Proceedings of Machine Learning Research 323:1–47, 2026 5th Conference on Causal Learning and Reasoning

Disentangling Dynamical Systems: Causal Representation Learning Meets Local Sparse Attention

Markus W. Baumgartner MARKUSB@ROBOTS.OX.AC.UK
Anson Lei ANSON@ROBOTS.OX.AC.UK
Joe Watson JOEWATSON@ROBOTS.OX.AC.UK
Ingmar Posner INGMAR@ROBOTS.OX.AC.UK
Applied Artificial Intelligence Lab, Oxford Robotics Institute, Oxford, UK

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Parametric system identification methods estimate the parameters of explicitly defined physical systems from data. Yet, they remain constrained by the need to provide an explicit function space, typically through a predefined library of candidate functions chosen via available domain knowledge. In contrast, deep learning can demonstrably model systems of broad complexity with high fidelity, but black-box function approximation typically fails to yield explicit descriptive or disentangled representations revealing the structure of a system. We develop a novel identifiability theorem, leveraging causal representation learning, to uncover disentangled representations of system parameters without structural assumptions. We derive a graphical criterion specifying when system parameters can be uniquely disentangled from raw trajectory data, up to permutation and diffeomorphism. Crucially, our analysis demonstrates that global causal structures provide a lower bound on the disentanglement guarantees achievable when considering local state-dependent causal structures. We instantiate system parameter identification as a variational inference problem, leveraging a sparsity-regularised transformer to uncover state-dependent causal structures. We empirically validate our approach across four synthetic domains, demonstrating its ability to recover highly disentangled representations that baselines fail to recover. Corroborating our theoretical analysis, our results confirm that enforcing local causal structure is often necessary for full identifiability.

Keywords: Causal representation learning, dynamical systems identification, sparse attention

1. Introduction

Parametric system identification methods aim to estimate the *parameters* that govern a complex system by observing the system's behaviour. While these methods have proven successful in finding interpretable and physically meaningful parameters that describe the underlying dynamical systems, classical methods often require substantial domain knowledge, such as a predetermined function space (Brunton et al., 2016; Gao et al., 2025). In stark contrast, deep learning methods adopt a domain-agnostic and fully data-driven approach to modelling dynamics. By bypassing explicit parameter estimation and directly predicting system trajectories, learning-based methods excel in highly complex settings where providing domain knowledge is prohibitively difficult, such as climate systems (Alet et al., 2025; Lam et al., 2023), biological systems (Abramson et al., 2024), and interactive environments like video games and autonomous driving (Hafner et al., 2025; Hu et al., 2023). However, these methods represent the dynamics of the underlying system *implicitly* within model weights, and therefore lack the insight and mechanistic understanding afforded by system identification techniques. Motivated by the complementary strengths of these paradigms, this work

© 2026 M.W. Baumgartner, A. Lei, J. Watson & I. Posner.

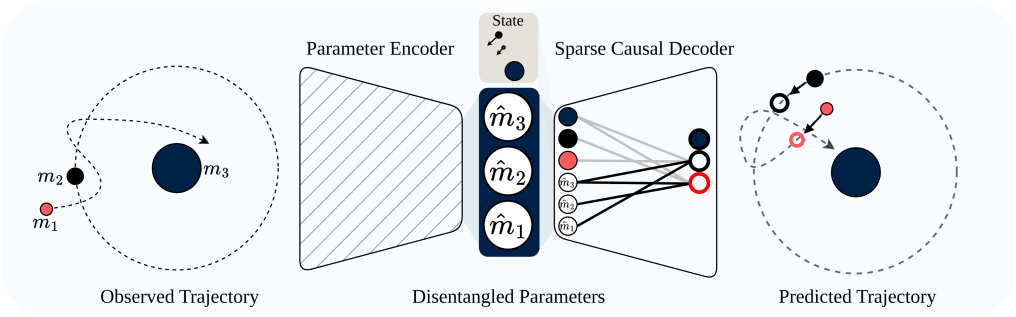


Figure 1: High-level overview of the developed theory. An observed trajectory (left) is encoded into a vector of latent system parameters (marked in dark blue). The developed theory shows that enforcing sparse causal relations between parameters and system components in the decoder (which performs one-step prediction) provably disentangles the system parameter representation.

investigates, both theoretically and empirically, *how* and *when* the underlying parameters of a physical system can be learned from observations without requiring prior functional knowledge.

The key intuition motivating our investigation is that physically meaningful parameters typically exert *sparse* causal influence on system states: a given parameter does not influence all system components at all times. Consider a system comprised of colliding objects. System parameters such as the masses and elasticity coefficients influence only the specific objects involved in a collision, and only at the moment of impact. We argue that this sparsity of influence serves as a practical model-selection criterion for discovering meaningful parameterisations of dynamical systems. We formalise this intuition within the framework of *parameter estimation as representation learning* (Yao et al., 2024a), where identifying physically meaningful parametrisations from system trajectories is analogous to learning disentangled representations from observations. This allows us to leverage causal representation learning theory to characterise the conditions under which parameters can be disentangled. Specifically, we extend prior results on mechanism sparsity (Lachapelle et al., 2024) to the context of dynamical systems and derive novel identifiability results that highlight the role of sparsity regularisation and elucidate the structural properties necessary for disentanglement.

To operationalise and empirically verify our theory, we propose a practical algorithm for learning parameters from observed trajectories. We instantiate our learning method as a sparsity regularised, VAE-style model that encodes observed trajectories into latent parameters and then decodes them, together with the initial state of the system, to reconstruct future trajectories (figure 1). A key aspect of our theory is that parameter identifiability can be significantly improved if the decoder model captures *local*, *state-dependent* causal structures. To reflect such structures, we employ a recently proposed sparsity-regularised transformer architecture as the decoder, designed to learn local causal dependencies. Empirically, we validate our approach across four synthetic domains and demonstrate that the proposed method successfully isolates the underlying system parameters where baseline representation methods fail. In sum, our core contributions are threefold:

- A novel theorem establishing identifiability of system parameters for local causal predictors, culminating in a graphical criterion.
- A practical implementation using a recent sparse transformer architecture for unsupervised local causal discovery and representation learning.
- An empirical validation across multiple domains, including object-centric and discontinuous systems, which corroborates the theory's predictions.

2. Identifiability of System Parameters

In this section, we present the main identifiability theorem of our work and highlight its practical implications. Our general approach extends prior theoretical results (Lachapelle et al., 2024), which uses mechanism sparsity to induce disentanglement, to the setting of learning parameterisations of dynamical systems from observed trajectories. In particular, we consider system parameters that change between the trajectories within a dataset. For example, consider a dataset containing planetary orbits. If each trajectory consists of a single instantiation of the system given an initial state and a particular set of parameters (e.g. planetary and solar masses), then the mass of each planet or sun strongly influences the evolution of the others. The theory we are establishing in this section defines the conditions on the underlying system and the learning method under which mechanism sparsity leads to disentangled system parameters.

2.1. Problem Setting

We consider nonlinear deterministic Markovian dynamical systems, which include the orbital example presented above, as well as the majority of natural systems that are not externally influenced.

