Template Matching, Not Time Learning: A Diagnostic for Self-Supervised Light Curve Encoders

Emma Chickles [1] **Kevin Burdge** [1]

Abstract

We propose period regression on catalogued variable stars as a diagnostic for whether a self-supervised encoder has learned each source's period from its raw irregularly-sampled brightness ("absolute time"), or has merely template-matched class identity. Decomposing predicted-log-period variance into between- and within-class components separates "the encoder identifies the class and predicts the class-mean period" from "the encoder reads the period off the individual source." Applied to eight encoders — including pretrained Chronos-T5 (46M) and MOMENT (110M), a 4.4M cadence-as-channel BiGRU, and a 4.8M continuous-time SSL transformer — the diagnostic returns the same answer: **60–70% of period-regression R^2 is between-class, and within-class Spearman ρ stays at** 0.20**–**0.26 (95% CIs over multiple seeds and data splits). A nonlinear MLP probe does not change this: it raises overall R^2 ($0.66 \to 0.77$ for the BiGRU) by sharpening class templates while ρ_{within} stays ≤ 0.31, whereas a direct Lomb–Scargle extractor reaches $\rho_{\text{within}} = 0.78$ on the same data — the within-source signal is recoverable, and the encoders are under-extracting it. As a control, retraining Chronos-T5 from scratch at matched scale collapses period R^2 from 0.685 to 0.186, isolating cross-domain pretraining as the dominant driver of its downstream win — but *not* of genuine time encoding. No encoder tested, including the published ASTROMER checkpoint, has learned absolute time on irregular ZTF photometry.

[1]MIT Kavli Institute for Astrophysics and Space Research, Cambridge, MA, USA. Correspondence to: Emma Chickles <echickle@mit.edu>.

Proceedings of the AI4Physics Workshop at the 43^{rd} International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026.

1. Introduction

Light curves — time series of stellar flux — are the primary data product of time-domain astronomical surveys, and the basis for inferring stellar variability, orbital periods, and transient phenomena (Aerts et al., 2010). Ground-based surveys observe a source only when it is above the horizon and the weather cooperates; the resulting time series has Δt that varies by an order of magnitude within a season and by two orders of magnitude across seasonal gaps. The Zwicky Transient Facility (Bellm, 2014; Masci et al., 2018) produces $\sim 10^{11}$ such epochs across $\sim 10^9$ stars on this irregular grid, and Rubin LSST (Ivezić et al., 2019) will exceed this by another 1–2 orders of magnitude. Hand-labeled subsets are vanishingly rare — the Chen et al. (2020) catalog of $\sim 25{,}000$ multi-class variables is fewer than 10^{-4} of ZTF — so self-supervised representation learning has emerged as the standard response (Morvan et al., 2022; Donoso-Oliva et al., 2023; Pan et al., 2024b; Zhang et al., 2024; Zuo et al., 2026).

Irregular sampling and the time-encoding problem. On a regular time grid, an architecture that learns a Fourier-like representation gets one essentially for free: position k corresponds to frequency k/N via the DFT. Mainstream time-series foundation models exploit this implicitly — CHRONOS (Ansari et al., 2024) and MOMENT (Goswami et al., 2024) both consume tokens at fixed stride. Astronomical light curves break the mapping: each observation i contributes phase $2\pi f t_i$ at candidate frequency f, and recovering periodicity from such data requires a non-uniform Fourier representation in which every observation carries its phase contribution at every frequency the model represents. The Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) solves this explicitly via closed-form regression at every candidate frequency; an SSL encoder that has learned period from irregular data is, in effect, learning a data-driven analogue. The astronomy-SSL line of work — ASTROMER (Donoso-Oliva et al., 2023; Donoso-Oliva, Cristobal et al., 2026), positional-encoding ablations (Moreno-Cartagena et al., 2023), contrastive light-curve embeddings (Ding et al., 2026), ATAT (Cabrera-Vives, G. et al., 2024), Astroconformer (Pan et al., 2024a), FALCO (Zuo et al., 2026) — can be read as a search for the

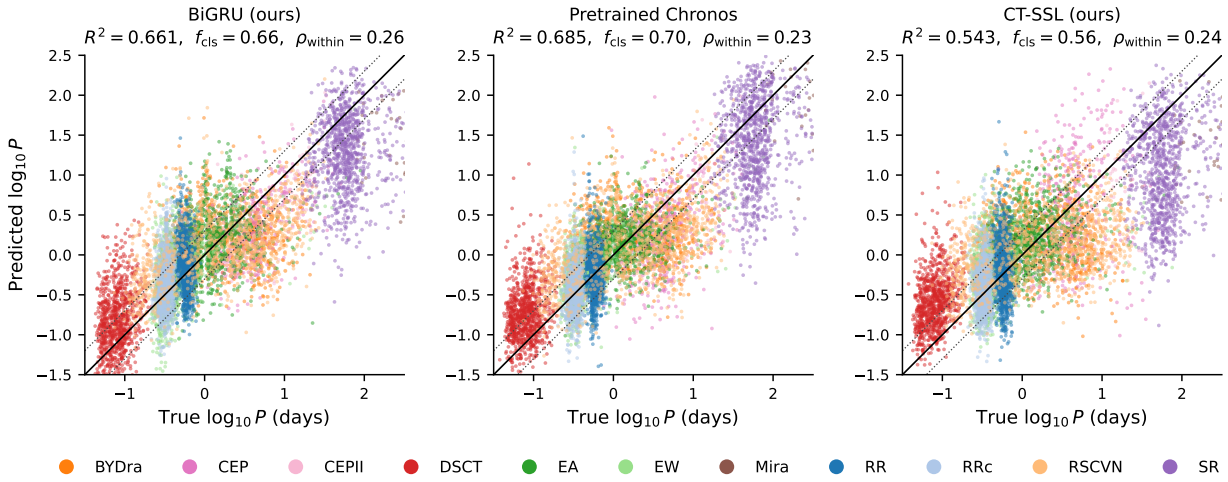

Figure 1. **Overall R^2 looks like time learning; the within-class structure shows it is not.** Predicted vs. true $\log_{10}(P)$ from a ridge probe on each method's z^{sig} embeddings, colored by catalog class. Solid black: $y = x$. Dotted: $P/2$ and $2P$ alias lines. Per-panel diagnostics: overall R^2 on $\log_{10}(P)$; f_{cls}, between-class share of predicted-log P variance (1.0 = pure class-template prediction); ρ_{within}, median within-class Spearman rank correlation (1.0 = perfect source-level period ordering inside each class). All three methods — a 4.4M cadence-as-channel BiGRU, the 46M pretrained Chronos checkpoint ($\sim 10^9$ pretraining timesteps), and our 4.8M continuous-time SSL transformer with explicit pairwise time- bias attention — show the same pattern: each class forms a near-vertical stripe at its class-typical period (RR Lyrae blue at $\log_{10} P \approx -0.2$, semiregulars purple ≈ 1.5–2.5), with weak within-stripe ordering ($\rho_{\mathrm{within}} \leq 0.26$). Cross-domain pretraining, an explicit cadence channel, and pairwise continuous-time attention all land in the same template-matching regime. Section 5.2 dissects the diagnostic; §5.3 reports the architectural-fairness control with Chronos-T5 retrained from scratch.

right time-encoding inductive bias on irregular data.

