REPRESENTATIONAL COLLAPSE AND SPECTRAL RECOVERY IN KOOPMAN-STRUCTURED TIME-SERIES JEPAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-embedding predictive architectures (JEPAs) are an effective framework for self-supervised learning on sequential data, but their latent representations remain difficult to interpret, particularly when a dynamical system generates the data. We study predictive self-supervised learning through Koopman operator theory and delay embeddings. Extending prior analyses of the eigenvalue-one case, we show that, under standard assumptions, training a linear latent predictor amounts to learning a finite-dimensional approximation to a Koopman-invariant subspace. Absent variance and covariance constraints on the latent space, both unconstrained and norm-preserving predictors converge toward degenerate representations rather than well-organized ones; this collapse, rather than the loss of oscillatory structure, is the dominant failure mode of predictive self-supervised learning on dynamical data. Standard variance-covariance regularization suffices to prevent it, and a Koopman-structured predictor adds interpretability: its learned rotation angles can be compared directly against the spectral content of the underlying system. We relate context length, latent dimension, and prediction horizon to observability, spectral capacity, and predictability. On synthetic dynamical systems, the learned representations recover the true oscillation frequencies, including combination frequencies not explicitly targeted by the model, while exhibiting the collapse and capacity trade-offs predicted by our analysis.

1 INTRODUCTION

In sequential settings, self-supervised learning (SSL) with Joint-Embedding Predictive Architectures (JEPAs) is a natural fit: an encoder maps past context to a latent state, and a predictor maps that state to a future latent target. For time series from physical processes, however, each observation is a noisy, partial, and often nonlinear view of an underlying trajectory, so it is unclear what structure such a model can recover and when its latent space preserves the geometry of the system. We study this question through Koopman operator theory: a JEPA with a linear latent predictor can be viewed as learning a finite-dimensional approximation to a Koopman-invariant subspace, in which the encoder learns observables that render the dynamics approximately linear and the predictor approximates the Koopman action on those coordinates.

This perspective also exposes a common failure mode of standard predictive objectives. We show that, without explicit variance and covariance constraints on the latent space, both unconstrained and norm-preserving predictors converge toward degenerate representations instead of dynamically meaningful ones. Preventing this collapse only requires standard variance-covariance regularization. What it does not provide is a way to verify that the latent dynamics correspond to the true dynamics of the underlying system, so we combine it with a Koopman-structured predictive architecture whose predictor is constrained to a block-diagonal orthogonal form and whose learned rotation angles can be compared directly against the spectral content of the system.

We analyze the resulting model along three axes, namely the temporal context needed for faithful state recovery, the latent dimension required to represent predictable modes, and the effect of the prediction horizon on the balance between predictable structure and mixing components, on synthetic systems from equilibria to chaotic attractors. The main contributions of this work are:

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 Koopman-based interpretation of JEPA-style predictive learning for dynamical data, extending prior analyses of the eigenvalue-one case to general invariant subspaces.
- A characterization of representational collapse in predictive self-supervised learning on dynamical data, and a demonstration that standard variance-covariance regularization is sufficient to prevent it regardless of predictor structure.
- Direct Koopman-JEPA, a Koopman-structured predictor whose learned angles can be compared directly against the spectral content of the underlying system, recovering true oscillation frequencies, including those arising from interactions between independent modes.
- An empirical study of how temporal context, latent dimension, and prediction horizon shape the learned representation.

2 Related Work

Predictive self-supervised learning and JEPAs. Self-supervised learning has shifted from invariant feature learning (Chen et al., 2020; Grill et al., 2020) to predictive objectives for sequential data. Joint-embedding predictive architectures (LeCun, 2022) pair an encoder with a predictor acting in latent space and have been instantiated in vision (Assran et al., 2023), in speech and multi-modal settings (Baevski et al., 2022), and for in-context time-series forecasting (Verdenius et al., 2024), in contrast to masked reconstruction in input space (Feichtenhofer et al., 2022). Existing analyses largely use information-theoretic criteria (Tschannen et al., 2020), and how learned coordinates relate to the underlying dynamics remains poorly understood.

Self-supervised learning for time series. Time-series SSL has been dominated by contrastive and reconstruction-based objectives, including instance- and temporal-level contrasting (Yue et al., 2022; Tonekaboni et al., 2021; Eldele et al., 2021), time-frequency consistency (Zhang et al., 2022), and masked-patch reconstruction (Nie et al., 2023); JEPA-style prediction has seen little uptake. Contrastive invariances rely on augmentations such as cropping, jittering, and time-warping, which do not commute with the underlying flow and so alter phase relationships between coordinates, whereas a JEPA predicts future latent states under the system's own evolution, preserving the correspondence between latent transitions and the generating dynamics (Section 3).

Collapse and regularization. Informational and dimensional collapse, where representations become trivial or low-rank (Jing et al., 2022), has motivated predictor asymmetry and variance/covariance regularization (Jing et al., 2022; van den Oord et al., 2019; Bardes et al., 2022; Zbontar et al., 2021). Such regularizers are effective, but they typically reflect batch-statistical properties rather than dynamical geometry, which raises the question of whether isotropic penalties are as costly for phase relationships and low-dimensional rotational modes as is commonly assumed, or whether generic regularization suffices once collapse is addressed.

Koopman learning and delay embedding. The Koopman operator (Mezić, 2005) underlies dynamic mode decomposition and its variants and deep Koopman methods (Lusch et al., 2018; Williams et al., 2015; Takeishi et al., 2017; Otto & Rowley, 2019), which seek embeddings with approximately linear dynamics; closest to our work, Ruiz-Morales et al. (2026) show that the loss of a JEPA with a linear predictor is minimized when the encoder represents the indicators of the ergodic components of the data (Koopman eigenfunctions at eigenvalue one) under an identity predictor, leaving oscillatory eigenfunctions to future work; we treat general invariant subspaces and rotational predictors, and add collapse, delay embedding, capacity, and horizon. Takens' theorem and its fractal extensions (Takens, 1981; Sauer et al., 1991), which underlie work on chaotic forecasting (Pathak et al., 2018; Gilpin, 2023), characterize when an observation window reconstructs the geometry of the attractor, and thus how much temporal context is needed to recover state information from partial observations.

3 Predictive Learning and Operator-Theoretic Perspective

Through the lens of ergodic theory, optimizing a linear latent predictor is an operator-theoretic system identification problem, and successful representation learning requires a context window that

resolves the topological manifold, sufficient algebraic rank in the latent space, and a prediction horizon that filters predictable modes from chaotic spectra.

3.1 Dynamical Setup and Predictive Objective

Assumption 3.1 (Underlying Dynamics)**.** *Let $(\mathcal{M}, g)$ be a compact m-dimensional Riemannian manifold, and let $T : \mathcal{M} \to \mathcal{M}$ be a C^r diffeomorphism ($r \geq 2$) generating the discrete-time evolution $u_{t+1} = T(u_t)$. We assume T preserves an ergodic invariant probability measure μ, allowing spatial expectations in $L^2(\mathcal{M}, \mu)$ to be equated with infinite-horizon empirical averages.*

Assumption 3.2 (Observation and JEPA Models)**.** *The system is observed through a generic scalar measurement function $\psi : \mathcal{M} \to \mathbb{R}$, yielding a time series $s_t = \psi(u_t)$. The JEPA processes a context window of length n_c, denoted $x_t = (s_t, \ldots, s_{t+n_c-1}) \in \mathbb{R}^{n_c}$. A continuous non-linear encoder $f_\theta : \mathbb{R}^{n_c} \to \mathbb{R}^k$ maps the window to a latent representation z_t. A linear Markovian predictor $M \in \mathbb{R}^{k \times k}$ is trained to approximate a target representation Δ steps into the future, optimized via the Mean Squared Error (MSE) loss:*

$$\mathcal{L}_{MSE} = \mathbb{E}_{x \sim \mathcal{X}} \left[\| M f_\theta(x_t) - f_{\bar{\theta}}(x_{t+\Delta}) \|_2^2 \right] \tag{1}$$

where $f_{\bar{\theta}}$ is a target encoder (often an Exponential Moving Average (EMA) of f_θ).

3.2 Predictive Learning as Approximate Koopman Subspace Identification

The context window x_t is strictly the evaluation of a delay-coordinate map $\Phi_{n_c} : \mathcal{M} \to \mathbb{R}^{n_c}$ at the true state u_t, defined as $\Phi_{n_c}(u) = [\psi(u), \psi(T(u)), \ldots, \psi(T^{n_c-1}(u))]^T$. Consequently, the learned latent representation is a vector-valued observable on the physical manifold: $F := f_\theta \circ \Phi_{n_c}$, where $F : \mathcal{M} \to \mathbb{R}^k$. Let $\mathcal{K}^\Delta : L^2(\mathcal{M}, \mu) \to L^2(\mathcal{M}, \mu)$ denote the Δ-step Koopman operator, defined by its composition with the flow: $(\mathcal{K}^\Delta h)(u) = h(T^\Delta(u))$.

Theorem 3.3 (Characterization of Zero Predictive Error)**.** *Assume an ideal target network ($f_{\bar{\theta}} = f_\theta$) and that the scalar components $\{h_j\}_{j=1}^k$ of F are linearly independent in $L^2(\mathcal{M}, \mu)$. Then the predictive loss $\mathcal{L}_{MSE}(F, M) = 0$ if and only if $\operatorname{span}\{h_1, \ldots, h_k\}$ is a k-dimensional invariant subspace of the Koopman operator $\mathcal{K}^\Delta$ in $L^2(\mathcal{M}, \mu)$, and M is the matrix representation of $\mathcal{K}^\Delta$ restricted to that basis.*

Proof sketch. Since $z_{t+\Delta} = (\mathcal{K}^\Delta F)(u_t)$, the loss equals $\mathbb{E}_\mu[\|MF - \mathcal{K}^\Delta F\|_2^2]$ and vanishes exactly when $\mathcal{K}^\Delta h_j \in \operatorname{span}\{h_1, \ldots, h_k\}$ for every j. Full proof in Appendix A.1. □

The condition $MF = \mathcal{K}^\Delta F$ is that of Ruiz-Morales et al. (2026), who treat the eigenvalue-one case with an identity predictor; we state it for deterministic dynamics and general M as an invariant-subspace characterization. The linear-independence hypothesis restricts which solutions the theorem describes; it is not a property equation 1 enforces on its own. The degenerate case $F \equiv \text{const}$, an eigenvalue-one observable in their sense, trivially attains zero loss with $M = I$, and nothing in equation 1 rules it out. Reaching a non-degenerate solution requires the regularizer of Section 4.3. Consequently, a linear predictor on a non-linear latent representation under an identical target network is equivalent to Deep Extended Dynamic Mode Decomposition (Deep EDMD), with the encoder acting as a dictionary learner.

3.3 Temporal Context and Observability

For the encoder to discover this invariant subspace, the context window n_c must contain sufficient geometric information to unroll the underlying manifold $\mathcal{M}$.

