SUSTAINED IS NOT STABLE: A REFERENCE-FREE TEST FOR ATTRACTORS IN FREE-RUNNING SEQUENCE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon generative sequence models are trained and scored under teacher forcing, which resets the state to ground truth at every step, and are deployed in closed loop, where nothing does. We show the objective is structurally silent on the closed loop: for a linear recurrence the one-step loss is invariant, at a perfect fit, under readout perturbations that move the closed-loop spectral radius, and at any residual its sublevel set contains both decaying and diverging loops. Across 64 identically trained single-layer state-space models the held-out loss is unrelated to that radius (Spearman $+0.06$, CI $[-0.21, +0.34]$), the radius alone classifies all 64 free-running outcomes, and free-running amplitude spans ten orders of magnitude. We give a reference-free test: scale the burned-in state by α, run the model on its own output, and fit settled amplitude against α on log–log axes. Slope 0 is an attractor. Slope k is a theorem: every homogeneous closed loop of degree k returns it at any width, depth or training length. Calibrated on twenty known systems ($25/26$ verdicts), the test agrees with Lyapunov exponents and Floquet multipliers on every informative trained loop inside its basin, needs no gradients or period, and reports finite basins that local methods cannot see. Across eight architectures it locates the attractor: not in depth but in whether a non-homogeneous operation lies on the closed loop—a biased readout fed back inside supplies it, outside does not, and a bias-free ReLU stack returns exactly 1.000. Mamba collapses despite having the lowest loss, and a deployed foundation model (TimesFM) lacks an attractor by design. On MIT-BIH a two-layer stack collapses on 26 of 32 seeds while beating persistence. The test certifies existence, not correctness; and existence cannot be assumed.

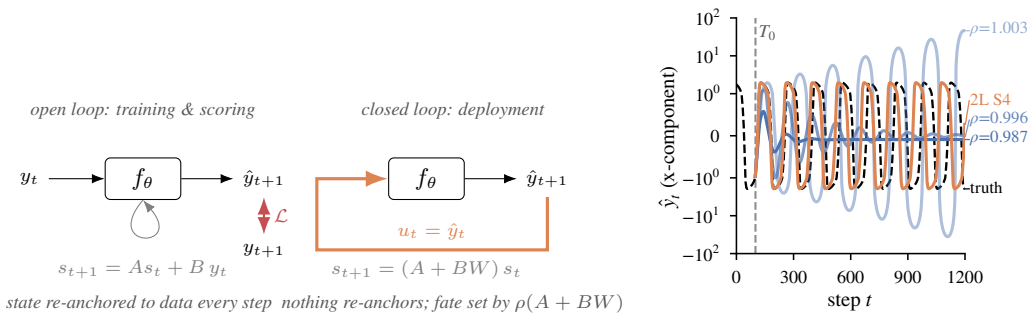
1 INTRODUCTION

A sequence model built for long-horizon generation is trained on one-step prediction with the true history as input—teacher forcing (Williams & Zipser, 1989)—and used by feeding its own output back as input. The first is an open loop, the second a closed loop (Figure 1a), and the objective never evaluates the closed loop. This is well known as the source of *drift*, and drift has remedies (Bengio et al., 2015; Lamb et al., 2016; Hess et al., 2023). This paper is about a failure no drift remedy can touch: whether the closed loop has an *attractor* at all.

The distinction is qualitative. Three seeds of one single-layer state-space model, trained identically to held-out losses of 2.80, 2.95 and 3.78×10^{-3}, run freely to $0.44\times$, $0.044\times$ and $4.7 \times 10^{4}\times$ the true amplitude (Table 1, first row). One undershoots, one dies, one explodes, and the amplitude does not order by loss (the best-fitting model in our study, a Mamba host, collapses completely). Nothing in the metric *can* see this: at a perfect one-step fit the loss of an affine closed loop is invariant under readout perturbations that move the closed-loop spectral radius, and at any residual its sublevel set contains both fates (Proposition 4). The quantity that decides the rollout lives where the objective does not look.

Blind to ρ, weakly related to the outcome. For a linear host with linear readout the autonomous map is affine and the fate is decided by the spectral radius ρ of the closed-loop Jacobian $A +$

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.



(a) (b)

Figure 1: **Sustained is not stable.** (a) Teacher forcing scores the open loop, whose state is re-anchored to data at every step; deployment runs the closed loop, whose fate for an affine host is set by $\rho(A + BW)$ (Proposition 2). (b) Four models trained to the same loss regime, run on their own output: three affine loops are marginal, decaying and diverging according to ρ; the two-layer stack sits on the target orbit.

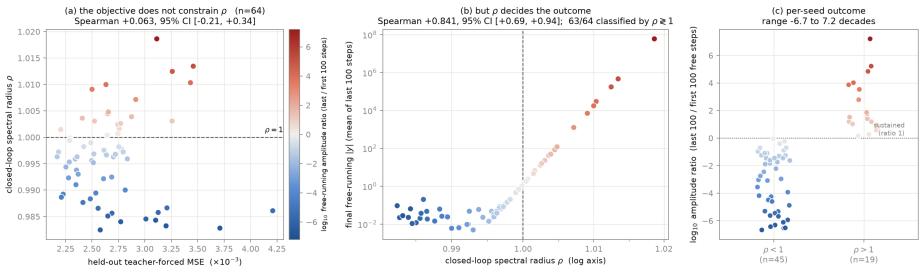


Figure 2: The objective does not constrain the quantity that decides the rollout: 64 single-layer models identical but for seed, coloured by the log amplitude ratio over the last and first 100 free steps. **(a)** loss against the exact closed-loop spectral radius: no correlation, diverging models (red) at low loss; **(b)** ρ against the outcome: a threshold at $\rho = 1$ classifies every model; **(c)** outcome by side of $\rho = 1$.

BW (Proposition 2). We trained 64 single-layer models differing only in seed and computed ρ analytically (Figure 2). No correlation is detected between held-out loss and ρ—Spearman $+0.063$, bootstrap 95% CI $[-0.215, +0.339]$—while ρ and the outcome are strongly related ($+0.841$, CI $[+0.695, +0.935]$) and the side of $\rho = 1$ classifies all 64: 45 decay, 19 diverge, none sustains a correct orbit. Loss spans a factor 1.94; amplitude spans 10.1 decades. Loss and outcome are weakly related ($+0.258$, CI $[+0.004, +0.483]$)—we do not claim independence—but the relation runs through ρ, and one of the three lowest-loss models diverges. The experiment was pre-registered with an explicit falsifier (a loss–ρ CI excluding zero) and hashed before the run (Appendix H).

Why not just compute ρ. For an affine loop one should. It stops being available when the loop is not affine: on two-layer stacks that sit on the target orbit to within 4% (§7), the linearisation at the origin gives $\rho \in [934, 6298]$ on every one of 24 models and carries no information about the outcome (Spearman -0.310, CI $[-0.654, +0.102]$); along the orbit ρ is 0.987–1.001. A single linearisation does not describe a global object.

A reference-free test with a known null. Rollout error needs a reference, which a generative model is meant to produce rather than consume, and conflates a stable wrong orbit (fidelity) with no orbit at all (stability). The defining property of an attractor needs no reference: nearby initial conditions converge to it. We scale the burned-in state by $\alpha \in [\frac{1}{4}, 4]$ and fit settled amplitude against α (Figure 3): an attractor gives slope 0. Slope k has a known null: if the closed loop commutes with positive scalings (degree-k homogeneous) it returns $\widehat{s} = k$ at every step (Proposition 1). A measured 1 is not undertraining; it is membership of a class that cannot have an attractor.

Contributions.

1. **The blind spot is structural.** For an affine closed loop the fate is fixed by ρ, which the one-step objective pins near one but cannot push to a side (Proposition 4); over 64 seeds loss is unrelated to ρ while ρ classifies every rollout.
2. **A test whose null is a theorem.** Positive homogeneity forces slope k at any finite horizon ($k = 1$ for LTI/ReLU, $k = 0$ for isolated normalisers). Calibrated on twenty known systems ($25/26$), the test agrees with Lyapunov exponents and Floquet multipliers on $17/20$ informative trained loops on the pre-registered grid and $20/20$ inside the basin; on loops that normalise their input without bypass (degree 0), a zero slope is forced by construction and must be confirmed by Floquet (§6).
3. **Where the attractor comes from.** A non-homogeneous operation *inside* the closed loop is necessary, not sufficient: a biased readout fed back at depth one supplies it, the same readout outside the loop does not, a two-layer bias-free ReLU stack returns 1.000 exactly, Mamba collapses despite perfect training loss, and a deployed foundation model (TimesFM) sustains an orbit while returning 1.0000 (§7).
4. **Existence is not correctness.** On MIT-BIH a two-layer stack collapses on $26/32$ seeds—including the seven best by loss—while beating persistence; the orbits that do exist beat at 52 and 177 against 73 bpm (§8).

2 Related work

Exposure bias. Scheduled sampling, sequence-level training and professor forcing (Bengio et al., 2015; Ranzato et al., 2016; Lamb et al., 2016), pushforward training and rollout benchmarks for emulators (Brandstetter et al., 2022; Lippe et al., 2023; Koehler et al., 2024), and generalised teacher forcing (Hess et al., 2023) all score the free run against a reference and call departure drift; Proposition 1 shows no such remedy can create an attractor for a homogeneous loop.

Certified stability and forgetting. Stable, antisymmetric, Lipschitz and equilibrium recurrences (Miller & Hardt, 2019; Chang et al., 2019; Erichson et al., 2021; Revay et al., 2024), learned Lyapunov functions (Manek & Kolter, 2019) and oscillator cells (Rusch & Mishra, 2021; Rusch & Rus, 2025; Boyer et al., 2025) certify boundedness or contraction of a block, not an isolated attracting orbit of a trained stack. The echo state property (Jaeger, 2001) is the driven-reservoir form of the question we ask autonomously—whether the state forgets its initial condition—and Jiang & Lai (2019) show a reservoir's spectral radius, not its fit, governs attractor reproduction.