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_t), \quad \boldsymbol{\theta} \sim \boldsymbol{p}_{\boldsymbol{\theta}}(\cdot), \quad \boldsymbol{\theta} \in \Theta, \quad \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{x}_0 \in \mathcal{X}_0 \subseteq \mathcal{X}, \tag{1}$$

$$\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta}) = [\boldsymbol{x}_0, \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_0), (\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{f}_{\boldsymbol{\theta}})(\boldsymbol{x}_0), ..., \boldsymbol{f}_{\boldsymbol{\theta}}^T(\boldsymbol{x}_0)], \quad \boldsymbol{x}_0, \boldsymbol{\theta} = \boldsymbol{h}^{-1}(\boldsymbol{\tau}), \tag{2}$$

where $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the state of a system at time t, $\boldsymbol{f}_{\boldsymbol{\theta}}$ denotes the Markovian dynamics model, characterised by a set of parameters $\boldsymbol{\theta} \in \Theta$, where Θ has non-zero measure. Trajectories are initialised from an initial distribution $\boldsymbol{x}_0 \in \mathcal{X}_0$. The set of systems $\boldsymbol{f}_{\boldsymbol{\theta}} \in \mathcal{F}$ defines an environment. For instance, $\mathcal{F}$ might represent the family of orbital mechanics. Then $\boldsymbol{f}_{\boldsymbol{\theta}}$ might be the instantiation of a system where a planet with mass θ_0 is orbiting a sun with mass θ_1. We assume that the parameters sparsely affect the state distribution via a causal graph $\mathcal{G}$ (assumption 3). For example, the planet's acceleration is affected by the mass of the sun θ_1 but not its own mass, and vice versa. Each trajectory is defined via a bijective trajectory decoder, $\boldsymbol{h}(\cdot)$, which, given a parameter and starting state, auto-regressively applies the Markovian dynamics model. Likewise, a parameter encoder, $\boldsymbol{h}^{-1}(\cdot)$, is defined as the inverse process (equation 21). This encoder-decoder setup is visualised in Figure 1 and forms a latent information bottleneck that forces learnt approximations to isolate the dynamical factors of variation. Let $\mathcal{S} = (\Theta, \mathcal{X}_0, \mathcal{X}, \mathcal{F}, \mathcal{G})$ denote the ground truth model of a dynamical family, and $\hat{\mathcal{S}} = (\hat{\Theta}, \mathcal{X}_0, \mathcal{X}, \hat{\mathcal{F}}, \hat{\mathcal{G}})$ denote the learnt latent parameter model. For uniqueness and without loss of generality, we enforce assumptions 1 and 2, which in combination also require the bijectivity of $\boldsymbol{h}(\cdot)$ and $\boldsymbol{h}^{-1}(\cdot)$ with respect to their image.

Assumption 1 (Existence and Parameter Influence (Yao et al., 2024a)) *For every $\boldsymbol{x}_0 \in \mathcal{X}_0, \boldsymbol{\theta} \in \Theta$ there exists a unique trajectory, $\boldsymbol{\tau}$, over horizon T, satisfying $\boldsymbol{x}_{t+1} = \boldsymbol{f}_\theta(\boldsymbol{x}_t)$. Equivalent to requiring $\boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta})$ to be injective.*

Assumption 2 (Observational Equivalence) *Assume that the two systems, the ground-truth $\mathcal{S}$ and learnt approximation $\hat{\mathcal{S}}$, are* observationally equivalent*, which posits that the modelled systems are equivalent up to invertible parameter representations. Observational equivalence requires, $\forall \boldsymbol{x}_0, \boldsymbol{\theta}: \quad \exists \hat{\boldsymbol{\theta}}$ s.t. $\boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta}) = \hat{\boldsymbol{h}}(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}})$ (Equivalently, $\hat{\boldsymbol{h}}(\cdot)$ is surjective).*

Assumption 3 (The Transition Model is Markov to a DAG (Lachapelle et al., 2024)) *The transition models of $\mathcal{S}$, and $\hat{\mathcal{S}}$, are causally and* faithfully *(Pearl, 2000; Spirtes et al., 2001) related to*

directed acyclic graphs (DAG), $\mathcal{G}$, and $\hat{\mathcal{G}}$, which fully define the dependencies between the ***output elements and latent parameters****. The edges of the graph are defined through the non-zero entries in the Jacobian of the transition model, requiring that the model and its first derivative are continuous. Simply, the following causal equivalence is defined, where $\boldsymbol{\theta}_{Pa_i^{\mathcal{G}}}$ denotes the subset of parameters that are parents of output i in graph $\mathcal{G}$: $f_i(\boldsymbol{x}_t, \boldsymbol{\theta}) \equiv f_i(\boldsymbol{x}_t, \boldsymbol{\theta}_{Pa_i^{\mathcal{G}}})$.*

$$(i,j) \notin \mathcal{G} \implies \frac{\partial(\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}))_i}{\partial \theta_j} = 0 \quad \forall \boldsymbol{x}_t, \theta. \tag{3}$$

Assumptions 1, 2 enforce separability and uniqueness of dynamical influences, as well as ensuring that the learnt approximation sufficiently captures the system dynamics. Assumption 3 formalises the intuition that parameters affect the system components sparsely by defining parameter influence as an inherent property of a system. All assumptions are commonly made in the non-linear ICA and causal representation learning literature (Schölkopf et al., 2012; Yao et al., 2024b). The fundamental investigation of this work is then to specify the conditions under which the representations of system parameters learnt by the parameter encoder, $\boldsymbol{h}(\cdot)$, align with the ground-truth system parameters, or equivalently, when the representation is identified (Hyvärinen et al., 2023; Carbonneau et al., 2022).

Relative to a ground-truth representation, a learnt representation is identified (definition 1)[1] if it is equivalent to the ground truth. Exact identifiability conditions are, in general, unobtainable in unsupervised settings (Locatello et al., 2019), due to a lack of defined scaling and bias of functional mappings (Lachapelle et al., 2022a). Consequently, we consider a looser form of identifiability, identifiability up to an equivalence operator (definition 2). Disentanglement is defined in this setting as identifiability up to permutation and element-wise diffeomorphism.

2.2. Graphical Criterion

We present our main identifiability result, which shows how mechanism sparsity, when applied to the context of learning parameters of dynamical systems, can lead to identifiable parameter representations up to permutation and element-wise diffeomorphism.

Theorem 1 (Disentanglement of Latent System Parameters) *Let $\mathcal{S}$, and $\hat{\mathcal{S}}$ correspond to two systems satisfying assumptions 1, 2, and 3. Further assume that,*

1. *The system, $\mathcal{S}$, is path connected (assumption 4).*
2. *$\hat{\mathcal{S}}$ satisfies $\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$ (assumption 5).*
3. *The Jacobian of the transition model, $\nabla_{\boldsymbol{\theta}} \boldsymbol{f}$,* varies sufficiently *(assumption 6).*

Then the dynamical representations of $\mathcal{S}$, and $\hat{\mathcal{S}}$ are equivalent up to permutation and element-wise diffeomorphism if $\mathcal{G}$ satisfies the following graphical criterion:

$$\forall i: \quad \cap_{a \in Ch_{\mathcal{G}}(i)} Pa_{\mathcal{G}}(a) = \{i\}, \tag{4}$$

where $Ch_{\mathcal{G}}(j)$, and $Pa_{\mathcal{G}}(j)$ denotes the set of children and parent nodes of a node j in a graph G.

1. All definitions can be found in appendix A.

Here, assumption 6, sometimes referred to as *assumption of variability* (Hyvarinen et al., 2019), is a standard assumption that requires the Jacobian of the ground truth model to depend sufficiently strongly on individual parameters such that it is possible to differentiate the parameters by observing their influence. Assumption 4 precludes the existence of disjoint partitions: in some systems, the state space can be partitioned into regions with boundaries that are not traversable by any trajectory[2]. In such cases, the theory will only hold locally within each partition, guaranteeing a weaker form of conditional disentanglement. Details of the assumptions and proofs are provided in appendix B.

The key intuition behind the theory and its practical implications can be understood through assumption 5 and the graphical criterion. Assumption 5 requires the learnt model to use a parameter representation with at most as many causal influences as the true system. To illustrate how this drives disentanglement, consider a simple 2D system with two parameters, each affecting exactly one state variable: $\theta_0 \to x_0, \theta_1 \to x_1$. Suppose an arbitrary learnt model uncovers an *entangled* representation. For example, a linear mixing of the ground truth parameters. $\hat{\theta}_0 = \frac{1}{2}(\theta_0 + \theta_1)$, $\hat{\theta}_1 = \frac{1}{2}(\theta_0 - \theta_1)$. In this case, correctly reproducing the system's behaviour would require each state variable to depend on both mixed parameters. The causal graph would therefore need to be *denser* than the true one. This illustrates a general principle: entangled representations typically require at least as many, and often more, causal edges to account for the same physical effects. By constraining how many causal relations the learned model may use, assumption 5 restricts the space of admissible representations, pushing the model toward disentangled parameters.

However, this only works if the underlying system is sufficiently sparse. If, for example, the underlying system graph were fully connected, then every representation would be fully dense and sparsity cannot induce disentanglement. The presented graphical criterion (equation 54) characterises exactly how sparse the underlying system needs to be: a ground-truth parameter can be identified if no other parameter also influences all of its children. In the following, we show that this criterion can be substantially relaxed by considering local causal graphs, such that a parameter is only unidentifiable if there is another parameter that influences all of its children *at all time*.

