MOSAIC: Module Discovery via Sparse Additive Identifiable Causal Learning for Scientific Time Series

Shicheng Fan[1] **Nour Elhendawy**[2] **Jianle Sun**[3] **Ke Fang**[1]
Kun Zhang[3,4] **Yihang Wang**[2] **Lu Cheng**[1]
[1]University of Illinois Chicago [2]Case Western Reserve University
[3]Carnegie Mellon University [4]MBZUAI
{sfan25,kfang10,lucheng}@uic.edu {nae47,yihang.wang4}@case.edu
jianles@andrew.cmu.edu kunz1@cmu.edu

Abstract

Causal representation learning (CRL) seeks to recover latent variables with identifiability guarantees, typically up to permutation and component-wise reparameterization under appropriate assumptions. However, identifiability does not imply interpretability: latent semantics are typically assigned post hoc by alignment with known ground-truth factors. This limitation is particularly acute in scientific time series, where underlying mechanisms are unknown and discovering interpretable structure is a primary goal. In contrast, scientific observations (such as residue-pair distances, climate indices, or process sensors) are inherently semantic, as they correspond to named physical quantities. This raises a key question: *can the interpretability of observations be transferred to the identifiable latent space?* We propose **MOSAIC** (**Mo**dule discovery via **S**parse **A**dditive **I**dentifiable **C**ausal learning), a sparse temporal VAE that integrates temporal CRL identifiability with support recovery over observed variables. MOSAIC identifies latent variables via regime-conditioned temporal variation, and recovers for each latent a sparse set of associated observations through an additive decoder, yielding module-level interpretability. We show that ANOVA main-effect supports are identifiable under general smooth mixing functions, with finite-sample recovery guarantees for a group-lasso estimator and empirical evaluation of MOSAIC's entropy regularizer. Empirically, MOSAIC recovers domain-consistent variable groups across RNA molecular dynamics, solar wind, ENSO climate, the Tennessee Eastman process, and a synthetic tokamak benchmark, enabling interpretable discovery of latent mechanisms in scientific time series.

1 Introduction

Causal representation learning (CRL) [Schölkopf et al., 2021, Zhang et al., 2024] aims to recover latent variables uniquely up to trivial transformations, a property known as identifiability. In particular, temporal CRL under regime-conditioned dynamics [Hyvarinen and Morioka, 2016, Yao et al., 2022] achieves identifiability up to permutation and component-wise transformations. However, identifiability alone does not guarantee scientific interpretability. In standard CRL settings, the meaning of learned latent variables is typically assigned post hoc by aligning them with known ground-truth factors, rather than being directly inferred by the model.

Time series in scientific domains present a complementary yet contrasting setting. Their observed variables are often named physical quantities (e.g., residue-pair distances in protein systems) and

[1]Open source code: https://github.com/shichengf/mosaic

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

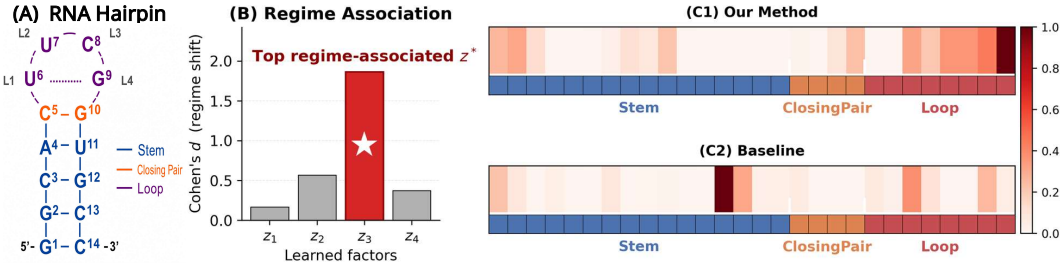

Figure 1: **RNA hairpin example. (A)** The cUUCGg RNA hairpin; colors mark three structural regions (Sterm, Closing Pair, Loop). **(B)** Regime-association score (Cohen's d) over four learned latents; z_3 is selected. **(C1, C2)** Influence column of z_3 under MOSAIC (C1) and TDRL [Yao et al., 2022] (C2); colored stripes match the regions in (A).

therefore inherently carry domain semantics. In contrast, the latents governing these observations correspond to underlying physical mechanisms that are typically unknown, and uncovering them is essential to scientific analysis. Existing scientific representation learning approaches (e.g.,[Wang et al., 2023]) leverage the semantic structure of observed variables, but do not provide identifiability guarantees for the uncovered latent mechanisms. As a result, scientific data offer interpretable observations but lack identifiable latents, whereas CRL provides identifiable latents without inherent scientific interpretation.

This suggests introducing CRL into scientific time series, but direct application of existing temporal CRL methods remains insufficient. Approaches such as TDRL [Yao et al., 2022, 2021, Song et al., 2024, Fan et al., 2026] identify latent variables, but rely on dense neural decoders that do not reveal which named observations each latent controls. As a result, the semantics of scientific observations are not transferred to the latent space. To address this gap, we propose **MOSAIC** (**Mo**dule discovery via **S**parse **A**dditive **I**dentifiable **C**ausal learning), a sparse temporal VAE that integrates temporal CRL with support recovery over named observations. MOSAIC first identifies latent variables via regime-conditioned temporal variation, and then recovers, for each latent, a sparse set of associated observations through an additive decoder. This transforms otherwise anonymous but identifiable latents into physically interpretable latent mechanisms.

The key idea is that MOSAIC uses an additive decoder that structurally separates each latent's contribution to the observations: each latent controls its own small network mapping to the observation space, so the set of observations it influences is well-defined and readable. We show that even when the true mixing function is non-additive, this additive fit recovers the correct per-latent influence pattern (Proposition 1), so the recovered supports are not artifacts of the modeling restriction. To scale to high-dimensional scientific data, MOSAIC also introduces a parallel transition prior removing the sequential bottleneck in prior temporal CRL [Yao et al., 2022, Song et al., 2024, Li et al., 2024].

Figure 1 illustrates this on an RNA molecule that folds into a hairpin shape. Its 14 residues group into three structural regions visible in Figure 1A: Stem (paired base), Closing Pair (hinge), and Loop (flexible tip). Each of the D=28 observed variables measures the distance behavior of one residue, so variable names carry physical meaning. MOSAIC ranks four learned latents by regime-association score (Figure 1B) and selects z_3, which is anonymous until we inspect its influence column. Under MOSAIC (Figure 1C1), influence concentrates on Loop residues, interpreting z_3 as tracking loop flexibility. Under TDRL (Figure 1C2), influence spreads across all groups, leaving z_3 uninterpretable. The key point is general: an identifiable latent must be anchored to named observed variables to yield a scientific interpretation.

Our contributions are fourfold. (1) **Problem formulation.** We formalize scientific time-series interpretability as recovering identifiable latent variables, their observed-variable supports, and the physical interpretation of regime-associated factors. (2) **Theory.** We prove population recovery of ANOVA main-effect supports, module and regime-association identifiability at $\rho = 0$, and a finite-sample bound for a group-lasso variant; for entropy regularization we give concentration and top-k statements. (3) **Architecture.** We introduce a sparse temporal VAE with two-stage curriculum and parallel transition prior. (4) **Experiments.** MOSAIC recovers localized modules on a controlled synthetic benchmark, a UUCG-loop correlate in RNA dynamics (Figure 1), and regime-associated factors on OMNI solar wind, ENSO climate, TEP, and a tokamak-inspired benchmark.

2 Related Work

Temporal causal representation learning. Nonlinear ICA is unidentifiable without auxiliary information [Hyvärinen and Pajunen, 1999, Locatello et al., 2019], and a now-mature literature restores identifiability through temporal structure [Hyvarinen and Morioka, 2016, Khemakhem et al., 2020, Yao et al., 2022, 2021, Brehmer et al., 2022], unknown nonstationarity [Song et al., 2023, 2024], instantaneous dependence [Li et al., 2024], continuous mechanism evolution [Fan et al., 2026], or mechanism and structural sparsity in the latent space [Lachapelle et al., 2022, 2024, Zheng et al., 2022, Zheng and Zhang, 2023]. All of these methods identify latent variables while keeping the latent-to-observed mapping flat and unstructured; none produces a module partition over named observed variables. Our experimental CRL baselines (TDRL [Yao et al., 2022], iVAE [Khemakhem et al., 2020], SlowVAE [Klindt et al., 2020], β-VAE [Higgins et al., 2017], and CtrlNS [Song et al., 2024]) span the most relevant settings: given clean regimes (TDRL, iVAE), weak temporal supervision (SlowVAE, β-VAE), and automatic regime discovery (CtrlNS); all exhibit the same support-recovery limitation in our experiments.

Causal discovery for scientific time series. Observed-space methods (Granger causality [Granger, 1969], transfer entropy [Schreiber, 2000], convergent cross mapping [Sugihara et al., 2012], PCMCI [Runge et al., 2019]; survey in [Gong et al., 2024]) recover pairwise dependencies between named variables, and LPCMCI [Gerhardus and Runge, 2020] detects latent confounders without naming them; recent work recovers latent causal graphs from spatiotemporal data [Wang et al., 2024] but does not solve the latent-to-observed support problem. These methods are complementary to MOSAIC: pairwise causal-discovery tools can operate within or between the modules MOSAIC identifies at reduced dimensionality.

Scientific representation learning and sparse statistical methods. VAMPnets [Mardt et al., 2018], SPIB [Wang and Tiwary, 2021], and TICA [Pérez-Hernández et al., 2013] produce useful low-dimensional summaries of scientific time series, but without identifiability, without explicit latent-to-observed support, and without a regime-association interpretation protocol. SPIB is our experimental baseline because it explicitly incorporates regime labels into its learned representation; VAMPnets and TICA target slow-mode kinetics rather than regime contrast. Sparse PCA [Zou et al., 2006] and ICA [Hyvärinen and Oja, 2000] yield readable sparse loadings under near-linear mixing, but a single linear projection cannot fit both saturating and expanding nonlinear responses from the same factor, and they lack temporal or regime-conditioned structure.

3 Problem Formulation

We work in a setting motivated by time series in scientific domains, where each observed variable is a scientific quantity carrying domain semantics. At each time step, we observe $\mathbf{x}_t \in \mathbb{R}^D$, where every x_i corresponds to one such named variable. This semantic structure on $\mathbf{x}_t$ enables interpretation of recovered latents through their observed-variable supports.

Let $\mathbf{x}_t$ be generated by n latent factors $\mathbf{z}_t \in \mathbb{R}^n$ through a smooth, injective mixing function $g(\cdot)$,

$$\mathbf{x}_t = g(\mathbf{z}_t) + \boldsymbol{\epsilon}_t, \qquad g : \mathbb{R}^n \to \mathbb{R}^D, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}), \tag{1}$$

with no additivity restriction on $g(\cdot)$. The latent modules follow a regime-dependent finite-lag transition model

$$z_{t,j} \mid \mathbf{z}_{t-1:t-L}, c_t \sim p_j(z_{t,j} \mid \mathbf{z}_{t-1:t-L}, c_t), \tag{2}$$

where $c_t \in \{0, 1\}$ indicates the regime (e.g., folded vs. unfolded in RNA dynamics). Here, "causal" follows temporal CRL, with c_t serving as an observed conditioning variable rather than an intervention. Modules whose transition dynamics differ across regimes are *regime-varying*; those with invariant dynamics are *regime-invariant*.

Main-effect supports as identifiable targets. Because $g(\cdot)$ need not be additive, a single variable x_i may depend on a latent factor z_j in a way that is irreducibly entangled with other factors. To define a recoverable target, we work under a product reference measure $\mu(\mathbf{z}) = \prod_{j=1}^{n} \mu_j(z_j)$ over normalized latent factors, the standard setting for Hoeffding ANOVA [Hoeffding, 1992]. Under this

reference measure, any square-integrable g admits the decomposition

$$g(\mathbf{z}) = g_0 + \underbrace{\sum_{j=1}^{n} g_j(z_j)}_{\text{main effects}} + \underbrace{\textstyle\sum_{j<k} g_{jk}(z_j, z_k) + \cdots}_{r(\mathbf{z})\text{: high-order interactions}}, \tag{3}$$

where $g_0 = \mathbb{E}_\mu[g(\mathbf{z})]$, each main effect $g_j : \mathbb{R} \to \mathbb{R}^D$ captures the marginal contribution of z_j (with per-coordinate components $g_j^{(i)} : \mathbb{R} \to \mathbb{R}$ for $i \in [D]$), and $r(\mathbf{z})$ collects all higher-order interactions. The expansion is $L^2(\mu)$-orthogonal and uniquely determined by the centering constraints $\mathbb{E}_{\mu_j}[g_j(z_j)] = 0$. MOSAIC's decoder targets the main effects $\{g_j\}$, a modeling choice analogous to Sparse PCA targeting the best linear projection. Section 4 shows the population-optimal additive fit recovers each g_j exactly.

For analysis, we quantify how much of the signal lives in interactions through the *interaction ratio*

$$\rho \;=\; \mathbb{E}_\mu\|r(\mathbf{z})\|^2 / \mathbb{E}_\mu\|g(\mathbf{z}) - g_0\|^2 \;\in\; [0, 1], \tag{4}$$

with $\rho = 0$ recovering exactly additive mixing. This quantity appears in the finite-sample bound of Section 4. Unless otherwise stated, population risks and variances in Section 4 are evaluated under this product reference measure and the independent observation-noise law.

Definition 1 (Main-Effect Support Structure)**.** *The main-effect support of the j-th component is $\mathcal{S}_j = \{i \in [D] : g_j^{(i)}$ is non-constant$\}$. When the $\{\mathcal{S}_j\}$ are mutually disjoint, they define a* module partition *of the observed variables, assigning each named x_i to a unique latent factor z_j. At $\rho = 0$ these coincide with the full supports of g; when $\rho > 0$, variables whose dependence on z_j lives entirely in higher-order terms lie outside $\mathcal{S}_j$.*

Recovery goals. From $\{(\mathbf{x}_t, c_t)\}$ we aim to recover (i) the latent factors $\mathbf{z}_t$ up to permutation and component-wise invertible transformation, (ii) the main-effect support $\mathcal{S}_j$ of each recovered latent factor, and (iii) the identity of the latent factor most associated with the regime contrast.

4 Identifiability Theory

This section establishes that the recovery goals of Section 3 are achievable in the appropriate sense: main-effect supports are identifiable for any smooth mixing function (Proposition 1), and at $\rho = 0$ the entire module partition together with the regime-associated identity is exactly recoverable (Theorem 1 and Corollary 1). For finite samples, we give a $O(N^{-4/5})$ estimation rate for the main effects (Proposition 3) and a per-channel sample-complexity bound for support recovery (Theorem 2).

Terminology. We use *regime-discriminative* and *regime-associated* interchangeably for the learned factor selected by the regime-discrepancy score, and *regime-varying* for the ground-truth factor whose dynamics differ across regimes.

Assumptions. (A1) Non-degenerate main effects, $|\mathcal{S}_j| \geq 1$ for all j. (A1′) Disjoint supports, $\mathcal{S}_j \cap \mathcal{S}_k = \emptyset$ for $j \neq k$, used when claiming a module partition or sparse-additive support recovery. (A2) Temporal sufficient variability [Yao et al., 2022]: the lag-vector cross-derivative matrix has full row rank, providing the auxiliary variation that identifies the latent factors up to permutation and component-wise reparametrization.

Proofs and the shared-support assignment rule (covering $|\mathcal{J}_i| \geq 2$) are in Appendix D.

4.1 Population-Level Identifiability

Proposition 1 (Main-Effect Identifiability Without False Positives)**.** *For any smooth and injective g, the centered additive fit $\hat{g}(\mathbf{z}) = \sum_j \hat{f}_j(z_j) + \hat{\mathbf{b}}$ minimizing the $L^2(\mu)$ population risk $\mathbb{E}_\mu\|\mathbf{x} - \hat{g}(\mathbf{z})\|^2$ subject to the centering constraints $\mathbb{E}_{\mu_j}[\hat{f}_j(z_j)] = 0$ for all j, satisfies $\hat{f}_j = g_j$ almost surely for all j. Hence $\hat{\mathcal{S}}_j = \mathcal{S}_j \subseteq \mathcal{S}_j^{\text{full}}$ is independent of ρ, where $\mathcal{S}_j^{\text{full}}$ denotes the full support of g in z_j.*

Remark. Proposition 1 is stated under the product reference measure $\mu = \prod_j \mu_j$. In practice, Stage 2 optimizes under the empirical encoder distribution, which approximates a product measure to the

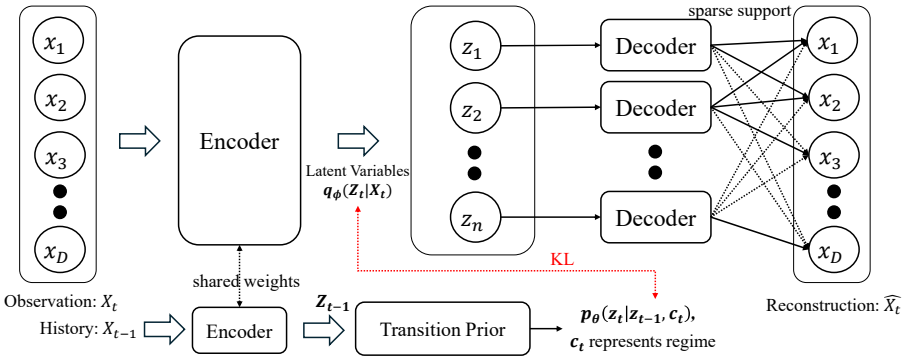

Figure 2: MOSAIC architecture. The encoder produces $\mathbf{z}_t$, which feeds per-factor additive decoders whose aggregated influence profiles form the support matrix $\mathbf{A}$. A regime-conditioned transition prior supplies the temporal identifiability signal via the posterior-prior KL. Stage 1 trains with a dense decoder; Stage 2 freezes the encoder and replaces it with the additive sparse head.

extent that Stage 1 identifiability succeeds. The gap between the two measures is absorbed by the interaction term $r(\mathbf{z})$ and does not alter the main-effect targets $\{g_j\}$; see Appendix D.2 for details.

Theorem 1 (Module Structure Identifiability at $\rho = 0$)**.** *If two models (g, p) and $(\tilde{g}, \tilde{p})$ both satisfy A1 and A2 with $\rho = 0$ and induce the same observational distribution $p(x_{1:T}, c_{1:T})$, then there exists a permutation π such that $\tilde{\mathcal{S}}_k = \mathcal{S}_{\pi(k)}$ for all k. If additionally A1′ holds, then the $\{\mathcal{S}_k\}$ form a module partition and this partition is identifiable.*

At $\rho = 0$, Theorem 1 sharpens Proposition 1: the entire module partition is identifiable, not just individual supports; when supports overlap ($|\mathcal{J}_i| \geq 2$), strict partition breaks down but support recovery remains sound:

Proposition 2 (Shared-Support Assignment Under Overlap)**.** *Let g satisfy $\rho = 0$ and define $\mathcal{J}_i = \{m \in [n] : i \in \mathcal{S}_m\}$. If $|\mathcal{J}_i| \geq 2$, then the population additive fit satisfies $\hat{f}_m^{(i)} = g_m^{(i)}$ for every $m \in \mathcal{J}_i$, and any hard-assignment rule $\arg\max_m A_{im}$ assigns i to a factor in $\mathcal{J}_i$ without false positives.*

Definition 2 (Regime-Discriminative Latent)**.** *Let $\Delta_j = \mathrm{KL}\big(p(z_j \mid c{=}0) \,\|\, p(z_j \mid c{=}1)\big)$. The set of regime-discriminative latents is $\arg\max_j \Delta_j$; when this maximizer is unique, we denote it j^*.*

Corollary 1 (Regime-Association Identifiability)**.** *Under the conditions of Theorem 1 with a unique maximizer, $\tilde{j}^* = \pi^{-1}(j^*)$ by KL invariance under invertible maps. Without uniqueness, the entire maximizer set is mapped by π^{-1}.*

The theory uses KL for invariance under component-wise invertible reparameterizations; in experiments, we use Cohen's d as a stable finite-sample proxy for mean shifts, with this substitution validated empirically in Appendix B.2.

4.2 Finite-Sample Behavior

Estimation efficiency is an active concern in CRL. Our population-level results above hold for any smooth estimator, while finite-sample guarantees are estimator-specific: for the classical group-lasso sparse additive estimator, we record a per-channel support-recovery bound in Appendix D.3 (Theorem 2), where D channels share the same N design points so the sample complexity is D-free. MOSAIC instead uses entropy regularization, for which finite-sample guarantees remain open, and we evaluate its performance empirically. We sweep N relative to D and interaction strength ρ on the synthetic benchmark (Appendices C.2, C.1) and calibrate $\hat{\rho}$ on real data (Appendix J); Section 6.4 discusses how to interpret these diagnostics on real data.

