PHASE DEPENDENCE IS A TASK PROPERTY: MEASURING WHAT CLASSIFIERS GAIN BEYOND THE FOURIER MAGNITUDE IN TIME-SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex-valued and other phase-aware architectures are regularly proposed for time-series tasks, yet their reported benefits are inconsistent. We argue the question "are phase-aware *models* better?" is underdetermined: phase sensitivity is first a measurable property of *tasks*. Because $x \mapsto |\mathcal{F}x|$ is not injective—its fibers contain all circular shifts, reflections and negations—any amplitude-spectrum classifier is blind to label structure varying within fibers, so the Bayes accuracy reachable from the global amplitude spectrum is a true ceiling for that class. We measure not that ceiling but how much of it classifiers reach—the *phase-accessibility gap* Δ_ϕ: the test-accuracy difference between the best model given the full signal and given only the amplitude spectrum, with model and representation selected by cross-validation on training data alone. It screens a model-agnostic target with a calibrated model family and is stable across disjoint estimator families. Against exact synthetic ground truth it gives no false positives in 30 amplitude-only tasks or 36 phase-uninformative nulls, but 11 of 12 on *redundancy* nulls—the label a function of $|\mathcal{F}x|$ yet the phase encoding it—so what it identifies is a finite-family bound, not a Bayes gap. It decomposes, relative to an explicit shift-invariant repertoire and up to a measured estimator term, into a *position* component a fixed-window spectrogram largely recovers (81–86%) and a *relative-phase* component it recovers much less of (54–66%).

Across 114 UCR datasets the estimated gap spans -0.20 to 0.53. Estimated *from training data alone* it predicts held-out gain across five model families with raw-signal access (1D CNN, MiniRocket, InceptionTime, a complex-valued CNN, 1-NN DTW; partial correlations 0.52–0.68 against untouched test performance, family-clustered intervals excluding zero) and is near zero where the gap is; an intervention randomizing phase while preserving amplitude drives the gap and an independent family's gain down together ($r = +0.91$). On the UEA multivariate archive, which took no part in its development, it again predicts held-out gain ($r = +0.70, +0.62$) and sorts it into the same semantic categories. Per dataset it is a screen: under train-selection uncertainty 33 of 114 datasets are confidently positive, 24 show no detected gap and 53 are inconclusive. As a case study, a real CNN given the analytic-signal input $[x, \mathcal{H}x]$ matches or beats its complex counterpart on the trunk and budget we test, so what helped *here* was a phase-bearing input, not arithmetic. The literature's inconsistency is what a task property, interacting with input and readout, predicts.

1 INTRODUCTION

A recurring proposal in time-series learning is to make models *phase-aware*—complex-valued networks carrying magnitude and phase through every layer (Trabelsi et al., 2018), spectral front-ends that retain the phase channel, position-sensitive architectures—on a decades-old intuition: in natural signals, Fourier phase carries most of the perceptually and semantically important structure (Oppenheim & Lim, 1981). Yet the record is inconsistent: complex networks beat real ones on some tasks

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.

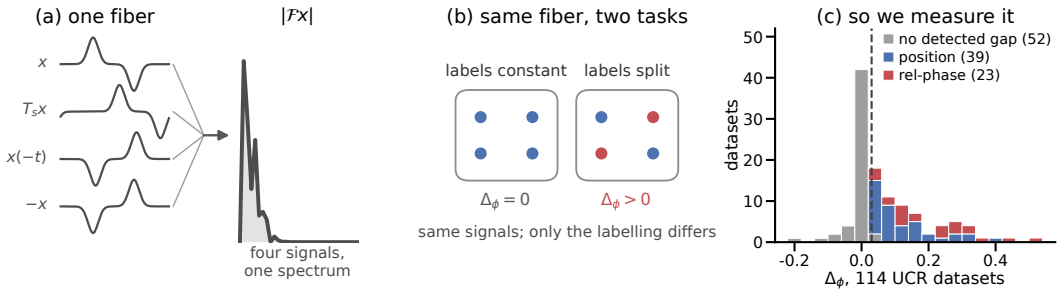

Figure 1: **Phase dependence is a property of the task, not of the signal. (a)** $x \mapsto |\mathcal{F}x|$ is many-to-one: a series, a circular shift of it, its time reversal and its negation look nothing alike yet share one amplitude spectrum (agreement to 10^{-14}), and every amplitude-only classifier is constant on such a fiber—Proposition 1, drawn. **(b)** Whether that costs anything depends on the *labelling*: constant on the fiber, the magnitude loses nothing; split within it, no amplitude-only model recovers the label at any capacity. The signals are the same in both boxes. **(c)** Estimating that quantity across 114 UCR datasets. Panels (a–b) are constructed: the claim is an identity, not a regularity.

and trail on others (Mönning & Manandhar, 2018), and domain gains (Dramsch et al., 2021) fail to transfer.

The question is underdetermined as posed, because it locates phase sensitivity in the *model*. Whether phase awareness can help is bounded by a logically prior quantity: *how much of the task's label information depends on phase at all*—a property of the joint distribution of signals and labels, measurable without committing to an architecture. Once measured it varies so widely (zero to fifty-plus accuracy points, §3) that averaging a model comparison over an arbitrary basket must produce unstable conclusions—the signature of a task property treated as a model property.

Our starting point is an algebraic fact (Figure 1a). The amplitude spectrum $x \mapsto |\mathcal{F}x|$ is not injective: circular shifts, reflection, and negation all preserve it, and in 1D the ambiguity is larger still (Beinert & Plonka, 2015; Shechtman et al., 2015). So *any* classifier factoring through amplitude—equivalently power spectrum or autocorrelation—is constant on each fiber, and within-fiber label structure is invisible to it *in principle*; the best amplitude-reachable Bayes accuracy is a true ceiling for that class. We operationalize not that distance but how much of it classifiers reach, the **phase-accessibility gap** Δ_ϕ—estimated with a fixed family of generic models under train-side CV selection with a single held-out evaluation, calibrated on synthetic tasks with known ground truth (§2, §6)—decomposed into a *position* and a *relative-phase* component—absolute location versus shift-invariant phase such as event *order*. What makes this a *task* property is that the same signals carry a gap or not depending only on how they are labelled (Figure 1b).

The empirical teeth: does a task's phase-accessibility gap, measured once with generic models, *predict* how much other model families gain over the amplitude spectrum? Across 114 fixed-length UCR datasets (Dau et al., 2019) and five model families—a 1D CNN, MiniRocket (Dempster et al., 2021), InceptionTime (Ismail Fawaz et al., 2020), a capacity-matched complex-valued CNN and training-free 1-NN DTW—the answer is yes for all five (partial correlations 0.53–0.77 controlling for length and class count), with gains near zero on the low-estimated-gap stratum ($\Delta_\phi < 0.03$) for four of the five, as predicted. The complex arm is the most instructive, and resolves into input and readout rather than arithmetic (§5): real-valued models on raw signals also access phase, so "complex vs. real" is not the axis along which phase value is realized—a concrete re-reading of the mixed results of Mönning & Manandhar (2018).

Contributions. (1) A reframing with a formal anchor: phase dependence as a per-task quantity, grounded in the non-injectivity of the amplitude representation, which makes the amplitude-reachable *Bayes* accuracy a principled ceiling. (2) A practical diagnostic—a model-family estimator of a model-agnostic target relative to the global Fourier magnitude—the phase-accessibility gap Δ_ϕ and its decomposition $\Delta_\phi = \Delta_{\text{align}} + \Delta_{\text{rel}} + \varepsilon$, calibrated on planted ground truth, including why the popular phase-scrambling surrogate must not be the amplitude reference (§2, §6). (3) To our knowledge the first task-side, repertoire-explicit phase-accessibility map of a major benchmark: across 114 UCR datasets the gap is strongly non-uniform (46% with no estimated gap above 0.03; tail to 0.53)

and the typology surfaces, with zero domain knowledge, tasks whose class semantics are defined by temporal organization, with a seven-dataset core led by TwoPatterns (§3). (3b) Per-dataset intervals that count training resampling and full re-selection: 33 of 114 datasets are confidently positive, so single-dataset use is a screen. (4) Prospective and external validation: estimated from training data alone, the gap predicts raw-signal gains across five families and on the unseen UEA archive, and the complex comparison resolves into input and readout rather than arithmetic (§4–§5).

2 A MODEL-FAMILY-CALIBRATED DIAGNOSTIC OF PHASE DEPENDENCE

A task is a distribution over pairs (x, y), $x \in \mathbb{R}^L$ a fixed-length, z-normalized series, y a class label, using each benchmark's standard split. Per-series normalization is part of the task definition: sufficiency claims are relative to the normalized task, scale and offset being removed for every arm equally. Let $\mathcal{F}$ be the discrete Fourier transform, $\mathrm{amp}(x) = |\mathcal{F}x| \in \mathbb{R}^{\lfloor L/2 \rfloor + 1}$ the *amplitude representation* and $[\,|\mathcal{F}x|, \cos\angle\mathcal{F}x, \sin\angle\mathcal{F}x\,]$ the *phase-augmented* one.

Proposition 1 (Amplitude fiber)**.** *For any $x \in \mathbb{R}^L$, any integer s, the circular shift T_s, the time reversal R and the negation $-x$ satisfy* $\mathrm{amp}(T_s x) = \mathrm{amp}(Rx) = \mathrm{amp}(-x) = \mathrm{amp}(x)$. *Any predictor $h = g \circ \mathrm{amp}$ is therefore constant on the orbit of x, so with $\eta_y(x) = P(Y{=}y \mid X{=}x)$, $Z = \mathrm{amp}(X)$, $A^*_{\mathrm{amp}} = \mathbb{E}\max_y \mathbb{E}[\eta_y(X) \mid Z] \le \mathbb{E}\max_y \eta_y(X) = A^*_{\mathrm{raw}}$ by Jensen's inequality, with* strict *inequality exactly when the conditional Jensen gap is positive on a set of positive probability—equivalently, when the argmax sets* $\arg\max_y \eta_y(x)$ *have empty intersection over a positive-measure set of fibres. For deterministic labels that is exactly: two points of a* common fibre *carry different labels with positive probability (orbit-level disagreement is a sufficient, not a necessary, witness).*

The shift theorem multiplies $F[k]$ by a unimodular factor, reversal conjugates it and negation flips its sign, so the modulus is unchanged. Strictness needs the argmax condition and not merely a varying Bayes *decision*: if some label maximises η_y everywhere on a fibre, an amplitude-only predictor can return it and loses nothing, even though the selected Bayes decision may differ across the fibre. The measure-theoretic statement and proof are in Appendix A. Two scope notes carry through the paper: the claim concerns the *global* DFT amplitude, the representation that amplitude/power-spectrum classifiers actually consume, not redundant magnitude measurements such as the STFT (Jaganathan et al., 2016); and it asserts the non-separability of a *representation*, not the impossibility of phase retrieval.

Proposition 1 concerns Bayes-optimal predictors: it defines A^*_{amp}, the Bayes accuracy reachable from $|\mathcal{F}x|$ and a true ceiling for every amplitude/power method, and a Bayes-level gap Δ^*_ϕ to the full-signal Bayes accuracy. Starred symbols are Bayes-level, unstarred ones empirical—formally $\Delta_\phi \equiv \Delta_{\phi;\,\mathcal{M},\mathcal{R}}$, the dependence on $\mathcal{M}$ and $\mathcal{R}$ suppressed in notation but never in interpretation. We name Δ_ϕ the *phase-accessibility gap*: accuracy a finite family can *reach* beyond $|\mathcal{F}x|$, not information provably absent from it—and part of it (position) localized magnitude recovers (§6). Three quantities stay apart: the *estimand* Δ^*_ϕ, the *estimator* Δ_ϕ, and the *predictive utility*—how well Δ_ϕ forecasts real gains (§4). The proposition anchors the first; it does not make the second consistent for it: Δ_ϕ can exceed zero when $\Delta^*_\phi = 0$ if the same magnitude information is merely harder for $\mathcal{M}$ to learn from $|\mathcal{F}x|$ than from raw—a representation-learnability confound the audits of §6 bound but cannot eliminate. Estimates use a fixed family $\mathcal{M}$ of strong generic classifiers under a protocol in which the test set is never used for any selection (Appendix B); writing $A[\mathcal{R}]$ for the test accuracy of the CV-selected pair over a representation set $\mathcal{R}$,

$$A_{\mathrm{amp}} = A\big[\{\,|\mathcal{F}x|\,\}\big], \qquad A_{\mathrm{full}} = A\big[\{\,x,\ |\mathcal{F}x|,\ [\,|\mathcal{F}x|, \cos\angle\mathcal{F}x, \sin\angle\mathcal{F}x\,],\ \Phi_2(x)\,\}\big], \tag{1}$$

$$\Delta_\phi = A_{\mathrm{full}} - A_{\mathrm{amp}}, \tag{2}$$

where Φ_2 is the strictly shift-invariant phase representation of §2.1 and the repertoire is fixed a priori. Because the amplitude arm lies inside A_{full}'s candidate set, $\Delta_\phi \ge 0$ holds *for the cross-validation objective by construction*—the selector can always return it—but *not* for held-out accuracy, since a larger candidate set adds selection variance; the empirical gap therefore has no guaranteed sign, and twelve estimates are negative and reported unclipped.

Estimator, not oracle. A_{amp} and A_{full} are each lower bounds on their Bayes counterparts, so their difference carries no one-sided guarantee: if $\mathcal{M}$ under-exploits phase representations Δ_ϕ *under*states true phase dependence, and if it under-fits the amplitude representation specifically it can

overstate it. The identification condition is therefore explicit: $\mathcal{M}$'s amplitude arm must not be the binding constraint. The redundancy nulls of §6 are exactly where it is, and there Δ_ϕ reports up to 0.533 against a truth of 0. We therefore treat Δ_ϕ as a *calibrated estimator*, not a certificate, and measure its error against exact synthetic ground truth (§6); the same family serves both arms, though nothing guarantees its errors on the two representations are equal (Appendix E). One reference we do *not* use is a time-domain model on *phase-scrambled* signals: it is confounded and overstates phase, which affects scrambling-style ablations common in the literature (Appendix C).

2.1 Structure: position versus relative phase

The fiber's generators differ in semantics (Figure 1a): circular shifts erase *absolute position*, while phase relations *between* frequencies—the temporal order and arrangement of events—survive. The two demand different remedies—padding or position encoding fixes the first, only phase-sensitive features the second—so the diagnostic separates them.

Define $\Phi_2(x) = [\,|\mathcal{F}x|,\ \cos\Delta^2\varphi,\ \sin\Delta^2\varphi\,]$, where $\Delta^2\varphi[k] = \angle\big(F[k{+}2]\,\overline{F[k{+}1]}^{\,2}\,F[k]\big)$ is the wrap-safe second difference of the spectral phase. It annihilates the linear-in-k ramp a circular shift adds and keeps inter-frequency relations, but fixes phases only up to $a + bk$, so it is also blind to sub-sample translation and to a constant rotation of every phase. It serves in A_{full}'s repertoire and as the decomposition's witness; its construction, mask and motivating failure are in Appendix M.

Let A_{shift} be the test accuracy of the CV-selected pair when every series—train and test alike—is independently circularly shifted, over the arms shifted-raw, $|\mathcal{F}x|$, Φ_2 and a selective translation *bispectrum* (Kakarala, 2012; Mataigne et al., 2024)—the last three invariant by construction, the first robust only because position is randomized. Randomizing position makes absolute location uninformative while preserving shift-invariant structure, the circular-shift group action being the exact fiber symmetry of Proposition 1, so

$$\begin{aligned}
\Delta_{\text{align}} &= A_{\text{full}} - A_{\text{shift}} && \text{(\textit{position}: circular-shift-sensitive)},\\
\Delta_{\text{rel}} &= A_{\text{shift}} - A^{\text{aug}}_{\text{amp}} && \text{(\textit{relative phase}: repertoire-extractable, shift-invariant)},\\
\varepsilon &= A^{\text{aug}}_{\text{amp}} - A_{\text{amp}} && \text{(a measured \textit{estimator} term)},\\
\Delta_\phi &= \Delta_{\text{align}} + \Delta_{\text{rel}} + \varepsilon.
\end{aligned} \tag{3}$$

Here $A^{\text{aug}}_{\text{amp}}$ is the amplitude arm refit under the *same* four-fold shift augmentation as A_{shift}, so the two sides of Δ_{rel} are matched in budget; A_{amp} stays unaugmented because Δ_ϕ is defined on the original task. For a shift-*invariant* arm the four copies are identical rows and add no information—but they are not neutral, and ε prices that: $+0.005$ on average, positive on 68 of 114 (sign test $p = 3{\times}10^{-4}$; Appendix M). It belongs to neither task component, so we write it separately. Enrichment raises what A_{shift} *selects on*, not held-out accuracy, where a larger candidate set adds selection variance: the bispectrum arm helps on 38 of 114 datasets and hurts 15, and $A_{\text{shift}} \geq A_{\text{amp}}$ likewise holds by selection only (24 negative Δ_{rel}, 35 negative Δ_{align}; Appendix M). We therefore read (3) as an algebraic identity plus a repertoire-relative *diagnostic*, not a monotone guarantee. No clamp is applied, small negative components are reported as estimation noise, and A_{shift} averages five shift realizations. We type each dataset as **no-detected-gap** if $\Delta_\phi < 0.03$, **rel-phase** if $\Delta_{\text{rel}} > 0.05$ and $\Delta_{\text{rel}} > 0.4\,\Delta_\phi$, and **position** otherwise—a sequential rule, so four datasets with $\Delta_{\text{rel}} > 0.05$ but $\Delta_\phi < 0.03$ (FaceAll, InsectEPGSmallTrain, LargeKitchenAppliances, RefrigerationDevices; only the second an estimator failure) are typed no-detected-gap. The first label states what was *observed*, not amplitude sufficiency, which no estimator can prove; all three are point estimates (§6 is stricter). "Position" and "relative phase" are shorthand for these group- and repertoire-relative quantities, not identified information-theoretic ones; since the two sides also differ in candidates and budget, both are pipeline-specific performance differences (Appendix M).

3 The phase-accessibility landscape of a standard benchmark

We ran the diagnostic on every fixed-length univariate dataset of the UCR 2018 archive (Dau et al., 2019)—114 of 128 datasets; the remainder are variable-length or contain missing values—using each dataset's standard train/test split (protocol details and compute in Appendix B).

Phase dependence is real, large, and wildly non-uniform. The distribution of Δ_ϕ (Figure 1c) has maximum 0.533 and interquartile range 0.132, with $52/114$ (45.6%) below the 0.03 threshold.

Twelve gaps are slightly negative, four beyond -0.05—selection noise on tiny training sets, where CV occasionally picks an arm that generalizes worse than amplitude (largest InsectEPGSmallTrain, -0.197)—and we report them unclipped. Those four are estimator *failures*, not evidence of sufficiency: excluding them leaves $48/114$ (42%) low-gap. Both facts matter: the negative-control fraction shows the diagnostic does not fire everywhere, and the heavy tail shows that on a large minority of tasks every amplitude pipeline *we audited*, the strongest amplitude learner included, forfeits tens of accuracy points; a claim about audited pipelines only (§6).

The decomposition finds the structure it was designed to find. Under the main (bispectrum-enriched) repertoire the archive splits into 52 no-detected-gap datasets (mean $\Delta_\phi = -0.010$, CNN gain -0.012), 39 **position** datasets (0.123, $+0.011$) and 23 **rel-phase** datasets (0.209, $+0.147$; every stratum mean in Table 3, the components per dataset in Figure 5). The position stratum carries essentially all of its gap in Δ_{align} (0.107 of 0.123)—location a spectrogram could recover, not shift-invariant phase, as §6 verifies directly. The rel-phase stratum is the framework's hardest prediction: given nothing but signals and labels, the diagnostic surfaces *TwoPatterns*—classes literally defined by the temporal order of two patterns—at the top ($\Delta_{\text{rel}} = 0.413$), then *CBF* ($0.396$; cylinder–bell–funnel *shape order*, invisible to sign-blind Φ_2, recovered by the bispectrum), then *BME*, *SmoothSubspace*, three *UWave* members, *SyntheticControl* and *UMD* (Table 5). A stable core of seven survives threshold perturbations, shift seeds and the conservative Φ_2-only variant; six keep the type under the budget-matched reference and five under a disjoint estimator family (Appendix M). We did not select these datasets; the estimator did, and the semantics agree. The enriched repertoire was chosen after inspecting UCR, so this membership is *exploratory* and threshold-sensitive—$32/114$ datasets sit within 0.02 of a typing boundary—while the confirmatory results are the continuous gap, the frozen Φ_2 variant and §4.1. Δ_{rel} is biased low by the repertoire: the typology is a reading aid, the continuous gap carries the weight (Appendix M).

"Average accuracy over UCR" therefore averages three qualitatively different regimes, and any phase-aware-versus-amplitude comparison inherits its conclusion from the basket's composition—one mechanism behind the instability of §1 and behind the mixed record of complex networks (§5).

4 The gap predicts raw-signal gains across model families

The predictive claim: Δ_ϕ, measured once with generic models, should forecast how much *other* model families gain from raw-signal access over the amplitude spectrum. Holding across unrelated families is what a task property predicts; holding for none sinks the framework. We test five.