2.3. Broadened Identifiability Guarantees with State-Dependent Local Causal Graphs

In practice, many dynamical systems exhibit interactions in which parameters exert influence *sparsely* along a trajectory. For example, parameters such as an object's mass and elastic coefficients influence only the specific objects involved in a collision and only at the moment of impact. A global graph must account for every possible interaction that might occur and, as a result, fails to capture the structure of a system that can be exploited to infer disentangled representations. We consider the setting in which local subsets of a global graph are expressible as functions of the state. Imagine there are b distinct local causal graphs, each being active in a non-trivial open partition of the state space $\boldsymbol{x} \in \mathcal{X}_{i \in [1,b]}$. The dynamics model, $\boldsymbol{f}(\cdot)$, can then be assumed as Markov (assumption 3) with respect to all local graphs within their respective partitions.

$$\mathcal{G}_l(\boldsymbol{x}) \subseteq \mathcal{G},\ \forall l \in [1, 2, ..., b]\,. \tag{5}$$

We can treat each partition as an independent representation learning problem. Within each partition, the space of admissible parameter representations is constrained separately by each unique graph. The path-connectedness assumption ensures that the system can traverse between distinct partitions of the state space along a trajectory. Therefore, a parameter representation must be

2. For an intuitive example of such a system, please refer to Appendix C

reusable across partitions and satisfy each representational constraint enforced by sparsity regularisation individually. This leads to the *local graphical criterion* (equation 6, derivation in appendix B.6), which formalises the conditions under which parameter reuse results in disentanglement.

$$\forall i: \quad \bigcap_{l\in[1,b]} \cap_{a\in \mathrm{Ch}_{\mathcal{G}_l}(i)} \mathrm{Pa}_{\mathcal{G}_l}(a) = \{i\}. \tag{6}$$

The local graphical criterion reveals that local sparsity constitutes a strictly stronger source of identifiability than global sparsity. As local graphs contribute to disentanglement through an intersection over graphs, the feasible space of parametrisations can only shrink as additional local graphs are introduced. One consequence is that if a parameter is disentangled in **any** individual sub-graph, then it must be disentangled globally. Identifiability may be attainable even when no individual local graph is sufficient to guarantee the disentanglement of any parameter on its own.

3. Learning Parameters with Sparse Attention

The identifiability guarantees of our theorem are expressed as constraints upon both the system, $\mathcal{S}$ and a model, $\hat{\mathcal{S}}$, where assumptions 1, 4, and 6 apply only to the system. In fact, only two constraints are imposed on the model: observational equivalence to the system and Markovian dynamics with respect to a causal graph with $\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$. In practice, these conditions are unenforceable directly. We recast these constraints as an optimisation problem by relaxing observational equivalence as a dynamical reconstruction loss and enforcing sparsity through regularisation.

3.1. Latent Representation Learning

In accordance with our overall framework of using representation learning techniques for latent parameter estimation, our method uses a Variational Autoencoder setup (Kingma and Welling, 2013). Specifically, in the dynamical systems setting, the goal is to maximise the likelihood of observed trajectories given the initial condition, $p(\boldsymbol{\tau} \mid \boldsymbol{x}_0)$, which requires a *conditional* VAE (Sohn et al., 2015). The overall model consists of two modules, parametrised by $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$: first, the encoder, $p_{\boldsymbol{\phi}}(\hat{\boldsymbol{\theta}}|\boldsymbol{\tau})$, encodes observed trajectories into latent parameters, and second, the decoder, $p_{\boldsymbol{\psi}}(\boldsymbol{\tau}|\hat{\boldsymbol{\theta}}, \boldsymbol{x}_0)$, generates the reconstructed trajectory conditioned on an initial state $\boldsymbol{x}_0$. In practice, conditioning on the initial state is implemented by concatenating the latent parameter with the initial state. In this work, all distributions are parameterised as Gaussian distributions with learnable mean and variance and a unit isotropic Gaussian distribution is used as the prior over the latent parameters. The encoder and the decoder are trained jointly by maximising the Evidence Lower bound, which consists of a reconstruction term, $\mathcal{L}_{\text{rec}}(\boldsymbol{\psi}, \boldsymbol{\phi})$, and a KL divergence term, $\mathcal{L}_{\text{KL}}(\boldsymbol{\phi})$. Following standard practice, we scale strength of the KL term using a hyperparameter (Higgins et al., 2017). In effect, forcing the model to reconstruct trajectories enforces the observational equivalence assumption.

3.2. Optimising for Sparse Local Causal Structure

The theory places a sparsity constraint on the *decoder*, which generates trajectories from parameters. To this end, we leverage the SPARTAN architecture introduced by Lei et al. (2025) to both measure and regularise the causal relationship between input parameters and system components. SPARTAN builds upon the transformer architecture by masking the flow of information between tokens and penalising the connections between them. Starting with the key, query, and value embeddings per token, i, for a layer l out of L layers, $\{\boldsymbol{k}_i, \boldsymbol{q}_i, \boldsymbol{v}_i\}$, SPARTAN samples an adjacency matrix, $\boldsymbol{A}^l$,

treating the sigmoid, $\sigma(\cdot)$, of the attention logit as the parameter of a Bernoulli distribution. The adjacency matrix is used as a mask to restrict information flow through the softmax when computing the next hidden state h_i. The adjacency matrix is a function of the current state and can be interpreted as a layer wise *local causal graph*. When stacking multiple transformer layers, the information flow between different tokens is tracked through the *number of paths*, $\bar{\boldsymbol{A}}_{i,j}$ connecting token j to i.

$$A^l_{i,j} \sim \text{Bern}(\sigma(\boldsymbol{q}_i^T \boldsymbol{k}_j)), \quad \boldsymbol{h}_i = \sum_j \frac{A_{ij}\exp(\frac{\boldsymbol{q}_i^T \boldsymbol{k}_j}{\sqrt{d_k}})\boldsymbol{v}_j}{\sum_i \exp(\frac{\boldsymbol{q}_i^T \boldsymbol{k}_j}{\sqrt{d_k}})}, \quad \bar{\boldsymbol{A}}(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}}) = \prod_{l=1}^{L} \left(\boldsymbol{A}^l + \mathbb{I}\right), \tag{7}$$

$$\mathcal{L}_{\text{path}}(\boldsymbol{\psi}, \boldsymbol{\phi}) = \mathbb{E}_{\hat{\boldsymbol{\theta}} \sim p_{\boldsymbol{\phi}}(\cdot|\boldsymbol{\tau}),\, \boldsymbol{x}_t \in \boldsymbol{\tau} \in D} \left[\|\bar{\boldsymbol{A}}(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}})\|_1.\right] \tag{8}$$

The identity matrix is added to account for residual connections. By inspecting the values of the path matrix, SPARTAN reveals a state-dependent causal graph between tokens that is differentiable and can be regularised using a path loss, $\mathcal{L}_{\text{path}}(\boldsymbol{\psi}, \boldsymbol{\psi})$.

3.3. Training Objective

Sparsity, reconstruction, and latent regularisation are jointly optimised in our proposed implementation of the theory. However, it is the sparsity of causal relations that drives disentanglement in our theory and not the other losses. To ensure the optimisation procedure learns the *sparsest* representation that models the dynamics, training is formulated as a constrained optimisation problem, where sparsity is minimised subject to a constraint on non-causal losses set by the target loss $\mathcal{L}_\star$. Learning the *sparsest* representation ensures that $\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$ without explicitly defining $\|\mathcal{G}\|_0$.

$$\min_{\boldsymbol{\phi}, \boldsymbol{\psi}} \mathcal{L}_{\text{path}}(\boldsymbol{\psi}, \boldsymbol{\phi}) \text{ subject to } \mathcal{L}_{\text{rec}}(\boldsymbol{\psi}, \boldsymbol{\phi}) + \mathcal{L}_{\text{KL}}(\boldsymbol{\phi}) + \mathcal{L}_{\text{logit}}(\boldsymbol{\psi}) \leq \mathcal{L}_\star. \tag{9}$$

Auxiliary losses are introduced in subsequent sections. The constrained optimisation problem is solved by introducing a Lagrange multiplier, λ_{dual}, and solving via a dual min-max optimisation scheme by taking alternating gradient steps on the parameters and λ_{dual} (Rezende and Viola, 2018).

