# PROCESS MEMORY BEYOND PREDICTION: AUDITING ALGEBRAIC STRUCTURE IN LEARNED DYNAMICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Does accurate prediction imply recovery of the process that generates the observations? We study this question in a controlled physical-representation benchmark with eight latent states and six primitive operations whose closure forms a 104-element transformation semigroup. The states are rendered as high-dimensional spatial fields under controlled carrier variation, letting us measure physical-state recoverability, predictive generalization, and process-structure recovery separately.

The physical states are highly recoverable from observation, with a frozen diagnostic decoder reaching $98.31\%$ mean validation accuracy, and learned models consistently outperform persistence across sealed distribution shifts. Yet the predictively trained Operation Channel Model (OCM), despite an explicit eight-state categorical representation and operation-specific transition channels, reaches only $15.72\%$ state alignment and $13.89\%$ sealed state-path accuracy against a $12.5\%$ chance reference. It exactly recovers only $1$ of $104$ transformations and none of the nine held-out transformations, and it does not outperform the validation-selected B4 baseline in the frozen primary joint-OOD comparison. An earlier simpler observation regime provides a positive control in which OCM recovers the full finite transformation structure at low-to-moderate noise, showing that the broader process-memory analysis can register exact structural recovery under less demanding conditions.

In Tier-C-v4, state recoverability and predictive learnability therefore do not, by themselves, guarantee recovery of the intended process algebra in the structured model tested here. When mechanistic or compositional structure is a modeling objective, process memory should be evaluated directly rather than inferred from forecasting accuracy.

## 1 INTRODUCTION

Predictive representations are widely used to model dynamical systems. World models and latent dynamics models compress high-dimensional observations into internal states that support forecasting, planning, or control Ha & Schmidhuber (2018); Hafner et al. (2019; 2020; 2025); Bardes et al. (2024). Accurate prediction, however, does not establish that the internal representation matches the process that generated the observations. Different latent organizations can support similar forecasts while encoding different states, transitions, or compositional structure.

This distinction matters when the generating process has a known structured description. Work on identifiable representation learning studies conditions under which latent variables or transition mechanisms can be recovered Locatello et al. (2019); Khemakhem et al. (2020); Hızlı et al. (2025); Zhang et al. (2026). Other approaches introduce relational, symmetry, Koopman, or semigroup structure directly into learned dynamics Kipf et al. (2018); Lusch et al. (2018); Li et al. (2020); Quessard et al. (2020); Chen & Wu (2023). These results show that structured dynamics can be recovered under suitable assumptions or encouraged through explicit constraints. We study a complementary question: when the underlying states are empirically recoverable and the dynamics are predictively learnable, does an explicitly structured model trained predictively recover the known process algebra?

We refer to this property as *process memory*. Process memory measures whether a learned representation preserves the state-transition structure that generates the observations, rather than only



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information sufficient for prediction. We construct a controlled benchmark with eight latent physical states and six primitive operations. Their closure contains 104 distinct transformations. Each state is rendered as a four-channel spatial field whose morphology varies across carriers. Because the complete transformation system is known, we can independently evaluate physical-state recoverability, predictive performance, state trajectories, exact transformations, held-out compositions, information classes, and pairwise information relations.

The benchmark is a controlled physical-representation diagnostic rather than a high-fidelity simulator of one material system. This design gives us exact process ground truth while retaining high-dimensional observations, physical carrier variation, composition shift, carrier shift, joint shift, and observation noise.

We study three questions. **RQ1:** Are the latent physical states empirically recoverable while the benchmark resists simple shortcuts? **RQ2:** Are the dynamics predictively learnable beyond persistence under distribution shift? **RQ3:** Does predictive training induce recovery of the known process algebra in OCM?

The results separate these questions. A frozen diagnostic decoder recovers the physical states with $98.31\%$ mean validation accuracy. All learned predictive models outperform persistence across the sealed test conditions. In contrast, OCM reaches only $15.72\%$ state alignment and $13.89\%$ state-path accuracy, compared with a $12.5\%$ chance reference. It exactly recovers $1/104$ transformations and $0/9$ held-out transformations. In the frozen primary comparison, OCM also does not outperform the validation-selected B4 predictive baseline.