For each dataset and each family f we compute a *model gain* $G_f = \text{acc}_f(\text{raw}) - A_{\text{amp}}$, the family's raw-signal accuracy in excess of the shared estimated amplitude reference of §2—a gain relative to that estimated reference, not an unconditional attribution. One baseline for every family keeps the definition uniform; the interpretive cost, a small family-specific offset for representations stronger than $\mathcal{M}$'s on amplitude, is quantified below, and a per-family baseline keeps every correlation positive, though weaker for the deep nets (Appendix E).

The families: a **1D CNN**, **MiniRocket** (Dempster et al., 2021), **InceptionTime** (Ismail Fawaz et al., 2020), a capacity-matched **complex CNN** (§5), and **1-NN DTW**, a training-free elastic matcher whose gain is against its own amplitude twin and so shares no measured quantity with any other row (Appendix B). Partial correlations residualize on length and class count; §6 adds A_{amp}, leaving 0.40–0.65. Because Δ_ϕ and most gains both subtract the shared A_{amp}, we also recompute every row against each family's *own* amplitude twin, which shares no measured quantity with Δ_ϕ: the relationship survives everywhere but not uniformly, *rising* for MiniRocket ($+0.720 \rightarrow +0.766$) and 1-NN DTW and falling for the deep networks ($+0.513 \rightarrow +0.237$ CNN, $+0.645 \rightarrow +0.450$ InceptionTime), which underfit the amplitude spectrum *as a sequence*—a limitation of those own-twin baselines rather than of the gap (Appendix E). UEA uses independent baselines throughout.

Five families, one relationship. For every family—three real-valued networks, a complex network and a training-free elastic matcher—gain rises with the gap: Pearson 0.51–0.72, partial 0.53–0.77 controlling for length and class count, all $p < 10^{-10}$, family-clustered intervals excluding zero (Table 4, Figure 3). These architectures have deliberately different inductive biases and share nothing but the task—though a learnability asymmetry common to all five would also produce agreement. That they track the same per-task scalar is the paper's central empirical fact—a claim about this *rep-*

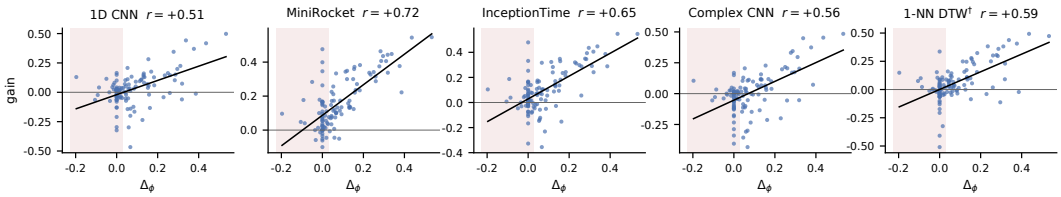

Figure 2: Held-out gain over the shared amplitude reference versus the phase-accessibility gap, per model family, all 114 datasets, no gating; shaded band = the no-detected-gap region ($\Delta_\phi < 0.03$). Five independently built families track the same per-task scalar. 1-NN DTW† is referenced to its own amplitude twin; y ranges differ. Coefficients: Table 4; full page: Figure 3.

resentation contrast, not about an identified Bayes-level property: the gap is cross-family predictive and weakens but survives under own-baseline references (CNN $+0.513 \to +0.237$; Appendix E).

The low-estimated-gap stratum. On the 52 datasets with estimated gap below 0.03, mean gains are -0.012 (CNN), $+0.036$ (InceptionTime), $+0.079$ (MiniRocket), -0.042 (complex CNN), and -0.019 (DTW): at or near zero, as predicted—where the instrument detects no gap, no family gains beyond a constant representation offset. Medians sit near zero (MiniRocket aside), so this is not a cancellation artifact. It is also not a negative control: membership is a *non-detection* selected by the same noisy estimate on the same test set, not a demonstration that the task carries no gap, so the true negative controls are the synthetic amplitude-only tasks of §6; a cross-fitted replication on disjoint halves behaves the same way (Appendix L). It is also *relative to the shared reference*: against each family's own amplitude twin the low-gap means are positive, not null ($+0.048$ CNN, $+0.021$ complex), so the stratum shows only that no family gains beyond the shared reference. MiniRocket's median $+0.050$ is a representation offset, not phase leakage: it is roughly constant in Δ_ϕ, and under its own baseline it falls to $+0.031$ while the correlation *strengthens* to $+0.766$ (Appendix E).

Two audits of this layer. The Bayes-level bound $G_f^* \le \Delta_\phi^*$ is inherited by no empirical pair, so we treat it as a *prediction*: for the CNN it holds on 85% of 114 datasets and 90% of the 103 with non-negative gap. Gating these correlations on $\Delta_\phi \ge 0.05$ would share the metric's noise and inflate them, so every headline number here is *ungated* (Appendices E and D).

4.1 A PROSPECTIVE DIAGNOSTIC FROM TRAINING DATA ALONE

Everything above uses a gap computed from official test accuracies, sound for *mapping* a benchmark but not for the practical proposal—measure the task before investing in phase-aware modeling—which needs a diagnostic available *before* the test set is touched. We therefore recompute the whole diagnostic inside the training set (stratified out-of-fold estimation, selection redone independently within every fold), yielding $\Delta_\phi^{\text{train}}$ without a single query to the test labels; the typology is then frozen and the five families' gains measured on the untouched split.

The prospective diagnostic uses the four-arm repertoire frozen *before* the UCR sweep, not the later bispectrum-enriched set, so these are already the conservative floor-repertoire results (Appendix B). It works. $\Delta_\phi^{\text{train}}$ tracks its test-side counterpart ($r = +0.83$, mean absolute difference 0.050) and—the operative result—predicts *held-out* gain for every family ($+0.524$ to $+0.657$; partials 0.52–0.68 controlling for length and class count; family-clustered intervals excluding zero; stratum separation at Mann–Whitney $p \le 5 \times 10^{-5}$ throughout), two of them better than the test-side estimate does. Disagreement concentrates where expected—small training sets ($r = -0.46$ against $\log n_{\text{train}}$).

5 THE COMPLEX CASE: INPUT AND READOUT, NOT ARITHMETIC

Complex-valued networks are the flagship phase-aware proposal, so we ask the question under matched conditions: a complex CNN capacity-matched to the real CNN of §4, same budget, five seeds. The *readout* matters more than the arithmetic. The standard architecture pools the *magnitude* of the final feature map, discarding the phase it spent every layer computing, and *loses* to its real twin (0.707 vs. 0.727, CI $[-0.035, -0.007]$); pooling $\mathrm{Re}\, z$ and $\mathrm{Im}\, z$ separately reverses the sign (0.744, CI $[+0.005, +0.031]$), concentrated on high-gap tasks. We label that reversal *exploratory*—best of

three readouts, margin not surviving a best-of-three adjustment, no replication on UEA—and keep only the *direction* the framework predicts: all three readouts give $r(\Delta_\phi, \Delta_{\text{cx}}) > 0$, gap interaction $p = 0.002$ with no head selection.

The control that isolates input from arithmetic. A complex network on the analytic signal has as its two real channels exactly $[x, \mathcal{H}x]$, and a *real* CNN given those same channels—matched in feature dimension and budget, parameters within 1.06–$1.14\times$, still differing in normalization, activation and weight sharing (Appendix H)—is the strongest model we test (mean 0.764; $+0.036$ over the real baseline, CI $[+0.022, +0.052]$; $+0.019$ over the complex `reim` network, CI $[+0.010, +0.032]$), tracking the gap more closely than the complex arm ($r = +0.416$; stratum interaction $p < 10^{-4}$). The Hilbert transform is deterministic in x and adds no *information*, only a phase-friendly inductive bias—the thesis in miniature. This is a case study, not a single-factor experiment—the arms differ in more than arithmetic—so the claim is narrow: *on the trunks and benchmarks tested here we observe no benefit from complex arithmetic over a real network given the same analytic-signal input*. We do not call complex networks inferior in general; what helped *here* was a phase-bearing input and a phase-preserving readout, both anticipated by the gap (Appendices F and H).

6 Calibration and robustness

Calibration against exact ground truth. We run the full diagnostic on generated tasks whose truth is exact by construction. On 3×30 tasks built from the fiber transformations there are no false positives in the 30 amplitude-only tasks (95% upper bound 0.095), power is $\mathbf{1.000}$ on both planted families with typing correct $30/30$, and mixed tasks recover on the diagonal (slope 1.00, $R^2 = 0.98$). A label that is a *hard nonlinear* function of the magnitudes alone, yet easier for the family in raw form, gives $0/36$ false positives at 0.03—**but only where phase is uninformative**. Where phase is *redundant* with an amplitude-determined label it fires: $5/6$ and $6/6$ on two such families (mean $+0.503$, max $+0.533$ against $\Delta^*_\phi = 0$). The null suite bounds the phase-uninformative case and no other, so the archive must be tested directly, which a training-size ladder does only in part (Appendix J). The strongest amplitude learner available to us lifts the amplitude side by only $+0.018$, preserving all 29 datasets that were interval-positive when that control was run and all core seven. On four *held-out* families with pre-registered predictions, detection is perfect on three ($20/20$ each; mean gaps $+0.50$, $+0.45$, $+0.24$) including one we predicted the repertoire would *miss*—that prediction was falsified and is reported as such—while *lowbin* is the failure (detection 0.35): phase in low-*magnitude* bins is washed out by standardization, a documented limit (Appendix J).

Phase, or the locality that global magnitude discards? Localized-magnitude references (spectrogram, wavelet) sharpen the decomposition rather than undercut it. Against a spectrogram at a *fixed* window the aggregate gap falls from 0.084 to 0.020–0.038 ($L/4$ to $L/16$), and the split is the point: the **position** stratum's gap drops $0.123 \rightarrow 0.017$–0.024 (81–86% recovered) while the **relative-phase** stratum's drops only $0.211 \rightarrow 0.073$–0.097 (54–66%). The components differ in *degree* of localized-magnitude recoverability, not in kind, and every claim is scoped to the global-magnitude reference (Appendix J relates this to in-principle spectrogram invertibility). Nor is the asymmetry an artifact of the *fixed* filterbank: a per-frame magnitude learned end-to-end, against a phase-preserving twin of identical trunk, costs 0.066 on relative phase and nothing detectable on position (difference $p = 0.012$; Appendix K).

Uncertainty, and an intervention. Conditional per-dataset intervals miss the dominant error source, train-side selection. Redoing the *entire* selection inside a class-stratified training bootstrap (with folds grouped so that resampling duplicates never straddle a fold) widens intervals about twofold ($0.055 \rightarrow 0.108$ median) and moves a third of the verdicts, always toward caution: **33 positive, 24 no-detected-gap, 53 inconclusive, 4 failure**; higher-B and joint bootstraps, run on the ungrouped construction, agree with it. So the instrument is a *benchmark-level diagnostic* and a *single-dataset screening tool only when the gap is large*. Cross-dataset correlation is still observational, so we also intervene: on six high-gap datasets each series is replaced—at nested per-series dose $q \in \{0, .25, .5, .75, 1\}$—by its phase-randomized surrogate, preserving every amplitude bin (deviation $< 10^{-14}$). The diagnostic and an independent family's realized gain fall *in step* (per-dataset $r = 0.92$–1.00, pooled $+0.91$; both sloping down with dose, -0.86 and -0.89) while the amplitude arms, whose input is unchanged, drift by at most 0.07—a noise floor, not a strict control. We read this as an interventional *consistency check*, not as proof that the quantity is label information

rather than learnability; the prospective and external validations remain load-bearing (Appendices J and I).

Three audits, and what the diagnostic is not. A disjoint estimator family reproduces the continuous gap ($r = 0.81$, 68% type agreement), independent pipelines agree in rank (Spearman $+0.48$ to $+0.70$), at most $8/114$ datasets change type across the threshold grid, and balanced accuracy leaves it almost unchanged (Spearman 0.98). Two negatives bound the claim: routing on $\Delta_\phi^{\text{train}}$ *loses* to always-raw for strong learners, and for a single pre-chosen model its own train-CV contrast predicts its gain better. The value is prognostic and comparative, not routing (Appendix I).

7 External validation on an untouched benchmark

Every design choice above was made on UCR, so we apply the *prospective* train-only diagnostic of §4.1 to the 30 datasets of the UEA multivariate archive (Bagnall et al., 2018), which took no part in the method's development and violates its univariate setting. The diagnostic extends channel-wise under a *common* circular shift of all channels, which remains in the fiber—so Proposition 1 applies unchanged and cross-channel lags survive in the data; but the per-channel invariant arms are invariant to independent per-channel shifts and do not encode them, so lags reach A_{shift} only through the shifted-raw arm (Appendix B). Gains use the two families with mature multivariate implementations, MiniRocket and 1-NN DTW, each against *its own* amplitude baseline.

The instrument transfers. The train-only gap tracks its test-side counterpart ($r = +0.80$) and predicts held-out gain on both families ($r = +0.696$ and $+0.622$; partials $+0.679$ and $+0.619$ controlling for length, class count, *channel count* and training size, $p = 1.9 \times 10^{-5}$ and 2.5×10^{-4}). The strata behave as on UCR: on the 10 datasets whose train-only gap falls below 0.03 mean gains are $+0.013$ and $+0.003$—nothing—and on the other 20, $+0.170$ and $+0.172$ (Mann–Whitney $p = 0.0001$ and 0.009); 33% of UEA shows no detectable gap, against 46% on UCR.

The semantics replicate too, which is the part no statistic can manufacture. Ranked by the train-only gap, the top of the UEA list is *UWaveGestureLibrary*—the multivariate sibling of the family that tops the UCR rel-phase stratum—then *ERing*, *CharacterTrajectories*, *PenDigits*, *SpokenArabicDigits* and *Libras*: trajectory and ordering tasks, every one. At the bottom sit *BasicMotions*, *EigenWorms*, *AtrialFibrillation* and *Epilepsy*, whose classes are defined by spectral content and periodicity. The frozen enriched-repertoire decomposition, run before the UCR mask and budget-matched reference existed and so without ε (Appendix G), replicates as well—14 datasets typed rel-phase, 7 position, 9 no-detected-gap, with *UWaveGestureLibrary* again at the top ($\Delta_{\text{rel}} = 0.221$)—and the components keep their division of labour: the external Δ_{rel} predicts realized gain ($r = +0.64$ MiniRocket, $+0.70$ DTW) while the position residual does not ($+0.32$ and -0.16). Finally the exploratory `reim` readout, frozen from UCR and rerun here, does *not* replicate (-0.005 overall; gap-stratum interaction $p = 0.47$) —the basis for §5's exploratory label. Per-dataset lists and the full multivariate protocol are in Appendices B and I.

8 Related work

Phase versus amplitude. That Fourier phase carries the dominant structure of natural signals is classical (Oppenheim & Lim, 1981); a modern line *exploits* it for domain generalization, treating phase as invariant semantics and amplitude as swappable style (Yang & Soatto, 2020; Xu et al., 2021; Lu et al., 2022), and PhASER (Mohapatra et al., 2025) adds Hilbert-phase augmentation and separate magnitude–phase encoding. None quantify per-task phase dependence. The closest (Hidaka et al., 2022) evaluates a phase-aware audio front-end on eight tasks—a model property on one domain, where ours is a task-side measurement across 114.

Complex-valued networks. Architectures and theory are well developed (Trabelsi et al., 2018; Bassey et al., 2021), with domain successes where phase is intrinsic (Dramsch et al., 2021) and mixed results otherwise (Mönning & Manandhar, 2018): that literature asks whether phase-aware *models* are better and gets unstable answers, the signature of an unmeasured task property (§5).

Shift invariance and phase-sensitive invariants. Convolutional TSC—ROCKET/MiniRocket (Dempster et al., 2020; 2021), InceptionTime (Ismail Fawaz et al., 2020)—and the scattering transform (Mallat, 2012) make position-invariance a design choice; our $\Delta_{\text{align}}/\Delta_{\text{rel}}$ decomposition is its

task-side counterpart. A caution: in the TSC ensemble literature "phase" means *subsequence alignment* (Bagnall et al., 2017; Middlehurst et al., 2021)—our *position* component, not Fourier phase as a whole. The complete shift-invariant phase-sensitive object is the *bispectrum* (Kakarala, 2012; Mataigne et al., 2024; Mathe et al., 2026), a slice of which we use as an arm (Appendix M).

Phase retrieval and dataset diagnostics. The magnitude's non-injectivity and its trivial and non-trivial 1D ambiguities are classical (Shechtman et al., 2015; Beinert & Plonka, 2015; Fannjiang & Strohmer, 2020); that literature studies *reconstruction* under side information (Jaganathan et al., 2016), while we import the impossibility direction to bound amplitude-consuming classifiers per task. The closest formal frame is information *usable* by a restricted predictive family (Xu et al., 2020), and conditional probing (Hewitt et al., 2021), which separates "easier to predict than a baseline" from "carrying information beyond it"—the distinction our redundancy nulls show Δ_ϕ cannot make. Δ_ϕ is an accuracy-scale, representation-contrast analogue for one fixed family; our additions are the task-side map, the shift decomposition and the cross-family evidence, not a new identification result. Generic feature libraries—catch22 (Lubba et al., 2019), tsfresh (Christ et al., 2018), hctsa (Fulcher & Jones, 2017)—characterize datasets to recommend models; ours is one property anchored to a provable Bayes contrast, decomposed and audited. Their spectral features are amplitude/power summaries, phase-blind by the non-injectivity we build on.

9 Limitations and scope

Our claims are bounded. *Estimator, not oracle* (§2, §6). *Accuracy, not bits:* it is in accuracy points, not information-theoretic ones. *Global DFT:* the ceiling concerns the global spectrum, and phase dependence is defined *relative to* it under z-normalization; richer magnitude representations induce smaller fibers, and Δ_{rel} is biased low by the repertoire but, as a difference of finite-model accuracies, has no one-sided guarantee (Appendix M). *Per-dataset reliability:* single-dataset estimates are noisy on small training sets (twelve negatives, four beyond -0.05); benchmark-level ordering is *measured*, per-task phase dependence only *screened*. *Blind spot:* phase in low-magnitude bins is missed by the standardized features, and the bins at the rounding floor are masked out of the phase block entirely (§2). *Metric scope:* balanced accuracy reproduces the typology; log-loss preserves the ranking but is limited by our classifiers' calibration (§6). *Model scope:* our families are diagnostic probes, not a leaderboard; strong ensembles are absent. *External validity:* UEA is external but shares a community and format with UCR. *Benchmark scope:* 114 univariate, fixed-length, z-normalized UCR datasets. *Correlational validation:* the cross-family layer is observational, supported by within-family ablations rather than task interventions. *The complex result binds one trunk:* five readouts on one capacity-matched architecture, winning margin $+0.017$ with three of five heads losing—so no general claim about complex networks is licensed (§5).

10 Conclusion

We proposed treating phase dependence—relative to the global Fourier magnitude—as a task property measurable at the benchmark level, grounded it in the non-injectivity of the amplitude representation, mapped a standard benchmark, and showed the resulting scalar predicts, prospectively and across model families and archives, how much raw-signal access pays. The complex-valued case illustrates the task-side view: the arithmetic was not the operative ingredient—a real network given the analytic signal used phase best of all, a complex network with a magnitude readout gained nothing from it here, and neither helped where the task has no gap. On UCR, nearly half the tasks show no gap we can detect, and on a fifth relative phase carries the largest share: before building a phase-aware architecture, it pays to measure whether the benchmark's tasks carry what the architecture is built to use.

Using the diagnostic. The typology suggests a workflow. Estimate Δ_ϕ and its train-selection interval from training data before choosing an architecture. With no detected gap, amplitude/power features are a sufficient baseline and phase machinery has little to use. A position-dominated gap is largely recovered by a fixed-window spectrogram (81–86%), so localized magnitude is the cheaper remedy. Where relative phase dominates, a spectrogram recovers least (54–66%); that is the case for phase-bearing inputs such as the analytic signal. An inconclusive interval calls for more training data or a direct comparison, not a type.

AI USE STATEMENT

In this work we used a generative AI assistant for tasks in the required-disclosure categories: implementing the analysis code, proposing and designing several control experiments (the adversarial amplitude-only nulls, the phase-destruction dose–response intervention, the own-contrast baseline, and the alternative-metric rescoring), generating the synthetic calibration and held-out task families, and assisting in the interpretation of results. We also used it for recommended-disclosure tasks: drafting and editing portions of the manuscript, brainstorming, searching for related literature, and formatting references. The assistant was also used for mathematical work, which the authors checked, including the characterization of what the invariant arms determine (phases up to an affine function of frequency for Φ_2; completeness only up to continuous translation for the bispectrum slice, with the one-term repair and its failure on gapped spectra; Appendix M) and the constructions in Figure 1. AI-simulated reviews of the manuscript and the supplementary package were also used to find errors, which the authors verified and corrected.