Definition 3.4 (Faithful Representation)**.** Let $A \subset \mathcal{M}$ denote the attractor supporting the invariant measure μ. Following Mezić (2021), specialized from the full manifold $\mathcal{M}$ to A because only states visited by the dynamics need to be distinguished, a representation $F : \mathcal{M} \to \mathbb{R}^k$ is faithful on A if it is injective there: $F(u_a) = F(u_b)$ implies $u_a = u_b$ for all $u_a, u_b \in A$. A representation that fails to be injective on A is topologically reduced (entangled) there.

Theorem 3.5 (Context for State Reconstruction)**.** *Injectivity of the delay map Φ_{n_c} on A is precisely Definition 3.4's faithfulness condition applied to Φ_{n_c} itself. Let m_A be A's topological dimension*

and $d_A \geq m_A$ its box-counting dimension. If $n_c < m_A$, Φ_{n_c} is not injective on A for any observation function ψ, and the predictive loss cannot vanish unless the target encoder merges the Δ-futures of every aliased pair, as a constant encoder does. If $m_A \leq n_c \leq 2d_A$, injectivity is neither guaranteed nor generically excluded. If $n_c > 2d_A$, Φ_{n_c} is injective on A for a generic ψ, and this obstruction to zero predictive loss is generically absent.

Proof sketch. Below m_A, dimension theory (Hurewicz & Wallman, 1941) rules out injectivity for every ψ, and, by continuity and the full support of μ, aliased states whose target futures are separated force an error bounded away from zero; above $2d_A$, the Fractal Delay Embedding Theorem (Sauer et al., 1991) gives generic injectivity. Full proof in Appendix A.2. □

This bound is an observability constraint on the input, not a limitation of optimization or of the latent dimension.

3.4 Latent Capacity and Spectral Representation

Even when the input window is observably identifiable, the latent dimension k may still be too small to represent the spectral content of interest.

Definition 3.6 (Target Spectral Components)**.** Let the predictable dynamics of the system be governed by a finite subset of the Koopman point spectrum, spanning a subspace $\mathcal{V}_{point} \subset L^2(\mathcal{M}, \mu)$. Under Assumption 3.1, μ is ergodic, so the only Koopman eigenfunctions with eigenvalue $\lambda = 1$ are the constants: $r = 1$, represented by the trivial indicator function $\chi_1 \equiv 1$. The targeted modes therefore consist of this single stationary component together with c independent limit cycles, associated with complex conjugate pairs $\lambda = e^{\pm i\omega}$, each corresponding to a two-dimensional real rotational subspace. The associated real dimension is $d_{spec} = 1 + 2c$.

Remark 3.7 (Mixtures of Ergodic Regimes)**.** Definition 3.6 is stated for a single ergodic system, where Assumption 3.1 forces $r = 1$. A mixture of several ergodic systems, such as several regimes pooled into one dataset, decomposes into as many ergodic components as there are regimes, each contributing its own indicator eigenfunction at $\lambda = 1$ (Ruiz-Morales et al., 2026), so a representation trained on it is expected to devote part of its capacity at eigenvalue 1 to distinguishing components rather than encoding within-component dynamics (Section 5.3).

Theorem 3.8 (Capacity Requirement for Spectral Representation)**.** *A necessary condition for exact representation of the targeted predictable subspace is $k \geq d_{spec}$. If $k < d_{spec}$, the latent space cannot contain a basis for the target Koopman subspace, and some modes must be merged or projected into shared directions.*

Proof sketch. The encoder's coordinate span has dimension at most k, so if $k < d_{spec}$ it cannot contain a d_{spec}-dimensional invariant subspace, and distinct invariant components are entangled within shared latent coordinates. Full proof in Appendix A.3. □

Remark 3.9 (Combination Frequencies and Countable Point Spectrum)**.** A torus with $r \geq 2$ incommensurate base frequencies $\omega_1, \ldots, \omega_r$ has a Koopman point spectrum containing every integer combination $n \cdot \omega = \sum_i n_i \omega_i$ for $n \in \mathbb{Z}^r$, each with its own two-dimensional rotational eigenspace, a countably infinite set; for example, $\omega_1 - \omega_2$, $\omega_1 + \omega_2$, and $2\omega_2$ are distinct from ω_1 and ω_2. No finite k can satisfy $k \geq d_{spec}$ for the full point spectrum, so Theorem 3.8, which concerns a target subspace $\mathcal{V}_{point}$ fixed in advance, is only ever a statement about a deliberately truncated target, not an intrinsic capacity requirement of the underlying dynamics.

3.5 Prediction Horizon and Spectral Bias

The prediction horizon determines which dynamical components dominate the learning signal: long horizons emphasize coherent, predictable structure and suppress rapidly decorrelating components.

Assumption 3.10 (Mixing continuous spectrum)**.** *Assume that the observable F admits an orthogonal decomposition $F = F_{point} + F_{cont}$, where F_{point} lies in the point-spectrum subspace and F_{cont} lies in the continuous-spectrum subspace. Assume further that the continuous component is strongly mixing, so that its temporal correlations decay asymptotically.*

Theorem 3.11 (Prediction Horizon and Representation of Mixing Components)**.** *Under Assumption 3.10, and assuming further that the latent coordinates split disjointly between the point and continuous components, meaning that each scalar coordinate h_j lies entirely in $\mathcal{H}_{point}$ or entirely in $\mathcal{H}_{cont}$, the best linear predictor for the continuous component becomes ineffective as $\Delta \to \infty$. More precisely, the optimal linear map restricted to F_{cont} converges to zero in the long-horizon limit, and the minimum achievable loss on that component approaches $||F_{cont}||^2_{L^2}$.*

Proof sketch. Correlations of F_{cont} vanish as the horizon grows, so the optimal linear estimator converges to zero and the remaining error is the variance of F_{cont}. Full proof, including the case without disjoint coordinates, in Appendix A.4. □

The theory suggests two empirical checks involving the horizon. By Theorem 3.3, M represents $\mathcal{K}^{\Delta}$, so a point-spectrum eigenfrequency ω appears in a structured predictor as a rotation by $\omega\,\Delta t\,\Delta$ (Δt the sampling interval), and learned angles should move with Δ as these analytic targets do (Section 5.2). The second, that continuous-spectrum content becomes harder to fit as Δ grows, is not tested directly in this work.

4 Structured Predictive Model

Building on the bounds of Section 3, we describe a Koopman-structured predictive architecture with two components: a block-diagonal orthogonal predictor that exposes static and rotational subspaces directly, and a variance/covariance regularizer that prevents the representational collapse to which unconstrained training is otherwise prone.

4.1 Representational Collapse in Unconstrained Predictors

Let $x_t \in \mathbb{R}^{n_c}$ denote a context window and $z_t = f_\theta(x_t) \in \mathbb{R}^k$ the corresponding latent representation. Given a target encoder $f_{\bar{\theta}}$ and a linear predictor $M \in \mathbb{R}^{k\times k}$, the standard JEPA objective is

$$\mathcal{L}_{MSE} = \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}_{x\sim\mathcal{X}}\left[\left|\left|M^h f_\theta(x_t) - f_{\bar{\theta}}(x_{t+h\Delta})\right|\right|_2^2\right] \tag{2}$$

where H is the rollout horizon and Δ is the prediction step.

In unconstrained form, however, minimizing the loss alone does not exclude the degenerate solution identified after Theorem 3.3: an encoder that maps every context window to approximately the same latent vector trivially satisfies the objective for any M that fixes that vector, including $M = I$. In practice, training an unconstrained linear predictor without additional regularization drives the latent representation toward this solution, a cloud that is narrow relative to its distance from the origin, instead of organizing itself around the system's geometry. Constraining the predictor's spectrum does not by itself prevent this (Sections 4.2 and 5.1); preventing collapse is the role of the regularizer of Section 4.3, not of the predictor's architecture.

4.2 Structured Predictors for Interpretable Latent Dynamics

To encode the assumption that some components of the dynamics are approximately measure-preserving or rotational, we parameterize the predictor using a block-diagonal orthogonal structure. Let

$$M_{dir} = \text{block_diag}\left(I_s,\ R(\omega_1), \ldots, R(\omega_c)\right), \tag{3}$$

where I_s is an identity matrix governing s static dimensions, and $R(\omega_i) \in SO(2)$ are 2×2 planar rotation matrices. This parameterization constrains the predictor to have eigenvalues on the unit circle for the rotational blocks, thereby ruling out contraction or expansion in those subspaces. In practice, the rotation angles $\{\omega_i\}_{i=1}^{c}$ are learned jointly with the encoder.

The split between static and rotational capacity follows a fixed rule: at least half of the latent budget is assigned to the identity block I_s, and the remaining coordinates are divided into as many 2×2 rotation blocks as fit, with any leftover coordinate returned to the identity block; for $k = 14$ this gives $s = 8$ and $c = 3$.

The motivation for this construction is interpretability: because each rotation angle ω_i occupies a fixed pair of coordinates, it can be read off directly and compared against the system's analytic Koopman eigenfrequencies, which an arbitrary dense M does not permit. Orthogonality is not needed to avoid collapse: an unconstrained predictor trained under the same regularizer does not collapse, and in most seeds its eigenvalues settle close to the unit circle without being constrained to (Section 5.2), although its matrix is not orthogonal. What the block-diagonal structure adds is a verifiable, per-coordinate readout of the learned dynamics, at the price of a strong prior.

4.3 Variance/Covariance Regularization

Preventing the collapse of Section 4.1 requires only that latent coordinates be kept from collapsing to a constant value and from becoming redundant with one another.

Let $Z \in \mathbb{R}^{B\times k}$ denote a minibatch of latent vectors, with batch size B, and let $C(Z) \in \mathbb{R}^{k\times k}$ be its empirical covariance matrix. The regularizer takes the variance-hinge and covariance-penalty form of VICReg (Bardes et al., 2022):

$$\mathcal{L}_{reg} = \lambda \frac{1}{k} \sum_{j=1}^{k} \max(0, 1 - \text{std}(Z_{\cdot,j})) + \mu \frac{1}{k} \sum_{i\neq j} C(Z)_{i,j}^2, \tag{4}$$

where $\lambda, \mu > 0$ are weighting coefficients. The first term prevents latent dimensions from collapsing to constant values; the second penalizes covariance between every pair of distinct coordinates.

Proposition 4.1 (Uncorrelated Rotational Coordinates)**.** Let $(\cos\theta, \sin\theta)$ be the coordinates of a rotational block whose phase θ is uniformly distributed on $[0, 2\pi)$, as holds for an ergodic limit cycle or quasi-periodic torus under its invariant measure. Then $\text{Cov}(\cos\theta, \sin\theta) = 0$.