Our contributions are threefold:

- We construct a high-dimensional benchmark with exact finite-process ground truth, held-out transformations, carrier variation, and controlled distribution shifts.
- We evaluate process memory at multiple levels, separating empirical state recoverability and predictive accuracy from symbolic state, transformation, and information-structure recovery.
- We show that highly recoverable states and learnable dynamics can coexist with weak recovery of the intended process algebra in a structured predictive model.

We do not claim that predictive models cannot recover structured dynamics. Our result is narrower: predictive success and empirical state recoverability do not, by themselves, guarantee the intended process representation under the architecture and objective studied here.

## 2 Related Work

**Predictive representations and world models.** World models and latent dynamics methods learn compact predictive representations for forecasting, planning, and control Ha & Schmidhuber (2018); Hafner et al. (2019; 2020; 2025); Bardes et al. (2024). Predictive objectives can also recover useful low-dimensional structure Recanatesi et al. (2021), and latent-operator methods support efficient prediction of complex physical systems Regazzoni et al. (2024); Kontolati et al. (2024). We ask a different question: whether predictive success recovers a specific known finite process rather than only a useful predictive representation.

**Identifiability and structured dynamics.** Latent variables are generally not identifiable without additional assumptions Locatello et al. (2019), while auxiliary information, temporal structure, and interventions can restore identifiability in suitable settings Khemakhem et al. (2020); Lippe et al. (2022; 2023). Recent work studies identification of latent states and transition mechanisms directly Hızlı et al. (2025); Nelson et al. (2026); Zhang et al. (2026). Structured dynamical models instead encode relational, symmetry, Koopman, or operator constraints Kipf et al. (2018); Lusch et al. (2018); Li et al. (2020); Quessard et al. (2020). Deep-OSG is particularly relevant because it explicitly incorporates semigroup consistency into learned dynamics Chen & Wu (2023). Our OCM does not receive the ground-truth semigroup multiplication table or transformation identities.

**Behavior versus internal structure.** Correct input-output behavior need not imply the expected internal computation. Liu et al. (2023) show that Transformers can implement finite-state behavior

through shortcut computations and analyze these alternatives using Krohn-Rhodes theory. We study a related issue using high-dimensional physical observations with complete finite-process ground truth. This lets us compare prediction, state recoverability, and exact process recovery within the same controlled system.

# 3 Benchmark and Problem Setup

## 3.1 Finite Process and Process Memory

Let

$$\mathcal{S} = \{s_1, \ldots, s_8\}$$

denote eight latent states and let

$$\mathcal{G} = \{\texttt{identity}, \texttt{cycle}, \texttt{pair_collapse}, \texttt{half_collapse}, \texttt{parity_collapse}, \texttt{reset}\}$$

denote six primitive operations. Each operation is a deterministic map $g : \mathcal{S} \to \mathcal{S}$.

For an operation word $w = (g_1, \ldots, g_L)$, the induced transformation is

$$T_w = g_L \circ \cdots \circ g_1.$$

Closing the primitive operations under composition gives

$$\mathcal{T} = \langle \mathcal{G} \rangle, \qquad |\mathcal{T}| = 104.$$

Different operation words can implement the same transformation. The 104 transformations also induce nine information classes according to the partitions they create on the eight-state space. Because the complete transformation system is enumerable, we can evaluate process memory through state alignment, state trajectories, exact transformations, information classes, and pairwise information relations. Additional algebraic details are given in Appendix A.

## 3.2 Physical Observations and Splits

For carrier $c$ and latent state $s$, we generate a four-channel spatial field

$$\mathbf{x}_{c,s} = \Phi(c, s) \in \mathbb{R}^{4\times 32\times 32}.$$

Carrier parameters vary morphology, including characteristic scale, interfaces, orientation statistics, and defect structure. The latent state controls factors including coarsening, anisotropy, and defect recovery. The same latent state must therefore remain recoverable across varied physical realizations.

We partition the 104 transformations into

$$86/9/9$$

train, validation, and test transformations. The complete symbolic catalog contains 3,144 operation sequences and 25,152 state trajectories. The balanced predictive split contains

$$1,376/144/144$$

train, validation, and test sequences.

We evaluate four sealed test conditions: `test_iid_pairing`, `test_composition`, `test_carrier`, and `test_joint`. The final test set contains 32 held-out carriers. Each condition is evaluated at noise fractions

$$\{0, 0.1, 0.25, 0.5, 1.0\}.$$

Before predictive evaluation, we verify that the states remain recoverable and that simple global-statistics and affine shortcuts do not reconstruct the full process. These audits are central to RQ1 because weak process recovery would otherwise be difficult to interpret.