All AI-assisted work was reviewed by the authors. All model-training experiments were run on the authors' own hardware, some launched and monitored by an AI coding agent under the authors' direction; the released entry point (`reproduce.py`) verifies a SHA-256 manifest and recomputes, from the released per-dataset CSVs, the headline statistics it lists, which the authors verified against the manuscript; a small number of reported figures are produced by other released scripts rather than by that entry point, and the archive says which; cited references were checked against their published sources. The research questions, the decision to freeze thresholds and protocols before running them, the disclosure of negative and falsified results, and all scientific claims are the authors' own. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use the public UCR 2018 archive with standard splits. The diagnostic, all model-family harnesses, the analysis scripts, and all per-dataset results (CSV) will be released; the released entry point (`reproduce.py`) verifies a SHA-256 manifest and recomputes the headline statistics it lists, each tagged with its location in the paper. The adversarial amplitude-only null generator is specified in Appendix B and released as `adversarial_nulls.py`, written to that specification later rather than being the original snippet; the archive lists which statistics the entry point covers and which regenerate from their own scripts. Appendix B specifies models, hyperparameters, seeds, and compute.

REFERENCES

Rima Alaifari and Matthias Wellershoff. Uniqueness of STFT phase retrieval for bandlimited functions. *Applied and Computational Harmonic Analysis*, 50:34–48, 2021.

Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Joshua Bassey, Xiangfang Li, and Lijun Qian. A survey of complex-valued neural networks. *arXiv preprint arXiv:2101.12249*, 2021.

Robert Beinert and Gerlind Plonka. Ambiguities in one-dimensional discrete phase retrieval from Fourier magnitudes. *Journal of Fourier Analysis and Applications*, 21(6):1169–1198, 2015.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 248–257, 2021.

Jesper Sören Dramsch, Mikael Lüthje, and Anders Nymark Christensen. Complex-valued neural networks for machine learning on non-stationary physical data. *Computers & Geosciences*, 146: 104632, 2021.

Albert Fannjiang and Thomas Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29: 125–228, 2020.

Ben D. Fulcher and Nick S. Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems*, 5(5):527–531, 2017.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

Philipp Grohs, Lukas Liehr, and Irina Shafkulovska. Multi-window STFT phase retrieval: Lattice uniqueness. *Journal of Functional Analysis*, 287(10):110620, 2024.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Shunsuke Hidaka, Kohei Wakamiya, and Tokihiko Kaburagi. An investigation of the effectiveness of phase for audio classification. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3708–3712, 2022. arXiv:2110.02878.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. STFT phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 770–781, 2016.

Ramakrishna Kakarala. The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: A group-theoretic approach. *Journal of Mathematical Imaging and Vision*, 44(3):341–353, 2012.

Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research*, 2022.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery*, 33(6): 1821–1852, 2019.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Simon Mataigne, Johan Mathe, Sophia Sanborn, Christopher Hillar, and Nina Miolane. The selective G-bispectrum and its inversion: Applications to G-invariant networks. *arXiv preprint arXiv:2407.07655*, 2024.

Johan Mathe, Adele Myers, Simon Mataigne, and Nina Miolane. bispectrum: Selective G-bispectra made practical. *arXiv preprint arXiv:2605.07270*, 2026.

Matthew Middlehurst, James Large, Michael Flynn, Jason Lines, Aaron Bostrom, and Anthony Bagnall. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning*, 110:3211–3243, 2021.

Payal Mohapatra, Lixu Wang, and Qi Zhu. Phase-driven generalizable representation learning for nonstationary time series classification. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2402.05960.

Nils Mönning and Suresh Manandhar. Evaluation of complex-valued neural networks on real-valued classification tasks. *arXiv preprint arXiv:1811.12351*, 2018.

Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A Fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14383–14392, 2021.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4085–4095, 2020.

A Proof of Proposition 1

We restate Proposition 1 in the measure-theoretic form the proof uses; it is the same statement, not a new one, so it carries the same number.

Proposition 1 (Amplitude fiber, measure-theoretic form)**.** *For any $x \in \mathbb{R}^L$, any integer s, the circular shift $(T_s x)[n] = x[(n-s) \bmod L]$, the time reversal $(Rx)[n] = x[(-n) \bmod L]$ and the negation $-x$ satisfy* $\mathrm{amp}(T_s x) = \mathrm{amp}(Rx) = \mathrm{amp}(-x) = \mathrm{amp}(x)$*. Consequently any predictor of the form $h = g \circ \mathrm{amp}$ is constant on the orbit of x under these transformations, and, with $\eta_y(x) = P(Y{=}y \mid X{=}x)$, $Z = \mathrm{amp}(X)$,*

$$A^*_{\mathrm{raw}} - A^*_{\mathrm{amp}} \;=\; \mathbb{E}\Big[\, \mathbb{E}\big[\, \max_y \eta_y(X) \,\big|\, Z\big] \;-\; \max_y \mathbb{E}\big[\eta_y(X) \,\big|\, Z\big] \,\Big] \;\geq\; 0$$

by Jensen's inequality applied to the convex map $v \mapsto \max_y v_y$, with equality if and only if the integrand vanishes almost surely. The integrand is strictly positive on a fibre precisely when $\bigcap_x \arg\max_y \eta_y(x)$, intersected over the fibre, is empty up to P_X-null sets; a Bayes decision that merely varies *across the fibre is not sufficient, since a label common to every argmax set can be returned by an amplitude-only predictor at no cost. For deterministic labels the argmax sets are singletons and the condition becomes: two points* of a common fibre *carry different labels with positive probability. The fibre strictly contains the orbit generated by shift, reversal and negation, so orbit-level disagreement is a sufficient witness, not the exact condition.*

Proof. Invariance. By the DFT shift theorem $(\mathcal{F}T_s x)[k] = e^{-2\pi iks/L}(\mathcal{F}x)[k]$ and the modulus removes the unimodular factor; for real x, time reversal conjugates the spectrum, $(\mathcal{F}Rx)[k] = \overline{(\mathcal{F}x)[k]}$, and negation flips its sign. In all three cases $|\mathcal{F}\cdot|$ is unchanged, so $h = g \circ \mathrm{amp}$ is constant on the orbit.

The inequality. Write $\eta_y(x) = P(Y{=}y \mid X{=}x)$ and $Z = \mathrm{amp}(X)$. A predictor measurable with respect to $\sigma(Z)$ attains at best $A^*_{\mathrm{amp}} = \mathbb{E}\max_y \mathbb{E}[\eta_y(X) \mid Z]$, while the raw-signal Bayes rule attains $A^*_{\mathrm{raw}} = \mathbb{E}\max_y \eta_y(X) = \mathbb{E}\,\mathbb{E}[\max_y \eta_y(X) \mid Z]$. Since $v \mapsto \max_y v_y$ is convex, conditional Jensen gives $\max_y \mathbb{E}[\eta_y \mid Z] \le \mathbb{E}[\max_y \eta_y \mid Z]$ pointwise, hence $A^*_{\mathrm{amp}} \le A^*_{\mathrm{raw}}$.

Equality. The difference is the expectation of a non-negative integrand, so it vanishes iff the integrand is zero almost surely. On a fibre the integrand is zero exactly when a single label $y^\star$ attains $\max_y \eta_y(x)$ for P_X-almost every x in that fibre—equivalently, when $\bigcap_x \arg\max_y \eta_y(x)$ over the fibre is non-empty—because an amplitude-only predictor may then return $y^\star$ and lose nothing. A selected Bayes *decision* that varies across the fibre is therefore not sufficient: with $P(Y \mid x) = (\frac{1}{2}, \frac{1}{2}, 0)$ and $P(Y \mid x') = (\frac{1}{2}, 0, \frac{1}{2})$ on a two-point fibre the decisions may be taken to differ, yet label 1 is a common maximiser and both accuracies equal $\frac{1}{2}$. Strictness holds precisely when the intersection is empty on a set of fibres of positive probability; for deterministic labels the argmax sets are singletons and this becomes: two points of a common fibre carry different labels with positive probability. □

The fiber is in fact larger than these *trivial ambiguities*—distinct 1D signals unrelated by shift or reflection can share an amplitude spectrum (Beinert & Plonka, 2015; Shechtman et al., 2015)—so Proposition 1 understates the collapse. Its scope: the *global* DFT amplitude, the representation amplitude/power classifiers actually consume, not redundant magnitude measurements such as the STFT, for which uniqueness results exist (Jaganathan et al., 2016); and it asserts the non-separability of a *representation*, not the impossibility of phase retrieval.

B Protocol and hyperparameters

Decimation is not driving the result. Deep-net inputs are decimated to $L \le 512$ by strided subsampling, which aliases high-frequency content and could in principle penalise exactly the tasks whose label lives in fine temporal structure. We therefore reran both deep arms with a proper anti-aliasing filter (`scipy.signal.decimate`) on the 36 datasets where decimation actually bites ($L > 512$; lengths 570–2844, median 1024). The conclusion is unchanged and if anything conservative: the gain–gap correlation *rises* under anti-aliasing, from $+0.294$ to $+0.319$ for the 1D CNN and from $+0.254$ to $+0.445$ for the complex `reim` arm, while per-dataset accuracy moves little (mean $|\Delta\mathrm{acc}|$ 0.038 and 0.062; means $0.629 \to 0.634$ and $0.582 \to 0.635$). Strided subsampling therefore understates rather than manufactures the relationship. Per-dataset values are in `antialias.csv`.

The pipeline end to end. For one dataset with its standard train/test split:

1. z-normalize every series. Fix the representation set $\mathcal{R}_{\mathrm{full}} = \{\, x,\ |\mathcal{F}x|,\ [\,|\mathcal{F}x|, \cos\angle\mathcal{F}x, \sin\angle\mathcal{F}x\,],\ \Phi_2(x)\,\}$ and the model family $\mathcal{M}$ (random forest, one-hidden-layer MLP, logistic regression).
2. For a representation set $\mathcal{R}$, define $A[\mathcal{R}]$: score every pair $(m, r) \in \mathcal{M} \times \mathcal{R}$ by stratified K-fold cross-validation *on the training set only*, refit the argmax on the full training set, and evaluate it *once* on the official test split. The test labels are read once per arm and never used for selection.
3. $A_{\mathrm{amp}} = A[\{|\mathcal{F}x|\}]$, $\quad A_{\mathrm{full}} = A[\mathcal{R}_{\mathrm{full}}]$, $\quad \Delta_\phi = A_{\mathrm{full}} - A_{\mathrm{amp}}$.
4. A_{shift}: repeat step 2 with every series independently circularly shifted (training side with $4\times$ shift augmentation) over the shift-invariant arms {shifted-raw, $|\mathcal{F}x|$, Φ_2, bispectrum}, averaged over five shift realizations. Also recompute the amplitude arm alone under the *same* $4\times$ augmentation and the same shift sequence, giving $A^{\mathrm{aug}}_{\mathrm{amp}}$. Then, exactly as in Equation (3), $\Delta_{\mathrm{align}} = A_{\mathrm{full}} - A_{\mathrm{shift}}$, $\Delta_{\mathrm{rel}} = A_{\mathrm{shift}} - A^{\mathrm{aug}}_{\mathrm{amp}}$ and $\varepsilon = A^{\mathrm{aug}}_{\mathrm{amp}} - A_{\mathrm{amp}}$, so that $\Delta_\phi = \Delta_{\mathrm{align}} + \Delta_{\mathrm{rel}} + \varepsilon$. Forming Δ_{rel} against the *unaugmented* A_{amp} instead folds ε into Δ_{rel} and is not the reported quantity.
5. Type the dataset: **no-detected-gap** if $\Delta_\phi < 0.03$; **rel-phase** if $\Delta_{\mathrm{rel}} > 0.05$ and $\Delta_{\mathrm{rel}} > 0.4\,\Delta_\phi$; **position** otherwise.

6. *Prospective variant* (§4.1): run steps 2–3 entirely inside the training set by stratified out-of-fold estimation, redoing the (m, r) selection independently within every fold, which yields $\Delta_\phi^{\text{train}}$ without reading a single official test label.

Everything downstream—the typology, the cross-family gain predictions, the intervals—is a function of these quantities alone.

Adversarial amplitude-only nulls. Three families of tasks whose labels are hard nonlinear functions of the Fourier *magnitudes* alone, so $\Delta_\phi^* = 0$ by construction while the raw waveform remains a plausible input. Each series is drawn as a magnitude spectrum $|F|[k] \propto (1+k)^{-0.3}$ with per-sample multiplicative jitter $\mathcal{U}(0.5, 1.8)$ and *i.i.d. uniform phases* on $[0, 2\pi)$ (DC and Nyquist real), so phase carries no label information at all. Band energies $e_b = \sum_{k \in b} |F|[k]^2$ are taken over six equal bands spanning bins $[3, L/2)$, and the label is a function of e only: *xor*, the exclusive-or of two random band-energy comparisons $[e_i > e_j] \oplus [e_k > e_l]$; *teacher*, the sign of a random two-layer $\tanh$ network (eight hidden units, standard-normal weights) applied to standardized $\log e$, median-thresholded; and *ratio*, the median-thresholded $\log\big((e_i/e_j)(e_k/e_l)\big)$. Twelve tasks per family span $L \in \{64, 128, 256\}$ and training sizes $\{60, 150, 300\}$ with a 300-sample test set, giving the 36 tasks of §6. The released `adversarial_nulls.py` implements this specification and drives the frozen estimator through the same harness as the other synthetic families; it was written to the specification later rather than being the original snippet, which we state so that any difference between a fresh run and the published count is attributable.

Data. UCR 2018 archive (Dau et al., 2019), all 114 univariate fixed-length datasets without missing values, standard train/test splits, per-series z-normalization. No re-splitting, no tuning on test.

Generic family $\mathcal{M}$ and train-side selection. Random forest (250 trees), MLP (one hidden layer, 128 units, max 400 iterations), multinomial logistic regression (max 2000 iterations), each on standardized features. Every reported $A[\cdot]$ is the accuracy on the official test split of the single (model, representation) pair selected by stratified K-fold cross-validation on the training set, $K = \min(5, \text{smallest class count})$; the test set is never used for any selection. Datasets in which some class has a single training example fall back to a pre-registered rule (model: random forest; representation: by training accuracy) and are flagged. A_{shift} applies the same train-side selection over the arms {shifted raw, $|\mathcal{F}x|$, Φ_2, bispectrum} on the shifted, $4\times$-augmented training set, with CV folds grouped by original sample so that augmentation copies never appear on both sides of a fold; it is averaged over five independent shift realizations (the Φ_2-only floor variant of Appendix M drops the bispectrum arm). **Arm sets per study.** The UCR gap, its bootstrap intervals and the train-only diagnostic use the four-arm A_{full} of §2 (raw, $|\mathcal{F}x|$, amplitude+phase, Φ_2); the decomposition's A_{shift} uses the four arms just listed, bispectrum included; the UEA studies, the magnitude-scope comparison and the intervention use the five-arm A_{full} that adds the bispectrum, frozen before those runs (Appendix G). No clamping is applied anywhere in the pipeline.

1D CNN (real). Two blocks: $\text{Conv1d}(1{\to}32,\ k{=}\min(7, L),$ same padding), BatchNorm, ReLU; then $\text{Conv1d}(32{\to}64,\ k{=}\min(5, L))$, BatchNorm, ReLU; adaptive max-pooling over time; linear head. Adam, learning rate 10^{-3}, weight decay 10^{-4}, cross-entropy, batch size $\min(32, n_{\text{train}})$, 250 epochs, five seeds averaged. Inputs decimated to length ≤ 512 by strided subsampling with factor $\lceil L/512 \rceil$.

Complex CNN. Consumes the analytic signal of the same z-normalized, decimated series: real part the series, imaginary part its Hilbert transform (computed via FFT). Complex convolutions are implemented as pairs of real convolutions, $(W_r + iW_i)(x_r + ix_i) = \big(c_r(x_r) - c_i(x_i)\big) + i\big(c_r(x_i) + c_i(x_r)\big)$; activations are modReLU with a trainable per-channel bias b, scaling both parts by $\text{ReLU}(|z| + b)/|z|$; no batch normalization; the head takes the channel magnitudes, global max-pools over time, and applies a linear layer. Channel widths $(22, 44)$ are chosen to match the real network's parameter count: 10,276 versus 10,882 at the reference configuration ($L \geq 7$, $C{=}2$); across the 114 UCR instantiations the complex/real parameter ratio is 0.88–0.94 (mean 0.94). Optimizer, budget, batch size, epochs, and seeds identical to the real twin.

MiniRocket. The `sktime` implementation with its default 10,000 kernels, `random_state` $= 0$; features standardized; `RidgeClassifierCV` readout with $\alpha \in \text{logspace}(-3, 3, 10)$ (Dempster et al., 2021). Inputs shorter than 9 samples (which occurs only for the amplitude-spectrum input of one 15-sample dataset) are zero-padded to length 9.

InceptionTime. Single network, depth 6 (Ismail Fawaz et al., 2020). Each module: a bottleneck $\mathrm{Conv1d}(\cdot\to 32, k{=}1)$ when the input has more than one channel (the first block has none); three convolutional branches with kernels $\{39, 19, 9\}$ (32 filters each, same padding, no bias) plus a $\mathrm{MaxPool}(3)$–1×1-conv branch; concatenation to 128 channels, BatchNorm, ReLU. Residual connections after the third and sixth modules; global average pooling; linear head. Adam, learning rate 10^{-3}, weight decay 10^{-4}, batch size $\min(128, n_{\mathrm{train}})$, 120 epochs, three seeds. The own-$|\mathcal{F}x|$ baseline of Appendix E trains the identical network on the z-normalized amplitude spectrum treated as a sequence.

1-NN DTW. Sakoe–Chiba band 10% of the (decimated, ≤ 256) length; the vectorized banded dynamic program is verified against a brute-force reference. Amplitude twin: 1-NN Euclidean on the z-normalized amplitude spectrum.

Train-only estimator (nested; §4.1). For outer folds $k = 1..K$ (stratified, on the training set only): (i) hold out fold k; (ii) on the remaining folds, run the *entire* inner selection independently—for each arm in $\{|\mathcal{F}x|\}$ and each arm in A_{full}'s set, score every model in $\mathcal{M}$ by inner CV and keep the best (model, arm) pair per arm-set; (iii) refit both winners on the outer-training part and predict fold k; (iv) accumulate out-of-fold correctness for the amplitude and full arms. Then $\Delta_\phi^{\mathrm{train}} = \overline{\mathrm{correct}}_{\mathrm{full}} - \overline{\mathrm{correct}}_{\mathrm{amp}}$. Official test labels are never read; the inner selection is repeated inside every outer fold, so no selection decision is shared across folds.

UEA preprocessing, and what it does to variable-length datasets. The UEA loader reads each series with `sktime`'s `load_from_tsfile` and truncates train and test to a common length. On the equal-length datasets that is a no-op; on the variable-length ones it is not, and the released results carry *JapaneseVowels* at length 7 (true range 7–29), *SpokenArabicDigits* at 4 (4–93) and *InsectWingbeat* at 2—truncation to the shortest series in the split, which discards most of the signal. For a paper about temporal organization that is a substantive choice and we flag it as one: a padded or resampled variant would define different tasks. It does not carry the external result. Dropping every dataset shorter than 30 samples—those three plus *PenDigits*, whose length 8 is genuine—moves the headline correlations from $+0.696/+0.622$ to $+0.697/+0.603$ on the remaining 26. We keep the all-30 numbers as the headline because the exclusion is post-hoc, and report this robustness beside them. The released `uea_external.py` states the truncation in its header; the paper did not, which this paragraph corrects.

Compute. The two-arm gap costs a few CPU-minutes per dataset (the full 114-dataset sweep runs overnight on a single 16-core workstation); the decomposition multiplies this by the five shift realizations, and the train-selection bootstrap by B. Deep-net arms used one consumer GPU (RTX 4080 Super), a few hours per family per pass; MiniRocket, DTW, the diagnostic, the UEA study, the intervention grid ($\approx$2 CPU-hours) and all bootstraps are CPU-only.

Robustness and calibration runs. The disjoint estimator family is {histogram gradient boosting, k-NN ($k{=}5$), RBF-SVM} under the same train-side selection. Dose-response tasks mix amplitude-carried and reversal-chirp-carried labels per sample with probability q (true gap $0.5(1-q)$; six replicates per q across lengths and training sizes); the SNR sweep varies chirp amplitude at fixed noise. Bootstrap intervals resample test indices ($B{=}1000$) over the cached, paired predictions of the train-side-selected arms.

Statistics. Partial correlations residualize both variables linearly on series length and class count. Stratum comparisons use two-sided or one-sided Mann–Whitney tests as stated. All correlations computed over all 114 datasets with no gating. The 79-family clustering groups name-obvious variants (the UWave, GunPoint, Phalanx, Cricket, SemgHand, MixedShapes, Freezer, ToeSegmentation, SonyAIBO, Lightning, Worms, NonInvasiveFetalECG, InsectEPG, Pig, EOG, Face, Ford and ECG groups); the full mapping ships with the code.

