Causal Emergent State Representation Learning for Control

Kenzo Clauw[1], **Fernando E. Rosas**[2,3,4], **Daniele Marinazzo**[1]
kenzoclauw@gmail.com

[1]**Department of Data Analysis, Ghent University**
[2]**Department of Informatics, University of Sussex**
[3]**Department of Brain Science, Imperial College London**
[4]**Centre for Eudaimonia and Human Flourishing, University of Oxford**

Abstract

Many real-world reinforcement learning problems involve controlling complex systems (swarming, fluid dynamics, chemistry) where the task-relevant state is an emergent collective variable not carried by any individual component. In this work, we introduce causal-emergent state representation learning (CESRL): the problem of learning a collective macro representation from micro-level particle observations, rather than from any individual part. Most state representation learning methods compress observations into a compact representation without explicitly targeting collective dynamics between particles, making them sample-inefficient when the task-relevant signal is encoded only at the collective level. We address the CESRL with a self-supervised information-theoretic objective trained on reward-free data. On a controlled discrete benchmark, our approach recovers the emergent representation where standard methods cannot. On a realistic physics benchmark, it outperforms existing baselines in sample efficiency.

1 Introduction

Imagine guiding a flock of birds. The agent does not observe the flock's centre of mass or heading directly. It only observes the bird-level microstate: the positions and velocities of the individual birds. At this micro level, each bird jitters, reacts to nearby neighbours, and can change direction abruptly, so no single bird provides a stable predictive signal for control. What moves smoothly is the flock as a whole: its centre of mass and heading. Figure 1 illustrates this setting: successful control depends on an emergent collective state, not on any one bird.

This example is illustrative of applying RL on real-world physical problems such as fluid dynamics (Xian et al., 2023), soft robotics (Lin et al., 2021), materials design (Govindarajan et al., 2025), and chemistry (Sridharan et al., 2024). In such environments, the agent observes a system of many interacting particles, while the task-relevant state is only visible collectively and not in any single component alone. In this work, we introduce causal-emergent state representation learning (CESRL) as an RL problem in which the agent must discover, from particle-level observations, a collective variable for control whose predictive structure is present only at the level of the whole system (Rosas et al., 2020).

State representation learning is often framed as learning a compact, task-relevant latent from complex observations, frequently visual observations (Botteghi et al., 2024; Echchahed & Castro, 2025). This view is well matched to settings in which the useful low-dimensional structure can be recovered directly from observation-level regularities, for example by decoding, filtering, or compressing an underlying state (Du et al., 2019). Our setting is different: the agent must infer a collective state that exists only at the level of the full particle system and not in any individual part alone. This

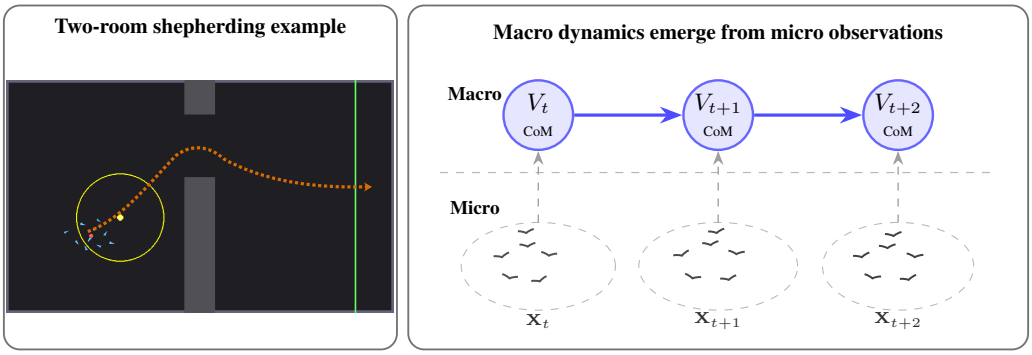


Figure 1: The agent observes the full particle system $\mathbf{X}_t$, but the control-relevant variable is a coarse-grained macro state $V_t = \phi(\mathbf{X}_t)$, such as the flock's centre of mass and heading. **Left:** a single shepherd must gather the flock and steer its centre of mass through the opening and behind the goal line. **Right:** after coarse-graining, the macro is predictive of its own future, whereas no individual bird is.

distinction is most accurate when the emergent macro is cleanly separated from the micro level, though more realistic environments may still leave residual observation-level cues that partially help standard methods.