5 Method

MOSAIC realizes the theoretical ordering of Section 4 as a two-stage training curriculum (Figure 2): Stage 1 identifies latent factors with a temporal prior, and Stage 2 freezes the encoder and recovers each latent's main-effect support through an additive sparse decoder.

Stage 1: Identifiable latent learning. A VAE [Khemakhem et al., 2020] with encoder $q_\phi(\mathbf{z}_t \mid \mathbf{x}_t)$ and a dense MLP decoder is trained jointly with a regime-conditioned temporal prior factorized across dimensions,

$$p(\mathbf{z}_t \mid \mathbf{z}_{t-L:t-1}, c_t) = \prod_{j=1}^{n} p_j(z_{t,j} \mid \mathbf{z}_{t-L:t-1}, c_t), \tag{5}$$

where each p_j is a Laplace density on the output of a per-dimension transition MLP receiving past L lag vectors and current factor $z_{t,j}$. The factorized form supplies auxiliary variation for temporal identifiability [Hyvarinen and Morioka, 2016, Yao et al., 2022]. The Stage 1 loss combines reconstruction, a standard-Gaussian KL, and the temporal-prior KL; the full form is in Appendix A.

Stage 2: Sparse support recovery. With the encoder and prior frozen, the dense decoder is replaced by an additive head $\hat{\mathbf{x}} = \sum_j f_j(z_j) + \mathbf{b}$, where each f_j is a small MLP estimating the ANOVA main effect g_j. After standardizing each latent factor, we form the *functional influence matrix* $\mathbf{A} \in \mathbb{R}^{D \times n}$ via the bias-invariant finite contrast $A_{ij} = |f_j(+1)_i - f_j(-1)_i|$, which cancels the shared bias $\mathbf{b}$ and gives equivalent supports for non-symmetric main effects; symmetric responses (e.g., z_j^2) require a variance- or range-based score (Appendix D). To enforce module structure, we apply an entropy penalty on each active factor's normalized influence column:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{|\mathcal{J}_{\text{alive}}|} \sum_{j \in \mathcal{J}_{\text{alive}}} H\big(A_{:,j} / \textstyle\sum_i A_{ij}\big)\,, \tag{6}$$

where $H(\cdot)$ is the Shannon entropy and $\mathcal{J}_{\text{alive}}$ masks near-dead factors. Low-entropy columns realize the supports $\{\mathcal{S}_j\}$ of Theorem 1: each z_j concentrates its influence on a small subset of observed variables. The Stage 2 objective adds $\lambda(t) \cdot \mathcal{L}_{\text{sparse}}$ to the frozen-encoder ELBO (schedule in Appendix A). Entropy is preferred over column-level magnitude penalties for module-partition alignment; Appendix E provides the mechanistic argument, an empirical comparison, and a discussion of how this column-level penalty relates to the basis-function group-lasso analyzed in Theorem 2. An alternative approach would directly sparsify a dense decoder's Jacobian $\partial\hat{\mathbf{x}}/\partial\mathbf{z}$; we discuss why the additive architecture is preferable in Appendix E.1.

Inference: regime-association interpretation. After training, MOSAIC ranks the learned latent factors by the regime KL divergence (Definition 2); under the mean-shift regime contrasts in our experiments, Cohen's d provides a stable finite-sample proxy [Cohen, 1988]. The selected z_{j^*} is interpreted through its influence column $A_{:,j^*}$, whose largest entries name the observed variables that give z_{j^*} its physical implication.

Parallel transition prior. The Stage 1 temporal prior requires component-wise Jacobians for the change-of-variables term. Reference implementations [Yao et al., 2021, 2022, Brehmer et al., 2022, Song et al., 2024] loop over n dimensions sequentially, which dominates wall-clock cost at scientific scale (synthetic training at $\hat{n} = 8$ takes $\sim$65 GPU-hours per seed). We replace the sequential loop with a single batched computation that yields the same log-det-Jacobian, giving $\geq 277\times$ wall-clock speedup on NVIDIA A40 (Table 6; numerical equivalence verified in Appendix A).

6 Experiments

We organize experiments around the **two abstract capabilities** of MOSAIC: identifying latent factors and recovering their observation-level supports.

Setup. We evaluate on six benchmarks across various domains with sample-to-dimension ratios spanning 13 to 14,000: synthetic ($D = 30$, $N = 420$K, full ground truth), RNA molecular dynamics (MD, $D = 28$, $N = 156$K), tokamak-inspired control ($D = 12$, $N = 4.6$K), OMNI solar wind [King and Papitashvili, 2005] ($D = 21$), climate ENSO [Huang et al., 2017b] ($D = 47$), and TEP [Downs and Vogel, 1993] ($D = 52$). The synthetic benchmark uses monotonic invertible nonlinear mixing with five function families (Appendix C). We run 5 seeds on synthetic and RNA, 3 on cross-domain. We additionally test structural localization on HP35 protein dynamics ($D = 132$; Section 6.3).

Baselines. We compare against three families (citations in Section 2): temporal CRL methods (TDRL, iVAE, SlowVAE, β-VAE, CtrlNS), linear sparse methods (Sparse PCA, ICA, PCA), and the regime-conditioned information bottleneck SPIB. For all methods, regime-association ranking

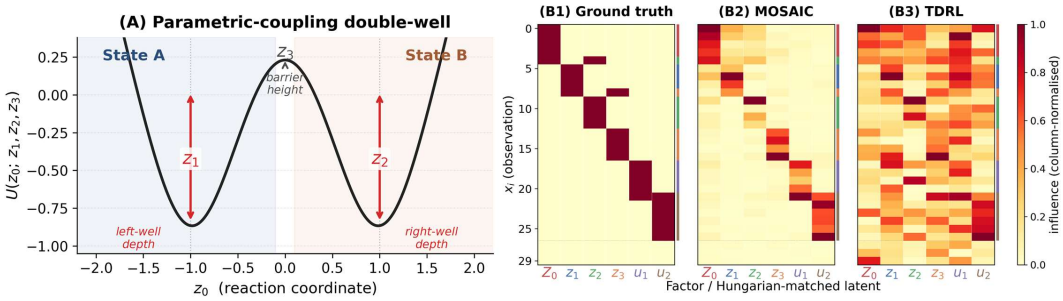

Figure 3: Physics-inspired double-well benchmark. **(A)** A reaction coordinate z_0 moves between two metastable basins (State A / State B); z_1, z_2, z_3 modulate left-well depth, right-well depth, and barrier height. **(B1–B3)** Column-normalized influence matrices over six ground-truth factors (z_0, z_1, z_2 regime-varying; z_3, u_1, u_2 regime-invariant). Learned latents in B2 and B3 are Hungarian-matched to the ground truth in B1. MOSAIC (B2) recovers the diagonal module structure; TDRL (B3) does not.

uses a shared Cohen's d pipeline; localization is read from additive decoder probes (MOSAIC), dense-decoder Jacobians (CRL methods), or linear loadings (Sparse PCA/ICA/PCA).

Metrics. Three metrics target different aspects of recovery: **MCC**, the mean correlation coefficient between learned and ground-truth latents under Hungarian matching, measures latent identifiability; **Z@top3** checks whether the three most regime-discriminative learned latents each correspond to a true regime-varying factor; and X_Z**@top3** measures support precision of those three latents against ground-truth observation supports, gated at top-3 mass ≥ 0.50 to penalize diffuse decoders whose argmax happens to overlap by chance. Full definitions are in Appendix B.

6.1 Physics-Inspired Synthetic Energy-Landscape Benchmark

We validate both capabilities on a controlled double-well benchmark with known ground-truth factors and observation-level supports. A reaction coordinate z_0 moves between two basins (local energy minima) while z_1, z_2, z_3 modulate left-well depth, right-well depth, and barrier height; State A/B occupancy defines the regime label. The benchmark generates $D = 30$ observations from $n_{\text{true}} = 6$ latents under monotonic nonlinear mixing, with shared-channel overlaps that violate the strict disjoint-support assumption A1$'$ to probe the more realistic partial-overlap regime of Proposition 2 (Figure 3; full specification in Appendix C).

MOSAIC is the only method that satisfies both abstract capabilities (Table 1): it identifies the latent factors with the best MCC (0.912), correctly identifies the regime-associated factors in all 5 seeds, and recovers their observation-level supports with the best X_Z@top3 (0.978). The baselines fall into three families that each cover only one capability. CRL methods (TDRL, iVAE, SlowVAE, β-VAE) use dense decoders, so although they recover useful latent factors (TDRL MCC 0.850), their supports are diffuse and gated X_Z@top3 is zero. Linear sparse methods (Sparse PCA, ICA, PCA) produce readable loadings (un-gated X_Z@top3 around 0.6–0.7) but cap MCC below 0.78 for lack of nonlinear capacity. SPIB's information-bottleneck objective does not align with driver recovery: Z@top3 and X_Z@top3 are uniformly zero across 36 runs (Appendix M). After correcting two tensor-shape errors in our CtrlNS adapter, a 3×3 search over transition weights and learning rates yields MCC 0.791, validation MSE 0.0389, and native gate accuracy 0.539 for the configuration selected by validation gate accuracy, with MSE breaking ties. CtrlNS learns useful latent structure, but its internally discovered partition need not align with the specified scientific regime. Post-hoc top-3 thresholding of TDRL's dense influence map only reaches X_Z@top3 $= 0.444$, confirming that sparsity alone cannot recover module structure without joint training. This X_Z@top3 of 0.978 is achieved on a benchmark that deliberately violates A1$'$ (Proposition 2).

6.2 RNA Molecular Dynamics: Interpreting a Regime-Associated Latent

We now apply MOSAIC to real scientific data where recovered supports can be read as substantive scientific claims. The mechanistic question is: *which learned latent factor is most associated with the C5-G10 closing-pair regime shift, and which physical residues give that factor its meaning?*

Setup. We perform MD simulations of the cUUCGg tetraloop (`GGCACUUCGGUGCC`) at 345K, 400K, and 500K (14 trajectories total; full simulation protocol in Appendix H). Per-residue distance features

Table 1: Synthetic benchmark with monotonic nonlinear mixing. X_Z@top3 uses the 0.50 concentration gate; stochastic methods report mean $\pm$ std over 5 seeds. iVAE uses the true regime label c_t. The tuned CtrlNS result is reported in the text; linear baselines report un-gated X_Z@top3. Detailed hyperparameters are in Appendix B; SPIB diagnostics are in Appendix M.

Method	MCC ↑	Z@top3 perfect seeds ↑	X_Z@top3 (gate 0.50) ↑
MOSAIC (ours)	$\mathbf{0.912 \pm 0.010}$	$\mathbf{5/5}$	$\mathbf{0.978 \pm 0.044}$
iVAE ($u = c_t$)	0.821 ± 0.007	3/5	0.556 ± 0.157
TDRL	0.850 ± 0.026	4/5	0.000 ± 0.000
TDRL + post-hoc top-3	0.850 ± 0.026	4/5	0.444 ± 0.211
SlowVAE	0.696 ± 0.046	2/5	0.000 ± 0.000
β-VAE (β=4, γ=0)	0.760 ± 0.060	0/5	0.000 ± 0.000
Sparse PCA (α=10)	0.775	—	0.610 (no gate)
ICA	0.765	—	0.610 (no gate)
PCA	0.684	—	0.670 (no gate)
SPIB	0.304 ± 0.019	0/5	0.000 ± 0.000

give D=28 named observed variables, which group into Loop (U6–G9), ClosingPair (C5, G10), and Stem residues for interpretation; these group labels are not used to supervise the latents. Regimes are defined by the closing-pair native-contact fraction Q_{cp} on contacts touching C5 or G10, with intact ($Q_{\text{cp}} > 0.8$) and disrupted ($Q_{\text{cp}} < 0.65$) frames forming a balanced dataset of 156,170 lag-2 windows ($N/D \approx 5{,}571$).

MOSAIC identifies a Loop-localized correlate of the closing-pair shift. The Cohen's-d ranking selects a single regime-discriminative latent z_3 (Figure 1B). Its influence column $A_{:,3}$ concentrates on Loop residues, with the top-4 entries 100% Loop (Figure 1C1). This is consistent with the cUUCGg folding pathway studied by Chen and García [2013], in which loop fluctuations couple to closing-pair disruption. Across 5 seeds, Loop ranks first in 4 runs and is the dominant residue group in all 5 (the fifth seed surfaces stem-loop coupled residues that remain physically consistent with loop dynamics; details in Appendix H). Regime accuracy is 0.912 ± 0.027.

Comparison to baselines. The same three baseline families from Section 6.1 reappear here. CRL methods (TDRL, iVAE, SlowVAE, β-VAE) yield a regime-associated latent but localize it to the wrong or diffuse residue groups; Figure 1C2 shows TDRL producing a sparse-looking column whose mass sits on Stem rather than Loop (X_Z@selected-latent ≤ 0.625 across CRL methods). This failure mode is informative: sparsity alone is insufficient without the additive main-effect projection that MOSAIC uses to align sparse mass with the correct latent. Table 2 summarizes the comparison; full per-seed details are in Appendix H. The Loop localization is not circular: redefining regimes by outer-stem contacts shifts the dominant residue group from Loop to Stem (Appendix H, Table 21).

Table 2: RNA baseline comparison (X_Z@selected-latent, top-8 precision against Loop).

Method family	Regime Acc	X_Z@sel ↑
MOSAIC (5s)	$\mathbf{0.912 \pm 0.027}$	$\mathbf{0.880 \pm 0.055}$
CRL best (TDRL, 4s)	0.912 ± 0.030	0.625 ± 0.102
Linear best (SpPCA α=10)	0.898	0.750
SPIB (5s)	0.645 ± 0.065	0.400 ± 0.137

6.3 HP35 Protein Dynamics: Localizing an H1 Folding Contrast

We use the HP35 Anton trajectory [Piana et al., 2012] to test whether a specified folding contrast can be localized to its structural region (Figure 4).

Setup. Each frame is represented by 132 backbone dihedral features. Before training, H1 folding is specified as the target contrast using native contacts from an independent distance representation. The methods receive the global dihedral inputs and regime labels, but neither the distances nor H1 annotations. We select the component with the largest absolute Cohen's d, then evaluate its separation from the second-best component and count how many of its top-40 features lie in H1, which contains 40 of the 132 input features. Features are ranked by decoder influence for MOSAIC and by loadings for linear methods.

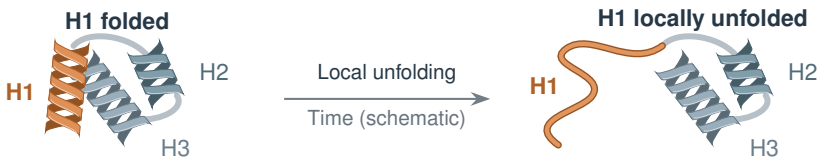

Figure 4: **HP35 protein dynamics.** Schematic time evolution from an H1-folded to an H1-locally-unfolded conformation, with H1 highlighted in orange.

MOSAIC identifies a distinct H1-associated latent. Its regime separation is 10.06 times that of the second-best component, and $30/40$ selected features lie in H1 (Table 3).

Comparison to baselines. PCA and FastICA also identify a leading component, but only 40% and 55% of their selected features lie in H1. Sparse PCA with $\alpha = 1$ places more than half outside H1, while $\alpha = 10$ improves H1 concentration but gives nearly tied H1 and H2 components, with a separation ratio of only 1.02. MOSAIC thus combines distinct regime association with concentrated H1 support, localizing the structural change associated with the specified folding contrast.

Table 3: HP35 H1 localization. Separation is the ratio of the largest to the second-largest absolute Cohen's d. H1 localization counts the H1 features among the selected component's top-40 features.

Method	Separation	H1 in top-40
MOSAIC	$10.06\times$	$30/40$ (75.0%)
PCA	$4.17\times$	$16/40$ (40.0%)
Sparse PCA ($\alpha = 1$)	$1.37\times$	$19/40$ (47.5%)
Sparse PCA ($\alpha = 10$)	$1.02\times$	$27/40$ (67.5%)
FastICA	$11.28\times$	$22/40$ (55.0%)

6.4 Cross-Domain Generalization

We now test whether the same pipeline generalizes across domains with different physics. We apply MOSAIC to three empirical datasets (OMNI solar wind [King and Papitashvili, 2005], ENSO climate [Huang et al., 2017b], Tennessee Eastman Process [Downs and Vogel, 1993]), and a physics-motivated synthetic tokamak benchmark.

Setup. Without ground-truth latent supports, we evaluate two coarser criteria. *Top-1 match*: does the regime-discriminative latent's largest-influence variable lie in the domain-expected group? *Winner group*: does the per-group mean influence rank the expected group first? We also report the *ratio* of the winner group's mean influence to the runner-up's, measuring localization sharpness. Expected groups are determined by domain knowledge (e.g., Niño bands for ENSO, MHD/density for tokamak) and used only for post hoc evaluation.

MOSAIC matches the domain-expected group in all 12 seeds across the four datasets (Table 4). Localization sharpness varies with how cleanly the regime label maps to a single physical mechanism: the disruption benchmark gives ratios up to 2.13, while TEP's multi-fault regime (IDV-1, 2, 4, 5) converges onto the shared downstream Stripper signature with ratio 1.07. As a reference point, SparsePCA and FastICA each achieve 3/3 winner-group recovery on three of the four datasets, confirming that the regime signal is detectable at group level by simple methods; however, they lack identifiable latents and concentrated supports. PCA fails on OMNI and Climate by selecting variance-dominant rather than regime-relevant variables. SPIB succeeds on 5/9 runs, failing all three OMNI seeds. MOSAIC is the only method achieving 12/12 across all four datasets. Full per-method tables are in Appendix I.

Case Study: low-N/D regime. Climate ($N/D = 13$) lies below the recovery range observed in our synthetic sweep (Appendix C.2). This sweep is a sample-size stress test for a fixed generator, not a transferable reliability threshold: recovery also depends on temporal dependence, signal strength, sparsity, interactions, and collinearity. The top regime-associated latent concentrates on equatorial Pacific SST cells (Figure 5), recovering a spatial region broader than the single grid cell defining the regime label, consistent with the underlying ENSO mode. Climate thus demonstrates regime prediction and coarse group localization, with sensor-level support ground truth unavailable. For new datasets without support labels, we propose checking effective sample size and regime transitions, the dense–additive residual gap on held-out temporal blocks, support concentration against a block-

Table 4: Cross-domain localization (MOSAIC, 3 seeds per dataset). Ratio = winner-group mean influence / runner-up.

Dataset	Domain	D	Regime Acc	Top-1 match	Winner group	Ratio
OMNI	Solar wind	21	0.940 ± 0.054	3/3	3/3 (Geomag/Coupling)	1.00–1.83
Disruption	Tokamak (synthetic)	12	0.911 ± 0.003	3/3	3/3 (MHD/Density)	1.38–2.13
Climate	ENSO	47	0.888 ± 0.027	3/3	3/3 (Niño bands)	1.16–2.00
TEP	Chemical proc.	52	0.729 ± 0.065	3/3	3/3 (Stripper)	1.07

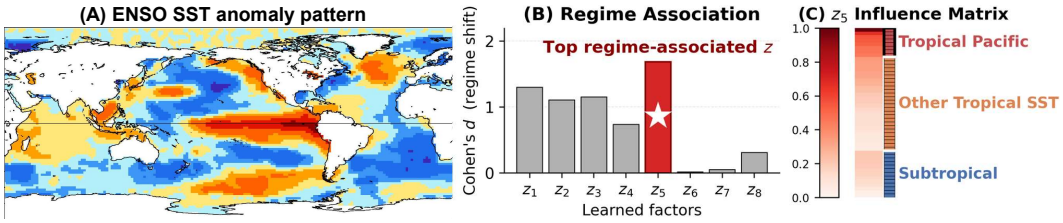

Figure 5: **Climate (ENSO) cross-domain localization** on NOAA ERSSTv5 sea surface temperature data [Huang et al., 2017b,a]. **(A)** The canonical ENSO SST anomaly pattern, with the warm equatorial Pacific band marking the physical mode underlying the regime label. **(B)** Cohen's d across 8 learned latents identifies z_5 as the top regime-associated factor. **(C)** Influence of z_5 over the 47 SST grid cells, ordered by group: support concentrates on the Tropical Pacific group corresponding to the equatorial Pacific band in (A), extending beyond the single grid cell that defines the regime label.

permutation null, and support stability across blocked refits against a random-support null. These diagnostics screen for failure modes without certifying exact recovery.

