information.

sampling frequency, numerical scale, number of channels, and noise characteristics. For example, scientific time series may range from short transient signals to ultra-long recordings, span sampling rates from micro-Hz to MHz, and vary from univariate measurements to high-dimensional multi-channel observations. In addition, foundation models for general time series are primarily trained to capture trends and seasonality for forecasting tasks. They focus on predicting overall temporal patterns rather than learning detailed representations of the underlying signal, limiting their performance on scientific tasks that require fine-grained, domain-specific information (Wu et al., 2026). Furthermore, scientific data are typically sparse: while the number of signal categories is large, each category often has limited labelled samples, making large-scale pretraining particularly difficult. Further discussion of scientific time-series characteristics and their implications for representation learning is provided in Appendix A.

Together, the extreme heterogeneity and data sparsity of scientific time series pose fundamental challenges for building encoders for scientific time series. In this work, we take a STEP towards investigating these challenges by studying encoder architectures capable of handling highly diverse signals. In addition, given the limited availability of scientific data, we investigate whether foundation models pretrained on related time series domains, such as audio, general time series, and neural signals, can provide transferable knowledge to improve performance on scientific tasks. We then propose **STEP**, a **s**cientific **t**ime-series **e**ncoder **p**retraining framework via cross-domain distillation. The STEP encoder introduces a novel adaptive patching mechanism to handle highly diverse sequence lengths and a statistics compensation scheme to accommodate varying numerical scales. It then leverages a cross-domain distillation framework to learn from multiple foundation models pretrained on different domains, effectively integrating their complementary knowledge. Experiments on seven scientific time series tasks demonstrate that STEP establishes an effective pretraining paradigm, achieving strong performance across diverse scientific domains.

The main contributions are summarized as follows: (i) We systematically analyze the transferability of foundation models pretrained on related time series domains (*i.e.*, audio, general time series, and neural signals), revealing the complementary strengths and limitations of different domains for scientific time series tasks. (ii) We design the STEP encoder with adaptive patching and a statistics compensation scheme to handle highly heterogeneous scientific signals. (iii) We adopt a cross-domain distillation framework for STEP that integrates knowledge from multiple foundation models, exploiting the complementary strengths of different domains.

2 Related Work

2.1 Time Series Encoder Structures

Recent advances in time series modelling commonly leverage Transformer-based architectures (Vaswani et al., 2017). However, the quadratic complexity of self-attention poses significant challenges for long-sequence processing. Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and FEDformer (Zhou et al., 2022) reduce this cost through sparse attention, decomposition, or

frequency-domain operations. These methods were developed primarily for forecasting and retain a largely point-wise view of the input, which can be inefficient for extremely long scientific recordings and may not provide the local abstractions required by discriminative tasks.

Patch-based tokenization offers a complementary route: a patch aggregates neighboring observations into one token, simultaneously shortening the attention sequence and exposing local temporal motifs. PatchTST (Nie et al., 2023) applies a fixed patch length and stride channel-independently, while Moirai (Woo et al., 2024) uses multiple patch sizes to accommodate different frequencies. TimeOmni (Wu et al., 2026) further routes each input to a patch size chosen from a predefined set. Although effective, these designs make temporal resolution a fixed hyperparameter or a discrete choice. A patch size suitable for a 24-point light curve may be inappropriate for a 2.4×10^5-point acoustic waveform; moreover, the informative duration of an event need not scale linearly with the sampling rate. STEP therefore treats patch formation as a differentiable, input-adaptive operation, allowing the effective receptive field and compression ratio to be learned from the downstream objective. Further discussion is provided in Appendix B.1.

2.2 Related Time Series Foundation Models

This section summarises foundation models of relevant time series domains such as audio, general time series, and neural signals. Extended discussion is provided in Appendix B.2.

Audio is a broad and fundamental class of time series signals, and its foundation models have been extensively studied across large-scale acoustic data. These models typically span speech (Hsu et al., 2021; Radford et al., 2023), general audio (*e.g.*, sound events) (Huang et al., 2022; Chen et al., 2023c), and music (LI et al., 2024; Zhu et al., 2025). Recent work has further explored multi-subdomain audio foundation models that unify speech and general audio within a single representation framework (Yang et al., 2025). Acoustic signals are typically relatively high-frequency (*e.g.*, 16kHz or 48kHz) and contain both periodic and aperiodic components. These models are usually pretrained on audio data at the order of hundreds of thousands of hours.

General time series represent a widely studied class of temporal signals, such as sales, electricity load, and weather data, typically sampled at coarse resolutions (*e.g.*, daily) and characterized by strong trends and seasonality. Foundation models for general time series are primarily focused on forecasting tasks, such as Moirai (Woo et al., 2024), Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), and TimeMoE (Shi et al., 2025). They are commonly pretrained on datasets containing tens to hundreds of billions of time points (Woo et al., 2024; Ansari et al., 2024; Shi et al., 2025).

Another important class of time series is neural signals, such as electroencephalography (EEG) and magnetoencephalography (MEG), which are high-dimensional, multivariate recordings capturing brain activity at fine temporal and frequency resolutions, often with substantial noise. Representative neural foundation models include LaBraM (Jiang et al., 2024) (pretrained on 2.5k hours EEG), CBraMod (Wang et al., 2025) (27k hours EEG), and BrainOmni (Xiao et al., 2025) (2k hours EEG + 656 hours MEG). Those models focus on capturing spatio-temporal and frequency-specific representations.

Although pretrained models in these domains are well-established, scientific time series are far more diverse in length, frequency, and modality, motivating our study of whether pretrained models from these related domains can be effectively leveraged for scientific time series tasks.

3 Methodology

3.1 Architecture of the STEP Encoder

To handle extremely diverse sequence lengths and heterogeneous channel dimensions, we design an encoder that adaptively compresses inputs into compact representations without handcrafted heuristics. The compression ratio is learned end-to-end, allowing the model to automatically determine an appropriate downsampling strategy for each sequence, while projecting inputs with different channel counts into a unified embedding space.

Our framework adopts a three-stage training pipeline and centres on a Learnable Adaptive Patching module for efficient long-sequence modelling. The architecture consists of three components:

a learnable adaptive patching block, a Transformer encoder backbone, and projection layers for distillation and downstream tasks. The overall structure is illustrated in Figure 1.

Learnable Adaptive Patching. To enable adaptive handling of highly diverse sequence lengths and channel dimensions, we introduce a learnable adaptive patching block that replaces fixed patching with a differentiable, learnable windowing mechanism.

Given an input time series $\mathbf{X} \in \mathbb{R}^{T\times C}$, where T denotes sequence length and C denotes number of channels, the module first applies a channel-independent 1D convolution to obtain $\mathbf{X}' \in \mathbb{R}^{T\times C\times D}$, where D denotes hidden dimension. A set of Gaussian windows $\mathbf{G} \in \mathbb{R}^{N\times W}$ is then used to segment $\mathbf{X}'$ into overlapped patches $\mathbf{X}_{\text{patch}} \in \mathbb{R}^{N\times W\times C\times D}$ via differentiable weighted aggregation, where N is the number of windows parameterized by learnable stride S and window length W. Axial attention (Ho et al., 2019) is applied sequentially along the temporal and channel dimensions within each patch to capture local spatio-temporal structure, followed by mean pooling to yield a downsampled embedding $\mathbf{E} \in \mathbb{R}^{N\times D}$.

Specifically, the stride learner directly maps the input sequence to the stride parameter via a multilayer perceptron (MLP):

$$S = 1 + \exp(\text{MLP}\,(\log(T)))\,, \tag{1}$$

with a logarithmic transformation applied to ensure numerical stability, flexible adaptation to diverse signal durations, and strictly positive strides for valid temporal downsampling. To prevent unbounded window length growth, motivated by the model's incentive to maximize per-step information, we fix the window length to twice the stride (*i.e.*, $\beta = 2$). To further ensure the stability of stride learning, we adopt a stride penalty

$$\mathcal{L}_{\text{stride}} = \text{huber}(\max(0,\, \log(S) - \log(S_{\max})))\,, \tag{2}$$

and a length penalty

$$\mathcal{L}_{\text{length}} = \text{huber}(\max(0,\, \log(N') - \log(N_{\max}))) + \text{huber}(\max(0,\, \log(N_{\min}) - \log(N')))\,, \tag{3}$$

where N' is a differentiable estimate of the embedding length N. In our experiments, $S_{\max}$ is set to 400, $N_{\max}$ to 200, and $N_{\min}$ to 5.