Detail relocated from the body. The estimator family $\mathcal{M}$ is a random forest, a one-hidden-layer MLP and logistic regression; for every candidate (model, representation) pair, stratified K-fold cross-validation on the *training* set selects the winner, which is refit on the full training set and evaluated *once* on the official test split. The shift-invariant arms are shifted-raw, $|\mathcal{F}x|$, Φ_2 and a selective translation bispectrum, the last sensitive to global sign and, for generic signals, complete only up to continuous (band-limited) translation; its construction, cost ($O(L)$ per series beyond one FFT), numerical stability near zero magnitude, the fold grouping and the cheaper Φ_2-only variant are described below and in Appendix M. The model families of §4 are a 1D CNN (two blocks, 32/64

channels, five seeds), MiniRocket, InceptionTime (three seeds), a capacity-matched complex CNN, and 1-NN DTW whose amplitude twin is 1-NN Euclidean on $|\mathcal{F}x|$; deep-net inputs are decimated to ≤ 512. The released analysis additionally residualizes on training size. The four-arm repertoire frozen before the UCR sweep is raw, $|\mathcal{F}x|$, amplitude+phase and Φ_2; the UCR-informed bispectrum enrichment enters only the decomposition and the external runs, never the prospective diagnostic.

C THE PHASE-SCRAMBLING SURROGATE IS CONFOUNDED

A tempting alternative ceiling is to train a time-domain model on *phase-scrambled* signals (random phases, amplitudes preserved). This estimator is confounded: scrambling destroys the waveform structure that time-domain models need in order to read *amplitude* information at all, so the surrogate underestimates the amplitude ceiling and overstates the gap. In controlled synthetic tasks where the ceiling is known exactly (labels planted in amplitude with probability q, in phase otherwise, so the amplitude-reachable accuracy is $0.5 + 0.5q$ by construction), the direct amplitude-feature estimator tracks the planted ceiling while the time-domain scrambling surrogate falls systematically below it; on real data the same confound manifests as large spurious "phase" effects on tasks with no true gap. All results below therefore use the amplitude *representation*, never scrambling, as the amplitude-only view. We flag this because scrambling-style ablations are common in the literature and, by this analysis, systematically overstate the role of phase.

The ceiling bound as a tested prediction, and the gating audit (full text from the body). A testable implication of the ceiling. At the Bayes level no family can extract more than the task offers, $G_f^* \leq \Delta_\phi^*$; the empirical pair inherits no such guarantee, so the bound is a *prediction to be tested*, not a property assumed. For the CNN—whose raw-signal strength is comparable to $\mathcal{M}$'s, so the shared baseline is unbiased for it (low-gap mean -0.008)—the bound $G_{\text{CNN}} \leq \Delta_\phi + 0.05$ holds on 83% of all 114 datasets ungated and 87% of the 102 with non-negative estimated gap, violations concentrating exactly where the gap estimate is noisy-negative. For families with a representation offset the bound is meaningful only against their own amplitude accuracy (MiniRocket 71%; Appendix E): the shortfall is the expected signature of the estimator's conservatism.

Gating audit. Gating the gain–gap correlations on $\Delta_\phi \geq 0.05$ is tempting, but the gate shares the metric's noise and mechanically inflates the gated correlation. Every headline number is therefore *ungated*; the coupling's algebra, magnitude, and a gate-free replication are in Appendix D.

The ceiling bound as a tested prediction (full text from the body). Two audits of this layer. The Bayes-level bound $G_f^* \leq \Delta_\phi^*$ is inherited by no empirical pair, so we treat it as a *prediction*: for the CNN, whose raw-signal strength matches $\mathcal{M}$'s, $G_{\text{CNN}} \leq \Delta_\phi + 0.05$ holds on 85% of all 114 datasets and 90% of the 103 with non-negative estimated gap, violations concentrating where the gap estimate is noisy-negative (Appendix E). And gating these correlations on $\Delta_\phi \geq 0.05$ would share the metric's noise and mechanically inflates them; every headline number here is *ungated*, with the coupling's algebra and a gate-free replication in Appendix D.

D THE GATE–METRIC COUPLING

The learned-task gate is $\text{acc} \geq A_{\text{full}} - 0.1$; the metric is $G = \text{acc} - A_{\text{amp}}$. On the gated subset, $G \geq (A_{\text{full}} - 0.1) - A_{\text{amp}} = \Delta_\phi - 0.1$: the gate imposes a floor on the gated gain that rises one-for-one with the gap, so a positive gated association is partly definitional. The gate is also not independent of the gap (excluded median $\Delta_\phi = 0.154$ vs. included 0.018; Mann–Whitney $p = 10^{-4}$): high-gap datasets are harder for the CNN precisely because more of their accuracy lives beyond the amplitude ceiling. Both facts disqualify the gate from inferential use; §4 lists the gate-free evidence the paper's claims rest on. The excluded datasets themselves are informative: they carry substantial phase information (mean $\Delta_\phi = +0.154$) that the CNN fails to realize (mean $G = -0.059$)—model weakness, not absence of phase, precisely the distinction a task-side diagnostic exists to draw. We report the coupling because it is easy to miss, and because analogous couplings (filtering results by a success criterion that shares terms with the reported metric) are easy to introduce silently in benchmark studies.

The original gating analysis, in full. A natural way to evaluate "phase gain" is under a learned-task gate (CNN within 0.10 of A_{full}), but the gate is mechanically coupled to its own metric and correlated with the gap—so we use it only as a descriptive statistic (Table 3, last column) and rested all inference on gate-free quantities: the ungated correlations and low-gap means of Table 4, the ungated upper-bound fraction above, and a replication under an independent ≤ 10-classes gate that cannot couple to accuracy (low-gap $+0.006$ vs. $+0.057$, Mann–Whitney $p = 0.02$). The coupling algebra, the excluded-set analysis, and why such gates arise silently in benchmark studies are in Appendix D.

Reading the position stratum. On position-type tasks the CNN realizes only $+0.011$ of an available 0.123 (9%); on rel-phase tasks, $+0.147$ of 0.209 (71%). That is what the decomposition predicts for pooling-based pipelines (Dempster et al., 2020): they exploit shift-invariant relative phase but discard absolute position. The decomposition thus offers, beyond *when* phase-aware modeling pays, a repertoire-relative estimate of the shift-invariant share—indicating which kind of architectural change a task is *likely* to reward under the current repertoire.

E CEILING CHOICE: SHARED VERSUS PER-FAMILY

The main text measures every family against the shared model-agnostic ceiling A_{amp} (uniform definition; small positive representation offsets for families stronger than $\mathcal{M}$ on amplitude information). The alternative subtracts each family's *own* amplitude-spectrum accuracy: $G_f^{\text{own}} = \text{acc}_f(\text{raw}) - \text{acc}_f(|\mathcal{F}x|)$. Table 1 reports both, for every family, on all 114 datasets.

Table 1: Robustness to the ceiling definition (114 datasets, no gating). "UB frac" is the fraction of datasets with gain $\leq \Delta_\phi + 0.05$. The final row is the complex-minus-real difference and is not an amplitude ceiling.

Family	Ceiling	Pearson r	Partial r	Low gap	High gap	UB frac
1D CNN	shared A_{amp}	$+0.513$	$+0.525$	-0.012	$+0.062$	0.851
1D CNN	own $\lvert\mathcal{F}x\rvert$	$+0.237$	$+0.230$	$+0.048$	$+0.044$	0.763
MiniRocket	shared A_{amp}	$+0.720$	$+0.769$	$+0.079$	$+0.226$	0.465
MiniRocket	own $\lvert\mathcal{F}x\rvert$	$+0.766$	$+0.790$	$+0.031$	$+0.167$	0.711
InceptionTime	shared A_{amp}	$+0.645$	$+0.658$	$+0.036$	$+0.156$	0.614
InceptionTime	own $\lvert\mathcal{F}x\rvert$	$+0.450$	$+0.464$	$+0.101$	$+0.186$	0.465
Complex CNN	shared A_{amp}	$+0.563$	$+0.580$	-0.042	$+0.050$	0.877
Complex CNN	own $\lvert\mathcal{F}x\rvert$	$+0.354$	$+0.349$	$+0.021$	$+0.053$	0.807
Complex − real	(paired difference)	$+0.276$	$+0.276$	-0.031	-0.012	—

Three observations. *(i) Robustness:* every family's correlation is positive and highly significant under either definition; no sign or significance changes anywhere, so the paper's central claim does not depend on the choice. *(ii) The offsets behave as the representation-advantage interpretation predicts:* switching MiniRocket to its own baseline collapses its low-gap offset from $+0.079$ to $+0.031$ and raises its upper-bound fraction from 0.465 to 0.711—the offset was MiniRocket's representational edge over $\mathcal{M}$, not phase leakage. *(iii) Neither definition is uniformly superior:* own-$|\mathcal{F}x|$ baselines are unreliable for the *deep* networks, which underfit the amplitude spectrum *as a sequence* (a global, non-local input at odds with their locality bias)—the pattern appears consistently for InceptionTime (correlation drops to $+0.450$, low-gap offset inflates to $+0.101$), the 1D CNN ($+0.237$, $+0.048$) and the complex CNN ($+0.354$)—so their shared-baseline rows are the trustworthy ones, while for the linear-readout family (MiniRocket) the own baseline is the cleaner choice. The remaining upper-bound shortfalls (MiniRocket 0.711 against its own baseline) are the expected signature of the estimator's conservatism (§2): Δ_ϕ, computed through the generic family $\mathcal{M}$, can understate true phase dependence, and the strongest phase-aware family can legitimately exceed the estimate. This is why the main text states the bound for the CNN, whose strength is comparable to $\mathcal{M}$'s.

Which CNN accuracy vector. The archive contains two 1D-CNN accuracy vectors: the run that §5, the complex-arm study and the own-baseline study use (mean 0.7273, in `step3_gains_complex.csv`), and a stale development copy left in `merged.csv` (mean 0.7409). Every CNN number in this paper uses the first. The second differs one-sidedly on six datasets, five of them high-gap (*OliveOil* $+0.36$, *EthanolLevel* $+0.16$, *HandOutlines* $+0.14$, *Adiac*

$+0.11$, *Yoga* $+0.08$, *Fish* $+0.07$), and would inflate every CNN correlation; `reproduce.py` recomputes each CNN-derived number from the first vector and fails if the second is used. With one run on both sides, the shared-versus-own comparison in this table is like-for-like: $+0.513 \rightarrow +0.237$ changes only the ceiling. Every row of both tables reproduces exactly from the released CSVs.

The one row that is not per-dataset reproducible. The per-dataset file behind the two own-$|\mathcal{F}x|$ rows for MiniRocket and InceptionTime was not preserved. Those two rows are corroborated at the *summary* level only, by an independent recomputation shipped as `ceilings_table.csv` ($+0.774$ and $+0.456$ against the $+0.766$ and $+0.450$ reported here).

F Complex-arm details

Capacity matching, seeds, and training budget as in Appendix B. The plain complex network consumes the analytic signal (real part the series, imaginary part its Hilbert transform) and carries 0.88–0.94 of the real network's parameter count ($10{,}276$ vs. $10{,}882$ at the reference configuration); the strengthened variant adds separate batch normalization of the real and imaginary parts after each complex convolution ($10{,}900$ parameters, i.e., matched within 0.2%) and selects its learning rate per dataset from $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$ on an 80/20 train-side validation split (two seeds per rate for selection; five seeds for the final model; the three rates are chosen on $34/36/44$ datasets respectively). The learning-rate grid was granted to the complex variants only; this asymmetry runs *against* the reported deficit being a tuning artifact—the tuned side still trails the untuned real twin. Mean accuracies: real 0.727, plain complex 0.707, strengthened complex 0.711; the strengthened variant beats the real twin on $50/114$ datasets—a coin flip. Cluster-level paired differences: plain, mean -0.020, 95% CI $[-0.035, -0.007]$; strengthened, mean -0.016, CI $[-0.032, -0.001]$; equivalence within ±0.02 fails for both (the interval is not contained in the margin), and within ±0.01 a fortiori. The weak positive correlation of the difference with the gap is family-robust (leave-one-family-out $[+0.19, +0.30]$) yet coexists with negative means in both strata. We do not condition on a "both-networks-learned" subset: that gate is a function of the accuracies and A_{full}, the collider structure Appendix D warns against.

G Development timeline and freeze points

Benchmark-level adaptive overfitting is a fair concern for any study that iterates on a public archive, so we state what was frozen when. *Frozen before the first full-archive sweep:* the typing thresholds (fixed in analysis code with a released file-history audit) and the estimator family; Φ_2 was added *before* that sweep, after synthetic calibration—not UCR results—exposed a relative-phase blind spot. *UCR-informed and disclosed as such:* the bispectrum-enriched repertoire was adopted after inspecting UCR results; its guards are that the pre-inspection floor variant reproduces every conclusion with an identical core (Appendix M) and that the enriched pipeline was then frozen *before* any external data was touched. *Frozen before running:* the train-only diagnostic protocol, the UEA study and its decomposition replication (archive downloaded only after the code was fixed; the enriched repertoire and shift action unchanged from UCR; the two later UCR corrections—the low-magnitude phase mask and the budget-matched amplitude reference—were *not* applied to it, so its decomposition is the pre-correction protocol, $\Delta_{\text{rel}} = A_{\text{shift}} - A_{\text{amp}}$ with no mask, whose UCR counterpart typed 28 rather than 23 datasets rel-phase), the held-out synthetic families, whose predictions were pre-registered—one prediction was falsified and is reported as such (§6)—and the phase-destruction intervention, the higher-B and joint bootstraps, and the own-contrast baseline, whose scripts were frozen and smoke-validated on planted ground truth before touching real data. *Exploratory and labeled so:* the `reim` readout (best of three pre-specified heads; its overall margin fails a best-of-three adjustment and did not replicate on UEA); the analytic-signal control was specified after the `reim` result—a targeted control rather than a pre-registered confirmation—and is reported with unadjusted intervals. *Post-hoc:* the train-selection bootstrap, which limits per-dataset use to screening. The complete audit trail ships with the code.

H The complex-valued case study in full

This appendix carries the statistics, adjustments and configuration details behind §5.

Why this is a case study for the thesis. The debate is usually about arithmetic; our evidence recasts it as two questions about information. *Does the task carry detectable phase?*—where none is (nearly half of UCR), no phase-aware model converts any, and negative results are expected. *Does the model receive and preserve it?*—a raw real CNN can, an analytic-input real CNN does best, a magnitude-head complex network markedly less well. Less well, *not* not at all: pooling $|z|$ of the final feature map is not the same operation as consuming $|\mathcal{F}x|$, and it is not phase-blind in the sense of Proposition 1. Intermediate complex layers can fold relative phase into magnitudes before the readout sees them—$|z_1 + z_2|^2 = |z_1|^2 + |z_2|^2 + 2\,\mathrm{Re}(z_1\overline{z_2})$ depends on the relative phase of its inputs—so Proposition 1, which is about the global amplitude spectrum of the *input*, does not transfer to a magnitude readout of a learned feature map. What the experiments support is that this readout performed worse on these trunks and benchmarks, and that a real network on the analytic signal beat every complex variant we tested; the mechanism sentence is scoped accordingly here and in §5. So the apparent advantage of complex networks is the advantage of phase-bearing inputs and phase-preserving readouts, both available to real models and both predicted by the gap.

The limits: one complex trunk, and the variants differ beyond arithmetic (complex weight sharing, modReLU, no batch-norm in the base variant, 0.88–0.94 the real twin's parameters, different optimization); the strengthened variant narrows but does not close this. The analytic-input real CNN is the same trunk as the real baseline with a two-channel input $[x, \mathcal{H}x]$: identical channel widths, kernel sizes, depth, batch normalization, optimizer, learning-rate schedule, epoch budget and five seeds as the real model, and 1.06–$1.14\times$ the complex model's parameter count (the complex trunk carries 0.88–$0.94\times$ the real one's); the only change from the real baseline is the extra input channel (Appendix B). The analytic-input control was designed *after* `reim`—targeted, not pre-registered, unadjusted intervals—though it replicates the gap-tracking direction and survives its controls, while `reim` is exploratory and did not transfer to UEA. The supported claim is narrow: *on the trunk, configuration and real-valued benchmarks tested here, complex arithmetic adds no value over a real network given the same analytic-signal input*. We do not call complex networks inferior in general—only that what helped *here* was input and readout, both anticipated by a task-side quantity measured in advance.

Full results, as originally reported. The customary readout loses; a phase-preserving one wins. The standard complex scaffold pools the *magnitude* of the final complex feature map before the classifier—discarding, at the last layer, the phase it spent every previous layer computing. Against the capacity-matched real CNN it loses (mean 0.707 vs. 0.727; clustered paired-difference CI $[-0.035, -0.007]$). Replacing the magnitude pool with a real/imaginary readout (`reim`: pool $\mathrm{Re}\,z$ and $\mathrm{Im}\,z$ separately) reverses the sign: 0.744 vs. 0.727, CI $[+0.005, +0.031]$, with the advantage concentrated on high-gap tasks ($+0.032$ vs. -0.001 on low-gap; stratum-interaction permutation $p = 0.002$). We report this as an *exploratory* finding: `reim` was the best of three readouts we tried, and its overall margin does not survive a best-of-three selection adjustment (max-statistic $p = 0.10$); on a frozen design applied to the independent UEA archive the overall effect does not replicate (§7). What *does* survive both the selection adjustment and every control is the direction we actually predict—that any complex-vs-real advantage grows with the phase-accessibility gap (all readouts give $r(\Delta_\phi, \Delta_{\mathrm{cx}}) > 0$; the three-readout average has a gap interaction at $p = 0.002$ with no head selection).

The key control: the analytic-signal input, not the arithmetic, carries the value here. A complex network built from the analytic signal receives, in its two real channels, exactly $[x, \mathcal{H}x]$ (the signal and its Hilbert transform). So we built the matching control: a *real-valued* CNN taking those same two channels, matched in feature dimension and training budget, with parameter count within 1.06–$1.14\times$ (not an exact match; the arms also differ in normalization, activation and weight sharing—this is a case study, not a single-factor experiment). The Hilbert channel is a deterministic transform of x and adds no *information*; what it adds is representation—a phase-friendly inductive bias for finite-sample learning, which is precisely the paper's point that the binding constraint is what a model can extract. It is the strongest network of all—mean 0.764, beating both the real baseline ($+0.036$, CI $[+0.022, +0.052]$) and the complex `reim` network ($+0.019$, CI $[+0.010, +0.032]$)—and its advantage tracks the gap more strongly than the complex network's ($r = +0.416$ vs. $+0.385$; gain over A_{amp} correlating with the gap at $r = +0.676$; stratum interaction $p < 10^{-4}$). On this trunk and benchmark we observe no benefit from complex arithmetic: what helps is the analytic signal—an explicitly phase-bearing input—which a real network uses at least as well as the complex one does. The uncontrolled structural differences above bound how far this can be pushed.

I AUDITS: BASELINE COUPLING, ESTIMATOR ROBUSTNESS, THRESHOLDS, AND DECISION VALUE

The following analyses support §6; they are audits of our own claims rather than findings, and are placed here for readability. Figure 4 defines the dataset subsets used throughout—the 52 low-gap, the 33 interval-positive, the 23 rel-phase, the 12-member Φ_2 floor and the stable core of seven—and their containments.

The shared-baseline coupling, quantified. Δ_ϕ and G_f share the term $-A_{\mathrm{amp}}$, so a positive correlation could arise mechanically. A null in which A_{full} and acc_f track A_{amp} but are otherwise independent (regressions on A_{amp}, independently permuted residuals) produces correlations of $+0.10$ (CNN), $+0.32$ (MiniRocket), $+0.26$ (InceptionTime) and $+0.12$ (complex)—so the coupling is real and non-negligible—but every observed value (Table 4) exceeds its own null's 97.5th percentile ($+0.29$ to $+0.47$; permutation $p < 2 \times 10^{-4}$, 5000 simulations). Partialing A_{amp} out of both axes leaves the correlations essentially intact ($+0.40$ to $+0.65$). Own-representation gains—sharing *no* measured quantity, not even A_{amp}—agree across families (MiniRocket×InceptionTime $r = +0.59$) and track the gap ($r(\Delta_\phi, \text{MiniRocket own}) = +0.77$, cluster CI $[+0.60, +0.87]$; moderate for the deep networks, whose own-amplitude arms suffer the mismatch of Appendix E). Re-referencing every family's gain to the *disjoint* family's amplitude estimate—so gap and gain share no measured quantity at all—leaves the correlations at $+0.40$ to $+0.54$ with clustered intervals excluding zero (e.g., CNN $[+0.15, +0.60]$, MiniRocket $[+0.34, +0.73]$). Finally, a cross-fitted replication on *disjoint* test halves covers two of the five families: MiniRocket reproduces to within hundredths ($+0.700$ against its same-half $+0.729$) and 1-NN DTW moves by 0.064 ($+0.526$ against $+0.590$); the three deep families need per-sample prediction dumps the run did not have and remain unmeasured (Appendix L). The coupling is real, quantified, and several times weaker than the signal.

