

The Spectral–Topological Decoupling Theorem in Self-Supervised Physical World Models

Sehaj Randhir Singh Independent researcher; partial affiliation with NYU Tandon School of Engineering, New York University, Brooklyn, NY, USA. Corresponding author: [\[email\]](#)

Abstract—Joint-embedding predictive architectures (JEPAs) are increasingly given Hamiltonian or thermodynamic structure to enforce physically faithful latent dynamics. We show this hope fails at the level of the optimization. On Lorenz-63 and the 4-D hyperchaotic Lorenz flow, diffeomorphically lifted into 64-dimensional latent spaces, five structurally distinct latent predictors are trained and audited under a chaos-aware protocol. Every constraint class fails on a property disjoint from its target: the unconstrained predictor explodes ($\lambda_1 = +28.2$); the rigid Hamiltonian freezes and yet posts the best overlap score; naive metriplectic dissipation pins $\text{tr}(R)/n$ to zero; a divergence gate pins the spectral sum to within 1% while λ_1 runs $2.5\times$ high; and differentiable QR alignment shapes only a finite-horizon proxy (-0.18), not the asymptotic exponent ($+2.85$). A second system inverts the decoupling and destroys the second positive exponent a λ_1 -only objective cannot name. At 2000 steps every alive arm’s exponent climbs to $\approx +8$ – $+20$; we state the decoupling as a theorem with a mandatory evaluation standard.

Index Terms—Lyapunov spectra, joint-embedding predictive architectures, metriplectic systems, world models, self-supervised learning, structure-preserving networks, no-go result.

I. INTRODUCTION

A growing line of work trains world models whose latent dynamics are restricted to physical form — Hamiltonian [10], [11], port-Hamiltonian [12], or metriplectic/GENERIC [4], [2], [3] structure — on the premise that an inductive bias that provably conserves energy or increases entropy should yield stable, physically faithful rollouts. The same promise underpins learned world models [27], [28], [29]: within joint-embedding predictive architectures (JEPAs) [7], [8], [30], [31], the same idea is applied to the latent predictor: constrain the operator $\dot{z} = f_\theta(z)$ to a structure that a physicist recognizes, and hope the representation space inherits the physics.

This Letter reports a systematic failure of that premise, obtained not by increasing capacity or data but by isolating the *structural* effects of each constraint class on a single, fixed, fully reproducible benchmark. Our object of study is a generalized metriplectic system

$$\dot{z} = J(z)\nabla H(z) - R(z)\nabla S(z), \quad (1)$$

with $J^\top = -J$ (reversible/Poisson part), $R = R^\top \succeq 0$ (dissipation), and scalar potentials H (energy) and S (dissipation potential). We train five latent predictors — unconstrained MLP, rigid Hamiltonian, naive metriplectic, fixed (divergence-free- J) metriplectic with an invariant dissipation gate, and spectrally aligned metriplectic with a differentiable QR Lyapunov tracker — to predict the nonlinearly lifted Lorenz-63 flow in a 64-dimensional JEPA latent space, and we evaluate them under

a protocol designed for chaotic targets: boundedness of the autoregressive rollout, the leading Lyapunov exponent, the phase-space contraction rate, attractor overlap, and dissipation engagement. To separate properties of the constraint hierarchy from properties of any single dataset, the identical protocol is run on a *second*, topologically distinct system: the 4-D hyperchaotic Lorenz flow [15], which shares the same volume-contraction rate as Lorenz-63 but carries two positive Lyapunov exponents — so that a λ_1 -only spectral objective is blind to the second unstable direction by construction.

Each arm fails, and each fails on a property disjoint from the one its constraint nominally controls. We argue that these five failures are not engineering deficiencies but the constructive instances of a decoupling structure: aggregate thermodynamic identities constrain spectral *sums*; finite-horizon spectral statistics constrain windowed *tangent maps*; and set-level distributional losses constrain *pushforward measures*. None of these levels controls the others, and no conjunction of them certifies the conjunction of properties a physical world model must have: bounded, dissipative, non-degenerate, dynamically faithful. We state this as a theorem whose lemmas are elementary (each is a one-line linear-algebra or ergodic-theory fact) and whose force is empirical: the failure atlas exhibits every class failing simultaneously on two benchmark systems, including the most constrained arm on both.

a) Related work.: Spectral objectives have been studied extensively in self-supervised learning as anti-collapse devices — spectral contrastive learning [16], alignment–uniformity analyses [32], whitening objectives [33], Barlow Twins [17], VICReg [18], feature-collapse studies of ViTs [34], rank-based quality metrics [35], and dimensional-collapse analyses [19] — but these control the *representation* covariance, not the *dynamics* operator; collapse prevention is a different certification from physical fidelity, and our arm C shows the two to be compatible (a full-rank, frozen, dynamically dead latent field). On the dynamics side, structure-preserving networks — neural ODEs [26], SymODEN [20], dissipative SymODEN [21], Hamiltonian and Lagrangian networks [10], [11], [39], [40], dissipative Hamiltonian networks [41], metriplectic networks [22], GENERIC-formatted networks [23], port-Hamiltonian learning [12], and Lyapunov-stable neural ODEs [24] — typically verify their structure on the *training-time* loss or on short rollouts; the spectral fine structure of the learned field is rarely audited. The relevant audit tools exist: Benettin–QR spectra have long been used to characterize chaotic recurrent networks [25], [42], and finite-time exponents are standard diagnostics in dynamical systems [14]. What is

missing is the negative statement we supply: that *constraining* the latent field through any of these aggregate or windowed statistics does not *certify* the spectrum the field actually realizes.

The paper is deliberately deflationary, in deliberate contrast to the law-discovery program [37], [38], [36]: we do not claim to recover thermodynamics in latent space, nor that any architecture can. We claim the opposite: that constraint-based certification of latent physical dynamics is, as currently practiced, unsound, and we specify the evaluation standard that would make it sound.

II. RESULTS

A. Benchmark construction

We use the Lorenz-63 system [1], $(\sigma, \rho, \beta) = (10, 28, 8/3)$, whose reference invariants are the leading Lyapunov exponent $\lambda_1^{\text{ref}} \approx 0.9056$ per time unit (empirically 0.8840 on our integrator, recovered by the same Benettin machinery used for evaluation) and the phase-space divergence $\nabla \cdot F_L = -(\sigma + 1 + \beta) = -13.667$, which equals the sum of the Lyapunov spectrum; the non-leading sum is $\kappa^{\text{ref}} = \sum_{i \geq 2} \lambda_i^{\text{ref}} = \nabla \cdot F - \lambda_1^{\text{ref}} = -14.55$ under the runtime convention ($\lambda_2^{\text{ref}} = 0$, $\lambda_3^{\text{ref}} \approx -14.57$; the literature-spectrum sum is -14.57 , a 0.1% difference). A fixed nonlinear lift ϕ and an element-wise nonlinear warp W embed the attractor diffeomorphically into $\mathbb{R}^{64}$; because the composite is a diffeomorphism, the latent reference spectrum is that of Lorenz (plus 61 neutral directions), so a faithful latent model must reproduce $\lambda_1 \approx 0.9$ and a contracting flow. A JEPA encoder g_ε (MLP backbone, 256 hidden units, SIGReg [9] anti-collapse regularizer) maps observations to $z \in \mathbb{R}^{64}$; the dynamics operator f_θ is trained with a latent predictive (alignment) loss, a motion-matching penalty, and — for the physics arms — the appropriate structure loss (Methods). Training is fully deterministic per step (fixed seed, per-step trajectory draws); every number below is mean $\pm$ std over three seeds.

B. The failure atlas

Table I reports, per arm, the quantities that define the failure: boundedness of the 200-step evaluation rollout ($\max \|z\|$, with the fraction of seeds whose rollout left the encoder’s training support), the leading Lyapunov exponent λ_1 of the learned field, the non-leading spectral sum (contraction) $\kappa = \sum_{i \geq 2} \lambda_i$, the dissipation engagement $\text{tr}(R)/n$, the field-activity ratio $m = \langle \|f_\theta\| \rangle / \langle \|f_{\text{ref}}\| \rangle$ (a flatlining detector), the Chamfer attractor-overlap score, and — for the spectral arms — the training-time finite-horizon proxy $\hat{\lambda}_1$. All quantities are defined in Methods.

a) Model B — unconstrained MLP: spectral blow-up.