Transformer Encoder Body: The architecture of the Transformer encoder body follows the design of the Whisper encoder (Radford et al., 2023). Specifically, the input time series subsampled features are first processed by two 1D convolutional layers. The resulting representations are then augmented with learnable positional embeddings and fed into 17 Transformer encoder blocks of hidden dimension 768.

Statistics Compensation Scheme: The numerical scales of the input model often vary by orders of magnitude. Global normalization would induce unstable training dynamics, whereas dataset-specific normalization requires prior knowledge of data source at inference time. Moreover, statistics such as mean and variance can encode useful task-relevant information. Therefore, we adopt a per-sample, per-channel normalization strategy: each input channel is independently standardized to zero mean and unit variance. The computed channel-wise statistics are log-scaled and concatenated as auxiliary features to the output of the initial convolutional layers, as shown in Figure 1. This design preserves discriminative signal-level statistics while ensuring consistent input scale across heterogenous signals.

3.2 Cross-Domain Knowledge Distillation

To transfer representations from related domains, we pretrain a shared STEP encoder using multiple frozen teachers, as illustrated in Figure 2. Each teacher supervises data from its corresponding domain, with the same input fed to that teacher and STEP.

A teacher-specific projection head maps the student representation into the corresponding teacher's feature space. During distillation, the stride is constrained so that the student's output sequence length matches that of the teacher. Joint supervision from these teachers encourages the shared encoder to integrate complementary representations.

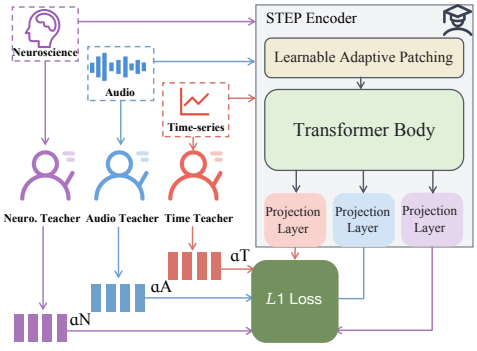


Figure 2: Multi-teacher distillation in STEP.

Table 1: Overview of the downstream scientific datasets. Length is reported in number of samples, and channels (# Ch.) denotes the number of input channels. A range indicates variable-length inputs.

Dataset	Discipline	Length	Freq. (Hz)	# Ch.	Task
GWOSC (Wu et al., 2026)	Astronomy	6.4×10^4	1.6×10^4	1	Gravitational-wave detection
LEAVES (Fei et al., 2024)	Astronomy	$24 \sim 1.3 \times 10^3$	1.2×10^{-5}	1	Light-curve classification
STEAD (Mousavi et al., 2019)	Earth science	6.0×10^3	1.0×10^2	3	Earthquake detection
MarmAudio (Lamothe et al., 2025)	Bioacoustics	$6.2 \times 10^4 \sim 2.4 \times 10^5$	9.6×10^4	1	Vocalisation classification
SleepEDF (Kemp et al., 2000)	Neuroscience	3.0×10^3	1.0×10^2	1	Sleep staging
TUAB (López de Diego, 2017)	Neuroscience	2.0×10^3	2.0×10^2	16	EEG abnormality classification
RadarComm (Jagannath & Jagannath, 2021)	Radar	1.28×10^2	1.0×10^7	2	Signal-type classification

For an input $\mathbf{X}_k$ from teacher k's domain, let $\mathbf{H}_{\mathrm{STEP}}(\mathbf{X}_k)$ and $\mathbf{H}_k(\mathbf{X}_k)$ denote the aligned student and teacher features, respectively. We minimize the weighted feature-matching objective

$$\mathcal{L}_{\mathrm{KD}} = \sum_{k=1}^{K} \lambda_k \left\| P_k(\mathbf{H}_{\mathrm{STEP}}(\mathbf{X}_k)) - \mathbf{H}_k(\mathbf{X}_k) \right\|_1 , \tag{4}$$

where K is the number of teachers, P_k is a learnable projection head, and λ_k weights each teacher's contribution. At each optimization step, supervision from all domains jointly updates the shared encoder and projection heads while the teachers remain frozen. Teacher–dataset pairings, loss weights, and the batch and optimization protocols are detailed in Appendix D.4.

4 Experimental Setup

4.1 Scientific Tasks and Datasets

We evaluate on seven scientific time-series tasks across five disciplines. Astronomy is represented by gravitational-wave detection (GWOSC, Wu et al. (2026)) and variable-star classification (LEAVES, Fei et al. (2024)), Earth science by earthquake detection (STEAD, Mousavi et al. (2019)), and bioacoustics by marmoset vocalisation classification (MarmAudio, Lamothe et al. (2025)). The remaining tasks are sleep staging (SleepEDF, Kemp et al. (2000)) and EEG abnormality classification (TUAB, López de Diego (2017)) in neuroscience, and communication signal-type classification (RadarComm, Jagannath & Jagannath (2021)) in radar. Together, these datasets span event detection and multi-class recognition, with sequence lengths ranging from 24 to 2.4×10^5 samples, sampling frequencies from 1.2×10^{-5} to 1.0×10^7 Hz, and channel counts from 1 to 16, with further details provided in Table 1. We cast all seven tasks as sequence classification and report macro-F1. Detailed descriptions of each dataset and task, are provided in Appendix C.

4.2 Baselines and Foundation Models Investigated

The proposed STEP encoder structure was compared to the following SOTA models: PatchTST (Nie et al., 2023), Informer (Zhou et al., 2021), Moirai (Woo et al., 2024), TimeMoE (Shi et al., 2025). We investigate the transferability of the following foundation models to scientific time series:

Audio foundation models: (i) Whisper[2] (Radford et al., 2023) pretrained on 680k hours of multilingual labelled speech data through weak supervision; (ii) SPEAR[3] (Yang et al., 2025) pretrained on a heterogeneous mixture of speech and general audio data.

General time series foundation models: (i) TimeMoE[4] (Shi et al., 2025), a decoder-based time series forecasting model pretrained on over 300 billion time points spans over 9 domains; (ii) Moirai[5] (Woo et al., 2024), an encoder-based time series forecasting model pretrained on 27B data spanning 9 domains.

[2] https://huggingface.co/openai/whisper-large-v3
[3] https://huggingface.co/marcoyang/spear-xlarge-speech-audio
[4] https://huggingface.co/Maple728/Time-MoE-200M
[5] https://huggingface.co/Salesforce/moirai-1.1-R-large

Table 2: Macro-F1 (%) of foundation models on downstream scientific time-series tasks. PT indicates whether pretrained weights are loaded; FT indicates whether the encoder is finetuned. The best and second-best results in each column are bolded and underlined, respectively.

Model	PT	FT	GWOSC	LEAVES	STEAD	MarmAudio	SleepEDF	TUAB	RadarComm
Whisper	√	×	67.06	34.68	97.90	69.42	49.60	61.03	8.61
	√	√	<u>98.92</u>	50.87	**99.95**	**98.78**	71.82	75.41	60.40
	×	√	97.07	53.56	99.71	<u>97.44</u>	63.20	72.42	2.78
SPEAR	√	×	91.21	41.87	92.33	77.84	64.37	73.41	55.12
	√	√	**99.41**	50.63	<u>99.85</u>	97.06	64.80	77.39	58.42
	×	√	96.97	61.63	99.66	93.42	**75.53**	75.87	59.69
TimeMoE	√	×	59.87	<u>80.53</u>	78.65	41.34	65.65	73.92	57.71
	√	√	97.45	**81.26**	97.61	60.79	<u>72.53</u>	**81.58**	**63.44**
	×	√	71.26	67.67	92.38	59.63	62.08	67.99	53.70
Moirai	√	×	56.25	64.76	94.53	41.90	48.43	69.79	42.79
	√	√	98.34	73.24	99.61	96.31	66.03	80.49	60.32
	×	√	86.52	52.34	99.56	86.73	63.55	75.90	<u>61.35</u>
CBraMod	√	×	46.57	14.89	87.96	16.65	63.56	74.30	13.55
	√	√	83.70	64.30	99.32	78.10	72.45	<u>80.81</u>	52.30
	×	√	52.74	50.88	94.72	29.87	68.50	72.25	51.09

Neural foundation models: (i) CBraMod[6] (Wang et al., 2025) an EEG foundation model pretrained on the Temple University Hospital EEG Corpus (TUEG) (Hamid et al., 2020); (ii) BrainOmni[7] (Xiao et al., 2025) which unifies EEG and MEG signals, pretrained on 2k hours of EEG and 656 hours of MEG data.