$$\max_{\lambda_{\text{dual}} > 0} \min_{\boldsymbol{\psi}, \boldsymbol{\phi}} \mathcal{L}_{path}(\boldsymbol{\psi}, \boldsymbol{\phi}) + \lambda_{\text{dual}} \Big(\mathcal{L}_{\text{rec}}(\boldsymbol{\psi}, \boldsymbol{\phi}) + \mathcal{L}_{\text{KL}}(\boldsymbol{\phi}) + \mathcal{L}_{\text{logit}}(\boldsymbol{\psi}) - \mathcal{L}_\star\Big). \tag{10}$$

Attention Logit Regularisation. The min-max optimisation scheme results in a two-phase learning process. First, the model learns to reconstruct whilst ignoring the sparsity regularisation loss ($\lambda_{\text{dual}} \to \infty$). At some point during training, the reconstruction threshold is reached, and the sparsity loss becomes dominant ($\lambda_{\text{dual}} \to 0$), which will push the entangled representation toward disentanglement. The representational *switch* poses a complex optimisation problem, particularly as transformers learn increasingly peaked and low-entropy attention patterns during training (Zhai et al., 2023), leading to vanishing gradients via the softmax, thereby hindering representational plasticity during the second phase of training. Vanishing gradients are combatted through a small loss that penalises large attention logits, $\mathcal{L}_{\text{logit}}(\boldsymbol{\psi})$. An ablation exemplifying the necessity of the logit losses is provided in appendix F.4. The functional form of the logit loss is shown below, where $l \in L$, $i \in T$, $j \in T$ represent the layer and two token dimensions respectively, and the query and key vector for each layer and token are functions of $\boldsymbol{x}_t$ and $\hat{\boldsymbol{\theta}}$.

$$\mathcal{L}_{\text{logit}}(\boldsymbol{\psi}) = \mathbb{E}_{\hat{\boldsymbol{\theta}} \sim p_{\boldsymbol{\phi}}(\cdot|\boldsymbol{\tau}),\, \boldsymbol{x}_t \in \boldsymbol{\tau} \in D} \left[\frac{\lambda_{logit}}{LT^2} \sum_{l,i,j=1}^{L,T,T} e^{\boldsymbol{q}_{l,i}^T \boldsymbol{k}_{l,j}} + e^{-\boldsymbol{q}_{l,i}^T \boldsymbol{k}_{l,j}}\right]. \tag{11}$$

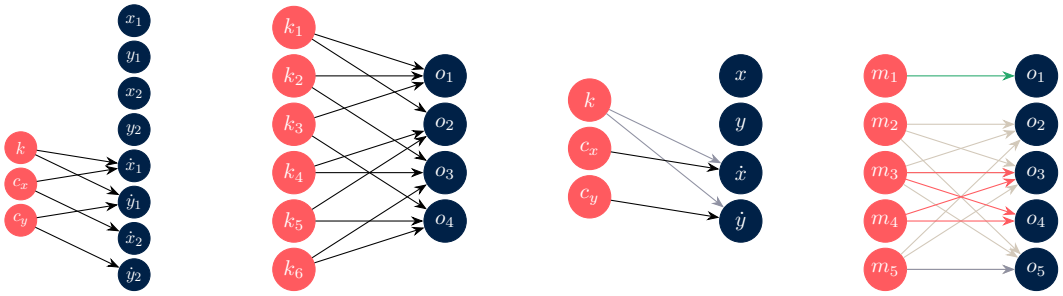


Figure 2: Depiction of the ground-truth DAGs of the evaluation environments. The *left* and *centre-left* correspond to the dual particle and springs environments, respectively, and both satisfy the global graphical criterion for disentanglement. The *centre-right* and *right* graphs correspond to the local particle and bounce environments, respectively. Coloured arrows indicate causal edges that are only active in subsets of the state space, and only a limited number of subsets are shown for the bounce environment. The local particle and bounce environments satisfy only the local, not the global, graphical criterion for disentanglement.

4. Experiments

Building on the local causal graph discovery method outlined in section 3 and the conditions specified in our main theorem, this section empirically examines the feasibility of recovering disentangled representations amidst modelling and approximation errors. Additionally, by testing in controlled domains, we assess the identifiability claims in section 2 and validate the theoretical framework through practical experiments.

Environments. We evaluate our method across four synthetic domains with known ground-truth parameters and causal structure. Disentanglement of a learnt representation up to permutation and element-wise diffeomorphism is measured by computing the *mean-max correlation coefficient* (MCC)[3] metric between the ground-truth and learnt parameters, as is common in related literature (Lachapelle et al., 2022a; Yao et al., 2022; Lachapelle and Lacoste-Julien, 2022).

The *dual particle* and *local particle* environments consist of modified 2D mass-spring-damper systems. The former contains two point masses with independent damping on the vertical and horizontal axes, where one point mass is additionally tethered to the origin via a spring. The latter models a single point mass under independent x-y damping and a spring which becomes *slack* near the origin. The *springs* environment features four point masses coupled by six springs, and the *bounce* environment simulates five variable mass balls that elastically collide, creating dynamic local causal graphs. Figure 2 depicts the corresponding causal structures, with expanded details on each environment provided in appendix E. Together, these environments span systems in which the global graphical criterion is sufficient for identifiability, the local criterion is adequate, and the global criterion is insufficient; the parameters causally influence scalar state values, and the parameters causally influence the system at the object level (modelled using vector representations). The *springs* and *bounce* environments are inspired by similar variations from prior works in causal discovery and representation learning (Löwe et al., 2022; Battaglia et al., 2016; Li et al., 2020).

3. Details on the MCC metric are provided in appendix F.1.

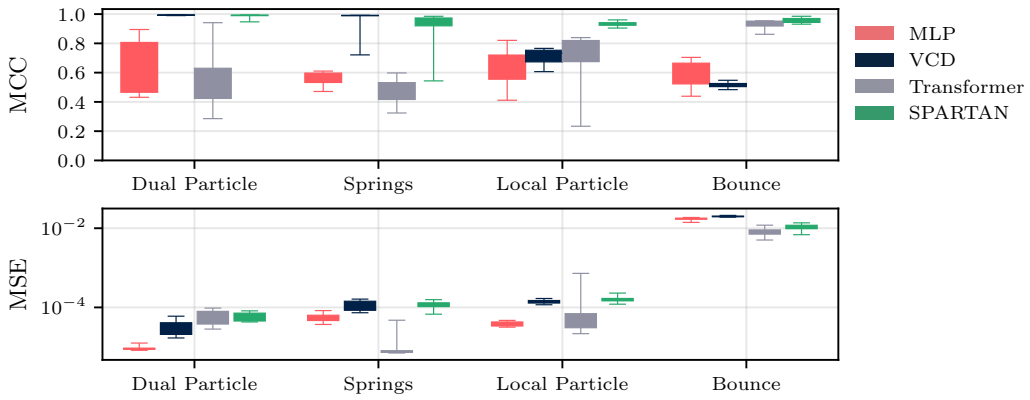


Figure 3: Comparison of disentanglement across the test environments, where an MCC of 1.0 represents perfect disentanglement. All trials are repeated over eight random seeds. Box plots display the minimum, lower quartile, upper quartile, and maximum values. The validation reconstruction loss is shown at the bottom, indicating that all models are approximately equiperformant in these environments. The VCD baseline, which learns static graphs, strongly disentangles in the first two environments, which satisfies the global graph criterion. In contrast, only SPARTAN, which learns state-dependent graphs, consistently disentangles in all environments.

Baselines. We compare against baselines chosen to isolate the role of causal structure and sparsity in parameter identifiability. All models share the same variational autoencoder learning objective and differ only in their inductive bias on dependencies between parameters and state, i.e. the decoder. Since the theory only places sparsity constraints on the decoder, the encoder architecture is held fixed across all baselines. An MLP is used as the encoder in all environments except the *bounce* environment, where a convolutional network is found to be more efficient for longer sequences. For the decoder, we compare against three baselines: 1. *MLP*, which imposes no structural constraints and serves as a control for disentanglement arising from the variational bottleneck alone; 2. *VCD* (Lei et al., 2023), which uses a masked MLP decoder based on a learned global causal graph and satisfies identifiability assumptions only when the global graphical criterion holds; 3. *Transformer*, which uses a transformer that does not explicitly impose sparsity constraints, but nonetheless can provide an implicit bias towards sparse information flow due to the softmax attention. Finally, our method, labelled SPARTAN, uses a sparsity-regularised transformer as the decoder, as described in section 3.