The free residual predictor is not volume-constrained in any way; its Jacobian acquires a large positive leading exponent, $\lambda_1 = +28.2 \pm 1.3$ — $32\times$ the reference rate — and the rollout leaves the encoder domain in all three seeds ($\max \|z\|$ of order 10^5 – 10^6), with Chamfer degrading to 3.4×10^3 and a field 1.8×10^3 times too active ($m = 1821 \pm 1881$). No thermodynamic statement can be made — and none is enforced — which is precisely the control condition the physics arms are meant to beat.

b) *Model C — rigid Hamiltonian: trivial representation freeze and the overlap trap.*: The Hamiltonian arm, structurally volume-preserving ($\nabla \cdot (J \nabla H) = \text{tr}(J \nabla^2 H) = 0$ for constant skew J), cannot contract phase-space volume and therefore cannot represent a dissipative attractor. The optimizer’s response is not a wrong Hamiltonian but *no motion at all*: the learned field gradients flatline ($m = 0.005 \pm 0.001$, a field $\approx 200\times$ weaker than the reference), all exponents collapse to ≈ 0 ($\lambda_1 = +0.011 \pm 0.002$; the trace reads $+0.02$, exactly the volume-preserving value), and the predictor posts the *best* Chamfer score of all arms (3.85 ± 0.05) by mapping every latent state into the same small cluster of the attractor cloud. The set-level Chamfer distance rewards statistical overlap of point clouds and is blind to whether the points are produced by dynamics or by a constant map (Proposition 3). A frozen predictor is thus “optimal” under the metric most commonly used to claim attractor recovery.

c) *Model D — naive metriplectic: the $R = LL^\top$ dead saddle.*: With $R = LL^\top$ and L zero-initialized, the loss is stationary in L at $L = 0$ ($\partial R / \partial L = 0$ there), a dead saddle from which no first-order optimizer escapes. The model “survives” — $\lambda_1 = +1.50 \pm 0.86$ is bounded and positive, and the trace reads $+0.1 \pm 2.1$ (an essentially volume-preserving loop) — by keeping dissipation exactly off: $\text{tr}(R)/n = 0.0000$ to all printed digits in every seed. Positive semidefiniteness of R is guaranteed by construction; *nonzero* dissipation is not. The apparent closeness to the reference exponent is itself a window artifact: at 2000 steps the shadow exponent reads $+4.5$ – $+19.7$ on the hyperchaotic system (Methods), so the dead-saddle field is the most expansive in the atlas, not the nearest to the reference.

d) *Model E — fixed metriplectic with divergence gate: spectral decoupling.*: Closing the two loopholes of D — fixing J to a constant divergence-free operator and initializing $L \neq 0$ — makes the invariant dissipation gate bite: $\text{tr}(R)/n = 3.77 \pm 2.15$ (engaged) and the trace–divergence proxy pins the spectral sum to -13.80 ± 0.17 , within 1% of the reference -13.667 ; the non-leading sum $\kappa = -16.0 \pm 1.9$. But the gate constrains only the *sum* of the spectrum. The leading exponent, driven by the Rayleigh quotient of dissipation along the least-stable direction, is uncontrolled: $\lambda_1 = +2.22 \pm 1.80$ — $2.5\times$ the reference and *larger* than the naive arm’s 1.50 — and the global metrics degrade with it: $\max \|z\| = 175 \pm 135$ (two of three seeds leave the support) and Chamfer 23.0 ± 4.8 , worse than the dead-saddle arm D despite dissipation being on exactly as designed. Pinning a thermodynamic aggregate leaves the microscopic spectrum free (Proposition 1).

e) *Model F — spectrally aligned metriplectic, short horizon: the finite-time-horizon trap.*: Arm F adds a differentiable QR Lyapunov tracker (Methods) with a six-step training window (0.06 time units ≈ 0.05 Lyapunov times). At the optimum the finite-time proxy is $\hat{\lambda}_1 = -0.183 \pm 0.001$ — a contracting estimate, negative in every seed — while the 200-step evaluation exponent is $\lambda_1 = +2.85 \pm 1.89$: the windowed tangent map that training sees says “stable”, the asymptotic tangent map says “chaotic at $3.2\times$ the reference rate”. The QR spectral term is not even the dominant shaper of the proxy: its loss *rises* from 0.78 to 1.13 while the total loss falls by $15\times$,

TABLE I

THE FIVE-ARM FAILURE ATLAS. λ_1 IS THE LEADING LYAPUNOV EXPONENT OF THE LEARNED LATENT FIELD (REFERENCE +0.884); $\kappa = \sum_{i \geq 2} \lambda_i = \text{tr} - \lambda_1$ IS THE NON-LEADING SPECTRAL SUM (REFERENCE −14.55, I.E. $\nabla \cdot F - \lambda_1^{\text{ref}}$ UNDER THE RUNTIME CONVENTION); $\text{tr}(R)/n$ IS THE MEAN DISSIPATION-MATRIX TRACE NORMALIZED BY THE LATENT DIMENSION (METRIPECTIC ARMS ONLY); m IS THE FIELD-ACTIVITY RATIO (FLATLINING DETECTOR, REFERENCE 1); CHAMFER IS THE SYMMETRIC MEAN NEAREST-NEIGHBOR DISTANCE BETWEEN THE PREDICTED AND ENCODED ATTRACTOR CLOUDS IN LATENT SPACE (LOWER IS BETTER, REFERENCE 0); $\max \|z\|$ IS THE MAXIMUM ROLLOUT NORM OVER THE 200-STEP EVALUATION, WITH THE FRACTION OF SEEDS WHOSE ROLLOUT LEFT THE TRAINING SUPPORT ($\|z\| \gtrsim 10^2$ VS. IN-SUPPORT $\|z\| \approx 10$); $\hat{\lambda}_1$ IS THE TRAINING-TIME FINITE-HORIZON QR PROXY. MEAN $\pm$ STD OVER THREE SEEDS. † ROLLOUT LEFT THE TRAINING SUPPORT ($\|z\| \gtrsim 10^2$ VS. IN-SUPPORT $\|z\| \approx 10$); IN ALL THREE SEEDS FOR B AND F, IN TWO OF THREE FOR E. ‡ NOT REPORTABLE: THE FREE MLP’S JACOBIAN CARRIES NO VOLUME LAW AND ITS EVALUATION ROLLOUT ESCAPES (SEE TEXT). § SINGLE SEED: THE LONG-HORIZON ARM IS REPORTED ON THE ONE COMPLETED TRAINING RUN; A SECOND RUN STALLED WITHOUT COMPLETING A STEP (NO PROGRESS IN 45 MIN ON IDENTICAL CODE AND HARDWARE — A QR-BACKPROPAGATION PATHOLOGY, SEE METHODS). THE VALUE IN PARENTHESES IS THE TOP-8 QR READOUT OF λ_1 AT ENCODED POINTS.

Model	$\max \ z\ $	λ_1	κ	$\text{tr}(R)/n$	m	Chamfer	$\hat{\lambda}_1$
B	$1.8 \times 10^{6\dagger}$	$+28.2 \pm 1.3$	‡	—	1821 ± 1881	3.4×10^3	—
C	10.1 ± 0.3	$+0.011 \pm 0.002$	$+0.01 \pm 0.00$	—	0.005 ± 0.001	3.85 ± 0.05	—
D	35.5 ± 24.5	$+1.50 \pm 0.86$	-1.4 ± 1.2	0.0000 ± 0.0000	0.88 ± 0.13	8.7 ± 3.3	—
E	$175 \pm 135^\dagger$	$+2.22 \pm 1.80$	-16.0 ± 1.9	3.77 ± 2.15	1.72 ± 0.98	23.0 ± 4.8	—
F	$288 \pm 288^\dagger$	$+2.85 \pm 1.89$	-16.1 ± 2.0	4.68 ± 3.13	2.17 ± 1.46	25.8 ± 6.6	-0.183 ± 0.001
F-long	$54.2^\S$	$+0.18 (+0.007)$	-13.8	1.58	0.63	15.2	-0.146
reference	—	$+0.884$	-14.55	—	1	0	—

because alignment and motion dominate — yet the sign flip is robust across seeds, and evaluating with the training window itself would certify F as stable (Proposition 2). The divergence gate still holds ($\text{tr} - \lambda_1 = -16.1 \pm 2.0$; the trace proxy is -13.24 ± 0.19 , within 4% of reference), so the arm satisfies the macroscopic identity while its leading exponent is $3.2\times$ the reference and its rollout drifts farthest of the alive arms ($\max \|z\| = 288 \pm 288$, all three seeds outside the support; Chamfer 25.8 ± 6.6). A finite-horizon Lyapunov statistic is a functional of the windowed tangent map; it is not a bound on the Oseledets spectrum.