Takeaway. Each experiment stresses a different aspect of the pipeline: the synthetic benchmark probes both capabilities under controlled nonlinear mixing with deliberate support overlap; RNA tests whether the recovered support constitutes a substantive scientific claim on a real molecular system, where MOSAIC's advantage is largest (Tables 1, 2); and the four cross-domain datasets are non-redundant (Climate: low N/D; TEP: multi-fault aggregation; Disruption: signal-strength separation; OMNI: geomagnetic regimes). MOSAIC is the only method that handles all six benchmarks with a single pipeline and shared hyperparameters. HP35 further tests localization of a specified folding contrast to a protein region.

7 Conclusion and Limitations

MOSAIC closes a gap between temporal CRL, which identifies latents but cannot determine what they represent, and scientific representation learning, which operates on named variables but lacks identifiability. By combining both, MOSAIC turns anonymous identifiable latents into module-level scientific interpretations, validated across diverse domains.

Ablation and robustness. Ablation on the synthetic benchmark (Appendix F.1) shows a clean decomposition: MCC depends only on the temporal prior, while X_Z requires all three components together; removing any single component breaks the alignment between sparse supports and correctly identified latents. Additional robustness experiments confirm graceful degradation under varying latent dimensionality (Appendix F.2), regime-label noise up to 30% (Appendix F.4), unsupervised regime discovery via k-means (Appendix C.4), and alternative regime definitions (Appendix C.3).

Scope and limitations. MOSAIC interprets a binary or one-vs-reference regime contrast at a time; multi-regime joint analysis is left to future work (Appendix F.3). Entropy regularization is used instead of the group-lasso variant of Theorem 2 because it directly enforces within-column concentration, at the cost of an open finite-sample theorem (Appendix E). Regime association is distributional, not interventional, restricted to mean shifts under Cohen's d; channels with weak mixing coefficients or purely higher-order dependence may fall outside the recovered support. MOSAIC complements rather than replaces existing tools: SPIB can supply regime labels, and pairwise causal-discovery methods can operate within recovered modules. Nonlinear optimization introduces seed-to-seed variance (reported as standard deviations throughout). The goal is hypothesis generation, not autonomous causal decisions, with potential applications in space weather forecasting, fusion diagnostics, climate monitoring, and chemical-process safety.

Acknowledgments

I thank my parents and friends for being there for me during a particularly difficult period in my life, when this work was carried out. Their companionship, care, and encouragement helped me through that time and gave me the strength to complete this work.

This work is supported by the National Science Foundation (NSF) Grant #2312862, NSF-Simons SkAI Institute, NSF CAREER #2440542, NSF #2533996, National Institutes of Health (NIH) #R01AG091762, NSF ACCESS Computing Resources, NAIRR, NRP, a Google Research Scholar Award, and Cisco gift grant.

Competing Interests

The authors declare no competing interests.

References

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Alan A Chen and Angel E García. High-resolution reversible folding of hyperstable rna tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 110 (42):16820–16825, 2013.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. Mvi-bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in lvlms. *arXiv preprint arXiv:2511.14159*, 2025.

Xuan Chen, Shuxun Wang, Shunli Yu, Yanzhong Cai, Zhiqiang Ma, Chenyang Gao, Ning Luo, Dibo Hou, and Yunqi Cao. A sequential end-of-line testing framework for waterproof micro-motor systems: Integrating hydro-load vibration analysis and airborne acoustic classification. *IEEE Sensors Journal*, 2026a.

Xuan Chen, Shuxun Wang, Yicheng Zhang, Yongping Ye, Jianbo Ye, Wei Li, Dibo Hou, and Yunqi Cao. Embedded-ai-driven on-site traumatic brain injury assessment using wireless flexible wearable sensors for real-time impact force and acceleration estimation. *IEEE Internet of Things Journal*, 2026b.

Xuan Chen, Xinjun Zuo, Yancheng Bi, Shunli Yu, and Yunqi Cao. A data-driven vibration analysis framework for micro-motor fault diagnosis and quality control. *arXiv preprint arXiv:2607.26519*, 2026c.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

PC De Vries, MF Johnson, B Alper, P Buratti, TC Hender, HR Koslowski, V Riccardo, and Jet-Efda Contributors. Survey of disruption causes at JET. *Nuclear fusion*, 51(5):053018, 2011.

James J Downs and Ernest F Vogel. A plant-wide industrial process control problem. *Computers & chemical engineering*, 17(3):245–255, 1993.

Shicheng Fan, Kun Zhang, and Lu Cheng. TRACE: Trajectory recovery for continuous mechanism evolution in causal representation learning. *arXiv preprint arXiv:2601.21135*, 2026.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in neural information processing systems*, 33:12615–12625, 2020.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and Yongjun Xu. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4):1–38, 2024.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Haochang Hao, Nour Elhendawy, Yihang Wang, and Lu Cheng. Confdock: Atom-specific uncertainty quantification for molecular docking via conformal prediction. *bioRxiv*, pages 2026–06, 2026.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pages 308–334. Springer, 1992.

Yifan Hu, Jie Yang, Xilin Dai, Wanxu Cai, Kuiye Ding, Yuante Li, Qinghua Liu, Enze Ma, Zhiyuan Qu, Yixin Wang, et al. The landscape of agentic time series systems: Architectures, reliability, and frontiers.

Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. *arXiv preprint arXiv:2509.14181*, 2025.

Boyin Huang, Peter W. Thorne, Viva F. Banzon, Tim Boyer, Gennady Chepurin, Jay H. Lawrimore, Matthew J. Menne, Thomas M. Smith, Russell S. Vose, and Huai-Min Zhang. NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5. Dataset, 2017a. Accessed: 2026.

Boyin Huang, Peter W Thorne, Viva F Banzon, Tim Boyer, Gennady Chepurin, Jay H Lawrimore, Matthew J Menne, Thomas M Smith, Russell S Vose, and Huai-Min Zhang. Extended reconstructed sea surface temperature, version 5 (ersstv5): upgrades, validations, and intercomparisons. *Journal of climate*, 30(20):8179–8205, 2017b.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. doi: 10.1016/S0893-6080(98)00140-3.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

J. H. King and N. E. Papitashvili. Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, 110(A2), 2005. doi: https://doi.org/10.1029/2004JA010649. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010649`.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporally causal representation with instantaneous dependence. *arXiv preprint arXiv:2405.15325*, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):5, 2018.

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1), 2013.

Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44): 17845–17850, 2012.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. *Advances in Neural Information Processing Systems*, 37:77098–77131, 2024.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13), 2021.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06221-2. URL `http://dx.doi.org/10.1038/s41586-023-06221-2`.

Kun Wang, Sumanth Varambally, Duncan Watson-Parris, Yi-An Ma, and Rose Yu. Discovering latent causal graphs from spatiotemporal data. *arXiv preprint arXiv:2411.05331*, 2024.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, and Hongwei Wang. GAM: Hierarchical graph-based agentic memory for LLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 34647–34664, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1600. URL https://aclanthology.org/2026.acl-long.1600/.

Jie Yang, Yifan Hu, Kexin Zhang, Luyang Niu, Philip S Yu, and Kaize Ding. Revisiting multivariate time series forecasting with missing values. *arXiv preprint arXiv:2509.23494*, 2025.

Jie Yang, Yifan Hu, Yuante Li, Kexin Zhang, Kaize Ding, and Philip S Yu. From observations to states: Latent time series forecasting. *arXiv preprint arXiv:2602.00297*, 2026a.

Jie Yang, Kexin Zhang, Guibin Zhang, Philip S Yu, and Kaize Ding. Glocal information bottleneck for time series imputation. *Advances in Neural Information Processing Systems*, 38:104452–104484, 2026b.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

Fangxu Yu, Tao Feng, Dehai Min, Lu Cheng, Ge Liu, and Tianyi Zhou. Tsrouter: Dynamic modality-model selection for time series reasoning. *arXiv preprint arXiv:2607.08940*, 2026.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=V4y0CpX4hK.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://openreview.net/forum?id=Wo1HF2wWNZb.

Henry Peng Zou, Chunyu Miao, Wei-Chieh Huang, Yankai Chen, Yue Zhou, Hanrong Zhang, Yaozu Wu, Liancheng Fang, Zhengyao Gu, Zhen Zhang, Kening Zheng, Fangxin Wang, Yi Nian, Shanghao Li, Wenzhe Fan, Langzhou He, Weizhi Zhang, Xue Liu, and Philip S. Yu. When users change their mind: Evaluating interruptible agents in long-horizon web navigation, 2026. URL https://arxiv.org/abs/2604.00892.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

A Architecture Details

Encoder. The encoder is a 4-layer MLP with LeakyReLU(0.2) activations and no normalization layers. The hidden dimension is domain-dependent (64 to 128); the final layer maps to $[\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2] \in \mathbb{R}^{2n}$, from which we reparametrize $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Stage 1 decoder (dense). A standard MLP $g_\theta : \mathbb{R}^n \to \mathbb{R}^D$ used during Stage 1 to learn identifiable latent factors. This decoder is discarded after Stage 1.

Stage 2 decoder (additive). Each of the n component functions f_j is a 2-layer MLP: $z_j \in \mathbb{R} \to \text{Linear}(1, H) \to \text{LeakyReLU} \to \text{Linear}(H, D)$, where H is the per-component hidden dimension (hidden_per_z). The total decoder output is $\hat{\mathbf{x}} = \sum_{j=1}^{n} f_j(z_j) + \mathbf{b}$, with a learnable bias $\mathbf{b} \in \mathbb{R}^D$. The encoder and transition prior are frozen during Stage 2; only the additive decoder parameters are optimized, using `AdamW` restricted to parameters with `requires_grad=True`.

Per-dimension residual prior. For dimension j, the Stage 1 prior assumes a Laplace density on the residual $r_{t,j} = h_j(\mathbf{z}_{t-L:t-1}, z_{t,j}, \mathbf{e}_{c_t})$, where h_j is a per-dimension transition MLP that receives the past L lag vectors, the current factor $z_{t,j}$ (but not $z_{t,k}$ for $k \neq j$), and the regime embedding. The prior density is

$$\log p_j(z_{t,j} \mid \mathbf{z}_{t-L:t-1}, c_t) = -\frac{|r_{t,j}|}{b_j} - \log(2b_j) + \log \left| \frac{\partial r_{t,j}}{\partial z_{t,j}} \right|, \tag{7}$$

where b_j is a learnable scale. Because h_j depends on $z_{t,j}$ through nonlinear layers, the Jacobian $\partial r_{t,j}/\partial z_{t,j}$ is not identically 1 and contributes a non-trivial log-determinant term, which is what motivates the batched autograd computation described below.

Training objectives. The Stage 1 loss combines reconstruction, a standard-Gaussian KL, and the temporal-prior KL,

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathrm{KL}\big(q_\phi(\mathbf{z}_t \mid \mathbf{x}_t) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) + \gamma \cdot \mathrm{KL}\big(q_\phi(\mathbf{z}_t \mid \mathbf{x}_t) \,\|\, p(\mathbf{z}_t \mid \mathbf{z}_{t-L:t-1}, c_t)\big), \tag{8}$$

with β weighting the standard-Gaussian KL and γ weighting the temporal-prior KL, evaluated for $t \geq L$. The two KL terms serve distinct roles: the standard-Gaussian term ($\beta = 2 \times 10^{-3}$) acts as an empirical regularizer that bounds latent magnitudes and prevents posterior collapse, while the temporal-prior term ($\gamma = 2 \times 10^{-2}$) supplies the regime-conditioned nonstationarity signal required for identifiability. This objective is not a single-prior ELBO; it follows the dual-regularization design used in prior temporal CRL implementations [Yao et al., 2022, 2021].

The Stage 2 loss is $\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{recon}} + \lambda(t) \cdot \mathcal{L}_{\text{sparse}}$. Because the encoder and prior are frozen, the temporal-prior KL is constant with respect to the additive decoder parameters and contributes no gradients; only reconstruction and the sparsity penalty shape the decoder. $\lambda(t)$ follows a three-stage schedule: zero during a 5-epoch warmup, linear ramp to $\lambda_{\max} = 50$ over the next 20 epochs, and constant $\lambda_{\max}$ thereafter. The active-factor mask $\mathcal{J}_{\text{alive}} = \{j : \sum_i A_{ij} > 0.01 \cdot \max_k \sum_i A_{ik}\}$ excludes near-dead factors from the entropy penalty in Eq. (6).

Transition prior. The parallel MLP takes the lagged latent history $\mathbf{z}_{t-L:t-1}$ together with the regime embedding $\mathbf{e}_{c_t}$ as input and outputs the per-dimension transition shift. For each target dimension j, the corresponding input width is $Ln + 1 + H_{\text{emb}}$, where H_{emb} is the regime-embedding hidden dimension, and all n dimensions are processed simultaneously via batched `einsum`: $\mathbf{h} = \sigma(\text{einsum}(\texttt{'bnf,nfh->bnh'}, \text{input}, \mathbf{W}_1) + \mathbf{b}_1)$, then $\mathbf{r} = \text{einsum}(\texttt{'bnh,nh->bn'}, \mathbf{h}, \mathbf{W}_2) + \mathbf{b}_2$.

The log-det-Jacobian required by the change-of-variables formula reduces to the diagonal $\partial r_t^{(j)}/\partial z_t^{(j)}$ for each j. Because $z_t^{(j)}$ enters only into $r_t^{(j)}$ in the additive prior, the scalar derivative equals the gradient of $\sum_{b,j} r_{b,j}$ with respect to the corresponding input coordinate, which we obtain in a single `torch.autograd.grad` call. This avoids instantiating the full $(BW, 1, BW, F)$-shaped Jacobian that TDRL's reference implementation builds per dimension.

Asymptotic complexity. Let $BW = B \cdot (T-L)$ denote the total number of time steps per batch, and $F = H_{\text{emb}} + Ln + 1$ the input width per transition MLP, where H_{emb} is the regime-embedding hidden dimension. Table 5 compares the two implementations along four axes.

Table 5: Asymptotic complexity of the transition prior. v1 = TDRL's reference implementation (sequential loop + `autograd.functional.jacobian`); v2 = ours (parallel MLP + scalar-sum autograd). The dominant savings are the elimination of the BW factor in the Jacobian cost and the elimination of $O(n)$ sequential GPU kernel launches. Here n denotes the latent dimension; D refers exclusively to the observation dimension throughout. The forward FLOPs include the input-to-hidden ($F \to H$) and hidden-to-hidden ($H \to H$) layers of the 3-layer transition MLP.

Quantity	v1 (TDRL)	v2 (ours)	Ratio
Forward FLOPs	$n \cdot BW \cdot (FH + H^2)$ (sequential)	$n \cdot BW \cdot (FH + H^2)$ (batched)	$1\times$ same work
Jacobian FLOPs	$n \cdot BW^2 \cdot F \cdot H$	$n \cdot BW \cdot F \cdot H$	$BW\times$ savings
GPU kernel calls	$O(n \cdot L)$ sequential	$O(L)$ batched	$n\times$ launch reduction
Peak memory	$O(n \cdot BW^2 \cdot F)$	$n \cdot BW \cdot F$	$BW\times$ savings

Two mechanisms account for the speedup. First, `autograd.functional.jacobian` constructs the full $(BW, 1, BW, F)$-shaped Jacobian even though only the BW diagonal entries are needed,

inflating both FLOPs and memory by a factor of BW; the scalar-sum trick extracts the diagonal directly in $O(BW)$ work. Second, the Python `for` loop over n latent dimensions serializes n small kernel launches, each under-utilizing the GPU; `einsum` fuses them into a single launch. The asymptotic speedup is therefore $\sim n \cdot \min(BW, \text{GPU-utilization gap})$, predicting that the ratio grows with both n and batch size, a prediction verified in the benchmark below.

Table 6: Transition prior wall-clock benchmark on NVIDIA A40 (30-iteration mean, 5 warmup iterations). Same symbols as Table 5; Z denotes the latent dimension (n in the complexity notation).

Configuration	Z	B	T	H	v1 time	v2 time	Speedup	v1 mem	v2 mem	Mem ratio
Synthetic	8	256	3	128	1635 ms	2.9 ms	563×	909 MB	38 MB	24×
RNA	4	256	3	128	814 ms	2.9 ms	277×	511 MB	27 MB	19×
Cross-domain	8	256	3	128	1637 ms	2.9 ms	560×	909 MB	38 MB	24×
Stress test	64	256	3	64	7327 ms	3.0 ms	2445×	3303 MB	122 MB	27×

The empirical scaling matches the complexity analysis: v1 runtime grows linearly with Z ($814 \to 1637 \to 7327$ ms as Z goes $4 \to 8 \to 64$, consistent with the $D\times$ kernel-launch overhead), while v2 runtime is essentially constant ($\sim$3 ms) because the einsum does not yet saturate the GPU at these sizes. The memory ratio of $\sim 24\times$ across configurations matches the predicted $BW = 256 \times (T-L) \approx 256$ savings, modulated by the width factor F. Both optimizations are algorithmic (not approximations): v2 computes the exact same log-det-Jacobian as v1, verified numerically to $2.6 \cdot 10^{-4}$ max absolute difference. End-to-end, synthetic training at the speedup-benchmark setting (Z=8, 80 epochs, $\sim$128K steps) drops from $\sim$65 hours to $\sim$90 minutes per seed. The main synthetic experiments use $\hat{n} = 12$ rather than the $Z = 8$ benchmark setting, so their wall-clock is moderately larger: the main $\hat{n} = 12$ synthetic setting takes on the order of 1–2 hours per seed, which is what makes the 5-seed cross-ablation experimental design feasible on a single A40.

B Hyperparameters

Table 7: Hyperparameter settings. All experiments share $\beta = 2\times10^{-3}$ (past-KL weight), $\gamma = 2\times10^{-2}$ (future-KL weight), lr $= 5\times10^{-4}$, weight decay $= 10^{-4}$, AdamW optimizer. Stage 1 uses a dense decoder; Stage 2 freezes the encoder and uses an additive decoder with the entropy-based sparsity penalty of Eq. (6).

Hyperparameter	Synthetic	RNA	OMNI	Disruption	Climate	TEP
Input dim D	30	28	21	12	47	52
Learned latent dim $\hat{n}$ (z_{dim})	12	4	5	5	8	8
Encoder hidden	128	128	64	64	128	128
Stage 2 H (hidden/z)	4	4	4	4	4	4
$\lambda_{\max}$ (sparsity)	50	50	50	50	50	50
Stage 1 epochs	80	80	80	80	80	80
Stage 2 epochs	80	80	80	80	80	80
Batch size	256	256	256	256	256	256

B.1 Experimental Protocol Details

Baselines include TDRL [Yao et al., 2022], CtrlNS [Song et al., 2024], β-VAE ($\beta = 4, \gamma = 0$), Sparse PCA ($\alpha = 10$), ICA, and PCA. Regime-association ranking uses a shared Cohen's d pipeline across all methods. For methods without additive decoders, localization is read from native structure: linear loadings for Sparse PCA, ICA, and PCA, and dense-decoder Jacobians for TDRL and β-VAE. For the latter, the per-sample Jacobian $\partial\hat{\mathbf{x}}_t/\partial\hat{\mathbf{z}}_t$ varies across the dataset; we aggregate by taking the mean of element-wise absolute values over all samples, $A_{ij} = \frac{1}{N}\sum_t |(\partial\hat{\mathbf{x}}_t/\partial\hat{\mathbf{z}}_t)_{ij}|$, to form the influence matrix. For SPIB, which lacks a generative decoder, we extract variable influence via a per-observation input-feature Lasso: for each learned latent $\hat{z}_k$, we fit $\hat{z}_k = \sum_i w_{ik}x_i + b_k$ with ℓ_1 regularization and set $A_{ik} = |w_{ik}|$.

Metric definitions. Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be the true latent factors ($M = 6$ on the synthetic benchmark: $z_0, z_1, z_2, z_3, u_1, u_2$), partitioned into regime-varying factors $\mathcal{D} \subseteq \mathcal{F}$ (those whose distribution shifts across regimes; here $\mathcal{D} = \{z_0, z_1, z_2\}$) and regime-invariant factors. Each true factor f_j has a ground-truth observation support $\mathcal{S}_j = \mathcal{P}_j \cup \mathcal{H}_j \subseteq [D]$, where $\mathcal{P}_j$ is the primary support (observations generated only from f_j) and $\mathcal{H}_j$ is the shared support (observations generated from f_j and at least one other factor). Let $\hat{Z} = \{\hat{z}_1, \ldots, \hat{z}_{\hat{n}}\}$ be the learned latent factors ($\hat{n} \geq M$), and let

$$d(\hat{z}_k) \;=\; \frac{\big|\,\mathbb{E}[\hat{z}_k \mid c{=}0] - \mathbb{E}[\hat{z}_k \mid c{=}1]\,\big|}{\sqrt{\frac{1}{2}\big(\mathrm{Var}[\hat{z}_k \mid c{=}0] + \mathrm{Var}[\hat{z}_k \mid c{=}1]\big)}} \tag{9}$$

denote the Cohen's d of $\hat{z}_k$ across the binary regime label c. We also form an influence matrix $A \in \mathbb{R}_{\geq 0}^{D \times \hat{n}}$ where A_{ij} quantifies the contribution of $\hat{z}_j$ to observation x_i (method-specific: decoder probe for MOSAIC, Jacobian for TDRL/β-VAE, or $|w_{ij}|$ from a per-observation Lasso for linear baselines).