Table 1: Model families used in the frozen Tier-C-v4 protocol.

| Model | Representation and transition model |
|---|---|
| B0 | Spatial persistence |
| B1 | PCA-128 field representation + affine operators |
| B2 | Operation-conditioned convolutional residual model |
| B3 | Convolutional recurrent sequence model |
| B4 | Patch-transformer sequence model |
| B5 | Continuous latent spatial operator |
| OCM | Eight-state operation-channel model |

# 4 Models and Evaluation

## 4.1 Predictive Models

All models receive the same normalized physical fields, but use different internal representations. We compare six predictive baselines with OCM.

OCM places an explicit eight-state categorical process between a convolutional encoder and decoder. Each primitive operation $g$ has a row-stochastic transition channel

$$\mathbf{M}_g \in [0,1]^{8\times 8}, \qquad \mathbf{M}_g\mathbf{1} = \mathbf{1}.$$

If $\mathbf{p}_t \in \Delta^7$ is the latent-state distribution, then

$$\mathbf{p}_{t+1} = \mathbf{p}_t\mathbf{M}_{g_t}.$$

The eight categories are not supervised with ground-truth state IDs.

The common predictive target is the noisy normalized physical field. OCM uses the frozen objective

$$\mathcal{L}_{\mathrm{OCM}} = \mathcal{L}_{\mathrm{field}} + \lambda_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{roll}}\mathcal{L}_{\mathrm{roll}} + \lambda_{\mathrm{det}}\mathcal{L}_{\mathrm{det}}.$$

The transition term encourages consistency under individual primitive operations, the rollout term encourages consistency across multi-step operation sequences, and the determinism term discourages diffuse transition channels. None of these terms reveals the ground-truth state permutation, the identity of the 104 transformations, equivalence between alternative operation words, the nine information classes, or the semigroup multiplication table. We therefore evaluate process-algebra recovery independently of the labels used to define the ground-truth structure.

We select configurations using validation noisy-target rollout MSE only. B2–B5 and OCM are evaluated across five frozen seeds

$$\{11, 23, 37, 53, 71\}$$

and five noise levels. B1 is deterministic and is fit once per noise level. This gives 130 frozen trainable fits. Full architecture, optimization, and configuration details appear in Appendix B.

## 4.2 Predictive and Structural Evaluation

Our primary predictive metric is noisy-target rollout MSE. The frozen confirmatory comparison evaluates OCM against validation-selected B4 on `test_joint` at noise fraction $0.25$. We average error across the five neural seeds for each of the 144 test sequences and use a paired 10,000-resample bootstrap over sequences. OCM superiority requires a negative mean OCM-minus-B4 difference and a negative upper endpoint of the $95\%$ confidence interval.

We audit process structure only for OCM because it exposes a native eight-state categorical representation and explicit primitive transition channels. We align OCM categories to ground-truth states using training carriers only. We then extract hard primitive maps from the frozen operation channels and compose them to obtain predicted maps for all 104 transformations.

We report state alignment, sealed state-path accuracy, exact transformation recovery, held-out transformation recovery, information-class accuracy, and pairwise information-order balanced accuracy. These metrics test different levels of the learned process representation. We do not claim that B1–B5

Table 2: Primary joint-OOD predictive results. Lower MSE is better. B4 is the validation-selected confirmatory comparator; B3 is the lowest descriptive test result.

| Model | Rollout MSE | Improvement vs. B0 |
|---|---|---|
| B0 | 1.216326 | – |
| B3 | **0.994846** | 18.21% |
| B4 | 1.000112 | 17.78% |
| OCM | 1.039329 | 14.55% |

contain no decodable process information because we do not train post-hoc structural probes on their continuous representations.

All model selection, checkpoints, evaluation interfaces, and the primary comparison were frozen before the single sealed test opening. Additional metric definitions and protocol details are provided in Appendix C.

## 5 RESULTS

**RQ1: Are the physical states recoverable without trivial shortcuts?** The final benchmark preserves a clear morphology shift, with mean coarse-to-fine characteristic-length ratio $2.844$ and mean interface-density ratio $0.471$. An independent frozen diagnostic decoder achieves $98.31\%$ mean validation state accuracy, with $97.66\%$ accuracy in the worst diagnostic seed. In contrast, a classifier using only 28 global field statistics obtains $12.5\%$, equal to chance for eight balanced states. A PCA-128 affine diagnostic exactly recovers only $40.49\%$ of primitive transitions and recovers all 104 transformations on $0\%$ of validation carriers. Thus, the state is highly recoverable while the predefined simple shortcuts do not reconstruct the complete process. Full RQ1 diagnostics are reported in Appendix D.