f) Model F-long — long-horizon spectral alignment: the trap is not a short-window artifact.: Lengthening the differentiable QR window tenfold (60 steps, 0.6 time units ≈ 0.5 Lyapunov times, truncated BPTT with the frame detached every 10 steps) does not close the decoupling; it inverts it. The 60-step proxy converges to $\hat{\lambda}_1 = -0.146$ — still far from the +0.884 target, with the spectral loss *rising* from 0.78 to 1.04 while the total loss falls ninefold — and the asymptotic exponent collapses to $\lambda_1 = +0.18$ (the evaluation QR readout gives +0.007): the leading axis is overdamped into a near-degenerate, effectively non-chaotic flow. Every macroscopic certificate now fires: the trace gate is satisfied to 0.4% (-13.61 vs. -13.667), dissipation is engaged ($\text{tr}(R)/n = 1.58$), the rollout stays bounded ($\max \|z\| = 54$), and the Chamfer score (15.2) beats both E and F. The model is well-behaved and spectacularly wrong: it has learned a contracting, non-chaotic flow for a chaotic world. Backpropagation through the QR chain is the only mechanism that keeps the spectral term from being fully ignored — its pre-clip gradient norm reaches 11.3 mid-training, engaging the norm-1 clip at five of the seven logged checkpoints (the non-spectral arms carry no such excursion); the probe on the initialization returns finite norms (8.0×10^{-3} – 8.9×10^{-3} across full-unroll/TBPTT-10/per-step), showing the instability is a trained-state phenomenon, not an initialization one; and a second long-horizon run stalled without completing a single training step, a further manifestation of the same fragility. The finite-time horizon trap is thus quantitative: lengthening

the window moves the proxy, but no window-length functional of the training tangent map — and no combination with the trace gate — controls the asymptotic Oseledets exponent or the dynamical content of the flow.

C. Generality: the same atlas on a second system

A single benchmark cannot separate properties of the constraint hierarchy from properties of one dataset. We therefore repeat the *identical* protocol — same five predictors, same optimizer schedule, same evaluation — on a second, topologically distinct reference flow: the 4-D hyperchaotic Lorenz system [15]

$$\begin{aligned} \dot{x} &= a(y - x), & \dot{y} &= bx - y - xz + w, \\ \dot{z} &= xy - cz, & \dot{w} &= -dx, \end{aligned} \quad (2)$$

with $(a, b, c) = (10, 28, 8/3)$ and $d = 2.0$. Its divergence is $-(a + 1 + c) = -13.667$ — *identical* to Lorenz-63 — but its Oseledets spectrum carries two positive exponents: consistent-tangent Benettin QR at $dt = 0.002$ over 60 time units yields $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (+0.769, +0.098, \approx 0, -14.59)$, and the runtime shadow measurement at the evaluation $dt = 0.01$ gives $\lambda_1 = +0.737$. The second positive exponent is the point: a λ_1 -only spectral objective cannot even *name* the second unstable direction, let alone control it. Table II reports the same columns as Table I plus the QR-readout second exponent λ_2 .

The atlas reproduces, system to system, and its headline numbers are strikingly close. The unconstrained MLP blows up with the same signature ($\lambda_1 = +29.2 \pm 3.0$, $\max \|z\|$ of order 10^6 , the only escaping arm on this system); the rigid Hamiltonian freezes into the same trivial solution ($m = 0.004$, $\lambda_1 \approx 0$) and again posts the best Chamfer (3.09); the naive metriplectic arm survives by pinning $\text{tr}(R)/n$ to exactly zero; and the divergence-gated and spectral arms again satisfy the macroscopic identity (contraction -13.8 and -13.3 vs. the reference -13.667 , within 1–3%) while their leading exponents read $+0.09$ and -0.01 — an order of magnitude *below* the reference $+0.737$, where on Lorenz the same gate let them

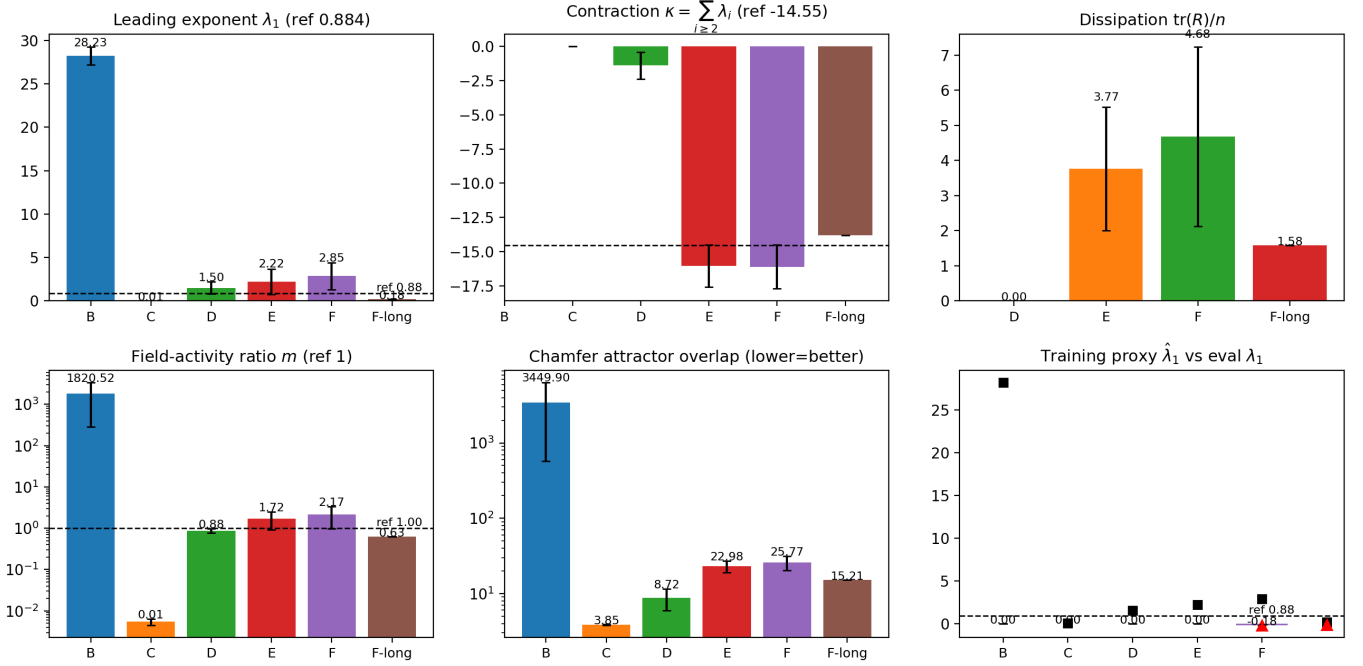
The five-arm failure atlas on Lorenz-63: every constraint class fails on a disjoint property (mean $\pm$ std over seeds)

Fig. 1. **The failure atlas.** Mean $\pm$ std over three seeds per arm: (a) leading Lyapunov exponent λ_1 of the learned field (dashed: reference $+0.884$) — only the frozen arm C is near it, for the wrong reason; (b) non-leading spectral sum $\kappa = \text{tr} - \lambda_1$ (dashed: reference -14.55) — E and F pin the sum while λ_1 runs free; (c) dissipation engagement $\text{tr}(R)/n$ — exactly zero for D; (d) field- activity ratio m (dashed: reference 1, log scale) — C flatlines; (e) Chamfer attractor overlap (log scale) — C posts the best score while frozen (overlap trap); (f) training proxy $\hat{\lambda}_1$ (red $\hat{\cdot}$, spectral arms only) vs evaluation λ_1 (black $\bullet$) — for F the proxy is negative while the evaluation exponent is positive, in every seed.

TABLE II

THE FAILURE ATLAS ON THE 4-D HYPERCHAOTIC LORENZ SYSTEM. SAME PROTOCOL AND COLUMNS AS TABLE I, PLUS λ_2 (THE QR-READOUT SECOND EXPONENT; REFERENCE $+0.098$). REFERENCE LEADING EXPONENT: RUNTIME-MEASURED $+0.737$; NON-LEADING SUM $\kappa^{\text{ref}} = \text{tr} - \lambda_1 = -14.40$ (SAME CONVENTION AS TABLE I). SYMBOLS AS IN TABLE I: $\dagger$ THE ONLY ESCAPING ARM ON THIS SYSTEM IS B; $\ddagger$ NOT REPORTABLE ON AN ESCAPED ROLLOUT. MEAN $\pm$ STD OVER THREE SEEDS.