4.1. Results

Figure 3 shows the distribution of MCC scores (higher is better) obtained when training each baseline in each environment across eight random seeds. Reconstruction error is reported to verify that differences in disentanglement are not attributable to failures in modelling the system's dynamics.

In environments where the global causal graphs satisfy the graphical criterion, namely *dual particle* and *springs*, models enforcing global causal sparsity recover disentangled parameter representations. Both the VCD and SPARTAN baselines consistently achieve MCC scores close to one (the upper bound), corroborating our theoretical analysis. In contrast, both the MLP and transformer baselines exhibit a substantially lower and broader distribution of achieved disentanglement. Figure

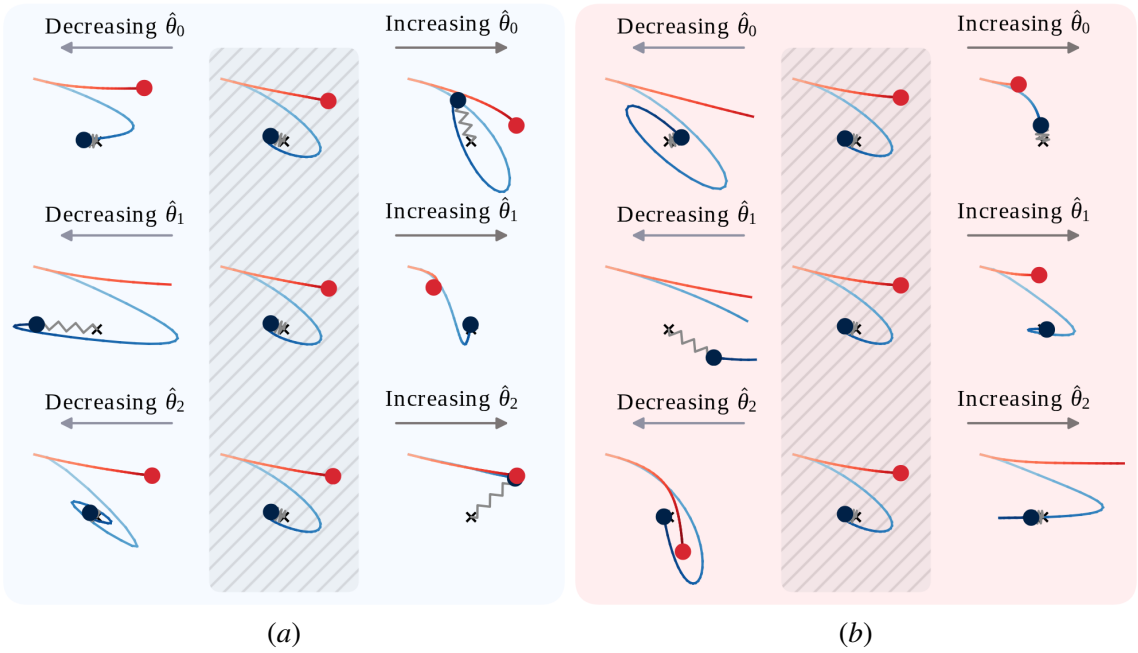


Figure 4: Autoregressive rollouts of the dual particle environment produced from the learnt (a) SPARTAN, and (b) MLP models. Starting from the same initial state, the red particle experiences resistive damping forces independently in the horizontal and vertical axes, whilst the blue particle additionally experiences a spring force to the origin. The ground-truth system parameters are the damping coefficients and the spring constant. (a) shows a clear separation of parameters, increasing $\hat{\theta}_2$ evidently reduces spring strength. As the red particle is unaffected by the spring, it is invariant to changes in $\hat{\theta}_2$. Increasing $\hat{\theta}_1$ *increases* horizontal damping, and increasing $\hat{\theta}_0$ *reduces* vertical axis damping. In contrast, the entangled representations produced by the MLP yield no such insight.

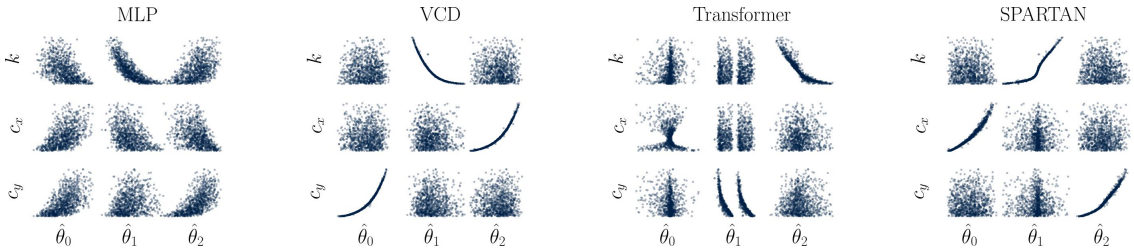


Figure 5: Plots of the entanglement mapping for all models in the dual particle environment. Each subplot shows the marginal distribution of a learnt parameter (x-axis) against the ground-truth (y-axis). Representative trials were chosen. Disentanglement is exemplified by the clear element-wise diffeomorphic structure learnt by both the SPARTAN and VCD architectures, whereas the MLP and Transformer architectures learn entangled representations. Extended results for all environments and representative samples of the causal graphs learnt are provided in appendix F.

5 illustrates samples of the latent representations learnt by each model in the *dual particle* environment. Here, we observe that VCD and SPARTAN both learn latent parameters that can be mapped to the ground-truth parameters in a one-to-one manner, i.e. each ground-truth parameter can be predicted from exactly one learned parameter, qualitatively illustrating the disentanglement induced by sparsity regularisation. Similar examples for other environments can be found in Appendix F.2. To demonstrate the *local* extension of our theory, in the *local particle* and *bounce* environments, the global graphical criterion is not satisfied, and local causal graphs are required for disentanglement. Consequently, the VCD baseline fails to recover disentangled parameters, while SPARTAN consistently learns highly disentangled representations. In the *local particle* environment, VCD is still able to attain stronger disentanglement compared to the other baselines, since the system graph still supports partial disentanglement. In the *bounce* environment, however, since the local causal graphs are significantly sparser (e.g. edges only exist for one timestep during collision), and therefore the VCD baseline is unable to recover any disentanglement. We note that the Transformer baseline also achieves high disentanglement in the *Bounce* environment. We hypothesise that this is due to the implicit bias from softmax attention, which has been shown to be sufficient for inducing local causal graphs in simple situations (Pitis et al., 2020). Nonetheless, SPARTAN is the only model that is able to learn disentangled parameters robustly across all tested environments. To further demonstrate the usefulness of the learned parameters, we perform latent *interventions* on the learned dynamical systems to generate counterfactual rollout trajectories. Figure 4 visualises how intervening on the learnt parameters in the *dual particle* environment results in physically meaningful changes to trajectories, whereas intervening on entangled parameters lead to uninterpretable changes in system behaviours that affect all states.

Overall, the empirical results corroborate the identifiability claims of section 2 and show that in settings where the relevant criteria are satisfied, disentangled representations are recoverable by realising sparse causal relations via appropriate sparsity regularisation.

5. Related Works

The theoretical backbone of our work draws inspiration from prior identifiability results in Causal Representation Learning (CRL) (Schölkopf et al., 2021) and Non-Linear Independent Component Analysis (Hyvärinen and Pajunen, 1999), which explore the conditions under which recovering unobserved variables from data is possible. Works in this area show that disentanglement can be achieved via various inductive biases (e.g., additive decoders (Lachapelle et al., 2023)), assumptions on the data-generating processes (e.g. equivariances (Ahuja et al., 2022), multi-view data (Yao et al., 2024c), and auxiliary information such as interventions (Lippe et al., 2022)). Variations exist within each subgenre; interventions can be known a priori (Ahuja et al., 2023), unknown (von Kügelgen et al., 2023; Varıcı et al., 2024), or differ in their assumptions about intervention targets. Furthermore, when distinguishing between non-temporal and temporal data, temporal CRL leverages distinct assumptions, including non-stationary time series (Hyvarinen and Morioka, 2016; Song et al., 2023), interventions on the temporal transition function (Lippe et al., 2022), or sparse relations between ground-truth latent factors (Li et al., 2025c).