**RQ2: Are the dynamics predictively learnable beyond persistence?** Across all 20 combinations of four test conditions and five noise levels, every learned model outperforms persistence in noisy-target rollout MSE. Table 2 reports the frozen primary condition, `test_joint` at noise fraction $0.25$.

The remaining learned models lie between $1.022196$ and $1.034443$. For the frozen OCM-versus-B4 comparison,

$$\Delta_{\mathrm{OCM-B4}} = +0.039217,$$

with $95\%$ paired-bootstrap confidence interval

$$[+0.034168, +0.044454].$$

OCM is better on only $12/144$ sequences. The confirmatory comparison therefore does not support OCM predictive superiority.

Figure 1(a) shows the joint-OOD noise sweep. Learned models retain a clear advantage over persistence as noise increases.

**RQ3: Does predictive training induce recovery of the known process algebra in OCM?** Figure 1(b) shows the central representation gap. The independent physical-state decoder reaches $98.31\%$, while OCM training-carrier state alignment is only $15.72\%$. Sealed state-path accuracy is $13.89\%$, close to the $12.5\%$ eight-state chance reference.

Table 3 summarizes state recoverability and process-memory recovery. The gap is not limited to state labels. OCM exactly recovers only

$$1/104 = 0.96\%$$

of the ground-truth transformations and

$$0/9$$

held-out transformations. Information-class accuracy is also $0.96\%$. Pairwise information-order balanced accuracy is $38.24\%$. This exceeds the $25\%$ balanced-chance reference for four relation

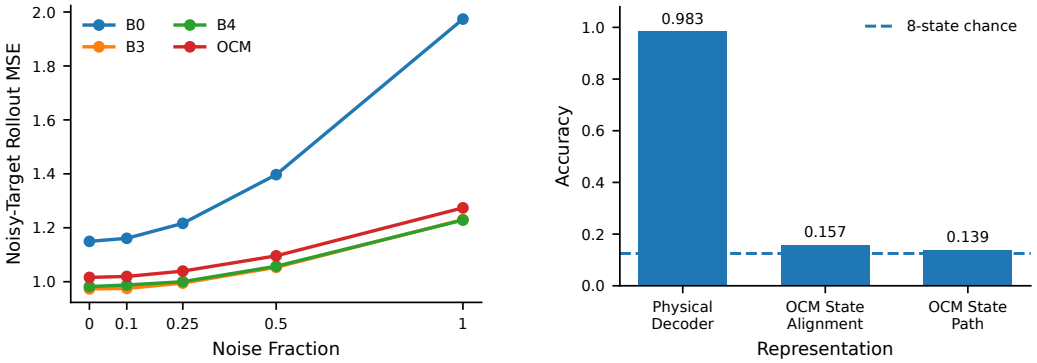


Figure 1: Key empirical results. (a) Joint-OOD noisy-target rollout MSE across noise levels. (b) Physical-state recoverability compared with OCM state alignment and sealed state-path recovery.

Table 3: State-recoverability and process-memory diagnostics for Tier-C-v4. The physical-state decoder is an independent frozen diagnostic; the remaining model rows evaluate the frozen OCM.

| Diagnostic | Result |
|---|---|
| Physical-state decoder | 98.31% |
| OCM state alignment | 15.72% |
| OCM state-path accuracy | 13.89% |
| Eight-state chance | 12.50% |
| Exact transformations | 1/104 (0.96%) |
| Held-out transformations | 0/9 |
| Information-class accuracy | 0.96% |
| Information-order balanced accuracy | 38.24% |
| Four-relation balanced chance | 25.00% |

classes, indicating that some coarse relation information is retained, but it remains far from complete recovery.

These results are not explained by the exact-map criterion alone. Exact transformation recovery requires every state destination to match and is therefore stringent. However, state-path recovery remains near the eight-state chance reference, information-class recovery remains at $0.96\%$, and the information-order audit reaches only $38.24\%$. Several levels of the structural audit therefore show that OCM does not recover the intended finite process representation.