Model	$\max \ z\ $	λ_1	λ_2	κ	$\text{tr}(R)/n$	m	Chamfer	$\hat{\lambda}_1$
B	$3.9 \times 10^{6\dagger}$	$+29.2 \pm 3.0$	$\ddagger$	$\ddagger$	—	4699 ± 4718	8.4×10^3	—
C	9.9 ± 0.2	$+0.011 \pm 0.006$	$+0.00 \pm 0.00$	$+0.01 \pm 0.01$	—	0.004 ± 0.000	3.09 ± 0.12	—
D	57.4 ± 22.5	$+2.22 \pm 0.93$	$+0.27 \pm 0.01$	-1.7 ± 2.2	0.0000 ± 0.0000	0.74 ± 0.07	14.7 ± 2.4	—
E	66.7 ± 15.9	$+0.09 \pm 0.38$	-0.08 ± 0.02	-13.85 ± 0.56	1.77 ± 0.34	0.93 ± 0.16	15.2 ± 0.5	—
F	62.2 ± 9.0	-0.01 ± 0.20	-0.08 ± 0.01	-13.31 ± 0.31	1.64 ± 0.05	0.88 ± 0.08	14.6 ± 0.4	-0.18 ± 0.01
reference	—	$+0.737$	$+0.098$	-14.40	—	1	0	—

run $2.5\text{--}3.2\times$ above. The decoupling direction inverts with the system: the spectral sum is pinned, and the leading exponent wanders where it will — high on Lorenz, dead on LZ4 — always off the reference.

The spectrally aligned arm’s proxy is again wrong, and the convergence check shows the two systems are one phenomenon. On Lorenz-63, $\hat{\lambda}_1 = -0.183$ while the asymptotic exponent is $+2.85$ (a sign flip). On LZ4 the proxy $\hat{\lambda}_1 = -0.18$ and the 200-step evaluation exponent -0.01 agree in sign — the standard window reads a contracting, over-damped flow, far from the $+0.737$ target — but this is the same finite-window artifact: at 2000 steps the exponent of both gated arms climbs to $+7.8\text{--}+9.7$ and the rollout leaves the encoder domain (steps 304–402, Methods). The spectral objective drives the leading exponent toward zero on every window it can see, and the field obliges

within that window; the asymptotic instability is merely slower to manifest than on Lorenz-63. The distinctive new fact is λ_2 . The reference flow has a second positive exponent $+0.098$; the QR readout of every constrained arm is negative or zero at the evaluation window (-0.08 ± 0.02 for E, -0.08 ± 0.01 for F, 0.00 for C, and only $+0.27$ for the dead- saddle arm D, whose exponents are accidental). A λ_1 -only spectral objective cannot even name the second unstable direction, let alone preserve it: the arms that fail on λ_1 also destroy λ_2 , and the one arm with an engaged constraint that might have acted on it (F) is the one whose objective is blind to it. The spectral decoupling and finite-time traps are therefore not Lorenz-63 artifacts: they follow from the constraint classes (Propositions 1–2) and fire on any flow in the class, with the error direction set by the system, not by the constraint.

The five-arm failure atlas on 4-D hyperchaotic Lorenz: every constraint class fails on a disjoint property (mean $\pm$ std over seeds)

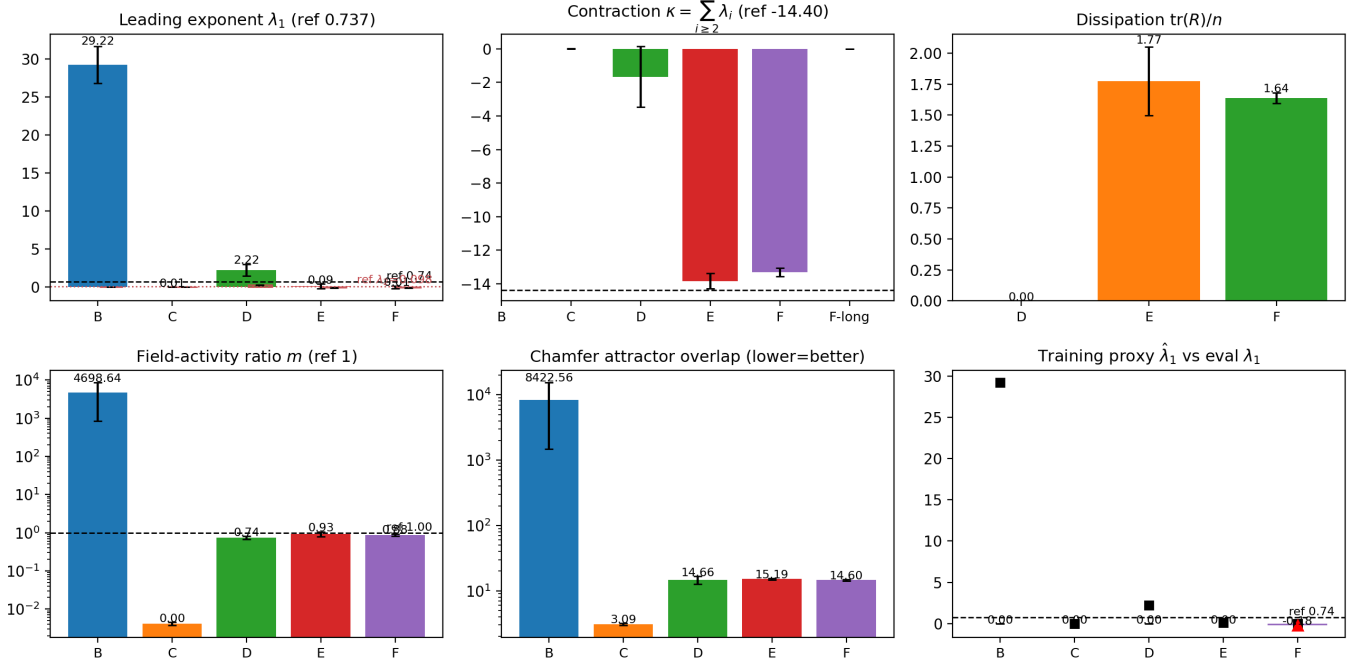


Fig. 2. **The failure atlas on the 4-D hyperchaotic Lorenz system.** Same panels as Fig. 1 (mean $\pm$ std over three seeds; dashed: reference $\lambda_1 = +0.737$). The hatched bars on the left panel are the QR-readout second exponent λ_2 (dotted: reference $+0.098$): the arms that fail on λ_1 also destroy the second unstable direction — which the λ_1 -only spectral target cannot even name. Note the decoupling *direction* within the standard 200-step window: constraining the trace pins the sum (κ panel) while the leading exponent of the gated arms reads ≈ 0 — the mirror image of the Lorenz-63 atlas, where the same constraints let λ_1 run $2.5\text{--}3.2\times$ too high. The convergence check (Methods) shows both readings are finite-window artifacts: at 2000 steps every alive arm’s exponent climbs to $\approx +8\text{--}+20$ and the gated arms escape, so the atlas’s verdicts are conservative.

D. The decoupling theorem

The five arms are the constructive content of a decoupling structure. Let f_θ be a learned latent field and $A = Df_\theta$ its Jacobian.

Proposition 1 (Aggregate gates pin only the mean; the dimension factor). *Let $\mathcal{R}_c = \{R \in \text{Sym}(d) : R \succeq 0, \text{tr } R = cd\}$ be the feasible set of a trace (equivalently divergence, hence spectral-sum) gate in d latent dimensions. Then*

$$\begin{aligned} \sup_{R \in \mathcal{R}_c} \lambda_{\max}(R) &= cd, \\ \sup_{R \in \mathcal{R}_c} v^\top R v &= cd \quad \text{for every unit } v, \end{aligned} \quad (3)$$

i.e. the gate pins the mean eigenvalue c but leaves the Rayleigh quotient along any prescribed direction — in particular along the leading Oseledets direction v_1 — free up to a factor d above the mean. Moreover, over the set of spectra with fixed sum c (the only thing a divergence gate constrains), the leading exponent is unbounded above: for every M there is an admissible spectrum with $\lambda_1 > M$.