Estimator-family robustness and clustered inference. Re-estimating the entire diagnostic with a *disjoint* generic family (gradient-boosted trees, k-NN, RBF-SVM) yields gaps tracking the main family's ($r = +0.81$; mean $|\Delta| = 0.045$): the diagnostic is a model-agnostic *target* estimated with a fixed model family, and the continuous estimate is family-robust. Pushing further, we compare against fully independent estimator *pipelines*—each strong family's own raw-versus-magnitude contrast (MiniRocket, InceptionTime, DTW), sharing neither classifiers nor representations with $\mathcal{M}$: rank agreement is substantial (Spearman $+0.48$ to $+0.70$) and the core-seven sit in the top decile of every pipeline's ranking (median percentile 91–96), but membership of the low-gap *set* is pipeline-dependent at the 0.03 line (25–37 of the 52 stay below it; 60–83% below 0.05), with MiniRocket's shortfall dominated by its constant representation offset (Appendix E). The estimand is model-agnostic; the estimator's threshold-level verdicts are not, and we say so. Thresholded typing is coarser—the mean absolute difference exceeds the 0.03 threshold, so type flips near the boundary are expected (68% agreement between the two families, both under the floor repertoire, which the disjoint family was run with); across the three settings we test (main-family floor, disjoint-family floor, and the main enriched repertoire), $66/114$ datasets keep an identical type and five of the seven core rel-phase members are rel-typed in all three. The archive also contains closely related variants (UWave×4, GunPoint×4, the Phalanx group, ...), so dataset-level p-values overstate independence: grouping the 114 datasets into 79 strict families and cluster-bootstrapping over families (the scheme barely matters: mechanical prefix-6 and prefix-8 groupings, and no clustering at all, move the CNN interval by at most 0.02 at either end), the 95% confidence intervals for every architecture's gain–gap correlation exclude zero (Table 4, last column; e.g., CNN $[+0.29, +0.69]$, MiniRocket $[+0.56, +0.86]$, 1-NN DTW $[+0.39, +0.74]$); all headline inference in this paper is reported at the cluster level. Leave-one-family-out correlations stay within $[+0.45, +0.55]$ (CNN), $[+0.71, +0.80]$ (MiniRocket), $[+0.62, +0.70]$ (InceptionTime), $[+0.52, +0.60]$ (complex) and $[+0.55, +0.62]$ (DTW): no single family drives the result.

Sensitivity and uncertainty. Across a threshold grid (0.02–0.04, 0.04–0.07, 0.3–0.5) at most 8 of 114 datasets change type—with the rel-phase set overlapping the default in 18–23 of 23 members. Six of the seven core members are rel-typed in all 27 configurations; the seventh, UWaveGestureLibraryAll, left the stratum under the budget-matched reference ($\Delta_{\mathrm{rel}} = 0.117$ of $\Delta_\phi = 0.319$, just under the 0.4 fraction) and is rel-typed in the 9 configurations with fraction 0.3. A_{shift}'s five-realization spread has median s.d. 0.003 (max 0.079); the shifted-raw arm wins only $75/570$ realizations—the invariant-by-construction arms carry the estimate. Degenerate-CV fallbacks: DiatomSizeReduction, FiftyWords, Fungi, Phoneme. **What the diagnostic is not.** A natural hope is an accuracy *router*:

use the raw pipeline when $\Delta_\phi^{\text{train}} \geq 0.03$ and the cheaper amplitude pipeline otherwise. On our data this policy *loses* to always-raw for the strong learners (MiniRocket 0.824 vs. 0.862; InceptionTime 0.792 vs. 0.803; it helps only 1-NN DTW, 0.743 vs. 0.734), because raw input rarely *hurts* even where it does not help—the amplitude pipeline wins on only 13–35% of datasets, and the router picks the per-dataset winner wrongly 28–37% of the time. Always-raw is the accuracy-safe default; if routing is the goal, a lightweight raw-versus-amplitude comparison of the target model itself is the right tool. The diagnostic's decision value is *prognostic*, not routing: it forecasts, prospectively and across archives, how much phase-aware effort (inputs, readouts, augmentation choices) will pay, and *which kind* of structure carries it. We then ran the strongest practical baseline head-to-head: each target model's *own* train-CV raw-versus-amplitude contrast. For predicting a model's own held-out gain, its own contrast wins, as it should—MiniRocket $r = 0.84$ versus the generic gap's 0.61, with the gap adding nothing further (partial $+0.12$, $p = 0.21$, $\Delta R^2 = 0.004$); DTW 0.75 versus 0.66, though there the gap retains significant incremental signal (partial $+0.30$, $p = 0.002$, $\Delta R^2 = 0.038$). Portability is where the generic instrument earns its keep: predicting a *different* family's gain, the gap ($r = +0.61/+0.66$) matches or beats one family's own contrast transferred across ($+0.46$ to $+0.65$), and it is the only quantity here with a principled target and a decomposition. Even a *single cheap* contrast carries much of the signal: 1-NN DTW raw versus Euclidean-$|\mathcal{F}x|$, essentially training-free, predicts the other families' gains at mean $|r| = 0.54$ against the full three-model, four-representation selection's 0.61, which adds $\Delta R^2 \approx 0.13$ on average (0.05–0.22). The practical guidance is therefore two-tier and deliberately modest: choosing inputs for one pre-selected model, compute that model's own contrast; mapping a benchmark, comparing families, or asking *what kind* of structure a task carries, use the gap—whose value is portability, decomposition and a principled target, not per-family prediction accuracy. Conditional intervals (the cheaper quantity, reported per dataset in Appendix N) have median width 0.055 and partition the archive into 42 datasets with upper bound below 0.03 (48 below 0.05; 38 of the 42 non-failures), 42 with lower bound above it and 30 ambiguous; $51/62$ high-gap datasets entirely above zero; all seven core rel-phase members exclude zero, while two boundary members (Car, OliveOil) straddle it—the intervals separate small-but-real structure from threshold noise. The four large negative gaps have entirely negative intervals: train-side selection error, not test noise.

Train-selection uncertainty, in full. Conditional per-dataset intervals—the arms selected and fitted once, only test indices resampled—cover test sampling alone, while the dominant error source is train-side selection. We therefore bootstrap, on all 114 datasets, the training set (class-stratified resampling with replacement), redoing the *entire* model-and-representation selection inside every replicate—with the inner cross-validation folds *grouped by original sample*, so that the duplicate rows a bootstrap replicate necessarily contains never land on both sides of a fold—evaluating on the fixed official test set, and taking percentile intervals over the replicate gaps (B=30; verdicts: positive if the interval lies above 0.03, no-detected-gap if below, failure if the point estimate is below -0.05, else inconclusive). This interval covers training resampling and the entire selection procedure at the *fixed* official test set; test-sampling uncertainty is quantified separately by the conditional interval, and folding the two together in normal quadrature yields 30 positive / 14 no-detected / 66 inconclusive / 4 failure—confident positives are robust to the extra width, confident absences are not. We label this a *sensitivity analysis*, not a coverage guarantee: percentile intervals can be asymmetric, accuracy differences need not be normal, train-selection and test-sampling effects need not be independent, and arm selection is discrete on small datasets; the clean construction is a joint bootstrap resampling both sides with full re-selection per replicate, which we release—and ran on the 12-dataset key subset (B=200): joint widths match the quadrature prediction (median 0.247 vs. 0.249; $r = 0.95$), and joint verdicts agree with the train-selection verdicts on $12/12$ (the quadrature's on $10/12$, erring once in each direction). The train-selection categories are thus what the joint construction supports on the tested subset, with the quadrature counts as the conservative sensitivity read. The result: train-selection intervals are about twice as wide (median $0.055 \rightarrow 0.108$), and the four-way per-dataset verdict changes on a third of them, always toward caution: under the train-selection interval only 33 datasets remain "positive" (from 42), 24 are "no detected gap", 53 are "inconclusive", and 4 are failures.

Fold grouping. A bootstrap replicate contains duplicated rows, and under a plain stratified inner split copies of one original sample can fall on both sides of a fold. That is not test leakage—

the test split is untouched—but it changes what the selection *inside* a replicate simulates, so every inner split groups copies by original sample. As a sensitivity check we also ran all 114 datasets ungrouped at the same budget (B=30, five folds, same seeds). The ungrouped construction gives 29/25/56/4; relative to it, grouping moves 15 verdicts in both directions and never removes a positive—six no-detected-gap→inconclusive, five inconclusive→no-detected-gap, four inconclusive→positive (*CricketZ*, *Lightning7*, *OliveOil*, *Yoga*)—and narrows intervals on balance (median $0.121 \rightarrow 0.108$; narrower on 61 datasets, wider on 42), the direction the mechanism predicts, since straddling copies corrupt within-replicate selection rather than the test-set evaluation. All seven core rel-phase members are positive under both. The ungrouped run was made in an earlier software environment, so we do not read the difference as a certified effect of grouping alone.

Two degenerate cases. Seven datasets return a zero-width interval $[0, 0]$, every replicate yielding an identical gap. Three of them (*FordB*, *OSULeaf*, *StarLightCurves*) have a frozen point estimate of exactly 0.000 as well: the magnitude arm wins the full-arm selection every time, so "no detected gap" is the right verdict and only the width is uninformative. The other four are precisely the degenerate-CV fallback datasets flagged above, where some class holds a single training example and a pre-registered rule takes over (Appendix B): *DiatomSizeReduction*, *FiftyWords*, *Fungi*, *Phoneme*. Under resampling that fallback fires in every replicate and deterministically selects the magnitude arm, collapsing the gap to zero even where the frozen estimate is large (0.284 for *FiftyWords*, 0.296 for *Fungi*). Their verdicts record the fallback rule, not evidence about the task. Separately, the three *Pig* datasets (104 training series, 52 classes—two examples per class) carry grouped intervals lying entirely below -0.05 ($[-0.12, -0.06]$, $[-0.36, -0.22]$, $[-0.27, -0.17]$) while their frozen point estimate is 0.000, because a class-stratified resample cannot preserve a two-per-class structure; they are typed no-detected-gap on the point estimate, and that verdict is not meaningful either. Of the 24 no-detected-gap datasets, then, 15 rest on a non-degenerate resample, and that is the number a reader should treat as a confident absence. We report all 114 verdicts unedited and mark these nine in the released CSV rather than dropping them. (Monte-Carlo checks at higher B: a 24-dataset subset at B=50 and a 12-dataset key subset at B=200 agree with the B=30 run in width—median $|\Delta\mathrm{width}|$ 0.008 and 0.014—and in verdict on 33 of 36 comparisons. The three changes run in both directions (two inconclusive-to-positive at B=50, one positive-to-inconclusive at B=200), so the budget does not systematically inflate or deflate positives, but it does leave the counts uncertain: an 8.3% per-dataset change rate implies roughly 10 of the 114 verdicts (Wilson 95% interval 3–25) could move under a much larger budget. The four counts should be read with that tolerance, and the qualitative split—a confident minority, a confident-null minority, and a large inconclusive middle—is what survives it.) Large gaps with adequate training data survive (TwoPatterns remains positive with interval width 0.057 even at B=200); borderline cases do not, and one showcase moves with the budget: UMD, positive at B=30, widens to inconclusive at B=200. We therefore state what the instrument is: *a benchmark-level diagnostic*, whose distributions, strata, and cross-family predictions are stable under every perturbation we test, and a *single-dataset screening tool only when the gap is large*. Per-dataset use requires train-selection intervals; we release the code, and the per-dataset train-only gaps with their train-selection intervals alongside every test-side estimate. The supplementary archive ships the scripts, every per-dataset CSV, the family-clustering map, the freeze-point audit and a SHA-256 manifest, with an entry point (`reproduce.py`) that verifies the manifest and recomputes, from the CSVs alone and with no archive download, model training or GPU, the headline statistics it lists—each tagged with its location in this paper. Statistics outside that list (the synthetic calibration and held-out families, the adversarial nulls, the GPU readout ablation, and the per-dataset bootstrap intervals) regenerate from their own released scripts, and the archive says which are which. Fold assignment, by contrast, is not the problem: the train-only gap has a median standard deviation of 0.018 across five independent fold seeds. Controlling jointly for training size, A_{amp} and the available headroom $1 - A_{\mathrm{amp}}$, the train-side gap's partial correlation with held-out gain stays positive for every family ($+0.49$ to $+0.61$ across CNN, MiniRocket, InceptionTime, DTW and complex; $+0.47$ to $+0.58$ for the test-side gap), so the association is not an artifact of ceiling geometry or sample size.

J CALIBRATION, ADVERSARIAL NULLS, HELD-OUT FAMILIES, AND THE INTERVENTION IN FULL

This appendix carries the full constructions and numbers behind the calibration paragraphs of §6; the body reports their outcomes.

Synthetic ground truth. We run the full diagnostic on three families of generated tasks whose truth is exact by construction—amplitude-only (per-sample phase randomization: true gap 0), planted relative-phase (chirp reversal: amplitude spectra identical by Proposition 1), and planted position—built from the fiber transformations themselves (recipes in Appendix B). Across 3×30 tasks spanning lengths, class counts and training sizes: no false positives occur in 30 amplitude-only tasks (one-sided 95% upper bound on the rate: 0.095; mean estimated gap $+0.0000$; all 30 typed no-detected-gap); power on the planted relative-phase and position tasks is $\mathbf{1.000}$ each, with type assignment correct on $30/30$ and $30/30$. Mixed tasks with true gap $0.5(1-q)$ are recovered on the diagonal ($\widehat{\Delta}_\phi = 1.00\,\Delta_\phi^{\text{true}} - 0.01$, $R^2 = 0.98$); an SNR sweep locates the operating range (power 1.00 down to chirp amplitude 0.25; 0.67 below). Typing thresholds were fixed in code before the full-archive sweep (file-history audit released); the sensitivity grid below removes their inferential weight regardless.

Adversarial amplitude-only nulls, and a stronger amplitude arm. The sharpest version of the estimator-confound worry is a task whose label is a *hard nonlinear* function of the magnitudes alone ($\Delta_\phi^* = 0$) that the family finds easier in raw form. We built three such families—XOR of band-energy comparisons, a random two-layer teacher boundary on band energies, and curved ratio products—with per-sample uniform-random phases, so the waveform carries the label only through $|\mathcal{F}x|$: across 36 tasks the false-positive rate at 0.03 is $0/36$ (max gap $+0.007$; the XOR family even shows the predicted *negative* bias, -0.034 mean, from selection variance on an uninformative full arm). On the real archive we replace A_{amp} by the strongest amplitude learner available to us—the maximum of $\mathcal{M}$'s arm, the disjoint family's arm, and MiniRocket applied to the magnitude sequence: the amplitude side rises by only $+0.018$ on average, the confound's signature correlation between gap and amplitude-side lift is nearly absent ($r = +0.11$), and every one of the 29 datasets that were interval-positive when this control was run—all of them positive under the grouped bootstrap as well, which adds four the control did not cover (Appendix I)—27 of the 28 rel-phase datasets as typed before the phase mask (this control was not recomputed under it; Appendix M), and all core seven (gaps 0.24–0.53) keep their gap above 0.03 against it. What the stronger arm does erode is the small-gap region (median $0.045 \to 0.029$)—consistent with the estimator framing, and stated as such.

What those nulls do not cover, and what happens when they do. All three families above draw phases i.i.d. uniform, so phase carries no label information and the tasks cannot exhibit *redundancy* between phase and the label. Redundancy is the harder case, and the $0/36$ above says nothing about it: the label can be an exact function of $|\mathcal{F}x|$—so $\Delta_\phi^* = 0$ still—while the phase is set *from* that label, handing a finite raw learner a shortcut the amplitude arm does not have. We added two such families, both with $\Delta_\phi^* = 0$ by construction. In *redundant* the label is the curved-ratio rule above and one bin's phase is set from it with probability p; in *parity* the label is the parity of ten magnitude comparisons and every phase is set to 0 or π by it, so the sign of $x[0]$ gives the label away while a finite amplitude model cannot learn a ten-bit parity. At the 0.03 threshold: *redundant* gives $5/6$ false positives (mean $+0.076$, max $+0.157$) and *parity* $6/6$ (mean $+0.503$, max $+0.533$). The dose-response is the informative part—sweeping p over four tasks each, the false-positive count is $0/4$ at $p \le 0.5$, $2/4$ at $p = 0.75$ and $4/4$ at $p = 1$, with mean gaps -0.008, $+0.000$, -0.014, $+0.020$, $+0.066$. Two ingredients are therefore needed for a large false gap, and they are separable: the redundancy must be near-deterministic, *and* the label must be one the family cannot learn from $|\mathcal{F}x|$ at all. Where the amplitude arm still succeeds (accuracy 0.90 in *redundant*) the false gap is bounded by what that arm leaves on the table; it is only when the amplitude arm collapses to chance (*parity*, 0.50) that the gap approaches 0.5. We record this as a correction to what the $0/36$ licenses: it bounds the phase-uninformative case and that case only. The discriminating property against real data is that an information gap cannot be closed by giving the amplitude side more resources whereas a learnability gap can, which is a learning-curve question and is settled next. Generator and results: `adversarial_nulls.py --leak-sweep` and `adv_nulls_redundant.csv`. **Provenance of the $0/36$ result.** The script that produced the

published $0/36$ was run locally and we were unable to recover it; a search of every run directory and archive we retain turned up only copies that self-describe as later reimplementations. What the supplement ships is therefore a reimplementation written from the specification above, which reuses the frozen estimator and reproduces the result, and and is labelled as such in the archive. The redundancy families of this paragraph, which are the ones that change what the nulls license, were generated by the shipped code.

Does the confound actually operate on this archive? The nulls establish that the estimator *can* be fooled; they cannot say whether it *is*. The discriminating property is that an information gap cannot be closed by giving the amplitude side more data, whereas a learnability gap can. We therefore refit both arms at five training-set sizes (15% to 100%, stratified, three seeds, the paper's own train-side selection at every rung, the official test set untouched) on the 88 datasets whose training sets support a ladder, and regressed the gap on $\log_2 n$. If the gap were representation-learnability it should shrink. It does not: the no-detected-gap stratum stays at zero (slope $+0.001$, CI $[-0.002, +0.004]$), the position stratum is flat ($0.091 \to 0.095$; slope $+0.0003$, CI $[-0.016, +0.015]$), and the relative-phase stratum moves the *wrong way* for the confound: the gap *widens* with data, $0.151 \to 0.201$, slope $+0.023$ with CI $[+0.009, +0.038]$, and it grows on 16 of its 22 datasets. That is what an amplitude ceiling would produce, but it does not identify one: two finite learners with different learning rates can also show a gap that widens with data before it closes, and a fixed-capacity amplitude learner can under-fit at every rung. The ladder therefore rules out the easy form of the confound—a gap that shrinks as the amplitude side gets more data—and nothing stronger. Two limits. The instrument's positive control (a redundancy family whose label the amplitude arm *can* learn) slopes down only weakly on the run that produced these numbers (-0.006; -0.030 on a second machine), so its demonstrated power is modest. And it has no power against the *parity* family, whose gap is flat by construction because a ten-bit parity needs exponentially more data than a fourfold increase supplies. A flat curve is therefore evidence against learnable-but-hard redundancy only; it cannot exclude redundancy that is infeasible on the amplitude side at any n we could reach. `learning_curve.py`, `learning_curve.csv`.

Held-out synthetic families. The calibration above shares design DNA with the estimator (indeed Φ_2 was *added* after an early calibration exposed a relative-phase blind spot), so we froze the estimator and generated four further families, with pre-registered predictions, that were never used to modify it: *sign* (label = global sign of an asymmetric motif; $|\mathcal{F}x|$ identical by Proposition 1), *revmotif* (a generic motif versus its time reversal), *coupling* (matched magnitudes; the label lives in the sign of a nonlinear phase-coupling triple), and *lowbin* (label-bearing phase confined to low-*magnitude* Fourier bins). Detection is perfect on the first three ($20/20$ each, all typed rel-phase; mean gaps $+0.50$, $+0.45$, $+0.24$)—including *sign*, which we had predicted the main repertoire would miss: the prediction was wrong, because although Φ_2 is sign-blind the raw arm inside A_{full} is not, so the blind spot lives in the *decomposition*, not in the gap. *lowbin* is the failure: detection 0.35, mean gap $+0.001$, never typed rel-phase—phase information carried by low-magnitude bins is washed out by feature standardization. We report it as a documented limit of the present estimator (a magnitude-weighted or whitened phase representation is the natural fix) rather than as evidence of amplitude sufficiency. Two further families we designed proved *ill-posed* and were discarded before the sweep, which is itself informative: a "which transient came first" task that is both shift-randomized and amplitude-matched collapses (with mirrored motifs the two classes are the same cyclic sequence read two ways; with distinct motifs the amplitude spectrum separates them outright), and a position-labeled task on a non-periodic background always leaked the label into $|\mathcal{F}x|$ through boundary effects—evidence that a shift-invariant "order" label must be carried by asymmetric *content*, exactly as in TwoPatterns.