An earlier simpler observation regime provides a separate positive control. This experiment predates the Tier-C-v4 sealed analysis and is reported only as a pre-existing capability control. Across five seeds, the OCM structural analysis achieves perfect primitive mapping, exact transformation, partition, and relation recovery at noise fractions $0$, $0.1$, and $0.25$. We do not treat those results as directly comparable Tier-C-v4 performance.

The positive control uses the same underlying eight-state transformation system and OCM model family, but a simpler observation regime. It therefore shows that the broader OCM process-memory analysis can register exact finite-process recovery when the representation aligns with the underlying process. Detailed positive-control results are reported in Appendix E.

## 6 Discussion and Limitations

Our results separate state recoverability, predictive learnability, and process-memory recovery. The first two are strongly supported in Tier-C-v4. The physical state is recoverable with $98.31\%$ diagnostic accuracy, and all learned predictive models outperform persistence. OCM nevertheless remains close to chance for symbolic state tracking and recovers almost none of the known transformation algebra.

Predictive performance therefore does not establish process equivalence in the structured model studied here.

The state-decoding result is important for interpreting the negative structural outcome. If the eight states were not distinguishable from the observations, weak OCM alignment could follow directly from missing information. Instead, the state is strongly recoverable by an independent diagnostic. The gap between $98.31\%$ external state recovery and $15.72\%$ OCM state alignment shows that information available in the observation need not be organized according to the same state variable inside a predictively trained representation.

The structural conclusion also does not depend only on exact transformation recovery. Exact recovery is strict because one wrong destination makes an entire map incorrect. We therefore also evaluate state trajectories, information partitions, and pairwise information relations. State-path accuracy remains close to the $12.5\%$ eight-state chance reference, and information-class accuracy is only $0.96\%$. Information-order balanced accuracy reaches $38.24\%$, above the $25\%$ four-class chance reference but well below complete relation recovery. These complementary diagnostics therefore provide a more nuanced but consistent picture: OCM retains limited structural information without recovering the intended finite process.

At the same time, our audit uses the predefined hard OCM readout. We therefore do not rule out every possible notion of approximate structure in the soft transition probabilities. Soft transition distances, posterior calibration, or probabilistic composition metrics could reveal additional partial structure, but they were not part of the frozen evaluation protocol.

The earlier positive control further narrows the interpretation. In a simpler observation regime, the OCM structural analysis recovers the full finite process exactly across five seeds at noise fractions up to $0.25$. Structural recovery then deteriorates at larger noise. This shows that the evaluation pipeline can register exact recovery when the learned representation aligns with the process. It does not establish that the same recovery should occur in Tier-C-v4, where the observation problem is substantially more demanding.

Our findings do not imply that architecture or prediction can never recover process structure. OCM provides the capacity for an eight-state process, but its objective does not supply true state IDs, equivalent-word constraints, transformation identities, or the semigroup multiplication table. Methods that explicitly enforce compositional or semigroup consistency Li et al. (2020); Chen & Wu (2023) may behave differently. Testing such interventions is a separate question that we leave for future work.

**Limitations.** Tier-C-v4 contains one finite eight-state process and one family of four-channel physical generators. OCM is also only one discrete process architecture. We structurally audit OCM because it exposes an intrinsic eight-state transition system, but we do not train standardized probes on the continuous representations of B1–B5. Our state-recoverability result is empirical rather than a theoretical identifiability theorem, and our structural metrics do not exhaust all possible soft notions of similarity. We therefore restrict our conclusion to the benchmark, architecture, and training objective studied here. Expanded limitations and protocol details are provided in the appendix.

## 7 Conclusion

We introduced a controlled benchmark in which state recoverability, predictive learning, and exact process-memory recovery can be evaluated separately. The physical states are highly recoverable and the dynamics are learnable beyond persistence, yet the structured OCM remains near chance for symbolic state tracking and recovers only $1/104$ transformations and no held-out transformations. We therefore find that predictive success and state recoverability do not guarantee recovery of the intended process algebra in the model studied here. When mechanistic or compositional structure matters, we argue that process memory should be evaluated directly rather than inferred from forecasting accuracy.

## Reproducibility Statement

We designed the experimental protocol to separate development from sealed evaluation and to preserve the full experimental record. The appendix reports the benchmark construction, transformation and sequence splits, model-family definitions, selected configurations, random seeds, noise levels, optimization settings, predictive metrics, structural-readout procedure, bootstrap comparison, and earlier-regime positive control. Model selection and the primary OCM-versus-B4 comparison were frozen before the single Tier-C-v4 test opening. The benchmark and implementation have been archived in public research repositories. Because the existing repository records contain author-identifying metadata, their canonical links are omitted during double-blind review and will be added after de-anonymization.