Diagnostics on learned dynamics. Lyapunov exponents (Benettin et al., 1980; Wolf et al., 1985; Rosenstein et al., 1993; Kantz, 1994), the 0–1 test (Gottwald & Melbourne, 2004) and Floquet multipliers (Guckenheimer & Holmes, 1983) are the classical tools; on trained RNNs they have been used to characterise the learned dynamics (Vogt et al., 2022; Engelken et al., 2023) and to score forecasters against the true attractor (Lu et al., 2018; Vlachas et al., 2020; Platt et al., 2022; Gilpin, 2021; Mikhaeil et al., 2022). Piecewise-linear RNNs reconstruct attractors precisely because their biases place them outside the homogeneous class (Koppe et al., 2019). All of these need a reference orbit, a long series or a Jacobian; ours needs five rollouts and one fit, and its null is a theorem.

Architectures. S4/S4D (Gu et al., 2022a;b), Mamba (Gu & Dao, 2024), LinOSS (Rusch & Rus, 2025) and parallel RNNs (Lim et al., 2024; Danieli et al., 2026) constrain *open-loop* eigenvalues; the closed loop is unconstrained.

3 The protocol

Measurement. Let f_θ predict the next observation from a history and take a burn-in window $y_{1:T_0}$ from held-out data. The rollout is autonomous, $\hat{y}_{t+1} = f_\theta(\hat{y}_{\le t})$, with no ground truth after T_0. After burn-in, scale the closed-loop state by α and continue; for α on a log-spaced grid $\mathcal{A}$ centred on 1

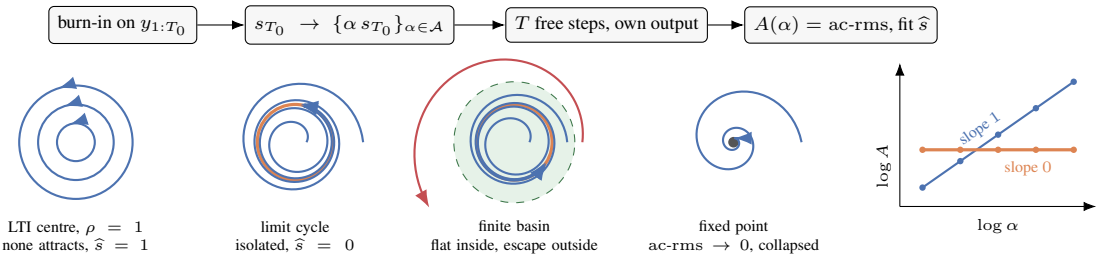


Figure 3: **The α-test and the fates it distinguishes.** Pipeline: burn in once, scale the state by α, run each copy on its own output, read the settled amplitude. Portraits, left to right: an LTI centre is a continuum of orbits, none attracting, so $A(\alpha) = \alpha A(1)$; a limit cycle is isolated, so A is flat; a finite basin is flat inside a boundary a Jacobian cannot see and escapes outside it; a fixed point scores zero oscillation. Right: the two limiting readings on the fitted axes. No reference trajectory enters.

(default $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$; Appendix E.4) record

$$A(\alpha) = \text{ac-rms}\left(\hat{y}_{T-W+1:T}\right), \qquad \text{ac-rms}(w) = \left\|w - \overline{w}\right\|_{\text{rms}}, \tag{1}$$

over a trailing window of W steps—mean-subtracted, so a settled constant scores 0—and fit the OLS slope

$$\widehat{s} = \frac{\sum_\alpha (\log\alpha - \overline{\log\alpha})(\log A(\alpha) - \overline{\log A})}{\sum_\alpha (\log\alpha - \overline{\log\alpha})^2}, \tag{2}$$

with a residual-bootstrap CI, R^2, the coefficient of variation $\text{CV}(A)$, the raw $A(\alpha)$, and the same slope $\widehat{s}_f$ of the dominant frequency (line spectra only). Rollouts run from n_{seq} burn-in windows and *every fate is decided per sequence*: a divergent sequence is counted, not averaged, since after clamping its ac-rms ≈ 0 is invisible to a mean slope. One case is degenerate by construction: if the first operation on the loop is a normalisation without bypass, it absorbs the scaling within one step and $A(\alpha)$ is flat whatever the dynamics (degree 0). On such scale-blind loops a zero slope is an identity of the wiring, not evidence, and must be confirmed by Floquet (§6).

Reading rule (in order). *Scale-blind*: de-scaled trajectories from all α coincide to numerical precision ($\leq 5 \times 10^{-4}$ over the horizon); the loop is degree-0, the zero is not read, and attractors must be certified by Floquet. *Divergent*: non-finite output in any sequence, or $\max_\alpha A > 10^6$ without exact homogeneity. *Collapsed*: all $A(\alpha) < 10^{-6}$. *Homogeneous*: $|\widehat{s} - 1| < 0.01$, $R^2 > 0.9999$, $|\widehat{s}_f| < 0.02$, tested before the amplitude cut-off so a growing homogeneous loop is still homogeneous. *Attractor*: $\text{CV}(A) < 0.10$—a flat line has no meaningful R^2, so constancy is decided first. *Non-power-law*: $R^2 < 0.90$, or fates that change with α (a basin boundary inside the grid). Otherwise *intermediate*, a nonlinear centre with amplitude-dependent period ($\widehat{s}_f \neq 0$). A stationarity flag is raised when the two halves of the window differ by more than 5% (Appendix B). Cost: $|\mathcal{A}|\, n_{\text{seq}}$ forward rollouts and a five-point regression.

4 Theory: The Null, The Necessity, and The Blind Spot

The protocol probes the autonomous map F, $s_{t+1} = F(s_t)$, with readout $y_t = G(s_t)$. Call H *positively homogeneous* if $H(\alpha s) = \alpha H(s)$ for all $\alpha > 0$, and the loop homogeneous if F and G are. Proofs not given here are in Appendix A.

Proposition 1 (Degree-k homogeneity forces $\widehat{s} = k$)**.** *If $F(\alpha s) = \alpha^k F(s)$ and $G(\alpha s) = \alpha^k G(s)$ for all $\alpha > 0$, and Φ is a scale-invariant window statistic, then $A(\alpha) = \alpha^k A(1)$ exactly. Thus $\widehat{s} = k$ with $R^2 = 1$ at every finite horizon. $k = 1$ is the LTI/bias-free-ReLU null, and any loop wrapped in RevIN (normalise-then-denormalise). $k = 0$ is any wiring whose first fed-back operation is a normaliser with no bypass: trajectories from αs_0 coincide after one step, yielding a flat $A(\alpha)$ as an identity of the wiring, not evidence of an attractor.*

Proof. If $s_t(\alpha) = \alpha^k s_t(1)$, induction and homogeneity of F, G, Φ give the rest. □

The statement is exact at every finite horizon, which is why the verdict is insensitive to burn-in, horizon and window (Appendix C). Homogeneity is preserved by linear maps, sums, compositions and positive parts, so LTI loops, bias-free rectifier networks without normalisation at any depth, their stacks, and Runge–Kutta steps of homogeneous vector fields all return $\widehat{s} = 1$ at any width, depth, parameter value and training length (Corollary A.1). What breaks it: any bias inside the loop; $\tanh$, sigmoid, GELU (Hendrycks & Gimpel, 2016), SiLU (Remark A.1).

Corollary 1.1 (The slope does not see the fate)**.** *Under Proposition 1, $\widehat{s} = 1$ whether the rollout decays, diverges or sustains bounded oscillation; $A(1)$ separates the three. Scaling is a neutral direction—a Floquet multiplier of exactly 1 (Corollary A.2)—so no periodic orbit of a homogeneous loop is exponentially attracting, and a 0.5% seed-to-seed difference in growth rate is a factor e^{10} in amplitude over 2000 steps.*

Proposition 2 (Fate of an affine closed loop)**.** *Let $F(z) = Mz + b$ with $1 \notin \operatorname{spec} M$. If $\rho(M) < 1$ every rollout decays to $(I - M)^{-1}b$; if $\rho(M) > 1$ generic rollouts diverge at rate ρ^t; if $\rho(M) = 1$ with semisimple unimodular eigenvalues, rollouts are bounded and do not decay. The fate is a function of $\rho(M)$ alone, discontinuous at $\rho = 1$; the early-window slope of a decaying affine loop is ≈ 1 (the* homogeneous transient*, §8). The trained single-layer S4 loops of this paper are affine through their projection biases (Appendix E.5).*

Proposition 3 (A slope-zero attractor requires a non-homogeneous loop)**.** *If (F, G) has an isolated attractor of positive amplitude—a compact invariant set with an open basin on which the settled* ac-rms *converges to one value $a_\Gamma \in (0, \infty)$—then (F, G) is not positively homogeneous.*

The converse fails: non-homogeneity only removes the obstruction. Affine loops decay or diverge by ρ, contracting cells collapse to a fixed point, the Neural ODE and single-layer Mamba hosts are nonlinear and fail, and a locally attracting orbit may have a basin that excludes the scaled states (Remark A.2, §6). A slope of one is *explained* by homogeneity; a slope of zero *certifies* non-homogeneity and settling, and is confirmed by a Floquet multiplier below one.

Proposition 4 (Teacher forcing pins ρ near one but not to a side)**.** *Let $L(G, c) = \mathbb{E}\|Gz + c - y^+\|^2$ be the teacher-forced loss of the open-loop map $f(z) = Gz + c$ on windows z of the target trajectory, and M the block-companion closed-loop matrix of f. (a) L depends on (G, c) only through the values of f on the support of the training windows: maps that agree there have identical loss whatever their closed-loop spectra. (b) If the target is a sustained periodic orbit and $L = 0$, then $\rho(M) \geq 1$. (c) If moreover $\rho(M) = 1$, every sublevel set $\{L \leq \varepsilon\}$ contains maps with $\rho < 1$ and maps with $\rho > 1$, realised inside the S4 parameterisation itself (Appendix A.4).*

So the loss pulls ρ toward the unit circle but, the fate being discontinuous at $\rho = 1$ while the loss is smooth there, cannot choose the side. Both predictions hold on the 64 seeds (Appendix A.5): training compresses the spread of ρ ($5.3\times$, to $[0.982, 1.019]$), and the side is predicted neither by the loss nor by the initial ρ (Spearman $+0.20$, CI $[-0.05, +0.43]$).