This work studies CESRL in a setting of offline training from reward-free data and downstream evaluation (Stooke et al., 2020). We learn these causal emergent representations based on a self-supervised information-theoretic objective (Rosas et al., 2020; McSharry et al., 2024). Our contributions are:

1. We introduce two benchmarks: the controlled discrete XOR Parity Chain, where the causal-emergent feature is known exactly, and the more realistic continuous Vortex Pair Wind CartPole.
2. In the XOR Parity Chain, we show that current methods fail to recover the causal-emergent state needed for control.
3. In Vortex Pair Wind CartPole, we show that useful causal-emergent representations can arise in a realistic physical system and improve downstream control over existing baselines, while the partial success of standard methods suggest that stronger realistic benchmarks are still needed.

2 Background and Problem Statement

We formalise CESRL in a Markovian setting with two coupled levels of description. The environment evolves at a micro level of interacting particles, while control depends on a lower-dimensional macro state induced by a coarse-graining of the full microstate. We then define the learning problem of recovering this macro state from reward-free data for downstream control.

2.1 Environment

We consider an infinite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P(s' \mid s, a)$ the transition kernel, $r(s, a)$ the reward function, ρ the initial-state distribution, and $\gamma \in (0, 1)$ the discount factor. A policy $\pi(a \mid s)$ maps states to action distributions, and the control objective is to maximise the expected discounted return $\mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}]$.

Micro MDP In our setting, the full state factorises as $S_t = (Z_t, X_t)$, where $X_t \in \mathcal{X}$ is a micro state and $Z_t \in \mathcal{Z}$ denotes any directly observed task-relevant variables outside the particle system. The micro state is itself factorised as $X_t = (X_t^1, \ldots, X_t^n)$ with $\mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^n$, where each X_t^i is one interacting micro-component. Thus the micro-level environment is the MDP $\mathcal{M}_{\text{micro}} =$

$(\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ with state space $\mathcal{S} = \mathcal{Z} \times \mathcal{X}$. A policy on the micro MDP acts on the full particle-level state, that is, as $\pi(a \mid z, x)$ with $x = (x^1, \ldots, x^n)$.

Coarse-graining Based on McSharry et al. (2024); Rosas et al. (2020), we define a coarse-graining as a map $\phi^\star : \mathcal{X} \to \mathcal{V}$ from the micro state to a lower-dimensional macro variable. The induced macro variable $V_t = \phi^\star(X_t)$ is supervenient on the micro state, in the sense that it is fully determined by the joint micro configuration. In the setting of interest here, it is also assumed to carry unique predictive information about future system evolution that is relevant for control and not available from any individual micro-component alone.

Macro MDP Given a coarse-graining $\phi^\star$, define the macro state $U_t = (Z_t, V_t)$ with $V_t = \phi^\star(X_t)$. This induces a macro-level MDP $\mathcal{M}_{\text{macro}} = (\mathcal{U}, \mathcal{A}, P_{\phi^\star}, r_{\phi^\star}, \rho_{\phi^\star}, \gamma)$ with state space $\mathcal{U} = \mathcal{Z} \times \mathcal{V}$. Here, $P_{\phi^\star}(u' \mid u, a)$ and $r_{\phi^\star}(u, a)$ denote the transition and reward model induced by the coarse-graining. In this work, we focus on the case where the coarse-graining is sufficiently self-contained for control, so that an optimal policy can be expressed in terms of U_t alone, as $\pi^\star(a \mid u)$ on the macro MDP. The agent, however, acts from the micro-level state (or observations derived from it), and the representation learning problem is to recover from that micro description a feature that supports macro-level control.

2.2 Reward-Free Pre-Training and Downstream Evaluation

Following the general state representation learning formulation (Echchahed & Castro, 2025), we seek an encoder $f_\theta : \mathcal{X} \to \hat{\mathcal{V}}$ that maps the current micro state to a learned macro representation $\hat{V}_t = f_\theta(X_t)$. The downstream policy then acts on the reduced state $\hat{U}_t = (Z_t, \hat{V}_t)$. For simplicity, we assume a single dominant causal-emergent macro variable, and do not study disentangling multiple emergent factors.