Proof. $\mathbb{E}[\cos\theta \sin\theta] = \frac{1}{2}\mathbb{E}[\sin 2\theta] = \frac{1}{4\pi}\int_0^{2\pi} \sin(2\theta)\, d\theta = 0$, and $\mathbb{E}[\cos\theta] = \mathbb{E}[\sin\theta] = 0$ by the same integral over a full period, so $\text{Cov}(\cos\theta, \sin\theta) = \mathbb{E}[\cos\theta \sin\theta] - \mathbb{E}[\cos\theta]\mathbb{E}[\sin\theta] = 0$. □

5 Experiments

We construct a benchmark of five continuous-time regimes: an Ornstein–Uhlenbeck process (stochastic equilibrium), a Stuart–Landau oscillator (noisy limit cycle), a quasi-periodic torus, and the chaotic Lorenz-63 and Rössler attractors. The torus, Lorenz-63, and Rössler are deterministic and instantiate Assumption 3.1 directly; the two stochastic regimes are simulated with injected noise and serve as stress tests of the theory beyond its literal scope. Multivariate states are projected to scalar sequences by a generic linear map, so latent structure must be inferred strictly from delay windows. We refer readers to Appendix B for further details.

Each subsection tests specific predictions of the preceding sections, stated as the results are discussed; numerical values not shown in the figures are collected in Appendix D.

5.1 Representational Collapse Without Regularization

Theorem 3.3 and the remark following it show that equation 1 has a degenerate global minimum, an encoder that outputs a constant, which any predictor containing the identity can accommodate, including the orthogonal predictor of equation 3. Collapse should therefore appear whatever the predictor's structure, and extra latent capacity should not remove it. We test this at the largest context and latent dimension of the grid, so that neither can be the limiting factor, training both predictors without the regularizer of Section 4.3 and measuring the collapse ratio $\|\text{mean}(z)\|/\text{mean}\|z\|$, near 1 for a latent cloud far from the origin relative to its width, together with the spread of the cloud itself, $\sqrt{\text{tr}\,\text{Cov}(z)}$, which does not depend on where it sits.

Figure 1 shows that both predictors collapse, as predicted, though to different degrees: Direct Koopman-JEPA collapses completely and identically across seeds, whereas the unconstrained predictor collapses strongly but less completely, with visible seed-to-seed variation (Table 1). The spread shows the same contrast, and the regularizer restores it. Neither parametrization protects

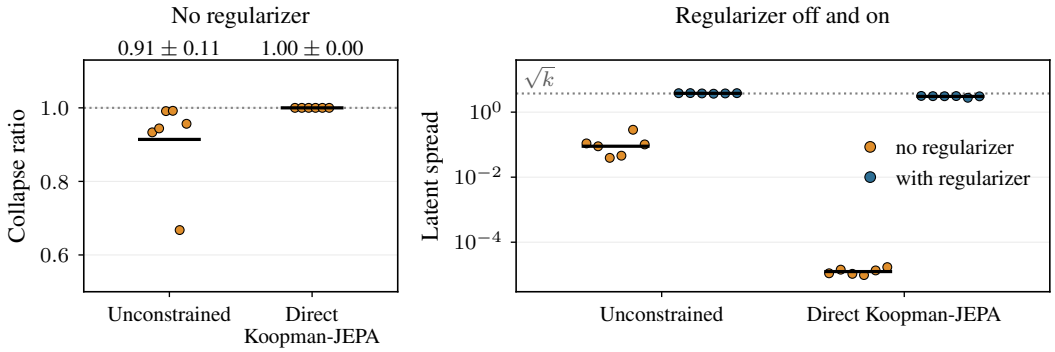


Figure 1: Representational collapse at $n_c = 15$, $k = 14$, one point per seed (bar: mean). Left: collapse ratio $\|\mathrm{mean}(z)\|/\mathrm{mean}\|z\|$ under equation 2 alone. Right: latent spread $\sqrt{\mathrm{tr}\,\mathrm{Cov}(z)}$ without and with the regularizer of Section 4.3 (dotted: $\sqrt{k}$).

against the degenerate minimum: constraining the spectrum to the unit circle does not prevent collapse, as anticipated in Section 4.2, and the failure is not one of capacity, since both models sit at the largest k of the grid. All remaining experiments use the regularizer of Section 4.3.

5.2 Predictor Ablation and Frequency Recovery

With collapse handled by the regularizer, we ask what the block-diagonal predictor contributes. Section 4.2 argued that it is not needed to keep the representation from degenerating and that what it adds is a readout; we test both halves.

For the ablation, we compare the unconstrained predictor under the regularizer with Direct Koopman-JEPA at the largest cell of the grid ($n_c = 15$, $k = 14$; Figure 3). It does not collapse, and its learned matrix settles with eigenvalues at or near the unit circle in most seeds without being constrained to (Tables 2 and 3); its separability is close to that of Direct Koopman-JEPA and its predictive error nearly three times higher, and both vary more across seeds. The block-diagonal structure is therefore not what prevents collapse, in line with the sufficiency of the regularizer (Section 4.3).

For the readout, each rotation angle ω_i is a single learned scalar. By Theorem 3.3, M represents $\mathcal{K}^{\Delta}$ on the learned subspace, so a block with angle ω_i corresponds to a Koopman eigenvalue $e^{i\omega_i}$ and hence to a system frequency ω with $\omega_i = \omega\,\Delta t\,\Delta$ modulo 2π, which predicts where each analytic frequency should sit and how it should move with Δ. Figure 2 compares the learned angles with these targets across a Δ-sensitivity sweep (Table 4 gives the counts): the two base frequencies of the Stuart-Landau and torus regimes and the three combination frequencies that Remark 3.9 predicts for the torus.

The learned angles do sit on analytic targets and move with Δ as the targets do: the base frequency ω_1 is recovered at every horizon in most seeds, and at the two longer horizons most angles fall almost exactly on a target. Some blocks also land on the combination frequencies $\omega_1 \pm \omega_2$ and $2\omega_2$, which are eigenfrequencies of the flow rather than parameters of the generating equations, so the network recovers structure of the dynamics rather than simulation constants. Recovery is nevertheless partial, as the theory anticipates: the torus has a countably infinite point spectrum (Remark 3.9) while $k = 14$ provides three rotation blocks (Section 4.2), so which modes a run selects depends on the seed. The shortest horizon is the weakest case: torus frequencies other than ω_1 are mostly missed, and several blocks lock onto a frequency outside the target set that matches the dominant spectral peak of the Rössler regime's observed signal (gray band in Figure 2), which we read as a genuine frequency of the data that the target set omits rather than a failure of the model.

5.3 Effects of Context Length and Latent Capacity

We now vary the context window n_c and the latent dimension k for both predictors under the regularizer, aggregated over 6 seeds (Figure 3), reporting k-NN accuracy for predicting the regime from

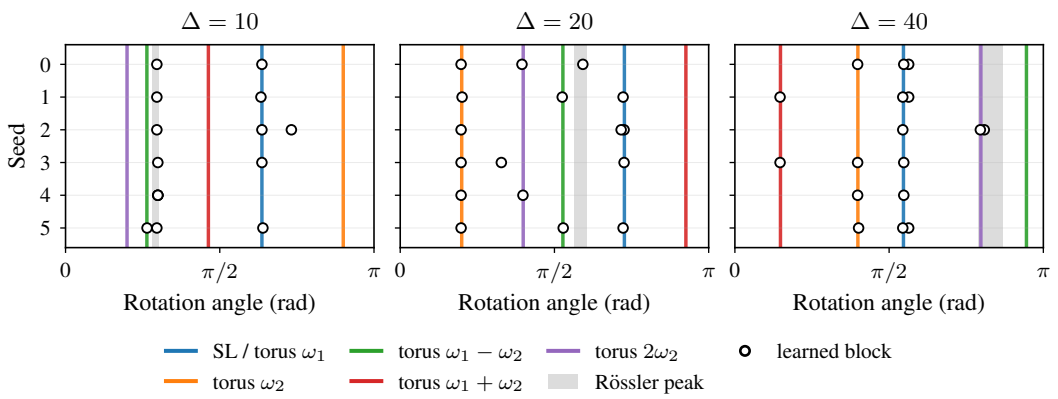


Figure 2: Rotation angles learned by Direct Koopman-JEPA, folded into $[0, \pi]$ (circles, one row per seed), against the analytic Koopman eigenfrequencies at each horizon Δ (vertical lines); a circle on a line is a recovered target. The gray band is the angle implied by the dominant spectral peak of the Rössler regime, which is not in the target set, with its estimation uncertainty.

the latent vector (regime separability), the false-nearest-neighbor rate (FNN) (Kennel et al., 1992) (topological entanglement), and scale-normalized predictive MSE.

Context length and observability. Theorem 3.5 predicts an obstruction below the topological dimension of the attractor, no guarantee in between, and generic injectivity above $2d_A$. At the shortest windows both predictors show the entanglement expected from pure aliasing, at the same level for every k, with a large error. For Direct Koopman-JEPA, separability and error then improve steeply and flatten near the largest embedding requirement in the benchmark, set by the Rössler regime ($n_c \gtrsim 6$, Appendix B) rather than by Lorenz. That the transition sits where delay-embedding theory places it supports reading short-window errors as an observability limit of the input rather than an optimization failure. The unconstrained predictor makes the transition later, its separability and error still improving until $n_c \approx 9$ (Figure 3); since observability is a property of the data, the extra context this arm needs points to an optimization cost rather than a different observability limit.

Comparison of the predictors. The bottom row of Figure 3 pairs the predictors seed by seed. Except at $n_c = 1$, where the unconstrained predictor has the lower error, Direct Koopman-JEPA has the higher k-NN accuracy and the lower error in 35 of the 42 cells with $k \geq 4$ and $n_c \geq 3$ (at least 5 of 6 seeds agreeing). For $k \geq 4$ and $n_c \geq 5$ the gain is 0.05 to 0.08 in accuracy and a factor of 1.6 to 2.1 in error, with the same sign in every seed. The k-NN gain is small and qualified by the identity-block analysis below. FNN does not separate the predictors: its difference changes sign along n_c and comes from the Ornstein-Uhlenbeck and Lorenz regimes (Appendix D.6).

The error floor and the continuous spectrum. Once the window resolves the topology, the error settles on a floor rather than vanishing. Theorem 3.11 identifies a source: the continuous-spectrum component of the chaotic regimes decorrelates from its own past, so a linear predictor cannot recover it, and the injected noise of the two stochastic regimes adds an unpredictable component. We read the floor in the MSE panel as this residual, consistent with the theory though not isolated by it, since the error pools all five regimes.

Latent capacity. Theorem 3.8 requires $k \geq d_{spec} = 1 + 2c$, so the number of rotation blocks c allocated by the split rule of Section 4.2 bounds what Direct Koopman-JEPA can express. At $k = 2$ the rule gives $c = 0$, so no rotational mode can be represented, and the dashed lines show a deficit in separability at moderate windows for Direct Koopman-JEPA, and in FNN for both predictors (more than twice its value at $k = 14$ from $n_c = 7$ on), so the latter is not specific to the missing rotation block. From $k = 4$ on ($c \geq 1$) performance is nearly flat (Table 6), in line with Remark 3.9: dimensions beyond those the model uses cannot help. The unconstrained predictor's k-NN deficit at $k = 2$ is larger and persists, with far larger variation across seeds, indicating an optimization outcome rather than an architectural one.