4.3 Implementation Details

For a fair comparison, all evaluated models use the same two-layer MLP classification head and are trained with cross-entropy. Checkpoints are selected by validation loss, and macro-F1 is computed on the held-out test set using the selected checkpoint. STEP is finetuned progressively: we first optimize the adaptive-patching module, then jointly train it with the classification head, and finally optimize the entire model end to end.

Complete implementation details are provided in Appendix D, including data splits and input processing, model-specific configurations, optimization hyperparameters, the three-stage finetuning schedule, and the protocols for transferability evaluation, architecture comparison, ablations, and single- and multi-teacher distillation.

5 Transferability of Pretrained Models to Scientific Tasks

We first examine how pretrained foundation models of related domain (see Section 4.2) perform on scientific time series tasks (introduced in Section 4.1). Specifically, we test three training setups: (i) Loading the pretrained weights but keep the encoder fixed, training only the downstream block; (ii) Loading the pretrained weights and finetune both the encoder and downstream block; (iii) All weights are randomly initialized and trained from scratch, serving as a baseline to evaluate the effectiveness of the backbone architecture. Results are shown in Table 2.

Pretrained encoder frozen *vs.* finetuned. First, comparing the first two rows for each model, we observe that finetuning the encoder generally improves performance over keeping it frozen, as expected. Keeping the pretrained weights frozen leads to near-chance performance on several tasks, indicating that existing pretrained models do not readily support direct transfer to heterogeneous scientific domains.

[6] https://huggingface.co/weighting666/CBraMod

[7] https://huggingface.co/OpenTSLab/BrainOmni

Table 3: Macro-F1 (%) comparison of encoder architectures trained from scratch. The best and second-best results in each column are bolded and underlined, respectively.

Model	GWOSC	LEAVES	STEAD	MarmAudio	SleepEDF	TUAB	RadarComm
PatchTST	72.85	<u>76.64</u>	94.95	66.24	<u>70.70</u>	77.64	60.13
Informer	54.70	71.09	66.31	75.90	62.58	**78.70**	60.19
TimeMoE	71.26	67.67	92.38	59.63	62.08	67.99	53.70
Moirai	<u>86.52</u>	52.34	**99.56**	<u>86.73</u>	63.55	75.90	**61.35**
STEP	**95.69**	**83.41**	<u>96.44</u>	**93.68**	**72.21**	<u>77.81</u>	<u>60.84</u>

Pretrained weights *vs.* random initialization. For the strategies that finetune the pretrained encoder versus training from scratch (comparing the last two rows for each model), the performance difference reflects the effectiveness of the pretrained weights. For SPEAR, finetuning the pretrained encoder improves performance on GWOSC, STEAD, MarmAudio and TUAB, indicating that the pretrained weights benefit these tasks. In contrast, performance on LEAVES and SleepEDF is lower than training from scratch. These results indicate that the effectiveness of pretrained weights varies across scientific tasks.

Intra-domain analysis. Among audio foundation models, SPEAR and Whisper excel on different tasks. Comparing general time series models, it is observed that TimeMoE gives much worse results on MarmAudio dataset compared to Moirai. It is possibly due to the input length restriction. Moirai divides the input into multiple patches for computation, allowing it to handle longer time series data while TimeMoE has a maximum input length of 4096.

Cross-domain comparison. Overall, pretrained models from different domains exhibit varying effectiveness on scientific time series tasks, reflecting task-specific biases. Whisper and SPEAR perform best on acoustic-related tasks such as marmoset vocalisation classification (MarmAudio). TimeMoE and Moirai are strong on relatively short sequences such as LEAVES and RadarComm, while TimeMoE achieves the best result on the multichannel TUAB task. CBraMod is also competitive on TUAB, consistent with its extensive pretraining on the TUEG dataset and its mechanism for integrating spatial and temporal representations. BrainOmni adopts a two-stage training pipeline consisting of a neural tokenizer and a foundation model, which prevents straightforward end-to-end finetuning across domains and is therefore excluded from the table.

6 Validation of the STEP Structure

6.1 Baseline Comparison

This section compares STEP against current SOTA time series structures in handling heterogenous scientific signals. Results are shown in Table 3. All models are trained from scratch to ensure fair comparison.

PatchTST demonstrates competitive performance on datasets characterized by relatively short sequences or strong local regularities (*e.g.*, LEAVES and SleepEDF), confirming the effectiveness of patch-level representations for capturing localized temporal patterns. However, its performance degrades markedly on long-range or weakly periodic signals such as GWOSC and MarmAudio. This suggests that fixed patching schemes impose a rigid temporal granularity, which limits adaptability across heterogeneous scientific signals with vastly different temporal scales.

Informer exhibits consistently weaker performance across most tasks, with particularly poor results on GWOSC and STEAD. While its ProbSparse attention mechanism is effective for reducing computational cost in forecasting scenarios, it appears less suitable for discriminative scientific tasks where fine-grained temporal cues and global context must be jointly preserved. This highlights a structural mismatch between forecasting-oriented sparse attention and general-purpose representation learning.

Moirai shows strong performance on GWOSC, STEAD, and MarmAudio, and surpass TimeMoE on almost downstream tasks. It is noted that although both designed for general time series forecasting,

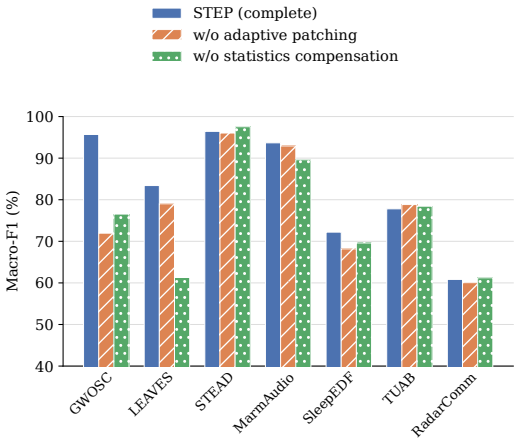


Figure 3: Macro-F1 (%) of STEP trained from scratch and variants without adaptive patching or statistics compensation.

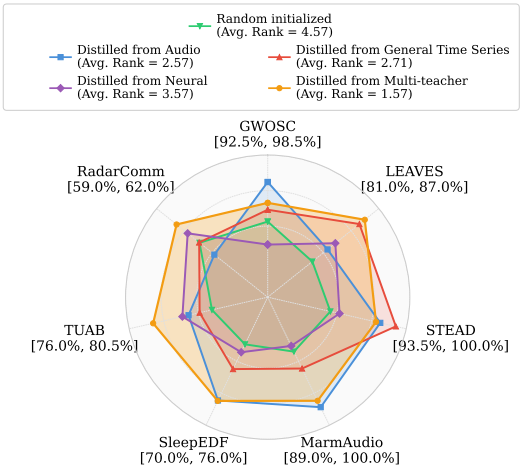


Figure 4: Downstream macro-F1 (%) of STEP after finetuning from random initialization vs. after distillation.

Moirai adopts an encoder-based structure[8] while TimeMoE uses a decoder-only structure. This indicates that encoder-only architectures are better at effectively adapting their representations to scientific data.

STEP achieves the overall best performance across downstream tasks, ranking first on four of them and remaining competitive on the rest. Effectiveness of the key design components are discussed in the following section.

6.2 Ablation Studies

We conduct ablation studies to evaluate the contribution of the two key components of the STEP encoder. Results are shown in Figure 3.

First, the learnable adaptive patching module is replaced by a fix length patching strategy. This variant employs rule-based patching, leaving sequences shorter than T_{thres} unchanged and downsampling longer sequences to a fixed length of T_{thres}. The downsampling ratio is set as the sequence length divided by T_{thres}, rounded up: $\lceil \frac{L}{T_{\text{thres}}} \rceil$, where L is the original sequence length. T_{thres} is set to 200 here. A marked performance drop is observed on GWOSC and SleepEDF, which have long sequences and challenging tasks. For short sequences, adaptive patching learns very small patches, effectively resembling the rule-based approach.