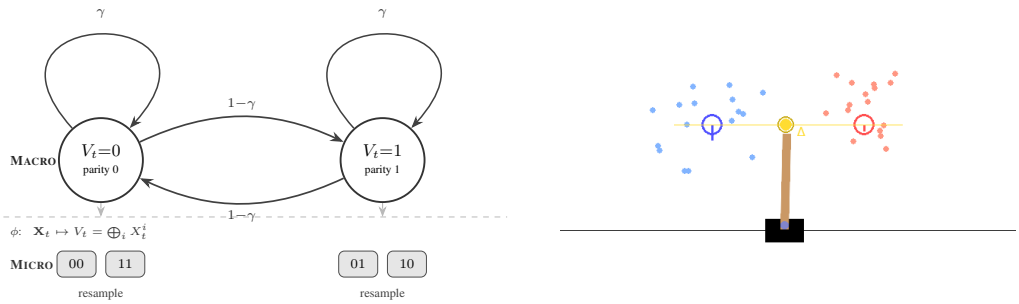
To learn this representation, we use an unsupervised pre-training setting with an offline dataset of reward-free transitions collected by a random or otherwise non-task-aware behaviour policy before downstream learning. We write this dataset as $\mathcal{D} = \{(X_t, A_t, X_{t+1})\}$, or more generally as $\mathcal{D} = \{(Z_t, X_t, A_t, Z_{t+1}, X_{t+1})\}$ when directly observed variables are present. The encoder is trained only on the micro observations X_t in this phase, with no reward information. After pre-training, the learned representation is used for downstream policy learning. In our experiments, the downstream task is based on PPO (see C for details).

3 Learning Causal-Emergent Representations

Following McSharry et al. (2024); Rosas et al. (2020), a good macro representation $V_t = \phi(\mathbf{X}_t)$ should be predictive of its own future, while that predictive power should not be reducible to any individual micro-component. We operationalise this with the Ψ objective:

$$\Psi(\phi) \;:=\; \underbrace{I(V_t; V_{t+1})}_{\text{predictive information}} - \beta \sum_{i=1}^{n} \underbrace{I(X_t^i; V_{t+1})}_{\text{pairwise micro contributions}} \quad , \tag{1}$$

where the first term rewards macro-level predictive information and the second penalises dependence on individual components. We introduce β as a hyperparameter to control the strength of this penalty. Setting $\beta = 1$ recovers the original formulation, while $\beta = 0$ reduces the objective to predictive information only. This flexibility is useful for the continuous benchmark, where the control-relevant macro is only approximately emergent and some individual-level cues can still be informative (see C.1 for details).



(a) (b)

Figure 2: **From latent macro states to observed micro dynamics. Left:** XOR Parity Chain, where a latent parity state evolves as a Markov chain and the micro state is resampled from the corresponding parity-consistent set. **Right:** Vortex Pair Wind CartPole, where a slowly varying hidden physical state drives the wind and is encoded collectively in the particle system. Full details are provided in Appendix A and Appendix B.

4 Can SRL Methods Learn Causal-Emergent Features?

The goal of this experiment is to isolate the role of a single causal-emergent feature: an agent that recovers it achieves the optimal policy. To do so, we introduce XOR Parity Chain (see Figure 2, left). The XOR parity $V_t = \bigoplus_{i=1}^{n} X_t^i$ is a canonical emergent feature: it cannot be predicted from any individual bit, only from all bits jointly. In the chain, V_t evolves as a slow two-state Markov chain, while the micro bits $\mathbf{X}_t$ are resampled uniformly from the parity-consistent set at each step. A standard SRL agent observing individual bits therefore receives a noisy, uninformative signal. An agent that recovers V_t can match its action $A_t \in \{0, 1\}$ to the current parity and solve the task.

We pretrain all encoders offline on data collected by a random policy, then freeze them for downstream PPO training. We compare three variants: Micro, which feeds the raw particle observations directly to the policy; Autoencoder, which replaces the particles with an autoencoder-compressed feature, and CESRL, which uses our causal-emergent encoder. We fix $\beta = 1$ as the parity variable is exact. Full details are in Appendix A and Appendix C.

Results Figure 3 (left) illustrates that reward-free pre-training with Ψ rapidly approaches the 1-bit ceiling of the latent parity process, showing that the encoder recovers the collective state almost exactly. Figure 3 (right) shows that this results in PPO reaching the maximum return for downstream evaluation. As expected, the micro and autoencoder baselines fail to recover the parity and do not solve the task.