5 Validation

The case that would confound the test. The natural objection is that the linear system simply decayed. The discriminating case is a closed loop at spectral radius exactly 1, where an affine loop oscillates indefinitely and every amplitude-retention metric reports success. We construct it on the single-layer S4 loop by scaling the readout by bisection until $\rho = 1.000000$ (Appendix E.5, Table 6). The model oscillates, stays bounded and retains true amplitude over 2000 steps, yet the protocol reports slope 1.00 on all five seeds (Figure 4): no orbit attracts the others.

Calibration on known systems. With no training involved, the reading rule was run on twenty systems with known dynamics in 26 configurations spanning every class of Figure 3 (Table 4): limit cycles and strange attractors read *attractor* with $\widehat{s}$ within 0.002 of 0; LTI centres, damped and growing linear loops and bias-free ReLU nets read *homogeneous* with $\widehat{s} = 1.000$, $R^2 = 1$; fixed points read *collapsed*; the undamped pendulum and Duffing oscillator read *intermediate* with $\widehat{s}_f = -0.20$ and $+0.28$. $25/26$ agree with ground truth; the miss is the Hopf normal form just below bifurcation, a decaying spiral at $R^2 = 0.8999$ against the 0.90 threshold with the stationarity flag raised—the resolution limit, threshold untouched. A sweep across the bifurcation resolves it to $\mu T \approx 10$ and the frequency slope separates nonlinear centres from the LTI centre (Figure 9): the protocol detects attractors, not linearity.

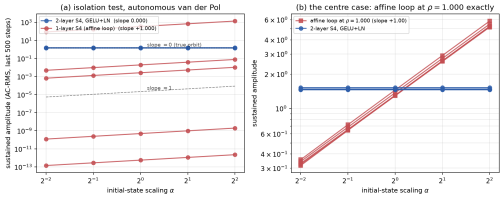


Figure 4: **(a)** settled amplitude against α, five seeds per arm: the affine single-layer S4 loop has unit slope whether it decays or diverges; the normalise-then-replace stack is flat by Proposition 1 (degree 0: shown for contrast, not as evidence). **(b)** the same affine loop projected to $\rho = 1$ sustains oscillation and the slope is still 1.

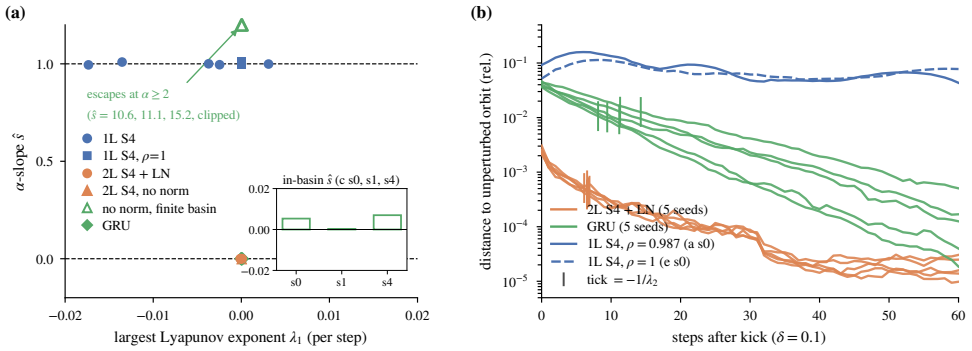


Figure 5: **Agreement with Lyapunov exponents and Floquet multipliers.** (a) $\widehat{s}$ against λ_1, one marker per seed: every loop with $\lambda_1 \approx 0, \lambda_2 < 0$ has $\widehat{s} \approx 0$, every loop with two zero exponents has $\widehat{s} \approx 1$; the three finite-basin seeds of (c) escape at $\alpha \geq 2$ (inset: in-basin $\widehat{s}$). (b) Median distance to the unperturbed orbit after a 10% kick: attractors recover at the rate $-1/\lambda_2$ predicts (ticks), affine loops do not return; the step near $t = 30$ is the orbit's relaxation jump.

6 Against Lyapunov, Floquet, and the measures that fail

On five closed loops—single-recurrence S4 (a), the two-layer S4 stack with (b) and without (c) LayerNorm, GRU (d), and (a) projected to $\rho = 1$ (e), five seeds each—we computed on the identical windowed free-running map in float64 the leading Lyapunov exponents, the Floquet multipliers of the settled orbit interpolated to its fractional period, the recovery time after a 10% kick, the local and on-orbit spectral radius, amplitude retention and KS (Figure 5, Table 7; Appendix E.6).

Agreement, and the two misjudgements. The α-verdict agrees with the Lyapunov/Floquet sign structure on 17 of 20 informative rows on the pre-registered grid and $20/20$ inside the basin. The five rows of (b) are degree-0 and certified by Floquet ($|\mu_2| \leq 8 \times 10^{-17}$, recovery 2.7 steps) but excluded from the α-count. For (a), $\lambda_1 = \log \rho$ to 2×10^{-5} and recovery matches $-1/\lambda_2$ within 6% where the loop decays. For (e), two zero exponents and both multipliers at 1.0000—a centre—while amplitude retention reads 0.99; and ρ at the origin reads 10^3 for (b) against 0.99 on the orbit. Each classical shortcut misjudges one row; the test misjudges none inside the basin.

What only the rollout test sees. Model (c) returns $\widehat{s} \approx 0$ on two seeds and escapes on three, while Lyapunov and Floquet report a locally attracting orbit on all five ($\lambda_2 \in [-0.039, -0.006]$, a contraction 4–25× weaker than with LayerNorm). Seeds 1 and 4 are flat on $\alpha \leq 1$ and leave the orbit at $\alpha \geq 2$; seed 0 loses sequences on both sides and the orbit it holds is the wrong one, 0.37× the true amplitude with period 159 against 143 (Appendix E.6). Linearisation certifies a local attractor; the rollout test reports its basin and, through the amplitude, its correctness. The discrepancy is a third piece of information—*attractor with finite basin*—and we keep both numbers. The test probes one ray of state space; whether other directions carry more information is left open (§9).

Table 1: Eight architectures on van der Pol, three seeds, host-scope protocol ($\alpha \in \{\frac{1}{4}, 1, 4\}$). Single affine recurrences are homogeneous with erratic amplitude; nonlinear cells that settle have an attractor; Mamba and the Neural ODE show non-homogeneity is necessary, not sufficient. Stacked rows use the normalise-then-replace block of Appendix E.2, whose first operation is a LayerNorm without bypass, making them degree 0: the two-layer S4 row is certified by Floquet (§6) and by the no-norm ablation; deeper rows are reported as measured and not independently certified. Cells: GRU (Cho et al., 2014), LSTM (Hochreiter & Schmidhuber, 1997), Transformer (Vaswani et al., 2017), Neural ODE (Chen et al., 2018), ParaGRU/ParaLSTM (Danieli et al., 2026).

architecture	configuration	$\widehat{s}$ (seeds)	class	amplitude / truth (seeds)
S4	single, linear readout	$0.995, 1.000, 0.999$	homogeneous	$0.44,\ 0.044,\ 4.7\times10^{4}$
LinOSS	single, linear readout	$0.996, 1.000, 1.000$	homogeneous	$0.42,\ 0.36,\ 0.93$ (seed 2 sustained)
S4	$2/3/4/6/8$ layers, norm-then-replace	0.000 (all)	scale-blind; 2L certified	0.97–1.02 at every depth
LinOSS	$2/3/4/6/8$ layers, norm-then-replace	0.000 (all)	scale-blind	0.97–1.01 at every depth
LinOSS	4 layers, no norm	$3.69,\ 0.25,\ 3.06$	basin / decay	—, $0.54,\ 0.02$
GRU	single	$-0.004, -0.002, +0.008$	attractor	≈ 1.00
GRU	4 layers, no norm	$7.25, -0.006, -0.007$	basin / attractor	$1.00, 0.99, 1.00$; seed 0 escapes at $\alpha{=}4$
Transformer	single	$-0.002, +0.038, +0.251$	attractor (2), interm. (1)	≈ 1.00; seed 2 at 0.83 for $\alpha{=}\frac{1}{4}$
ParaGRU	single	$+0.008, +0.002, +0.004$	attractor	0.98
ParaLSTM	single	$-0.003, +0.004, -0.004$	attractor	0.98–0.99
Neural ODE	single	—	divergent	non-finite
Mamba	single, 2000 steps	—	divergent (1), collapsed (5)	—

Cost. At $W = 32$, the only window at which both were run, Lyapunov costs 5.4–$7.9\times$ and Floquet 4.0–$5.7\times$ the α-slope (Table 8); the ratio grows with W since the Jacobian of a W-output state is $(WC)^2$ and its QR cubic. No gradients makes the test usable on a black-box checkpoint, and no state Jacobian makes it tractable for a Transformer whose state is a key–value cache; on a scale-blind loop one Floquet computation is still required to confirm an orbit.

7 Where the attractor comes from

Table 1 applies the protocol to eight architectures on the autonomous van der Pol oscillator (van der Pol, 1926), as a single recurrence (what analyses treat) and as a stack of normalise-then-replace blocks (what is deployed). The single affine recurrences behave as Proposition 2 requires, realising undershoot, collapse and divergence and never a correct orbit; every nonlinear cell that settles has an attractor; the Neural ODE diverges; a single-layer Mamba trained $5\times$ longer reaches the lowest teacher-forced loss in the study (1.4×10^{-5}, $30\times$ below the successful S4 stack) and still collapses or diverges on all $16/16$ sequences across six variants. Non-homogeneity is necessary, not sufficient, and the best-fitting model cannot free-run. The stacked rows are flat because their first operation is a LayerNorm; *what* supplies the attractor is answered by the loss-matched ablation below.