Matching via Hungarian assignment [Kuhn, 1955]. Given samples of $(\hat{\mathbf{z}}, \mathbf{z}_{\text{true}})$, we build the affinity matrix $C_{jk} = |\rho(f_j, \hat{z}_k)|$ using Pearson correlation and solve $\max_\pi \sum_j C_{j,\pi(j)}$ via the Hungarian algorithm. We then threshold the assignment: $\hat{z}_k$ is declared *matched* to f_j iff $\pi(j) = k$ *and* $C_{jk} \geq \tau$ (we use $\tau = 0.5$); otherwise $\hat{z}_k$ is UNMATCHED. Without τ, the Hungarian solution would always return a full permutation, so low-quality correlations near zero would falsely count as recovered factors.

MCC. The standard mean correlation coefficient over the thresholded Hungarian matching, averaged over matched pairs.

Z@top3 (regime-association recovery). Let $\mathcal{T}_3$ be the indices of the three learned latents with the largest Cohen's d. Then

$$\text{Z@top3} \;=\; \tfrac{1}{3}\,\big|\{\hat{z}_k \in \mathcal{T}_3 : \hat{z}_k \text{ is matched to some } f_j \in \mathcal{D}\}\big|. \tag{10}$$

We report Z@top3 as a fraction $a/3$, and call a run *perfect* if Z@top3 $= 3/3$. This asks whether the three most regime-discriminative learned factors each correspond to a true regime-associated factor.

X_Z@top3 (support recovery for regime-associated factors). For each $\hat{z}_k \in \mathcal{T}_3$, let $m(k) \in \mathcal{F} \cup \{\varnothing\}$ denote its Hungarian match. Define the top-3 mass ratio $\mu(k) = \sum_{i \in \mathrm{Top3}(A_{\cdot,k})} |A_{i,k}| \big/ \sum_{i=1}^{D} |A_{i,k}|$, measuring the fraction of the influence column concentrated in its top-3 entries. The per-latent precision is

$$p(k) \;=\; \begin{cases} \dfrac{\big|\,\mathrm{Top3}(A_{\cdot,k}) \cap \mathcal{S}_{m(k)}\,\big|}{3}, & m(k) \in \mathcal{D} \text{ and } \mu(k) \geq 0.50, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

The "@top3" suffix in the metric name refers to this fixed Top-3 selection. A hit on a shared observation in $\mathcal{S}_{m(k)}$ counts as a correct recovery regardless of which of the overlapping factors $m(k)$ is. Then

$$X_Z\text{@top3} \;=\; \tfrac{1}{3} \sum_{\hat{z}_k \in \mathcal{T}_3} p(k). \tag{12}$$

The concentration gate $\mu(k) \geq 0.50$ reflects the module-discovery objective: a regime-discriminative latent that spreads its influence uniformly across all observed variables, even when its top-3 argmax happens to overlap the true support, has not identified a module. The threshold 0.50 is permissive — MOSAIC satisfies it with mean top-3 mass 0.69 ± 0.03 on this benchmark (Table 15) — and penalizes only decoders substantially more dispersed than MOSAIC's. Appendix L reports sensitivity to threshold choice. Unmatched or regime-invariant latents contribute 0.

A note on the Hungarian threshold. The Hungarian threshold $\tau = 0.5$ is necessary because, without it, runs whose top-3 $\hat{z}$ by Cohen's d happen to correlate weakly-but-better-than-random with regime-associated factors would look artificially good.

Regime accuracy is a logistic probe on $\hat{\mathbf{z}}$. For X_Z@top3, when ground-truth support is not available, we recover per-latent support by the largest consecutive ratio gap in sorted A_{ij} values (threshold 1.5), with fallback threshold $0.1 \cdot \max_i A_{ij}$ when no gap is detected.

B.2 Cohen's d versus KL ranking

The theory defines the regime-discriminative coordinate using KL divergence because KL compares the full conditional distributions and is invariant under componentwise invertible reparametrizations. In experiments, we use Cohen's d as a stable finite-sample proxy for two-regime mean-shift contrasts. This proxy is not part of the identifiability proof.

As a diagnostic, we compared Cohen's d against a one-dimensional histogram estimator of forward KL on the evaluated Synthetic, RNA, OMNI, and TEP runs. The KL estimator uses a 64-bin shared-support histogram with $q_{0.005}/q_{0.995}$ trimming and Laplace smoothing, and we report forward KL $p_0 \| p_1$ as the primary ranking. Exact top-1 agreement is not guaranteed, but the top-set and module-level decisions are stable: the mean top-3 overlap is 86.1%, exact top-1 agreement holds in 8 of 12 evaluated runs, and the selected-module question agrees in 9 of 12 runs (Table 8). Climate was omitted because its evaluation pipeline differs from the other datasets and was not compatible with the shared diagnostic. We therefore use Cohen's d only as a finite-sample ranking proxy for the experimental protocol, not as a replacement for the KL-based identifiability statement.

A broader limitation applies to both Cohen's d and the marginal KL criterion of Definition 2: they target marginal distributional shifts, which is sufficient but not necessary for regime-varying dynamics. A factor whose transition dynamics change across regimes while its stationary marginal remains invariant would not be detected by either criterion. In the scientific settings we evaluate, regime contrasts (folded vs. unfolded, storm vs. quiet, faulty vs. normal) produce clear marginal mean shifts, so this gap does not arise empirically. For regime changes that manifest primarily as variance or higher-order moment shifts, Cohen's d would need to be replaced by a nonparametric two-sample statistic such as an MMD or energy distance; all six datasets in our evaluation are mean-shift dominated.

Table 8: Diagnostic comparison of Cohen's d and one-dimensional histogram forward-KL ranking.

Scope	Top-1 agreement	Mean top-3 overlap	Module agreement
12 evaluated runs	8/12	86.1%	9/12

C Synthetic Benchmark Details

Generative process. The synthetic benchmark generates $D = 30$ observations from $n_{\text{true}} = 6$ latent factors via monotonic invertible nonlinear mixing. Each observation channel $i \in [D]$ is assigned a primary parent $j \in [n_{\text{true}}]$ and a mixing function g_{ij} drawn (cyclically across i) from

$$\mathcal{F} = \big\{\, z \mapsto \tanh(5z),\ z \mapsto z^3/4,\ z \mapsto \operatorname{erf}(3z),\ z \mapsto \arctan(5z),\ z \mapsto \operatorname{sign}(z)\,|z|^{1/3} \big\}. \quad (13)$$

The observation is $x_i = \sum_{j \in S_i} g_{ij}(z_j) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The five functions span saturating ($\tanh$, $\arctan$), expanding ($z^3/4$, $\operatorname{sign} \cdot |z|^{1/3}$), and symmetric sigmoidal (erf) shapes, defeating any single linear projection that simultaneously fits all channels generated by one factor.

Linear-correlation calibration. The generator is calibrated so that each observation remains strongly associated with its assigned parent latent while limiting spurious cross-factor linear correlations; exact calibration statistics depend on the synthetic variant and are not used in the evaluation metrics. For the main synthetic benchmark used in Table 1, the generator produces a near-linear parent association (mean parent $|r| \approx 0.99$) before nonlinear monotonic transformations, while support recovery is evaluated against the known parent/support map rather than against correlation values. After applying the nonlinear transforms in $\mathcal{F}$, the channel-parent linear correlation is substantially reduced, consistent with the MCC ceiling of 0.68–0.78 achieved by linear baselines (Table 1).

Latent dynamics. The latent factors evolve under a fixed potential with regime-dependent coupling (full specification in the release code). Regime labels are assigned by the well occupancy of the reaction-coordinate factor, yielding 420 K balanced samples (210 K per regime).

Ground-truth support structure. The 30 observed channels are partitioned into 6 ground-truth factors with 3 additional pure-noise channels. Table 9 lists the primary and shared assignments used to score module recovery. Critically, the benchmark includes shared-channel overlaps (channels 4,

8, 21) by design: the strict disjoint-support assumption A1′ used in Theorem 1 and Theorem 2 is unrealistic for real scientific data, where physical observables routinely couple to multiple latent mechanisms. The shared channels deliberately violate A1′, placing the benchmark in the relaxed partial-overlap regime characterized by Proposition 2. The MCC, X_Z@top3, and module-recovery results reported in Section 6.1 should therefore be read as evidence that MOSAIC operates beyond the strict A1′ setting and recovers structure under the more realistic partial-overlap conditions that Proposition 2 covers.

Table 9: Ground-truth support structure for the synthetic benchmark. The combined support $\mathcal{S}_{m(k)} = \mathcal{P}_{m(k)} \cup \mathcal{H}_{m(k)}$ (primary plus shared channels) is used to score Top-3 recovery in X_Z@top3; shared channels count as hits for either listed factor. Channels 27, 28, and 29 are pure noise.

Factor	Role	Primary channels	Shared channels
z_0	Regime-varying	$\{0, 1, 2, 3\}$	$\{4\}$ with z_2
z_1	Regime-varying	$\{5, 6, 7\}$	$\{8\}$ with z_3
z_2	Regime-varying	$\{9, 10, 11, 12\}$	$\{4\}$ with z_0
z_3	Invariant	$\{13, 14, 15, 16\}$	$\{8\}$ with z_1
u_1	Invariant	$\{17, 18, 19, 20\}$	$\{21\}$ with u_2
u_2	Invariant	$\{22, 23, 24, 25, 26\}$	$\{21\}$ with u_1

Module recovery thresholding. For each learned latent z_j, we sort A_{ij} in descending order and cut at the largest consecutive ratio gap exceeding 1.5; if no such gap exists in the top 15 entries, we fall back to the relative threshold $0.1 \cdot \max_i A_{ij}$. This procedure is parameter-free per column and adapts to different sparsity levels across factors.

C.1 Interaction-strength sweep

We sweep the interaction strength α from 0 to 2.0, adding pairwise cross-factor interaction terms $\alpha \cdot \tanh(z_a \cdot z_b)$ while preserving the same main effects and regime labels (5 seeds per α, z_dim=12). This probes Theorem 2's effective-noise picture: larger interaction residuals should make support recovery more sample-demanding.

Table 10: Interaction strength sweep on the synthetic benchmark (5 seeds per α, z_dim=12). As α increases from 0 (additive, paper assumption) to 2 (strong non-additive coupling), support recovery X_Z degrades monotonically from 0.978 to 0.538, while MCC remains comparatively stable. The $\alpha = 0$ row is the additive-mixing limit, trained and evaluated on a separate α-sweep dataset using the same protocol as Table 1. The MCC gap (0.934 vs Table 1's 0.912) reflects the two independent training trajectories; the per-seed X_Z pattern (four perfect seeds plus one at $8/9 = 0.889$) coincides between the two runs under the discrete fixed-Top-3 metric, hence the matching aggregate 0.978 ± 0.044.

α	MCC ↑	Z@top3 perfect (out of 5) ↑	X_Z@top3 (gate 0.50) ↑
0.0	0.934 ± 0.006	5/5	0.978 ± 0.044
0.1	0.936 ± 0.005	5/5	0.911 ± 0.130
0.3	0.920 ± 0.021	4/5	0.744 ± 0.276
0.5	0.941 ± 0.011	4/5	0.713 ± 0.216
0.7	0.945 ± 0.014	4/5	0.683 ± 0.178
1.0	0.894 ± 0.045	3/5	0.651 ± 0.286
1.5	0.907 ± 0.037	3/5	0.544 ± 0.070
2.0	0.889 ± 0.041	3/5	0.538 ± 0.166

The sweep is consistent with Theorem 2's effective-noise picture: as α increases, the non-additive variance contribution grows and the additive decoder's support recovery becomes harder, while the encoder's latent recovery (MCC) remains robust in a wider range. Support recovery leaves the high-reliability range between $\alpha = 0.1$ and $\alpha = 0.3$, then degrades more gradually, falling below 0.6 only at the strongest interaction levels ($\alpha \geq 1.5$).

C.2 Sample-complexity sweep (N/D)

We subsample the synthetic benchmark by retaining every k-th frame (preserving temporal ordering) to vary N/D from 5 to 5000, training MOSAIC with the same hyperparameters across all levels (3 seeds per level). The main benchmark uses the full $N = 420$K samples; the subsampled levels below use up to $N = 150$K.

Table 11: N/D sweep on the synthetic benchmark (3 seeds per level). MCC and X_Z@top3 are seed means $\pm$ std.

N/D	N	MCC ↑	Regime Acc	X_Z@top3 ↑
5	150	0.558 ± 0.051	0.938 ± 0.050	0.022 ± 0.039
10	300	0.568 ± 0.079	0.956 ± 0.017	0.044 ± 0.077
25	750	0.561 ± 0.066	0.998 ± 0.003	0.044 ± 0.077
50	1500	0.663 ± 0.038	0.999 ± 0.001	0.072 ± 0.075
100	3000	0.657 ± 0.061	0.999 ± 0.001	0.219 ± 0.038
250	7500	0.688 ± 0.043	1.000	0.089 ± 0.102
500	15000	0.727 ± 0.044	1.000	0.503 ± 0.067
1000	30000	0.728 ± 0.005	1.000	0.724 ± 0.164
5000	150000	0.814 ± 0.048	1.000	0.830 ± 0.058

Two patterns emerge. First, MCC climbs smoothly from 0.56 to 0.81 with no phase transition, indicating that latent identifiability degrades gradually at low N. Second, X_Z@top3 exhibits a sharper transition: near zero for $N/D \leq 50$, rising through 0.50 around N/D=500, and reaching 0.83 at N/D=5000. Support recovery is more data-hungry than latent recovery, consistent with Theorem 2's prediction that the sample requirement scales with the signal-to-noise ratio $(\sigma_\epsilon^2 + \sigma_{r,\max}^2)/\sigma_{\min}^2$, which is a fixed population-level quantity that determines the N threshold for reliable recovery.

Regime accuracy saturates at $N/D \geq 25$ (above 0.998), confirming that regime-label signal is not the bottleneck. Note that temporal subsampling reduces not only N but also the density of temporal transitions, which may additionally degrade the temporal prior's nonstationarity signal at very low N/D.

C.3 Case study: regime definition determines the scientific question

The regime label encodes the scientific question that MOSAIC answers. We demonstrate this on the synthetic benchmark by applying two regime definitions to the same data, each targeting a different physical contrast, and showing that MOSAIC returns a different (and correct) regime-associated factor in each case.

Definition A: which latent modules distinguish State A from State B? Regime labels are assigned by well occupancy (State A vs. State B). Under this definition, MOSAIC recovers the full module structure (MCC $\approx$ 0.85, X_Z@top3 $\approx$ 0.88) and correctly identifies the reaction-coordinate factors as regime-varying. This is the standard setting reported in Section 6.1.

Definition B: what signals precede a transition from A to B? We redefine regimes as stable-A (particle in well A with no transition within 50 frames) versus pre-transition-A (particle in well A but about to transition within 50 frames). This asks a fundamentally different question: not "what differs between states" but "what predicts an upcoming transition."

Under Definition B, MCC drops from 0.85 to 0.565 ± 0.042 and X_Z@top3 drops from $\sim$0.88 to 0.022 ± 0.039 (Table 12). This degradation is physically expected: the pre-transition signal depends on future behavior, but MOSAIC's temporal prior conditions only on past observations. The temporal prior can detect distributional differences between regimes that are visible in the current and past frames; it cannot anticipate future transitions that have not yet manifested in observable dynamics.

Takeaway. The contrast between Definitions A and B clarifies the scope of MOSAIC's regime-association concept: it identifies factors whose distributions differ most across the user-specified regimes, not factors that predict future regime transitions. This is analogous to the RNA cross-regime result (Section 6.2), where switching the regime definition from closing-pair contacts to outer-stem contacts shifts the top regime-associated factor from Loop to Stem. In both cases, MOSAIC faithfully

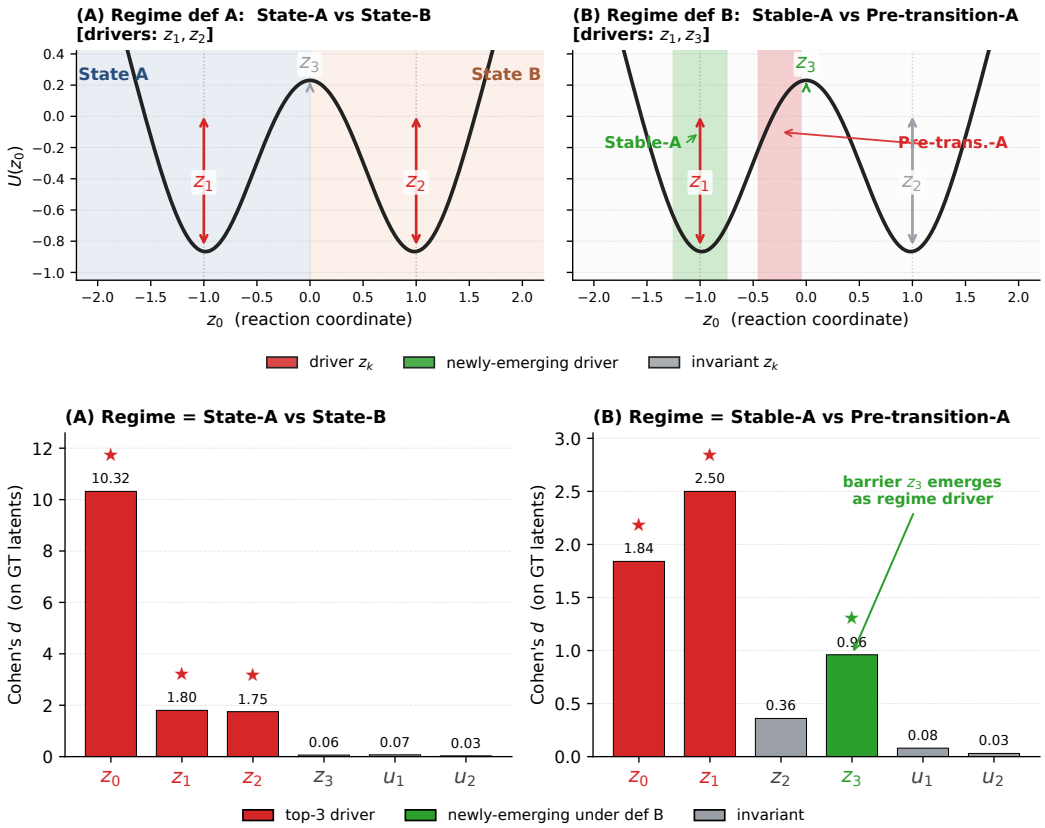

Figure 6: Regime definition determines the scientific question. **Top row**: the two regime partitions overlaid on the same parametric-coupling double-well, with z_1 (left-well depth), z_2 (right-well depth), and z_3 (barrier height) annotated in each panel; colour marks each factor as regime-varying (red), newly regime-associated (green), or invariant (gray) under the regime shown. (A) Def A splits the reaction coordinate into the left well (State A) and right well (State B); both well-depth factors z_1 and z_2 are regime-varying. (B) Def B splits the left well temporally into Stable-A (near the well bottom) and Pre-transition-A (approaching the barrier); the barrier-height factor z_3 becomes most regime-associated while z_2 becomes invariant. **Bottom row**: Cohen's d of each ground-truth latent confirms the shift. The same latent system yields a different regime-association set because the scientific question encoded in the regime label has changed.

Table 12: Regime Definition B (pre-transition labeling) on the synthetic benchmark, 3 seeds. Definition A (well-occupancy) achieves MCC ≈ 0.85, X_Z@top3 ≈ 0.88.

Seed	Regime Acc	MCC	Z@top3	X_Z@top3
0	0.901	0.528	0/3	0.000
42	0.931	0.609	2/3	0.067
123	0.936	0.559	1/3	0.000
Mean	0.923	0.565	∼1/3	0.022

answers the question encoded in the regime label. Users should choose regime definitions that reflect the scientific contrast of interest.

C.4 Unsupervised regime discovery via k-means

MOSAIC requires binary regime labels. We test whether unsupervised regime discovery can substitute for ground-truth labels by running k-means (k=2) on the raw observations to generate pseudo-regime labels, then training MOSAIC with these labels (3 seeds).