5 Do Realistic Benchmarks Contain Useful Causal Emergence?

In this experiment, we consider a more realistic scenario in which the causal-emergent feature is not directly known and coexists with the original system state. We introduce Vortex Pair Wind CartPole, an extension of the classical CartPole benchmark. A pair of hidden wind vortices generates a lateral force on the cart; the agent observes the standard CartPole state combined with a field of particles orbiting the two vortex centres. The emergent feature - the net wind imbalance - changes slowly, while each particle's trajectory is individually noisy.

The encoder is trained only on the particle observations, and its output is concatenated with the CartPole state for downstream control. Unlike XOR, the emergent feature here is only approximately collective: individual particles carry partial information about their vortex, so the strict penalty ($\beta = 1$) may over-suppress useful cues. We therefore sweep β to find the right level of collective bias without discarding individually informative signal. We additionally evaluate generalisation by

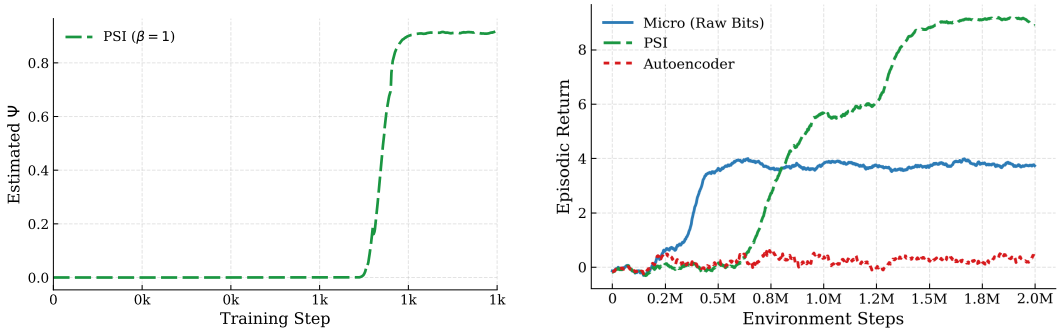


Figure 3: **XOR Parity Chain results. Left:** reward-free representation pre-training with the Ψ objective ($\beta = 1$). The objective rapidly approaches the 1-bit ceiling of the latent parity process. **Right:** downstream PPO learning curves. Ψ solves the task near the oracle ceiling, while *Micro* plateaus far below optimum and *Autoencoder* fails to recover the parity signal.

testing the frozen encoder under higher particle noise and a longer horizon, without retraining (full details in Appendix B and Appendix C).

Results From 4, we observe that the learned representation is sensitive to the choice of β. As β increases, the measured causal emergence under the Ψ objective decreases, suggesting that in this benchmark the learned representation relies less on strongly causal-emergent structure as the penalty on micro-level dependence is increased. The best-performing setting yields somewhat more sample-efficient downstream control than both Micro and Autoencoder, with the smallest tested value, $\beta = 0.1$, performing best in our sweep. However, the gains are modest, and both baselines are still able to solve the task. This indicates that the benchmark contains useful causal-emergent structure, but that successful control does not depend on it as cleanly as in XOR. A full downstream sweep over β is shown in Appendix B.1.

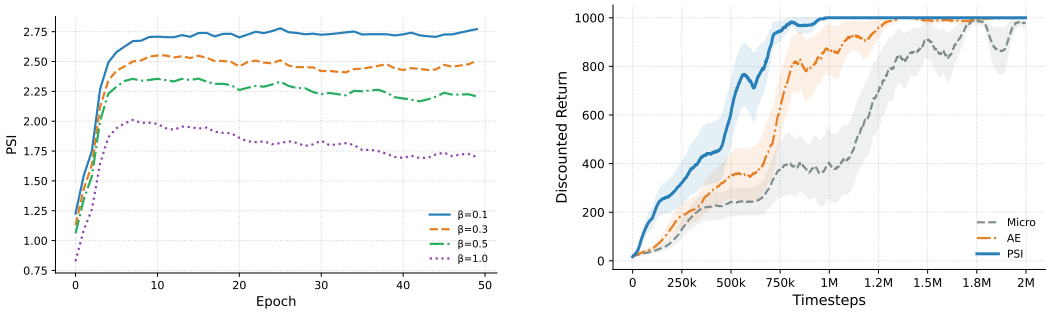


Figure 4: **Vortex Pair Wind CartPole results. Left:** representation-side sweep over β for the Ψ objective, showing that the measured causal emergence decreases as β increases. **Right:** best downstream PPO result, where Ψ reaches high return earlier than both the Micro and Autoencoder.