Depth is incidental; the loop is what matters. Figure 6 shows loss-matched S4 variants, five seeds each, none normalised unless stated (Appendix F). Feeding a biased MLP readout back *inside* the loop—no second recurrence—gives $\widehat{s} = 0.000 \pm 0.003$ at $W = 32$ and $|\widehat{s}| \leq 0.010$ on every seed that holds an orbit (two on target, one at $1.4\times$, one near a fixed point, one divergent: the fates of the deep rows, at depth one). The identical head *outside* the loop gives 0.998 ± 0.004, homogeneous again. The converse passes: a two-layer stack of bias-free ReLU blocks without normalisation—nonlinear but homogeneous by Corollary A.1—returns $\widehat{s} = +1.000$ on $5/5$ seeds with $R^2 = 1.000$ (biases and LayerNorms verified absent). Restoring any one non-homogeneous ingredient leaves the class: bias or GELU alone. "Depth" in the standard block is simply where a non-homogeneous operation first lands on the loop.

What normalisation does and does not do. Without LayerNorm the stack still has a local attractor on $4/5$ seeds, so normalisation does not create it, but its contraction is 4–$25\times$ weaker (§6). Whether normalisation widens the basin cannot be read from the normalise-then-replace rows, whose α range is never explored (Proposition 1). Proposition 1 adds one thing: drift remedies can move a homogeneous model's amplitude, never its class. Whether a deployed stack has an attractor cannot be assumed, which is why it has to be measured.

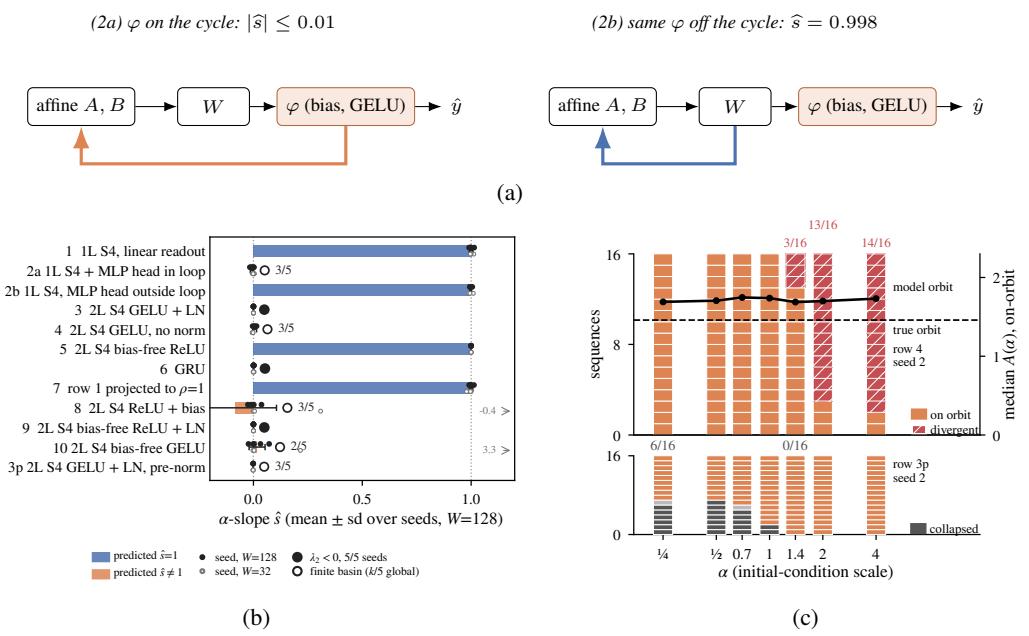


Figure 6: **The attractor is decided inside the loop.** (a) The same recurrence and the same biased MLP head; only the feedback edge differs. (b) Loss-matched S4 ablations, mean ± sd over five seeds: homogeneous rows return 1, non-homogeneous rows 0; open markers mark rows with a finite basin on some seeds; the only rows containing LayerNorm are 3, 3p and 9. (c) Per-sequence fates against α for the finite-basin seed of the no-norm stack (escape at $\alpha \geq 1.4$) and, below, the pre-norm seed that collapses at small α.

A released foundation model. The protocol needs no gradients, so we ran it on TimesFM 2.5 (200M parameters, public checkpoint). Its released code documents $\mathrm{TimesFM}(aX + b) = a\,\mathrm{TimesFM}(X) + b$ for $a \geq 0$, implemented by RevIN: the closed loop is degree-1 by construction (Proposition 1). On van der Pol ($T_0 = 1024$, 2048 free steps) it sustains a period-locked orbit at $1.45\times$ the true amplitude, and returns $A(\frac{1}{4}, \frac{1}{2}, 1, 2, 4) = 0.51, 1.02, 2.04, 4.08, 8.16$: exactly $\hat{s} = 1.0000$, $R^2 = 1.000000$, frequency slope 0.0000. A deployed forecaster lacks an attractor by design.

8 Existence is not correctness: a physiological recording

We apply the protocol to the MIT-BIH Arrhythmia Database (Moody & Mark, 2001; Goldberger et al., 2000) with the inter-patient DS1/DS2 split, three architectures at ≈ 150k parameters (Table 2, Figure 7; Appendix G). All beat persistence on one-step test MSE. The two-layer stack uses the pre-norm wiring.

Reading it, and whether loss is blind here. The verdicts match the theory (Table 2); the test is what reveals the two-layer collapse, and the orbits that do exist are not the right ones. Within one configuration, loss never says which model fails. For the two-layer stack ($n = 32$) every loss–outcome CI contains zero ($|\log$ amplitude error$|$: $+0.09$ $[-0.24, +0.42]$; retention -0.25 $[-0.54, +0.09]$; $\hat{s}$ $+0.01$ $[-0.37, +0.37]$) and the 26 collapsed seeds include loss ranks 1–7. For the GRU ($n = 32$) loss is weakly to moderately related to amplitude error ($+0.47$ $[+0.14, +0.68]$) and to nothing else; the eight failing seeds sit at loss ranks $4, 21, 22, 25, 27, 29, 30, 31$. Pooled across architectures the correlation is $+0.75$, a between-architecture confound that drops to $+0.35$ $[+0.13, +0.52]$ when stratified. This is the regime of van der Pol at $n = 64$.

Table 2: MIT-BIH, DS1→DS2. The affine loop decays on every seed with the homogeneous transient of Proposition 2; the GRU holds an ECG-like orbit at the wrong amplitude and rate; the two-layer stack collapses on most seeds while its loss is three times better than persistence (0.110).

architecture	n, lr	TF test MSE	fates
1L S4, linear readout (affine)	$5, 3\times10^{-3}$	0.056	collapsed $5/5$; $\rho_{\mathrm{orb}} = 0.91$–$0.98$; early $\widehat{s} = 0.89$–0.99
same, 8k steps	$5, 3\times10^{-3}$	—	collapsed $5/5$; $\rho = 0.887$–0.913
GRU	$5, 10^{-3}$	0.022	attractor $4/5$; amp 0.74 ± 0.33; retention 0.90 ± 0.12
GRU	$32, 10^{-3}$	0.0215–0.0251	attractor 21, non-power-law 11; $8/32$ below $0.1\times$; median amp 0.57
2L S4, pre-norm	$5, 3\times10^{-3}$	0.032–0.042	collapsed $3/5$, attractor $1/5$, non-power-law $1/5$
2L S4, pre-norm	$32, 10^{-3}$	0.0293–0.0355	$26/32$ below $0.1\times$ (fixed point, α-dependent basin); median amp 0.001

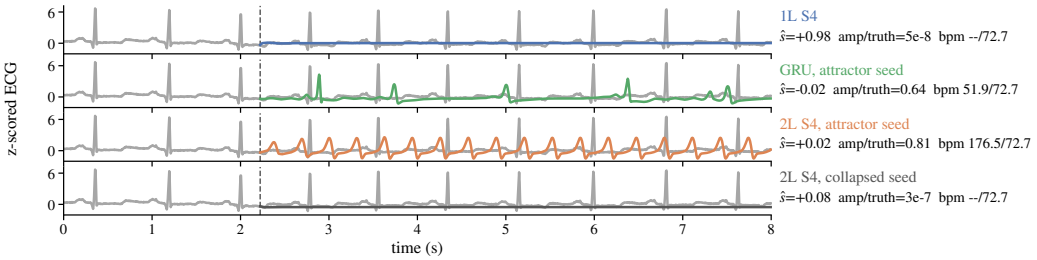


Figure 7: **A real recording.** Free-running rollouts of four MIT-BIH checkpoints on one held-out segment (grey: the recording; dashed: end of burn-in). The single-layer S4 decays with a homogeneous transient; a GRU seed holds a beat-like orbit at $0.64\times$ the amplitude and 52 against 73 bpm; the one two-layer seed that holds an orbit beats at 177 bpm; the typical two-layer seed collapses.

9 Limitations

Existence, not correctness. The test certifies an attractor exists, not that it is right (§8; model (c) seed 0); which orbit is learned needs a reference. **Normalised loops.** A normalise-then-replace wiring returns a flat $A(\alpha)$ by construction (degree 0); a zero is confirmed by Floquet. Perturbing only the x-channel (which LN cannot erase) moves the state and returns to the orbit ($\widehat{s} = -4\times10^{-3}$), pointing to directional tests. **One ray.** The test perturbs along the radial direction only, the one with a theorem for its null; whether random or Jacobian-aligned directions carry more basin information is untested. **Resolution.** Contraction slower than $\approx 10/T$ is unresolved at horizon T, and small-amplitude nonlinear centres give $\widehat{s} \approx 1$, separated only by the frequency slope, which is undefined on broadband spectra. **Scale.** Models are $\leq$ 200k parameters, trained from scratch on one oscillator and one recording; TimesFM demonstrates the test scales without gradients.

10 Conclusion

Teacher forcing evaluates a loop that is never deployed, and for an affine recurrence it is silent exactly where the deployed loop is decided. We gave a five-rollout, reference-free test whose null is a theorem, calibrated it on twenty known systems, matched it against Lyapunov exponents and Floquet multipliers, and used it to locate the attractor: not in depth, not in the cell, but in whether a non-homogeneous operation lies on the closed loop. The test says whether there is an orbit. Which orbit is the question we would most like to see measured next.