An intervention on real data. Cross-dataset correlation, however well controlled, shows agreement about a representation contrast; the causal reading needs an *intervention*. We built one: on six high-gap datasets, each series is replaced—with nested per-series dose $q \in \{0, .25, .5, .75, 1\}$—by its phase-randomized surrogate, preserving every amplitude spectrum bin-for-bin ($\max \big| |\mathcal{F}x| - |\mathcal{F}x_{\text{sur}}| \big| < 10^{-14}$), so only phase organization decays. The diagnostic and an independent family's realized gain fall *in step*: per-dataset $r(\Delta_\phi(q), \text{gain}_{\text{MiniRocket}}(q)) = 0.92$–$1.00$ (pooled $+0.91$ over the 30-point grid), both sloping down with dose ($r(q, \Delta_\phi) = -0.86$; $r(q, \text{gain}) = -0.89$), while both amplitude arms stay flat (max drift 0.07—the built-in negative control)—though flat is weaker than what should hold: the surrogate preserves $|\mathcal{F}x|$ bin-for-bin to 10^{-14}, so the amplitude arm's input is unchanged across doses and its accuracy should be *identical*, not merely close. The resid-

ual sits on two datasets (*SmoothSubspace* 0.073, *UMD* 0.042; the other four are ≤ 0.033) and is re-selection and refitting variance—the arm is re-selected and refit at every dose—not test sampling, which cannot move a quantity computed from identical inputs. Fitting the amplitude arm once per dataset and reusing it across doses would make this control exact. At full destruction both collapse to approximately zero or slightly below, the below being the documented small-train selection-error mode at its worst (CBF, -0.19). Moving a real task's phase organization moves the diagnostic and the downstream gain together, with the amplitude arms flat. Design details, since they bound the claim: each series draws one uniform u_i and one surrogate phase set, both fixed across doses, so the dose masks are nested and the path is monotone by construction; we ran one randomization per dataset, not several, so the reported correlations carry sampling and selection variance but *not* randomization variance, and they pool doses within datasets rather than modelling the dataset hierarchy. Because the surrogate preserves every magnitude bin exactly, the amplitude arms' dose-to-dose movement (median 0.032, max 0.073) is by construction pure estimator and test-sampling noise, and serves as a direct read of that noise floor rather than as a confound. We read this as an interventional *consistency check*—the estimator responds to phase organization specifically—not as proof by itself that the quantity is label information rather than learnability, which is what the adversarial-null and strong-amplitude audits above address; the prospective and external validations remain the load-bearing evidence.

Localized magnitude and phase retrieval. A continuous spectrogram is almost surely invertible up to a global phase (Griffin & Lim, 1984; Jaganathan et al., 2016; Alaifari & Wellershoff, 2021), though not from lattice samples at any density (Grohs et al., 2024). What a finite learner extracts from a *discretized* spectrogram is a different and strictly weaker quantity, and it is what the localized-magnitude arms of §6 measure.

External validation: what the body does not have room for

The body carries every headline number of the UEA study; this section adds only what it omits, so that the two never have to be kept in sync.

Ranking detail. *UWaveGestureLibrary* tops the train-only ranking at $\Delta_\phi^{\mathrm{train}} = 0.333$. The rel-phase share of the external archive is higher than UCR's (47% against 25%), which is what the archive's trajectory-heavy composition predicts.

The multivariate estimand. The diagnostic extends channel-wise: the DFT and its magnitude are taken per channel and concatenated, and a *common* circular shift of all channels stays in the fiber, so Proposition 1 applies unchanged. That action deliberately preserves cross-channel lags, so class information carried by inter-channel delay is retained rather than averaged away—independent per-channel shifts would define a coarser and different estimand that destroys it. That preservation is in the *data*, not in the features: the per-channel invariant arms are invariant to independent per-channel shifts too, so they cannot encode a lag. Cross-channel relative phase therefore reaches A_{shift} only when the shifted-raw arm is selected and is otherwise lost to Δ_{rel}. An inter-channel phase-difference arm (invariant to a common shift, not to independent ones) would carry it; we have not added one, and contrasting common- with per-channel-shift actions would isolate the effect.

How exposed is the real archive to the redundancy mechanism? An operational screen. The redundancy nulls above reach a false gap of $+0.533$—numerically the archive's own maximum—but they reach it only in a specific configuration: the label must be one the amplitude arm cannot learn *at all*, so that arm collapses to chance (0.50 on a two-class task) and the entire remaining headroom is available to be mistaken for phase. That gives a screen that costs nothing to run and that we report per dataset: how far above chance the amplitude arm actually sits. A large gap on a dataset whose amplitude arm is already far above chance cannot be produced by the parity mechanism at full strength, because the mechanism's size is bounded by what that arm leaves on the table. The screen bounds that one mechanism and no other: a task mixing amplitude-easy examples with examples whose amplitude encoding is hard and whose phase encoding is easy keeps the amplitude arm far above chance while still carrying a zero Bayes gap and a large empirical one, so a wide margin above chance is not evidence against redundancy in general. Across the 62 datasets with $\Delta_\phi \geq 0.03$ the amplitude arm sits a median $+0.359$ above the uniform-chance rate, and only three sit within 0.10 of it: *Wine* ($+0.037$), *Ham* ($+0.091$) and *SmoothSubspace* ($+0.093$). Six of the seven core relative-phase datasets sit $+0.20$ to $+0.51$ above chance. The exception: *SmoothSubspace* carries

the archive's largest gap ($+0.533$) with an amplitude arm only 0.093 above chance, which is exactly the regime in which our own null fires. We therefore do not claim that dataset as evidence of Bayes-level phase dependence; it is the clearest single case where a redundant, more-learnable encoding and a genuine information gap cannot be separated by anything we measure.

UEA: does the decomposition survive the variable-length truncation? Four UEA datasets are truncated hard by the common-length rule (*InsectWingbeat* to 2, *SpokenArabicDigits* to 4, *JapaneseVowels* to 7, *PenDigits* to 8), and all four are typed relative-phase, so that stratum is the one truncation could inflate. Dropping every dataset below length 30 leaves the other two strata untouched and shrinks only that one: $14/7/9$ becomes $10/7/9$. The predictive results are unchanged—$r(\Delta_{\mathrm{rel}}, \text{MiniRocket gain})$ $+0.636 \rightarrow +0.627$, $r(\Delta_{\mathrm{rel}}, \text{DTW gain})$ $+0.700 \rightarrow +0.689$, and the train-side headline $+0.696/+0.622 \rightarrow +0.697/+0.603$—as is the top of the semantic ranking (*UWaveGestureLibrary*, *CharacterTrajectories*, *SelfRegulationSCP1* in that order under both). What the truncation does affect is the size of the relative-phase stratum, and we report the count both ways.

The calibration paragraphs above are the unabridged versions of §6; the audits are in Appendix I and the complex-valued case study in Appendix H.

K A LEARNED LOCALIZED-MAGNITUDE ADVERSARY

§6 measures localized-magnitude recoverability against *fixed* filterbanks. The natural objection is that a fixed filterbank is a weak adversary, and that a filterbank learned end-to-end for the task would close the relative-phase-accessibility gap as well as the position one—which would make the asymmetry an artifact of how hard we tried on the amplitude side rather than a property of the representation. This appendix builds that adversary, together with a control that makes the comparison self-contained, and states the outcome rule before the numbers.

Construction. Both arms frame the series identically—window $W = cL$ with $c \in \{1/4, 1/8, 1/16\}$ and hop $W/2$, the windows of §6—and are identical after the front end. Writing u_t for the t-th frame and $\mathcal{R}$ for the real DFT,

$$\begin{aligned} \text{arm M:} \quad & u_t \mapsto \mathbf{W} \log(1 + |\mathcal{R}u_t|), \\ \text{arm R:} \quad & u_t \mapsto \mathbf{W}' \operatorname{sgn}(v_t) \log(1 + |v_t|), \qquad v_t = [\Re\mathcal{R}u_t;\, \Im\mathcal{R}u_t], \end{aligned}$$

with $\mathbf{W} \in \mathbb{R}^{K \times n_b}$, $\mathbf{W}' \in \mathbb{R}^{K \times 2n_b}$, $n_b = \lfloor W/2 \rfloor + 1$, both learned. Because v_t is a linear bijection of u_t, arm R is the same architecture with the magnitude removed and nothing else changed: it is the phase-preserving twin, not a different model. Every parameter of arm M is downstream of $|\cdot|$, so arm M is a function of the per-frame magnitude spectrum alone. That invariance is *imposed*, not hoped for, and is asserted numerically: the maximum change of arm M's output when the contents of an analysis frame are circularly rotated is 5.0×10^{-16} in double precision and 3.6×10^{-7} in the single precision that trains, against a control parameterization that must and does fail the same test by 0.53. Arm R must also fail it, and does (1.36).

One consequence deserves stating plainly, because it makes the adversary stronger than the phrase "learned filterbank" suggests. When $n_b \leq K$—102 of the 114 datasets—$\mathbf{W}$ has rank n_b almost surely, so the composition of $\mathbf{W}$ with the first trunk convolution spans every convolution that could be applied to $\log(1 + |\mathcal{R}u_t|)$ directly. The mix is then redundant: the trunk receives the *complete* per-frame magnitude spectrum losslessly, with no filterbank bottleneck at all, and gradient descent has no reason to move $\mathbf{W}$ from its initialization. Inspection of the released weights confirms exactly that—on the datasets we examined the entries are statistically indistinguishable from their initial distribution. This is the intended behaviour, not a training failure. On the remaining 12 datasets $n_b > K$ and the mix does discard information; none of those 12 is in the relative-phase stratum (seven no-detected-gap, five position), and excluding them leaves the result below unchanged ($C(\text{rel}) - C(\text{pos}) = +0.069$, $P(\leq 0) = 0.022$).

Protocol. All 114 datasets at full resolution. The diagnostic's A_{amp} and A_{full} are computed on full-resolution signals, so any decimation here would compare two different signals; an automatic guard refuses every dataset whose length disagrees with the diagnostic's, and it excluded none. The configuration (c, K), $K \in \{32, 64\}$, is chosen *per arm* by three-fold cross-validation on the training set alone; five seeds; 250 epochs; Adam at 10^{-3}, weight decay 10^{-4}, batch $\min(32, n)$—

the settings of the paper's 1D CNN arm, so that only the front end differs. The trunk pools with `AdaptiveMaxPool1d(4)` rather than to a single value: a global maximum after shift-equivariant convolutions is shift-invariant, which would blind the trunk to precisely what the position stratum measures. A planted pure-position task gates that, alongside a planted amplitude-only task that gates learnability—a model that cannot learn also reports no cost, for the wrong reason.

Outcome rule, fixed in advance. The headline is the arm-internal cost of the magnitude, $C = \overline{A}_{\mathrm{R}} - \overline{A}_{\mathrm{M}}$ per stratum, which involves no ceiling and no model zoo. C clearly larger on the relative-phase stratum than on the position stratum confirms the scoping of §6; the two indistinguishable would mean the scoping is too narrow and must be widened.

Table 2: Arm-internal cost of discarding within-frame phase, 114 datasets, five seeds per arm. C is a mean over a heavy-tailed per-dataset distribution; the median and the count of datasets with a positive cost are given beside it, and the bootstrap interval is over datasets.

Stratum	n	$\overline{A}_{\mathrm{R}}$	$\overline{A}_{\mathrm{M}}$	C	median	$C > 0$
no detected gap	52	0.730	0.749	-0.019	-0.008	21
position	39	0.768	0.778	-0.009	$+0.000$	18
rel-phase	23	0.814	0.748	$+0.066$	$+0.034$	15

Result (Table 2). Arm R carries twice arm M's front-end parameters and tends to lose where phase is irrelevant, which is what the no-detected-gap stratum shows and why we use it as the control stratum. Differencing against it assumes the architectural handicap is the same in every stratum—that parameter count, optimization and inductive bias cost arm R equally on position, relative-phase and no-detected-gap tasks—which we have not verified, and no-detected-gap does not mean phase-free. With that assumption stated, $C(\text{position}) - C(\text{no-detected-gap}) = +0.010$ (bootstrap 95% CI $[-0.018, +0.036]$, $P(\leq 0) = 0.24$) and $C(\text{rel-phase}) - C(\text{no-detected-gap}) = +0.085$ ($[+0.017, +0.166]$, $P(\leq 0) = 0.005$); directly, $C(\text{rel-phase}) - C(\text{position}) = +0.075$ ($[+0.005, +0.158]$, $P(\leq 0) = 0.016$; label permutation $p = 0.012$; rank test $p = 0.041$). On the position stratum we detect no cost beyond the no-detected-gap stratum's (two-sided $p = 0.50$); that is an absence of evidence, not a demonstrated equivalence, and the interval is wide enough to hold a modest cost. On the relative-phase stratum the cost is significantly larger. None of this depends on A_{full}, on a ceiling, or on the model zoo.

The sharpest single instance is the dataset the diagnostic ranks first for Δ_{rel}. On *TwoPatterns*, a front end that discards within-frame phase reaches 0.315 (five seeds, 0.303–0.329; chance 0.25), while the identical architecture keeping phase reaches 0.999 (0.9985–0.9990)—same trunk, same framing family, same optimizer, same seeds, same train-side selection, one operation removed.

What the number does not support. The effect is heavy-tailed. The five largest relative-phase datasets carry 110% of that stratum's total cost (*TwoPatterns* $+0.684$, *SyntheticControl* $+0.342$, *SmoothSubspace* $+0.251$, *UMD* $+0.221$, *UWaveGestureLibraryX* $+0.175$), and dropping the two largest leaves $C(\text{rel-phase}) = +0.024$ with an interval crossing zero. What keeps this from being an outlier artifact is that the whole distribution moves: the median is $+0.034$ against position's $+0.000$, 15 of 23 are positive, the 10%-trimmed mean is $+0.042$ against -0.003, and the rank test—which ignores magnitudes—gives $p = 0.041$. The mean should be quoted with the median and the sign count, never alone. Restricting to the 54 datasets where both arms' cross-validations happened to select the same (c, K) gives $+0.030$ with an interval crossing zero; that conditioning is post-hoc, since the selections differ precisely because the arms differ, and we report it. Finally, $\overline{A}_{\mathrm{M}}$ is a model, not a ceiling: it falls below the global magnitude reference A_{amp} on 33 of 114 datasets, but 29 of those are in the no-detected-gap stratum (the two 52-class *Pig* datasets with very small training sets are the extremes), against 3 of 39 position and 1 of 23 relative-phase. Read against the zoo instead of against arm R, the comparison breaks its own ceiling—arm M exceeds A_{full} on 48 of 114—which is why C and not the zoo-anchored fraction is the headline.

The window-selected variant. Choosing the spectrogram window per dataset by *test* accuracy sharpens the fixed-window contrast of §6 further, to -7% against 31%. That is a test-side selection; §6 reports fixed windows, and the variant is recorded here only so that the sensitivity of the contrast to the adversary's strength is visible in one place.

Reproducibility. Non-deterministic GPU kernels propagate into the cross-validation $\arg\max$: *TwoPatterns* selected $c = 1/4$ in a smoke run and $c = 1/8$ in the full run from identical code, data and seed. Selection remains train-side only and the two configurations put arm M at 0.315 and 0.341, but the released CSV is not bit-reproducible from a rerun.

A pilot we discarded, and why it is reported. A first version of this experiment used a front end of the form $u_t \mapsto \sqrt{\langle u_t, a_k\rangle^2 + \langle u_t, b_k\rangle^2}$ with a_k, b_k learned independently—the shape a reader is likely to write down for "magnitude of a learned filtering". Nothing ties (a_k, b_k) into a quadrature pair, so its only invariance is $x \mapsto -x$: under within-frame rotation a true quadrature filterbank changes by 0.000 and this one by 0.48. Worse, the hypothesis class contains the rectified raw signal—taking a_k a unit impulse at tap k and $b_k = 0$ yields exactly $\log(1 + |x_t|)$ at full temporal resolution—and that setting passes a negation-invariance self-test at exactly zero, while a linear probe on $\log(1 + |x|)$ separates a chirp from its time reversal at 1.000 where the global magnitude reaches 0.521. The pilot also decimated series longer than 512 samples against full-resolution baselines. Its numbers are not reported because they measure a sign-invariant near-raw model rather than a localized magnitude. We record it because the failure is easy to reproduce and hard to see: a front end most readers would call magnitude-only can retain nearly all relative phase, and the obvious self-test does not detect it. Both halves of that claim regenerate from released scripts that need no data and no GPU (`pilot_frontend_check.py`, `verify_v2_invariance.py`). Within-frame rotation invariance, not sign invariance, is the property worth asserting.

Table 3: The phase-dependence typology of the UCR archive (114 datasets; main bispectrum-enriched repertoire—UCR-informed, frozen before every external run, Appendix G; the Φ_2-only floor variant is Appendix M). Formally Δ_{rel} is the repertoire-recoverable shift-invariant phase component—biased low, with no one-sided guarantee—and Δ_{align} the circular-shift-sensitive one, with ε the estimator term of Equation (3); "rel-phase"/"position" are their shorthands (§2.1). Because the enriched repertoire was chosen after inspecting UCR, this membership is partly exploratory; the confirmatory evidence is the frozen Φ_2 variant, the held-out synthetic families, the intervention and UEA. Train-selection verdicts: 33 positive / 24 no-gap / 53 inconclusive / 4 failure (§6). Columns are stratum means; "CNN gain" is over the shared A_{amp}, and "CNN reaches A_{full}" is descriptive only—no inferential claim uses it (§4).

Type	n	Δ_ϕ	Δ_{align}	Δ_{rel}	ε	CNN gain	CNN reaches A_{full}
no detected gap	52	-0.010	-0.014	0.000	$+0.004$	-0.012	90%
position	39	0.123	0.107	0.008	$+0.008$	$+0.011$	59%
rel-phase	23	0.209	0.024	0.181	$+0.004$	$+0.147$	74%

Table 4: Table 4 in full—Spearman ρ and the two paired-difference rows restored: the gap predicts gain over the estimated amplitude reference across model families (114 datasets, no gating; partials control length and class count, $p \le 1.1\times10^{-11}$; last column: family-clustered bootstrap CIs, §6). "Low"/"high gap": mean gain for $\Delta_\phi < 0.03$ (n=52) / rest (n=62). Rows use the shared A_{amp} except 1-NN DTW† (own amplitude twin, sharing no quantity with other rows); the last two rows are paired *differences* against the real CNN.

Family	Pearson r	Spearman ρ	Partial r	Low gap	High gap	Cluster 95% CI
1D CNN	$+0.513$	$+0.386$	$+0.525$	-0.012	$+0.062$	$[+0.29, +0.69]$
MiniRocket	$+0.720$	$+0.623$	$+0.769$	$+0.079$	$+0.226$	$[+0.56, +0.86]$
InceptionTime	$+0.645$	$+0.520$	$+0.658$	$+0.036$	$+0.156$	$[+0.46, +0.79]$
Complex CNN	$+0.563$	$+0.433$	$+0.580$	-0.042	$+0.050$	$[+0.34, +0.72]$
1-NN DTW†	$+0.593$	$+0.495$	$+0.599$	-0.019	$+0.132$	$[+0.39, +0.74]$
Complex − real (mag. head)	$+0.276$	$+0.220$	$+0.276$	-0.031	-0.012	$[+0.07, +0.44]$
Analytic real − real	$+0.416$	$+0.430$	$+0.427$	$+0.008$	$+0.060$	$[+0.22, +0.60]$

L Cross-fitted replication

Both axes of §4 are evaluated on the same official test split, so correlated test-sampling noise could in principle contribute to the association even after the calibrations of §6. Here every dataset's test

set is split into stratified halves; Δ_ϕ is estimated on one half and every family's gain on the other, both directions averaged, over 20 random splits (arms and models are fit once on the training set; only the evaluation samples are disjoint).

The correlation is computed across datasets *within* each split and direction and then aggregated, with the low-gap stratum defined on the estimation half and evaluated on the other. Over 20 splits in both directions on 113 datasets (***SmoothSubspace*** is too short for MiniRocket's kernels): MiniRocket $+0.700$ (sd 0.030 over split-directions, $[+0.625, +0.744]$) against its same-half $+0.729$, and 1-NN DTW $+0.526$ (sd 0.027) against $+0.590$. Both rows use the pre-mask decomposition, so the same-half figures are its Δ_ϕ, not the masked one of Appendix M; the mask moves every same-half correlation here by at most 0.009. The cross-fitted low-gap stratum still separates cleanly: MiniRocket $+0.035$ low against $+0.161$ high, DTW -0.018 against $+0.130$. So the association survives the separation, and a few hundredths of the same-half correlation is shared test-sampling noise—the direction one should expect, and smaller than the effect. Only these two families' per-sample predictions are recomputable on CPU; the three deep families need prediction dumps this run did not have, so their cross-fitted values are unmeasured. One pitfall for reimplementations: averaging the gap and the gain over both halves before correlating reproduces the full-test-set quantities exactly and defeats the control, which is why `crossfit.py` keeps one row per (dataset, split, direction).

M Decomposition robustness: invariant repertoire and shift protocol

Two identification concerns from §2.1 are tested. *(i) Repertoire:* the main decomposition uses the bispectrum-enriched invariant set; here we evaluate the cheaper Φ_2-only *floor* variant obtained by dropping the bispectrum arm, quantifying how much of the relative-phase estimate the richer, sign-sensitive representation contributes (the bispectrum slice is exactly shift-invariant and, unlike Φ_2, sensitive to global sign and to phase rotation; it is generically complete up to continuous translation, not up to the cyclic group, as the next paragraph shows). *(ii) Shift protocol:* circular shifting is replaced by reflection-padded translation (no wrap-around).