6 Related Work

State representation learning for RL Common SRL methods for RL include reconstruction-based objectives, contrastive objectives such as CURL (Laskin et al., 2022), predictive world models such as Dreamer (Hafner et al., 2020), and information-bottleneck approaches such as RepDIB (Islam et al., 2022). These methods typically aim to compress observations into features that preserve task-relevant information for control. Our setting differs in that the representation must recover a collective state that is only visible at the level of the full system rather than in any individual component alone.

Mutual-information objectives for representation learning A prominent approach to reward-free representation learning maximises mutual information between current and future observations, originally proposed for sequential data by Oord et al. (2019) and adapted to RL by Stooke et al. (2020); Laskin et al. (2020). (Rakelly et al., 2019) show that among MI-based objectives, forward predictive criteria are theoretically sufficient for control. In our setting, however, a pure predictive objective may collapse to solutions tracking any individual component rather than collective structure. Our Ψ objective is a constrained variant: it maximises predictive MI while penalising information attributable to individual micro-components.

Causal Emergent Representations (McSharry et al., 2024) introduce Ψ-maximisation for unsupervised causal-emergent representation learning from time-series data. This approach has been applied to improve robustness and generalisation (Yadav, 2025). Our work aims to understand if these representations are useful for training a downstream task in RL.

7 Conclusion

In this work we introduced the CESRL problem and showed that a self-supervised information-theoretic objective trained on reward-free data recovers the emergent collective representation on a controlled discrete benchmark where standard methods cannot, and improves sample efficiency on a realistic physics benchmark.

As this is ongoing work, we see the following future research directions. The first is a stronger evaluation. Our current benchmarks are intentionally controlled and leave residual observation-level shortcuts, so they do not yet test causal emergence as cleanly as we would like. Future work should move to more complex physical domains where collective structure is genuinely central to control, such as fluid dynamics (Xian et al., 2023; Lagemann et al., 2025), chemistry (Govindarajan et al., 2025), and bird flocking navigation (Durve et al., 2020; Perrin et al., 2021).

The second direction is robustness and generalisation. Our results suggest that representations biased toward collective predictive structure may transfer more reliably under perturbation, and this may matter beyond strictly causally emergent problems. A natural next step is to evaluate this idea on contextual and adaptive RL benchmarks in the spirit of CARL (Benjamins et al., 2021), where hidden environment factors vary across tasks. This direction is also motivated by prior work showing that predictive mutual-information objectives are the ones most closely aligned with control, while richer state-based objectives can preserve irrelevant detail (Rakelly et al., 2021). We therefore expect CESRL-style constraints to be especially promising when combined with recent self-supervised latent objectives for history-dependent control problems (Ni et al., 2024).

The third direction is unsupervised exploration. The current work assumes a pre-collected offline dataset with sufficient coverage. Unsupervised RL methods (Laskin et al., 2021) address this via intrinsic signals that gather informative transitions without reward (Pathak et al., 2017; Eysenbach et al., 2018). The Ψ objective connects naturally to empowerment-based exploration (Klyubin et al., 2005), and formalising this link is a promising direction for future work.

The fourth direction is to move beyond a single dominant macro variable. This paper studies the simplest case in which one coarse-graining is sufficient for control, but realistic systems may contain several interacting causal-emergent factors at different scales (Rosas et al., 2020). Extending CESRL to identify, factorise, and disentangle multiple emergent variables, while separating control-relevant collective structure from nuisance collective dynamics, is an important step toward realistic many-body control.

References

Carolin Benjamins, Theresa Eimer, Frederik Schubert, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. CARL: A Benchmark for Contextual and Adaptive Reinforcement

Learning, October 2021. URL http://arxiv.org/abs/2110.02102. arXiv:2110.02102 [cs].

Nicolò Botteghi, Mannes Poel, and Christoph Brune. Unsupervised Representation Learning in Deep Reinforcement Learning: A Review, May 2024. URL http://arxiv.org/abs/2208.14226. arXiv:2208.14226 [cs].