REPRODUCIBILITY STATEMENT

Code, checkpoints, calibration integrators, the Lyapunov/Floquet implementation with fractional-period interpolation, DS1/DS2 record lists, a manifest with the MD5 of every result file quoted, and the timestamped, hashed pre-registration files are in the supplementary material; Appendix E traces each table to its script and seeds.

AI USE STATEMENT

Generative AI assistance was used for code implementation, experiment orchestration, and drafting and editing of manuscript text. All claims, experimental designs and numerical results were verified by the authors.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean-Marie Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 1: Theory. *Meccanica*, 15(1):9–20, 1980. doi: 10.1007/BF02128236. Part 2: Numerical application, same volume, pp. 21–30, doi:10.1007/BF02128237.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015. arXiv:1506.03099.

Jared Boyer, T. Konstantin Rusch, and Daniela Rus. Learning to dissipate energy in oscillatory state-space models. *arXiv preprint arXiv:2505.12171*, 2025. D-LinOSS.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations (ICLR)*, 2022. Spotlight. Pushforward trick. arXiv:2202.03376.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1902.09689.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. arXiv:1806.07366.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. GRU. arXiv:1406.1078.

Federico Danieli, Pau Rodriguez, Miguel Sarabia, Xavier Suau, and Luca Zappella. ParaRNN: Unlocking parallel training of nonlinear RNNs for large language models. In *International Conference on Learning Representations (ICLR)*, 2026. Oral. arXiv:2510.21450. Code: github.com/apple/ml-pararnn.

Georg Duffing. *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. Sammlung Vieweg, Heft 41/42. Friedrich Vieweg & Sohn, Braunschweig, 1918.

Rainer Engelken, Fred Wolf, and L. F. Abbott. Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, 5:043044, 2023.

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2006.12070.

Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1(6):445–466, 1961. doi: 10.1016/S0006-3495(61)86902-6.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. dysts. arXiv:2110.05266.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Georg A. Gottwald and Ian Melbourne. A new test for chaos in deterministic systems. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460 (2042):603–611, 2004. doi: 10.1098/rspa.2003.1183.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022a. S4. arXiv:2111.00396.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022b. S4D. arXiv:2206.11893.

John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, volume 42 of *Applied Mathematical Sciences*. Springer, New York, 1983.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Florian Hess, Zahra Monfared, Manuel Brenner, and Daniel Durstewitz. Generalized teacher forcing for learning chaotic dynamics. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13017–13049, 2023. arXiv:2306.04406.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report 148, GMD, 2001.

Junjie Jiang and Ying-Cheng Lai. Model-free prediction of spatiotemporal dynamical systems with recurrent neural networks: Role of network spectral radius. *Physical Review Research*, 1(3): 033056, 2019. doi: 10.1103/PhysRevResearch.1.033056.

Holger Kantz. A robust method to estimate the maximal Lyapunov exponent of a time series. *Physics Letters A*, 185(1):77–87, 1994. doi: 10.1016/0375-9601(94)90991-1.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2411.00180.

Georgia Koppe, Hazem Toutounji, Peter Kirsch, Stefanie Lis, and Daniel Durstewitz. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLoS Computational Biology*, 15(8):e1007263, 2019.

Alex M. Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. arXiv:1610.09038.

Yi Heng Lim, Qi Zhu, Joshua Selfridge, and Muhammad Firmansyah Kasim. Parallelizing nonlinear sequential models over the sequence length. In *International Conference on Learning Representations (ICLR)*, 2024. DEER. arXiv:2309.12252.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2308.05732.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. AdamW. arXiv:1711.05101.

Zhixin Lu, Brian R. Hunt, and Edward Ott. Attractor reconstruction by machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):061104, 2018. doi: 10.1063/1.5039667.

Gaurav Manek and J. Zico Kolter. Learning stable deep dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. arXiv:2001.06116.

Jonas M. Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with RNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2110.07238.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1805.10369.

George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning (ICML)*, 2010.

Jason A. Platt, Stephen G. Penny, Timothy A. Smith, Tse-Chun Chen, and Henry D. I. Abarbanel. A systematic exploration of reservoir computing for forecasting complex spatiotemporal dynamics. *Neural Networks*, 153:530–552, 2022. doi: 10.1016/j.neunet.2022.06.025.

Ilya Prigogine and Grégoire Nicolis. On symmetry-breaking instabilities in dissipative systems. *The Journal of Chemical Physics*, 46(9):3542–3550, 1967. doi: 10.1063/1.1841255.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. arXiv:1511.06732.

Max Revay, Ruigang Wang, and Ian R. Manchester. Recurrent equilibrium networks: Flexible dynamic models with guaranteed stability and robustness. *IEEE Transactions on Automatic Control*, 69(5):2855–2870, 2024. doi: 10.1109/TAC.2023.3294101. Early access July 2023. arXiv:2104.05942.

Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1–2): 117–134, 1993. doi: 10.1016/0167-2789(93)90009-P.

Otto E. Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5):397–398, 1976. doi: 10.1016/0375-9601(76)90101-8.

T. Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.00951.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *International Conference on Learning Representations (ICLR)*, 2025. Oral. LinOSS. arXiv:2410.03943.

J. T. Stuart. On the non-linear mechanics of wave disturbances in stable and unstable parallel flows. Part 1. *Journal of Fluid Mechanics*, 9(3):353–370, 1960. doi: 10.1017/S002211206000116X.

Balthasar van der Pol. On "relaxation-oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):978–992, 1926. doi: 10.1080/14786442608564127.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. arXiv:1706.03762.

Pantelis R. Vlachas, Jaideep Pathak, Brian R. Hunt, Themistoklis P. Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126:191–217, 2020. doi: 10.1016/j.neunet.2020.02.016.

Ryan Vogt, Maximilian Puelma Touzel, Eli Shlizerman, and Guillaume Lajoie. On Lyapunov exponents for RNNs: Understanding information propagation using dynamical systems tools. *Frontiers in Applied Mathematics and Statistics*, 8:818799, 2022.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Alan Wolf, Jack B. Swift, Harry L. Swinney, and John A. Vastano. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985. doi: 10.1016/0167-2789(85)90011-9.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. arXiv:1910.07467.

A Proofs and further theory

A.1 The homogeneous class

Corollary A.1 (The homogeneous class)**.** *Positive homogeneity of degree one is preserved by linear maps, sums, compositions and positive parts. Hence each of the following returns $\widehat{s} = 1$ at any width, depth, parameter value and training length: (i) LTI loops $F(s) = Ms$, in particular $M = A + BW$; (ii) networks of bias-free linear maps and ReLU (Nair & Hinton, 2010), LeakyReLU, PReLU, absolute value, max-pooling or maxout units without biases and without normalisation; (iii) stacks and residual compositions of (i)–(ii); (iv) the flow, and any Runge–Kutta step, of a locally Lipschitz homogeneous vector field.*

Proof. $\max(\alpha u, 0) = \alpha \max(u, 0)$ for $\alpha > 0$; linear maps commute with scalars; sums and compositions of homogeneous maps are homogeneous. For (iv), if $z(t)$ solves $\dot{z} = f(z)$ with $f(\alpha z) = \alpha f(z)$ then $\alpha z(t)$ solves it from $\alpha z(0)$; uniqueness gives homogeneity of the flow, and each RK stage is a composition of sums of homogeneous maps. □

Remark A.1 (What breaks homogeneity). Any additive bias inside the loop; $\tanh$, sigmoid, GELU, SiLU, softplus; and LayerNorm/RMSNorm (Ba et al., 2016; Zhang & Sennrich, 2019), homogeneous of degree zero, which moreover absorb a scaling of their input entirely. A bias alone does not create an attractor: if $F(s) = Ms + c$ and $1 \notin \operatorname{spec} M$ the orbits come in scaled families about $p = (I - M)^{-1}c$ and $\widehat{s} \to 1$ as α grows or $c \to 0$. The single-recurrence S4 is affine in this sense; its slopes 0.995–1.000 fall short of 1.000 by the bias.

A.2 Scaling is a neutral direction

Corollary A.2. *If F is homogeneous then $F(s + \epsilon s) = (1 + \epsilon)F(s)$ exactly, so a radial perturbation propagates without growth or decay relative to the trajectory, and for a p-periodic orbit s_0 is an eigenvector of the monodromy $D(F^p)(s_0)$ with eigenvalue exactly 1; no periodic orbit of a homogeneous loop is exponentially attracting.*

Proof. Differentiate $F(\alpha s) = \alpha F(s)$ at $\alpha = 1$: $DF(s)s = F(s)$; by the chain rule $D(F^t)(s_0)s_0 = s_t$, and $s_p = s_0$. □

Writing $s_t = r_t u_t$, the angular part evolves autonomously and $\log r_T/r_0 = \sum_t \log \|F(u_t)\| \approx T\lambda(\theta)$: nothing selects an amplitude, and seed-to-seed differences in λ invisible to a bounded loss are amplified by $e^{T\delta\lambda}$.

A.3 Proposition 3 and why its converse fails

Proof of Proposition 3. For s_0 in the open basin, αs_0 is in it for α near 1; homogeneity would give $a_\Gamma = \alpha a_\Gamma$, forcing $a_\Gamma \in \{0, \infty\}$. □

Remark A.2. (i) *Affine loops*: $Ms + c$ with $c \neq 0$ is non-homogeneous yet decays to p if $\rho < 1$ and diverges if $\rho > 1$ (45 and 19 of the 64 gate models). (ii) *Contracting cells*: $\tanh(Js)$ with $\|J\| < 1$, or a GRU with small weights, contracts to a unique fixed point. (iii) *Measured*: the Neural ODE host diverges on every seed and the single-layer Mamba host diverges on one and collapses on two, at 400 training steps. (iv) *Finite basins*: a locally attracting orbit may exclude the scaled states (model (c), §6; GRU depth four without normalisation, Table 1).