We then remove the concatenation of per-sample mean and standard deviation statistics from the input representation. A distinct performance decrease is observed for LEAVES and MarmAudio, where sequences are fairly irregular in length. In these cases, per-sample normalization alone is unreliable, as the computed mean and standard deviation carry little statistical significance. Appending the mean and standard deviation as auxiliary features mitigates this issue and preserves informative signal statistics.

7 Cross Domain Knowledge Distillation

Based on the results in Section 5, we select SPEAR as the teacher from audio domain, TimeMoE as the teacher from general time series domain, and BrainOmni as the teacher from neural signal domain. Due to TimeMoE's input length constraint, each sequence is truncated to 2048 time step segment. The target representation of BrainOmni is formed by concatenating its final compressed channel output. A complementary analysis of teacher–student representation alignment using CKA is provided in

[8]Moirai 1.1 used in this paper is encoder-based while its subsequent Moirai 2.0 uses decoder-only structure.

Appendix E.1. A comparison between the STEP encoder trained from random initialization and that pretrained via distillation with different teacher models is presented in Figure 4.

Effectiveness of distillation on teacher-favoured tasks. We first observe that knowledge distillation consistently improves student performance on tasks where the teacher model itself exhibits strong downstream performance. For instance, when distilled from SPEAR, the student encoder achieves gains on GWOSC, STEAD, and MarmAudio, all of which are tasks where the SPEAR teacher demonstrates high accuracy. Similarly, distillation from TimeMoE yields clear improvements on LEAVES, where the TimeMoE teacher likewise maintains strong performance.

Limited benefits on tasks where the teacher model struggles. In contrast, for tasks where the teacher model performs poorly, such as LEAVES, RadarComm, and TUAB for SPEAR, the distilled encoder does not consistently outperform its random initialized counterpart. In these cases, performance is largely determined by the student's own architectural inductive biases rather than the distilled knowledge. An illustrative case is the MarmAudio dataset. While TimeMoE shows limited performance due to its fixed input length constraint during distillation, STEP encoder, free from such constraints, still benefits from TimeMoE supervision and achieves stronger downstream results after finetuning. This shows that distillation effectiveness is jointly determined by teacher quality and student capacity.

Cross-teacher complementarity. Figure 4 visualizes the complementary performance profiles of the distilled initializations. SPEAR is particularly effective on long-range audio-like signals such as marmoset vocalization (MarmAudio), whereas TimeMoE excels on low-frequency time series such as light curve (LEAVES). and BrainOmni excels on high dimensional signals (TUAB). We observe that student gains generally reflect the relative strengths of individual teachers, which motivates leveraging multiple teachers to capture complementary knowledge.

Multi-teacher distillation. Multi-teacher distillation improves overall performance by integrating complementary strengths from different teachers. As shown in Figure 4, multi-teacher distillation improves over random initialization across all seven tasks and achieves the largest average gain among the distillation-based initializations. Although the resulting student falls short of the best single-teacher-distilled student on GWOSC, STEAD, and MarmAudio, it delivers consistent gains across the full task suite. Its advantage therefore lies in broad cross-task effectiveness rather than optimal performance on every individual task, supporting its use for cross-domain encoder initialization. Additional experiments with varying downstream training set sizes are presented in Appendix E.2.

Overall, these results show that (i) direct transfer performance and distillation utility are related but not equivalent; (ii) single-teacher supervision produces task-dependent gains and can degrade performance when the transferred representation is poorly matched to the student or downstream task; and (iii) combining heterogeneous teachers broadens task coverage and provides a consistently beneficial initialization across all evaluated domains. These findings support cross-domain multi-teacher distillation as an effective pretraining strategy for scientific time series while highlighting the importance of both teacher diversity and student architecture.

8 Conclusions

In this work, we investigate the challenges of modelling scientific time series, which are characterized by extreme heterogeneity and data sparsity. We first analyze the transferability of foundation models pretrained on related time series domains such as audio, general time series, and neural signals and reveal their complementary strengths and limitations for scientific tasks. Based on these insights, we propose STEP encoder that combines adaptive patching, statistics compensation, and cross-domain distillation to integrate knowledge from multiple foundation models. Experiments on seven diverse scientific time series tasks demonstrate that STEP establishes an effective pretraining paradigm, consistently improving performance across domains. Our results highlight that scientific time series are fundamentally different from well-established time series domains while cross-domain foundation models hold great potential to accelerate representation learning for scientific time series, taking a STEP toward domain-adaptive pretraining and unified scientific signal modelling.

AI USE STATEMENT

We used generative AI tools solely to polish the language of the manuscript, improving grammar, clarity, and readability. These tools were not used for translation or for any scientific component of the work, including research ideation, methodology or experimental design, theoretical development, code implementation, data generation or processing, or interpretation of results. The authors take full responsibility for the final content of the paper, including all AI-assisted language edits.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *Proc. ICML*, Honolulu, 2023c.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proc. ICML*, Vienna, 2024.

Maximilian Dax, Stephen R Green, Jonathan Gair, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time gravitational wave science with neural posterior estimation. *Physical review letters*, 127(24):241103, 2021.

Maximilian Dax, Stephen R Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Neural importance sampling for rapid and reliable gravitational-wave inference. *Physical review letters*, 130(17):171403, 2023.

Maximilian Dax, Stephen R Green, Jonathan Gair, Nihar Gupte, Michael Pürrer, Vivien Raymond, Jonas Wildberger, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time inference for binary neutron star mergers using machine learning. *Nature*, 639(8053):49–53, 2025.

Ya Fei, Ce Yu, Kun Li, Xiaodian Chen, Yajie Zhang, Chenzhou Cui, Jian Xiao, Yunfei Xu, and Yihan Tao. LEAVES: An Expandable Light-curve Data Set for Automatic Classification of Variable Stars. *The Astrophysical Journal Supplement Series*, 275(1):10, 2024.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash Time Series Forecasting Archive. In *Proc. NeurIPS Datasets and Benchmarks*, Virtual, 2021.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A Family of Open Time-series Foundation Models. In *Proc. ICML*, Vienna, 2024.

A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, and J. Picone. The Temple University Artifact Corpus: An Annotated Corpus of EEG Artifacts. In *Proc. SPMB*, Philadelphia, Pennsylvania, 2020.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation. In *Proc. NeurIPS*, New Orleans, 2023.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial Attention in Multidimensional Transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked Autoencoders that Listen. In *Proc. NeurIPS*, New Orleans, 2022.

Anu Jagannath and Jithin Jagannath. Dataset for modulation classification and signal type classification for multi-task and single task learning. *Computer Networks*, 199:108441, 2021.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *Proc. ICLR*, Vienna, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

B Kemp, AH Zwinderman, B Tuk, HAC Kamphuisen, and JJL Oberyé. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Charly Lamothe, Manon Obliger-Debouche, Paul Best, Régis Trapeau, Sabrina Ravel, Thierry Artières, Ricard Marxer, and Pascal Belin. A large annotated dataset of vocalizations by common marmosets. *Scientific Data*, 12(1):782, 2025.

Chenqi Li, Yu Liu, Shuo Zhang, Timothy Denison, and Tingting Zhu. Standing on the Shoulders of Giants: Rethinking EEG Foundation Model Pretraining via Multi-Teacher Distillation. *arXiv preprint arXiv:2603.04478*, 2026.

Yizhi LI, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In *Proc. ICLR*, Vienna, 2024.

Silvia López de Diego. Automated Interpretation of Abnormal Adult Electroencephalograms. Master's thesis, Temple University, 2017.

Michael W. McCracken and Serena Ng. FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

S. Mostafa Mousavi, Yixiao Sheng, Weiqiang Zhu, and Gregory C. Beroza. STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access*, 7:179464–179476, 2019.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *Proc. ICLR*, Kigali, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, South Brisbane, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proc. ICML*, Honolulu, 2023.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One. In *Proc. CVPR*, Seattle, 2024.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *Proc. ICLR*, Singapore, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, Long Beach, 2017.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *Proc. ICLR*, Singapore, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Proc. ICML*, Vienna, 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.

Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Junlin Liu, Ke Fan, Qitan Lv, Jimin Zhuang, Chen Zhang, Zheqi Yuan, Siyuan Hou, Tianyi Lin, Kai Chen, Bowen Zhou, and Chao Zhang. SciTS: Scientific Time Series Understanding and Generation with LLMs. In *Proc. ICLR*, Rio de Janeiro, 2026.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A Brain Foundation Model for Unified EEG and MEG Signals. In *Proc. NeurIPS*, San Diego, 2025.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland. SPEAR: A Unified SSL Framework for Learning Speech and Audio Representations. *arXiv preprint arXiv:2510.25955*, 2025.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution Time-Series Transformer for Long-term Forecasting. In *Proc. AISTATS*, Valencia, 2024.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled Knowledge Distillation. In *Proc. CVPR*, New Orleans, 2022.