What the invariant arms determine, and what they do not. Both invariants fix the phases of $\mathcal{F}x$ only up to the transformations that leave them unchanged, and those form a larger set than the cyclic group. The bispectrum slice $B_k = F_k F_1 \overline{F_{k+1}}$, used over the non-negative frequencies only, determines every phase recursively from φ_1, so generically (no vanishing bin in the chain) it pins the signal down to a phase ramp e^{ibk}—a translation—but no term in it ever enforces $e^{ibL} = 1$. For even length a non-vanishing Nyquist bin must stay real, which forces $b \in 2\pi\mathbb{Z}/L$ and leaves only integer circular shifts; for odd length—29 of the 114 datasets, including four of the core seven—any real b is admissible, and the fiber also contains band-limited *sub-sample* translations. Φ_2, a second difference, fixes phases only up to $a + bk$ and is additionally blind to a constant rotation $x \mapsto x\cos c + \mathcal{H}x \sin c$ of every phase, of which negation ($c = \pi$) is one case; the bispectrum slice is not. `invariance_counterexample.py` verifies both on the released feature functions: at $L = 9$, a 0.53-sample translation leaves $|\mathcal{F}x|$, Φ_2 and the slice unchanged to 4×10^{-16} while the signal lies 2.24 in L_2 from every integer shift and sign of the original, and a rotation by $c = 1.1$ leaves Φ_2 unchanged while moving the slice by 1.0. What this changes is attribution, and in a known direction. For a spectrum whose support has no gap along the chain, the only information the main repertoire's invariant arms discard beyond the cyclic group is sub-sample *position*—same events, same order—since the phase-rotation blind spot of Φ_2 is available through the bispectrum arm; that information enters Δ_{align}, where position belongs. The condition is not generic in practice: the 10^{-10} mask removes near-empty bins, and where the support has gaps the recursion breaks and genuine inter-frequency phase changes become invisible as well. A construction shows it: at $L = 17$ with support $\{2, 5, 8\}$, moving the phase of bin 5 alone leaves $|\mathcal{F}x|$, Φ_2 and the slice unchanged to 2×10^{-15} while the signal lies 1.96 from every integer shift and sign of the original. Relative-phase structure on sparse or gapped spectra can therefore be missed by the invariant arms and reaches A_{shift} only if the shifted-raw arm recovers it, which nothing guarantees. The bias on Δ_{rel} is still downward, but attributing the residue to position is no longer safe, and we have not calibrated it on gapped-support tasks. In the Φ_2-only floor variant the rotation blind spot is real, which makes that variant more conservative about Δ_{rel} than we described, not less. One triple product whose three non-negative-frequency indices sum to L—for odd L, $F_1 F_m^2$ with $m = (L-1)/2$—carries the ramp e^{ibL} and restores the cyclic group: the same script shows integer shifts move it by 0 and the

sub-sample one by 2.95 rad at $L = 9$. We release it but did not rerun the decomposition with it, so every result above is for the repertoire as run.

The vanishing-magnitude convention is a choice. Φ_2 takes $\angle 0 = 0$ on bins whose magnitude is numerically zero, with no thresholding. This makes the map total and threshold-free, but it is a convention rather than a neutral default—and it is not a continuity fix: $(\cos \angle z, \sin \angle z)$ has no continuous extension at $z = 0$, so the convention selects one value at the singularity rather than removing it. Two consequences follow. *First*, on signals whose label-bearing structure concentrates in low-magnitude bins, those bins contribute a constant rather than their (ill-conditioned) phase, so the invariant is biased toward the high-energy part of the spectrum. The *lowbin* held-out family of §6 was designed to expose exactly this and does (detection 0.35); a magnitude-weighted or whitened phase representation is the natural fix and is left to future work. That bias is one-sided—it can only cause Δ_{rel} to *understate* relative-phase structure, never to manufacture it—*provided* the repertoire really is invariant on the data at hand, which is the second consequence. *Second*, the algebraic shift-invariance of Φ_2 and of the bispectrum arm does not survive in floating point wherever a bin sits at the rounding floor, and on this archive that is the common case rather than the exception. We measured it: circularly shifting every series and recording the largest change in each representation, $|\mathcal{F}x|$ moves by at most 9×10^{-13} on all 114 datasets, while Φ_2 and the bispectrum arm move by 2.000—the full range of a $(\cos, \sin)$ feature—on **91 of 114**. The mechanism is not the $\angle 0 = 0$ convention, which fires only on exactly-zero bins and is harmless because a circular shift leaves those bins exactly zero: it is bins that are near zero but *not* zero, whose phase is rounding noise and therefore does not rotate correctly under the shift. Reproducing that condition synthetically, the damage is confined but real—six of Φ_2's 95 features move and four of the bispectrum arm's 97, namely the three difference terms touching the offending bin in each of the $\cos$ and $\sin$ blocks, while every other feature stays invariant to 7×10^{-15}; in norm the feature vector moves by 3.8% and 2.7%. **The one-sided reading does not survive on the unmasked repertoire**: Δ_{rel} can in principle be inflated there, because a classifier is free to use a feature that varies with the shift it is supposed to be invariant to. That is why the decomposition reported in this paper is the *masked* one; the rest of this section is the measurement. The fix is the magnitude-weighted representation named above, now not optional: mask bins below a relative floor out of the phase block, leaving the magnitude block untouched. Because the mask is computed from $|\mathcal{F}x|$, which is exactly shift-invariant, the mask is shift-invariant too and the masked entries take a determinate value rather than a noisy one—which is what makes the masked representation invariant in floating point rather than only in algebra. On the archive it works: at relative floors of 10^{-12}, 10^{-10} and 10^{-8} the count of datasets above 10^{-6} is $0/114$ in every case. Two measurements say the floor is not a tuning knob and the mask is not discarding signal. First, the masked fraction of all bins is 0.4963% at 10^{-12} and 10^{-10} and 0.4964% at 10^{-8}: only 39 bins in the whole archive lie between 10^{-10} and 10^{-8} of their sample's largest, so there is a clean gap between real spectral content and the rounding floor, and any floor across four orders of magnitude selects the same set. Second, that fraction is almost entirely the DC bin, which z-normalization forces to zero: masking DC alone would give 0.4822%, and the measured excess, 0.0142%, is the part that matters. The DC term was in fact already inert—both Φ_2 and the bispectrum arm form their phase from a product in which the DC factor appears, so that product is exactly zero before and after a shift—which locates the defect precisely: $5{,}414$ non-DC bins archive-wide, 0.029 per sample, so at most 2.9% of samples carry one, each contaminating six of Φ_2's 95 features.

Does it change the decomposition? Masked features were noise and a classifier should be indifferent to them, but a model trained on shift-augmented data could in principle use a feature that varies with the augmentation, so we recomputed A_{shift} and the typology under the mask rather than argue the point, fixing in advance that whatever the masked run returned would be adopted. It moves the typology: $52/34/28$ becomes $52/38/24$ (no-detected-gap / position / rel-phase), with seven datasets changing type—*CricketZ* into the rel-phase stratum, *Chinatown*, *EOGVerticalSignal*, *PhalangesOutlinesCorrect* and *WordSynonyms* out of it, *LargeKitchenAppliances* to no-detected-gap and *ProximalPhalanxTW* to position. Those are the numbers this paper reports.

But the movement is threshold sensitivity, not a corrected inflation, and the distinction matters for how much weight the typology can bear. Four measurements say so. *(i)* The change has no direction: Δ_{rel} falls on 46 datasets and rises on 32 (sign test $p = 0.14$), and the mean change is -0.0002 against a mean *absolute* change of 0.0075—two orders of magnitude smaller, so the mask jitters Δ_{rel} rather than deflating it. *(ii)* Every one of the seven movers was already within

0.030 of a typing threshold and four within 0.006; archive-wide $11/114$ datasets sit within 0.005 of one, 18 within 0.01 and 40 within 0.02, so a perturbation of this size moving a handful is what a hard-thresholded rule does. *(iii)* The movers are not the datasets the mask moved most: *Wine*'s Δ_{rel} changes by $+0.063$, the largest in the archive, and it stays rel-phase, while 14 non-movers change by more than 6 of the 7 movers do. *(iv)* There is no dose–response. *ElectricDevices* carries $5{,}224$ of the $5{,}414$ non-DC masked bins—96% of all the masking in the archive—and its Δ_{rel} moves by $+0.002$ with no change of type, while *Chinatown*, which has exactly *one* non-DC bin masked out of $4{,}719$, changes type. The mechanism is not information removal but discrete re-selection: a perturbation below the resolution of the cross-validation objective flips which $(\text{model}, \text{representation})$ pair wins.

Two further checks bound what else could have moved. A_{amp} is mask-independent by construction and differs from the frozen run on 1 of 114 datasets by 2×10^{-4}, which bounds environment and refit drift far below the Δ_{rel} changes; A_{full} legitimately does depend on the mask, through its `d2phi` slot, and differs on 10 datasets (largest *LargeKitchenAppliances*, $0.771 \to 0.723$, whose full arm had selected `d2phi`—the one place in the archive where masking plausibly cost real predictive content rather than noise). And the conclusions that do not route through the typology are unchanged: the seven-member core is rel-typed $7/7$, every rel-phase example named in §3 keeps its type, and re-stratifying the learning curve of Appendix J by the masked types leaves its verdict intact (rel-phase slope $+0.025$, CI $[+0.008, +0.042]$, against $+0.023$ before). `invariance_check.py` and `invariance.csv` carry the per-dataset errors and the 91 names; the check exists because exact invariance in algebra and realised invariance on data are different claims.

Are the two arms of the decomposition matched? A_{full} selects over $\{\texttt{raw}, |\mathcal{F}x|, \texttt{ampphase}, \Phi_2\}$ on n unaugmented rows; the shift arm selects over $\{\texttt{raw}, |\mathcal{F}x|, \Phi_2, \texttt{bispec}\}$ on $4n$ shift-augmented rows. So A_{shift} can exceed A_{full} for reasons unrelated to position, and $\Delta_{\text{align}} < 0$ on 35 of 114 datasets. We recomputed A_{shift} under three further configurations, everything else held: the repertoire matched to A_{full}'s at $4\times$ augmentation, the main repertoire at $1\times$, and both. Negative Δ_{align} falls $35 \to 30 \to 27 \to 21$; of the original 35, 22 disappear, 13 survive and 8 new ones appear, so the net reduction is $14/35$ and the mismatch explains part but not all of it. *Matching the repertoire is not a correction*: replacing the bispectrum with `ampphase`, which the shift randomization destroys, leaves the Φ_2-only floor, and indeed reproduces its archive mean to three decimals ($+0.0252$ against $+0.025$). Dropping an invariant arm makes Δ_{rel} smaller by construction, not fairer. The budget is the part that needed fixing, and is fixed next. `matched_repertoire.py`, `matched_repertoire.csv`.

The amplitude reference was trained on a quarter of the shift arm's rows. For a shift-*invariant* arm the four augmented copies are not four views of the data: the representation is invariant, so the four feature matrices are identical (at most 5.3×10^{-9} apart under the mask). The augmentation adds no information there. It is not a no-op either. `LogisticRegression` minimizes $C\sum_i \ell_i + \frac{1}{2}\|w\|^2$, so four times the rows against a fixed penalty is four times weaker regularization—on a synthetic control with literally duplicated rows the fitted $\|w\|$ goes $3.98 \to 6.09$—and the MLP gets roughly $2.4\times$ the gradient steps. A_{amp} received none of it. The symptom is visible in the paper's own negative control: on the 52 no-detected-gap datasets, where A_{full}—which *has* Φ_2 and trains on unshifted data—could not beat A_{amp}, the unmatched Δ_{rel} is positive on 31 and negative on 14 (sign test $p = 0.016$, mean $+0.0039$). We therefore refit the amplitude arm under the same four-fold augmentation and the same shifted test set, drawing the same shifts in the same order, so the subtraction is paired dataset by dataset and seed by seed. It rises by $+0.005$ (68 of 114 up, sign test $p = 3 \times 10^{-4}$, CI $[+0.0014, +0.0086]$); refitting on shift-randomized data *without* the duplication moves it by -0.002 (40 up, 48 down, $p = 0.46$), so the effect is the duplication, not the randomization. Against that reference the null stratum centres exactly—mean $+0.0002$, 13 positive, 13 negative, 26 *exactly* zero, sign test $p = 1.00$—and the archive mean Δ_{rel} falls from $+0.045$ to $+0.039$. That is the decomposition this paper reports, and ε in Equation (3) is the difference, written separately because it belongs to neither task component. The typology moves by five datasets ($52/38/24 \to 52/39/23$: *Chinatown* and *Meat* in, *Car*, *CricketZ* and *UWaveGestureLibraryAll* out); six of the seven core members keep the type, *UWaveGestureLibraryAll* falling under the ratio threshold with $\Delta_{\text{rel}} = +0.117$ intact. Every stratum contrast survives (Table 3). `amp_budget.py`, `amp_budget.csv`.

What that control also measures: the selection-noise floor. The shift arm's candidate set *contains* $|\mathcal{F}x|$, so a budget-matched amplitude-only arm can beat it only by the cross-validation picking a

worse arm on the training folds. It does, on 24 of 114 datasets, by at most 0.110 and by 0.029 on average among those—an archive-mean cost of 0.006, which is a direct measurement of what a twelve-candidate selection pays over a three-candidate one, and the floor against which every difference above should be read. It lands where it matters least: 13 of the 24 are in the no-detected-gap stratum and 11 in the position stratum, and *none* is in the relative-phase stratum. So Δ_{rel} is biased low for a second, now-quantified reason—it charges the shift arm for its own selection variance—and that charge falls entirely outside the stratum carrying the signal. One tempting cut must be resisted: excluding the 24 leaves the null stratum positive on 13 and negative on 0, which looks like residual bias but is not evidence at all, because a non-violator is by definition a dataset with $\Delta_{\mathrm{rel}} \geq 0$. Conditioning on it conditions on the sign. The valid statement is the full-stratum one above.

Why the circular shift, and what a padded translation would measure instead. The decomposition is defined with respect to the circular-shift group because that is exactly the symmetry Proposition 1 establishes for $|\mathcal{F}x|$: T_s is a group action on $\mathbb{R}^L$ that leaves the amplitude spectrum invariant, so averaging over it isolates the shift-invariant part of the signal's label information without changing the reference A_{amp}. A padded (non-circular) translation does not have that property. It is not a group action on fixed-length series—composing two translations is not a translation once content reaches an edge—and, more importantly, it *removes information*: whatever leaves the crop is destroyed rather than relocated, so the resulting accuracy drop confounds "position was uninformative" with "part of the signal was deleted". *TwoPatterns* makes this concrete: its classes are defined by the order of two transients, and a padded translation that pushes the first transient off the left edge changes the label-relevant content, whereas a circular shift moves it to the right edge and preserves it. We therefore report the circular-shift decomposition throughout, and note that a boundary-aware variant would define a different—and strictly lossy—estimand. Other invariant repertoires or group actions may attribute the components differently; Δ_{rel} is biased low relative to the repertoire we fix, with no one-sided guarantee (§2.1).

Repertoire (the conservative floor). Removing the bispectrum arm leaves the cheaper Φ_2-only repertoire used during development; it yields a strictly conservative variant of the main decomposition: $\mathrm{corr}(\Delta_{\mathrm{rel}}^{\Phi_2}, \Delta_{\mathrm{rel}}^{\mathrm{main}}) = +0.78$, archive mean $+0.025$ versus $+0.039$, and a 12-member rel-phase set of which 10 are in the budget-matched main 23 (the cross-setting core of seven is inside it). The floor variant was computed before the budget matching of §2.1, so the subset relation is no longer exact: *Car* and *UWaveGestureLibraryAll* are the two exceptions, both of which lose the type at the ratio threshold rather than by Δ_{rel} vanishing. The largest difference is CBF ($+0.13$ under Φ_2 versus $+0.40$ under the main repertoire), whose cylinder–bell–funnel *shape order* is exactly the global-sign-bearing structure the sign-blind Φ_2 cannot represent; BME and Wine behave likewise, and all three are typed *position* under the floor—the floor is conservative in exactly the way its blindness predicts. The direction is the typical one but not a guarantee: enriching the repertoire usually moves mass from Δ_{align} to Δ_{rel}, and on this archive the bispectrum raises held-out A_{shift} on 38 datasets, lowers it on 15 and leaves 61 unchanged (§2.1). Finite-model accuracy differences carry no one-sided guarantee, so "biased low" is a statement about the repertoire-relative estimand, not a claim that the empirical Δ_{rel} can only rise; the floor variant is the right default only when compute is tight, and every conclusion of the paper holds under it.

Protocol. Reflection-padded translation behaves qualitatively differently ($\mathrm{corr} = +0.46$ with the circular estimate; archive mean falls to -0.048; core five of seven retained, with TwoPatterns and UWaveGestureLibraryAll lost), and the mechanism is instructive rather than worrying: padded translation is *not a group action on the observed window*—content genuinely exits the crop, destroying label information for reasons unrelated to phase structure (for TwoPatterns, the class-defining pattern pair can partially leave the window, making the label ill-posed on some crops), which depresses *every* arm, amplitude included. Circular shifting, by contrast, is exactly the fiber transformation of Proposition 1: content-preserving, and the correct symmetry for the decomposition. The idealized target it estimates is A_{inv}^{*}, the Bayes accuracy achievable from the maximal invariant of the circular-shift orbit; A_{shift} approaches it only as the invariant repertoire approaches that maximal invariant, so $\Delta_{\mathrm{rel}} = A_{\mathrm{shift}} - A_{\mathrm{amp}}^{\mathrm{aug}}$ targets from below, as far as the repertoire is concerned, the shift-invariant Bayes gap $A_{\mathrm{inv}}^{*} - A_{\mathrm{amp}}^{*}$ and Δ_{align} the residual up to global-magnitude Bayes accuracy—a decomposition of *estimates*, exact in algebra, repertoire-relative in semantics. The padded variant therefore does not measure the same quantity; we report it as a demonstration that Δ_{align} prices *circular*-shift-destroyed information—matching the theory—rather than lossy-translation robustness.

N PER-DATASET RESULTS

Table 5: The 23 rel-phase datasets under the main (bispectrum-enriched) repertoire with the low-magnitude phase mask and the budget-matched amplitude reference of Appendix M, sorted by Δ_{rel}. Bold: the seven-member cross-setting core (identical under thresholds, shift seeds and the Φ_2-only conservative variant; six of seven here, five under a disjoint estimator family—*UWaveGestureLibraryAll* falls just under the ratio threshold with $\Delta_{\mathrm{rel}} = +0.117$). The Φ_2-only variant yields 12 members, 10 of them here (Appendix M).

Dataset	Δ_{rel}	Δ_{ϕ}
TwoPatterns	+0.413	+0.436
CBF	+0.396	+0.338
BME	+0.360	+0.287
SmoothSubspace	+0.340	+0.533
Wine	+0.322	+0.259
Lightning7	+0.274	+0.260
UWaveGestureLibraryX	+0.251	+0.311
SyntheticControl	+0.225	+0.377
UWaveGestureLibraryY	+0.177	+0.235
UWaveGestureLibraryZ	+0.175	+0.250
UMD	+0.161	+0.278
FreezerSmallTrain	+0.150	+0.053
FreezerRegularTrain	+0.122	+0.132
EOGHorizontalSignal	+0.112	+0.155
OliveOil	+0.100	+0.067
SmallKitchenAppliances	+0.094	+0.056
Chinatown	+0.093	+0.131
CricketY	+0.092	+0.156
Crop	+0.068	+0.133
Meat	+0.067	+0.067
MelbournePedestrian	+0.065	+0.136
CricketX	+0.060	+0.113
Trace	+0.050	+0.040

The complete per-dataset table follows; the same numbers are released as `amp_budget.csv` (gap and decomposition) and `step3_models.csv`, `step3_gains_complex.csv`, `dtw_arm.csv` (family accuracies) with the code. Column definitions: A_{amp}, A_{shiftinv}, A_{full}, the gap, its position and relative-phase components, and the type are as in §2 (decomposition columns use the main bispectrum-enriched repertoire; the floor variant ships in `shift_robust.csv`). The CNN gain is measured against the shared estimated reference A_{amp} (matching Table 4); the MiniRocket and InceptionTime gains against their own $|\mathcal{F}x|$ references (the per-family definition of Appendix E); the Complex column is the capacity-matched complex-minus-real difference of §5.

Table 6: Full per-dataset results (114 UCR datasets; v2 nested-CV protocol). Decomposition columns use the bispectrum-enriched repertoire under the low-magnitude phase mask and the budget-matched amplitude reference (Appendix M); $\Delta_{\mathrm{al}} + \Delta_{\mathrm{rel}}$ falls short of Δ_{ϕ} by the estimator term ε of Equation (3). Gain definitions: CNN vs. the shared estimate A_{amp}, matching Figure 2 and Table 4; Rocket, Incep. and DTW vs. their own $|\mathcal{F}x|$ accuracies (the per-family definition of Appendix E), so those three columns are not comparable with the CNN column; Complex is the capacity-matched complex−real difference.