A.4 The affine closed loop: Propositions 2 and 4

For the windowed hosts the state is the window $z_t = (\hat{y}_{t-W+1}, \ldots, \hat{y}_t) \in \mathbb{R}^{WC}$ and the map shifts and appends: $F(z) = (Sz;\, f(z))$ with $S = [\,0 \;\; I_{(W-1)C}\,]$. For $f(z) = Gz + c$ this is $Mz + b$ with M the block companion matrix, the matrix $A + BW$ of §1 in recurrent-state coordinates. That the trained loops are affine is checked: over all 64 gate models the Jacobian of F at random states equals its value at $z = 0$ to double precision.

Proof of Proposition 2. $z_{t+1} - p = M(z_t - p)$; the Jordan form gives the three cases. The early-window slope of a decaying loop is ≈ 1 because the readout window is $\alpha\tilde{a} - \tilde{b}$ with $\tilde{a}, \tilde{b}$ independent of α, and $\tilde{b}$ is negligible before the bias-driven component dominates. □

Proof of Proposition 4. (a) is the definition of L. (b) $L = 0$ means $z_{t+1} = Mz_t + b$ along the windows; q-periodicity gives $M^q(z_0 - p) = z_0 - p$ with $z_0 \neq p$, so $\lambda^q = 1$ for some eigenvalue. (c) Write $f(z) = \sum_{j=1}^{W} G_j \hat{y}_{t+1-j} + c$ and set $G_{r,j} = r^j G_j$; the companion determinant gives $\operatorname{spec} M_r = r \operatorname{spec} M_1$, so $\rho(M_r) = r$, while the residual $\sum_j (r^j - 1) G_j \hat{y}_{t+1-j}$ is continuous in r and vanishes at $r = 1$. For the S4 host the dilation is realised by $(W_{\text{in}}, a) \mapsto (rW_{\text{in}}, ra)$, which multiplies G_1 by r and G_j by r^j including the residual feedthrough (Appendix E.2). At a perfect fit part (a) makes L invariant under any ΔG vanishing on the data support while ρ moves; at $L > 0$ the statement is the sublevel-set one of (c), not first-order flatness. □

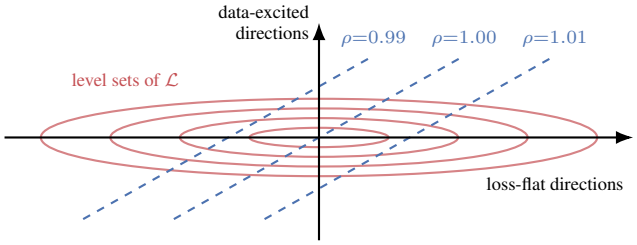


Figure 8: Proposition 4(a): a parameter change that vanishes on the data support leaves $\mathcal{L}$ unchanged and moves ρ across one.

A.5 The $n = 64$ population and initialisation

$\rho(M)$ is computed from the companion matrix at $W = 32$; Spearman CIs are percentile bootstraps over 10^4 resamples. Loss vs. ρ: $+0.063$ $[-0.215, +0.339]$ (Pearson $+0.021$, $p = 0.87$). ρ vs. log final amplitude: $+0.841$ $[+0.695, +0.935]$; radii $[0.982, 1.019]$, 45 below one, 19 above; side of one classifies $64/64$; amplitudes span 10.1 decades; the rollout's own growth factor agrees with ρ to 10^{-5}. Loss vs. outcome: $+0.258$ $[+0.004, +0.483]$, Mann–Whitney $p = 0.024$; the diverging models span loss ranks 3 to 62. Initialisation: initial radii $[0.843, 1.049]$, sd falls from 0.044 to 0.008 under training; Spearman of final on initial ρ $+0.200$ $[-0.048, +0.430]$, $r^2 = 0.016$; initial ρ is unrelated to loss (-0.009) and outcome ($+0.137$). Rebuilding the Jacobian at $W \in \{16, 64, 128\}$ changes ρ by at most 0.010; for six seeds the exact recurrent-state closed loop agrees with the $W = 256$ companion to 2×10^{-7}.

B THE ESTIMATOR AND THE READING RULE

Under Proposition 1 the points lie on a line and $R^2 = 1$ exactly. Under an attractor $A(\alpha)$ is constant up to noise, so R^2 is a ratio of two noise terms (Lorenz and Rössler return $R^2 \in [0, 0.35]$ with $\widehat{s}$ within 0.002 of zero); constancy is therefore decided first by $\mathrm{CV}(A)$. The two witnesses cannot both fire: a homogeneous loop has $\mathrm{CV}(A) = \mathrm{CV}(\alpha) = 0.88$ on the five-point grid regardless of the model. Slopes are reported to three decimals; $|\widehat{s}| < 5 \times 10^{-4}$ is written 0.000. The frequency slope is the OLS slope of the log dominant FFT frequency (Hann window, parabolic peak) on $\log \alpha$, defined for line spectra only; on the broadband Lorenz spectrum the peak-picker returns $+0.28$ and the spectral-centroid slope $+0.015$, neither meaningful. Threshold robustness: halving or doubling each of the seven thresholds changes no verdict in Table 4 except the Hopf miss, which moves to *intermediate* when the R^2 cut-off is lowered to 0.85. A period-two orbit reads intermediate on the amplitude axis and is identified by the fate table; for chaotic or fixed-point targets the statistic (1) must change and we have not validated that variant.

C SENSITIVITY

Table 3: $\widehat{s}$ across the protocol's own hyperparameters (S4 depth 1 and 2, seed 0; one knob varied about $(T_0, T, W) = (100, 1000, 200)$).

setting	single recurrence	two-layer stack
burn-in $T_0 \in \{50, 100, 200, 400\}$	$+1.002$, $+1.000$, $+1.003$, $+0.998$	0.000
rollout $T \in \{500, 1000, 2000\}$	$+1.000$	0.000
window $W \in \{100, 200, 500\}$	$+1.000$	0.000
window $W \in \{32, 64, 128\}$ (baseline protocol)	$+0.995$ to $+1.013$	± 0.000
range	$[+0.998, +1.003]$	$\lvert\widehat{s}\rvert < 7 \times 10^{-7}$

D CALIBRATION ON SYSTEMS WITH KNOWN DYNAMICS

Each system is integrated by batched RK4 from αz_0 on seven log-spaced $\alpha \in [2^{-2}, 2^2]$; $\Delta t = 0.01$ (0.005 Lorenz; 0.02 Rössler, FHN, ReLU nets); $T = 600$, $W = 80$ time units, lengthened where the period demands (FHN $4000/2000$, Rössler $2000/400$, Lorenz window 200); if the half-window amplitudes differ by more than 5%, T and W are doubled at most twice and the stationarity flag is raised if the last attempt fails. A fixed-row replication with seven fresh runs (van der Pol $\mu = 3$, LTI $\rho = 0.99$ and 1.01, a 3-layer ReLU net, Hopf $\mu \in \{-0.1, 0.05, 0.5\}$) gives $19/20$ with the same single miss.

E IMPLEMENTATION

Single RTX 4090, fp32 training, fp64 Jacobians and baselines, PyTorch 2.12; seeds fixed per run; every result file's MD5 is in the manifest.

E.1 DATA

The generator integrates $\ddot{x} = \mu(1 - x^2)\dot{x} - x$, $\mu = 2$, by semi-implicit Euler at $\Delta t = 0.05$ from $x_0, v_0 \sim \mathcal{U}[-2, 2]$, discards 2000 burn-in steps, observes $(x, \dot{x})$ noise-free; no forcing anywhere. Training 192×300, held-out 64×300, rollout contexts $16 \times (T_0{+}10)$, all regenerated from seed. "Truth" amplitude is the $\mathrm{ac\text{-}rms}$ of the true orbit, 1.46.

E.2 ARCHITECTURES AND WIRING

Every host is wrapped as $\hat{y} = W_{\mathrm{out}}\,\mathrm{host}(W_{\mathrm{in}}u)$, $D = 64$, $C = 2$, bias-free projections. At depth 1 nothing else is applied. For S4, $s_t = a \odot s_{t-1} + Bh_t$, $y_t = h_t + Cs_t$ with B, C full linear maps with bias and $a = \exp(-e^{\log a})$, so the depth-1 closed loop is affine. At depth $k > 1$ each block

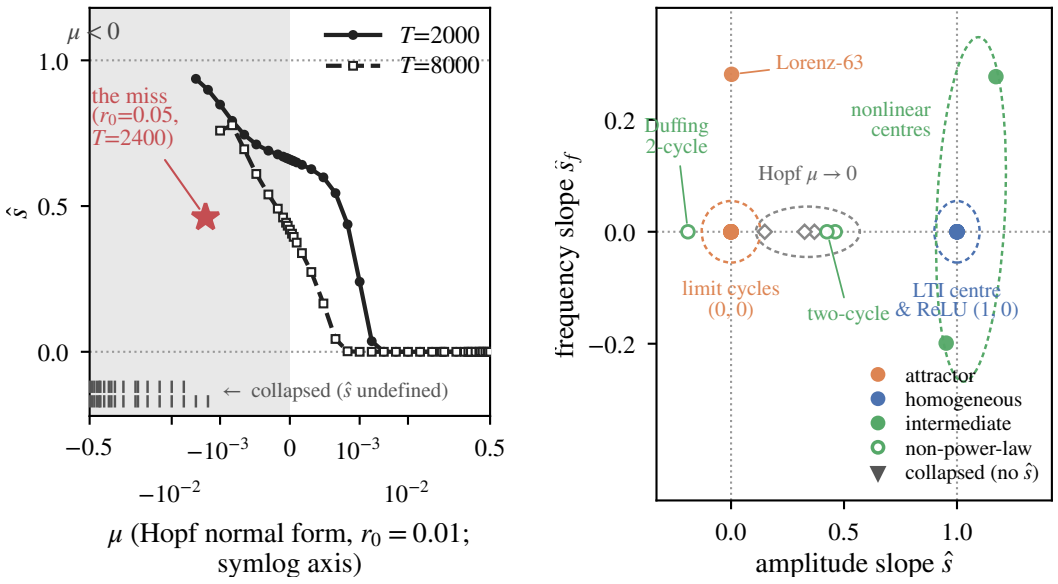


Figure 9: **Calibration.** (a) $\hat{s}$ across the Hopf bifurcation ($r_0 = 0.01$, $T = 2000$ and 8000); the star is the single miss. (b) Amplitude slope against frequency slope for the rows with a finite slope; limit cycles sit at $(0, 0)$, homogeneous loops at $(1, 0)$, nonlinear centres off the $\hat{s}_f = 0$ axis. Lorenz-63 is broadband and its peak-picked $\hat{s}_f$ is not meaningful.