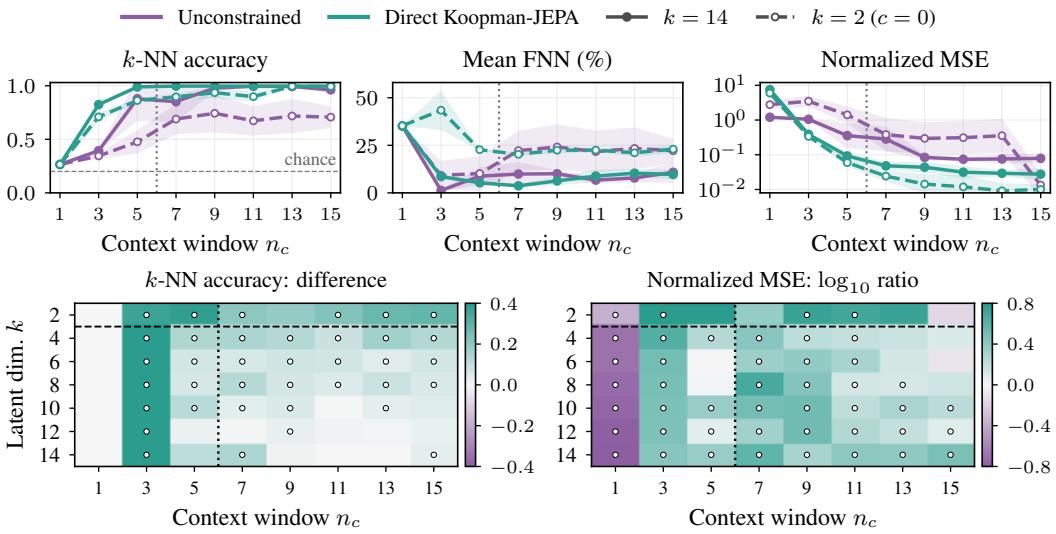


Figure 3: Effect of n_c and k (6 seeds). Top: both predictors at $k = 14$ and $k = 2$ (mean and one standard deviation). Bottom: difference between the predictors, paired by seed (teal: Direct Koopman-JEPA better); dots: at least 5 of 6 seeds agree, difference not negligible (Appendix D). Dotted lines: $n_c = 6$; dashed: $k = 2$ vs. $k \geq 4$.

Mixtures of ergodic regimes. Remark 3.7 predicts that, because the benchmark pools several ergodic systems, part of the capacity at eigenvalue one is spent distinguishing regimes rather than encoding within-regime dynamics. Restricting the analysis of Direct Koopman-JEPA at the largest cell to each block of the latent space (Table 7), the identity block alone separates the regimes essentially as well as the full representation, and almost all of its variance lies between regimes: it behaves as a discrete regime label. The rotation blocks carry within-regime structure and separate regimes less well. This confirms the prediction and qualifies the k-NN panel of Figure 3, which measures regime separability at least as much as recovery of within-regime dynamics; we rely on the frequency analysis for the latter. This agrees with the regime clustering that Ruiz-Morales et al. (2026) explain through indicator eigenfunctions.

6 Conclusion

We studied predictive self-supervised learning for dynamical data through Koopman operator theory and delay embedding. A linear predictor in latent space admits a natural interpretation as a finite-dimensional approximation to a Koopman-invariant subspace, conditional on the encoder's coordinates remaining linearly independent.

Direct Koopman-JEPA was not necessary to prevent collapse, since a regularized unconstrained predictor did not collapse, though it gave a consistent gain in k-NN accuracy and predictive error at all but the shortest windows (Figure 3). Its main addition was verifiability: its learned rotation angles could be compared directly against the system's analytic Koopman eigenfrequencies, including combination frequencies, and followed these targets as they moved with Δ, except that at the shortest horizon several blocks instead matched a frequency of the Rössler regime that the target set omits. Among the limitations, which frequencies a given run recovers depends on the seed, and the identity block's role as a regime classifier is an artifact of pooling ergodic systems (Appendix E discusses these and further limitations).

AI use statement

In this work, we used generative AI tools for generating synthetic datasets consisting of simulated trajectories from different dynamical systems.

We have not used generative AI tools for designing or conducting research methodology or experiments, and the other required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for editing software code used to generate the synthetic dataset and figures, editing figure-generation scripts, and condensing and improving the readability of portions of the paper text. We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested for correctness by the authors, and the generated simulations and figures were checked by the authors. The AI-assisted text was reviewed and revised by the authors for accuracy and consistency with the underlying work. We take responsibility for the final content of this work, including text, claims, code, figures, datasets, and other artifacts produced with the aid of generative AI.

Reproducibility statement

All theoretical results are stated with their assumptions in Section 3 and proved in full in Appendix A. The synthetic benchmark is specified completely in Appendix B (system equations and parameters, sampling interval, burn-in, observation map, and normalization), and the encoder, target network, predictor parameterizations, optimizer, regularizer weights, and evaluation protocols (trajectory-level train/test splits, k-NN accuracy, false-nearest-neighbor rate, and normalized MSE) are given in Appendix C. Numerical values that the main text refers to but does not show in the figures are tabulated in Appendix D. Anonymized code is provided as supplementary material: from a fixed seed it generates the dataset, trains all models for six seeds, aggregates the results, and regenerates Figures 1 to 3, the tables of Appendix D, and the correlation-dimension estimate of Appendix B. Its README lists the exact commands and the measured compute cost, and the aggregated result files are included so that the reported numbers can be checked without retraining. Unit tests cover the evaluation protocol and the structure of the block-diagonal predictor.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, pp. 1–12, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2352–2359, 2021.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35946–35958. Curran Associates, Inc., 2022. URL

https://proceedings.neurips.cc/paper_files/paper/2022/file/e97d1081481a4017df96b51be31001d3-Paper-Conference.pdf.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling, 2023. URL https://arxiv.org/abs/2110.05266.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Witold Hurewicz and Henry Wallman. *Dimension Theory*, volume 4 of *Princeton Mathematical Series*. Princeton University Press, Princeton, 1941.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning, 2022. URL https://arxiv.org/abs/2110.09348.

Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411, Mar 1992. doi: 10.1103/PhysRevA.45.3403.

Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nat. Commun.*, 9(1):4950, November 2018. doi: https://doi.org/10.1038/s41467-018-07210-0.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005. doi: https://doi.org/10.1007/s11071-005-2824-x.

Igor Mezić. Koopman operator, geometry, and learning of dynamical systems. *Notices of the American Mathematical Society*, 68(7):1087–1105, 2021.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Samuel E Otto and Clarence W Rowley. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019.

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120:024102, Jan 2018. doi: 10.1103/PhysRevLett.120.024102. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.024102.

Pablo Ruiz-Morales, Dries Vanoost, Davy Pissoort, and Mathias Verbeke. Koopman invariants as drivers of emergent time-series clustering in joint-embedding predictive architectures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25176–25183, 2026. doi: 10.1609/aaai.v40i30.39708. URL https://ojs.aaai.org/index.php/AAAI/article/view/39708.

Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65:3-4, 11 1991. doi: 10.1007/BF01053745.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1130–1140, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366–381. Springer, 1981.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning, 2020. URL https://arxiv.org/abs/1907.13625.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Stijn Verdenius, Andrea Zerio, and Roy L. M. Wang. Lat-pfn: A joint embedding predictive architecture for in-context time-series forecasting, 2024. URL https://arxiv.org/abs/2405.10093.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: https://doi.org/10.1007/s00332-015-9258-5.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8980–8987, Jun. 2022. doi: 10.1609/aaai.v36i8.20881. URL https://ojs.aaai.org/index.php/AAAI/article/view/20881.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.08496.

A Detailed Proofs

This appendix provides the detailed mathematical derivations for the lemmas, theorems, and loss decomposition presented in Section 3 of the main text.

Notation: Let $(\mathcal{M}, g)$ be a compact m-dimensional Riemannian manifold, and $T : \mathcal{M} \to \mathcal{M}$ a discrete-time C^r diffeomorphism ($r \geq 2$) preserving an ergodic probability measure μ. We denote the Hilbert space of scalar, square-integrable observables on $\mathcal{M}$ as $\mathcal{H} = L^2(\mathcal{M}, \mu)$. The inner product is $\langle h_1, h_2 \rangle = \int_{\mathcal{M}} h_1(u) h_2(u) d\mu(u)$ with induced norm $\|h\|_{L^2}$. The Koopman operator $\mathcal{K} : \mathcal{H} \to \mathcal{H}$ is defined by composition with the flow: $(\mathcal{K}h)(u) = h(T(u))$. Because T is measure-preserving, $\mathcal{K}$ is a unitary operator on $\mathcal{H}$. For vector-valued observables $F : \mathcal{M} \to \mathbb{R}^k$, $\mathcal{K}$ acts component-wise.

A.1 Proof 1: Predictive Learning as Approximate Koopman Subspace Identification

To understand what the JEPA encoder is capable of learning, we first express the input sequence space $\mathbb{R}^{n_c}$ and the latent space $\mathbb{R}^k$ strictly in terms of the physical manifold $\mathcal{M}$.

Lemma A.1 (JEPA Delay Mapping)**.** *The context window $x_t \in \mathbb{R}^{n_c}$ fed into the JEPA encoder is the evaluation of a delay-coordinate map $\Phi_{n_c} : \mathcal{M} \to \mathbb{R}^{n_c}$ at the true underlying state u_t. Consequently, the latent representation is a vector-valued observable $F \in [L^2(\mathcal{M}, \mu)]^k$.*

Proof. By Assumption 3.2, the scalar observation at time $t+i$ is exactly $s_{t+i} = \psi(T^i(u_t))$. Therefore, the context window $x_t = (s_t, s_{t+1}, \ldots, s_{t+n_c-1})$ can be written as

$$x_t = \left[\psi(u_t), \psi(T(u_t)), \ldots, \psi(T^{n_c-1}(u_t))\right]^\top.$$

Defining the delay map $\Phi_{n_c}(u) := [\psi(u), \psi(T(u)), \ldots, \psi(T^{n_c-1}(u))]^\top$, we have $x_t = \Phi_{n_c}(u_t)$.