Table 13: Auto-regime on the synthetic benchmark: k-means pseudo-labels vs. ground-truth labels. NMI and agreement are between pseudo and GT labels.

Seed	NMI	Agreement	MCC	Z@top3	X_Z@top3
0	0.959	99.5%	0.877	3/3	0.717
42	0.959	99.5%	0.858	2/3	0.600
123	0.959	99.5%	0.823	2/3	0.600
Mean			0.853 ± 0.027		0.639 ± 0.068

On the synthetic benchmark, k-means nearly perfectly recovers the ground-truth regime (NMI = 0.959, 99.5% agreement), and MOSAIC trained on pseudo-labels achieves MCC = 0.853, matching the supervised result ($\sim$0.85). X_Z@top3 is lower (0.64 vs. $\sim$0.88), possibly because minor label noise at regime boundaries degrades the sparsity penalty signal. This validates unsupervised regime discovery on well-separated synthetic data; real-data domains with less clear-cut regime structure may require iterative label refinement.

D Theory Proofs and Finite-Sample Rates

D.1 Population Identifiability Proofs

Proof of Proposition 1 (Main-Effect Identifiability Without False Positives). The ANOVA decomposition $g(\mathbf{z}) = g_0 + \sum_j g_j(z_j) + r(\mathbf{z})$ is the unique $L^2(\mu)$-orthogonal expansion satisfying $\mathbb{E}_{\mu_j}[g_j(z_j)] = 0$ and r orthogonal to every univariate function of a single z_j [Hoeffding, 1992]. Hence the population MSE decomposes as $\mathbb{E}_\mu\|\mathbf{x} - \hat{g}(\mathbf{z})\|^2 = \mathbb{E}\|\boldsymbol{\epsilon}\|^2 + \sum_j \mathbb{E}_{\mu_j}\|g_j(z_j) - \hat{f}_j(z_j)\|^2 + \mathbb{E}_\mu\|r(\mathbf{z})\|^2$, with no cross terms. The sum is minimized independently in each $\hat{f}_j$ by $\hat{f}_j = g_j$, and the residual $\mathbb{E}_\mu\|r\|^2$ is irreducible by any additive model. Therefore $\hat{f}_j^{(i)}$ is non-constant iff $g_j^{(i)}$ is non-constant, so $\hat{\mathcal{S}}_j = \mathcal{S}_j \subseteq \mathcal{S}_j^{\text{full}}$, where $\mathcal{S}_j^{\text{full}}$ denotes the full support of g with respect to z_j. □

Proof of Theorem 1 (Module Structure Identifiability). Under the theorem's hypothesis both models have $\rho = 0$, so $g(\mathbf{z}) = g_0 + \sum_j g_j(z_j)$ and $\tilde{g}(\mathbf{z}) = \tilde{g}_0 + \sum_k \tilde{g}_k(\tilde{z}_k)$ are exactly additive.

Step 1 (Latent identifiability). By Assumption A2 and the identifiability result of Yao et al. [2022] (Theorem 1), the latent representations of the two models are related by $\tilde{z}_k = \phi_{\pi(k)}(z_{\pi(k)})$ for all $k \in [n]$, where π is a permutation and each $\phi_j : \mathbb{R} \to \mathbb{R}$ is a smooth diffeomorphism.

Step 2 (Additive structure constrains the mapping). At the population optimum the two models have the same conditional mean, so $\sum_{j=1}^n g_j(z_j) + g_0 = \sum_{k=1}^n \tilde{g}_k(\tilde{z}_k) + \tilde{g}_0$. Substituting $\tilde{z}_k = \phi_{\pi(k)}(z_{\pi(k)})$ and re-indexing via $k = \pi^{-1}(j)$ gives two additive decompositions of the same function of $\mathbf{z}$. After absorbing component means into the intercept, the centered Hoeffding decomposition under the product reference measure is unique. Therefore $g_j(z_j) = \tilde{g}_{\pi^{-1}(j)}(\phi_j(z_j)) + c_j$ for all $j \in [n]$, where $c_j \in \mathbb{R}^D$ are constants satisfying $\sum_j c_j = \tilde{g}_0 - g_0$.

Step 3 (Support preservation). Examining coordinate i: $g_j^{(i)}(z_j) = \tilde{g}_{\pi^{-1}(j)}^{(i)}(\phi_j(z_j)) + c_j^{(i)}$. If $i \in \mathcal{S}_j$ (i.e., $g_j^{(i)}$ is non-constant), then since ϕ_j is a diffeomorphism, the composition $\tilde{g}_{\pi^{-1}(j)}^{(i)} \circ \phi_j$ is also non-constant, implying $i \in \tilde{\mathcal{S}}_{\pi^{-1}(j)}$. The converse follows by applying the same argument with ϕ_j^{-1}. Therefore $\mathcal{S}_j = \tilde{\mathcal{S}}_{\pi^{-1}(j)}$ for all j. □

Proof of Corollary 1 (Regime-Association Identifiability). By Step 1 of Theorem 1, $\tilde{z}_k = \phi_{\pi(k)}(z_{\pi(k)})$ where each ϕ_j is invertible. The KL divergence is invariant under invertible transformations: $\mathrm{KL}(p(\tilde{z}_k \mid c{=}0)\|p(\tilde{z}_k \mid c{=}1)) = \mathrm{KL}(p(z_{\pi(k)} \mid c{=}0)\|p(z_{\pi(k)} \mid c{=}1))$. Therefore the ranking, and hence the argmax set, is preserved under π^{-1}; if the maximizer is unique, $\tilde{j}^* = \pi^{-1}(j^*)$. □

D.2 Product Measure Versus Data Distribution

Proposition 1 defines main-effect identifiability under the product reference measure $\mu = \prod_j \mu_j$, while the Stage 2 decoder is optimized under the empirical distribution induced by the frozen encoder. Two observations bridge this gap.

First, the ANOVA main effects $\{g_j\}$ are population-level objects uniquely determined by g and μ; they serve as *targets* for the additive decoder rather than quantities that the decoder defines.

Second, by the time Stage 2 begins, Stage 1 identifiability (Assumption A2) has already pushed the encoder toward a representation whose marginal factorizes across dimensions up to permutation and component-wise transformation. Any residual statistical dependence among the encoded coordinates is absorbed by the interaction term $r(\mathbf{z})$ in the ANOVA expansion, leaving the per-component main effects g_j unchanged. The practical consequence is that the Stage 2 additive decoder targets the correct main effects to the extent that the encoder's posterior marginals approximate independent factors, a condition that improves as Stage 1 identifiability tightens.

We emphasize that this product-versus-empirical gap is not specific to MOSAIC: all temporal CRL methods that invoke nonlinear ICA identifiability face the same discrepancy between the theoretical reference measure and the learned encoder distribution. Proposition 1 is a statement about the population optimum under μ; the finite-sample gap between μ and the empirical encoder distribution is part of the overall estimation error characterized empirically in Appendices C.2 and C.1.

D.3 Finite-Sample Rates for Sparse Additive Estimators

The population identifiability results above hold for any smooth estimator. The two results in this subsection, together with their proofs, instantiate finite-sample guarantees for one specific estimator class: the classical sparse additive model with B-spline main effects and group-lasso regularization [Ravikumar et al., 2009]. They serve as a reference point for what is achievable in the additive ANOVA framework, but do not directly govern the MLP-based decoder used by MOSAIC. The empirical finite-sample behavior of MOSAIC's implementation is characterized in Appendices C.2, C.1, and J; the design rationale and empirical comparison between entropy and group-lasso regularization are given in Appendix E.

Proposition 3 (Finite-Sample Main-Effect Estimation)**.** *Let $\hat{f}_j^{(N)}$ be the additive fit to N i.i.d. samples using a per-component B-spline basis with $m_N = O(N^{1/5})$. Under standard smoothness and sub-Gaussian design conditions [Ravikumar et al., 2009],*

$$\mathbb{E}_\mu \big\| \hat{f}_j^{(N),(i)} - g_j^{(i)} \big\|^2 = O(N^{-4/5}) \quad \forall j \in [n],\, i \in [D], \tag{14}$$

with constants depending on smoothness and on $\sigma_\epsilon^2 + \mathbb{E}_\mu \|r(\mathbf{z})\|^2$. This is the standard univariate spline rate for nonparametric regression. For time series, N refers to independent windows, or to the effective sample size under mixing.

Proof of Proposition 3. Apply Theorem 3 of Ravikumar et al. [2009] to each output channel $i \in [D]$ separately. Under sub-Gaussian design and the standard smoothness assumption that each $g_j^{(i)}$ lies in a second-order Sobolev class, spline smoothing with $m_N = O(N^{1/5})$ basis functions attains the univariate minimax rate $\mathbb{E}_\mu \|\hat{f}_j^{(N),(i)} - g_j^{(i)}\|^2 = O(N^{-4/5})$. The $L^2(\mu)$-orthogonality of the ANOVA decomposition makes the population MSE separable across channels. Each channel independently achieves the $O(N^{-4/5})$ rate, so the per-channel convergence exponent is unaffected by D; the total squared error summed over all D channels scales as $O(D \cdot N^{-4/5})$, but the rate exponent remains $-4/5$ per channel. The interaction residual $r(\mathbf{z})$ contributes to the effective noise variance $\sigma_\epsilon^2 + \mathbb{E}_\mu \|r\|^2$ by the same orthogonality argument used in the proof of Theorem 2. □

Theorem 2 (Finite-Sample Support Recovery, Group-Lasso Variant)**.** *Let $\sigma^2_{\min} = \min_{j, i \in \mathcal{S}_j} \mathrm{Var}_{\mu_j}[g_j^{(i)}(z_j)]$ denote the minimum main-effect signal, let $\sigma^2_{r,\max} = \max_{i \in [D]} \mathrm{Var}_\mu[r^{(i)}(\mathbf{z})]$ denote the maximum per-channel interaction variance, and let N denote the number of independent samples (or effective sample size under mixing). Under Assumption A1′, sub-Gaussian design, and the sparse-additive incoherence and restricted eigenvalue conditions of Ravikumar et al. [2009] applied per output channel with $m_N = O(N^{1/5})$ B-spline*

basis functions, the group-lasso-regularized additive decoder recovers $\mathcal{S}_j$ for all $j \in [n]$ with high probability provided

$$N \gtrsim C(s_{\max}, n, m_N) \cdot \frac{\sigma_\epsilon^2 + \sigma_{r,\max}^2}{\sigma_{\min}^2}, \tag{15}$$

where $s_{\max} = \max_{i \in [D]} |\{j : i \in \mathcal{S}_j\}|$ is the maximum number of active factors per output channel and $C(s_{\max}, n, m_N)$ collects the sparsity- and basis-dependent constants from Ravikumar et al. [2009]. Under A1′, $s_{\max} = 1$ for all primary-support channels.

The D output channels share the same N design points (the latent samples $\hat{\mathbf{z}}_t$), so the sample complexity does not scale with D; it is governed by the per-channel signal-to-noise ratio $\sigma_{\min}^2/(\sigma_\epsilon^2 + \sigma_{r,\max}^2)$ and the sparsity level $s_{\max}$. Higher-order interactions inflate $\sigma_{r,\max}^2$ and therefore the sample requirement, while stronger main-effect signals $\sigma_{\min}^2$ relax it.

Proof sketch of Theorem 2. We apply the support recovery analysis of Ravikumar et al. [2009] per output channel. For each channel $i \in [D]$, write the additive regression as $x_i = \sum_j g_j^{(i)}(z_j) + \eta_i$, where $\eta_i = r^{(i)}(\mathbf{z}) + \epsilon_i$ is the effective residual. Two properties make this reduction valid. First, $L^2(\mu)$-orthogonality of the ANOVA decomposition ensures $\mathbb{E}_\mu[\eta_i \cdot h(z_j)] = \mathbb{E}_\mu[\epsilon_i] \cdot \mathbb{E}_\mu[h(z_j)] = 0$ for any univariate h with $\mathbb{E}_{\mu_j}[h] = 0$, so η_i is uncorrelated with the additive regressors under the product measure. Second, because $r^{(i)}$ is a smooth bounded function of $\mathbf{z}$ and ϵ_i is Gaussian, the sum η_i is sub-Gaussian with parameter bounded by $\sigma_\epsilon^2 + \mathrm{Var}_\mu[r^{(i)}(\mathbf{z})]$, provided the latent distribution μ has sub-Gaussian marginals (a condition satisfied by the Laplace transition prior). We note that η_i is not independent of $\mathbf{z}$, since $r^{(i)}$ is a deterministic function of $\mathbf{z}$; the application of Ravikumar et al. [2009] therefore requires their fixed-design variant or an additional assumption that the empirical covariance of the B-spline basis concentrates despite the dependence, which holds under the sub-Gaussian design condition stated in the theorem. Applying the per-channel support recovery bound, the worst-case channel determines the overall sample requirement, giving the stated bound with $\sigma_{r,\max}^2 = \max_i \mathrm{Var}_\mu[r^{(i)}(\mathbf{z})]$. At $\rho = 0$, $r \equiv 0$ and the bound reduces to the standard sparse additive model guarantee with Gaussian noise alone. Assumption A1′ ensures $s_{\max} = 1$ for all primary-support channels, simplifying the incoherence requirement. Proposition 2 below records the hard-assignment behavior when supports overlap. □

D.4 Shared-Support Assignment and Entropy-Variant Lemmas

Proof of Proposition 2. (i) By $L^2(\mu)$-orthogonality of the ANOVA decomposition applied per channel, the population MSE for channel i decomposes as $\sum_{j' : i \in \mathcal{S}_{j'}} \mathbb{E}_{\mu_{j'}} \|g_{j'}^{(i)} - \hat{f}_{j'}^{(i)}\|^2 + \text{const}$, minimized independently in each summand by $\hat{f}_{j'}^{(i)} = g_{j'}^{(i)}$. The argument is identical to Proposition 1, applied channel-wise and in parallel to all factors whose support contains i.

(ii) For any factor $\ell \notin \mathcal{J}_i$, $g_\ell^{(i)} \equiv 0$, so any valid influence score has $A_{i\ell} = 0$. If the in-support maximum is strictly positive and tie-breaking does not prefer zero-score factors, the argmax cannot select $\ell \notin \mathcal{J}_i$. Among overlapping factors, the rule selects the largest observed influence score by definition. The finite contrast $|f_m(+1)_i - f_m(-1)_i|$ used in MOSAIC is one such score when the contrast is non-degenerate; integrated or quantile contrasts can replace it for symmetric responses. □

Scores for symmetric responses. The finite contrast $|f_j(+1)_i - f_j(-1)_i|$ vanishes for even main effects such as $g_j^{(i)}(z_j) = a z_j^2$. Two alternative scores handle this case. The *variance score* $A_{ij}^{\mathrm{var}} = \mathrm{Var}_{z_j \sim \hat{\mu}_j}[f_j(z_j)_i]$, estimated from the empirical encoder marginal, is nonzero whenever $f_j^{(i)}$ is non-constant regardless of symmetry. The *range score* $A_{ij}^{\mathrm{range}} = \max_{z_j \in [\hat{q}_{0.05}, \hat{q}_{0.95}]} f_j(z_j)_i - \min_{z_j \in [\hat{q}_{0.05}, \hat{q}_{0.95}]} f_j(z_j)_i$ provides a robust finite-sample alternative. In our experiments, the monotonic nonlinear mixing functions (Section 6.1) and the physical distance features (Section 6.2) produce non-symmetric main effects, so the default finite contrast suffices; the variance and range scores are provided for generality.

Implementation-relevant statements for the entropy penalty. The following lemma and proposition describe properties of the entropy-regularized implementation. They do not supply a finite-sample support-recovery theorem (that is an open problem; see Appendix E), but they record the weaker

facts that entropy biases the decoder toward concentrated influence columns and that exact top-k support recovery follows under an influence-gap condition on the estimated column.

Lemma 1 (Entropy Penalizes Diffuse Influence Columns)**.** *Let $a \in \mathbb{R}^D_{\geq 0}$ be a nonzero influence column and define $p_i = a_i / \sum_{\ell=1}^D a_\ell$. Then*

$$0 \leq H(p) \leq \log D,$$

where the lower bound is attained by one-hot columns and the upper bound is attained by the uniform column. Thus, for fixed column magnitude, minimizing $H(p)$ biases the decoder toward concentrated rather than diffuse influence columns.

Proof. This is the standard entropy bound on a discrete distribution over finite D elements. The minimum entropy is attained by any point mass, and the maximum entropy is attained by the uniform distribution. □

Proposition 4 (Top-k Support Consistency from Influence Separation)**.** *For a latent factor j, let $A^*_{:,j}$ be the population influence column induced by the additive main effect g_j, and let $\mathcal{S}_j$ be its support with $k_j = |\mathcal{S}_j|$. Assume an influence gap*

$$\Delta_j = \min_{i \in \mathcal{S}_j} A^*_{ij} - \max_{i \notin \mathcal{S}_j} A^*_{ij} > 0.$$

If an estimated additive decoder satisfies

$$\|\hat{A}_{:,j} - A^*_{:,j}\|_\infty < \Delta_j / 2,$$

then the top-k_j entries of $\hat{A}_{:,j}$ recover $\mathcal{S}_j$ exactly.

Proof. For any $i \in \mathcal{S}_j$ and $\ell \notin \mathcal{S}_j$, the influence gap definition gives $A^*_{ij} \geq \min_{i' \in \mathcal{S}_j} A^*_{i'j} = \max_{\ell' \notin \mathcal{S}_j} A^*_{\ell'j} + \Delta_j \geq A^*_{\ell j} + \Delta_j$. Combined with $\|\hat{A}_{:,j} - A^*_{:,j}\|_\infty < \Delta_j/2$:

$$\hat{A}_{ij} > A^*_{ij} - \Delta_j/2 \geq A^*_{\ell j} + \Delta_j - \Delta_j/2 = A^*_{\ell j} + \Delta_j/2 > \hat{A}_{\ell j}.$$

Thus every true-support coordinate has strictly larger estimated influence than every false-support coordinate, so selecting the top k_j entries recovers $\mathcal{S}_j$. □

E Entropy Sparsity Penalty: Mechanism and Design

The main text introduces the entropy penalty (Eq. 6) as MOSAIC's mechanism for recovering module structure, but defers the question of why entropy works and why alternative sparsity penalties do not. This appendix gives a mechanistic account in four parts: (i) what entropy regularizes once the influence column is normalized, (ii) why this design is compatible with reconstruction in a way that magnitude-based penalties are not, (iii) how entropy interacts with the alive-mask to handle over-capacity, and (iv) the role of entropy relative to the population identifiability results of Section 4. An important terminological distinction: Theorem 2 analyzes the *basis-function* group-lasso of Ravikumar et al. [2009], which groups B-spline coefficients for each (latent, channel) pair to perform per-channel variable selection. The column-level penalty compared empirically below, $\sum_j \|A_{:,j}\|_2$, groups across output channels for a given latent and targets a different object: within-column concentration of the influence matrix. Theorem 2 serves as a reference point for what the basis-function group-lasso achieves in the ANOVA framework; the empirical comparison below evaluates whether the column-level penalty or entropy is more effective at recovering module structure from MOSAIC's additive decoder.

What entropy regularizes. Fix a factor index j with influence column $A_{:,j} \in \mathbb{R}^D_{\geq 0}$ and let $p_{:,j} = A_{:,j} / \|A_{:,j}\|_1$ denote its normalization onto the probability simplex. The penalty

$$R_{\text{ent}}(A) = \frac{1}{|\mathcal{J}_{\text{alive}}|} \sum_{j \in \mathcal{J}_{\text{alive}}} H(p_{:,j}) \tag{16}$$

is a function of the *shape* of $A_{:,j}$, not its total magnitude. Concretely, R_{ent} is invariant under positive rescaling of any single column: $A_{:,j} \mapsto c \cdot A_{:,j}$ leaves $p_{:,j}$ and hence $H(p_{:,j})$ unchanged for any

$c > 0$. Lemma 1 bounds $H(p_{:,j}) \in [0, \log D]$, with the lower bound attained at one-hot columns and the upper bound at uniform columns. Minimizing R_{ent} therefore biases each column toward concentration on a small subset of observed variables, the support set $\mathcal{S}_j$ of Definition 1.

Decoupling from reconstruction. The shape-magnitude split is what makes entropy compatible with reconstruction. The Stage 2 objective $\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{ELBO}} + \lambda(t) \cdot \mathcal{L}_{\text{sparse}}$ asks each f_j to carry enough signal that $\hat{\mathbf{x}} = \sum_j f_j(z_j) + \mathbf{b}$ reconstructs $\mathbf{x}$. This pins down the magnitude $\|A_{:,j}\|_1$, but says nothing about how that magnitude is distributed across the D observation channels. Entropy operates exactly on this remaining degree of freedom: it rearranges mass within a column without contesting the column's overall scale. Reconstruction and entropy therefore act on orthogonal degrees of freedom, and an additive decoder can satisfy both: $\|A_{:,j}\|_1$ is set by the strength of the main effect g_j, and the shape of $A_{:,j}$ collapses onto $\mathcal{S}_j$.