of the van der Pol stacks is *normalise-then-replace*: $h \leftarrow \mathrm{host}_i(\mathrm{LN}_i^{(1)} h)$, where the host returns its normalised input plus its body and the un-normalised h is dropped, then $h \leftarrow h + \mathrm{FF}_i(\mathrm{LN}_i^{(2)} h)$, $\mathrm{FF} = \mathrm{Linear}(D, 2D) \to \mathrm{GELU} \to \mathrm{Linear}(2D, D)$. No un-normalised stream bypasses $\mathrm{LN}_1^{(1)}$, so the closed loop is degree-0 (Proposition 1); measured on model (b) seed 0 in float64, rollouts from all five α differ by $\leq 5 \times 10^{-4}$. The ECG stack and ablation rows 3p/9p use the standard *pre-norm residual* $h \leftarrow h + (\mathrm{host}(\mathrm{LN}\, h) - \mathrm{LN}\, h)$, in which the stream survives and α is a genuine perturbation. `--nonorm` replaces every LayerNorm by the identity. Parameter counts (depth 1): S4 8,640; LinOSS 12,800; Mamba 16,896; Neural ODE 25,024; GRU 25,216; Transformer 67,200; ParaGRU 12,928; ParaLSTM 17,344. S4 stacks $k = 2/4/8$: 50,688/101,120/201,984. Ablation rows: 1L S4 + MLP head 17,090 (inside) / 17,218 (outside); 2L bias-free ReLU 49,536; 2L no norm 50,176.

E.3 Training

One-step teacher-forced MSE, AdamW (Loshchilov & Hutter, 2019) 3×10^{-3} (ECG also 10^{-3}), weight decay 10^{-5}, gradient clip 1.0, batch 32; no schedule, dropout, noise, scheduled sampling or rollout loss. Table 1: 400 steps, seeds 0–2. Gate and baselines: 800 steps. Loss-matched ablation: stop at the first 20-step check with held-out MSE $\leq 3 \times 10^{-3}$ (deep rows reach 1–2 $\times 10^{-3}$ at step 60; affine rows 2.9–3.0 $\times 10^{-3}$ at 260–440); fixed-budget and 10^{-3}-target variants give the same verdicts. Two-layer S4 loss-matched to the single recurrence on three seeds: $\hat{s} = 0.000$ at 80–100 steps against $+0.996$ for a 3000-step one-layer model at the same loss.

E.4 Protocol configurations

E.5 Closed-loop spectral radius and the projected loops

For the windowed model the autonomous map on $z \in \mathbb{R}^{WC}$ drops the oldest observation and appends $f(z)$; its Jacobian is the companion matrix with bottom block $\partial f / \partial z$ from autodiff at $z = 0$; ρ is the largest eigenvalue modulus in double precision. For depth 1 the map is affine so J does not depend on z (verified at three random windows per seed: $\max |J(z) - J(0)| = 0$). The projected

Table 4: Calibration on 26 configurations (no training). Systems: van der Pol (van der Pol, 1926), Hopf and Stuart–Landau (Stuart, 1960), FitzHugh–Nagumo (FitzHugh, 1961), Brusselator (Prigogine & Nicolis, 1967), Lorenz (Lorenz, 1963), Rössler (Rössler, 1976), Duffing (Duffing, 1918). CV: coefficient of variation of $A(\alpha)$; assigned category follows the reading rule of §3; trend is the half-window amplitude ratio. $25/26$ match ground truth.

System	Regime	Expected	$\hat{s}$	CV	R^2	Trend	Assigned	
Van der Pol	$\mu = 0.5$	attractor	$+0.000$	0.0000	0.537	stationary	attractor	✓
Van der Pol	$\mu = 1$	attractor	$+0.000$	0.0000	0.507	stationary	attractor	✓
Van der Pol	$\mu = 2$	attractor	$+0.000$	0.0007	0.118	stationary	attractor	✓
Van der Pol	$\mu = 4$	attractor	-0.002	0.0045	0.104	stationary	attractor	✓
Hopf normal form	$\mu = 1$	attractor	$+0.000$	0.0000	–	stationary	attractor	✓
Stuart–Landau (shear)	$c_2 = 1.5$	attractor	$+0.000$	0.0000	–	stationary	attractor	✓
FitzHugh–Nagumo	$I = 0.5$	attractor	$+0.000$	0.0005	0.861	stationary	attractor	✓
Brusselator	$A = 1, B = 3$	attractor	-0.000	0.0009	0.014	stationary	attractor	✓
Lorenz-63	chaotic	attractor	$+0.002$	0.0060	0.138	stationary	attractor	✓
Rössler	chaotic	attractor	-0.002	0.0055	0.069	stationary	attractor	✓
Linear centre	$\rho = 1$	homogeneous	$+1.000$	0.8482	1.000	stationary	homogeneous	✓
Linear centre 4-D	two incommensurate	homogeneous	$+1.000$	0.8482	1.000	stationary	homogeneous	✓
Damped linear	$\zeta = 0.1, T = 60$	homogeneous	$+1.000$	0.8482	1.000	decaying†	homogeneous	✓
Unstable linear	$\sigma = 0.02$	homogeneous	$+1.000$	0.8482	1.000	growing†	homogeneous	✓
ReLU recurrent net	bias-free, seed 0	homogeneous	$+1.000$	0.8482	1.000	growing†	homogeneous	✓
ReLU recurrent net	bias-free, seed 1	homogeneous	$+1.000$	0.8482	1.000	growing†	homogeneous	✓
Damped linear	$\zeta = 0.1, T = 600$	collapsed	–	–	–	collapsed	collapsed	✓
Overdamped linear	$\ddot{x} + 4\dot{x} + x = 0$	collapsed	–	–	–	collapsed	collapsed	✓
Damped nonlinear	fixed pt at $(1, 0)$	collapsed	–	–	–	collapsed	collapsed	✓
Duffing (damped)	double well	collapsed	–	–	–	collapsed	collapsed	✓
Duffing (damped)	tilted double well	collapsed	–	–	–	collapsed	collapsed	✓
Pendulum	conservative, softening	intermediate	$+0.952$	0.7890	0.999	stationary	intermediate	✓
Duffing (undamped)	conservative, hardening	intermediate	$+1.172$	1.0676	0.994	stationary	intermediate	✓
Hopf below bifurcation	$\mu = -0.002, r_0 = 0.05$	intermediate	$+0.461$	0.3718	0.900	decaying†	non-power-law	**disagree**
Two-cycle radial	cycles $r = 1, 3$	non-power-law	$+0.425$	0.5750	0.625	stationary	non-power-law	✓
Duffing (pumped)	two cycles	non-power-law	-0.191	0.2134	0.375	stationary	non-power-law	✓

† stationarity flag raised. The miss: a decaying spiral whose concave log–log fit gives $R^2 = 0.8999$ against the 0.90 cutoff; threshold not adjusted.

Table 5: Grids and horizons per experiment. In the windowed hosts the sliding window of the model's last W outputs *is* the closed-loop state, so α multiplies the burned-in window before the input projection; every recurrent state restarts from zero at each re-forward.

experiment	α grid	T	W	n_{seq}
calibration (§5)	$2^{\text{linspace}(-2,2,7)}$	600 (ext. to 2400)	$T/7.5$	1
host scope (Table 1)	$\{\frac{1}{4}, 1, 4\}$	1000	128 / trailing 200	16
gate (Figure 2)	—	1000	32 / trailing 100	8
baselines (§6, Fig. 4)	$\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$	2000	32 and 128 / trailing 500	16
ablation, ECG	$\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$	2000	128 (ECG 256)	16
basin check	$\{\frac{1}{4}, \frac{1}{2}, 0.7, 1, 1.4, 2, 4\}$	4000	128	16

rows of Table 6 scale the readout $W_{\text{out}} \leftarrow gW_{\text{out}}$ by bisection until the $W = 32$ companion ρ equals 1.000000 or 0.990. Per-seed as-trained radii: 0.9866, 0.9963, 1.0031, 0.9975, 0.9828.

E.6 Baselines: implementation, full table and cost

All baselines run on the same windowed map ($W = 32$, 16 trajectories) in float64 with autodiff Jacobians (finite-difference check $\leq 1.1 \times 10^{-6}$ on every row). Lyapunov: Benettin/QR (Benettin et al., 1980) over 6000 steps after 500 warm-up; halves agree to 3×10^{-4}. Floquet: the integer-period monodromy depends on starting phase (raw $|\mu_1|$ swings 0.25–1.51, because the period 142.58 is not an integer and the relaxation jump changes the tangent speed by up to 0.41/step); we compute the monodromy at $\lfloor P \rfloor$ and $\lceil P \rceil$ over eight phases and interpolate $\log|\mu|$, which returns 0.9993 $[0.9988, 0.9997]$ for the true oscillator; $|\mu_2|$ is cross-checked against $e^{\lambda_2 P}$, agreeing to two digits

Table 6: The confounding case (five seeds per row, §5). KS: two-sample Kolmogorov–Smirnov statistic of the $|x|$ marginal against the target's.

closed loop	sustains bounded oscillation?	KS $\lvert x\rvert$	measured $\widehat{s}$
$\rho = 0.990$ (projected)	no, decays ($A(1) \approx 8 \times 10^{-8}$ of truth)	0.98–1.00	—[†]
$\rho = 1.000$ (projected)	**yes, indefinitely** (0.88–0.99× truth)	0.14–0.19	**1.00** (+0.996 to +1.013)
$\rho \in [0.983, 1.003]$ (as trained)	no: four decay, one diverges	0.98–1.00	+0.995 to +1.009
two-layer S4, norm-then-replace	yes (1.00–1.04× truth)	0.02–0.09	0.000 (5/5; degree 0, certified)

[†]Decayed to 10^{-7}; the residual's fitted slope is +1.00, the homogeneous transient.

wherever above 10^{-10}. Recovery: 10% kick at 8 phases × 16 trajectories, distance to the unperturbed orbit fitted log-linearly; for slowly decaying affine loops the same-index estimator with 1500 post-kick steps. Model (c) per seed: seeds 2 and 3 flat ($\widehat{s} = -6 \times 10^{-4}$, -7×10^{-4}); seeds 1 and 4 flat to three decimals for $\alpha \leq 1$ then 10^{10}–10^{15} at $\alpha \geq 2$; seed 0 has finals 0.075, 0.38, 1.20, 0.98 at $\alpha = \frac{1}{4}, \frac{1}{2}, 1, 2$ and 4×10^{14} at 4, the small-α means carried by $1/16$ and $5/16$ sustained sequences at the same ≈ 1.20 amplitude while the rest decay, so its basin is bounded on both sides; its orbit has period 159.5 against the true 142.6.