The neural network encoder $f_\theta : \mathbb{R}^{n_c} \to \mathbb{R}^k$ maps this window to a latent representation. The composition $F := f_\theta \circ \Phi_{n_c}$ defines a vector-valued function $F : \mathcal{M} \to \mathbb{R}^k$. Assuming ψ and f_θ are sufficiently regular (e.g., continuous on a compact $\mathcal{M}$), the scalar components of F, denoted $\{h_j\}_{j=1}^k$, are bounded and therefore belong to $L^2(\mathcal{M}, \mu)$. □

Theorem A.2 (Zero Predictive Error; restatement of Theorem 3.3)**.** *Assume an ideal target network ($f_{\bar{\theta}} = f_\theta$) and that the scalar components $\{h_j\}_{j=1}^k$ of F are linearly independent in $L^2(\mathcal{M}, \mu)$. Then the predictive loss $\mathcal{L}_{MSE}(F, M) = 0$ if and only if $\mathrm{span}\{h_1, \ldots, h_k\}$ is a k-dimensional invariant subspace of the Koopman operator $\mathcal{K}^\Delta$ in $L^2(\mathcal{M}, \mu)$, and M is the exact matrix representation of $\mathcal{K}^\Delta$ restricted to that basis.*

Proof. By Lemma A.1, the current latent state is $z_t = F(u_t)$. The target latent state, advanced by Δ steps, is

$$z_{t+\Delta} = f_\theta\big(\Phi_{n_c}(T^\Delta(u_t))\big) = F\big(T^\Delta(u_t)\big).$$

By the definition of the Koopman operator, $F(T^\Delta(u_t)) = (\mathcal{K}^\Delta F)(u_t)$. Substituting these into the JEPA MSE objective, and using the ergodic theorem to equate the expectation over the data distribution $x \sim \mathcal{X}$ with the spatial integral over the invariant measure μ, we obtain

$$\mathcal{L}_{MSE}(F, M) = \mathbb{E}_{u\sim\mu}\Big[\big\|MF(u) - (\mathcal{K}^\Delta F)(u)\big\|_2^2\Big] = \sum_{j=1}^k \big\|(MF)_j - \mathcal{K}^\Delta h_j\big\|_{L^2}^2.$$

For deterministic dynamics this is the decomposition of Ruiz-Morales et al. (2026) with its stochasticity term equal to zero. This expected loss reaches its global minimum $\mathcal{L}_{MSE} = 0$ if and only if, for μ-almost every $u \in \mathcal{M}$, the vector equality

$$MF(u) = (\mathcal{K}^\Delta F)(u) \quad \text{in } \mathbb{R}^k$$

holds. Because norms are non-negative, $\mathcal{L}_{MSE} = 0$ if and only if $(MF)_j = \mathcal{K}^\Delta h_j$ for all $j = 1, \ldots, k$, holding μ-almost everywhere. Writing this in terms of the components, for each j,

$$\mathcal{K}^\Delta h_j = \sum_{i=1}^k M_{j,i}\, h_i.$$

This implies that the action of $\mathcal{K}^\Delta$ on any basis function h_j results in a linear combination of the functions $\{h_1, \ldots, h_k\}$. By definition, this means $\mathrm{span}\{h_1, \ldots, h_k\}$ is a $\mathcal{K}^\Delta$-invariant subspace, and M is the restriction of the operator to this finite-dimensional basis. Conversely, if $\mathcal{K}^\Delta h_j = \sum_i M_{j,i} h_i$ for every j, the integrand above vanishes μ-almost everywhere and $\mathcal{L}_{MSE} = 0$. Linear independence of $\{h_j\}$ is what makes "$\mathrm{span}\{h_1, \ldots, h_k\}$ is k-dimensional" meaningful and makes the coefficients $M_{j,i}$ unique: without it, the span has dimension less than k and the representation is degenerate rather than a genuine k-dimensional invariant subspace. □

Remark A.3 (Degenerate Global Minima)**.** The hypothesis of linear independence is a restriction on which solutions Theorem A.2 describes, not a property that equation 1 enforces. Any constant vector $F \equiv c \in \mathbb{R}^k$ satisfies $\mathcal{K}^\Delta F = F$ trivially, since constants are fixed points of $\mathcal{K}$; this is the constant case of the eigenvalue-one solutions of Ruiz-Morales et al. (2026), and what we add is that unregularized training reaches it as a collapsed representation. Taking $M = I$ gives $\mathcal{L}_{MSE} = 0$, while the components $h_j \equiv c_j$ are all multiples of the constant function 1 and hence linearly dependent whenever $k \geq 2$. This degenerate global minimum is not excluded by the objective in equation 1 alone, and unregularized training is driven toward it in practice (Section 5): the encoder's output shrinks toward a constant regardless of the true state, rather than forming an injective embedding of $\mathcal{M}$. Theorem A.2 therefore characterizes the non-degenerate zero-loss solutions; reaching one requires supplementing the objective with a term that rules out the degenerate case, which is the role of the variance/covariance regularizer of Section 4.3.

Remark A.4 (Deep EDMD Connection)**.** Exact finite-dimensional invariant subspaces rarely exist in systems with continuous spectrum, so $\mathcal{L}_{MSE} = 0$ is an idealization. In practice, the network identifies an approximate invariant subspace such that $\mathcal{K}^{\Delta}F = MF + R$, where R is a residual function. With an ideal target network, minimizing the MSE minimizes the $L^2(\mu)$ norm of this residual, which makes the JEPA mathematically equivalent to Deep Extended Dynamic Mode Decomposition (Deep EDMD) operating on a learned dictionary.

A.2 Proof 2: Temporal Context and Observability

We now establish the topological lower bound on the context window size n_c. For the neural network to learn the dynamical features of the point spectrum, the input window must provide a sufficient geometric embedding of the physical state.

Theorem A.5 (Context Length and State Reconstruction; restatement of Theorem 3.5)**.** *Let $A \subset \mathcal{M}$ denote the attractor supporting the observed dynamics, let m_A be its topological dimension, and let $d_A \geq m_A$ be its box-counting dimension.*

(i) If $n_c < m_A$, the delay map Φ_{n_c} is not injective on A for any *choice of observation function ψ. If moreover there are $u_a \neq u_b$ in A with $\Phi_{n_c}(u_a) = \Phi_{n_c}(u_b)$ and $F_{\bar{\theta}}(T^{\Delta}u_a) \neq F_{\bar{\theta}}(T^{\Delta}u_b)$, where $F_{\bar{\theta}} = f_{\bar{\theta}} \circ \Phi_{n_c}$ is the target representation, then $\mathcal{L}_{MSE} > 0$ for every M.*

(ii) If $m_A \leq n_c \leq 2d_A$, injectivity of Φ_{n_c} on A is neither guaranteed nor generically excluded; whether the predictive loss can vanish is not settled by this theorem.

(iii) If $n_c > 2d_A$, Φ_{n_c} is injective on A for a generic observation function ψ, and this topological-aliasing obstruction to zero predictive loss is generically absent.

Proof. (i) Dimension-counting obstruction. By classical dimension theory (Hurewicz & Wallman, 1941), a continuous injective image of a set of topological dimension m_A must lie in an ambient space of dimension at least m_A. Since Φ_{n_c} maps A into $\mathbb{R}^{n_c}$, $n_c < m_A$ rules out injectivity for *every* ψ, not merely a generic one. This is the first claim. For the second, take ψ, f_θ and $f_{\bar{\theta}}$ continuous, T a homeomorphism, and $A = \operatorname{supp}\mu$ (which T leaves invariant), and let (u_a, u_b) be a pair as in the statement. Put $\delta = \|F_{\bar{\theta}}(T^{\Delta}u_a) - F_{\bar{\theta}}(T^{\Delta}u_b)\| > 0$ and $e(u) = \|Mf_\theta(\Phi_{n_c}(u)) - F_{\bar{\theta}}(T^{\Delta}u)\|$, so that $\mathcal{L}_{MSE} = \mathbb{E}_\mu[e^2]$ and e is continuous on A. The predictions $Mf_\theta(\Phi_{n_c}(\cdot))$ coincide at u_a and u_b, because their windows do. By continuity there are neighborhoods $N_a, N_b \subset A$ of u_a, u_b on which the prediction and the target $F_{\bar{\theta}}(T^{\Delta}\,\cdot)$ each stay within $\delta/8$ of their values at u_a, respectively u_b. For $u \in N_a$ and $u' \in N_b$ the targets then differ by more than $3\delta/4$ and the predictions by less than $\delta/4$, so $e(u) + e(u') \geq \delta/2$ and $e(u)^2 + e(u')^2 \geq \delta^2/8$. Since $A = \operatorname{supp}\mu$, $p = \min\{\mu(N_a), \mu(N_b)\} > 0$, and averaging over $\mu(\cdot \mid N_a) \otimes \mu(\cdot \mid N_b)$ gives

$$\mathcal{L}_{MSE} \;\geq\; \int_{N_a} e^2\, d\mu + \int_{N_b} e^2\, d\mu \;\geq\; p\,\delta^2/8 \;>\; 0.$$

No positive-measure set of aliased states is needed: continuity and full support turn a single aliased pair into two sets of positive measure. The extra hypothesis cannot be dropped, since a constant target encoder attains zero loss (Remark A.3). It holds if $f_{\bar{\theta}}$ is injective on $\Phi_{n_c}(A)$ and not every aliased pair stays aliased after Δ. For $\Delta \leq n_c$ the latter follows when ψ separates the orbits of distinct points of A, since otherwise the two states would produce identical windows at all times $j\Delta$, hence identical observations forever; for $\Delta > n_c$, as in our experiments ($\Delta = 20$), it remains an assumption.

(ii) No general obstruction, and no general guarantee. Neither the classical smooth-embedding bound (which controls dimension via m) nor the fractal delay-embedding theorem (which requires $n_c > 2d_A$) applies in this band. Both injective and non-injective delay maps occur here depending on the system and the choice of ψ; the aliasing argument of part (i) cannot be invoked as a general claim, and no embedding guarantee analogous to part (iii) is available either. This range is a genuine gap in the theory, not a result we can state either way.

(iii) Generic embedding above $2d_A$. By the Fractal Delay Embedding Theorem (Sauer et al., 1991), for $n_c > 2d_A$ the delay map Φ_{n_c} is injective on A for almost every observation function ψ and generic delay parameters. For such ψ there is no positive-measure set on which distinct states alias,

so the argument of part (i) does not apply, and nothing in this theorem prevents $\mathcal{L}_{MSE}$ from reaching zero. □

A.3 Proof 3: Latent Capacity and Spectral Representation

Theorem A.6 (Capacity Requirement for Spectral Representation; restatement of Theorem 3.8)**.** *A necessary condition for exact representation of the targeted predictable subspace is $k \geq d_{spec}$. If $k < d_{spec}$, the latent space cannot contain a basis for the target Koopman subspace, and some modes must be merged or projected into shared directions.*

Proof. By Theorem A.2, achieving $\mathcal{L}_{MSE} \to 0$ requires the scalar components $\{h_j\}_{j=1}^{k}$ of the learned observable F to form a basis that fully spans the targeted invariant subspace $\mathcal{V}_{point}$.

The encoder output $F(u)$ maps to $\mathbb{R}^k$. Therefore, the functional subspace $S_F = \operatorname{span}\{h_1, \ldots, h_k\}$ generated by its coordinates has dimension at most k. By Definition 3.6, faithfully representing the single ergodic component ($r = 1$ under Assumption 3.1) and c independent phase oscillators requires a functional subspace of dimension exactly d_{spec}.

If the latent dimension $k < d_{spec}$, then $\dim(S_F) \leq k < d_{spec}$. Since a subspace of dimension less than d_{spec} cannot contain a subspace of dimension d_{spec}, S_F cannot contain $\mathcal{V}_{point}$.

Consequently, the network is forced to project the d_{spec}-dimensional physical dynamics onto a lower-dimensional latent space. Because Koopman eigenfunctions of distinct eigenvalues are linearly independent (and orthogonal, as $\mathcal{K}$ is unitary), this lower-rank projection necessarily maps linearly independent dynamical modes into shared latent directions. Geometrically, this manifests as topological squashing (e.g., projecting two-dimensional limit cycles into one-dimensional lines), permanently destroying linear phase predictability. □

A.4 Proof 4: Prediction Horizon and Spectral Bias

Finally, we formalize how the prediction horizon Δ acts as a spectral filter that isolates specific Koopman modes.