Our approach builds specifically on the mechanism sparsity principle (Lachapelle and Lacoste-Julien, 2022) to achieve identifiability. The key distinction between our approach and existing CRL efforts is that, whereas prior works explore *state* representation for observations such as images (Buchholz et al., 2023; Lachapelle et al., 2024), our setting concerns learning dynamical *parameters*.

In this sense, our work can be viewed as an instantiation of the broader framework of *parameter learning as representation learning*, as advocated in recent literature (Yao et al., 2024a).

Central to our theory and proposed method is the notion of *local causal graphs*, which capture the causal dependencies between variables at a fine-grained temporal resolution. This nascent idea has recently been explored in the context of world models (Zhao et al., 2025; Lei et al., 2025) and reinforcement learning (Hwang et al., 2024; Seitzer et al., 2021), where state-dependent local structures are shown to improve robustness and adaptation efficiency. To the best of our knowledge, our work is the first to theoretically demonstrate how local causal graphs can extend identifiability guarantees for parameter learning and, more broadly, representation learning.

On a conceptual level, our work shares the motivation of System Identification methods, which aim to estimate the underlying parameters of a system given observed trajectories. Here, classical methods require significant prior knowledge of the system, such as known functional forms in PINN-based ODE learning methods (Raissi et al., 2019), and a pre-specified library of functions in the SINDy family of works (Brunton et al., 2016; Rudy et al., 2017). In contrast, our work relaxes these requirements and relies only on the general assumption of mechanism sparsity. Beyond System Identification, the notion of extracting latent descriptions of systems from observed trajectories also appears in adjacent fields such as meta-RL (Wang et al., 2024; Wen et al., 2024), where latent task embeddings are inferred from past trajectories, and in-context inductive reasoning (Macfarlane and Bonnet, 2025), where the system rules are inferred from context examples as latent program codes. We believe the insights developed in this work are readily transferable to these lines of work.

6. Conclusion

In this work, we analyse the identifiability of dynamical system parameters by extending prior results in causal representation learning. Our main result shows that mechanism sparsity can induce identifiable representation of dynamical parameters and make explicit the conditions on the class of dynamical systems where this is possible. In settings where parameters exert *temporally sparse* causal influence on the system, we extend our theory to show that state-dependent local causal graphs strictly strengthen these identifiability guarantees, enlarging the class of identifiable systems.

Motivated by the theory, we propose a practical algorithm for identifying latent parameters using variational inference and sparse transformers. Experiments across four synthetic domains corroborate the theoretical analysis: mechanism sparsity allows consistent learning of disentangled representations, and in scenarios where globally sparse methods are insufficient, local sparse attention enables consistent disentanglement. Furthermore, we empirically show that the disentangled parameters learnt under these conditions enable physically meaningful and localised interventions on system behaviour, as well as counterfactual trajectory generation.

Limitations and Future work. The present work has several limitations that point toward promising directions for future investigation. It has long been conjectured that disentangled representations play a central role in improving generalisation and robustness (Lachapelle et al., 2022b; Bengio et al., 2013). While our experiments demonstrate that disentanglement is possible, future work should evaluate its generalisation properties in downstream tasks. Moreover, in its current form, our theory requires a priori knowledge of the dimension of the parameter space, which can be relaxed in future work by, for example, performing model selection on the latent dimension. Finally, the theorem makes explicit *what* can be learnt from a given dataset and system, and, as such, future work can exploit the graphical criterion for active data acquisition strategies.

Acknowledgments

This research was supported by an EPSRC Programme Grant (EP/V000748/1). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558) and the SCAN computing cluster in carrying out this work. Ingmar Posner holds concurrent appointments as a Professor of Applied AI at the University of Oxford and as an Amazon Scholar. This paper describes work performed at the University of Oxford and is not associated with Amazon.

The authors would also like to thank Alexander Mitchell and Frederik Nolte, whose keen insight and invaluable discussions were instrumental in making this work possible.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*. Proceedings of Machine Learning Research, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023.

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1517384113.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462, 2023.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 35(7):8747–8761, 2022.

Michael Chang, Tomer Ullman, Antonio Torralba, and Joshua Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019.

Minghao Fu, Biwei Huang, Zijian Li, Yujia Zheng, Ignavier Ng, Yingyao Hu, and Kun Zhang. Identification of nonparametric dynamic causal structure and latent process in climate system. *arXiv preprint arXiv:2501.12500*, 2025.

Mars Liyao Gao, Jan P Williams, and J Nathan Kutz. Sparse identification of nonlinear dynamics and koopman operators with shallow recurrent decoder networks. *arXiv preprint arXiv:2501.13329*, 2025.

Jan E. Gerken, Jimmy Aronsson, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Geometric deep learning and equivariant neural networks. *Artificial Intelligence Review*, 56(12):14605–14662, Dec 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10502-7. URL https://doi.org/10.1007/s10462-023-10502-7.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

Boyang Gu and Anastasia Borovykh. On original and latent space connectivity in deep neural networks. *arXiv preprint arXiv:2311.06816*, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

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*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv e-prints*, pages arXiv–2309, 2023.

Inwoo Hwang, Yunhyeok Kwak, Suhyung Choi, Byoung-Tak Zhang, and Sanghack Lee. Fine-grained causal dynamics learning with quantization for improving robustness in reinforcement learning. *Proceedings of Machine Learning Research*, 235:20842–20870, 2024.

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 Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2023.100844. URL https://www.sciencedirect.com/science/article/pii/S2666389923002234.

Jindong Jiang, Fei Deng, Gautam Singh, Minseung Lee, and Sungjin Ahn. Slot state space models. *Advances in Neural Information Processing Systems*, 37:11602–11633, 2024.

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.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. Pmlr, 2018.

Sebastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. *arXiv preprint arXiv:2211.14666*, 2022a.

Sebastien 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 Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 428–484. PMLR, 11–13 Apr 2022b. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36:25112–25150, 2023.

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.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

John M. Lee. *Introduction to smooth manifolds*. Graduate texts in mathematics, 218. Springer, New York, second edition. edition, 2012. ISBN 9781489994752.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. Variational causal dynamics: Discovering modular world models from interventions. *Transactions on Machine Learning Research*, 2023.

Anson Lei, Ingmar Posner, and Bernhard Schölkopf. Spartan: A sparse transformer learning local causation. *Advances in Neural Information Processing Systems*, 2025.

Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled representation learning in non-markovian causal systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=uLGyoBn7hm.

Shen Li, Bryan Hooi, and Gim Hee Lee. Identifying through flows for recovering latent representations. In *International Conference on Learning Representations*, 2019.

Yuke Li, Yujia Zheng, Tianyi Xiong, Zhenyi Wang, and Heng Huang. Understanding catastrophic interference on the identifibility of latent representations. *arXiv preprint arXiv:2509.23027*, 2025a.

Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33: 9180–9192, 2020.

Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards identifiability of hierarchical temporal causal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=0J0y9vSCWf.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. International Conference on Learning Representations, ICLR, 2025c.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. URL http://proceedings.mlr.press/v97/locatello19a.html. Best Paper Award.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pages 509–525. PMLR, 2022.

Matthew V Macfarlane and Clement Bonnet. Searching latent program spaces, 2025. URL https://arxiv.org/abs/2411.08706.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International conference on machine learning*, pages 4402–4412. PMLR, 2019.

Judea Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, 2nd edition edition, 2000. ISBN 1-139-63780-0.

Adeel Pervez, Francesco Locatello, and Stratis Gavves. Mechanistic neural networks for scientific machine learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40484–40501. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pervez24a.html.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.

org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017. doi: 10.1126/sciadv.1602614. URL https://www.science.org/doi/abs/10.1126/sciadv.1602614.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 459–466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

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.

Simon Schug, Seijin Kobayashi, Yassir Akram, Maciej Wolczyk, Alexandra Maria Proca, Johannes von Oswald, Razvan Pascanu, Joao Sacramento, and Angelika Steger. Discovering modular solutions that generalize compositionally. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=H98CVcX1eh.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 22905–22918, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=V8GHCGYLkf.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158. doi: 10.7551/mitpress/1754.001.0001. URL https://doi.org/10.7551/mitpress/1754.001.0001.

Sjoerd Van Steenkiste, Klaus Greff, Michael Chang, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2018.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *Advances in Neural Information Processing Systems*, 37:111614–111648, 2024.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36: 48603–48638, 2023.