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1665–1674. PMLR, May 2019. URL https://proceedings.mlr.press/v97/du19b.html.

Mihir Durve, Fernando Peruani, and Antonio Celani. Learning to flock through reinforcement. *Physical Review E*, 102(1):012601, July 2020. ISSN 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.102.012601. URL http://arxiv.org/abs/1911.01697. arXiv:1911.01697 [physics.soc-ph].

Ayoub Echchahed and Pablo Samuel Castro. A Survey of State Representation Learning for Deep Reinforcement Learning, 2025. URL https://arxiv.org/abs/2506.17518. Version Number: 1.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is All You Need: Learning Skills without a Reward Function, October 2018. URL http://arxiv.org/abs/1802.06070. arXiv:1802.06070 [cs].

Prashant Govindarajan, Mathieu Reymond, Antoine Clavaud, Mariano Phielipp, Santiago Miret, and Sarath Chandar. CrystalGym: A New Benchmark for Materials Discovery Using Reinforcement Learning, September 2025. URL http://arxiv.org/abs/2509.23156. arXiv:2509.23156 [cs].

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination, March 2020. URL http://arxiv.org/abs/1912.01603. arXiv:1912.01603 [cs].

Riashat Islam, Hongyu Zang, Manan Tomar, Aniket Didolkar, Md Mofijul Islam, Samin Yeasar Arnob, Tariq Iqbal, Xin Li, Anirudh Goyal, and Nicolas Heess. Representation Learning in Deep RL via Discrete Information Bottleneck. *arXiv preprint arXiv:2212.13835*, 2022.

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135, Edinburgh, Scotland, UK, 2005. IEEE. ISBN 978-0-7803-9363-9. DOI: 10.1109/CEC.2005.1554676. URL http://ieeexplore.ieee.org/document/1554676/.

Christian Lagemann, Sajeda Mokbel, Miro Gondrum, Mario Rüttgers, Jared Callaham, Ludger Paehler, Samuel Ahnert, Nicholas Zolman, Kai Lagemann, Nikolaus Adams, Matthias Meinke, Wolfgang Schröder, Jean-Christophe Loiseau, Esther Lagemann, and Steven L. Brunton. HydroGym: A Reinforcement Learning Platform for Fluid Dynamics, December 2025. URL http://arxiv.org/abs/2512.17534. arXiv:2512.17534 [physics].

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: Unsupervised Reinforcement Learning Benchmark, October 2021. URL http://arxiv.org/abs/2110.15191. arXiv:2110.15191 [cs].

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery, March 2022. URL http://arxiv.org/abs/2202.00161. arXiv:2202.00161 [cs].

Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. SoftGym: Benchmarking Deep Reinforcement Learning for Deformable Object Manipulation. In *Proceedings of the 2020 Conference on Robot Learning*, pp. 432–448. PMLR, October 2021. URL https://proceedings.mlr.press/v155/lin21a.html.

David McSharry, Christos Kaplanis, Fernando E Rosas, and Pedro A M Mediano. Learning diverse causally emergent representations from time series data. 2024.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging State and History Representations: Understanding Self-Predictive RL, April 2024. URL http://arxiv.org/abs/2401.08898. arXiv:2401.08898 [cs].

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2778–2787. PMLR, July 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. Mean Field Games Flock! The Reinforcement Learning Way, May 2021. URL http://arxiv.org/abs/2105.07933. arXiv:2105.07933 [cs.MA].

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5331–5340. PMLR, May 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which Mutual-Information Representation Learning Objectives are Sufficient for Control?, June 2021. URL http://arxiv.org/abs/2106.07278. arXiv:2106.07278 [cs].

Fernando E. Rosas, Pedro A. M. Mediano, Henrik J. Jensen, Anil K. Seth, Adam B. Barrett, Robin L. Carhart-Harris, and Daniel Bor. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, 16(12):e1008289, December 2020. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1008289. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008289.

Jiaming Song and Stefano Ermon. Understanding the Limitations of Variational Mutual Information Estimators, March 2020. URL http://arxiv.org/abs/1910.06222. arXiv:1910.06222 [cs, math, stat].