Dataset	#cls	L	A_{amp}	A_{shift}	A_{full}	Δ_{ϕ}	Δ_{al}	Δ_{rel}	type	CNN	Rocket	Incep.	DTW	Cx
ACSF1	10	1460	0.620	0.650	0.770	0.150	0.120	0.000	position	0.070	0.300	0.093	0.100	-0.120
Adiac	37	176	0.737	0.751	0.737	0.000	-0.014	0.000	no-gap	-0.130	0.079	-0.327	-0.130	-0.172
ArrowHead	3	251	0.526	0.518	0.754	0.229	0.237	-0.000	position	-0.042	0.274	0.257	-0.040	-0.149
BME	3	128	0.593	0.953	0.880	0.287	-0.073	0.360	rel-phase	0.151	0.367	0.353	0.360	+0.193
Beef	5	470	0.667	0.700	0.667	0.000	-0.033	0.040	no-gap	-0.044	0.033	-0.100	-0.067	-0.078
BeetleFly	2	512	0.950	0.950	0.950	0.000	0.000	0.000	no-gap	-0.067	-0.100	-0.167	-0.050	-0.117
BirdChicken	2	512	0.900	0.900	0.900	0.000	0.000	0.000	no-gap	0.017	-0.050	-0.033	-0.100	-0.100
CBF	3	128	0.551	0.950	0.889	0.338	-0.061	0.396	rel-phase	0.442	0.448	0.447	0.438	+0.007
Car	4	577	0.583	0.640	0.650	0.067	0.010	0.000	position	0.011	0.333	0.317	0.017	+0.094
Chinatown	2	24	0.846	0.891	0.977	0.131	0.085	0.093	rel-phase	0.133	0.108	0.105	0.172	-0.016
ChlorineConcentration	3	166	0.805	0.832	0.887	0.082	0.055	0.000	position	-0.124	-0.051	-0.231	-0.009	-0.111
CinCECGTorso	4	1639	0.902	0.944	0.961	0.059	0.017	0.035	position	-0.299	-0.073	-0.150	-0.049	-0.010
Coffee	2	286	1.000	1.000	0.964	-0.036	-0.036	0.000	no-gap	0.000	0.000	0.000	0.000	+0.000

Dataset	#cls	L	A_{amp}	A_{shift}	A_{full}	Δ_ϕ	Δ_{al}	Δ_{rel}	type	CNN	Rocket	Incep.	DTW	Cx
Computers	2	720	0.640	0.654	0.624	-0.016	-0.030	0.016	no-gap	0.063	0.100	0.052	-0.020	-0.036
CricketX	12	300	0.477	0.558	0.590	0.113	0.032	0.060	rel-phase	0.095	0.338	0.318	0.249	-0.025
CricketY	12	300	0.456	0.550	0.613	0.156	0.063	0.092	rel-phase	0.056	0.372	0.329	0.269	-0.009
CricketZ	12	300	0.479	0.546	0.600	0.120	0.054	0.031	position	0.108	0.349	0.283	0.156	-0.054
Crop	24	46	0.601	0.678	0.734	0.133	0.056	0.068	rel-phase	0.082	0.133	0.128	0.111	-0.015
DiatomSizeReduction	4	345	0.853	0.876	0.899	0.046	0.023	-0.000	position	-0.020	0.075	0.074	0.029	+0.149
DistalPhalanxOutlineAgeGroup	3	80	0.727	0.758	0.784	0.058	0.026	0.023	position	0.029	0.007	-0.058	0.144	-0.024
DistalPhalanxOutlineCorrect	2	80	0.754	0.770	0.754	0.000	-0.017	0.015	no-gap	0.018	0.025	0.008	0.000	+0.000
DistalPhalanxTW	6	80	0.662	0.656	0.698	0.036	0.042	-0.000	position	0.005	-0.007	-0.019	-0.022	+0.007
ECG200	2	96	0.840	0.840	0.840	0.000	0.000	-0.010	no-gap	-0.010	0.070	0.073	-0.060	-0.023
ECG5000	5	140	0.927	0.932	0.936	0.009	0.004	0.000	no-gap	0.011	0.015	0.012	0.000	+0.003
ECGFiveDays	2	136	0.998	0.998	0.998	0.000	0.000	0.000	no-gap	-0.130	0.002	-0.001	-0.207	+0.001
EOGHorizontalSignal	12	1250	0.246	0.399	0.401	0.155	0.001	0.112	rel-phase	0.282	0.307	0.296	0.279	-0.054
EOGVerticalSignal	12	1250	0.271	0.334	0.439	0.169	0.105	0.056	position	0.176	0.271	0.235	0.224	-0.058
Earthquakes	2	512	0.748	0.747	0.748	0.000	0.001	-0.001	no-gap	0.017	-0.022	-0.029	-0.029	-0.012
ElectricDevices	7	96	0.595	0.625	0.637	0.042	0.012	0.029	position	0.103	0.148	0.108	0.080	+0.000
EthanolLevel	4	1751	0.354	0.354	0.740	0.386	0.386	0.000	position	-0.012	0.206	0.273	-0.022	+0.084
FaceAll	14	131	0.820	0.887	0.809	-0.011	-0.077	0.058	no-gap	0.081	-0.014	-0.026	0.105	-0.032
FaceFour	4	350	0.545	0.511	0.864	0.318	0.352	0.023	position	-0.061	0.443	0.405	0.091	+0.008
FacesUCR	14	131	0.782	0.806	0.768	-0.014	-0.038	0.027	no-gap	0.024	0.179	0.116	0.154	-0.040
FiftyWords	50	270	0.374	0.397	0.657	0.283	0.260	-0.000	position	0.103	0.409	0.291	0.297	+0.067
Fish	7	463	0.806	0.806	0.851	0.046	0.046	-0.022	position	-0.072	0.171	0.175	0.074	-0.284
FordA	2	500	0.936	0.957	0.952	0.015	-0.005	0.019	no-gap	-0.045	0.011	0.015	-0.324	+0.001
FordB	2	500	0.817	0.816	0.817	0.000	0.001	0.000	no-gap	-0.080	0.002	0.013	-0.114	+0.018
FreezerRegularTrain	2	301	0.827	0.966	0.959	0.132	-0.007	0.122	rel-phase	0.160	0.173	0.170	0.103	-0.004
FreezerSmallTrain	2	301	0.712	0.875	0.765	0.053	-0.110	0.150	rel-phase	0.207	0.256	0.168	0.085	-0.002
Fungi	18	201	0.565	0.566	0.860	0.296	0.295	0.000	position	0.199	0.435	0.435	-0.113	-0.025
GunPoint	2	150	0.880	0.861	0.933	0.053	0.072	-0.005	position	0.111	0.113	0.118	-0.027	-0.069
GunPointAgeSpan	2	150	0.864	0.859	0.937	0.073	0.078	-0.034	position	0.133	0.133	0.105	0.035	-0.020
GunPointMaleVersusFemale	2	150	0.981	0.987	0.981	0.000	-0.006	0.006	no-gap	0.017	0.019	0.005	0.019	-0.015
GunPointOldVersusYoung	2	150	0.914	0.890	0.946	0.032	0.056	-0.034	position	0.037	0.076	-0.047	0.089	+0.016
Ham	2	431	0.591	0.583	0.743	0.152	0.160	-0.006	position	0.038	0.114	0.127	-0.086	+0.029
HandOutlines	2	2709	0.832	0.841	0.900	0.068	0.059	-0.010	position	-0.140	0.103	0.105	0.041	+0.086
Haptics	5	1092	0.419	0.439	0.438	0.019	-0.001	-0.001	no-gap	-0.043	0.110	0.076	-0.033	+0.051
Herring	2	512	0.578	0.578	0.578	0.000	0.000	-0.000	no-gap	0.016	0.109	0.068	-0.062	-0.005
HouseTwenty	2	2000	0.798	0.815	0.798	0.000	-0.017	-0.000	no-gap	-0.025	0.118	0.112	0.336	-0.003
InlineSkate	7	1882	0.382	0.395	0.382	0.000	-0.013	-0.000	no-gap	-0.064	0.095	0.093	0.084	-0.111
InsectEPGRegularTrain	3	601	0.827	0.841	0.827	0.000	-0.014	0.000	no-gap	0.165	0.108	0.115	0.052	-0.015
InsectEPGSmallTrain	3	601	0.839	0.940	0.643	-0.197	-0.297	0.100	no-gap	0.130	0.024	0.032	0.149	-0.025
InsectWingbeatSound	11	256	0.515	0.524	0.640	0.126	0.117	-0.007	position	-0.236	0.158	0.088	-0.001	+0.101
ItalyPowerDemand	2	24	0.870	0.911	0.963	0.093	0.052	0.033	position	0.090	0.070	0.075	0.089	+0.006
LargeKitchenAppliances	3	720	0.723	0.813	0.723	0.000	-0.090	0.110	no-gap	0.100	0.147	0.160	-0.064	-0.091
Lightning2	2	637	0.639	0.672	0.770	0.131	0.098	-0.016	position	0.148	0.131	0.142	0.246	-0.033
Lightning7	7	319	0.493	0.767	0.754	0.260	-0.014	0.274	rel-phase	0.146	0.274	0.274	0.301	+0.059
Mallat	8	1024	0.893	0.897	0.893	0.000	-0.003	-0.000	no-gap	0.025	0.035	0.031	0.035	+0.016
Meat	3	448	0.867	0.917	0.933	0.067	0.017	0.067	rel-phase	-0.200	0.100	0.067	0.033	-0.100
MedicalImages	10	99	0.671	0.683	0.730	0.059	0.047	0.027	position	0.067	0.120	0.074	0.113	-0.028
MelbournePedestrian	10	24	0.756	0.820	0.891	0.136	0.072	0.065	rel-phase	0.129	0.135	-0.028	0.088	-0.015
MiddlePhalanxOutlineAgeGroup	3	80	0.610	0.562	0.578	-0.033	0.016	-0.044	no-gap	-0.074	-0.058	-0.030	0.006	+0.061
MiddlePhalanxOutlineCorrect	2	80	0.797	0.802	0.815	0.017	0.012	0.000	no-gap	0.018	0.055	-0.081	-0.048	+0.003
MiddlePhalanxTW	6	80	0.571	0.591	0.591	0.019	0.000	-0.003	no-gap	-0.032	-0.045	-0.074	0.039	+0.054
MixedShapesRegularTrain	5	1024	0.886	0.896	0.899	0.013	0.003	0.009	no-gap	0.018	0.085	0.079	0.086	-0.021
MixedShapesSmallTrain	5	1024	0.797	0.827	0.797	0.000	-0.030	-0.000	no-gap	0.030	0.121	0.071	0.089	-0.082
MoteStrain	2	84	0.692	0.607	0.874	0.182	0.267	-0.000	position	0.128	0.234	0.214	0.155	-0.028
NonInvasiveFatalECGThorax1	42	750	0.815	0.831	0.908	0.093	0.077	0.001	position	0.049	0.134	0.055	0.014	-0.057
NonInvasiveFatalECGThorax2	42	750	0.850	0.877	0.935	0.085	0.058	0.009	position	0.031	0.114	0.017	-0.007	-0.034
OSULeaf	6	427	0.793	0.788	0.793	0.000	0.006	-0.000	no-gap	0.041	0.153	0.044	-0.037	-0.139
OliveOil	4	570	0.833	0.933	0.900	0.067	-0.033	0.100	rel-phase	-0.467	0.100	-0.356	0.067	+0.033
PhalangesOutlinesCorrect	2	80	0.763	0.804	0.826	0.063	0.022	0.030	position	0.033	0.069	0.020	-0.016	-0.025
Phoneme	39	1024	0.246	0.271	0.141	-0.105	-0.130	0.000	no-gap	-0.065	0.048	0.007	0.075	+0.010
PigAirwayPressure	52	2000	0.269	0.276	0.269	0.000	-0.007	0.000	no-gap	-0.210	0.394	0.135	-0.173	-0.024
PigArtPressure	52	2000	0.505	0.562	0.505	0.000	-0.058	0.000	no-gap	-0.296	0.476	0.479	-0.409	-0.043
PigCVP	52	2000	0.596	0.582	0.596	0.000	0.015	-0.000	no-gap	-0.324	0.346	0.332	-0.510	-0.082
Plane	7	144	1.000	1.000	1.000	0.000	0.000	0.000	no-gap	0.000	0.000	0.000	0.038	-0.010
PowerCons	2	144	0.900	0.838	0.944	0.044	0.107	-0.054	position	-0.052	0.083	0.059	0.067	-0.061
ProximalPhalanxOutlineAgeGroup	3	80	0.839	0.835	0.839	0.000	0.004	0.004	no-gap	0.026	-0.005	-0.007	0.049	-0.002
ProximalPhalanxOutlineCorrect	2	80	0.863	0.873	0.887	0.024	0.014	0.008	no-gap	0.036	0.038	-0.008	-0.041	-0.040
ProximalPhalanxTW	6	80	0.766	0.786	0.810	0.044	0.024	-0.013	position	-0.002	0.000	-0.047	0.024	-0.018
RefrigerationDevices	3	720	0.517	0.518	0.517	0.000	-0.001	0.079	no-gap	-0.023	-0.035	-0.012	-0.008	-0.044
Rock	4	2844	0.580	0.580	0.700	0.120	0.120	0.000	position	0.107	0.200	0.140	0.040	-0.113
ScreenType	3	720	0.525	0.509	0.525	0.000	0.017	-0.030	no-gap	-0.036	-0.061	-0.030	-0.027	-0.084
SemgHandGenderCh2	2	1500	0.747	0.740	0.833	0.087	0.093	-0.012	position	-0.165	0.173	-0.051	-0.035	+0.017
SemgHandMovementCh2	6	1500	0.413	0.437	0.467	0.053	0.029	0.004	position	-0.143	0.211	-0.147	0.122	+0.034
SemgHandSubjectCh2	5	1500	0.640	0.690	0.802	0.162	0.112	0.052	position	-0.133	0.236	-0.153	0.244	-0.067
ShapeletSim	2	500	0.717	0.656	0.667	-0.050	0.011	-0.084	no-gap	0.115	0.228	-0.102	-0.011	-0.154
ShapesAll	60	512	0.698	0.725	0.743	0.045	0.018	0.012	position	0.051	0.225	0.162	0.048	-0.137
SmallKitchenAppliances	3	720	0.659	0.737	0.715	0.056	-0.022	0.094	rel-phase	-0.052	0.173	-0.136	0.013	-0.049
SmoothSubspace	3	15	0.427	0.764	0.960	0.533	0.196	0.340	rel-phase	0.498	0.547	0.547	0.473	-0.013
SonyAIBORobotSurface1	2	70	0.940	0.927	0.940	0.000	0.013	-0.012	no-gap	0.011	-0.052	-0.109	0.065	+0.003

Dataset	#cls	L	A_{amp}	A_{shift}	A_{full}	Δ_{ϕ}	Δ_{al}	Δ_{rel}	type	CNN	Rocket	Incep.	DTW	Cx
SonyAIBORobotSurface2	2	65	0.908	0.887	0.887	-0.021	0.000	-0.020	no-gap	0.043	-0.005	0.024	-0.005	-0.019
StarLightCurves	3	1024	0.957	0.961	0.957	0.000	-0.004	-0.000	no-gap	-0.029	0.021	-0.097	-0.046	+0.029
Strawberry	2	235	0.946	0.948	0.973	0.027	0.025	-0.000	no-gap	-0.049	0.035	-0.034	0.019	-0.023
SwedishLeaf	15	128	0.904	0.917	0.904	0.000	-0.013	0.000	no-gap	0.026	0.058	-0.019	-0.038	-0.022
Symbols	6	398	0.844	0.794	0.857	0.013	0.063	-0.053	no-gap	0.039	0.139	0.139	0.108	-0.072
SyntheticControl	6	60	0.610	0.830	0.987	0.377	0.157	0.225	rel-phase	0.370	0.377	0.387	0.407	+0.004
ToeSegmentation1	2	277	0.697	0.698	0.697	0.000	-0.001	-0.000	no-gap	0.149	0.263	0.253	0.026	+0.019
ToeSegmentation2	2	343	0.746	0.654	0.654	-0.092	0.000	-0.086	no-gap	-0.023	0.177	0.187	0.100	+0.003
Trace	4	275	0.930	0.990	0.970	0.040	-0.020	0.050	rel-phase	0.070	0.070	0.070	0.160	-0.217
TwoLeadECG	2	82	0.958	0.854	0.872	-0.086	0.018	-0.110	no-gap	0.041	0.040	0.039	0.023	-0.008
TwoPatterns	4	128	0.453	0.878	0.889	0.436	0.010	0.413	rel-phase	0.417	0.544	0.547	0.493	+0.035
UMD	3	150	0.618	0.779	0.896	0.278	0.117	0.161	rel-phase	0.345	0.375	0.350	0.306	-0.014
UWaveGestureLibraryAll	8	945	0.633	0.762	0.952	0.319	0.190	0.117	position	0.074	0.338	0.241	0.230	+0.129
UWaveGestureLibraryX	8	315	0.444	0.688	0.755	0.311	0.066	0.251	rel-phase	0.225	0.406	0.291	0.382	+0.072
UWaveGestureLibraryY	8	315	0.453	0.637	0.687	0.235	0.051	0.177	rel-phase	0.144	0.327	0.311	0.383	+0.054
UWaveGestureLibraryZ	8	315	0.457	0.643	0.707	0.250	0.064	0.175	rel-phase	0.163	0.338	0.294	0.250	+0.052
Wafer	2	152	0.996	0.996	0.996	0.000	-0.001	0.000	no-gap	-0.001	0.003	0.001	-0.012	-0.006
Wine	2	234	0.537	0.889	0.796	0.259	-0.093	0.322	rel-phase	-0.006	0.056	-0.204	0.018	-0.049
WordSynonyms	25	270	0.378	0.422	0.561	0.183	0.139	0.049	position	0.060	0.382	0.242	0.323	-0.021
Worms	5	900	0.714	0.678	0.714	0.000	0.036	-0.029	no-gap	-0.039	0.026	-0.074	0.026	-0.074
WormsTwoClass	2	900	0.753	0.784	0.753	0.000	-0.031	-0.000	no-gap	-0.061	0.000	-0.139	-0.039	-0.043
Yoga	2	426	0.690	0.697	0.848	0.158	0.151	0.002	position	-0.020	0.220	0.182	0.055	+0.050

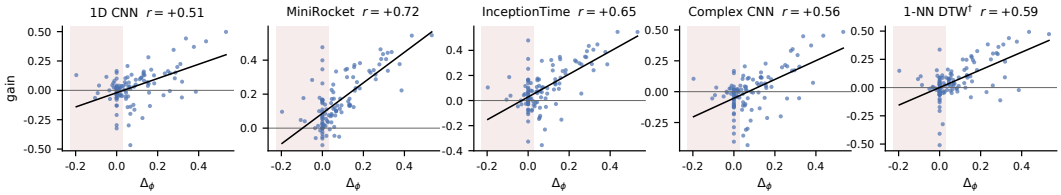

Figure 3: Gain versus phase-accessibility gap per model family, over all 114 datasets (no gating; shaded band = no-detected-gap region). Five model families with access to the raw signal—three real-valued networks, a complex-valued network, and training-free 1-NN DTW† (gain against its own amplitude twin, as in Table 4)—reproduce the same relationship. The capacity-matched complex-minus-real difference is reported in Table 4 and §5.

The dataset sets, and how they nest. All are subsets of the 114 UCR datasets.

- **Low-gap (52)** — point estimate $\Delta_\phi < 0.03$. Descriptive; the benchmark map.
- **Interval-positive (33)** — grouped train-selection interval lies entirely above 0.03; the confident-signal set, disjoint from low-gap.
- **rel-phase (23)** — typed relative-phase under the main bispectrum-enriched repertoire (Δ_{rel} dominant). Overlaps interval-positive but is defined by the point decomposition, so not nested in it.
- **Φ_2-only floor (12)** — rel-phase under the conservative sign-blind repertoire; 10 of them in the 23.
- **Stable core (7)** — rel-phase under the Φ_2-only floor, across shift seeds, and (five of seven) under a disjoint family. core-7 $\subset$ floor-12, but neither is inside rel-23: six of the core and ten of the floor are, UWaveGestureLibraryAll having left it under the budget-matched reference.

Figure 4: Definitions of the dataset subsets used throughout, and their containments.

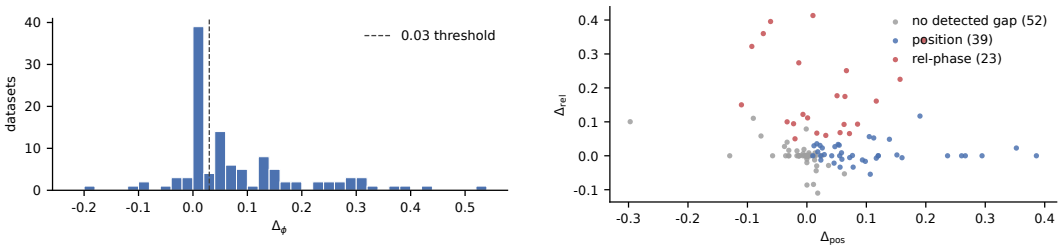

Figure 5: **Left:** phase-gap distribution over 114 UCR datasets: a large mass at zero (52/114 below 0.03, dashed) and a heavy tail to 0.53—for nearly half the benchmark the estimator detects no gap above 0.03, while the tail makes a blanket denial of phase equally wrong. **Right:** the decomposition: most aggregate gap is position (Δ_{align}); a distinct minority carries substantial relative-phase structure (Δ_{rel}).