Min Wang, Xin Li, Leiji Zhang, and Mingzhong Wang. Metacard: Meta-reinforcement learning with task uncertainty feedback via decoupled context-aware reward and dynamics components. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15555–15562, Mar. 2024. doi: 10.1609/aaai.v38i14.29482. URL https://ojs.aaai.org/index.php/AAAI/article/view/29482.

Lu Wen, Eric H. Tseng, Huei Peng, and Songan Zhang. Dream to adapt: Meta reinforcement learning by latent context imagination and mdp imagination. *IEEE Robotics and Automation Letters*, 9(11):9701–9708, 2024. doi: 10.1109/LRA.2024.3417114.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems*, 37: 71705–71736, 2024a.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *38th Conference on Neural Information Processing Systems*, volume 37, 2024b.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *12th International Conference on Learning Representations*, 2024c.

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

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, and Mengyue Yang. Curious causality-seeking agents learn meta causal world, 2025. URL https://arxiv.org/abs/2506.23068.

Appendix A. Useful Definitions

Definition 1 (Identifiable Models (Khemakhem et al., 2020)) *Two systems $\mathcal{S}$, $\hat{\mathcal{S}}$, are said to be identifiable if and only if*

$$\forall(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}_0): \quad \boldsymbol{\tau} = \hat{\boldsymbol{\tau}} \implies \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}. \tag{12}$$

Definition 2 (Identifiable up to Equivalence) *Two systems $\mathcal{S}$, $\hat{\mathcal{S}}$, are said to be identifiable up to some equivalence operator if and only if*

$$\forall(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}_0): \quad \boldsymbol{\tau} = \hat{\boldsymbol{\tau}} \implies \boldsymbol{\theta} \sim_{equiv} \hat{\boldsymbol{\theta}}. \tag{13}$$

Definition 3 (Diffeomorphism) *A map $\boldsymbol{f} : \mathbb{R}^M \to \mathbb{R}^N$ between two differentiable manifolds $\mathcal{M}$ and $\mathcal{N}$ is called a* diffeomorphism *if it satisfies the following equivalent conditions:*

1. *$\boldsymbol{f}$ is a one-to-one continuously differentiable mapping of $\mathcal{M}$ onto $\mathcal{N}$ with a continuously differentiable inverse mapping $\boldsymbol{f}^{-1}$.*
2. *$\boldsymbol{f}$ is a C^1 bijection (see definition 5), meaning that:*
   - *$\boldsymbol{f}$ is injective (one-to-one): $\boldsymbol{f}(\boldsymbol{z}_1) = f(\boldsymbol{z}_2)$ implies $\boldsymbol{z}_1 = \boldsymbol{z}_2$ for all $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}$,*
   - *$\boldsymbol{f}$ is surjective: for every $\boldsymbol{x} \in \mathcal{N}$, there exists a $\boldsymbol{z} \in \mathcal{M}$ such that $\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{x}$.*

Definition 4 (Surjectivity) *A function, $\boldsymbol{f}$, with codomain and domain $\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}$, is surjective if for every $\boldsymbol{y} \in \mathcal{Y}$ there exists $\boldsymbol{x} \in \mathcal{X}$ such that $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$.*

Definition 5 (C^k Functions) *A function $\boldsymbol{f} : U \to \mathbb{R}^m$ defined on an open subset $U \subset \mathbb{R}^n$ is said to be of class C^k if it is k-times continuously differentiable on U. In other words, the partial derivatives of $\boldsymbol{f}$ up to order k on U are continuous functions.*

Definition 6 (Equiv. up to Elementwise Perm. & Diffeomorphism (Lachapelle et al., 2022a)) *A representation, $\hat{\boldsymbol{\theta}}$, is said to be equivalent with respect to a ground truth representation, $\boldsymbol{\theta}$ up to permutation and elementwise diffeomorphism (see definition 3) if and only if there exists a permutation matrix, P, and a set of one-to-one diffeomorphisms, $c_i(\cdot)$, such that*

$$\boldsymbol{\theta} \sim_{diff} \hat{\boldsymbol{\theta}} \implies \forall(i, \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) : \theta_i = c_i([P\hat{\boldsymbol{\theta}}]_i). \tag{14}$$

Definition 7 (Disentanglement) *Two representations, $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, are provably disentangled if they are identifiable up to elementwise diffeomorphism and permutation equivalence.*

Definition 8 (Subsets of Graphs)

$$\mathcal{G}_a \subseteq \mathcal{G}_b \Leftrightarrow \left[A_{\mathcal{G}_b}\right]_{i,j} = 0 \implies \left[A_{\mathcal{G}_a}\right]_{i,j} = 0. \tag{15}$$

Graphs subsets are defined as follows. Every zero in the adjacency matrix of the superset graph must also be a zero in the adjacency matrix of the subset graph.

Definition 9 (Graph Arithmetic)

Graph arithmetic allows the product and sum of Jacobians (and consequently their respective graphs) to be expressed in the adjacency matrix domain. Graph operations are defined via Boolean matrix algebra. All summation operations are overloaded by element-wise OR operations, and all scalar multiplication with the AND operation.

Definition 10 (Right and Left Graph Consistency) *Any graph $\mathcal{C}$ is said to be left graph consistent with another graph $\mathcal{G}$ if and only if the graph is invariant to adjacency matrix multiplication:*

$$A_{\mathcal{G}} = A_{\mathcal{C}} A_{\mathcal{G}}. \tag{16}$$

The same applies for right consistency:

$$A_{\mathcal{G}} = A_{\mathcal{G}} A_{\mathcal{C}}. \tag{17}$$

Definition 11 (Real Subset) *The real subset relative to a directed acyclic graph, $\mathcal{G}$, with n input nodes and m output nodes is denoted as $\mathbb{R}_{\mathcal{G}}$. It is defined as the set of real matrices where all elements and only the elements with edges in $\mathcal{G}$ can take any non-zero value, and all elements with no edges in $\mathcal{G}$ are zero.*

$$M \in \mathbb{R}_{\mathcal{G}} \implies \Big(M_{i,j} \neq 0 \iff (i,j) \in \mathcal{G} \Big). \tag{18}$$

An alternative definition defines $\mathbb{R}_{\mathcal{G}}$ as the set of matrices for which there exists a $\boldsymbol{\lambda}$ which satisfies:

$$\mathbb{R}_{\mathcal{G}} : \quad \{ M | \exists \boldsymbol{\lambda} \in \mathbb{R}^{\|\mathcal{G}\|_0} \text{ s.t. } \boldsymbol{M} = \sum_{(i,j) \in \mathcal{G}} \lambda_{i,j} \boldsymbol{e}_{i,j} \}. \tag{19}$$

Where $\boldsymbol{e}_{i,j}$ is an indicator matrix.

Appendix B. Formal Proofs and Derivation of Graphical Criteria

Contents

B.1 Entanglement Map and Entanglement Graph . . . 23
B.2 Disentanglement and Entanglement Graphs . . . 23
B.3 Assumptions . . . 24
B.4 Definition and Proofs of Supporting Lemmas . . . 25
B.5 Proof of Theorem 1 . . . 29
B.6 Extension to Local Causal Graphs . . . 31
B.6.1 Pointwise Consistency of the Entanglement Mapping. . . . 31
B.6.2 Allowed Structure of V under local graphs. . . . 31
B.6.3 Corollaries and Edge Cases. . . . 32

This appendix provides the formal mathematical foundations and proofs for the identifiability of system parameters in latent dynamical systems. We recapitulate the main theorem and problem setting from the main text to keep the proof self-contained. The proof of the main theorem is found in section B.5. Below, we provide the precise setting and the required lemmas. We consider a non-linear deterministic Markovian dynamical system defined as the tuple $\mathcal{S} = (\Theta, \mathcal{X}_0, \mathcal{X}, \mathcal{F}, \mathcal{G})$ with state variables $\boldsymbol{x}_t$ and dynamical parameters $\boldsymbol{\theta}$, following the evolution:

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_t), \quad \boldsymbol{\theta} \in \Theta, \quad \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{x}_0 \in \mathcal{X}_0 \subseteq \mathcal{X}, \tag{20}$$

$$\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta}) = [\boldsymbol{x}_0, \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_0), (\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{f}_{\boldsymbol{\theta}})(\boldsymbol{x}_0), ..., \boldsymbol{f}_{\boldsymbol{\theta}}^T(\boldsymbol{x}_0)], \quad \boldsymbol{x}_0, \boldsymbol{\theta} = \boldsymbol{h}^{-1}(\boldsymbol{\tau}), \tag{21}$$