Period regression as a diagnostic. If an encoder has learned a useful non-uniform Fourier representation, the orbital period should be recoverable from its embedding. The naive metric — ridge- probe R^2 on $\log_{10}(P)$ — conflates two very different behaviors: (a) the encoder identifies the source's variability class (categories of variable star, e.g. RR Lyrae, eclipsing binaries, or Miras), and the probe predicts the class-mean period, versus (b) the encoder actually reads the period off this individual source's photometry. The distinction matters because astronomical variability classes have wildly different mean periods (RR Lyrae ~ 0.5 d, Mira ~ 100 d), so a model that learns class identity gets period for free without ever building a time representation. We therefore propose decomposing predicted-log-P variance into between-class and within-class components, summarized by two numbers we use throughout: the *class-driven fraction* f_{cls} (between-class share of predicted-period variance), which isolates the template-matching contribution and runs from 0 to 1; and the *within-class Spearman* ρ_{within}, the median rank correlation between true and predicted period *inside* each class, which measures source-level period ordering. Only $\rho_{\mathrm{within}} \to 1$ indicates the encoder has read each source's own period — what we mean by "learning absolute time" — as opposed to merely recalling its class's typical period. We use this as the primary diagnostic, with overall R^2 as a context number.

The architectural-fairness confound. Existing astronomy benchmarks compare a 46M-parameter foundation model pretrained on 10^9 cross-domain timesteps against a 5M astronomy-specific model pretrained on $\sim 10^5$ light curves. Strong performance from the foundation model could reflect either architectural fit *or* cross-domain pretraining — two very different stories for practitioners deciding what to deploy. We close the gap by retraining Chronos-T5 from scratch on the same 400,000-source ZTF pool at parameter-matched scale ($\approx$ 4.9M). **LogR balanced accuracy collapses from 0.650 (pretrained, 46M, 10^9 timesteps) to 0.317 (from-scratch, 4.9M, 4×10^5 sources) and period-regression R^2 from 0.685 to 0.186.** The collapse isolates cross-domain pretraining as the dominant driver of pretrained-Chronos performance, and suggests that whatever Fourier behavior the pretrained checkpoint exhibits was acquired during the 10^9-timestep regime, not by the architecture itself.

The diagnostic returns the same answer for every method tested: template matching, not time learning. Across eight encoders — including pretrained Chronos-T5 (46M, $\sim 10^9$ timesteps) and pretrained MOMENT (110M), our 4.8M continuous-time SSL transformer, and a 4.4M cadence-as-channel BiGRU — the class-driven fraction of predicted-log-P variance sits between 0.60 and 0.70, and within- class Spearman ρ between 0.20 and 0.26. An overall $R^2 = 0.685$ on $\log_{10}(P)$ that looks like time learning is, on

decomposition, $\sim 70\%$ class-template matching. Even pretrained Chronos — the highest-overall-R^2 method we tested — has $\rho_{\text{within}} = 0.23$, lower than our cadence-as-channel BiGRU (0.26). For Chronos-T5 retrained from scratch at matched scale, ρ_{within} collapses to 0.03 — no source-level signal at all.

Cadence-as-channel as one control among others. A 4.4M bidirectional GRU on $(\text{flux}, \log_{10}\sigma, \log_{10}\Delta t)$, trained with plain InfoNCE on 4×10^5 ZTF sources, matches the 46M pretrained Chronos checkpoint on classification (0.651 vs. 0.650 LogR) and on overall period R^2 (0.661 vs. 0.685). Pretrained Chronos sees flux on a regular token grid with no access to Δt, so the parity is a statement about *how* time information enters the model, not about recurrent vs. attention. But the diagnostic shows that this parity is also parity in template- matching: neither method has actually learned to read period off an individual source.

Contributions. We contribute: (i) a **diagnostic** decomposing period-regression predictions into between- (f_{cls}) and within-class (ρ_{within}) components, separating class-template matching from genuine per-source period reading; (ii) a **negative result** — across eight encoders (including 46M/110M pretrained checkpoints and our continuous-time transformer), 60–70% of period R^2 is between-class and $\rho_{\text{within}} \leq 0.26$, so high R^2 is not evidence of learned time encoding; (iii) an **architectural-fairness benchmark** — retraining Chronos-T5 from scratch at matched scale drops period R^2 from 0.685 to 0.186 and ρ_{within} to 0.03, isolating cross-domain pretraining as the driver of its win but not of any time-encoding capability; and (iv) a **recommendation** that astronomy-SSL papers report f_{cls} and ρ_{within} alongside period R^2, treating ρ_{within} as the metric of merit.

2. Background and Related Work

ZTF photometry. The Zwicky Transient Facility (Bellm, 2014; Masci et al., 2018) produces irregularly sampled g, r, i photometry at median cadence $\sim$ 1 day, yielding $\sim$ 200–1000 epochs per source across multiple seasons and per-epoch flux uncertainties spanning two orders of magnitude. Standard preprocessing for SSL applies per-source MAD normalization and calibration to a reference magnitude.

Astro-SSL on light curves. The astronomy-specific SSL literature began with masked-token pretraining of transformer encoders (Morvan et al., 2022; Donoso-Oliva et al., 2023) and has expanded to scalable variants (Moreno-Cartagena et al., 2023), contrastive formulations (Ding et al., 2026), conformer architectures with mixed local/global attention (Pan et al., 2024a), multi-survey generalization (Donoso-Oliva, Cristobal et al., 2026), transient classification transformers (Allam & McEwen, 2024), and, most recently, multi-modal photometry+spectroscopy alignment for supernovae (Zhang et al., 2024) and large-scale foundation pretraining (Zuo et al., 2026). Several of these report scaling-law-like behavior on stellar surface parameters (Pan et al., 2024b). Our work fits in this lineage but focuses on a question these papers do not directly answer: how much of *cross-domain* foundation-model performance carries over to the astronomy domain when pretraining data is held fixed?

General time-series foundation models. CHRONOS (Ansari et al., 2024) discretizes flux into 4096 bins with a mean-scaling tokenizer and runs T5 encoder–decoder as a causal LM. MOMENT (Goswami et al., 2024) runs masked-patch reconstruction on fixed-grid sequences. PATCHTST (Nie et al., 2023) patchifies multi-channel inputs and applies channel-independent attention. All three release pretrained checkpoints at $\geq 10^8$ params, pretrained on $\geq 10^9$ cross-domain timesteps.

Architectural fairness. We define a *fair* architectural comparison as: same labeled probe set, same pretraining data pool, parameter counts within 5%, same SSL objective when feasible, same probe protocol. The frozen-pretrained-foundation-model comparisons violate the second and third by 4–5 orders of magnitude in pretraining-data scale and $10\times$ in parameter count.

3. Method

3.1. Encoder zoo

We compare five encoder families at $\approx$ 4.4–4.9M parameters (Table 3):

- **Continuous-time SSL transformer (CT-SSL, ours, 4.82M).** A six-layer attention encoder with continuous-time bias $B(\Delta t_{ij}) = \sum_k a_k \sin(\omega_k \Delta t_{ij}) + b_k \cos(\omega_k \Delta t_{ij})$ added to the attention logits. Time enters via random Fourier features (64-dim) on a per-token t_i embedding plus this pairwise bias with $K = 16$ frequency pairs. Built on ZTF $g{+}r{+}i$ with shared MAD scaling. Reference for the headline "ours" baseline.
- **1D CNN ($\sim$ 4.4M).** Six `(Conv1d` k=7 $\rightarrow$ `GroupNorm` $\rightarrow$ `GELU` $\rightarrow$ `Dropout)` blocks at d=352 channels, with `[flux,` $\log_{10}\sigma$`,` $\log_{10}\Delta t$`]` as a 3-channel sequence. Mean-pool over valid positions, two-head projection.
- **Bidirectional GRU ($\sim$ 4.4M).** Three-layer `nn.GRU` at hidden= 320 in both directions on the same 3-channel sequence. Mean-pool of $(B, L, 2d)$ outputs. fp32 training (bf16 caused gradient explosion in early experiments).