## AI Use Statement

We used OpenAI ChatGPT as an assistive tool during this project. Its use included research brainstorming and refinement of research questions, code drafting and debugging assistance, literature search and organization, and drafting and editing of manuscript text.

All experiments were executed by the authors on their own computing infrastructure. The authors inspected the experimental outputs, verified the reported numerical results against the frozen artifacts, checked the cited literature and bibliographic information, and reviewed all AI-assisted code and manuscript text. Generative AI was not used as a substitute for executing the reported experiments or for fabricating experimental measurements. The authors take full responsibility for the methodology, results, claims, citations, code, and final manuscript.

## References


Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

Junfeng Chen and Kailiang Wu. Deep-OSG: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493:112498, 2023. doi: 10.1016/j.jcp.2023.112498.

David Ha and Jürgen Schmidhuber. World models, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Çağlar Hızlı, Çağatay Yıldız, Matthias Bethge, S. T. John, and Pekka Marttinen. Identifying latent state transitions in non-linear dynamical systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d16mJDyQN6.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020.

Thomas N. Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697. PMLR, 2018.

Katiana Kontolati, Somdatta Goswami, George Em Karniadakis, and Michael D. Shields. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. *Nature Communications*, 15:5101, 2024. doi: 10.1038/s41467-024-49411-w.

Yunzhu Li, Hao He, Jiajun Wu, Dina Katabi, and Antonio Torralba. Learning compositional Koopman operators for model-based control. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1ldzA4tPr.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603. PMLR, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=itZ6ggvMnzS.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning, 2026. URL https://arxiv.org/abs/2603.11970.

Robin Quessard, Thomas D. Barrett, and William R. Clements. Learning disentangled representations and group structure of dynamical environments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19727–19737, 2020.

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12:1417, 2021. doi: 10.1038/s41467-021-21696-1.

Francesco Regazzoni, Stefano Pagani, Matteo Salvador, Luca Dede', and Alfio Quarteroni. Learning the intrinsic dynamics of spatio-temporal processes through latent dynamics networks. *Nature Communications*, 15:1834, 2024. doi: 10.1038/s41467-024-45323-x.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models, 2026. URL https://arxiv.org/abs/2607.22430.

## A Additional Benchmark Details

The Tier-C-v4 benchmark contains eight latent states and six primitive operations:

$$\mathcal{G} = \{\texttt{identity}, \texttt{cycle}, \texttt{pair_collapse}, \texttt{half_collapse}, \texttt{parity_collapse}, \texttt{reset}\}.$$

Their closure contains 104 transformations. For transformation $T \in \mathcal{T}$, we define the induced information partition

$$\Pi(T) = \left\{T^{-1}(s) : s \in \mathcal{S},\ T^{-1}(s) \neq \emptyset\right\}.$$

The 104 transformations form nine information classes. Across ordered distinct transformation pairs there are 10,712 relations: 2,008 equivalent, 1,728 more informative, 1,728 less informative, and 5,248 incomparable.

The complete symbolic catalog contains 3,144 operation sequences and 25,152 state trajectories. The transformation split is 86/9/9 for train/validation/test, while the balanced predictive sequence split is 1,376/144/144.

The final sealed test set contains 32 held-out carriers and four evaluation conditions: `test_iid_pairing`, `test_composition`, `test_carrier`, and `test_joint`. All four are evaluated at noise fractions

$$0,\ 0.1,\ 0.25,\ 0.5,\ 1.0.$$

## B Model and Training Details

The frozen model registry contains the following families:

Table 4: Frozen model registry.

| Model | Description |
|---|---|
| B0 | Spatial persistence; returns the initial field as the prediction. |
| B1 | Training-only PCA-128 representation followed by one ridge-regularized affine operator per primitive operation. |
| B2 | Convolutional encoder, operation embedding, residual spatial transition, and convolutional decoder. |
| B3 | Convolutional field encoder with recurrent operation-sequence processing. |
| B4 | Patch-token spatial encoder with operation tokens and Transformer sequence processing. |
| B5 | Continuous latent field representation with learned operation-conditioned latent operators. |
| OCM | Convolutional encoder and decoder surrounding eight-state row-stochastic operation channels. |

The frozen selected configurations are

$$\text{B1-c02},\ \text{B2-c01},\ \text{B3-c02},\ \text{B4-c03},\ \text{B5-c02},\ \text{OCM-c03}.$$

For B2-B5 and OCM, the frozen neural training contract uses five seeds

$$\{11, 23, 37, 53, 71\}.$$

Training uses effective batches of 256 examples through micro-batches of four and gradient accumulation over 64 micro-batches. Runs use at most 200 epochs, early-stopping patience 25, gradient-norm clipping at 5, and no mixed precision. B1 is deterministic and is fit once per noise level.