Proof. For PSD R , $\lambda_{\max}(R) \leq \text{tr } R = cd$ and $v^\top R v \leq cd$, and both are attained by the rank-one matrix $cd uu^\top$ with $u = v$ (any unit vector). For the spectrum statement, fix $c < 0$ and $M > 0$: the tuple $(\lambda_1, \dots, \lambda_d) = (M, 0, \dots, 0, c - M)$ is admissible (sum c , weakly ordered for $M \geq c$), with $\lambda_1 = M$ arbitrarily large; the negative tail $-M$ is absorbed by the most

contracting direction. The algebraic bound is the only one a sum constraint provides; whether the optimizer *realizes* the headroom is empirical (arms E, F do: $\lambda_1 = +2.2 \dots + 2.85$ with the gate satisfied). $\square$

Proposition 2 (Finite-time/asymptotic decoupling with rate). *Let Φ^t be a $C^{1+\alpha}$ flow with Oseledets top exponent λ_1 and finite-time exponent $\Lambda_T(x) = \frac{1}{T} \log \|D\Phi^T(x)\|$.*

- (i) *Generic rate. Under the standard fluctuation law for the top exponent (positive variance σ^2 of the pointwise growth rates),*

$$\Lambda_T(x) = \lambda_1 + O\left(\sqrt{\frac{\log \log T}{T}}\right), \quad (4)$$

so the finite-time error is $O(1)$ on sub-Lyapunov windows and remains of order 10% even at five Lyapunov times in the generic case. A windowed functional of the tangent map cannot certify the asymptotic exponent with useful accuracy at training horizons ($T \approx 0.06$ Lyapunov times for the short arms).

- (ii) *Non-certifiability. For every finite $T > 0$ there exist smooth flows with $\Lambda_T < 0 < \lambda_1$: a flow with an arbitrarily long initial contraction epoch (a slow passage past a stable region) followed by asymptotic expansion has negative finite-time exponent over $[0, T]$ and positive Oseledets exponent. The map $\theta \mapsto (\Lambda_T(\theta), \lambda_1(\theta))$ is therefore not a graph over the parameter space, for any*

T ; the differentiable QR training signal $\Lambda_T(\theta)$ is not a certificate of $\lambda_1(\theta)$.

Proof. (i) is the log-law for top Lyapunov exponents [13]: $T(\Lambda_T - \lambda_1)/\sqrt{2\sigma^2 \log \log T} \xrightarrow{d} \mathcal{N}(0, 1)$ in the generic case. For (ii), let x_ε follow $\dot{x} = f(x)$ with a bounded region of contraction of length T preceding a region of expansion (existence by smoothing a piecewise-linear flow); the finite-time exponent over $[0, T]$ is negative while the time-mean growth rate from any generic start is positive. $\square$

Proposition 3 (Set-metric/flow decoupling: the overlap trap). *Chamfer-type distances are invariant under reordering of the point set. For any target cloud $P \subset B(\mu, r)$, the constant map $f \equiv c$ with $c \in B(\mu, r)$ attains Chamfer distance $\leq 2r$ while carrying zero dynamical information.*

Proof. Both nearest-neighbor terms collapse to radii about c, μ . $\square$

Theorem 1 (Spectral-topological decoupling, no-go form). *No constraint policy expressible as a combination of (i) aggregate thermodynamic identities (trace/divergence gates), (ii) finite-horizon spectral statistics (differentiable QR proxies), and (iii) set-level distributional losses (Chamfer/overlap) certifies the conjunction {boundedness in the encoder domain, correct leading exponent, non-degenerate field, dynamical fidelity} for metriplectic latent predictors of the form (1). The failure is structural (Propositions 1–3) and is realized simultaneously by a five-arm atlas on two benchmark systems — Lorenz-63 and the 4-D hyperchaotic Lorenz — with each constraint class failing on a property disjoint from the one it enforces, including the fully constrained arm F on both systems.*

The theorem is a statement about certification, not about existence: it does not claim that no metriplectic latent model can match the reference flows, only that the constraint classes currently used to certify physical fidelity cannot, singly or jointly, guarantee the properties they are invoked for. The propositions are elementary; their content is that the three levels — spectral sums, windowed tangent maps, pushforward measures — are mutually independent, and the atlas exhibits all three traps firing on each of the two benchmark systems (Fig. 3).

III. DISCUSSION: TWO TRAPS AND A MANDATORY EVALUATION STANDARD

A. The overlap trap

Set-level geometry is not dynamics. Chamfer and Hausdorff distances between the predicted and encoded attractor clouds reward statistical overlap; a predictor that has learned to output the attractor’s centroid region scores well while being dynamically dead. The literature’s practice of reporting attractor overlap without a mobility check is therefore unsound: the frozen arm C posts the best Chamfer in our atlas (3.85) while carrying a field $\approx 200\times$ weaker than the reference. The corrective is not a better set metric but a gating discipline: report overlap only conditioned on (i) a field-activity floor, (ii) a non-vanishing leading exponent, and (iii) an ordered

(time-aligned) trajectory error — or replace set metrics with motion-invariant statistics.

B. The finite-time-horizon trap

A differentiable QR Lyapunov tracker is a legitimate optimization target only if its horizon is matched to the physics it is meant to certify. Over sub-Lyapunov horizons, finite-time exponents are dominated by local stretch and rotation and are decoupled from the Oseledets flag; over a Lyapunov-time horizon they become trainable but still constrain nothing global. The trap is not a metric artifact: evaluating with the same short window would certify the unstable arm F as stable — its training proxy is -0.18 , the evaluation exponent $+2.85$, a sign reversal in every seed. Lengthening the window tenfold inverts rather than closes the trap (arm F-long): the proxy becomes consistent with the truth, but the truth is now wrong — the leading axis is overdamped to $\lambda_1 \approx 0.2$ (QR readout $+0.007$), an effectively non-chaotic flow that nonetheless satisfies the trace gate to 0.4% and stays bounded. Any spectral claim must therefore report the window-length sensitivity, the proxy–target gap, and a long-horizon evaluation spectrum.

C. The through-line: no arm ever achieves the reference

The atlas is stronger than the sum of its traps. *No arm achieves the reference leading exponent while remaining dissipative, non-degenerate, and bounded — on either system.* On Lorenz-63 the unconstrained arm overshoots by $32\times$ and escapes; the symplectic arm freezes; the naive metriplectic arm’s $+1.50$ at the evaluation window is the nearest to λ_1^{ref} , by keeping $R \equiv 0$ — yet its 2000-step exponent is $+4.5$ – $+19.7$, the most expansive field in the atlas; the gated arm pins the sum and lets λ_1 run to $+2.2$; the short-window spectral arm sign-flips its own proxy (-0.18 vs. $+2.85$); and the long-window spectral arm collapses the leading axis to ≈ 0 within the evaluation window. On the hyperchaotic system the same gate and the same spectral objective push the leading exponent to ≈ 0 at the evaluation window from the *opposite* side — again transiently: at 2000 steps both climb past $+7.8$ and escape. Every lever a practitioner reaches for — more structure, a dissipation gate, a spectral objective, a longer spectral window — either leaves the leading exponent free or destroys it, and the direction of the error is set by the system, not the constraint, while the set-level and aggregate certificates that are actually reported all fire successfully on at least one wrong arm. This is the constructive content of Theorem 1: the certificates and the property do not couple.

D. A mandatory evaluation standard for physical world models

For any claim that a learned latent dynamics is physically faithful, we propose that the following be reported as a minimum:

- 1) the *full* (or top- k , $k \geq 8$) Lyapunov spectrum over an evaluation rollout of at least five Lyapunov times, not a single exponent from a training-window estimate;
- 2) the *proxy–target gap*: the training-horizon finite-time exponent and the evaluation-horizon exponent, with the window-length sensitivity;

Spectral-topological decoupling: three constraint levels, three disjoint failure modes