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 *Proc. AAAI*, Vancouver, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *Proc. ICML*, Baltimore, Maryland, 2022.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-Supervised Music Representation Learning with Mel Residual Vector Quantization. *arXiv preprint arXiv:2501.01108*, 2025.

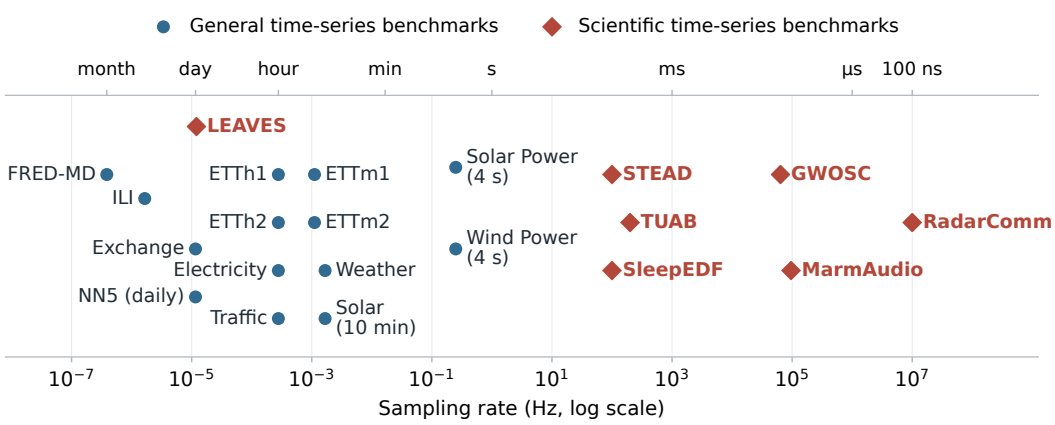


Figure 5: Sampling rates of selected general time-series benchmarks and the seven scientific tasks studied in this work, shown on a logarithmic scale. General benchmark rates follow the ETT dataset (Zhou et al., 2021), the Autoformer benchmark (Wu et al., 2021), and the Monash forecasting archive (Godahewa et al., 2021); scientific task rates are summarized in Table 1. The upper axis shows corresponding sampling intervals. The LEAVES value indicates a representative daily scale for irregular light curves, rather than uniform sampling.

A Scientific Time Series and Their Modeling Characteristics

Scientific inquiry frequently relies on time-resolved measurements to characterize processes whose dynamics cannot be captured by isolated observations (Wu et al., 2026; Mousavi et al., 2019; Kemp et al., 2000). In this work, we focus on temporal signals produced by scientific instruments, controlled experiments, and physics-grounded simulations, where temporal structure provides evidence for events, system states, or signal-generating mechanisms of scientific interest (Dax et al., 2021; Fei et al., 2024; Lamothe et al., 2025). Our study covers seven such tasks across astronomy (GWOSC (Wu et al., 2026) and LEAVES (Fei et al., 2024)), Earth science (STEAD (Mousavi et al., 2019)), bioacoustics (MarmAudio (Lamothe et al., 2025)), neuroscience (SleepEDF (Kemp et al., 2000) and TUAB (López de Diego, 2017)), and radar (RadarComm (Jagannath & Jagannath, 2021)). They involve gravitational-wave strain, stellar light curves, seismic motion, animal vocalisations, electrophysiological activity, and radar signals, thereby representing diverse forms of physical and biological observation.

The temporal structure of an observed signal depends jointly on the underlying phenomenon and the measurement process. Figure 5 illustrates this variation through a comparison with selected forecasting benchmarks. FRED-MD contains monthly macroeconomic observations (McCracken & Ng, 2016); Exchange and ILI use daily and weekly observations, respectively, while Electricity and Traffic are hourly and Weather is recorded every ten minutes (Wu et al., 2021); ETTh1/2 and ETTm1/2 use hourly and fifteen-minute intervals (Zhou et al., 2021); the Monash archive provides Solar and Wind Power series recorded every four seconds (Godahewa et al., 2021). Our tasks span a representative daily scale for LEAVES, 100 Hz for STEAD and SleepEDF, 96 kHz for MarmAudio, and 10 MHz for RadarComm, with input lengths from 24 to 2.4×10^5 samples (Table 1). Thus, these scientific tasks are not uniformly high-frequency, but collectively cover a much wider range of sampling scales than the selected general benchmarks. Equal sample counts can represent very different physical durations, making a single fixed patch size difficult to apply across tasks.

The information encoded by scientific time series is also closely tied to the mechanisms that generate and measure them. Relevant evidence may appear as a weak transient embedded in detector noise, characteristic waveform morphology, oscillatory or phase structure, or coordinated activity across sensors. Channel counts range from 1 to 16 and reflect different measurement configurations: STEAD records three orthogonal components of ground motion, TUAB captures spatially distributed electrical activity through multiple EEG electrodes, and RadarComm represents radar signals using paired channels. Numerical statistics can likewise carry physical or acquisition-related information. Although per-sample normalization is needed to reconcile the substantially different scales of heterogeneous

instruments, removing the original amplitude and dispersion entirely may eliminate information useful for distinguishing the underlying events or states.

The associated learning problems extend beyond predicting future values. The seven tasks require identifying merger and earthquake events, recognizing stellar and bioacoustic signal types, determining physiological states and EEG abnormalities, and classifying radar modulations. Reliable labels often depend on expert interpretation, controlled acquisition, or scientific simulation. Moreover, a long recording may contain only a small number of independent events, subjects, or experimental instances, so large numbers of sampled time points do not necessarily imply abundant supervision. Scientific time-series representation learning must therefore preserve fine-grained evidence across heterogeneous measurement scales while remaining effective under limited task-specific labels.

These considerations motivate the three central components of STEP. Adaptive patching accommodates the absence of a universal temporal resolution, statistics compensation restores potentially informative scale information removed by normalization, and cross-domain distillation exploits transferable temporal knowledge when unified scientific pretraining data are scarce. Together, these components address the heterogeneity and supervision constraints of the tasks studied here.

B Extended Related Work

B.1 Patching as Temporal Compression

Patching changes both the computational cost and the inductive bias of a time-series encoder. Replacing T point tokens by roughly T/S patch tokens, where S is the stride, reduces the dominant self-attention cost from $\mathcal{O}(T^2)$ to approximately $\mathcal{O}((T/S)^2)$. More importantly, each token summarizes a local interval and therefore defines the temporal granularity visible to subsequent layers. PatchTST (Nie et al., 2023) demonstrated that fixed, channel-independent patches are effective for long-horizon forecasting. This design preserves local ordering within each patch and substantially shortens the token sequence, but its patch length and stride remain dataset-level hyperparameters.

Subsequent work has introduced multiple temporal resolutions rather than relying on one patch scale. MTST (Zhang et al., 2024) uses parallel branches with different patch sizes to capture short, high-frequency patterns together with longer seasonal structure. Moirai (Woo et al., 2024) similarly supports multiple patch sizes when learning across datasets with different frequencies. More recently, TimeOmni (Wu et al., 2026) routes an input to a patch size selected from a predefined candidate set. These methods establish that patch resolution is consequential and that using several resolutions can improve coverage of heterogeneous temporal patterns.

However, fixed patches, parallel multi-resolution branches, and routing over a discrete patch bank still require the available resolutions to be chosen in advance. Scientific signals make this requirement particularly restrictive. Equal numbers of samples correspond to very different physical durations across datasets, while task-relevant structures may range from isolated impulses to long oscillatory envelopes. Selecting a very small patch preserves detail but produces an impractically long attention sequence; selecting a large patch is efficient but can average away weak transients. STEP replaces this manual trade-off with differentiable Gaussian windows and a learned stride. Its contribution is therefore not patching itself, but making compression continuously adaptable to the input length and task loss while constraining the resulting token count.

B.2 Foundation Models across Related Time-Series Domains

The main text introduces foundation models from audio, general time series, and neural signals in terms of their data domains and representative models. Here, we expand this overview by examining how pretraining objectives and input representations shape the temporal information these models learn. This perspective also helps distinguish generalisation within a domain from transfer across different measurement processes and task requirements.