Bhuvanesh Sridharan, Animesh Sinha, Jai Bardhan, Rohit Modee, Masahiro Ehara, and U. Deva Priyakumar. Deep reinforcement learning in chemistry: A review. *Journal of Computational Chemistry*, 45(22):1886–1898, 2024. ISSN 1096-987X. DOI: 10.1002/jcc.27354. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.27354. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.27354.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling Representation Learning from Reinforcement Learning, September 2020. URL https://arxiv.org/abs/2009.08319v3.

Zhou Xian, Bo Zhu, Zhenjia Xu, Hsiao-Yu Tung, Antonio Torralba, Katerina Fragkiadaki, and Chuang Gan. FluidLab: A Differentiable Environment for Benchmarking Complex Fluid Manipulation, March 2023. URL http://arxiv.org/abs/2303.02346. arXiv:2303.02346 [cs.RO].

Shashank Yadav. Causal Emergent Representation Learning Under Distribution Shift in Critical Care Time Series. 2025.

Supplementary Materials

The following content was not necessarily subject to peer review.

A XOR Causal Parity Lock: Full Details

A.1 Environment Specification

Formal dynamics The micro-state $\mathbf{X}_t = (X_t^1, \ldots, X_t^N) \in \{0,1\}^N$ evolves in two coupled stages at each step. First, the macro parity transitions:

$$V_{t+1} = \begin{cases} V_t & \text{w.p. } \gamma, \\ 1 - V_t & \text{w.p. } 1-\gamma. \end{cases} \tag{2}$$

Second, the new micro-state is drawn uniformly from the consistent parity class:

$$\mathbf{X}_{t+1} \sim \mathcal{U}(V_{t+1}) = \mathrm{Uniform}\big(\{\mathbf{x} \in \{0,1\}^N : \textstyle\bigoplus_j x_j = V_{t+1}\}\big). \tag{3}$$

The observation supplied to the agent is $o_t = (t/L,\, X_t^1, \ldots, X_t^N) \in \mathbb{R}^{N+1}$, where $t/L \in [0,1]$ is a normalised step counter. The reward at each step is $r_t = +1$ if $A_t = V_t$, else -1. Episodes terminate after L steps.

Table 1: Default configuration for the XOR Causal Parity Lock.

Parameter	Symbol	Default
Number of parity bits	N	5
Macro persistence	γ	0.99
Episode length	L	10
Offline dataset size	-	10^6 steps
Encoder feature size	d	1
PSI pre-training epochs	-	100
AE pre-training epochs	-	10
PPO training steps	-	2×10^6
PPO learning rate	-	2.5×10^{-4}
PPO parallel environments	-	16
PPO steps per rollout	-	128
Random seeds	-	1 (encoder), 5 (policy)

Experimental parameters

Offline dataset The 10^6 transition pairs used for encoder pre-training are collected under a fixed no-op policy ($A_t = 0$ at every step) over a single uninterrupted episode. Because state transitions in this environment are action-independent (actions affect only the reward signal), any fixed policy yields the same transition distribution. We use $A_t = 0$ as the simplest choice.

B Vortex Pair Wind CartPole

The Vortex Pair Wind CartPole environment wraps ‘CartPole-v1‘ with a hidden fluid process that generates a lateral wind on the cart. The full environment observation contains the standard 4-dimensional CartPole state together with $N = 32$ tracer-particle positions, giving a 68-dimensional observation when the CartPole state is included, which is the setting used in our experiments. However, in the CESRL pipeline the encoder is trained only on the 64 particle coordinates; the 4 CartPole state variables are passed directly to the downstream PPO policy and concatenated with the learned latent.

Hidden vortex dynamics The hidden process contains two vortex circulations, $\Gamma_t^{(1)}$ and $\Gamma_t^{(2)}$, each evolving as an AR(1) process:

$$\Gamma_{t+1}^{(k)} = \rho_v \, \Gamma_t^{(k)} + \mathcal{N}(0, \sigma_v^2), \quad k \in \{1, 2\}.$$

The signed circulation imbalance $\Delta_t = \Gamma_t^{(1)} - \Gamma_t^{(2)}$ drives a lateral wind $F_{\text{wind}} = \alpha \Delta_t$ on the cart, while the rest of the CartPole dynamics and reward remain those of standard CartPole; the fluid process adds no task-specific shaping reward. The implementation defaults are $\rho_v = 0.95$, $\sigma_v = 0.04$, initialisation std 0.3, and vortex separation $d = 2.0$. The benchmark experiments use the more persistent regime $\rho_v = 0.99$.