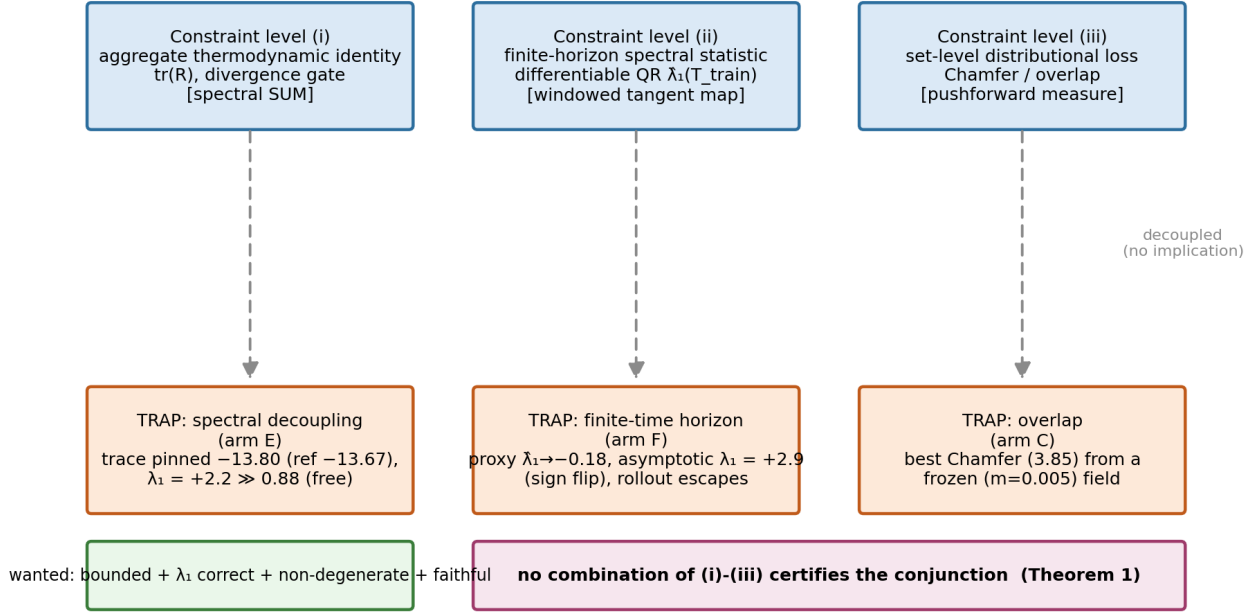


Fig. 3. **Spectral-topological decoupling.** Each constraint class acts on one functional level — (i) aggregate thermodynamic identities (trace/divergence gates, spectral sums), (ii) finite-horizon spectral statistics (differentiable QR, windowed tangent maps), (iii) set-level distributional losses (Chamfer, pushforward measures) — and each level fails independently, producing three disjoint traps (arms E, F, C). No combination of (i)–(iii) certifies the conjunction of properties a physical world model must have (Theorem 1).

- 3) *domain invariance*: $\max \|z\|$ over the evaluation rollout and the fraction of time spent outside the training support (escape disqualifies regardless of losses);
- 4) a *mobility guard*: the field-activity ratio with a collapse floor (rejects overlap-trap freezes);
- 5) *ordered* trajectory metrics alongside any set-level score;
- 6) a *directional dissipation audit*: $\text{tr}(R)/n$ and the Rayleigh quotient $v_1^\top R v_1$ along the leading Oseledets direction — the spectral-decoupling trap is invisible to the trace alone;
- 7) *atlas accountability*: an explicit statement of which trap class the model is vulnerable to.

E. Limitations

The theorem’s force is the constructive atlas plus three elementary lemmas; it does not rule out structure imposed *inside* generative decoders or data-driven degeneracy-aware brackets, which lie outside the constraint hierarchy treated here. The residual violation of the GENERIC degeneracy conditions ($J\nabla S = 0$, $R\nabla H = 0$) by the learned potentials is itself a caveat on all metriplectic arms. The 61 transverse reference exponents are neutral by construction, not by measurement. All conclusions are for two dissipative chaotic flows of the same family (Lorenz-type polynomial vector fields with identical divergence), one encoder family (MLP), and one optimizer regime; the atlas is evidence, and the propositions are the theorem. A reference flow from a different dynamical class (e.g. a coupled-oscillator network with a conserved quantity)

would strengthen — or, if the traps failed to fire, weaken — the claim, and is the immediate next experiment.

IV. METHODS

A. Setting and reference dynamics

Let $M = \mathbb{R}^3$ carry the Lorenz-63 flow $\dot{x} = F_L(x)$, $F_L = (\sigma(y-x), x(\rho-z)-y, xy-\beta z)$ with $(\sigma, \rho, \beta) = (10, 28, 8/3)$. The reference invariants are the leading Lyapunov exponent $\lambda_1^{\text{ref}} \approx 0.9056$ per time unit and the constant divergence $\nabla \cdot F_L = -(\sigma + 1 + \beta) = -13.667$, which equals the sum of the full spectrum (the remaining exponents are $\lambda_2 = 0$ and $\lambda_3 \approx -14.57$). One Lyapunov time is $T_L = \lambda_1^{-1} \approx 1.13$ time units.

Observations are produced by a fixed nonlinear lift ϕ and an element-wise nonlinear warp W ,

$$\phi(x)_j = m_{j0} \sin(\omega_j \tilde{x}_j) + m_{j1} \tanh(\tilde{x}_j^3) + m_{j2} \tanh(e^{\tilde{x}_j} - 1), \quad \tilde{x} = (x - \mu)/s, \quad (5)$$

$$W(u) = u + 0.2 \sin(2u) \cos(u^2), \quad (6)$$

with fixed random weights (m_j, ω_j) and normalization chosen so that each lifted feature lies in $[-1, 1]$, on which W has positive derivative and is therefore a diffeomorphism of the bounded domain. The composite $W \circ \phi$ embeds the attractor into $\mathbb{R}^{64}$, because the lift is a diffeomorphism on the invariant manifold, the reference spectrum of the latent flow is exactly that of Lorenz augmented by 61 neutral directions (Lyapunov spectra are invariant under smooth conjugacies [6], [13]). A faithful latent model must therefore reproduce $\lambda_1 \approx 0.9$, zero

transverse exponents, and a volume-contracting flow with total contraction -13.667 .

For the second system, $M = \mathbb{R}^4$ carries the 4-D hyperchaotic Lorenz flow (2) with $(a, b, c, d) = (10, 28, 8/3, 2.0)$, whose divergence $-(a+1+c) = -13.667$ coincides with Lorenz-63 and whose Oseledets spectrum $\lambda \approx (+0.769, +0.098, 0, -14.59)$ (consistent-tangent Benettin QR, $dt = 0.002$, 60 time units, float64) carries two positive exponents. The identical lift ϕ and warp W map the 4-D attractor into $\mathbb{R}^{64}$ (with fixed normalization and lift parameters seeded separately), so the latent reference spectrum is the 4-D spectrum augmented by 60 neutral directions. The F arm’s spectral target is the runtime-measured $\lambda_1 = +0.737$ at the evaluation $dt = 0.01$, exactly as for Lorenz. Everything else — encoder, losses, curriculum, evaluation horizon, seeds — is bit-identical between the two systems; only the vector field, its normalization, and the reference invariants differ. This is the sense in which the two atlases test the constraint hierarchy rather than the dataset.

B. Encoder, training, and regularization

The encoder g_ξ is an MLP backbone (vector mode, input dimension 32, 256 hidden units, GELU, LayerNorm) followed by a projector onto the $d = 64$ latent space. Anti-collapse is handled by SIGReg [9]: for $P = 64$ random unit projections and $F = 16$ frequencies t , the empirical characteristic function of the projected batch is matched to the Gaussian characteristic function $\exp(-t^2/2)$, an Epps–Pulley-style test that penalizes collapse to a point or a subspace without any teacher or momentum encoder. The prediction target is the latent $g_\xi(x_{t+\tau})$ from a momentum-copied encoder; the dynamics operator is trained on the multi-step alignment loss

$$\mathcal{L}_{\text{align}} = \mathbb{E} \|f_\theta^{\circ k}(z_t) - z_{t+k}\|^2, \quad z_t = g_\xi(x_t), \quad (7)$$

over a progressive-horizon curriculum $k \in \{1, 3, 6\}$ (30%/60% of the training schedule). A motion-matching penalty with a detached target magnitude prevents the trivial “do-not-move” solution that minimizes alignment by freezing. Training: AdamW ($\text{lr} = 3 \times 10^{-4}$), gradient clipping at norm 1, batch 128, 250 steps per seed, three seeds, with fully deterministic per-step trajectory draws (per-step data seed), so every reported number is bit-reproducible from (seed, schedule).

C. The metriplectic tensor layout

Physics arms parameterize the latent field on the generalized metriplectic template (1): $J(z) = A(z) - A(z)^\top$ skew (reversible/Poisson part), $R(z) = L(z)L(z)^\top/n$ positive semidefinite (dissipation), and $H(z), S(z)$ scalar potentials realized as Softplus MLPs. Two structural facts anchor the analysis:

- For constant skew J , the reversible field is exactly divergence-free, $\nabla \cdot (J \nabla H) = \text{tr}(J \nabla^2 H) = 0$ (skew contracted with a symmetric Hessian), so all phase-space contraction comes from the dissipative part and is controlled by R alone.
- The dissipation matrix is PSD by construction ($R = LL^\top$), but its engagement is not: with L zero-initialized the loss

is stationary in L at $L = 0$, a dead saddle from which no first-order optimizer escapes and $R \equiv 0$ forever (Arm D).