This decoupling is the essential property a sparsity penalty needs in order not to fight reconstruction, and it is precisely what magnitude-based penalties lack. Element-wise ℓ_1 ($\sum_{i,j} |A_{ij}|$) and group-lasso ($\sum_j \|A_{:,j}\|_2$) both penalize magnitude. They reduce in the same way regardless of whether the model (a) drops zero-support entries while preserving in-support magnitude, which is what we want, or (b) shrinks the entire column toward zero, which kills the latent. Under (a) reconstruction stays good but the penalty does not fall as fast; under (b) the penalty falls but reconstruction degrades. The optimization typically lands at a compromise, alive-but-diffuse columns with reduced magnitude, which violates module structure. Consider two columns with identical ℓ_1 norm α: $A^{(\text{conc})}_{:,j} = (\alpha, 0, \ldots, 0)$ and $A^{(\text{diff})}_{:,j} = (\alpha/D, \ldots, \alpha/D)$. Both carry the same total influence magnitude, so reconstruction treats them as equivalent alternatives. However, the column-level group-lasso penalty $\|A_{:,j}\|_2$ is α for the concentrated column and $\alpha/\sqrt{D}$ for the diffuse column, so it *prefers* the diffuse solution. Entropy gives $H = 0$ for the concentrated column and $H = \log D$ for the diffuse column, correctly preferring concentration. This illustrates the fundamental misalignment: magnitude-based penalties reward spreading mass across channels, while entropy rewards concentrating it.

Empirical comparison on the synthetic benchmark. We compare entropy and group-lasso head-to-head on the synthetic energy-landscape benchmark with $z_{\dim} = 12$, sweeping the group-lasso coefficient across $\lambda \in \{10, 50, 100, 1000\}$ to identify its best operating point. Entropy uses its main-text setting $\lambda = 50$. Table 14 reports MCC for latent identifiability, X_Z@top3 for support recovery, and top-3 mass for column concentration.

Table 14: Entropy vs group-lasso on the synthetic benchmark ($z_{\dim} = 12$, 5 seeds). Group-lasso is swept across four coefficients to identify its best operating point. MCC is unchanged across all variants because Stage 1 identifiability is independent of the Stage 2 sparsity choice.

Variant	λ	MCC	X_Z@top3	Top-3 mass
Entropy	**50**	$\mathbf{0.912 \pm 0.009}$	$\mathbf{0.978 \pm 0.044}$	$\mathbf{0.685 \pm 0.028}$
Group-lasso	10	0.912 ± 0.009	0.822 ± 0.113	0.632 ± 0.026
Group-lasso	50	0.912 ± 0.009	0.867 ± 0.130	0.643 ± 0.020
Group-lasso	100	0.912 ± 0.009	0.911 ± 0.083	0.642 ± 0.015
Group-lasso	1000	0.912 ± 0.009	0.556 ± 0.199	0.683 ± 0.036

Three patterns emerge. First, MCC is identical at 0.912 ± 0.009 across all variants. The choice of sparsity penalty does not affect Stage 1 latent identifiability; that signal comes entirely from the temporal prior, and the penalty only shapes the Stage 2 decoder. Second, at its best setting $\lambda = 100$, group-lasso reaches X_Z@top3 of 0.911, still below entropy's 0.978 and with substantially higher seed variance (± 0.083 vs ± 0.044). Figure 7(C) shows why: at this λ the group-lasso decoder concentrates the diagonal entries correctly but leaves visible off-support mass in nearly every column, particularly on z_2, z_3, and u_1. This residual mass drives X_Z variance through unstable top-k selection. Entropy at $\lambda = 50$ produces the visibly cleaner diagonal of Figure 7(B). Third, pushing group-lasso to $\lambda = 1000$ collapses X_Z@top3 to 0.556 ± 0.199 even though MCC is unchanged. This is the magnitude-shrinkage failure mode anticipated earlier: with strong enough magnitude pressure, group-lasso drives column norms toward zero rather than concentrating mass within columns, and the resulting low-magnitude columns no longer carry a clean support signal. Notably, at $\lambda = 1000$ the top-3 mass (0.683) matches entropy's (0.685), yet X_Z@top3 collapses: the column does concentrate on three channels, but those channels are no longer the true support. Once magnitude shrinks below

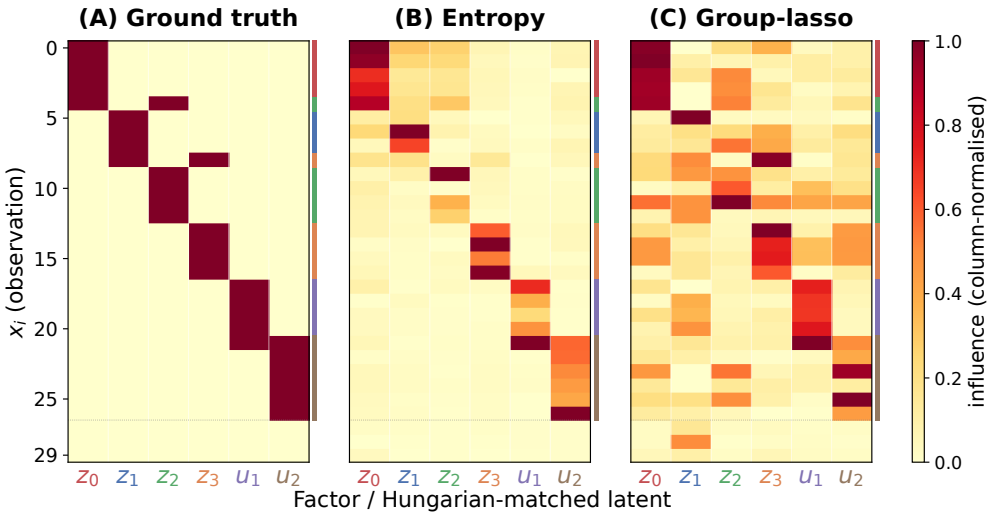

Figure 7: Influence matrices on the synthetic benchmark, column-normalized. **(A)** Ground-truth support structure: each true factor maps to a disjoint block of observed variables, with channels 27–29 as pure noise. **(B)** Entropy variant ($\lambda = 50$): the diagonal block structure is recovered with concentrated mass on the true support and minimal off-support leakage. **(C)** Group-lasso variant at its best operating point ($\lambda = 100$): the diagonal is recovered but every column carries non-trivial off-support mass, illustrating the magnitude-not-shape character of the penalty. Both panels use the same Hungarian alignment of learned latents to ground truth.

the noise floor, top-k selection picks up spurious entries. Entropy is invariant to this collapse by construction, since it only acts on normalized column shapes.

Why the alive mask is not optional. Scale invariance has a flip side: entropy alone cannot kill a latent. Multiplying a column by $c \to 0^+$ leaves $H(p_{:,j})$ unchanged, so a near-dead latent looks the same to R_{ent} as an active one. When $z_{\text{dim}} > n_{\text{true}}$, this would force the entropy penalty to concentrate excess latents onto some channel, producing spurious modules anchored to whichever channel the optimizer happened to pick first.

The alive mask $\mathcal{J}_{\text{alive}} = \{j : \sum_i A_{ij} > 0.01 \cdot \max_k \sum_i A_{ik}\}$ supplies the missing on/off mechanism. Columns whose total magnitude has collapsed are excluded from the entropy sum, so the optimizer is free to leave them at zero rather than forced to spike them. The two mechanisms partition the work cleanly: the alive mask handles the binary alive/dead decision (driven by reconstruction, since a latent stays alive only if it carries reconstruction signal), and entropy handles the shape decision conditional on being alive. Appendix F.2 quantifies the consequences when this partition is stressed: at $z_{\text{dim}} = 20$ the alive-mask threshold is near its operating limit, support recovery becomes seed-dependent, and X_Z@top3 drops from 0.978 at $z_{\text{dim}} = 12$ to 0.733.

Relation to population identifiability. Proposition 1 is unconditional on the sparsity penalty: at the population optimum the additive fit recovers each main effect g_j exactly, so $\hat{\mathcal{S}}_j = \mathcal{S}_j$ holds without any sparsity regularization. Entropy is therefore not part of the identifiability machinery. Its role is finite-sample readability. With finite N, the estimated $\hat{f}_j$ carries small but non-zero residual mass on channels outside $\mathcal{S}_j$, due to a combination of optimization noise and finite-sample bias. Without an explicit concentration mechanism, the influence matrix A contains many small entries everywhere, and any thresholding rule (top-k, ratio-gap, $\arg\max$) becomes seed-dependent. Entropy compresses this spurious mass back toward zero, making $\mathcal{S}_j$ readable from a finite-sample A. Proposition 4 formalizes this readout step: an estimated column within $\Delta_j/2$ of the population column $A^*_{:,j}$ in ℓ_∞ exactly recovers $\mathcal{S}_j$ via top-k_j selection. The entropy penalty is the mechanism that drives the estimated column close enough to its population counterpart for the influence-gap condition to bite.

Theoretical status. A tight finite-sample support recovery guarantee for entropy-regularized sparse additive models, analogous to Theorem 2 for group-lasso, is not currently available. The obstruction is geometric: H is concave in p on the simplex, but the composition $A \mapsto p(A) \mapsto H(p(A))$ is

neither convex nor concave in the unnormalized parameter A, so the convex-analysis tools that drive the group-lasso proof do not transfer. Lemma 1 and Proposition 4 together give a partial conditional statement: if column estimation is accurate enough to satisfy the influence gap, support recovery follows. A fully unconditional finite-sample statement remains open. We treat MOSAIC's synthetic evaluations (Appendices C.2 and C.1) as empirical validation of the same qualitative trends predicted by Theorem 2: higher interaction residuals and lower sample-to-dimension ratios both make support recovery harder.

Summary. Entropy is not a surrogate for group-lasso. The two penalties target different objects: group-lasso regulates column magnitude and is indifferent to internal shape, while entropy regulates column shape and is indifferent to magnitude. The shape-magnitude decoupling is what allows entropy to coexist with reconstruction; the alive mask supplies the on/off decision that scale-invariance prevents entropy from making; and entropy's role in the overall pipeline is finite-sample readout of the population-identifiable supports that Proposition 1 guarantees.

E.1 Relation to Direct Jacobian Sparsity

An alternative to MOSAIC's additive decoder is to keep a dense decoder $g_\theta(\mathbf{z})$ and directly regularize its Jacobian $\partial g_\theta/\partial \mathbf{z}$ toward sparsity, e.g., via ℓ_1 on the averaged absolute Jacobian entries $\frac{1}{N}\sum_t |(\partial \hat{\mathbf{x}}_t/\partial \hat{\mathbf{z}}_t)_{ij}|$. This approach has two limitations that the additive architecture avoids. First, the Jacobian of a dense MLP is a local quantity: $\partial g_\theta(\mathbf{z})/\partial z_j$ depends on the evaluation point $\mathbf{z}$, so a single factor z_j can produce a non-zero Jacobian entry at some $\mathbf{z}$ values and a near-zero entry at others. Aggregating over the dataset (e.g., by averaging) loses this local structure and may mask or fabricate support entries. The additive decoder's influence $A_{ij} = |f_j(+1)_i - f_j(-1)_i|$ is a global functional property of f_j, independent of where in the latent space it is evaluated. Second, Jacobian sparsity on a dense decoder does not structurally prevent cross-factor entanglement in the decoder: even if $(\partial g/\partial z_j)_i \approx 0$ at most points, the decoder can still route information from z_j to x_i through higher-order paths that the first-order Jacobian does not capture. The additive architecture $\hat{\mathbf{x}} = \sum_j f_j(z_j) + \mathbf{b}$ eliminates cross-factor interaction by construction, so sparsity in A directly corresponds to the ANOVA main-effect support of Proposition 1. The trade-off is representational capacity: the additive decoder cannot fit interaction terms in the mixing function, which is why MOSAIC targets main effects rather than full supports (Section 3).

F Robustness and Sensitivity Studies

F.1 Ablation and Robustness Studies

We ablate the three core components on the synthetic benchmark (5 seeds, $z_{\text{dim}} = 12$). Stage 2 ablations share the Stage 1 encoder, so their MCC is identical by construction.

Table 15: Ablation (5 seeds, $z_{\text{dim}} = 12$). X_Z@top3 uses the 0.50 concentration gate.

Config	MCC ↑	X_Z@top3 (gate 0.50) ↑	Top-3 mass ↑
Full MOSAIC	$\mathbf{0.912 \pm 0.009}$	$\mathbf{0.978 \pm 0.044}$	0.685 ± 0.028
w/o sparsity ($\lambda = 0$)	0.912 ± 0.009	0.378 ± 0.259	0.472 ± 0.031
w/o temporal ($\gamma = 0$)	0.765 ± 0.020	0.400 ± 0.181	0.771 ± 0.062
w/o additive (dense)	0.912 ± 0.009	0.000 ± 0.000	0.391 ± 0.008

The two capabilities decompose cleanly: latent identifiability (MCC) depends only on the temporal prior, while support recovery (X_Z) requires all three components together. The w/o temporal row is informative: its top-3 mass (0.771) exceeds full MOSAIC's (0.685), yet X_Z is only 0.400. Sparse columns alone are insufficient; they must align with correctly identified latents, the same failure mode TDRL exhibits on RNA (Section 6.2).

Robustness checks. We further verify MOSAIC's behavior under varying conditions: latent dimensionality z_{dim} sweep (Appendix F.2), regime-label noise up to 30% flip rate (Appendix F.4), unsupervised regime discovery via k-means (Appendix C.4), and a case study showing how the regime definition determines the scientific question MOSAIC answers (Appendix C.3). MOSAIC degrades gracefully in all four settings.

F.2 z_dim Sensitivity

The synthetic benchmark uses $n_{\text{true}} = 6$ true latent factors. MOSAIC's main-text results use $z_dim = 12$, chosen to be large enough to host all true factors with margin for dead latents. Here we sweep z_dim across under-, matched-, and over-capacity regimes to characterize MOSAIC's sensitivity to this hyperparameter.

Table 16: MOSAIC z_{dim} sensitivity on the synthetic benchmark ($n_{\text{true}} = 6$, 5 seeds). We sweep representative under-, matched-, and over-capacity settings under the canonical Stage 2 configuration ($\lambda = 50$, $H = 4$, 80 epochs, $\gamma = 0.02$). "Pass" denotes seeds with MCC ≥ 0.85 and X_Z@top3 (gate 0.5) ≥ 0.85.

z_{dim}	MCC	X_Z@top3 (gate 0.5)	Mean top-3 mass	Seeds passing
4	0.769 ± 0.054	0.133 ± 0.130	0.624 ± 0.028	0 / 5
6	0.900 ± 0.017	0.889 ± 0.099	0.692 ± 0.034	4 / 5
8	0.912 ± 0.010	0.844 ± 0.151	0.698 ± 0.043	5 / 5
12	$\mathbf{0.912 \pm 0.009}$	$\mathbf{0.978 \pm 0.044}$	0.685 ± 0.028	**5/5**
16	0.906 ± 0.010	0.733 ± 0.133	0.751 ± 0.031	3 / 5
20	0.885 ± 0.017	0.733 ± 0.249	0.606 ± 0.085	3 / 5

Under-capacity ($z_{\text{dim}} = 4$) breaks both latent recovery and module localization: MCC drops to 0.769 and no seed clears the joint threshold, consistent with the fact that four latent dimensions cannot represent the six true factors. At matched-to-moderately-over capacity ($z_{\text{dim}} \in \{6, 8, 12\}$), MCC is stable at 0.90–0.91 and the joint pass rate is high (4/5, 5/5, and 5/5 respectively). Both $z_{\text{dim}} = 8$ and $z_{\text{dim}} = 12$ achieve full 5/5 seed passing, but $z_{\text{dim}} = 12$ reaches the highest X_Z@top3 (0.978 vs. 0.844) with lower seed-to-seed variance, which is why we adopt it as the main synthetic default.

For stronger over-capacity ($z_{\text{dim}} \in \{16, 20\}$), MCC degrades only mildly, indicating that the temporal representation remains largely stable, but X_Z@top3 drops to 0.733 and the joint pass rate falls to 3/5. We observe a recurring failure mode behind this drop. With excess latent capacity, a subset of the extra factors carry a strong regime-association signal (high Cohen's d) but receive an influence column that is nearly uniform across the D observed variables, with per-row entries close to $1/D$ and no readable top-k support. These factors are precisely the kind that the alive mask of Appendix E is designed to neutralize: their column magnitude has not collapsed, so they are not dropped from the entropy sum, yet they have nothing localized to say. Under the canonical $\lambda_{\max} = 50$ schedule the entropy term is not strong enough to either concentrate them onto a coherent support or to push their magnitudes below the alive-mask threshold, so they coexist alongside the genuinely localized factors and inflate seed variance.

In practice this failure mode is benign. The diffuse high-d factors are easy to identify by the same diagnostic that drives MOSAIC's interpretation protocol: their influence column has near-uniform mass and no readable top-k support. We therefore recommend, when operating at $z_{\text{dim}} \gg n_{\text{true}}$, ranking factors jointly by regime-association score *and* top-3 mass, and skipping any factor whose top-3 mass is close to the uniform baseline $3/D$. The first regime-discriminative factor with a concentrated influence column carries the scientific signal; the diffuse high-d factors do not introduce false support claims, they simply cannot be interpreted. The 3/5 pass rate at $z_{\text{dim}} \in \{16, 20\}$ reflects strict joint-threshold accounting under fixed λ rather than an inability to recover the localized factor. Practitioners who do not know n_{true} in advance should therefore use a moderately overcomplete latent dimension as the structural default, and apply the concentration filter at interpretation time as a safeguard against the over-capacity regime.

F.3 Multi-Regime Extension

MOSAIC's interpretation protocol is contrastive. Given a scientifically specified regime contrast, it identifies the latent factor most associated with that contrast and interprets the factor through its sparse observation support. When data contain more than two regimes, the intended use is therefore one-vs-reference or pairwise analysis, depending on the scientific question. For example, a user may ask which module distinguishes a particular fault state from normal operation, or which structural module differs between two molecular regimes.

A joint K-way analysis is a distinct problem: it requires specifying what object should be recovered across all regimes simultaneously, such as a single latent per regime, a shared latent across multiple contrasts, or groups of latents with overlapping supports. Without this additional target definition, simultaneously explaining all $\binom{K}{2}$ pairwise differences is not the same task as the binary regime-association problem studied in this paper. Richer regime labels can provide stronger auxiliary variation for latent identifiability, but support localization additionally requires resolving assignment, sparsity, and possible shared-mechanism conflicts across contrasts.

Preliminary $K = 8$ separable-regime experiments support this distinction: some one-vs-reference contrasts localize well, while others remain diffuse or are assigned to the wrong module under a single global sparsity coefficient. This suggests that robust joint multi-regime localization requires additional machinery, such as contrast-specific assignment, per-regime adaptive sparsity, latent-group selection, or interaction-aware decoders. We therefore leave joint K-way support localization to future work and use MOSAIC in this paper as a contrastive interpretation method for binary or user-specified one-vs-reference scientific questions.

F.4 Regime Label Noise Robustness

Regime labels in real applications come from heuristics (SPIB assignments, thresholding, expert annotation) and may be noisy. We test MOSAIC's degradation by stratified random flips of the canonical binary synthetic benchmark's regime labels at rates $p \in \{0.05, 0.10, 0.20, 0.30\}$, holding the underlying data (dynamics, observations, ground-truth support) fixed. Training uses the noisy labels; evaluation uses the clean ground-truth regime-association identity and support.

Table 17: Regime label noise robustness (3 seeds per noise rate, z_dim=12). Labels are perturbed by stratified random flipping at rate p within each regime class. acc_noisy is the logistic-probe accuracy measured against the noisy labels (the model saw); acc_clean is the same probe measured against ground-truth clean labels.

Noise rate p	MCC	X_Z@top3 (gate 0.50)	acc_noisy	acc_clean
0.00	0.912 ± 0.010	0.978 ± 0.044	0.995	0.995
0.05	0.891 ± 0.012	0.960 ± 0.040	0.942	0.992
0.10	0.870 ± 0.031	0.952 ± 0.044	0.896	0.992
0.20	0.856 ± 0.045	0.863 ± 0.082	0.812	0.942
0.30	0.808 ± 0.062	0.827 ± 0.113	0.732	0.892

MOSAIC degrades gracefully under label noise. At moderate noise ($p = 0.10$), MCC drops from 0.912 to 0.870 and X_Z from 0.978 to 0.952. At heavy noise ($p = 0.30$), MCC drops to 0.808 (11% relative) and X_Z to 0.827 (15% relative). For the practitioner, the 5–10% noise regime typical of imperfect real-world regime labeling has minimal impact on MOSAIC's outputs.