- **Chronos-T5 from scratch (4.92M).** MEANSCALEUNIFORMBINS tokenizer at vocab=4096, $\pm 15\sigma$ range, identical to released checkpoints. T5 encoder+decoder at d=192, 4+4 layers, d_{ff}=768, four heads. Univariate flux only (σ, t discarded by the tokenizer's design). Causal LM objective via the standard seq2seq cross-entropy on tokenized future flux (context=320, prediction=64).

- **PatchTST (fs).** HF `PatchTSTForPretraining` at 16×16 patches, d=128, three layers, three input channels (time/flux/err). 100K-source pretraining (per Nie et al. (2023)'s masked-patch recipe).

For comparison we also report two *frozen pretrained* baselines: CHRONOS-SMALL (46M, the public `amazon/chronos-t5-small` checkpoint, embedded via mean pool of encoder hidden states) and MOMENT (110M, similarly mean-pooled).

3.2. Data pipeline

Pretraining pool. We stream g+r+i light curves directly from ZTF matchfiles (`/orcd/data/kburdge/001/ZTF/matchfiles/`), applying calibration, quality flags, MAD normalization, and median-matching across bands per source. Sources with < 10 valid epochs are dropped. Sequences are random-windowed/padded to length 384. Fig. 3 (Appendix) shows a representative light curve and the aliasing that makes period recovery hard.

Holdout. The 25,565 sources from the Chen et al. 2020 labeled subset (Chen et al., 2020) are held out from pretraining via per-tile $(\mathrm{RA}, \mathrm{Dec})$ (location on the sky) matching at $1''$, ensuring that no source the encoder embeds at probe time was seen during SSL.

Augmentation. Two non-overlapping random windows per source (disjoint-window augmentation) form the InfoNCE positive pair, with flux-error-aware noise equalization at p=0.5 on the second view.

3.3. Loss

For all four contrastive baselines (CT-SSL, CNN, RNN, PatchTST) we use the same disentangled loss, following the spirit of Audenaert et al. (2025)'s dual-encoder structured-contrastive setup that separates physics from instrument properties: a signal latent z^{sig} trained with InfoNCE, and a quality latent z^{qual} regressed against per-source uncertainty so the InfoNCE encoder is not forced to encode quality/systematics in the signal latent.

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}}\Big(P(z_1^{\text{sig}}), P(z_2^{\text{sig}})\Big) + \lambda_q \cdot \tfrac{1}{2} \sum_{v\in\{1,2\}} \text{MSE}\big(z_v^{\text{qual}}, \log_{10}\tilde{\sigma}\big)\,, \tag{1}$$

where P is a 2-layer MLP projection head, τ=0.07 for InfoNCE, λ_q=1.0, and $\tilde{\sigma}$ is the per-source median flux uncertainty. The qual head is required for DDP correctness (otherwise its parameters never participate in the loss and the reducer raises "parameters which did not receive grad").

3.4. Probe protocol

For each method we (i) embed all 25,565 labeled sources via mean-pool of the encoder, (ii) standardize each dimension to unit variance, and (iii) run two probes:

- **LogR:** L_2-regularized logistic regression (C=1.0), 5-fold stratified CV, balanced accuracy.

- **kNN:** k=10, cosine distance, same CV split.

We report *balanced* accuracy because the labeled subset is severely class- imbalanced (3000 each for the eight common classes; 1132 CEP, 316 CEPII, 117 Mira). All confusion matrices and per-class precision/recall/F1 are in Appendix B.

4. Experimental Setup

Hardware. Training on 4× NVIDIA A100-80GB or 4× L40S-46GB nodes via DDP. Probes on 1× A100/L40S, $\sim$ 1 minute per method.

Optimizer. AdamW at 3×10^{-4} (CT-SSL, CNN, Chronos-T5) or 1×10^{-4} (BiGRU; lower learning rate to prevent gradient explosion in fp32 recurrent path), weight decay 10^{-4}, warmup-cosine schedule with 2 warmup epochs over 40 total. Patience-10 early stopping on validation loss. bf16 mixed precision throughout, except BiGRU which uses fp32 to avoid recurrent gradient overflow.

Effective batch size. CT-SSL: $96 \times 4 = 384$. CNN/RNN: $96 \times 4 = 384$. Chronos-T5: $256 \times 4 = 1024$ (more memory-efficient seq2seq path).

Pretraining budget. Each model trains for ≤ 40 epochs or until early-stopping fires. The CT-SSL/CNN/RNN runs see $\approx 1.6\times10^7$ contrastive pair examples per epoch ($4\times$ 400K independent dataloader iterations under DDP). Chronos-T5 sees $4\times$1M source-windows.

Reproducibility. Seed 42 for all encoder training. Code, the diagnostic script, and the probe protocol are released at `https://github.com/emmachickles/`

period-diagnostic, and the embedding files and labeled subset used for every table are archived on Zenodo (DOI to be assigned). The probe results in Table 4 use 5 CV-split seeds ($\{42, 0, 1, 7, 2024\}$) and 1000-resample source bootstraps; re-running scripts/diag_mlp_and_ci.py on the released embeddings reproduces every CI. The labeled probe set is the 25,565-source Chen et al. (2020) subset; periods are that catalog's LS values.

5. Results

5.1. Classification probe: a surface-level puzzle

Table 1 reports linear-probe and k-NN balanced accuracy on the 11-class subset (Fig. 4, Appendix, visualizes the same ranking). The classical Random Forest (RF; Breiman, 2001) baseline on hand-crafted Chen et al. features (~ 60 engineered statistics including median, MAD, rolling autocorrelation, and catalog period) sets a soft ceiling at 0.882 that no SSL method reaches.

Two patterns in the SSL methods are worth registering before we interpret them. First, the 4.4M cadence-as-channel BiGRU and the 46M pretrained Chronos checkpoint reach near-identical balanced accuracy (0.6514 vs. 0.6503), despite a $10\times$ parameter gap and a $\sim 2{,}500\times$ pretraining-data gap. Second, when we retrain Chronos-T5 from scratch at parameter-matched scale (4.92M) on the same 400K-source ZTF pool, balanced accuracy collapses to 0.3170.

Read at face value, these two observations support a clean architecture-vs.-pretraining narrative: a small recurrent network with cadence as a feature channel is competitive with a 46M cross-domain pretrained model, and removing the cross-domain pretraining is what kills foundation-model performance. We will show in §5.2 that this reading is incomplete — in particular, that the BiGRU/Chronos parity on classification is parity *in template matching*, not parity in any deeper time-encoding sense — but the classification numbers themselves are what they are, and they motivate the question the diagnostic is built to answer: when two methods look equivalent on a downstream label, what are they actually equivalent at?

A few smaller observations from the table that we will not dwell on: the 1D CNN at 4.41M parameters reaches 0.6174 LogR / 0.5598 kNN, between the BiGRU and the CT-SSL transformer; PatchTST trails the field at 0.15 LogR despite a comparable parameter budget; and the BiGRU's gain over Chronos-T5-from-scratch is broad across all 11 classes (per-class precision/recall/F1 in Appendix B), not driven by a single dominant class. We carry the BiGRU, the pretrained Chronos checkpoint, and Chronos-T5-from-scratch forward as the three reference points for the diagnostic in §5.2; the others appear in the period-regression table for completeness.

5.2. The diagnostic: template matching, not time learning

The classification parity in §5.1 raises a question that overall accuracy cannot answer: when the BiGRU and pretrained Chronos reach the same balanced accuracy, are they doing the same thing? We use period regression to find out, and design the probe specifically to separate two behaviors that overall R^2 on $\log_{10}(P)$ would otherwise conflate.