Particle micro-observations Each particle is assigned to one of the two vortices and orbits a fixed centre at horizontal offset $\pm d/2$, with particles alternating between the left and right vortex. If particle i is assigned to vortex $\nu(i)$, then its angular speed is

$$\omega_t^{(i)} = \Gamma_t^{(\nu(i))} / \max(r^{(i)}, 0.1),$$

and its angle evolves as

$$\theta_{t+1}^{(i)} = \theta_t^{(i)} + 0.05 \, \omega_t^{(i)} + \mathcal{N}(0, \sigma_p^2).$$

The observed particle position is then $\mathbf{p}_t^{(i)} = (c^{(i)} + r^{(i)} \cos \theta_t^{(i)}, \; r^{(i)} \sin \theta_t^{(i)})$, with radii sampled uniformly in $[0.2, 0.8]$ at reset. Individual particle trajectories are ambiguous, while the collective signed mean x-position tracks the hidden circulation imbalance. In the implementation, this coarse-grained macro is the particle-field mean in the x-direction.

Table 2: Key parameters of the Vortex Pair Wind environment.

Parameter	Symbol	Default
Particles	N	32
Vortex separation	d	2.0 m
AR(1) persistence	ρ_v	0.95 default; 0.99 in experiments
Circulation noise std	σ_v	0.04
Initial circulation std		0.3
Coupling scale	α	2.0 (reliability), 5.0 (peak)
Particle angular noise	σ_p	0.1 (reliability), 0.3 (peak)
Orbital radius range	$r^{(i)}$	$\mathcal{U}(0.2, 0.8)$ m
Include CartPole state		true
Encoder input		particle field only (64 dims)
Oracle observation		false in main experiments

B.1 Downstream β Sweep

C Implementation Details

C.1 Ψ Training Hyperparameters

We follow the training procedure of McSharry et al. (2024) directly, including the SMILE mutual information estimator (Song & Ermon, 2020), the three-network architecture (representation network f_θ, macro critic g_φ, per-component critics $\{h_{\xi^i}\}$), the decoupled update scheme, and the gradient reversal layer for the micro critics. Table 3 lists our hyperparameters; the only experiment-specific setting is β (fixed at 1 for XOR, swept over $\{0.1, 0.5, 1.0\}$ for Vortex Pair Wind CartPole).

Autoencoder baseline Standard MLP autoencoder with symmetric encoder/decoder, trained with MSE loss for 10 epochs on the same offline dataset. Feature size is 1 (same as PSI encoder) to ensure a fair comparison.

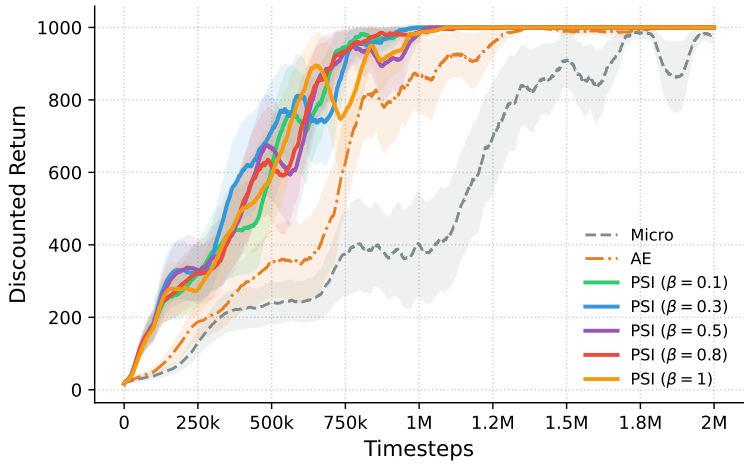


Figure 5: **Downstream performance across β in Vortex Pair Wind CartPole.** Comparison of downstream PPO learning curves for different settings of the reducibility weight in the Ψ objective.

Table 3: Ψ training hyperparameters (shared across experiments).

Hyperparameter	**Value**
Encoder hidden sizes	$[256, 256]$
Critic hidden sizes	$[512, 512, 512]$
Encoder learning rate	10^{-4}
Critic learning rate	10^{-3}
SMILE clip value	5.0
Encoder update frequency	every 5 steps
Batch size	2000
Training epochs	100