G Related Applications and Future Directions

Scientific sensing and industrial monitoring. Wearable and industrial sensing provide natural settings for extending module discovery. Recent applications include estimating impact force and acceleration with wireless wearable sensors [Chen et al., 2026b], combining vibration and acoustic measurements for micro-motor testing [Chen et al., 2026a], and using vibration analysis for fault diagnosis and quality control [Chen et al., 2026c]. These tasks connect measured signals to physical quantities or operating conditions. A future application of MOSAIC could examine whether distinct latent factors correspond to compact groups of sensor channels or derived features. This would complement prediction and classification by identifying which measurements carry each factor, with physical calibration or known operating changes providing an external basis for interpreting the recovered modules.

Latent states for forecasting. LatentTSF forecasts in a learned latent state space [Yang et al., 2026a], while TimeAlign aligns historical and future representations to address their distributional gap [Hu et al., 2025]. These approaches motivate testing whether MOSAIC's sparse latent modules can provide useful predictive states while retaining their links to named observations.

Robustness under input perturbations. MVI-Bench studies robustness to misleading visual inputs in large vision-language models [Chen et al., 2025]. Its focus on input reliability motivates a related evaluation direction for module discovery: testing how recovered supports change under controlled perturbations of observed time series. For scientific sensing, such perturbations could include channel noise, missing measurements, or calibration shifts. Glocal-IB aligns latent representations of masked and originally observed inputs for time-series imputation [Yang et al., 2026b], while CRIB uses consistency-regularized representations to forecast directly from partially observed inputs [Yang et al., 2025]. Comparing module stability across these conditions could help distinguish sensitivity to measurement quality from changes associated with the underlying physical process. ConfDock's atom-specific uncertainty quantification for molecular docking [Hao et al., 2026] also motivates studying calibrated uncertainty estimates for recovered module assignments in molecular applications.

Time-series reasoning systems. TSRouter selects a modality and model for each time-series reasoning query [Yu et al., 2026], while the survey by Hu et al. places such reasoning within the broader development of agentic time-series systems. MOSAIC could supply these systems with compact descriptions of latent modules and their associated sensor channels, allowing downstream reasoning to refer back to specific measured variables.

Interactive scientific analysis. A further direction is to integrate MOSAIC as a module-discovery tool in an interactive assistant. GAM's graph-based memory [Wu et al., 2026] could organize prior hypotheses and module comparisons across an analysis session. InterruptBench, which evaluates changing user instructions during web navigation [Zou et al., 2026], motivates testing how such an assistant revises its analysis when a researcher changes the target variables or time interval. Agent Security Bench [Zhang et al., 2025] offers a starting point for assessing attacks involving tool outputs and agent memory in these workflows.

H RNA Molecular Dynamics: Extended Analysis

System description. The cUUCGg tetraloop (14 nucleotides: `GGCACUUCGGUGCC`) is a canonical RNA hairpin. We perform MD simulations of this hairpin at three temperatures: 345K (near melting, partial unfolding), 400K (above melting, frequent transitions), and 500K (rapid unfolding); the full simulation protocol is given below in the "MD protocol" paragraph. The trajectory inventory consists of 14 input trajectories: 13 curated replicas plus one auxiliary 500K unfolding trajectory. The 500K data does not dominate the disrupted class — the disrupted frames come predominantly from the long 400K trajectories containing loop/stem disruption events — but we do not claim temperature-invariant localization in this manuscript.

MD protocol. All trajectories were generated in OpenMM under the AMBER14 force field with OL3 corrections for RNA, in TIP3P explicit water with neutralizing counterions in a periodic cubic box with at least 10 Å of solvent padding. The initial structure was the cUUCGg tetraloop hairpin (PDB 2KOC), prepared once via energy minimization followed by NPT equilibration at 300 K and 1 bar; the post-equilibration checkpoint, together with the serialized system and topology, was reused as the launch state for every replica so that all inter-replica divergence reflects the integrator's stochastic stream rather than initialization noise. Production runs used the Langevin middle integrator ($\gamma = 1\,\mathrm{ps}^{-1}$) with hydrogen-mass repartitioning ($\mathrm{hydrogenMass} = 3$ amu) enabling a 4 fs timestep, a Monte Carlo barostat at 1 bar with update frequency 25 steps, particle-mesh Ewald electrostatics, and a 10 Å nonbonded cutoff with HBond constraints. Each replica was assigned an independent integrator random seed and velocities re-drawn from the Maxwell–Boltzmann distribution at the target temperature. Trajectories spanned 100 ns to 1 μs in length and were generated at 345 K, 400 K, and 500 K (14 trajectories total). Two trajectory streams were recorded per replica: an RNA-only DCD written every 10 ps used for downstream feature extraction, and a full-system DCD written every 100 ps retained as an archival reference. Compute resources were the Bridges2 cluster (PSC) and Google Colab Pro (A100 GPU).

Feature extraction. For each of the 14 residues, we compute the mean and standard deviation of its minimum heavy-atom distance to every other residue at each frame. This yields $D = 28$ per-residue distance features (14 mean + 14 std). Distance features directly capture base-pairing module structure: paired residues (e.g., G1–C14) co-vary in distance space, while unpaired loop residues (U6–G9) fluctuate independently.

Regime definition. Regimes are defined by the closing-pair native-contact fraction Q_{cp}. Candidate residue pairs are sequence-separated ($|i - j| \geq 3$) closest-heavy contacts that touch C5 or G10. Native contacts are selected from frame 0 of each trajectory, not from an external 2KOC reference structure. A pair is included as native if $d_{ab}(0) < 0.45\,\text{nm}$ ($4.5\,\text{Å}$), and it is present at frame t iff $d_{ab}(t) < 1.2\,d_{ab}(0)$. Thus $Q_{\text{cp}}(t)$ is the mean of binary contact indicators over the native C5/G10-touching contact set; contacts are binary, not distance-weighted.

A 51-frame rolling mean is applied before thresholding to suppress thermal noise. Frames are labelled intact when $Q_{\text{cp}} > 0.8$ (regime 0) and disrupted when $Q_{\text{cp}} < 0.65$ (regime 1); intermediate frames and a ± 50-frame buffer around label transitions are discarded. From stored labels, 24.1% of frames are discarded after smoothing and transition-buffer removal; this combines intermediate-band frames and transition-buffer frames because Q_{cp} itself was not stored in the diagnostics file. After lag-2 windowing (sequence length 3) and class balancing, the dataset contains $156{,}170$ windows ($78{,}085$ per regime, $N/D \approx 5{,}571$). For cross-regime validation, we also train with an outer-stem contact criterion Q_{stem} using torsion-angle features ($D = 56$).

Scope of RNA experiments. We have not run threshold-sensitivity, three-regime, or temperature-specific MOSAIC experiments in this manuscript. The reported RNA result is a pooled-temperature analysis; we do not claim temperature-invariant localization.

Ground-truth support structure. Based on established RNA structural biology, we define three modules corresponding to the physical regions of the hairpin:

Table 18: Ground-truth support structure for the RNA tetraloop. Each module corresponds to a structural region of the hairpin.

Module	Residues	Variables	Physical role
Stem	G1, G2, C3, A4, U11, G12, C13, C14	16	Base-paired stem (last to unfold)
ClosingPair	C5, G10	4	Stem-loop interface (commitment gate)
Loop	U6, U7, C8, G9	8	UUCG tetraloop (first to fluctuate)

Extended results. With $n = 4$ latent factors, MOSAIC achieves regime accuracy 0.912 ± 0.027 (5 seeds) and mean concentration 0.617 ± 0.052. The top regime-associated latent (highest Cohen's d, mean $d = 1.48 \pm 0.24$) localizes to Loop in 4 of 5 seeds and Stem in the remaining seed. Across all 5 seeds, the rank ordering of module influence is stable: Loop $>$ Stem $>$ ClosingPair, consistent with the known thermodynamic coupling of stem unpairing and loop melting in RNA hairpin unfolding [Chen and García, 2013].

Table 19: RNA regime-association summary for MOSAIC (5 seeds), moved from the main text for compactness.

Setting	Regime Acc ↑	Top regime-associated module
MOSAIC (cp_Q, 5 seeds)	$\mathbf{0.912 \pm 0.027}$	Loop (#1 in 4/5)

Seed-level diagnostic for the non-Loop-first run. The only non-Loop-first MOSAIC seed is not a failure to capture the regime signal: the selected regime-associated latent still has strong regime association (Cohen's $d = 1.395$). Its support, however, is routed through stem and stem-junction residues rather than purely through Loop residues. In this seed, five of the top-8 non-Loop entries come from stem positions, consistent with stem–loop co-variation during closing-pair disruption. This suggests a boundary case of the discrete Loop/ClosingPair/Stem annotation rather than an unrelated localization: the transition signal remains physically coupled to the loop-opening process, but the sparse decoder assigns part of the support to correlated stem residues. Under the enhanced gap-adaptive support diagnostic, this seed retains Loop recall 0.75, indicating that most Loop residues remain present in the learned support even though they do not dominate the top-8 precision metric.

Baseline comparison. Table 20 compares MOSAIC to CRL, linear, and information-bottleneck baselines under the unified X_Z@selected-latent metric (precision of the top regime-associated factor's top-8 influence entries against Loop's 8 ground-truth variables). MOSAIC achieves 0.880 ± 0.055, exceeding every baseline by at least 0.13. All baselines except PCA still rank Loop as their top regime-associated module by Cohen's d, confirming that the Loop signal is strong, but only MOSAIC

concentrates its influence cleanly on the ground-truth Loop support. PCA is a useful failure case: it ranks Stem first rather than Loop, consistent with variance-dominant behavior in unregularized linear decomposition, where high-variance structural coordinates dominate principal components even when they are not the main regime-shift factor. SlowVAE achieves the highest regime accuracy (0.955) but the lowest X_Z@selected-latent among CRL methods (0.562), illustrating the distinction between regime separability and module localization.

Table 20: RNA baseline comparison using the canonical X_Z@selected-latent metric (precision of the top regime-associated factor's top-8 influence entries against Loop's 8 variables). MOSAIC: 5-seed mean $\pm$ std; CRL baselines: 4 seeds (42/123/456/789); linear baselines and SPIB: single seed or 5-seed mean as noted.

Method	Regime Acc ↑	X_Z@selected (k=8) ↑
MOSAIC (ours, 5s)	$\mathbf{0.912 \pm 0.027}$	$\mathbf{0.880 \pm 0.055}$
TDRL (4s)	0.912 ± 0.030	0.625 ± 0.102
iVAE (4s)	0.930 ± 0.003	0.594 ± 0.157
SlowVAE (4s)	0.955 ± 0.001	0.562 ± 0.217
CtrlNS (4s)	0.929 ± 0.011	0.562 ± 0.161
β-VAE (4s)	0.858 ± 0.028	0.469 ± 0.120
Sparse PCA (α=10)	0.898	0.750
Sparse PCA (α=1)	0.902	0.625
FastICA	0.901	0.625
PCA	0.901	0.625
SPIB (5s)	0.645 ± 0.065	0.400 ± 0.137

For cross-regime validation, retraining with Q_{stem} labels shifts the dominant module from Loop to Stem (Table 21). This supports a non-circular interpretation: the selected regime-associated latent changes its localized support with the physical regime definition, rather than always pointing to a fixed residue group.

Regime design and scientific target. The cross-regime result also clarifies a practical use pattern: the regime definition determines the scientific question that the regime-association ranking answers. If the goal is to compare two steady states (State A vs. State B), regimes should be defined directly by those states. If the goal is to study transition dynamics (how State A moves toward State B), regimes can be defined as stable State A versus pre-transition State A. Under either choice, MOSAIC returns the factor with the strongest cross-regime shift under that task-specific definition.

Table 21: Cross-regime validation on RNA: the top regime-associated module shifts with the physical regime definition, tracking the structural layer adjacent to the contact set used for regime labeling.

Regime definition	Feature space	Localized module of selected latent
Closing-pair contacts (Q_{cp})	Per-residue distances (D=28)	Loop
Outer-stem contacts (Q_{stem})	Torsion angles (D=56)	Stem

I Cross-Domain Details

Each cross-domain dataset is preprocessed per Appendix B. Regime definitions are as follows. OMNI [King and Papitashvili, 2005] uses a geomagnetic-activity threshold on the Kp index. The tokamak disruption benchmark is synthetic, generated by an AR(1) process ($\rho = 0.95$, seed=42) over four latent sources (MHD instability, Density, Energy confinement, Plasma shape), mapped to 12 observables through a sparse mixing matrix; the pre-disruptive regime (regime 1) activates the MHD and Density sources with a ramp in the second half of each shot. Construction details are in Appendix I.1, and variable selection follows tokamak physics [De Vries et al., 2011]. Climate uses ENSO phase derived from NOAA ERSSTv5 sea surface temperature data [Huang et al., 2017b,a]. TEP [Downs and Vogel, 1993] uses a binary fault-vs-normal label where fault instances are drawn from IDV-1, IDV-2, IDV-4, and IDV-5 (feed-ratio, B-composition, reactor-cooling, and condenser-

cooling faults respectively). This multi-fault regime design tests whether MOSAIC can recover a shared downstream signature when multiple distinct upstream disturbances are labeled jointly.

Throughout these diagnostics, the selected latent is the one with largest distributional discrepancy across the specified regimes. This association-based selection should not be interpreted as an interventional causal effect of the latent on the transition.

Table 22: Additional numerical details for the cross-domain benchmarks (MOSAIC, 3 seeds per dataset). "Top d" reports the observed range of the leading Cohen's d across seeds, and "Expected regime-associated variables" summarizes the variables on which the dominant factor loads.

Dataset	Expected regime-associated variables	Regime Acc	Top d	Consistent
OMNI	Coupling-related solar-wind parameters	0.87–0.98	1.54–1.98	3/3
Disruption	MHD/Density group (q_{95}, greenwald_frac)	0.91	1.96–2.08	3/3
Climate	Niño 3.4 and tropical Pacific indices	0.85–0.92	1.44–1.83	3/3
TEP	Stripper group (product composition XMEAS_37–41)	0.729 ± 0.065	0.743 (range 0.554–0.862)	3/3

Table 23: Signal-strength separation on the synthetic tokamak benchmark. Both q_{95} and locked-mode amplitude are generated from the same MHD latent source, so both receive the pre-disruptive ramp scaled by their respective mixing coefficients. However, locked-mode's smaller mixing coefficient produces a much weaker ramp signal (Cohen's $d = 0.24$ vs. 2.08), causing MOSAIC to assign it to a separate, weakly regime-associated factor rather than grouping it with q_{95}. This is a limitation: MOSAIC's support recovery is signal-strength dependent, and weak-signal channels from the same source may be split into separate factors rather than recovered as a single module.

Latent factor	Cohen's d	Top loaded variable	Role
z_3	2.08	q_{95}	Regime-varying (strong signal)
z_0	0.24	Locked-mode amp.	Weakly regime-associated (weak signal)

Across the four cross-domain benchmarks where the additive-sparse assumption is met (OMNI, Disruption, Climate, TEP), the dominant factor loads on the domain-expected variables in all three seeds per dataset, for 12 of 12 seeds in aggregate.

Linear baseline comparison. Table 24 reports the available deterministic linear baselines on three cross-domain benchmarks (OMNI, Disruption, Climate). SparsePCA (α=1) achieves domain-consistent regime-associated localization on all three reported datasets, matching MOSAIC on this coarse domain-knowledge criterion for the same datasets. This means the cross-domain result should be read as transfer evidence for MOSAIC's interpretation protocol rather than as a universal dominance claim; MOSAIC's added value is identifiable nonlinear temporal latents and cleaner separation of regime-associated from variance-dominant factors. PCA fails on OMNI and Climate, each identifying variance-dominant rather than regime-relevant variables.

On Disruption, all linear methods include locked-mode amplitude in their top-3 regime-associated variables alongside q_{95}. MOSAIC instead assigns locked-mode to a separate, weakly regime-associated factor (Table 23). Since both observables are generated from the same MHD source, including locked-mode alongside q_{95} (as linear methods do) is in fact consistent with the ground-truth module structure. MOSAIC's separation reflects signal-strength thresholding rather than true module discovery in this case: the weak ramp signal on locked-mode falls below the concentration threshold, splitting a single ground-truth module across two learned factors.

SPIB on cross-domain. SPIB achieves domain-consistent regime-associated localization in 5 of 9 reported runs (Table 25), compared to 9/9 for MOSAIC on the same three datasets (OMNI, Disruption, Climate). On OMNI, SPIB fails all three seeds; its early-stopping convergence criterion triggers within 2 to 4 epochs.

I.1 Synthetic Tokamak Disruption Benchmark: Construction Details

The tokamak disruption benchmark is a physics-motivated synthetic dataset. We use synthetic construction because accessing per-shot time-series diagnostic data from operational tokamaks (JET, DIII-D, ASDEX Upgrade) requires facility-specific data access agreements not within the scope of this work; the ITPA disruption database contains aggregate shot-level summary statistics rather

Table 24: Available linear baselines on three cross-domain benchmarks (single seed, deterministic). "Consistent" uses the same domain-knowledge criteria as Table 4.

Dataset	Method	Regime Acc	Consistent?
OMNI	SparsePCA α=1	0.992	YES
OMNI	SparsePCA α=10	0.992	YES
OMNI	FastICA	0.992	YES
OMNI	PCA	0.992	NO
Disruption	SparsePCA α=1	0.924	YES
Disruption	SparsePCA α=10	0.924	YES
Disruption	FastICA	0.925	YES
Disruption	PCA	0.924	YES
Climate	SparsePCA α=1	0.956	YES
Climate	SparsePCA α=10	0.975	YES
Climate	FastICA	0.936	YES
Climate	PCA	0.934	NO

Table 25: SPIB on cross-domain benchmarks (3 seeds each).

Dataset	Regime Acc	Consistent?
OMNI	0.696 ± 0.105	0/3
Disruption	0.763 ± 0.053	3/3
Climate	0.750 ± 0.066	2/3

than the per-shot time series required for temporal representation learning. We instead construct a benchmark whose variable names and module structure follow established tokamak physics to preserve the scientific interpretation of group-mean winners.

Generation. Four latent sources (MHD instability, Density, Energy confinement, Plasma shape) evolve as AR(1) processes with autocorrelation $\rho = 0.95$ and innovation standard deviation 0.3, seed $= 42$. A sparse block-structured mixing matrix $M \in \mathbb{R}^{4\times 12}$ maps sources to observables: MHD source maps to q_{95}, l_i, locked_mode; Density source maps to $\bar{n}$, greenwald_frac; Energy source maps to w_{MHD}, β_N, $p_{\mathrm{rad}}/p_{\mathrm{in}}$; Shape source maps to I_p, elongation, triangularity, δ. Mixing coefficients are drawn uniformly from $[0.6, 1.5]$. Gaussian observation noise ($\sigma = 0.15$) is added.

Regime labels. In regime 0 (healthy), sources evolve under the AR(1) dynamics alone. In regime 1 (pre-disruptive), the MHD and Density sources receive an additive ramp in the second half of each shot: $s_t^{(\mathrm{MHD})} \mathrel{+}= 1.5 \cdot \mathrm{progress}_t$ and $s_t^{(\mathrm{Density})} \mathrel{+}= 0.8 \cdot \mathrm{progress}_t$, where $\mathrm{progress}_t = (t - T/2)/T$ for $t > T/2$ and zero otherwise. This reflects the physics-literature pattern in which MHD mode growth and density-limit approach precede disruption onset [De Vries et al., 2011]. MOSAIC's recovery of MHD or Density as the regime-associated group is consistent with this ground-truth construction.

Sample size. The processed synthetic tokamak disruption benchmark contains $N = 4{,}600$ lagged windows with sequence length 3 and $D = 12$ variables, balanced across healthy and pre-disruptive regimes (2,300 windows each). It was generated with 100 trajectories per class, shot length 50, and lag 2; the reported $N/D = 383$ is the sample-to-dimension ratio, not the number of samples per regime.

Scope. The variable names, module grouping, and pre-disruptive ramp dynamics are drawn from published tokamak physics. The sparse mixing structure matches MOSAIC's additive-sparse assumption more cleanly than raw tokamak diagnostics might; this benchmark therefore provides a controlled evaluation of MOSAIC on tokamak-relevant physical structure, not a substitute for deployment on real disruption prediction.