Audio: diverse acoustic data and spectrotemporal patterns. Audio pretraining uses reconstruction, discrete-token prediction, or speech–text supervision, which reward different aspects of acoustic structure (Table 4). Models trained on similar signals therefore need not retain the same information.

Table 4: Pretraining mechanisms of representative foundation models. This comparison summarises objectives and input structure, rather than ranking transfer performance.

Model	Learning objective / pipeline	Representation design
Audio		
HuBERT (Hsu et al., 2021)	Masked hidden-unit prediction	Cluster-derived speech units
Whisper (Radford et al., 2023)	Weakly supervised speech–text learning	Acoustic representations linked to textual supervision
AudioMAE (Huang et al., 2022)	Masked reconstruction	Spectrogram patches
BEATs (Chen et al., 2023c)	Discrete acoustic-token prediction	Iteratively refined tokenizer and representations
SPEAR (Yang et al., 2025)	Masked fine-grained token prediction	Multi-codebook quantisation; speech and general audio
General time series		
Moirai (Woo et al., 2024)	Probabilistic forecasting	Multiple patch sizes, variable-variate attention, mixture output
Chronos (Ansari et al., 2024)	Cross-entropy over numerical tokens	Scaling and quantisation of values
TimesFM (Das et al., 2024)	Forecasting	Patched decoder architecture
TimeMoE (Shi et al., 2025)	Autoregressive forecasting	Sparse mixture of experts
MOMENT (Goswami et al., 2024)	Masked patch reconstruction	Time Series Pile; evaluation across multiple tasks
Neural signals		
LaBraM (Jiang et al., 2024)	Masked neural-code prediction	Channel patches; vector-quantised spectrum targets
CBraMod (Wang et al., 2025)	Masked EEG reconstruction	Spatial/temporal attention; asymmetric positional encoding
BrainOmni (Xiao et al., 2025)	Tokenizer and foundation-model pretraining	EEG/MEG; sensor layout, orientation, and type

The coverage of pretraining data provides a second axis of variation. Speech, environmental sounds, and music share temporal and spectral structure, but differ in the patterns relevant to their downstream tasks. SPEAR (Yang et al., 2025) brings speech and general audio into a shared framework through masked prediction of fine-grained tokens obtained using multi-codebook vector quantisation. Such approaches broaden acoustic coverage, while leaving open how far the learned representations extend beyond acoustic measurements. In particular, a waveform or spectrogram interface carries assumptions about sampling rate and frequency structure. Applying that interface to another signal type does not by itself ensure that physically meaningful durations or frequency bands retain their original interpretation.

General time series: heterogeneous temporal data and predictive patterns. General time-series models differ in how they encode numerical values, aggregate observations, and allocate model capacity (Table 4). Their objectives include both forecasting and masked reconstruction.

Forecasting provides supervision for temporal dependencies useful in predicting future observations. This can include local variation as well as trends and seasonality; it does not restrict a model to coarse temporal patterns. However, predictive usefulness and discriminative usefulness are different criteria: evidence that distinguishes an event type or physiological state need not be the evidence most useful for reducing forecasting error. General-purpose pretraining broadens this task perspective. MOMENT (Goswami et al., 2024), for example, uses masked patch reconstruction on the Time Series Pile and evaluates representations on forecasting, classification, anomaly detection, and imputation. The literature therefore encompasses both forecasting-oriented models and encoders evaluated across multiple analytical tasks. Their suitability for a particular application depends on the interaction between pretraining data, learning objective, and the temporal detail retained by the input representation.

Neural signals: electrophysiological recordings and spatiotemporal structure. Neural foundation models combine temporal dynamics with sensor organisation. Table 4 contrasts approaches that incorporate spectral codes, spatial and temporal attention, or explicit sensor geometry.

Sensor information introduces a distinctive consideration for generalisation. A channel represents a measurement at a particular location and, depending on the modality, with a particular orientation and sensor type. Accommodating different channel configurations therefore involves more than accepting a variable number of input sequences: the model must also account for how those measurements relate to one another. Encoding this structure can support learning across neural recording setups, but the same spatial assumptions may not apply to other multivariate signals. Generalisation across EEG or MEG configurations and transfer to a different sensing modality consequently involve different forms of distribution shift.

Cross-domain perspective. Across these three research areas, reusable representations are shaped by three interacting factors: the signal distributions covered during pretraining, the information rewarded by the learning objective, and the temporal or spatial structure preserved by the input interface. Domain labels alone provide an incomplete description of these factors. Models from the same domain can learn different representations under reconstruction, token prediction, or supervised objectives, while models from different domains may share mechanisms for capturing local patterns and longer temporal dependencies. A useful comparison therefore considers which signal properties are retained, which variations are treated as irrelevant, and how closely these choices match the evidence required by a downstream task. This view offers a basis for analysing transferability without assuming that pretraining scale or success within the original domain guarantees performance across scientific measurements.

B.3 Distillation as Knowledge Integration

Knowledge distillation transfers information from a pretrained teacher to a student, commonly by matching predictions or intermediate representations. Recent work has focused on making this transfer more reliable when the teacher and student differ. Decoupled Knowledge Distillation (Zhao et al., 2022) separates supervision associated with the target class from that associated with the remaining classes, allowing the student to use the teacher's output structure more selectively. One-for-All (Hao et al., 2023) instead introduces task-aware projectors between heterogeneous teacher and student architectures. These methods improve the transfer interface, but they generally assume that the models solve the same task and share an output space.

A related direction uses distillation to combine complementary foundation models rather than only to compress one model. AM-RADIO (Ranzinger et al., 2024), for example, distills several vision foundation models into a single encoder and shows that their different representation strengths can be consolidated. More closely related to scientific signals, MTDP (Li et al., 2026) uses a two-stage strategy to pretrain an EEG foundation model: it first fuses representations from vision and time-series teachers through a learned gating network, and then distills the fused representation into the EEG encoder. These studies demonstrate the potential of combining heterogeneous foundation models, but scientific time-series teachers still introduce a practical alignment challenge. An audio model, a forecasting model, and a neural-signal model are trained on different data and produce representations with different sequence lengths and feature dimensions. Their outputs therefore cannot be directly averaged or matched through a common prediction head.

STEP addresses this mismatch through domain-specific feature distillation. Each frozen teacher supervises data from its own domain, while a teacher-specific projection maps the shared STEP representation into that teacher's feature space. The distillation losses are then combined to update one encoder. This design does not require the teachers to share labels, inputs, or representation semantics. Instead, it uses them as complementary sources of temporal supervision, allowing STEP to retain a unified architecture while learning from audio, general time-series, and neural-signal pretraining.

C Datasets and Tasks

The proposed method is evaluated on seven downstream scientific tasks spanning astronomy, Earth science, bioacoustics, neuroscience, and radar. These tasks cover a broad range of signal modalities

and temporal dynamics, as summarized in Table 1. They span sequence lengths from 24 to 2.4×10^5, sampling frequencies from 1.2×10^{-5} to 1.0×10^7 Hz, and channel counts from 1 to 16. The held-out test sets contain 1,024 examples each for GWOSC, LEAVES, and SleepEDF; 2,048 each for STEAD and TUAB; 900 for MarmAudio; and 1,280 for RadarComm.

GWOSC (Wu et al., 2026) is a synthetic gravitational-wave dataset generated following established simulation-and-injection pipelines (Dax et al., 2021; 2023; 2025), in which numerically simulated gravitational waves are injected into real strain data from the GW Open Science Center. The downstream task is binary detection of merger events.

LEAVES (Fei et al., 2024) is a light-curve dataset developed for automatic variable star classification. It integrates publicly available observations from three major astronomical surveys— the All-Sky Automated Survey for SuperNovae (ASAS-SN), Gaia, and the Zwicky Transient Facility (ZTF). The downstream task involves categorizing variable stars into seven superclasses based on the observed light curve: Eclipsing binaries, Rotational variables, RR Lyrae, Cepheids, Long-period variables, Delta Scuti, and Non-variable.

STEAD (Mousavi et al., 2019) is a global dataset of earthquakes recorded between January 1984 and August 2018. Each waveform contains three channels, spans one minute, and is sampled at 100 Hz. The downstream task is binary earthquake detection.

MarmAudio (Lamothe et al., 2025) is a dataset of common marmoset vocalizations, recorded over 40 months at a sampling rate of 96 kHz in a soundproofed animal facility room containing three cages. The dataset is annotated with six vocalization types: Infantcry, Phee, Seep, Trill, Tsik, and Twitter.