The strict GENERIC degeneracy conditions $J \nabla S = 0$ and $R \nabla H = 0$ [2], [3] are deliberately not imposed: the target is an open system that must lose mechanical energy, and forcing $dH/dt = 0$ would re-break dissipation. The residual degeneracy violation is a documented caveat of every metriplectic arm. The macroscopic control is the *trace–divergence gate*,

$$\mathcal{L}_{\text{gate}} = \beta_g \left(\frac{1}{n} \text{tr} R - r^* \right)^2, \quad (8)$$

which, via the divergence identity, pins the spectral *sum*; it cannot pin any individual exponent (Proposition 1).

D. The five arms

All arms share the encoder, optimizer, curriculum, and data; only the structural constraint on f_θ differs.

- B Unconstrained MLP:** $f_\theta(z) = z + \text{MLP}(z)$, residual map, no structure, no physics loss (control arm).
- C Rigid Hamiltonian:** $f_\theta(z) = J_c \nabla H_c(z)$ with fixed constant skew J_c , quadratic H_c , and a flat energy-conservation penalty (a closed, volume-preserving prior — structurally incapable of contraction).
- D Naive metriplectic:** full (1) with learned J , $R = LL^\top$ zero-initialized, no gate. The dead saddle keeps $\text{tr}(R)/n$ at exactly zero.
- E Fixed metriplectic:** constant divergence-free J , nonzero L initialization, trace–divergence gate. The gate bites; the sum is pinned; λ_1 is not.
- F Spectral-aligned (short horizon):** Arm E plus the differentiable continuous-QR loss $\beta_\lambda (\hat{\lambda}_1 - \lambda_1^{\text{ref}})^2$ with a six-step window (0.06 time units ≈ 0.05 Lyapunov times).
- F-long Spectral-aligned (long horizon):** Arm E plus a 60-step differentiable QR loss (0.6 time units ≈ 0.5 Lyapunov times) with truncated BPTT (frame detached every 10 steps).

The physics arms use the dissipative energy-balance consistency penalty (energy decay matched to dissipated power), and Arm C uses the flat variant. The spectral weight is 0.10 (F) and 0.20 (F-long); the gate weight is 0.02.

E. Differentiable continuous QR (Benettin)

Let $A(t) = \partial_z f_\theta(z(t)) \in \mathbb{R}^{64 \times 64}$ be the tangent operator and X the fundamental matrix of $\delta \dot{z} = A(t) \delta z$, factorized $X = QR$ with $Q \in O(64)$ and R upper triangular, $R_{ii} > 0$. Substituting into $\dot{X} = AX$ and left-multiplying by $Q^\top$ gives $Q^\top \dot{A} Q = Q^\top \dot{Q} + \dot{R} R^{-1}$, a unique decomposition into skew and upper-triangular parts, hence the continuous-QR system

$$\dot{Q} = Q \Omega, \quad \Omega_{ij} = \begin{cases} (Q^\top \dot{A} Q)_{ij}, & i > j \\ 0, & i = j \\ -(Q^\top \dot{A} Q)_{ji}, & i < j \end{cases} \quad (9)$$

$$\frac{d}{dt} \log R_{ii} = (Q^\top \dot{A} Q)_{ii}.$$

The diagonal of the pulled-back operator drives the exponents; the strictly-lower part rotates the frame. The implementable discrete scheme (Benettin–Galgani–Giorgilli–Strelcyn

[5]) propagates an orthonormal frame Q_k , forms the Jacobian action $J_{\text{local}}Q_k$ by finite-difference JVP per column, advances $V = Q_k + dt J_{\text{local}}Q_k$, re-orthonormalizes $V = Q_{k+1}R_k$ by Householder QR, and accumulates

$$\hat{\lambda}_i(T) = \frac{1}{N dt} \sum_{k=1}^N \log |(R_k)_{ii}|, \quad T = N dt. \quad (10)$$

The chain is differentiable end to end (“torch.linalg.qr” backprops through the Householder reflections; the sign ambiguity of QR is fixed by $\log |\cdot|$ with a floor). For the long-horizon arm the frame is detached every $k = 10$ steps (truncated BPTT). We probed backpropagation through a 100-step QR chain under three regimes — full unroll, TBPTT-10, and per-step detach — on the initialization used here; the total gradient norms were 8.0×10^{-3} , 8.7×10^{-3} , and 8.9×10^{-3} respectively (all finite): on this initialization the failure is not gradient explosion but the proxy–truth gap documented in the Results. Two caveats are explicit: gradients are ill-conditioned near $R_{ii} \rightarrow 0$ or eigenvalue collisions, and — decisively — $\hat{\lambda}_i(T)$ is a *finite-time* exponent, not a bound on the asymptotic Oseledets spectrum (Proposition 2).

F. Evaluation metrics

All evaluation is on held-out initial conditions (fixed seed), with a 200-step autoregressive rollout (2 time units ≈ 1.8 Lyapunov times) in the learned latent space:

- *Boundedness/domain invariance*: $\max \|z\|$ over the rollout; escape from the training support ($\|z\| \gtrsim 10^2$ vs. in-support $\|z\| \approx 10$) is scored per seed.
- *Leading Lyapunov exponent λ_1* : shadow-trajectory Benettin on the one-step latent map, invariant under the coordinate warp, directly comparable to 0.884.
- *Contraction*: Hutchinson trace estimate $\text{tr}(J_{\text{map}} - I)/dt = \mathbb{E}_\varepsilon[\varepsilon^\top (J_{\text{map}} - I)\varepsilon]/dt$ (the divergence proxy, equal to the spectral sum) and its non-leading part $\kappa = \text{tr} - \lambda_1 = \sum_{i \geq 2} \lambda_i$.
- *Dissipation engagement*: $\text{tr}(R)/n$ averaged over the rollout (metriplectic arms).
- *Field-activity ratio m* : mean per-step displacement of the predicted rollout divided by the (detached) target displacement — a flatlining detector; $m \approx 1$ for a healthy field, $m \ll 1$ for a frozen one.
- *Attractor overlap*: symmetric Chamfer and Hausdorff distances in latent space between the predicted rollout cloud and the encoded ground-truth cloud.
- *Finite-horizon proxy*: the training-time $\hat{\lambda}_1$ read out at the end of training.

A full-spectrum (top-8) continuous-QR readout over the evaluation rollout cross-checks λ_1 and the non-leading sum.

G. Reproducibility

Every number in Tables I and II is generated by `benchmark.py` on CPU (PyTorch 2.10, Python 3.14): three seeds, 250 training steps per arm, batch 128, deterministic per-step data. The two systems are selected by `--system lorenz` and `--system lz4`; the protocol and

code path are otherwise identical. Training is bit-reproducible for a fixed thread count; a four-vs-eight-thread rerun of arm E shifted its exponent by $|\Delta \lambda_1| \lesssim 0.7$, an order of magnitude below the seed-to-seed spread reported in the tables, so the conclusions are insensitive to CPU float jitter. A fully independent retraining of arms E and F on Lorenz-63 — different wall-clock schedule and thread count — reproduced the atlas within seed-to-seed spread (E: $\lambda_1 = 2.18 \pm 1.86$ vs. 2.22 ± 1.80 ; F: 2.93 ± 1.85 vs. 2.85 ± 1.89 ; $\text{tr}(R)/n$ and Chamfer within 2%), and a third independent retraining of the frozen arm C reproduced its statistics (Chamfer 3.82–3.91 vs. 3.85 ± 0.05 ; $\lambda_1 \approx 0.008$ vs. 0.011 ± 0.002). The convergence check above reruns the shadow exponent on the saved weights, so the tables, the convergence paragraph, and the reproducibility numbers are drawn from independent artifacts of the same protocol.

a) *Convergence of the reported exponents.*: The leading exponents in Tables I–II are 200-step estimates (1.8 Lyapunov times). To certify that they are asymptotic rather than transient, the shadow-trajectory exponent was recomputed on the same final weights at 2000 steps (20 Lyapunov times) for every checkpointed arm. The frozen arm C is the only converged one: its exponents are ≈ 0 at both horizons on *both* systems ($|\lambda_1(2000)| \leq 0.018$ and $|\lambda_1(200)| \leq 0.018$, all rollouts finite to step 2000). Every alive arm’s 200-step value understates its 2000-step value by an order of magnitude. On Lorenz-63, E grows from $+0.8$ – $+4.6$ to $+8.8$ – $+9.4$ and F from $+1.7$ – $+5.5$ to $+9.8$ – $+10.0$, with the latent trajectory leaving the encoder domain at steps 213–300 of 2000 in all six seeds. On the hyperchaotic system the pattern fires with a longer calm phase: the gated arms read ≈ 0 at 200 steps but climb to $+7.8$ – $+10.1$ and escape at steps 304–407, and the dead-saddle arm D — whose 200-step value is the nearest to the reference — reads $+4.5$ – $+19.7$ at 2000 steps (bounded trajectory, strongly expanding tangent map). The reported 200-step values therefore *understate* the instability of every alive arm on both systems, and the atlas’s verdicts are conservative, not optimistic: any evaluation window shorter than the instability’s onset time can certify a pathological field.