I.2 TEP: Multi-Fault Regime

The Tennessee Eastman Process (TEP) is a well-studied chemical-process benchmark with 52 observed variables (41 measurements `XMEAS_1`–`XMEAS_41` and 11 manipulated variables `XMV_1`–

XMV_11) grouped into Reactor, Condenser, Separator, Stripper, Compressor, and Feed/Global physical clusters. The normal-vs.-faulty regime split uses $N = 4{,}176$ post-windowing samples ($D = 52$, $N/D \approx 80$). We label regime 0 as normal operation (IDV-0) and regime 1 as jointly drawn from IDV-1 (stream-A feed ratio fault), IDV-2 (stream-B composition fault), IDV-4 (reactor cooling fault), and IDV-5 (condenser cooling fault). Under this multi-fault labeling, MOSAIC's top-1 influence variable on the most regime-associated latent is XMEAS_1 (stream-A feed flow, the IDV-1 primary fault variable) across all 3 seeds (seeds 0, 42, 123; selected latents z_4, z_5, z_3 respectively), and the per-group mean winner is the Stripper group (containing Stream-11 product composition variables XMEAS_37–XMEAS_41, which are the downstream response to all four fault types). Regime accuracy is 0.729 ± 0.065 and the top Cohen's d ranges from 0.554 to 0.862 (mean 0.743). The winner-vs-runner-up group-mean ratio is 1.07, substantially smaller than Disruption's 1.38–2.13: this reflects the multi-fault regime structure, in which no single fault's primary variable group dominates because the regime-associated latent must encode four distinct fault types jointly, and the Stream-11 product-composition cluster is the common downstream observable of all four faults. Under a single-fault regime (e.g., IDV-1 only), we would expect a sharper group-mean separation; the multi-fault benchmark represents a harder setting.

J Interaction Ratio Calibration on Real Data

Theorem 2 predicts that support recovery becomes more sample-demanding as the interaction ratio ρ grows. To locate RNA and the four cross-domain benchmarks on this empirical scale, we estimate $\hat{\rho}$ using only quantities already computed during MOSAIC training.

Estimator. Let $\hat{\mathbf{z}}_t$ denote the Stage 1 encoder posterior mean and $\hat{g}^{\mathrm{add}}(\hat{\mathbf{z}}_t) = \sum_j \hat{f}_j(\hat{z}_{t,j}) + \hat{\mathbf{b}}$ the Stage 2 additive-decoder reconstruction. Write $\hat{\sigma}^2_{\epsilon,\mathrm{ub}} = \frac{1}{N}\sum_t \|\mathbf{x}_t - \hat{g}^{\mathrm{dense}}(\hat{\mathbf{z}}_t)\|_2^2$ for the Stage 1 dense-decoder reconstruction MSE, which serves as an upper-bound proxy for the true observation-noise variance σ^2_ϵ. The empirical interaction-residual variance is

$$\hat{\sigma}^2_r \;=\; \frac{1}{N}\sum_{t=1}^{N}\big\|\mathbf{x}_t - \hat{g}^{\mathrm{add}}(\hat{\mathbf{z}}_t)\big\|_2^2 \;-\; \hat{\sigma}^2_{\epsilon,\mathrm{ub}}, \tag{17}$$

obtained by subtracting the noise-floor proxy from the additive-decoder residual. Writing $\hat{\sigma}^2_g = \frac{1}{N}\sum_t \|\mathbf{x}_t - \bar{\mathbf{x}}\|_2^2 - \hat{\sigma}^2_{\epsilon,\mathrm{ub}}$ for the total signal variance (with $\bar{\mathbf{x}} = \frac{1}{N}\sum_t \mathbf{x}_t$), the interaction ratio is estimated by $\hat{\rho} = \hat{\sigma}^2_r/\hat{\sigma}^2_g$. Because $\hat{\sigma}^2_{\epsilon,\mathrm{ub}}$ is an upper bound on true noise, subtracting it reduces both numerator and denominator. Since the numerator (interaction variance) is smaller than the denominator (total signal variance), the fractional reduction is larger in the numerator, so $\hat{\rho}$ is a lower bound on the true ρ. This makes the estimator optimistic about additivity: if anything, the true interaction contribution is larger than the reported $\hat{\rho}$, which means MOSAIC's empirical support recovery on these datasets is achieved under interaction levels at least as strong as those reported in Table 26.

Results. Table 26 reports $\hat{\rho}$ on the five evaluated domains, together with the input dimensionality D, sample-to-dimension ratio N/D, and the primary recovery metric for each dataset. For RNA, the recovery metric is X_Z@selected-latent from Appendix H; for the four cross-domain benchmarks, ground-truth support is not available, so we report regime accuracy from Table 4 where available.

Calibration anchor. On the synthetic benchmark, where the ground-truth mixing is additive ($\rho = 0$ by construction), the same estimator returns $\hat{\rho}_{\mathrm{synth}} = 0.024$ (median over knot/CV grid), establishing a finite-sample calibration floor. All real-domain estimates above sit at or modestly above this floor, consistent with the additive-mixing assumption being empirically defensible.

Comparison with the synthetic sweep. On the controlled interaction sweep in Appendix C.1, three of the four estimable evaluated domains (OMNI at 0.045, TEP at 0.018, Disruption at 0.003) fall inside the calibrated range with $\hat{\rho} \le 0.05$. RNA's $\hat{\rho} = 0.161$ lies beyond it, reflecting allosteric coupling richer than the synthetic benchmark's pairwise $\tanh$ construction can produce; MOSAIC still localizes the selected latent to Loop in 4 of 5 seeds, showing that the empirical method can remain useful outside the lowest-interaction calibrated range.

Figure 8 summarizes the joint behavior of $\hat{\rho}$, synthetic interaction strength α, and MOSAIC's support-recovery and regime-association metrics, with the four estimable evaluated domains marked on the empirical degradation curve.

Table 26: Estimated interaction ratio $\hat{\rho}$ on the five evaluated domains. Smaller $\hat{\rho}$ means the mixing is closer to exactly additive, and main-effect recovery is expected to be more accurate. TEP is a 3-seed estimate (mean $\pm$ std); other estimable domains are single-ckpt estimates. †Climate ($N/D = 13$) is marked "—" because its N/D ratio is too low for the nonparametric additive-projection estimator to yield a reliable $\hat{\rho}$; MOSAIC's regime accuracy on this dataset (0.89) confirms it is not in a catastrophic-interaction regime.

Dataset	D	N/D	$\hat{\rho}$	Recovery metric
RNA	28	5,571	0.161	0.880 ± 0.055 (X_Z@selected-latent)
OMNI	21	1,277	0.045	0.940 ± 0.054 (Regime Acc)
TEP	52	80	$\mathbf{0.018 \pm 0.006}$	0.729 ± 0.065 (Regime Acc)
Disruption	12	383	0.003	0.911 ± 0.003 (Regime Acc)
Climate	47	13	—†	0.888 ± 0.027 (Regime Acc)

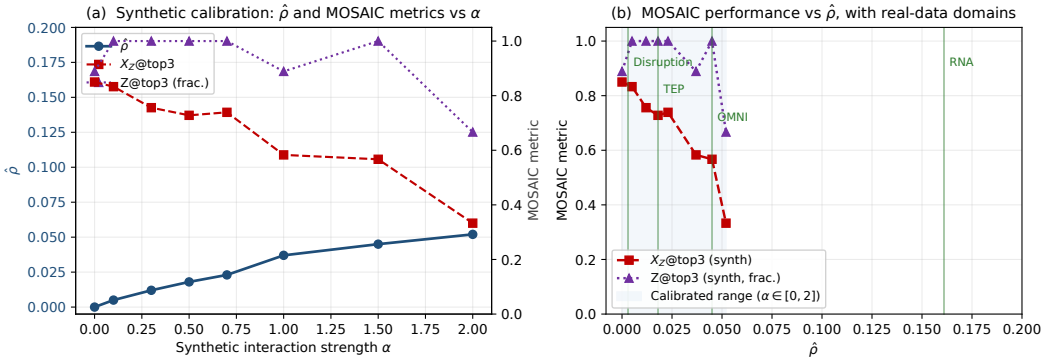

Figure 8: Interaction-ratio calibration. (a) On the synthetic interaction sweep, $\hat{\rho}$ (left axis) grows with the interaction strength α, while MOSAIC metrics (right axis) show that the stricter module-localization score X_Z@top3 degrades earlier than the latent-level regime-association score Z@top3. (b) Plotting MOSAIC metrics against $\hat{\rho}$ places the estimable evaluated domains on the empirical degradation curve: OMNI, TEP, and Disruption fall inside the synthetic calibrated range (shaded), while RNA ($\hat{\rho} = 0.161$) lies beyond it but still yields the expected Loop-associated localization. Climate is omitted because its N/D ratio is insufficient for the $\hat{\rho}$ estimator used here (Table 26).

K Concentration Metric Details

The concentration metric $\text{conc}(i) = \max_j A_{ij} / \sum_j A_{ij}$ measures how strongly each variable is dominated by a single latent factor. At the uniform baseline ($1/n$), every factor contributes equally and no module structure is present. At concentration 1.0, the variable is perfectly assigned to one factor. On RNA, the mean concentration is 0.617 ± 0.052 (5 seeds), indicating that most variables are predominantly governed by a single latent factor. Consistent with Theorem 2, concentration correlates with N/D across domains: datasets with higher N/D ratios achieve sharper support assignments.

L Concentration Gate Sensitivity

X_Z@top3 in Table 1 uses a concentration gate at top-3 mass threshold 0.50: a regime-discriminative latent contributes to the score only if its top-3 influence mass accounts for at least half of the total column mass. Ungated X_Z@top3 measures top-rank overlap with the ground-truth support, whereas the concentration gate tests whether a factor yields a localized module-level explanation; a dense decoder with partial top-rank overlap can still fail the gate because its influence is spread across many observed variables. TDRL illustrates this pattern: its ungated score is 0.519, so there is nonzero overlap with the true support, but its top-3 mass averages only $\approx 25\%$ and the gate correctly reports diffuse mass rather than a zero-overlap failure. Table 27 reports X_Z@top3 across a range of thresholds. The gate at 0.50 is the permissive choice: MOSAIC's mean top-3 mass is 0.70, so it satisfies the gate on all seeds. Gates at 0.60 and 0.70 penalize MOSAIC's own moderate seeds as

well, reducing discriminative power. Without any gate, iVAE's dense decoder scores 0.863 purely from argmax alignment, masking the absence of actual support concentration.

Table 27: X_Z@top3 sensitivity to the concentration-gate threshold μ_0 on the synthetic benchmark. Gate at 0.50 (used in Table 1) cleanly separates sparse-decoder methods (MOSAIC, iVAE) from dense-decoder methods (TDRL, SlowVAE, β-VAE). MOSAIC is unaffected at 0.50 ($X_Z = 0.978$), moderately penalized at 0.60 ($X_Z = 0.844$), and more heavily penalized at 0.70 ($X_Z = 0.511$) — indicating its top-3 captures approximately 65%–70% of column mass on average. iVAE's decoder, despite argmax correctness under no gate, places only $\approx 50\%$ of mass in the top-3, causing sharp degradation.

Method	gate 0.50	gate 0.60	gate 0.70	no gate
MOSAIC	0.978	0.844	0.511	0.978
iVAE	0.556	0.222	0.000	0.863
TDRL	0.000	0.000	0.000	0.519
SlowVAE	0.000	0.000	0.000	0.407
β-VAE	0.000	0.000	0.000	0.407

M SPIB Hyperparameter Sweep on the Synthetic Benchmark

SPIB underperforms MOSAIC and most other baselines on the synthetic benchmark (Table 1), with regime accuracy 0.304 ± 0.019 and zero gated X_Z@top3. We swept SPIB hyperparameters to verify that this is a structural mismatch with the support-recovery objective rather than a poorly tuned baseline.

Sweep design. We swept $\beta \in \{0.001, 0.01, 0.1, 1.0\}$ and $K \in \{2, 4, 8\}$, the two hyperparameters that govern SPIB's information-bottleneck strength and the number of metastable states. For $K > 2$, we enabled SPIB's iterative label refinement (`refinements=5`) so that extra output slots could be populated from the binary ground-truth regime labels. All 12 configurations were trained on the synthetic benchmark at seed 42.

Table 28: SPIB hyperparameter sweep on the synthetic benchmark, seed 42. Regime accuracy is the logistic-probe accuracy on $\hat{\mathbf{z}}$; MCC, Z@top3, and X_Z@top3 are defined identically to Table 1. "Driver $\hat{z}^*$" is the learned latent index with largest Cohen's d under each configuration; these are learned latent indices, not ground-truth factor labels.

β	K	Regime acc	MCC	Z@top3	X_Z@top3	Driver z^*	Cohen's d
0.001	2	0.705	0.274	0/3	0.000	z_0	0.538
0.001	4	**0.814**	**0.464**	0/3	0.000	z_2	1.361
0.001	8	0.791	0.267	0/3	0.000	z_5	1.030
0.01	2	0.724	0.320	0/3	0.000	z_3	0.788
0.01	4	0.676	0.191	0/3	0.000	z_6	0.647
0.01	8	0.667	0.208	0/3	0.000	z_1	0.485
0.1	2	0.698	0.287	0/3	0.000	z_3	0.602
0.1	4	0.717	0.303	0/3	0.000	z_0	0.776
0.1	8	0.630	0.222	0/3	0.000	z_5	0.477
1.0	2	0.711	0.306	0/3	0.000	z_3	0.683
1.0	4	0.701	0.271	0/3	0.000	z_3	0.644
1.0	8	0.699	0.233	0/3	0.000	z_7	0.619

Two patterns emerge. First, regime accuracy is moderately tunable: 7 of 12 configurations exceed 0.70, and the best ($\beta = 0.001$, $K = 4$) reaches 0.814. The Table 1 number of 0.304 used SPIB's default hyperparameters, which fall outside this tunable range. Second, and more importantly, identification metrics are uniformly zero: every one of the 12 configurations achieves Z@top3 $= 0/3$ and X_Z@top3 $= 0.000$. Even the strongest-tuned configuration in regime accuracy, with Cohen's d on its top latent reaching 1.361, fails to place any of its top-3 latents on the ground-truth regime-varying factors $\{z_0, z_1, z_2\}$.

Seed stability of the best configuration. We re-ran the best configuration ($\beta = 0.001$, $K = 4$) at two additional seeds. Regime accuracy is reproducible at 0.843 ± 0.028 (seeds 0/42/123), confirming

that SPIB is genuinely tunable on regime prediction. MCC is more variable (0.315 ± 0.113), but Z@top3 and X_Z@top3 remain identically zero across all three seeds, with different "winner" latents (z_7, z_2, z_0) selected as most regime-discriminative in each run.

Table 29: Seed stability of the best SPIB configuration ($\beta = 0.001$, $K = 4$).

Seed	Regime acc	MCC	Z@top3	X_Z@top3	Driver z^*	Cohen's d
0	0.881	0.199	0/3	0.000	z_7	0.568
42	0.814	0.464	0/3	0.000	z_2	1.361
123	0.834	0.281	0/3	0.000	z_0	0.958
Mean $\pm$ std	0.843 ± 0.028	0.315 ± 0.113	0/3 (all)	0.000 (all)	—	—

Interpretation. The dissociation between regime accuracy (tunable, reaches ≈ 0.84) and identification metrics (uniformly zero across 36 runs total) is consistent with SPIB's design objective. SPIB's information bottleneck optimizes the latent representation for predicting future state under regime conditioning, which yields embeddings that distinguish regimes from spurious correlates without requiring the latents to align with the underlying causal factors. This is a different goal from MOSAIC's: MOSAIC requires the latents themselves to be identifiable causal coordinates, and the additive sparse decoder then ties each one to a localized observed-variable support. Regime accuracy alone does not imply driver recovery, and Table 28 shows the gap quantitatively: even the best-tuned SPIB run is ≈ 13 points below MOSAIC on regime accuracy and entirely absent on Z@top3 and X_Z@top3.

This complementarity is consistent with how Section 7 positions SPIB: methods that automatically discover regime structure can supply the regime label that MOSAIC then interprets, but they do not themselves recover module-level causal supports.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All quantitative claims in the abstract and introduction are supported by experimental results in Section 6 and Appendix C. Limitations are discussed in Section 7.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the main limitations: regime association is distributional rather than interventional, support recovery targets main effects and localized modules, the current definition is binary, and nonlinear optimization has seed-to-seed variance. Computational scaling is discussed in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theorem 1 (module structure identifiability) and Theorem 2 (finite-sample support recovery) are proved in Section 4 and Appendix D. Assumptions are stated explicitly (A1, A1′, A2, sparse-additive incoherence, and uniqueness of the regime-discrepancy maximizer when a single latent is claimed).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All hyperparameters, architecture details, data generation procedures, and training protocols are provided in Sections 6, A, B, and C. The synthetic benchmark and synthetic tokamak disruption benchmark are fully specified and reproducible from the paper alone. Dataset sources for OMNI, Climate, and TEP are cited. The RNA dataset is generated by us, with full details provided in Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository (`https://github.com/shichengf/mosaic`) includes the MOSAIC model and training code, the synthetic energy-landscape data generator, and the RNA preprocessing pipeline (per-residue distance feature extraction, Q_{cp} computation, smoothing, transition-buffer removal, and windowing). This is sufficient to reproduce the synthetic benchmark in Section 6.1 end-to-end. Baseline configurations and hyperparameters are documented in Appendix B.1 and Table 7. Real-world datasets use publicly available sources: OMNI [King and Papitashvili, 2005], NOAA ERSSTv5 [Huang et al., 2017b,a], the Tennessee Eastman Process simulator [Downs and Vogel, 1993], and HP35 [Piana et al., 2012]; the synthetic tokamak disruption benchmark is generated by the construction in Appendix I.1 with seed 42.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All training details including data splits, optimizer, learning rate, batch size, early stopping criteria, and sparsity schedule are provided in Section 6 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Synthetic and MOSAIC RNA results use 5 seeds; RNA CRL/iVAE baselines use 4 seeds (42/123/456/789), as reported in Appendix H. Cross-domain results use 3 seeds per benchmark (seeds 0, 42, 123), with mean $\pm$ standard deviation reported in Table 4 and Appendix I.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on NVIDIA A40 GPUs (memory benchmark in Appendix A, Table 6). The synthetic benchmark takes on the order of 1–2 hours per seed at $\hat{n} = 12$ (approximately 90 minutes per seed at the $Z = 8$ speedup-benchmark setting); each RNA or cross-domain benchmark takes 1 to 8 hours per configuration. The reported experiments used approximately 5 days on 4 NVIDIA A40 GPUs, for an upper-bound total of about 480 GPU-hours. Preliminary experiments during method development consumed additional compute of comparable order; these exploratory runs are not included in the 480 GPU-hour estimate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work analyzes publicly available scientific datasets and does not involve human subjects, personal data, or dual-use concerns.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 7 discusses broader impact, including potential applications in space weather forecasting, fusion reactor diagnostics, climate monitoring, and chemical-process monitoring. No negative societal impacts are foreseen for this basic research contribution.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The models and data in this paper (regime-switching time series analysis) pose no risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers for the methods and datasets used in this work. The TDRL [Yao et al., 2022] and SPIB [Wang and Tiwary, 2021] implementations are released under the MIT License. For CtrlNS [Song et al., 2024] (`https://github.com/xiangchensong/ctrlns`), we obtained permission from one of its authors, Kun Zhang, to use, modify, and redistribute the code. OMNI [King and Papitashvili, 2005] is provided under NASA SPDF's CC0/public-domain data-use policy. NOAA ERSSTv5 [Huang et al., 2017b,a] permits unrestricted data access and use, with citation of the dataset and accompanying publication. The Braatz distribution of the Tennessee Eastman Process [Downs and Vogel, 1993] is provided under the University of Illinois permissive license included in its README. HP35 [Piana et al., 2012] is available under the D. E. Shaw Research data license for non-commercial research. RNA MD trajectories of the cUUCGg tetraloop were generated by the authors for the present study; prior structural-biology work on this system is cited where relevant scientific facts are referenced [Chen and García, 2013]. The tokamak disruption benchmark is a physics-motivated synthetic construction (seed=42), released with the code; variable names and module structure follow tokamak physics literature [De Vries et al., 2011], but the data is not drawn from any operational tokamak. See Appendix I.1 for construction details.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The MOSAIC implementation, synthetic data generator, and training and evaluation scripts are publicly available at `https://github.com/shichengf/mosaic`. The repository README documents installation, data preparation, training, evaluation, and licensing.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No, not as part of the research methodology. LLMs were not used in the proposed method, experiments, data generation, evaluation, or analysis pipeline. Any LLM use was limited to writing and editing assistance, which the checklist excludes from mandatory declaration.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook (`https://neurips.cc/Conferences/2026/MainTrackHandbook`) for what should or should not be described.