SleepEDF (Kemp et al., 2000) is a comprehensive dataset containing overnight polysomnography (PSG) recordings. Each recording includes signals of two-channel EEG (Fpz-Cz and Pz-Cz) and one EOG, sampled at 100Hz. We follow prior work and process the data by utilizing only the Fpz-Cz EEG channel and perform 5-class sleep staging: W, N1, N2, N3, and REM. The EEG signals are processed with a 0.5-45Hz bandpass Butterworth filter to remove background noise.

TUH Abnormal EEG Corpus (TUAB) (López de Diego, 2017) is a subset of the clinical EEG archive collected at Temple University Hospital, developed for automatic identification of abnormal brain activity. Recordings are labelled as normal or abnormal according to clinical interpretation, providing recording-level supervision for EEG abnormality classification. The multichannel signals capture electrical activity at different scalp locations. In this work, the downstream task is binary classification using 16-channel EEG segments sampled at 200 Hz, with 2,000 samples per channel corresponding to 10 seconds.

RadarComm (Jagannath & Jagannath, 2021) contains synthetic radar and communication waveforms generated using GNU Radio for modulation and signal-type classification. The source dataset provides both modulation-format and signal-class labels, and incorporates channel effects and radio hardware impairments under varying signal-to-noise ratios. These effects include multipath fading and carrier-frequency and sampling-rate offsets. Each example used in this work contains 128 complex samples represented by two channels, the in-phase (I) and quadrature (Q) components, at a sampling rate of 10 MHz. We formulate the task as single-label modulation classification with eight classes: Airborne-detection, Airborne-range, Air-Ground-MTI, Ground mapping, Radar-Altimeter, Satcom, AM radio, and short-range.

D Extended Experimental Setup

D.1 Common Data and Evaluation Protocol

All seven downstream problems are cast as sequence classification. We construct disjoint training, validation, and test splits. GWOSC, LEAVES, STEAD, MarmAudio, SleepEDF, and TUAB use 20,000 training and 4,000 validation examples. Because RadarComm is smaller, it uses 10,000 training and 2,000 validation examples. The held-out test-set sizes are reported in Appendix C. Checkpoints are selected solely by validation loss, after which macro-F1 is computed once on the held-out test set. For binary tasks, macro-F1 is the unweighted mean of the F1 scores of the two classes.

Table 5: Training regimes for the transferability study. All five foundation models use the same regime-specific settings.

Regime	Backbone initialization	Trainable modules	Learning rate
Frozen	Pretrained	Classifier only	10^{-4}
Full finetuning	Pretrained	Backbone and classifier	5×10^{-5}
From scratch	Random	Backbone and classifier	10^{-5}

Variable-length inputs are padded within a batch and valid lengths are retained by the input pipeline. For STEP, an input channel x is normalized independently within each sample as $\hat{x} = (x - \mu)/(\sigma + 10^{-7})$. The statistics-compensation branch represents the removed mean by $\mathrm{sign}(\mu)\log(1 + |\mu|)$ and the standard deviation by $\log(\sigma + 10^{-7})$, and appends these features to the initial convolutional representation.

STEP, Whisper, and SPEAR process each complete input, including the full MarmAudio recordings. PatchTST and TimeMoE cannot encode the longest inputs within a single context window. For these models, long signals are divided into context-length windows because the model input is bounded by a maximum length. When training, a new window is selected randomly at each epoch; while inference covers the signal with sliding windows and obtains the sequence-level prediction by voting over window-level predictions. Crops for event-detection tasks are selected using the event label so that positive training windows contain the annotated event; other training windows are sampled uniformly. Whisper and SPEAR are supplied with a nominal sampling rate of 16 kHz without resampling the original measurements. Models without native multichannel support receive channels concatenated along the time dimension in a fixed channel order, whereas models with native channel support preserve the channel axis.

D.2 Transferability of Foundation Models

For Whisper, SPEAR, TimeMoE, Moirai, and CBraMod, we replace the original task head with the common two-layer classifier $1024 \to 512 \to K$, using ReLU between the two linear layers. Temporal representations are mean-pooled before classification. Table 5 specifies the three training regimes; each runs for at most 30 epochs. We use Adam and a StepLR schedule that multiplies the learning rate by 0.5 every five epochs. This study uses cross-entropy weighted by the square-root inverse class frequency. Its nominal global batch size is 32 and is divided across the participating devices in distributed runs.

D.3 STEP and Architecture Baselines

The from-scratch architecture comparison uses PatchTST, Informer, TimeMoE, Moirai, and STEP with randomly initialized weights. All methods use the same data splits, standard cross-entropy, validation-loss model selection, and a maximum of 30 epochs. Model-specific input limits and batch sizes are retained because their memory costs differ substantially.

PatchTST uses three encoder layers, hidden size 128, 16 attention heads, encoder and head dropout 0.2, AdamW with learning rate 2×10^{-4} and weight decay 10^{-2}, three warm-up epochs followed by cosine decay, and early stopping with patience five. Patch length/stride are 16/8 for LEAVES, 8/4 for RadarComm, and 32/16 for the other five datasets. Context length is 1,318 for LEAVES, 128 for RadarComm, 2,000 for TUAB, and 2,048 otherwise. Its batch size is 32 except for the 16-channel TUAB task, for which it is 16.

Informer uses two encoder layers, hidden size 512, feed-forward size 2,048, eight attention heads, dropout 0.05, and ProbSparse factor 5. It is optimized with AdamW at 2×10^{-4}, weight decay 10^{-2}, three warm-up epochs followed by cosine decay, and early-stopping patience five.

STEP contains 125,466,216 parameters. Its Transformer body contains 17 blocks with hidden size 768, and the classifier is a two-layer projection $1024 \to 512 \to K$ with ReLU. The adaptive-patching constraints are $S_{\max} = 400$, $N_{\min} = 5$, and $N_{\max} = 200$; the stride and length penalties both have unit weight. We optimize STEP using ScaledAdam with base learning rate 10^{-3}, clipping scale 2.0,

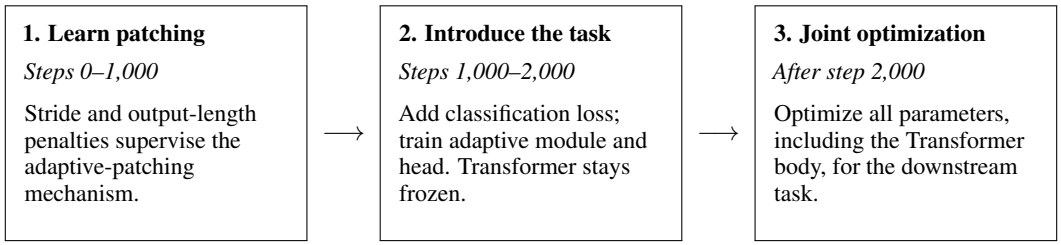


Figure 6: STEP downstream training schedule. Distilled and randomly initialized encoders follow the same stages, enabling comparison of their initializations.

and the Eden learning-rate schedule with 2,000 warm-up batches. The stride-learning parameter groups use a learning-rate scale of 0.5. Per-device batch size is one, reflecting the memory cost of processing complete high-resolution inputs. Training runs for at most 30 epochs, with early stopping after five epochs without an improvement in validation loss.

The same three-stage schedule is used for randomly initialized and distilled STEP encoders (Figure 6).

D.4 Single- and Multi-Teacher Distillation

We select SPEAR, TimeMoE, and BrainOmni as the audio, general-time-series, and neural-signal teachers, respectively. Teachers remain frozen. SPEAR is paired with LibriSpeech 960h (Panayotov et al., 2015); TimeMoE is paired with Time-300B (Shi et al., 2025), with each series cropped to 2,048 points; and BrainOmni is paired with its original pretraining corpus (Xiao et al., 2025). For BrainOmni, compressed channel outputs are concatenated to form the target. The STEP output length is constrained to match the corresponding teacher, and a teacher-specific projection maps the 768-dimensional student representation into the teacher feature space.