Theorem A.7 (Prediction Horizon and Representation of Mixing Components; restatement of Theorem 3.11)**.** *Under Assumption 3.10, and assuming further that the latent coordinates split disjointly between the point and continuous components, meaning that each scalar coordinate h_j lies entirely in $\mathcal{H}_{point}$ or entirely in $\mathcal{H}_{cont}$, the best linear predictor for the continuous component becomes ineffective as $\Delta \to \infty$. More precisely, the optimal linear map restricted to F_{cont} converges to zero in the long-horizon limit, and the minimum achievable loss on that component approaches $\|F_{cont}\|_{L^2}^2$.*

Proof. The proof has five steps.

1. **Spectral orthogonality.** Because $\mathcal{K}$ is unitary on $L^2(\mathcal{M}, \mu)$, the spectral decomposition of a unitary operator (Mezić, 2005) splits the Hilbert space orthogonally into a point-spectrum subspace $\mathcal{H}_{point}$ and a continuous-spectrum subspace $\mathcal{H}_{cont}$. Thus the learned observable decomposes as $F = F_{point} + F_{cont}$, with $\langle F_{point}, F_{cont} \rangle = 0$ componentwise. Because $\mathcal{K}^{\Delta}$ preserves this subspace structure, $\mathcal{K}^{\Delta} F_{point}$ remains orthogonal to $\mathcal{K}^{\Delta} F_{cont}$.

2. **Loss decomposition.** Due to this orthogonality, the MSE objective splits into the sum of the predictive errors on the two subspaces:

$$\mathcal{L}_{MSE}(F, M) = \mathbb{E}_{\mu} \left\| M F_{point} - \mathcal{K}^{\Delta} F_{point} \right\|_2^2 + \mathbb{E}_{\mu} \left\| M F_{cont} - \mathcal{K}^{\Delta} F_{cont} \right\|_2^2.$$

3. **Decay of correlations.** For observables in $\mathcal{H}_{cont}$ of a strongly mixing system, the temporal cross-correlation decays to zero asymptotically. Formally, for every pair of components,

$$\lim_{\Delta \to \infty} \langle \mathcal{K}^{\Delta} F_{cont}, F_{cont} \rangle_{L^2(\mu)} = 0.$$

   While the cross-correlation vanishes, the L^2 norm is strictly preserved by the unitary operator: $\|\mathcal{K}^{\Delta} F_{cont}\|_{L^2}^2 = \|F_{cont}\|_{L^2}^2$.

4. **Irreducible residual.** Let $\mathcal{I}_{point}$ and $\mathcal{I}_{cont}$ be the index sets of the coordinates lying in $\mathcal{H}_{point}$ and $\mathcal{H}_{cont}$, which partition $\{1, \dots, k\}$ under the disjoint-coordinates hypothesis, and permute coordinates so that $\mathcal{I}_{point}$ comes first. The self-covariance $\Sigma_0 = \mathbb{E}_\mu[FF^\top]$ is then block diagonal, $\Sigma_0 = \Sigma_{0,point} \oplus \Sigma_{0,cont}$, because $\langle h_i, h_j \rangle_{L^2} = 0$ whenever $i \in \mathcal{I}_{point}$ and $j \in \mathcal{I}_{cont}$. Likewise the target cross-covariance $\Sigma_{cross}(\Delta) = \mathbb{E}_\mu[(\mathcal{K}^\Delta F)F^\top]$ decomposes as $\Sigma_{cross,point}(\Delta) \oplus \Sigma_{cross,cont}(\Delta)$, because $\mathcal{K}^\Delta$ preserves the point/continuous split. Assuming the coordinates are linearly independent, so that Σ_0 is invertible, the least-squares optimal predictor $M^*(\Delta) = \Sigma_{cross}(\Delta)\,\Sigma_0^{-1}$ is block diagonal, and its continuous block
$$M_{cont}(\Delta) = \Sigma_{cross,cont}(\Delta)\,\Sigma_{0,cont}^{-1}$$
can be optimized independently of the action of M on the point-spectrum coordinates. By step 3, $\Sigma_{cross,cont}(\Delta) = \mathbb{E}_\mu[(\mathcal{K}^\Delta F_{cont})F_{cont}^\top]$ approaches the zero matrix as $\Delta \to \infty$, so $M_{cont}(\Delta) \to 0$. The partial loss on the continuous spectrum therefore converges to
$$\mathcal{L}_{cont}(\Delta) \;\longrightarrow\; \mathbb{E}_\mu \big\| 0 - \mathcal{K}^\Delta F_{cont} \big\|_2^2 = \|F_{cont}\|_{L^2}^2.$$

5. **Global optimum.** Under the disjoint-coordinates hypothesis, minimizing the continuous-spectrum block alone drives the encoder toward $\|F_{cont}\|_{L^2} \to 0$ as $\Delta \to \infty$, since retaining continuous-spectrum content contributes an irreducible penalty without improving predictive accuracy. This does not, however, imply pressure to preserve F_{point}: by Remark A.3, $F \equiv \text{const}$ is itself a global minimum of $\mathcal{L}_{MSE}$ alone for every Δ, not only in the long-horizon limit. The horizon argument shows that continuous-spectrum content becomes strictly more costly to retain than point-spectrum content as Δ grows; it does not show that the objective prefers retaining point-spectrum content over retaining nothing at all. Whatever pressure keeps F_{point} nonzero, rather than collapsing along with F_{cont}, comes from outside equation 1: the variance/covariance regularizer of Section 4.3.

Steps 1 to 4 establish the stated limits; step 5 records what they imply for the encoder. □

Remark A.8 (Without Disjoint Coordinates)**.** If F_{point} and F_{cont} instead share latent coordinates, Σ_0 no longer block-decomposes, and the jointly optimal $M^*(\Delta)$ cannot be split into independently optimized point and continuous parts. The continuous component's contribution to the target cross-covariance still vanishes as $\Delta \to \infty$, but its contribution to the shared self-covariance Σ_0 does not, so retaining nonzero F_{cont} still degrades achievable point-spectrum prediction even at large Δ. The disjoint-coordinates hypothesis is what licenses the clean isolation in step 4 of the proof. It is a property of the encoder's output, not of the predictor: a block-diagonal M removes coupling between blocks in the prediction, so the analysis applies block by block, but it does not by itself force the coordinates of a block to lie entirely in $\mathcal{H}_{point}$ or entirely in $\mathcal{H}_{cont}$.

Remark A.9 (Structured Predictors)**.** Under the disjoint-coordinates hypothesis, suppose the predictor is additionally constrained to be block orthogonal, as in Direct Koopman-JEPA; finding M then becomes a non-convex optimization. More importantly, an orthogonal M cannot decay to zero: its action on the continuous coordinates preserves the norm, $\|MF_{cont}\|_{L^2} = \|F_{cont}\|_{L^2}$, while the cross term $\langle MF_{cont}, \mathcal{K}^\Delta F_{cont} \rangle$ vanishes as $\Delta \to \infty$ by step 3. Thus the asymptotic loss on the continuous component approaches $2\|F_{cont}\|_{L^2}^2$ rather than $\|F_{cont}\|_{L^2}^2$. This stricter penalty provides even stronger optimization pressure for the encoder to eliminate unpredictable continuous spectra from the rotational subspaces.

B Experimental Implementation Details

This section provides the exact mathematical formulations and hyperparameter configurations used to generate the synthetic benchmark dataset. This ensures full reproducibility of the topological and spectral bounds discussed in the main text.

B.1 Data Generation and Dynamical Regimes

The dataset consists of 25,000 total trajectories (5,000 per regime) spanning five distinct dynamical systems. Each trajectory is simulated with a time step of $\Delta t = 0.10$. To ensure all sequences are

sampled directly from the systems' invariant measures (attractors), a burn-in period of 2,048 steps is simulated and discarded for every trajectory. The remaining 1,024 steps constitute the observed sequence.

B.1.1 ORNSTEIN–UHLENBECK PROCESS (STOCHASTIC EQUILIBRIUM)

We simulate a 3D mean-reverting stochastic process defined by $du = -\theta u\, dt + \sigma dW_t$. The reversion speed is set to $\theta = 1.0$ and the noise standard deviation to $\sigma = 0.1$. The system is integrated using the Euler-Maruyama method.

B.1.2 STUART-LANDAU OSCILLATOR (NOISY LIMIT CYCLE)

This system exhibits a stable limit cycle at $r = 1$. The state $u = [x, y, z]^\top$ evolves according to:

$$\begin{aligned} dx &= ((1 - r^2)x - \omega y)dt + \sigma dW_t^x \\ dy &= ((1 - r^2)y + \omega x)dt + \sigma dW_t^y \\ dz &= -z dt + \sigma dW_t^z \end{aligned}$$

where $r^2 = x^2 + y^2$, $\omega = 2.0$, and the noise scale is $\sigma = 0.2$. The z-dimension exponentially decays to zero. It is integrated via Euler-Maruyama.

B.1.3 QUASI-PERIODIC TORUS

A deterministic, quasi-periodic system is generated using two incommensurate frequencies, $\omega_1 = 2.0$ and $\omega_2 = 2\sqrt{2}$. The continuous trajectory is defined as $x_0(t) = \sin(\omega_1 t + \phi_1)$, $x_1(t) = \cos(\omega_1 t + \phi_1)$, and $x_2(t) = \sin(\omega_2 t + \phi_2)$, where $\phi_1, \phi_2 \sim \mathcal{U}(0, 2\pi)$ are random phase offsets. The 3D physical state is defined as $u_t = [x_0(t) + x_2(t), x_1(t), x_2(t)]^\top$.

B.1.4 LORENZ-63 ATTRACTOR (CHAOTIC, POINT AND CONTINUOUS SPECTRA)

The classical 3D Lorenz system is defined by:

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z$$

We use the standard chaotic parameters $\sigma = 10.0$, $\beta = 8/3$, and $\rho = 28.0$. The system is integrated using the explicit Runge-Kutta method of order 4(5) (`scipy.integrate.solve_ivp`). The underlying box-counting dimension is $d_A \approx 2.06$, theoretically requiring a context window $n_c \geq 5$ for topological unrolling.

B.1.5 HYPERCHAOTIC RÖSSLER (HIGH-DIMENSIONAL CHAOS)

A 4D hyperchaotic system defined by:

$$\dot{x} = -y - z, \quad \dot{y} = x + ay + w, \quad \dot{z} = b + xz, \quad \dot{w} = -cz + dw$$

Parameters are set to $a = 0.25$, $b = 3.0$, $c = 0.5$, and $d = 0.05$. It is similarly integrated via RK45. No literature value of d_A is available for this exact parameterization; we estimate it numerically via the Grassberger-Procaccia correlation-dimension algorithm, which gives $d_A \approx 2.5$, stable across sample sizes, sampling strides, and random seeds. This yields $2d_A \approx 5.1$, so $n_c \geq 6$ suffices for generic topological unrolling.