The full five-arm run takes approximately 6.2 wall-clock hours on a 20-core CPU under load. The long-horizon arm is `benchmark_long.py` with a 60-step horizon and truncated BPTT; it is reported on the one completed seed (the five short-horizon arms are three seeds each). A second long-horizon run stalled without completing a training step (45 min without progress, CPU-bound, on identical code and hardware), which we attribute to the QR-backpropagation fragility documented in the Results — an additional, if anecdotal, instance of the same instability. The empirical reference $\lambda_1 = +0.8840$ is recovered by the same Benettin machinery on the true Lorenz flow.

DATA AVAILABILITY

All benchmark data (rollouts, checkpoints, evaluation JSONs, and convergence artifacts for both systems) are generated by the released code and are reproducible bit-for-bit from the stated seed schedules.

CODE AVAILABILITY

The complete benchmark (five arms, evaluation protocol, and the long-horizon variant) is available at <https://github.com/sehajr-singhs/spectral-topological-decoupling>. The five-arm atlas is reproduced by two commands: `python benchmark.py --seeds 0 1 2 --steps 250 (Lorenz-63)` and `python benchmark.py --system lz4 --seeds 0 1 2 --steps 250 (4-D hyperchaotic Lorenz)`; the long-horizon variant by `python benchmark_long.py` with `--seeds 0 --steps 120 --horizon 60`.

AUTHOR CONTRIBUTIONS

S.R.S. conceived the study, designed and implemented the benchmark suite, ran all experiments, performed the analysis, and wrote the manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ACKNOWLEDGMENTS

□

REFERENCES

- [1] E. N. Lorenz, Deterministic nonperiodic flow, *J. Atmos. Sci.* **20**, 130 (1963).
- [2] M. Grmela, H. C. Öttinger, Dynamics and thermodynamics of complex fluids. I. Development of a general formalism, *Phys. Rev. E* **56**, 6620 (1997).
- [3] H. C. Öttinger, M. Grmela, Dynamics and thermodynamics of complex fluids. II. Illustrations of a general formalism, *Phys. Rev. E* **56**, 6633 (1997).
- [4] P. J. Morrison, A paradigm for joined Hamiltonian and dissipative systems, *Physica D* **18**, 410 (1986).
- [5] G. Benettin, L. Galgani, A. Giorgilli, J.-M. Strelcyn, Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them, *Meccanica* **15**, 9 (1980).
- [6] V. I. Oseledets, A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems, *Trans. Moscow Math. Soc.* **19**, 197 (1968).
- [7] Y. LeCun, A path towards autonomous machine intelligence, preprint (2022).
- [8] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabinovich, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, *CVPR* (2023).
- [9] A. Bardes, J. Ponce, Y. LeCun, SIGReg: Sigmoidal Gaussian regularization for preventing collapse in joint-embedding predictive architectures, *preprint* (2024).
- [10] S. Greydanus, M. Dzamba, J. Yosinski, Hamiltonian neural networks, *NeurIPS* (2019).
- [11] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, S. Ho, Lagrangian neural networks, *ICLR* (2020).
- [12] S. A. van der Schaft, Port-Hamiltonian systems, *Encyclopedia of Systems and Control* (2020).
- [13] L. Barreira, Y. Pesin, Lyapunov exponents and smooth ergodic theory, *AMS Surveys and Monographs* **23** (2002).
- [14] G. Haller, Distinguished material surfaces and coherent structures in three-dimensional fluid flows, *Physica D* **149**, 248 (2001).
- [15] G. Qi, G. Chen, S. Du, Z. Chen, Z. Yuan, Analysis of a new chaotic system, *Physica A* **352**, 295 (2005). [4-D Lorenz-type hyperchaotic flow; the system used here is its standard (10, 28, 8/3, d) form.]
- [16] J. HaoChen, C. Wei, A. Gaidon, J. Ma, Provable guarantees for self-supervised deep learning with spectral contrastive loss, *NeurIPS* (2021).
- [17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow Twins: self-supervised learning via redundancy reduction, *ICML* (2021).
- [18] A. Bardes, J. Ponce, Y. LeCun, VICReg: variance-invariance-covariance regularization for self-supervised learning, *ICLR* (2022).
- [19] L. Jing, P. Vincent, Y. LeCun, Y. Tian, Understanding dimensional collapse in contrastive self-supervised learning, *ICLR* (2022).
- [20] Y. D. Zhong, B. Dey, A. K. Chakraborty, Symplectic ODE-net: learning Hamiltonian dynamics with control, *ICLR* (2020).
- [21] Y. D. Zhong, B. Dey, A. K. Chakraborty, Dissipative SymODEN: encoding Hamiltonian dynamics with dissipation and control into deep learning, *ICLR* (2021).
- [22] Q. Hernández, A. Badias, D. Gonzalez, F. Chinesta, E. Cueto, Deep learning of thermodynamics-aware reduced-order models from spatiotemporal data, *Comput. Methods Appl. Mech. Eng.* **379**, 113749 (2021).
- [23] Z. Zhang, S. Shin, M. Yacoub, J. C. Bringham, GENERIC formalism-based neural networks for learning dissipative and reversible dynamics, *J. Comput. Phys.* **443**, 110520 (2021).
- [24] J. Z. Kolter, G. Manek, Learning stable deep dynamics models, *NeurIPS* (2019).
- [25] R. Engelken, F. Wolf, L. F. Abbott, Lyapunov spectra of chaotic recurrent neural networks, *Phys. Rev. Lett.* **124**, 028301 (2020).
- [26] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, Neural ordinary differential equations, *NeurIPS* (2018).
- [27] D. Ha, J. Schmidhuber, World models, *arXiv:1803.10122* (2018).
- [28] D. Hafner, T. Lillicrap, M. Norouzi, J. Ba, Mastering Atari with discrete world models, *ICLR* (2021).
- [29] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Mathet, A. Beevi, F. Hasselt, M. Heess *et al.*, Genie: generative interactive environments, *ICML* (2024).
- [30] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, N. Ballas, Revisiting feature prediction for learning visual representations from video (V-JEPA), *Trans. Mach. Learn. Res.* (2024).
- [31] Q. Garrido, N. Ballas, A. Joulin, Y. LeCun, Learning and leveraging world models in visual representation learning, *arXiv:2403.00504* (2024).
- [32] T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, *ICML* (2020).
- [33] A. Ermolov, F. Siomos, N. Sebe, J. Kannala, Whitening for self-supervised representation learning, *ICML* (2021).
- [34] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, *ICCV* (2021).
- [35] Q. Garrido, R. Balestrierio, L. Najman, Y. LeCun, RankMe: assessing the downstream performance of pretrained self-supervised representations by their rank, *ICML* (2023).
- [36] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, *Nat. Rev. Phys.* **3**, 422–440 (2021).
- [37] S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proc. Natl Acad. Sci. USA* **113**, 3932–3937 (2016).
- [38] R. Iten, T. Metger, H. Wilming, L. del Rio, R. Renner, Discovering physical concepts with neural networks, *Phys. Rev. Lett.* **124**, 010508 (2020).
- [39] P. Toth, D. J. Rezende, A. Jaegle, S. Racanière, A. Botev, I. Higgins, Hamiltonian generative networks, *ICLR* (2020).
- [40] M. Lutter, C. Ritter, J. Peters, Deep Lagrangian networks: using physics as model prior for deep learning, *ICLR* (2019).
- [41] A. Sosanya, S. Greydanus, Dissipative Hamiltonian neural networks: learning from chaotic trajectories with limited data, *arXiv:2209.15832* (2022).
- [42] R. Vogt, M. Puelma Touzel, E. Shlizerman, R. Engelken, On Lyapunov exponents for RNNs: understanding information propagation using dynamical systems theory, *ICML* (2022).