Probe and metrics. We evaluate each method's embedding for direct period inference via ridge regression from z^{sig} to $\log_{10}(P)$, with P taken from the Chen et al. (2020) catalog (Lomb-Scargle periods, with the catalog's own purity cuts already applied; 5-fold CV with RidgeCV over $\alpha \in [10^{-2}, 10^{4}]$). We report: (i) R^2 *on* $\log_{10}(P)$ — the conventional summary; (ii) *class-driven fraction* (f_{cls}) — the between-class share of predicted-log P variance, $f_{\text{cls}} = \frac{\sum_c n_c(\bar{\hat{y}}_c - \bar{\hat{y}})^2}{\sum n \cdot \text{Var}(\hat{y})}$, which goes to 1.0 when predictions are constant within each class (pure class-template) and to 0.0 when predictions are uncorrelated with class; (iii) *within-class Spearman* (ρ_{within}) — median across classes of the rank correlation between true and predicted $\log P$ inside each class, which goes to 1.0 for an encoder that ranks sources within a class by their actual period. A model that has learned absolute time has high R^2 *and* low f_{cls} *and* ρ_{within} near 1. A model that has template-matched has high R^2, f_{cls} near 1, and ρ_{within} near 0.

The negative result. Every encoder we tested template-matches. Across the five methods with non-trivial R^2 (≥ 0.54) — BiGRU, 1D CNN, our CT-SSL transformer, pretrained Chronos, pretrained MOMENT — the class-driven fraction sits between 0.59 and 0.70. That is, two thirds of the apparent "time learning" is the encoder identifying the class and the probe predicting that class's mean period. Within-class Spearman ρ tops out at 0.26 (BiGRU) — a *weak* source-level period signal, far from the $\rho_{\text{within}} \to 1$ a genuine non-uniform Fourier learner would have. Pretrained Chronos, despite the highest overall R^2, has $\rho_{\text{within}} = 0.23$, lower than our cadence-as-channel BiGRU (0.26): the foundation model is *more* class-template-driven, not less. For Chronos-T5 retrained from scratch at matched scale, ρ_{within} collapses to 0.03 — no source-level period signal at all. Pretraining at 10^9 cross-domain timesteps gives Chronos a competent class-template encoder; it does not give it absolute time.

Fig. 1 makes this visible: predictions colored by catalog class form near-vertical stripes at each class's typical period — RR Lyrae (blue) at $\log_{10} P \approx -0.2$, δ-Sct (red) at $\log_{10} P \approx -0.7$, semiregulars (purple) at $\log_{10} P \approx 1.5$–

Table 1. Linear and k-NN probe balanced accuracy on the 25K-source 11-class subset subset, 11 classes. Best in bold; second-best italic. Pretrained baselines shaded grey for visual separation.

Method	Params	LogR bal-acc	kNN bal-acc
RF on Chen+ features	—	*0.882*	*0.882*
Pretrained Chronos-small	46M	0.6503	0.5611
Pretrained MOMENT-base	110M	0.5239	0.4580
Pretrained ASTROMER (frozen)	0.66M	0.327	0.263
CT-SSL (ours)	4.82M	0.5692	0.5043
CT-SSL no-disjoint ablation	4.82M	0.5341	0.4844
CT-SSL Fourier ablation	4.82M	0.5120	0.4851
BiGRU from scratch	4.42M	**0.6514**	**0.5956**
1D CNN from scratch	4.41M	0.6174	0.5598
Chronos-T5 from scratch	4.92M	0.3170	0.2340
PatchTST (fs)	∼5M	0.1500	0.0993

Table 2. Period regression on $\log_{10}(P)$. Conventional R^2 ranks methods as one would expect, but the diagnostic columns show *every* learned method we tested with non-trivial R^2 has $f_{\rm cls} \geq 0.60$ and $\rho_{\rm within} \leq 0.26$ — predictions are dominated by class identity, with weak source-level signal on top. Even the 46M pretrained Chronos checkpoint, the highest-R^2 method, has $\rho_{\rm within} = 0.23$. Chronos-T5 from scratch has $\rho_{\rm within} = 0.03$ — no source-level signal at all. The classical positive control — a 40-dimensional embedding built from the top-20 Lomb-Scargle peaks of each light curve — sits at $\rho_{\rm within} = 0.23$ under a linear probe. This is a probe limitation, not a label ceiling: the direct LS top peak reaches $\rho_{\rm within} = 0.78$ (Table 4), so the within-class signal is recoverable and the SSL encoders are under-extracting it.

Method	R^2	$f_{\rm cls}$	$\rho_{\rm within}$	MAE	Exact 5%	Alias 5%
Pretrained Chronos-small	0.685	0.70	0.23	0.369	3.92%	10.03%
BiGRU from scratch	**0.661**	0.66	**0.26**	**0.388**	3.86%	9.69%
1D CNN from scratch	0.601	0.60	0.25	0.418	3.62%	9.09%
Pretrained MOMENT-base	0.597	0.65	0.21	0.419	3.71%	9.11%
CT-SSL (ours, v3)	0.543	0.56	0.24	0.446	3.34%	8.76%
Pretrained ASTROMER (cross-survey)	0.399	0.45	0.118	0.504	3.48%	5.21%
Chronos-T5 from scratch	0.186	0.33	0.03	0.604	2.64%	6.80%
PatchTST (fs)	0.127	0.21	0.00	0.645	2.22%	6.50%
Classical positive control (hand-crafted period extractor):						
LS top-20 peaks (per source)	0.209	0.32	0.23	0.642	2.41%	7.13%

2.5 — with weak within-stripe ordering. A 0.5 d RR Lyrae and a 0.6 d RR Lyrae are mapped to nearly indistinguishable predicted periods; that's $\rho_{\rm within} = 0.26$ in pictures.

Robustness: a nonlinear probe does not rescue the within-class signal. A linear ridge probe cannot, by construction, recover period information that the encoder has stored nonlinearly. To test whether our negative result is an artifact of probe linearity (raised by both reviewers), we re-ran the full diagnostic with a nonlinear MLP probe (one 256-unit hidden layer, early stopping; same 5-fold protocol) on each method's frozen embedding. Table 4 reports both probes with bootstrap 95% confidence intervals. The MLP *does* extract more overall variance — BiGRU R^2 rises from 0.66 to 0.77, pretrained Chronos from 0.69 to 0.72, the 1D CNN from 0.60 to 0.71 — confirming the embeddings carry nonlinearly-decodable structure that the linear probe misses. But the extra variance is almost entirely *between-class*: $f_{\rm cls}$ stays at 0.70–0.75 and within-class ρ rises only marginally, topping out at 0.31 (BiGRU) and remaining at 0.27 for pretrained Chronos. The nonlinear probe sharpens class templates; it does not reveal a hidden source-level period code. The negative result is therefore a property of the encoders, not of the probe.

Nor is it an artifact of pooling. A second probe-side concern is that mean-pooling the final layer could discard period information held in an intermediate layer. Probing each of the CT-SSL transformer's six layers separately (Table 7, Appendix C) shows $\rho_{\rm within}$ rising monotonically but only from 0.19 (layer 0) to 0.25 (layer 5): no intermediate representation encodes period appreciably better than the pooled output we report.

A direct Lomb–Scargle ceiling: the signal is recoverable. The remaining question (Reviewer uEcd) is whether $\rho_{\rm within} \approx 0.26$ is close to a label-noise floor or far below what is recoverable. We answer it directly: running a Lomb–Scargle periodogram on each light curve and using its top peak as the prediction yields $\rho_{\rm within} = \mathbf{0.78}$ against the same catalog periods (Table 4, last row; $R^2 = 0.81$, 60.9% of sources within 5% of the catalog period, 22.3% at the

$P/2$ alias). A classical extractor recovers within-source period ordering far above every SSL encoder. Because the catalog periods are themselves LS-derived, 0.78 is best read as a near-ceiling for LS-style estimation rather than an absolute bound, but the conclusion is unambiguous in either reading: the within-class signal is present and recoverable from the photometry, and the SSL encoders capture roughly one-third of it. They are under-extracting, not saturating a noise floor.