Table 7: **Reference-free α-slope against gradient-based baselines on the same closed loop.** Five van der Pol imitators (μ=2, 800 steps, 5 seeds; mean ± sd), float64, W=32, 16 trajectories. The α-verdict agrees with the Lyapunov/Floquet sign structure on **17 of 20** informative rows on the pre-registered grid; the exceptions are model-(c) seeds whose $\alpha \geq 2$ states leave a finite basin. The five (b) rows are degree-0 and certified by Floquet ($|\mu_2| \leq 8 \times 10^{-17}$) but excluded from the α count. Grey: misjudgements by ρ_0 (b) and by retention (e).

	(a) 1L S4 lin. readout	(b) 2L S4 norm-then-replace	(c) 2L S4 GELU, no LN	(d) GRU	(e) (a) at ρ=1.000
loop	affine	nonlinear	nonlinear	nonlinear	affine
$\hat{s}$ (W=32)	1.000±.005	-2×10^{-8}	esc. at $\alpha\geq2$ (3/5)[§]	−.0002±.0002	1.003±.007
$\hat{s}$ (W=128)	1.005±.006	2×10^{-8}	same[§]	−.0002±.0005	1.005±.006
λ_1	−.0068±.0075	−.00002	.00005	.00003	4×10^{-6}
λ_2	$= \lambda_1$	−.152±.006	−.025±.017	−.095±.018	-5×10^{-6}
$\lvert\lambda_1 - \log\rho_0\rvert$	$< 2\cdot10^{-5}$	–	–	–	$< 2\cdot10^{-5}$
$\lvert\mu_1\rvert$	0.64±0.58	1.0000±.0003	1.0000±.0001	1.0001±.0002	1.0000
$\lvert\mu_2\rvert$	$= \lvert\mu_1\rvert$	6×10^{-9}	0.15±0.17	1×10^{-5}	1.0000
$\tau_{\rm rec}$ / $-1/\lambda_2$	0.94–1.00[‡]	3.9/6.6	40/72	9.0/10.9	$> 10^4/\infty$
ρ_0	0.99±.01	3412±1521	1.06±.02	1.08±.04	1.0000
$\rho_{\rm orb}$	0.99±.01	0.996±.005	1.026±.004	1.035±.004	1.0000
amp / truth	0 or ∞	1.013±.016	1.03±.16	0.999±.002	0.92±.04
amp. retention	0 or ∞	1.000±.003	1.01±.03	1.000±.001	0.99±.00
KS($\lvert x\rvert$)	0.99±.01	.052±.021	.23±.06	.025±.012	.17±.02
$\hat{s}$ vs Lyap./Floquet	5/5	5/5	2/5[§] (5/5 in basin)	5/5	5/5

[‡]decaying seeds, ratio of recovery time to $-1/\lambda_2$ (268 vs 269; 402 vs 405 steps). [§]seeds 0, 1, 4 escape at $\alpha\geq2$ by the rule of §3; in-basin $\hat{s} = 0.000$ on all three.

F Ablation: rows and fate tables

Rows of Figure 6b (five seeds, loss-matched): 1 single S4, linear readout, $\widehat{s} = 1.000 \pm 0.008$; 2a biased MLP head inside the loop, 0.000 ± 0.003 at $W = 32$; 2b same head outside, 0.998 ± 0.004; 3 two-layer post-norm GELU+LN, 0.000 ± 0.000, retention 0.997, amplitude 1.003; 3p two-layer pre-norm, 0.000 on $3/5$, one finite basin, one fixed point; 4 two-layer GELU no norm, 0.000 on $3/5$, one finite basin, one fixed point; 5 two-layer bias-free ReLU no norm, $+1.000 \pm 0.000$, $R^2 = 1.000$; 6 GRU 0.001; 7 row 1 projected to $\rho = 1$, 1.000±0.007; 8 ReLU with bias no norm, -0.083 ± 0.169; 9 bias-free ReLU + affine-free LN, 0.000 ± 0.000; 10 bias-free GELU no norm, 0.017 ± 0.033. Row 2a per seed (4000 steps, 16 sequences per α): seed 0, 14–16/16 at a residual amplitude below 5% of

Table 8: Cost per model (wall-clock seconds, median over 5 seeds; one RTX 4090, float64, 16 trajectories, $W{=}32$). G: gradients of the loop map; R: reference continuation; P: period estimate.

Method	needs	(a)	(b)	(c)	(d)	(e)	$\div\, \hat{s}_{W32}$
α-slope, $W{=}32$	–	7.2	17.2	16.3	10.0	8.1	$1\times$
α-slope, $W{=}128$	–	23.1	52.9	52.9	34.4	25.3	3.1–$3.5\times$
Lyapunov (Jacobians + QR)	G	56.8	97.2	94.7	53.6	62.1	5.4–$7.9\times$
Floquet (Jacobians + monodromy)	G, P	40.8	81.8	73.7	39.6	44.8	4.0–$5.7\times$
Perturbation recovery	P	4.4	7.7	7.6	5.4	4.7	0.5–$0.6\times$
0–1 test	P	0.2	0.3	0.3	0.3	0.3	$0.03\times$
$\rho_{\mathrm{local}}(0)$	G (one)	0.02	0.03	0.02	0.02	0.02	$0.003\times$
MSE@k, KS, Hausdorff	**R**	0.2	0.2	0.2	0.2	0.2	$0.02\times$

the orbit; seed 1, $16/16$ divergent; seed 2, $16/16$ on an orbit at $1.4\times$; seeds 3–4, $16/16$ on the target orbit ($1.15\times$, $1.10\times$). Row 4 seed 2: $0/16$ divergent at $\alpha \le 1$ ($A = 1.72$–1.75), $3/16$ at 1.4, $13/16$ at 2, $14/16$ at 4; seed 3 fixed point. Row 3p seed 2: $6/16$ collapsed at $\alpha = \frac{1}{4}$, $0/16$ at $\alpha \ge 1.4$; seed 4 fixed point. Row 5 verification: bias parameters $\varnothing$, LayerNorms $\varnothing$, activations $\{\mathrm{ReLU}\}$, no S4 D term, homogeneity error 3.5×10^{-7}. GRU depth 4 no norm seed 0: amplitude 1.11, 1.00, 7.7×10^{8} at $\alpha = \frac{1}{4}, 1, 4$. Transformer seed 2: 0.83, 1.00, 1.03 (three-point grid, no CI).

G MIT-BIH

DS1/DS2 inter-patient split (train records 101, 106, 108, 109, 112, 114–116, 118, 119, 122, 124, 201, 203, 205, 207–209, 215, 220, 223, 230; test 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212–214, 219, 221, 222, 228, 231–234), channel 0 at 180 Hz, per-record z-scored, $L = 600$, batch 32, $W = 256$, context 400, 2000 free steps, per-sequence fates on 16 test contexts. Learning rate 3×10^{-3} ($n = 5$) and 10^{-3} ($n = 32$). The 1L S4 8000-step rerun: $\rho = 0.887$–0.913, collapsed $5/5$. Post-norm 2L S4 at 3×10^{-3}: validation MSE 0.614 at step 450, test 0.598 against persistence 0.110 and predict-mean 1.000; not used. GRU $n = 32$: 21 attractor, 11 non-power-law; 6 seeds with partial α-dependent collapse. 2L S4 $n = 32$: 3 attractor, 1 intermediate, 18 non-power-law, 10 collapsed; of 27 seeds with per-sequence dumps, 0 fully on orbit, 5 collapse in all 80 rollouts, 22 partially; no sequence diverged. Bootstrap Spearman (10^4 draws, pairs resampled) as in §8. The heart rate in Figure 7 is the dominant FFT period of the last 1000 free samples against the segment's 72.7 bpm.

H Pre-registration record

Eleven timestamped prediction files (epochs 1789950317–1790325982) are published unchanged with SHA-256 digests. Two hashed predictions bear on the claims. (1) The seed gate: predicted $|\mathrm{Spearman}(\rho, \log \mathrm{fin})| > 0.7$ and a loss–ρ CI containing zero, with the falsifier that a CI excluding zero would collapse the central claim; confirmed at $n = 24$ and $n = 64$; its loss–outcome corollary, confirmed at $n = 24$, fails marginally at $n = 64$ ($+0.26$ $[+0.004, +0.48]$) and is reported as such. (2) The first autonomous linear arm: predicted decay, observed divergence, which is why all three fates of Corollary 1.1 are enumerated and the centre case of §5 exists. Also wrong before data: that four-layer stacks would keep $\widehat{s} \ge 0.5$ (measured 0.000; now understood as forced by the normalise-then-replace wiring); that the two-layer non-trivial Floquet multiplier would lie in $[0.3, 0.95]$ (it is $\approx 10^{-17}$); that the integer-period monodromy was adequate (it is not). The bias-free ReLU stack, the ablation matrix and the sensitivity sweep were expectations in experiment briefs, not hashed predictions.