Across five noise levels, this yields 125 neural final fits for B2-B5 and OCM plus five deterministic B1 fits, for 130 frozen trainable fits in total.

## C Evaluation Details

For a valid rollout of length $L$, noisy-target rollout MSE is

$$\mathrm{MSE}_{\mathrm{roll}} = \frac{1}{L}\sum_{k=1}^{L}\frac{1}{|\Omega|}\,\|\widehat{\mathbf{x}}_{t+k} - \widetilde{\mathbf{x}}_{t+k}\|_2^2\,.$$

The primary comparison uses OCM and B4 at

$$\texttt{test_joint}, \qquad \sigma = 0.25.$$

For each of 144 sequences, we average error across the five neural seeds and compute

$$d_i = \mathrm{MSE}_i^{\mathrm{OCM}} - \mathrm{MSE}_i^{\mathrm{B4}}.$$

We estimate uncertainty using 10,000 paired bootstrap resamples with frozen seed 73011.

For OCM state alignment, we take

$$\widehat{z} = \arg\max_j p_\theta(z = j \mid \mathbf{x})$$

and obtain the best one-to-one assignment between OCM categories and ground-truth states using training carriers only.

State-path accuracy is

$$\mathrm{Acc}_{\mathrm{path}} = \frac{\sum_t \mathbb{I}[\widehat{s}_t = s_t]}{\text{number of valid trajectory states}}.$$

We extract a deterministic primitive map from each OCM operation channel by taking the maximum-probability destination for every source state. We compose these primitive maps according to canonical operation words. Transformation $T$ is exactly recovered only when

$$\widehat{T}(s) = T(s) \qquad \forall s \in \mathcal{S}.$$

Information-class accuracy compares $\Pi(\widehat{T})$ with $\Pi(T)$. For the four pairwise information-relation classes, we report balanced accuracy,

$$\mathrm{BA}_{\mathrm{info}} = \frac{1}{4}\sum_{r=1}^{4}\mathrm{Recall}_r.$$

## D Full Tier-C-v4 Results

### D.1 Benchmark Validity

Table 5: Tier-C-v4 validity and state-recoverability diagnostics.

| Diagnostic | Result | Interpretation |
|---|---|---|
| Mean characteristic-length ratio | 2.844 | Fine/coarse morphology separation |
| Minimum characteristic-length ratio | 2.535 | Separation holds across carriers |
| Mean interface-density ratio | 0.471 | Controlled coarsening shift |
| Carrier-scale association (Spearman $\rho$) | -0.07596 | Weak residual association |
| State-decoder validation accuracy | 98.31% | Eight states highly recoverable |
| Worst-seed state-decoder accuracy | 97.66% | Recovery is seed-stable |
| Coarsening-factor validation accuracy | 99.87% | Morphology factor is recoverable |
| Global-statistics shortcut accuracy | 12.50% | At 8-state chance |
| PCA-128 affine latent MSE | 10.4335 | Affine shortcut fails |
| Exact primitive recovery | 40.49% | Incomplete affine recovery |
| Validation carriers with all 104 exact | 0% | No exact full-algebra shortcut |

Table 6: Noisy-target rollout MSE at the frozen primary joint-OOD condition.

| Model | Noisy-target rollout MSE | Improvement vs. B0 |
|---|---|---|
| B0 | 1.216326 | – |
| B1 | 1.022196 | 15.96% |
| B2 | 1.030627 | 15.27% |
| **B3** | **0.994846** | 18.21% |
| B4 | 1.000112 | 17.78% |
| B5 | 1.034443 | 14.95% |
| OCM | 1.039329 | 14.55% |

### D.2 Primary Predictive Condition

The frozen OCM-minus-B4 difference is

$$+0.039217$$

with $95\%$ paired-bootstrap interval

$$[+0.034168, +0.044454].$$

OCM is better on 12 of 144 sequences.