A published astronomy-SSL checkpoint. To check that the result is not an artifact of our own reimplementations (Reviewer uEcd), we additionally embed the held-out subset with the released ASTROMER (Donoso-Oliva et al., 2023) checkpoint (the 0.66M single-band MACHO-pretrained model, frozen; flux→magnitude conversion and windowing in Appendix D) and run the identical probe. Applied cross-survey to ZTF, ASTROMER reaches $R^2 = 0.40$, $f_{\rm cls} = 0.45$, and $\rho_{\rm within} = 0.12$ (ridge; the MLP probe is statistically identical at $0.40/0.11$; Tables 1–4). Its modest overall R^2 partly reflects the MACHO→ZTF domain shift (different survey, cadence, and band), so we do not read it as a peer of the ZTF-trained encoders on overall accuracy. But on the question the diagnostic is built to answer it gives the same verdict as every other model: $\rho_{\rm within} = 0.12$ is *the lowest of any non-degenerate encoder we tested*, far below the recoverable 0.78 ceiling — no within-class time learning. A published, astronomy-specific SSL checkpoint does not escape the diagnostic; if anything, transferred to a new survey it template-matches more weakly, not more genuinely. A within-survey test would require pretraining ASTROMER on ZTF, which we leave to future work; our point is only that the released checkpoint, as a practitioner would download it, does not transfer within-class period structure to ZTF.

Diagnostic characterization on simple controls. Two controls confirm the diagnostic is responding to a real signal, not a measurement artifact (Fig. 2, Appendix A). *(i) Synthetic oracle.* An embedding $\hat{z}_i = \log_{10} P_i + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, recovers the expected response ($\rho_{\rm within} = 0.97, 0.89, 0.43$ at $\sigma = 0.05, 0.10, 0.50$ dex): the diagnostic scales mechanically with the period information present and returns $\rho_{\rm within} > 0.7$ when it is there. Read through this curve, the SSL cluster at $\rho_{\rm within} \approx 0.21$–$0.26$ sits near "catalog $\log P$ corrupted by ~ 0.7–1.0 dex of noise" — one order of magnitude in P. *(ii) Classical period extractor.* A hand-crafted 40-dim embedding consisting of the top-20 Lomb-Scargle peaks (each $[\log_{10} P_k, \log_{10} \text{power}_k]$) sits at $\rho_{\rm within} = 0.23$ on the same Chen+ 2020 sources (Table 2, last row) under a *linear* ridge probe. We initially read this as a label-noise ceiling, but the direct Lomb–Scargle extractor above ($\rho_{\rm within} = 0.78$, Table 4) shows otherwise: the low number here reflects the linear probe's inability to *select* the correct peak among 20 candidates, not a ceiling on recoverable period information. With the correct peak chosen (i.e. the LS top peak itself), within-class ordering is recovered at $\rho_{\rm within} = 0.78$. The classical signal is strong; it is the SSL encoders that fall short.

What the catalog labels mean. We predict the Chen et al. (2020) catalog's Lomb-Scargle periods as labels rather than running LS ourselves, so the experiment measures how well an encoder aligns with the catalog's LS-validated period, not de novo recovery from photometry. The catalog has $f_{\rm cls} = 0.84$ at zero label noise — 84% of log-P variance is between-class, an inherent property of classes spanning ~ 3.5 orders of magnitude in mean period — but the within-class ceiling is *not* near 0.23: the direct LS extractor reaches $\rho_{\rm within} = 0.78$ on these same labels (Table 4), so the recoverable within-class headroom is real and the SSL encoders are not capturing it. An encoder that mirrors the catalog also inherits its window-function biases ($P/2$ aliases for symmetric EBs, daily-cadence aliases).

5.3. Architectural fairness control: cross-domain pretraining gives class templates, not time

Is even the class-template ability a property of the T5 architecture on ZTF, or is it acquired during cross-domain pretraining? We retrain Chronos-T5 from scratch at parameter-matched scale (4.92M) on the same 400K-source ZTF pool, with identical tokenizer, objective, and probe protocol.

Chronos-from-scratch collapses on every metric: classification balanced accuracy $0.6503 \to 0.3170$, period R^2 $0.685 \to 0.186$, $f_{\rm cls}$ $0.70 \to 0.33$, and $\rho_{\rm within}$ $0.23 \to 0.03$. Validation cross-entropy plateaus within ~ 1 epoch at ≈ 6.42 nats (uniform prior over the 4096-bin tokenizer is $\ln 4096 = 8.32$ nats).

The reading is clean: with neither t nor σ available and a next-token CE dominated by per-token photometric noise, the from-scratch run captures only the marginal flux-distribution shape. Cross-domain pretraining endows the released checkpoint with generic time-series structure (heteroskedasticity, autocorrelation, trend/seasonality) enough to template-match ZTF classes — but its $\rho_{\rm within} = 0.23$ is no better than the cadence-as-channel BiGRU's with no pretraining at all. **Cross-domain pretraining buys class templates; it does not buy time.**

5.4. Cadence-as-channel: handing the model its phase clock

Our three from-scratch contrastive encoders share an objective (Eq. 1) and pretraining pool but differ in how time enters: the CNN and BiGRU consume $\log_{10} \Delta t$ as a per-step input channel, while the CT-SSL transformer encodes time via random Fourier features plus a pairwise continuous-

time attention bias. The cadence-as-channel encoders beat CT-SSL on classification (0.65 BiGRU, 0.62 CNN vs. 0.57 LogR) and on overall period R^2 — but on the diagnostic the three are tied ($\rho_{\text{within}} = 0.26/0.25/0.24$): an explicit cadence channel tightens class templates without moving the within-class metric. The architecture gap is thus not load-bearing for the main claim; all three land in the same template-matching regime.

6. Discussion

Why we trust the diagnostic. The two natural objections — that the catalog labels cap ρ_{within}, or that a linear probe hides a nonlinear period code — are both ruled out empirically in §5.2: a direct LS extractor reaches $\rho_{\text{within}} = 0.78$ on these same labels (so the headroom is real, not label noise), and a nonlinear MLP probe still leaves $\rho_{\text{within}} \leq 0.31$ (so the signal is largely absent from the embeddings, not merely linearly inaccessible).

Open questions. The information needed is in the data — Lomb-Scargle extracts it in closed form — so the empirical question is whether SGD on a self-supervised loss converges there at realistic scales. The natural follow-ups: whether astronomy-pretrained foundation models at 10^7+ sources (Zuo et al., 2026; Zhang et al., 2024) move ρ_{within} where our matched-scale controls cannot; and whether period-sensitive objectives (period-shifted contrastive negatives, a differentiable phase-fold head, photometry–spectroscopy alignment) push the encoder beyond template matching — each a candidate the diagnostic is designed to score.

What this changes for the field. The diagnostic itself — f_{cls} and ρ_{within} alongside overall R^2 — is the contribution we expect to be most useful: it converts a misleading scalar into a 2-axis readout separating class-template matching from genuine time learning. Astronomy-SSL papers reporting period R^2 as evidence of time encoding are, on our data, reading off class-template structure any competent classifier inherits for free.

7. Limitations

1. **Encoder-training-seed robustness.** All metrics carry bootstrap 95% CIs over 5 probe data-splits (Table 4; split-level ρ_{within} s.d. ≤ 0.02). Retraining the BiGRU with 4 encoder seeds gives $\rho_{\text{within}} = 0.245 \pm 0.018$, $f_{\text{cls}} = 0.656 \pm 0.009$ — not a seed artifact, though retraining all eight encoders at multiple seeds remains future work. Single labeled subset, single survey; cross-survey transfer (TESS, ATLAS) not tested.