For teacher k, we minimize $\mathcal{L}_k = \|P_k(\mathbf{H}_{\mathrm{STEP}}) - \mathbf{H}_k\|_1$ over aligned time steps. Multi-teacher pretraining uses $\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{SPEAR}} + \mathcal{L}_{\mathrm{TimeMoE}} + 0.05\mathcal{L}_{\mathrm{BrainOmni}}$. Rather than sampling a single domain probabilistically, every optimization step draws one batch from each of the three domain loaders, accumulates their gradients, and then updates STEP once. The per-device batch sizes are 1, 25, and 1 for LibriSpeech, Time-300B, and BrainOmni data, respectively. The teachers remain frozen throughout.

Multi-teacher distillation is run for at most 15 epochs with ScaledAdam, base learning rate 1×10^{-3}, the Eden schedule, and 2,000 warm-up batches; the best checkpoint is selected by validation loss. Single-teacher SPEAR and BrainOmni runs are trained for 10 epochs and selected by validation loss. Because a full pass through Time-300B is much larger, TimeMoE distillation is step-based; the validation-selected checkpoint at 1.32 million steps is used. After pretraining, each distilled checkpoint is finetuned on every downstream task using exactly the same three-stage STEP protocol as the randomly initialized control. Thus, the single- and multi-teacher comparisons isolate the effect of pretrained initialization.

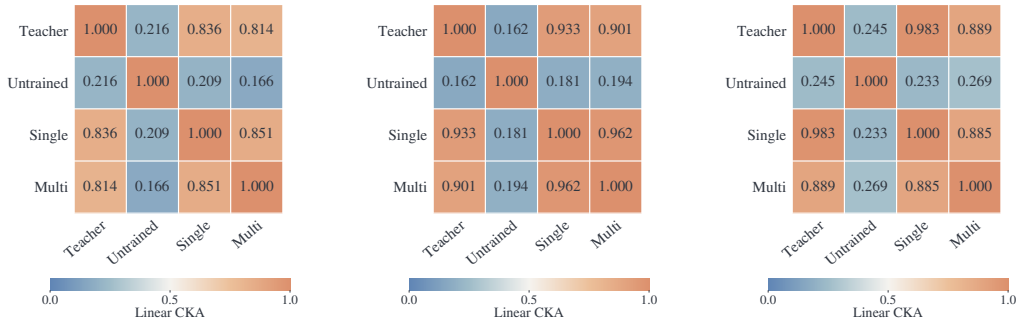
D.5 Compute and Reproducibility

Experiments are implemented in PyTorch and run on NVIDIA H200 GPUs, using DDP or DeepSpeed for distributed training as appropriate. Long-input tasks, including GWOSC and MarmAudio, use eight GPUs, while the remaining downstream tasks use four. The reported STEP runs use full-precision training. We will release the code, preprocessing scripts, split files, configurations, and checkpoints upon acceptance.

E Extended Experiments

E.1 Representation Alignment after Distillation

We complement the downstream experiments with a direct analysis of the representations learned through distillation. Specifically, we examine whether the STEP encoder acquires the sample



(a) LibriSpeech / SPEAR (b) Time-300B / TimeMoE (c) BrainOmni corpus / BrainOmni

Figure 7: Sample-level linear CKA on 500 examples per distillation domain. Each matrix compares the domain teacher, a untrained STEP encoder, the corresponding single-teacher encoder, and the shared multi-teacher encoder. Axis labels Teacher, Untrained, Single, and Multi denote these four models, respectively. All panels use the same color scale from 0 to 1.

similarity structure of its teachers by comparing sample-level linear centered kernel alignment (CKA) between each teacher and three STEP variants: untrained (the undistilled control), distillation from the corresponding single teacher, and multi-teacher distillation. All distilled encoders are evaluated before downstream finetuning, so this analysis probes representation transfer during pretraining.

Evaluation protocol. We sample 500 examples without replacement from each of the configured distillation training pools using seed 42: LibriSpeech for SPEAR, Time-300B for TimeMoE, and the BrainOmni pretraining corpus for BrainOmni. Within a domain, all four models receive the same saved inputs in the same order. The same multi-teacher checkpoint is evaluated in all three domains. Models run in evaluation mode, with one example per forward pass. We mean-pool valid temporal tokens to obtain one vector per example. Student features are extracted before the teacher-specific distillation projection heads, while teacher features are taken from the wrappers that supply the distillation targets. Thus, the comparison probes the encoder itself rather than its learned projection heads.

For feature matrices $X \in \mathbb{R}^{n \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$, we compute

$$\mathrm{CKA}(X, Y) = \frac{\langle K_c, L_c \rangle_F}{\|K_c\|_F \|L_c\|_F}. \tag{5}$$

Here, $K_c = HXX^\top H$ and $L_c = HYY^\top H$ are centered sample Gram matrices, with $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $n = 500$. This is the biased linear CKA estimator. Higher CKA indicates more similar relationships among the same examples. The two feature dimensions may differ, so no additional projection or dimensionality reduction is needed for this comparison.

Evidence of representation transfer. Figure 7 shows a consistent increase in teacher alignment after distillation. On LibriSpeech, teacher–student CKA increases from 0.216 for random initialization to 0.836 for single-teacher and 0.814 for multi-teacher distillation. The corresponding values are 0.162, 0.933, and 0.901 on Time-300B, and 0.245, 0.983, and 0.889 on the BrainOmni corpus. Multi-teacher distillation therefore exceeds the random baseline by 0.598, 0.739, and 0.645, respectively. Because these similarities are measured before the distillation projection heads, they support the conclusion that teacher supervision shapes the shared encoder's representation structure. The same multi-teacher encoder aligns with each teacher on its respective domain, providing evidence that one encoder can retain representation structure from all three sources.

Interpretation and scope. Single-teacher distillation achieves higher teacher alignment than multi-teacher distillation in every domain, with gaps of 0.022, 0.032, and 0.094. These observations support effective multi-domain representation transfer, but do not imply that multi-teacher training maximizes similarity to each individual teacher. CKA measures representation similarity rather than classification quality; together with the downstream gains, it supplies complementary evidence for the effectiveness of distillation.

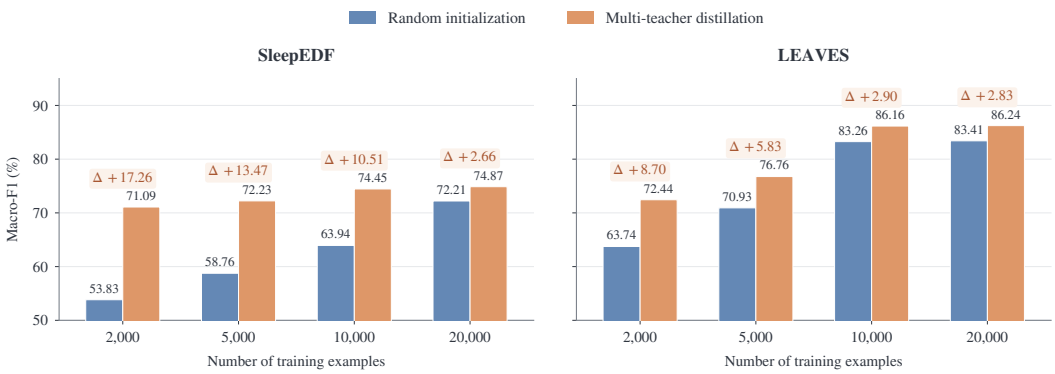


Figure 8: Macro-F1 as a function of the number of downstream training examples on SleepEDF and LEAVES. Bars compare random initialization with multi-teacher distillation; Δ denotes the absolute macro-F1 gain from distillation.

E.2 Impact of Data Scale on Distillation Efficacy

We further examine whether multi-teacher distillation is particularly beneficial when downstream labels are limited. On SleepEDF and LEAVES, we subsample the training set to 2,000, 5,000, 10,000, and 20,000 examples and compare STEP initialized randomly with STEP initialized by multi-teacher distillation. The class proportions in each subset are kept approximately consistent with those of the full 20,000-example training set. The results with 20,000 training examples are taken from the main experiments reported above. The validation and test sets remain fixed across all training set sizes. Except for the training set size, all finetuning and evaluation settings remain the same as in the main experiments.

As shown in Figure 8, multi-teacher initialization improves macro-F1 at every data scale on both tasks, with the largest gains in the lowest-data regime. On SleepEDF, the gain decreases from 17.26 points with 2,000 examples to 2.66 points with 20,000 examples. On LEAVES, it decreases from 8.70 to 2.83 points. These results indicate that cross-domain distillation improves data efficiency: its advantage is most pronounced when downstream supervision is scarce and narrows as more labelled examples become available.