### D.3 Structural Diagnostics

Table 7: State recoverability and Tier-C-v4 OCM structural recovery.

| Diagnostic | Result |
|---|---|
| Physical decoder validation accuracy | 98.31% |
| OCM training-carrier state alignment | 15.72% |
| OCM sealed state-path accuracy | 13.89% |
| 8-state chance reference | 12.50% |
| Exact transformation recovery | 1/104 (0.96%) |
| Held-out exact transformation recovery | 0/9 (0%) |
| Information-class accuracy | 0.96% |
| Information-order balanced accuracy | 38.24% |

Averaged across noise levels, state-path accuracy is approximately $15.0\%$ for IID pairing, $15.3\%$ for composition shift, $12.7\%$ for carrier shift, and $12.6\%$ for joint shift.

## E Earlier-Regime Structural Positive Control

We report an earlier, simpler observation regime only as a positive control for structural recovery. It is not directly comparable with Tier-C-v4.

Across five seeds, the OCM structural analysis reports perfect primitive mapping, exact transformation recovery, partition recovery, and relation recovery at noise fractions $0$, $0.1$, and $0.25$.

Table 8: Earlier-regime OCM positive control. Values are means across five frozen seeds. Test relation BA denotes balanced accuracy on test-involved transformation relations.

| Noise | Primitive Map | Exact Transform. | Test Exact | Test Relation BA |
|---|---|---|---|---|
| 0.00 | 1.000 | 1.0000 | 1.000 | 1.0000 |
| 0.10 | 1.000 | 1.0000 | 1.000 | 1.0000 |
| 0.25 | 1.000 | 1.0000 | 1.000 | 1.0000 |
| 0.50 | 0.900 | 0.0615 | 0.000 | 0.2697 |
| 1.00 | 0.5125 | 0.0288 | 0.000 | 0.2574 |

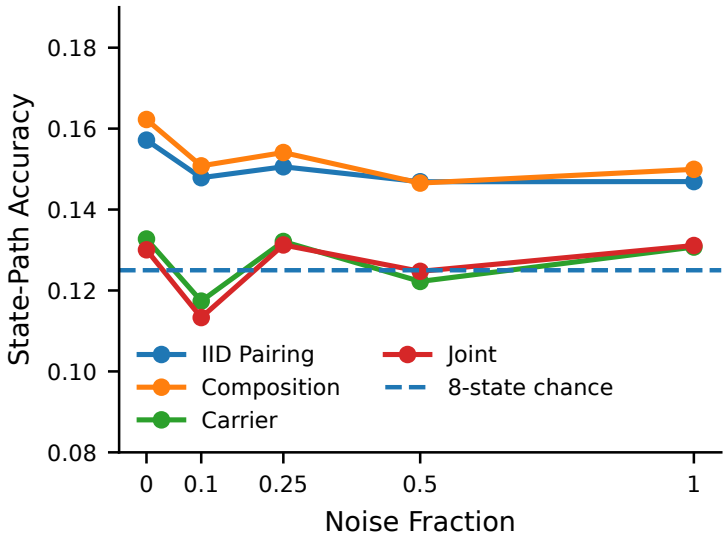


Figure 2: OCM state-path recovery across test conditions and noise levels. The eight-state chance reference is $12.5\%$.

The positive control establishes that the process-memory analysis can record exact structural recovery in a simpler regime. It does not imply that Tier-C-v4 should admit the same recovery, and we do not use it in the frozen Tier-C-v4 confirmatory comparison.

## F Additional Limitations

Tier-C-v4 studies one eight-state finite process and one family of four-channel physical generators. Additional continuous-state, stochastic, or partially observed systems would be needed to establish broader generality.

OCM is one parameterization of a discrete process-memory model. We do not claim that other architectures or objectives would produce the same structural result.

We do not train a Tier-C-v4 variant with explicit semigroup supervision, equivalent-word consistency, or carrier-invariant state supervision. Consequently, the present experiments do not determine which structural constraint would be sufficient to improve recovery.

We also do not train standardized post-hoc probes on B1-B5. Their continuous representations may contain decodable state or process information even though they do not expose a native finite-state transition system.

Finally, our structural audit uses a predefined hard readout. Soft transition-probability metrics could reveal approximate information that is not captured by hard symbolic recovery. The positive-control experiment shows that the hard analysis can register exact structure, but it does not remove this limitation.