B.2 OBSERVATION MODEL AND TAKENS' EMBEDDING

To strictly enforce a partial-observability bottleneck, the multi-dimensional physical state $u_t \in \mathbb{R}^m$ is projected into a 1D scalar observation $s_t \in \mathbb{R}$ via a fixed linear measurement function: $s_t = w^\top u_t$.

To satisfy the genericity conditions of the delay-embedding theorems, we utilize a weight vector populated with distinct irrational constants to prevent symmetrical overlaps, truncated to the physical state's dimension m (3 for every regime except the 4D hyperchaotic Rössler): $w = [1.0, \pi/4, \sqrt{2}/3, e/5][:\, m]$. Finally, to prevent the architecture from trivializing the classification/separation of regimes purely by sequence magnitude, every 1D sequence is individually Z-score

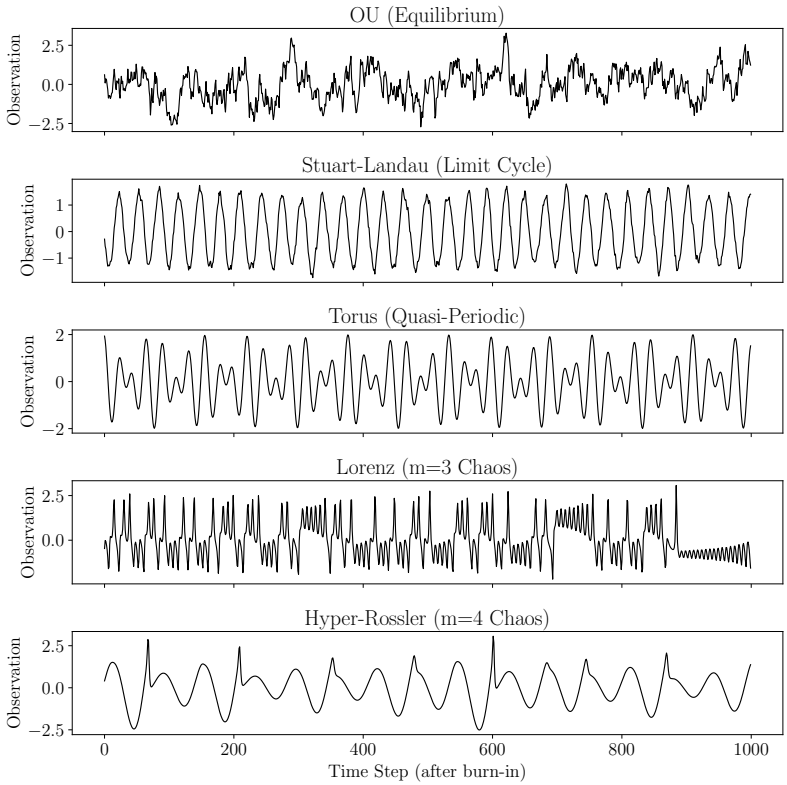


Figure 4: Samples for every regime scalar projection.

normalized (zero mean, unit variance) prior to being passed to the network. The model is thereby forced to organize its latent space strictly based on temporal frequency, phase geometry, and invariant measures.

C Architecture and Optimization Details

To evaluate the theoretical delay-embedding and capacity bounds without confounding variables introduced by complex sequence models (e.g., Transformers or deep CNNs), we parameterize the encoder f_θ using a straightforward Multi-Layer Perceptron (MLP). This explicitly forces the network to learn the topological mapping strictly from the context window $x_t \in \mathbb{R}^{n_c}$.

C.1 Encoder Architecture

The online encoder $f_\theta : \mathbb{R}^{n_c} \to \mathbb{R}^k$ consists of three hidden layers with a uniform width of 64 units, utilizing GELU activation functions. The final linear layer maps directly to the latent space of dimension k without a terminal activation function. The specific forward pass is defined as:

$$\text{Linear}(n_c \to 64) \to \text{GELU} \to \text{Linear}(64 \to 64) \to$$
$$\text{GELU} \to \text{Linear}(64 \to 64) \to \text{GELU} \to \text{Linear}(64 \to k)$$

C.2 Target Network

In accordance with standard JEPA training paradigms, the target encoder $\bar{f}_\theta$ shares the exact structural architecture as the online encoder but does not receive gradient updates. Instead, its weights are updated via an Exponential Moving Average (EMA) of the online network's weights:

$$\theta_{target} \leftarrow \tau\theta_{target} + (1-\tau)\theta_{online}$$

The EMA momentum parameter is fixed at $\tau = 0.996$ across all experiments.

C.3 Predictor Architecture

For the baseline standard JEPA, the predictor M is an unconstrained learned weight matrix $M \in \mathbb{R}^{k\times k}$. For the proposed Koopman-structured architecture, M is constrained to the block-diagonal orthogonal parameterization, where the planar rotation angles $\{\omega_i\}_{i=1}^{c}$ are learned jointly with the encoder via standard backpropagation. The split between static and rotational capacity follows: at least half of k is assigned to the identity block, and the remaining coordinates are divided into as many 2×2 rotation blocks as fit, with any leftover coordinate returned to the identity block.

C.4 Optimization and Hyperparameters

All models are trained using the Adam optimizer with a learning rate of 1×10^{-3} for 600 epochs, a mini-batch size of $B = 128$, and a rollout horizon H of 3. For the variance/covariance regularizer, the weighting coefficients controlling the variance-preservation term and the covariance-penalty term are set symmetrically to $\lambda = 0.3$ and $\mu = 0.3$. The unconstrained and Direct Koopman-JEPA arms use identical values of λ and μ, differing only in the predictor's parameterization.

For the k-NN classification metric, we draw 50 continuous stride-1 sliding windows from each of 10 held-out test trajectories per regime (500 windows total per regime). The resulting latent embeddings are split 70/30 by whole trajectory rather than by individual window, so that no trajectory contributes windows to both the training and test sets. A 5-Nearest-Neighbors classifier is then trained to predict the ground-truth regime label.

For the FNN metric, adapted from the dimension-increment test of Kennel et al. (1992) to a fixed-Δ latent trajectory, each held-out trajectory's embeddings are processed independently; concatenating trajectories would treat the boundary between them as a valid one-step transition, which it is not. For each point i, its nearest neighbor in latent space is found under Euclidean distance, excluding any candidate within a Theiler window of ± 10 samples in time (without this exclusion, a point's own close-in-time successor is almost always its nearest neighbor purely from temporal autocorrelation, which trivially passes the test and probes nothing about the embedding's spatial geometry). The pair is flagged false if $\|z_{i+1} - z_{\mathrm{nn}(i)+1}\| / \|z_i - z_{\mathrm{nn}(i)}\| > 10$, where the forward step is one sample at the trajectory's native sampling rate. The reported FNN% pools these per-trajectory percentages across all held-out trajectories and regimes, weighted by each trajectory's number of evaluated points.

D Additional Diagnostics

This appendix reports the quantities that Section 5 refers to but does not tabulate. Unless stated otherwise, values are means and standard deviations over the 6 seeds, evaluated on held-out test trajectories, and the variance/covariance regularizer is on ($\lambda = \mu = 0.3$).

D.1 Collapse Without Regularization

Table 1 gives, for each seed, the collapse ratio of Figure 1 (here without the regularizer), the relative spread of the latent cloud, defined as the square root of its total variance divided by $\|\mathrm{mean}(z)\|$, and its absolute spread, the square root of the total variance. All are computed at $n_c = 15$ and $k = 14$ on 500 stride-1 windows per regime. The collapse ratio is high whenever the cloud is narrow relative to its distance from the origin, so it does not by itself distinguish a cloud that has shrunk to a point from one that has merely moved away from the origin. The absolute spread, plotted in Figure 1, separates the two cases: the unconstrained predictor's is thousands of times that of Direct Koopman-JEPA in every seed, and in seed 4 its spread exceeds the offset of its mean, so that seed is not collapsed by this measure.

D.2 Predictor Ablation

Table 2 compares the unconstrained and the Direct Koopman-JEPA predictor under the regularizer at the largest cell of the grid ($n_c = 15$, $k = 14$). The coordinate standard deviation is the mean over latent coordinates of their standard deviation on held-out windows; the variance term of the regularizer pushes it toward 1. The unconstrained predictor sits at that target, while Direct Koopman-

Table 1: Collapse diagnostics per seed, without the regularizer ($n_c = 15$, $k = 14$). Relative spread is the total standard deviation of the latent cloud divided by $\|\mathrm{mean}(z)\|$.

	Unconstrained			Direct Koopman-JEPA		
Seed	Ratio	Rel. spread	Abs. spread	Ratio	Rel. spread	Abs. spread
0	0.93	0.40	0.109	1.00	4.3×10^{-5}	1.1×10^{-5}
1	0.94	0.36	0.089	1.00	8.2×10^{-5}	1.4×10^{-5}
2	0.99	0.14	0.039	1.00	6.2×10^{-5}	1.1×10^{-5}
3	0.99	0.13	0.046	1.00	3.4×10^{-5}	9.8×10^{-6}
4	0.67	1.38	0.287	1.00	6.9×10^{-5}	1.4×10^{-5}
5	0.96	0.31	0.102	1.00	6.4×10^{-5}	1.7×10^{-5}

JEPA sits below it but far from collapse. The normalized MSE panel of Figure 3 is drawn on a log scale, so its values for this cell are given here.

Table 2: Predictor ablation under the regularizer at $n_c = 15$, $k = 14$ (mean $\pm$ standard deviation over 6 seeds).

	Unconstrained	Direct Koopman-JEPA
Coordinate standard deviation	1.01 ± 0.01	0.76 ± 0.04
k-NN accuracy	0.962 ± 0.037	0.997 ± 0.002
FNN (%)	10.8 ± 5.5	9.7 ± 2.4
Normalized MSE	0.079 ± 0.029	0.028 ± 0.013

Table 3 reports the spectrum of the unconstrained predictor's learned matrix M. In four of the six seeds all 14 eigenvalues have modulus at least 0.95 and none exceeds 1.00 by more than 0.005, so the matrix is close to unitary in spectrum without being constrained to be. In the other two seeds one eigenvalue sits near 0.4. The matrix is not orthogonal in general: its singular values range from about 0.4 to 1.6, and in seed 0 they spread widely (0.61 to 1.58) although all eigenvalues lie near the unit circle, so M is far from a normal matrix there.

Table 3: Spectrum of the unconstrained predictor's matrix M under the regularizer ($n_c = 15$, $k = 14$).

Seed	$\min \lvert\lambda\rvert$	$\max \lvert\lambda\rvert$	$\#\,\lvert\lambda\rvert < 0.9$	$\sigma_{\min}$	$\sigma_{\max}$
0	0.95	1.00	0	0.61	1.58
1	0.38	1.00	1	0.38	1.05
2	0.37	1.00	1	0.37	1.08
3	0.95	1.00	0	0.90	1.05
4	0.95	1.00	0	0.89	1.05
5	0.96	1.00	0	0.90	1.06

D.3 Frequency Recovery

Table 4 summarizes Figure 2. A learned rotation angle is called active if it exceeds 0.05 rad; the number of active blocks is out of 18 (3 blocks per seed, 6 seeds). The error of an active angle is its distance to the nearest of the five analytic targets folded into $[0, \pi]$, and the chance error is the same quantity for angles drawn uniformly from $[0, \pi]$. The five target columns count the seeds, out of 6, in which some active block lies within 0.05 rad of the target. The smallest gap between two analytic targets is 0.20, 0.40 and 0.46 rad at $\Delta = 10$, 20 and 40.