2. **ASTROMER is evaluated cross-survey.** The one published astronomy-SSL checkpoint we run (ASTROMER (Donoso-Oliva et al., 2023)) was pretrained on MACHO and applied to ZTF without adaptation; a within-survey test would require pretraining it on ZTF (future work). Its low overall R^2 partly reflects this transfer, though $\rho_{\text{within}} = 0.12$ still indicates no within-class signal.

3. Architectural-fairness retraining is at $N{=}1$ architecture (Chronos-T5). The from-scratch CNN and BiGRU *do* ingest explicit time channels and still fail the within-class test, separating the "missing time input" confound from the "missing pretraining" one. Whether other foundation-model architectures (MOMENT, PatchTST, TimesFM) collapse to the same ceiling at matched scale is the natural follow-up.

4. BiGRU best is reported at epoch 6; further training may move the number ± 0.01 bal-acc. The within-class conclusion holds under both linear and MLP probes (Table 4); a probe with explicit period-inference structure could extract still more, but would no longer separate "the encoder learned time" from "the probe can extract it."

8. Conclusion

Period regression on a multi-class catalog looks like a strong test of learned time encoding, but the conventional R^2 conflates identifying the class and predicting its mean period with reading period off an individual source. Decomposing predicted-log-P variance into between- and within-class components separates these. Across eight encoders, the diagnostic returns the same answer: 60–70% of period R^2 is class-template matching and ρ_{within} tops out at 0.26 — even pretrained Chronos sits at 0.23 — while a direct Lomb–Scargle extractor reaches 0.78 on the same labels. None of the encoders we tested have learned absolute time on irregular ZTF photometry. We propose that future astronomy-SSL work report the within-class diagnostic alongside overall period R^2, and treat ρ_{within} as the metric of merit for the time-encoding question.

Acknowledgements

The authors thank Jeroen Audenaert for the dual-encoder structured-contrastive framework underlying the disentangled loss and for manuscript feedback. This work used the MIT ORCD/Engaging cluster.

References

Aerts, C., Christensen-Dalsgaard, J., and Kurtz, D. W. *Asteroseismology*. 2010. doi: 10.1007/978-1-4020-5803-5.

Allam, Tarek, J. and McEwen, J. D. Paying attention to astronomical transients: introducing the time-series

transformer for photometric classification. *RAS Techniques and Instruments*, 3(1):209–223, 01 2024. ISSN 2752-8200. doi: 10.1093/rasti/rzad046. URL https://doi.org/10.1093/rasti/rzad046.

Ansari, A. F., Stella, L., Turkmen, A. C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, B. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. Expert Certification.

Audenaert, J., Muthukrishna, D., Gregory, P. F., Hogg, D. W., and Villar, V. A. Causal Foundation Models: Disentangling Physics from Instrument Properties. *arXiv e-prints*, art. arXiv:2507.05333, July 2025. doi: 10.48550/arXiv.2507.05333.

Bellm, E. The Zwicky Transient Facility. In Wozniak, P. R., Graham, M. J., Mahabal, A. A., and Seaman, R. (eds.), *The Third Hot-wiring the Transient Universe Workshop*, pp. 27–33, January 2014. doi: 10.48550/arXiv.1410.8185. URL https://arxiv.org/abs/1410.8185.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001. doi: 10.1023/A:1010933404324.

Cabrera-Vives, G., Moreno-Cartagena, D., Astorga, N., Reyes-Jainaga, I., Förster, F., Huijse, P., Arredondo, J., Muñoz Arancibia, A. M., Bayo, A., Catelan, M., Estévez, P. A., Sánchez-Sáez, P., Álvarez, A., Castellanos, P., Gallardo, P., Moya, A., and Rodriguez-Mancini, D. Atat: Astronomical transformer for time series and tabular data. *A&A*, 689:A289, 2024. doi: 10.1051/0004-6361/202449475. URL https://doi.org/10.1051/0004-6361/202449475.

Chen, X., Wang, S., Deng, L., de Grijs, R., Yang, M., and Tian, H. The zwicky transient facility catalog of periodic variable stars. *The Astrophysical Journal Supplement Series*, 249(1):18, jul 2020. doi: 10.3847/1538-4365/ab9cae. URL https://doi.org/10.3847/1538-4365/ab9cae.

Ding, J., Chen, X., Gao, X., Tang, X., Wang, S., Huang, Y., Qi, X., Xue, G., Luo, A., and Liu, J. Starclr: Contrastive learning representation for astronomical light curves. *arXiv preprint arXiv:2604.24516*, 2026. doi: 10.48550/arXiv.2604.24516. URL https://arxiv.org/abs/2604.24516. Accepted for publication in Astronomy & Astrophysics.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., and Vardhan, H. ASTROMER. A transformer-based embedding for the representation of light curves. *A&A*, 670:A54, February 2023. doi: 10.1051/0004-6361/202243928. URL https://doi.org/10.1051/0004-6361/202243928.

Donoso-Oliva, Cristobal, Becker, Ignacio, Protopapas, Pavlos, Cabrera-Vives, Guillermo, Cádiz-Leyton, Martina, and Moreno-Cartagena, Daniel. Generalizing across astronomical surveys: Few-shot light curve classification with astromer 2. *A&A*, 707:A170, 2026. doi: 10.1051/0004-6361/202554026. URL https://doi.org/10.1051/0004-6361/202554026.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c. URL https://doi.org/10.3847/1538-4357/ab042c.

Lomb, N. R. Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophys. Space Sci.*, 39(2):447–462, February 1976. doi: 10.1007/BF00648343.

Masci, F. J., Laher, R. R., Rusholme, B., Shupe, D. L., Groom, S., Surace, J., Jackson, E., Monkewitz, S., Beck, R., Flynn, D., Terek, S., Landry, W., Hacopians, E., Desai, V., Howell, J., Brooke, T., Imel, D., Wachter, S., Ye, Q.-Z., Lin, H.-W., Cenko, S. B., Cunningham, V., Rebbapragada, U., Bue, B., Miller, A. A., Mahabal, A., Bellm, E. C., Patterson, M. T., Jurić, M., Golkhou, V. Z., Ofek, E. O., Walters, R., Graham, M., Kasliwal, M. M., Dekany, R. G., Kupfer, T., Burdge, K., Cannella, C. B., Barlow, T., Sistine, A. V., Giomi, M., Fremling, C., Blagorodnova, N., Levitan, D., Riddle, R., Smith, R. M., Helou, G., Prince, T. A., and Kulkarni, S. R. The zwicky transient facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific*, 131(995):018003, dec 2018. doi: 10.1088/1538-3873/aae8ac. URL https://doi.org/10.1088/1538-3873/aae8ac.

Moreno-Cartagena, D., Cabrera-Vives, G., Protopapas, P., Donoso-Oliva, C., Pérez-Carrasco, M., and Cádiz-Leyton, M. Positional encodings for light curve transformers: Playing with positions and attention. In *ICML 2023 Workshop on Machine Learning for Astrophysics*, 2023. doi: 10.48550/arXiv.2308.06404. URL https://arxiv.org/abs/2308.06404. PMLR 202, Honolulu, Hawaii, USA.

Morvan, M., Nikolaou, N., Yip, K. H., and Waldmann, I. P. Don't pay attention to the noise: Learning self-supervised representations of light curves with a denoising time series transformer. In *Machine Learning for Astrophysics Workshop at ICML 2022*, 2022. URL https://arxiv.org/abs/2207.02777. arXiv:2207.02777.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Pan, J.-S., Ting, Y.-S., and Yu, J. Astroconformer: The prospects of analysing stellar light curves with transformer-based deep learning models. *Monthly Notices of the Royal Astronomical Society*, 528(4): 5890–5903, 03 2024a. ISSN 0035-8711. doi: 10.1093/mnras/stae068. URL https://doi.org/10.1093/mnras/stae068.