The last column counts the seeds with an active block at the angle $\omega_R \, \Delta t \, \Delta$ (folded into $[0, \pi]$) implied by the dominant spectral peak $\omega_R \approx 0.92$ rad per time unit of the Rössler regime's observed signal. The peak is estimated from the mean periodogram of 400 trajectories per regime, which locates it only to within half a frequency bin, $\pi/(N \, \Delta t)$ for trajectories of N samples, so the tolerance on the angle is $\pi\Delta/N$: 0.03, 0.06 and 0.12 rad at $\Delta = 10$, 20 and 40. The same procedure returns

peaks near 2.0 rad for the Stuart-Landau regime and near 2.8 and 2.0 rad for the torus, matching their known frequencies 2 and $2\sqrt{2} \approx 2.83$. At $\Delta = 40$ the blocks at 2.50 and 2.54 rad of one seed fall inside both this tolerance and the 0.05 rad tolerance of $2\omega_2$, so they cannot be attributed to either and no count is given.

Table 4: Frequency recovery across the Δ-sweep (6 seeds per Δ). Errors are in radians.

		Error			Seeds recovering target					
Δ	Active	Mean	Median	Chance	ω_1	ω_2	$\omega_1-\omega_2$	$\omega_1+\omega_2$	$2\omega_2$	Rössler
10	15	0.078	0.102	0.19	5/6	0/6	1/6	0/6	0/6	6/6
20	17	0.032	0.006	0.18	4/6	6/6	2/6	0/6	2/6	1/6
40	17	0.015	0.005	0.17	6/6	4/6	0/6	2/6	1/6	–

D.4 Context Length and Latent Capacity

Tables 5 and 6 summarize the grid of Figure 3 along each axis. Table 5 averages over k and gives the mean FNN and the normalized MSE at each context length; for Direct Koopman-JEPA the mean FNN reaches its minimum at $n_c = 7$ and drifts up mildly for larger windows, which we do not analyze. Table 6 averages over the cells with $n_c \geq 5$ and gives, for each latent dimension, the k-NN accuracy, the mean FNN and the seed-to-seed standard deviation of FNN (averaged over those cells). In the bottom row of Figure 3, a dot requires, besides agreement of at least 5 of 6 seeds on the sign, a mean difference of at least 0.02 in k-NN accuracy or 0.05 in the $\log_{10}$ error ratio.

Table 5: Effect of the context length n_c, averaged over k. U: unconstrained predictor; K: Direct Koopman-JEPA.

	FNN (%)		Normalized MSE	
n_c	U	K	U	K
1	35.2	35.2	1.5	6.96
3	6.7	16.2	1.57	0.382
5	9.2	9.0	0.355	0.0832
7	10.2	7.0	0.177	0.0332
9	11.5	9.2	0.0991	0.026
11	10.5	10.2	0.0839	0.0239
13	10.6	11.6	0.0823	0.0193
15	10.1	12.3	0.0325	0.0216

Table 6: Effect of the latent dimension k, averaged over the cells with $n_c \geq 5$. U: unconstrained predictor; K: Direct Koopman-JEPA.

	k-NN accuracy		FNN (%)		FNN seed std (%)	
k	U	K	U	K	U	K
2	0.667	0.930	20.6	22.0	10.0	1.4
4	0.851	0.987	13.8	9.8	2.5	2.4
6	0.931	0.996	6.3	7.1	2.6	1.1
8	0.918	0.995	6.2	8.6	2.8	1.9
10	0.944	0.995	8.0	7.4	3.7	1.6
12	0.977	0.995	8.6	7.1	3.5	1.9
14	0.943	0.995	9.0	7.4	3.7	2.0

D.5 Identity-Block Analysis

Table 7 restricts the analysis of Direct Koopman-JEPA at $n_c = 15$, $k = 14$ to each block of the latent space. The identity block comprises the first $s = 8$ coordinates and the rotation blocks the remaining 6. The k-NN and nearest-centroid classifiers predict the regime from 50 stride-1 windows of each

of 10 held-out trajectories per regime, the windows of the main k-NN metric, under the same $70/30$ split by trajectory: every trajectory lies entirely on one side, so neighboring windows of a trajectory never appear in both training and test. The between-regime variance share of a coordinate is the fraction of its total variance explained by the regime label, averaged over the coordinates of the block. Directions of the pooled covariance with eigenvalue below 10^{-6} are counted as near-zero (between 2 and 4 per seed), and the last row gives the share of their squared loading that falls on the identity block.

Seed 4 is an outlier in three rows (the nearest-centroid accuracy of the rotation blocks, their between-regime variance share, and the share of near-zero directions in the identity block) and drives most of their spread. Its rotation blocks carry 33.2% of their variance between regimes and reach 96.1% nearest-centroid accuracy, and 84.4% of its near-zero directions lie in the identity block. In the other five seeds the rotation blocks carry at most 2.1% of their variance between regimes, their nearest-centroid accuracy lies between 52% and 68%, and all near-zero directions lie in the identity block.

Table 7: Identity-block analysis of Direct Koopman-JEPA ($n_c = 15$, $k = 14$). Values are percentages: mean $\pm$ standard deviation, and range, over 6 seeds.

	Mean $\pm$ std	Range
k-NN accuracy, full representation	99.7 ± 0.2	99.3 to 100.0
k-NN accuracy, identity block only	99.7 ± 0.2	99.5 to 100.0
k-NN accuracy, rotation blocks only	96.2 ± 2.6	91.7 to 99.7
Nearest-centroid accuracy, identity block only	99.6 ± 0.3	99.1 to 99.9
Nearest-centroid accuracy, rotation blocks only	64.9 ± 15.0	52.0 to 96.1
Between-regime variance share, identity block	98.9 ± 0.7	97.4 to 99.3
Between-regime variance share, rotation blocks	6.2 ± 12.1	0.1 to 33.2
Near-zero covariance directions lying in the identity block	97.4 ± 5.8	84.4 to 100.0

D.6 FNN as a Diagnostic

FNN is computed from the encoder's outputs alone and one native sample ahead, so it probes how smoothly the encoder maps a trajectory, not what the predictor achieves at the horizon Δ. Three checks qualify how it should be read (all reproduced by `scripts/fnn_diagnostics.py`). First, its level depends on the protocol: the same models give 35% at $n_c = 1$ with the 50 windows per trajectory of the main protocol and 89% with 500, because more windows bring nearer neighbors, so only comparisons at a fixed protocol are meaningful. Second, the raw windows used as embeddings have an FNN of 0% in every regime at $n_c = 3$ and 7 (35% at $n_c = 1$), so the FNN of the learned latent spaces above that reflects roughness added by the encoder, not aliasing of the input. It comes from two regimes: at $k = 14$ and $n_c = 7$ it is 26.6 ± 15.8 and 21.9 ± 15.5 for the unconstrained predictor and 10.4 ± 2.8 and 7.3 ± 2.2 for Direct Koopman-JEPA in the Ornstein-Uhlenbeck and Lorenz regimes, and at most 1% in the other three, which explains the large seed-to-seed spread (Table 6). Third, the difference between the predictors, averaged over $k \geq 4$, changes sign along n_c: unconstrained minus Direct Koopman-JEPA it is -5.4 points at $n_c = 3$ (Direct Koopman-JEPA higher in 5 of 6 seeds), between $+2.3$ and $+3.4$ for $n_c = 5, 7, 9$ and -1.6 and -2.4 for $n_c = 13, 15$; over the grid, at least 5 of 6 seeds favor Direct Koopman-JEPA in 14 cells and the unconstrained predictor in 18.

With the forward step set to the training horizon Δ instead of one sample, FNN favors Direct Koopman-JEPA at $n_c = 3$ (27% against 59% at $k = 14$) but splits the grid evenly (19 cells against 17). We do not use this variant to compare the predictors: for an orthogonal predictor, a zero prediction loss leaves the distance between two latents unchanged after Δ, so the variant is close to the training loss itself and favors that predictor by construction.

E Limitations and Broader Impacts

E.1 Limitations of the Proposed Framework

While the operator-theoretic perspective provides rigorous bounds for predictive self-supervised learning, several limitations remain:

1. Fixed Capacity Allocation: The split between static and rotational capacity is a deterministic function of k alone, not a choice adapted to the data's actual oscillatory content: at $k = 2$ it gives $c = 0$, and Section 5.3 shows this produces a persistent entanglement deficit with little variation across seeds (Table 6). On a benchmark pooling several ergodic regimes, this same rigidity shows up differently: most of the identity block's capacity ends up spent distinguishing which regime a window came from rather than encoding within-regime dynamics (on average 98.9% between-regime variance share, versus 6.2% for the rotation blocks; Table 7), a predictable consequence of training on a mixture of ergodic systems rather than the single ergodic system Assumption 3.1 describes.
2. Continuous Spectra: A linear predictor's ability to represent strongly mixing content decays as the prediction horizon grows, but this does not mean the architecture automatically isolates predictable structure while discarding the rest. For this reason, generating long-horizon forecasts for genuinely turbulent or strongly chaotic content remains out of reach for a linear predictor by construction.
3. Limits on Recovering the Full Spectrum: Which analytic frequencies a trained Direct Koopman-JEPA recovers depends on the random initialization of the rotation angles, since fitting them is a local, non-convex search: Figure 2 shows one block at $\Delta = 20$ settling near 1.86 rad, a Rössler-associated frequency, instead of any SL/torus target. No finite k could represent the full spectrum regardless of the seed, since a quasi-periodic system's point spectrum is a countably infinite set of combination frequencies once two or more base frequencies are present.
4. Stationarity Assumptions: The delay-embedding theorems assume that the underlying diffeomorphism T and the measurement function ψ are stationary. In applications involving severe non-stationarity or highly heteroscedastic measurement noise, the required context window n_c may effectively diverge, undermining the topological guarantees of Section 3.

E.2 Broader Impacts

Self-supervised learning has rapidly become the foundational paradigm for large-scale time-series models. However, standard predictive architectures frequently act as black-box feature extractors, optimizing statistical objectives that are disconnected from the physical reality generating the data.

We anticipate this framework will be particularly beneficial in scientific and medical domains—such as climatology forecasting, fluid dynamics, and the analysis of complex bio-signals (e.g., ECG/EEG) where preserving the true underlying dynamical geometry is critical for safety and downstream analysis.

Conversely, a potential negative impact arises from misapplying deterministic operator bounds to inherently stochastic, non-physical data (e.g., financial markets or social sentiment sequences). Forcing an orthogonal, norm-preserving prior onto systems entirely dominated by noise or exogenous shocks may yield overconfident but physically meaningless representations. Practitioners must ensure that the targeted datasets exhibit at least partial deterministic or oscillatory structure before aggressively applying Koopman-constrained inductive biases.