Pan, J.-S., Ting, Y.-S., Yu, J., Huang, Y., and Liu, J.-F. The scaling law in astronomical time series data. In *ICML 2024 AI for Science Workshop*, 2024b. URL https://openreview.net/forum?id=2fGbh0f15l.

Scargle, J. D. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *ApJ*, 263:835–853, December 1982. doi: 10.1086/160554.

Zhang, G., Helfer, T., Gagliano, A. T., Mishra-Sharma, S., and Ashley Villar, V. Maven: a multimodal foundation model for supernova science. *Machine Learning: Science and Technology*, 5(4):045069, dec 2024. doi: 10.1088/2632-2153/ad990d. URL https://doi.org/10.1088/2632-2153/ad990d.

Zuo, X., Tao, Y., Huang, Y., Kang, Z., Chen, H., Cui, C., Pan, J., Kong, X., Ting, Y.-S., Tang, X., Han, H., Mu, H., Xu, Y., Fan, D., Xue, G., Luo, A., and Liu, J. FALCO: Foundation Model of Astronomical Light Curves for Time Domain Astronomy. Implementation and Applications on Kepler Data. *AJ*, 171(1):10, January 2026. doi: 10.3847/1538-3881/ae1467. URL https://doi.org/10.3847/1538-3881/ae1467.

A. Architecture, data, and probe-comparison details

Table 3 lists the encoder zoo; Fig. 3 shows a representative ZTF light curve and the aliasing that makes period recovery hard; Fig. 4 visualizes the classification / period-R^2 ranking summarized in Tables 1 and 2.

Table 3. Architecture zoo. Pretrained foundation models shown for context.

Model	Params	Pretrain	Inputs
CT-SSL (ours)	4.82M	400K–1M ZTF	t, f, σ + mask
CNN (1D, 6 layers)	4.41M	400K ZTF	$f, \log\sigma, \log\Delta t$
BiGRU (3 layers)	4.42M	400K ZTF	$f, \log\sigma, \log\Delta t$
Chronos-T5 (fs)	4.92M	1M ZTF	f (univariate)
PatchTST (fs)	$\sim$ 5M	100K ZTF	t, f, σ
Chronos-small (frozen)	46M	$\sim 10^9$ generic	f
MOMENT-base (frozen)	110M	$\sim 10^9$ generic	f (regular grid)

B. Per-class confusion matrices and breakdowns

Tables 5 and 6 report per-class precision/recall/F1 for the two from-scratch baselines that fit in the body. The 11 classes are alphabetically: BYDra, CEP, CEPII, DSCT, EA, EW, Mira, RR, RRc, RSCVN, SR.

C. Per-layer period probe

To rule out that mean-pooling the final layer discards period information present in an intermediate representation, we run the ridge-probe diagnostic on each of the CT-SSL transformer's six layers separately (mean-pooled per layer; same protocol as §3.4). Table 7 shows ρ_{within} increasing monotonically with depth but saturating at 0.25 — no layer recovers source-level period appreciably better than the pooled output reported in the main text.

D. ASTROMER embedding procedure

We evaluate the released ASTROMER checkpoint (`macho_a0`: a 2-layer, d=256, single-band encoder, 0.66M parameters, pretrained on MACHO R-band) on the same 25,565-source held-out set. ASTROMER consumes single-band $(t, \text{mag}, \sigma_{\text{mag}})$ sequences, so we convert each ZTF light curve from the cached flux stream via $\text{mag} = -2.5\log_{10} f$ and $\sigma_{\text{mag}} = (2.5/\ln 10)\,\sigma_f/f$, dropping non-positive (forced-photometry) flux; all 25,565 sources retain $\geq$ 10 valid epochs (median 502). We feed sources through ASTROMER's encoder with its native windowing ($\leq$ 200 observations per window), group windows back to sources via object IDs, and mean-pool the per-observation attention output to a 256-dim per-source embedding, which then enters the identical probe protocol (§3.4). Because ASTROMER was pretrained on a different survey (MACHO), this is a cross-survey transfer; its modest overall R^2 is expected, but its $\rho_{\text{within}} = 0.12$ speaks directly to the within-class time-learning question regardless of overall scale.

Table 4. **Linear vs. nonlinear probe, with 95% CIs.** Period regression on $\log_{10}(P)$ under a linear ridge probe and a nonlinear MLP probe. Point estimates are means over 5 data-split seeds ($\pm$ standard deviation); brackets are 2.5–97.5% bootstrap intervals over sources. The MLP raises overall R^2 but leaves ρ_{within} weak (≤ 0.31) and f_{cls} high, so the missing signal is in the encoders, not the probe. The direct Lomb–Scargle extractor (bottom) reaches $\rho_{\text{within}} = 0.78$, establishing that the within-class signal is recoverable. Encoder training seed is fixed at 42 (see §7); CIs quantify split- and sample-level uncertainty.

	Ridge (linear)		MLP (nonlinear)	
Method	R^2	ρ_{within}	R^2	ρ_{within}
Pretrained Chronos-small	0.686	0.228 [0.20, 0.26]	0.718	0.274 [0.26, 0.32]
BiGRU (ours)	0.661	0.258 [0.22, 0.29]	0.772	**0.308** [0.28, 0.34]
1D CNN (ours)	0.601	0.252 [0.22, 0.27]	0.710	0.293 [0.27, 0.33]
Pretrained MOMENT-base	0.596	0.206 [0.18, 0.25]	0.587	0.214 [0.20, 0.26]
CT-SSL (ours)	0.543	0.237 [0.20, 0.28]	0.606	0.290 [0.23, 0.29]
Pretrained ASTROMER (xsurv.)	0.399	0.118	0.395	0.112
Chronos-T5 (from scratch)	0.187	0.036 [0.01, 0.06]	0.172	0.037 [0.02, 0.07]
PatchTST (fs)	0.127	0.002 [-0.02, 0.08]	−0.010	0.015 [-0.00, 0.04]
Classical reference ceiling (direct period extractor on the same light curves):				
LS top peak (per source)		—	0.810	**0.779**

Table 5. Chronos-T5 from scratch, epoch-10 best checkpoint. Probe protocol in Section 3.4.

Class	P	R	F1	N
BYDra	0.234	0.200	0.216	3000
CEP	0.209	0.042	0.069	1132
CEPII	0.000	0.000	0.000	316
DSCT	0.276	0.276	0.276	3000
EA	0.666	0.745	0.703	3000
EW	0.323	0.387	0.352	3000
Mira	0.474	0.077	0.132	117
RR	0.460	0.540	0.497	3000
RRc	0.466	0.659	0.546	3000
RSCVN	0.198	0.120	0.150	3000
SR	0.412	0.443	0.427	3000

E. Compute budget

The full benchmark sits comfortably under 200 GPU-h on a small academic cluster, well below the $\sim 10^4$ GPU-h required to pretrain a Chronos- small from scratch on the original cross-domain corpus.

F. Per-model training hyperparameters

Table 9 consolidates the per-model training configuration described in prose in §4, one row per model (addressing a reviewer request). All from-scratch encoders use seed 42 and patience-10 early stopping on validation loss; the contrastive baselines share the InfoNCE+qual loss of Eq. 1 (τ=0.07, λ_q=1.0).

Table 6. BiGRU from scratch, epoch-6 best checkpoint (val 1.8747). Note that precision/recall/F1 are well above Chronos-T5 from-scratch across every class — the gap is broad, not driven by a single dominant class.

Class	P	R	F1	N
BYDra	0.567	0.582	0.574	3000
CEP	0.696	0.648	0.672	1132
CEPII	0.452	0.149	0.224	316
DSCT	0.818	0.850	0.834	3000
EA	0.871	0.868	0.869	3000
EW	0.668	0.683	0.676	3000
Mira	0.714	0.513	0.597	117
RR	0.798	0.823	0.810	3000
RRc	0.723	0.788	0.754	3000
RSCVN	0.405	0.337	0.368	3000
SR	0.865	0.924	0.894	3000

Table 7. Per-layer ridge-probe diagnostic for the CT-SSL transformer (6 layers). ρ_{within} peaks at the final layer (0.25); no intermediate layer hides a stronger period signal.

Layer	R^2	f_{cls}	ρ_{within}
0	0.402	0.446	0.193
1	0.424	0.470	0.192
2	0.445	0.488	0.208
3	0.454	0.496	0.234
4	0.456	0.498	0.240
5 (pooled)	0.460	0.504	0.254

Table 8. Wall-clock and GPU-hour costs.

Run	Hardware	Wall (h)	GPU-h
CT-SSL v5 (1M sources, 5 epochs)	4×A100-80	12	48
Chronos-T5 from-scratch (10 epochs)	4×A100-80	16	64
CNN from-scratch (∼10 epochs)	4×L40S	12	48
BiGRU from-scratch (∼10 epochs)	4×A100-80	9	36
Probes (each)	1×L40S	0.05	0.05
Total reported in main results		49	196

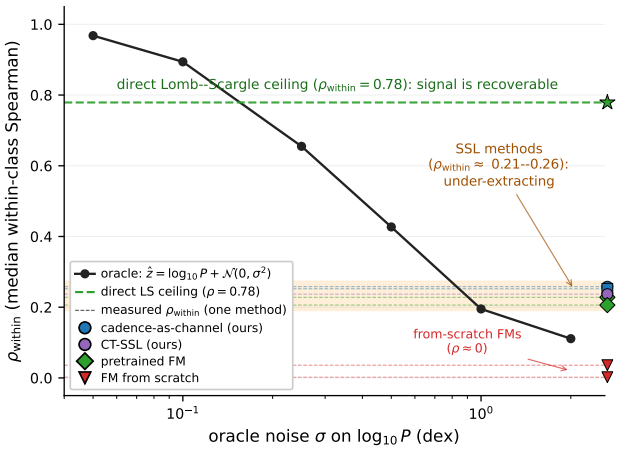

Figure 2. **Diagnostic response to known period information.** Solid black curve: ρ_{within} for the synthetic oracle embedding $\hat{z} = \log_{10} P + \mathcal{N}(0, \sigma^2)$ as a function of noise level σ (dex on $\log P$; star at left marks the $\sigma = 0$ point, $\rho_{\text{within}} = 1.0$). Dashed horizontal lines: measured ρ_{within} for each tested method (markers at right, each labelled). The cluster of SSL methods at $\rho_{\text{within}} \approx 0.21$–$0.26$ corresponds to "perfect period information corrupted by ~ 0.7–1.0 dex of noise." Chronos-T5 from-scratch and PatchTST sit near zero, indicating no source-level period signal at all. The upper dashed green line marks the direct Lomb–Scargle ceiling ($\rho_{\text{within}} = 0.78$): the achievable within-class signal sits far above every SSL encoder, so the methods are under-extracting rather than saturating a label-noise floor.

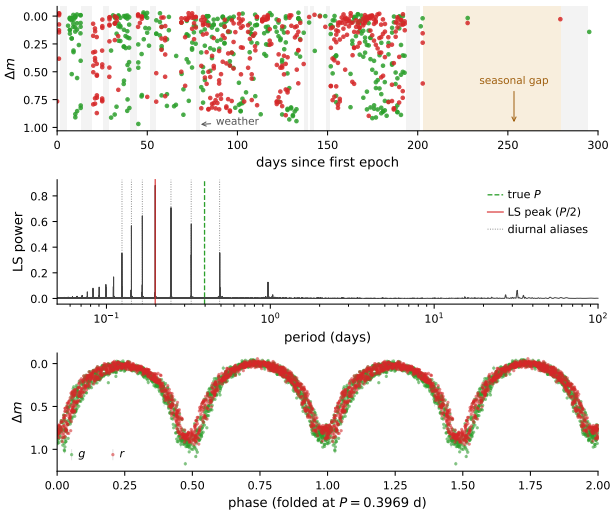

Figure 3. One ZTF eclipsing-binary light curve in g (green) and r (red), forced photometry. *Top:* the first 300 days, in differential magnitude relative to peak (axis inverted; brighter is up). Grey bands mark within-season gaps of 3–15 days (weather, lunar avoidance, scheduling); the orange band is the seasonal exit when the source sets behind the Sun (~ 50 d in this window, several months in subsequent seasons). *Middle:* Lomb-Scargle periodogram with n_{terms}=1 on the full ~ 5.4-yr baseline (time-centred input, $df = 1/(3 \times \text{baseline})$, $f_{\min} = 4/\text{baseline}$; cadence-alias mask deliberately disabled so the window-function structure is visible). The dominant peak sits at $P/2$, not the catalog P, because the eclipsing geometry produces two minima per orbital cycle and the 1- and 2-harmonic Fourier basis cannot distinguish that from a single dip at half the period; secondary peaks at 1 d and 0.5 d are pure window-function aliases. Recovering the true P requires raising $n_{\text{terms}} \geq 4$ a priori — BLS, prewhitening, sigma-clipping, conditional entropy, and PDM all return the half-period alias on this source. *Bottom:* all ~ 3000 epochs phase-folded at the catalog period (no binning); two consecutive eclipses per cycle confirm the orbital period despite the irregular sampling. **The encoder receives only the per-epoch tuple** $(\text{flux}, \log_{10} \sigma, \log_{10} \Delta t)$ **— never the periodogram, period, phase, or band identity.**

Table 9. Per-model training hyperparameters. "eff. batch" is the DDP-aggregated batch over 4 GPUs (single-process for PatchTST); LR is the AdamW peak after warmup-cosine (cosine-annealing for PatchTST). BiGRU uses fp32 to avoid recurrent-gradient overflow; all others bf16.

Model	LR	eff. batch	epochs	prec.
CT-SSL (ours)	3×10^{-4}	384	≤ 40	bf16
1D CNN	3×10^{-4}	384	≤ 40	bf16
BiGRU	1×10^{-4}	384	≤ 40	fp32
Chronos-T5 (fs)	3×10^{-4}	1024	≤ 40	bf16
PatchTST (fs)	1×10^{-3}	64	20	bf16

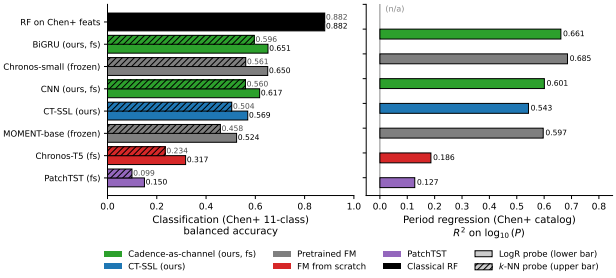

Figure 4. **Two downstream tasks, same ranking.** *Left:* linear-probe (LogR, solid lower bars) and k-NN (hatched upper bars) balanced accuracy on the 11-class subset of Chen et al. (2020). *Right:* ridge-probe R^2 for $\log_{10}(P)$ on the same sources. The method ordering is preserved across tasks, but §5.2 shows the two top-tier methods are equivalent in template matching, not in time learning.