Where $\boldsymbol{f}(\cdot)$ is the dynamics model which describes the system's evolution, $\boldsymbol{\tau}$ is a trajectory generated by the autoregressive application of the dynamics model. The dynamics model is faithful to a directed acyclic graph (DAG) $\mathcal{G}$ in modelling how parameters influence the state variables. $\boldsymbol{h}(\cdot)$ is a trajectory *decoder* which encapsulates the autoregressive process, and $\boldsymbol{h}^{-1}(\cdot)$ is the trajectory *encoder* which recovers a latent parameter representation from a trajectory. Our primary objective is to establish the conditions under which a learnt representation of dynamics $\hat{\boldsymbol{\theta}}$ (belonging to a learnt approximation $\hat{\mathcal{S}}$) is equivalent to the ground truth parameters $\boldsymbol{\theta}$ up to permutation and element-wise diffeomorphism. For uniqueness and without loss of generality, we enforce assumptions 1 and 2, which in combination also require the bijectivity of $\boldsymbol{h}(\cdot)$ and $\boldsymbol{h}^{-1}(\cdot)$ with respect to their image. We further assume that the parameters sparsely affect the state distribution via a causal graph $\mathcal{G}$.

Assumption 1 (Existence and Parameter Influence (Yao et al., 2024a)) *For every $\boldsymbol{x}_0 \in \mathcal{X}_0, \boldsymbol{\theta} \in \Theta$ there exists a unique trajectory, $\boldsymbol{\tau}$, over horizon T, satisfying $\boldsymbol{x}_{t+1} = \boldsymbol{f}_\theta(\boldsymbol{x}_t)$. Equivalent to requiring $\boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta})$ to be injective.*

Assumption 2 (Observational Equivalence) *Assume that the two systems, the ground-truth $\mathcal{S}$ and learnt approximation $\hat{\mathcal{S}}$, are* observationally equivalent*, which posits that the modelled systems are equivalent up to invertible parameter representations. Observational equivalence requires, $\forall \boldsymbol{x}_0, \boldsymbol{\theta}: \quad \exists \hat{\boldsymbol{\theta}}$ s.t. $\boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta}) = \hat{\boldsymbol{h}}(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}})$ (Equivalently, $\hat{\boldsymbol{h}}(\cdot)$ is surjective).*

Assumption 3 (The Transition Model is Markov to a DAG (Lachapelle et al., 2024)) *The transition models of $\mathcal{S}$, and $\hat{\mathcal{S}}$, are causally and* faithfully *(Pearl, 2000; Spirtes et al., 2001) related to directed acyclic graphs (DAG), $\mathcal{G}$, and $\hat{\mathcal{G}}$, which fully define the dependencies between the* ***output***

elements and latent parameters. *The edges of the graph are defined through the non-zero entries in the Jacobian of the transition model, requiring that the model and its first derivative are continuous. Simply, the following causal equivalence is defined, where $\boldsymbol{\theta}_{Pa_i^{\mathcal{G}}}$ denotes the subset of parameters that are parents of output i in graph $\mathcal{G}$: $f_i(\boldsymbol{x}_t, \boldsymbol{\theta}) \equiv f_i(\boldsymbol{x}_t, \boldsymbol{\theta}_{Pa_i^{\mathcal{G}}})$.*

$$(i,j) \notin \mathcal{G} \implies \frac{\partial(\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}))_i}{\partial \theta_j} = 0 \quad \forall \boldsymbol{x}_t, \theta. \tag{22}$$

These assumptions establish a framework in which multiple parametrisations can yield identical observations, allowing the study of the conditions under which additional assumptions lead to disentanglement. We begin by introducing the concept of an entanglement mapping and an entanglement graph as necessary prerequisites to understand the theory. Subsequently, we build intermediate results in forms of lemmas. Section B.5 combines the lemmas into the proof of Theorem 1. Finally, section B.6 extends the main result to state-dependent local causal graphs and demonstrates that state-dependent sparsity improves identifiability guarantees.

B.1. Entanglement Map and Entanglement Graph

The trajectory encoder, $\boldsymbol{\theta} = \boldsymbol{h}^{-1}(\boldsymbol{\tau})$, allows us to define an *entanglement map*, $\boldsymbol{v}(\cdot)$, which relates the latent representations learnt by $\hat{\mathcal{S}}$, with the parameters in $\hat{\mathcal{S}}$. Given that the two systems are modelling the same underlying system, it follows that for each parameter of the ground truth system, $\boldsymbol{\theta}$, there must exist a learnt representation $\hat{\boldsymbol{\theta}}$ such that the modelled process is equivalent $\boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{\theta}) = \hat{\boldsymbol{f}}(\boldsymbol{x}_0, \boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0))$. The equivalence of the system is formalised in the next section as assumption 2. Observational equivalence requires the injectivity of the *entanglement map*.

$$\hat{\boldsymbol{\theta}} = \boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0) = (\hat{\boldsymbol{h}}^{-1} \circ \boldsymbol{h})(\boldsymbol{\theta}, \boldsymbol{x}_0). \tag{23}$$

The dependence of the entanglement mapping upon a trajectory's starting state, $\boldsymbol{x}_0$, is a consequence of the application to dynamical domains. This dependence allows the dynamical representation to become entangled with the state representation. The entanglement map is also associated with an *entanglement graph*, $\mathcal{V}$, which quantifies the dependencies between the two representations using the same fundamental definition used for the system graphs $\mathcal{G}$, and $\hat{\mathcal{G}}$. The *entanglement graph* denotes which system parameters, $\boldsymbol{\theta}$, are required to map to the learnt representation. Equivalently, $\hat{\boldsymbol{\theta}}_i = v_i(\boldsymbol{\theta}_{\mathrm{Pa}_i^{\mathcal{V}}}, \boldsymbol{x}_0)$ where $\boldsymbol{\theta}_{\mathrm{Pa}_i^{\mathcal{V}}}$ is the subset of parameters that are parents of node i in the *entanglement graph*.

$$(i,j) \notin \mathcal{V} \implies \frac{\partial[\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0)]_i}{\partial \theta_j} = 0 \quad \forall \boldsymbol{\theta}, \boldsymbol{x}_0. \tag{24}$$

Given, $\mathbf{h}(\cdot)$ is bijective (and therefore so is its inverse), the entanglement map must also be bijective, as the composition of two bijective functions is also bijective. Consequently, $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are reparametrisations of the same underlying manifold.

B.2. Disentanglement and Entanglement Graphs

The entanglement graph provides a mathematical description of the relationship between a known ground-truth parametrisation and a representation learnt by a neural network. If we show that the adjacency matrix of the entanglement graph must be a permutation of the identity matrix, then we have also equivalently demonstrated that the model has disentangled up to elementwise diffeomorphism.

To illustrate the point, consider $\mathcal{V} = \mathbf{I}$ and a two-dimensional parametrisation. This entanglement graph forces the entanglement mapping dependencies to be element-wise, since each output latent can depend on at most one input latent. Consequently, $\hat{\theta}_0 = v_0(\theta_0), \hat{\theta}_1 = v_1(\theta_1)$, or some permutation thereof. Note that the entanglement mapping $\boldsymbol{v}(\cdot)$ can be highly non-linear, and constraining the entanglement graph does not allow us to make any assumptions about the functional form of the mapping, only that it is smooth, differentiable, and invertible, or in other words, a diffeomorphism.

Hence, we have established that showing an entanglement graph is either the identity matrix or a permutation of the identity matrix is sufficient to establish that a learnt representation is disentangled up to an element-wise diffeomorphism (see definition 3) and permutation.

Without further assumptions upon the systems $\mathcal{S}$ and $\hat{\mathcal{S}}$, the entanglement graph connecting them is unconstrained. The following section introduces the necessary assumptions and the theorem under which the entanglement graph is constrained to be a permutation of the identity.

B.3. Assumptions

We formalize three core assumptions which in conjunction sufficiently constrain the entanglement graph and form the primary identifiability result (restated below). These assumptions are applied on top of the system preliminaries (assumptions 1, 2, and 3). Assumption 4 ensures that the representation of dynamics does not entangle with the state representation, by precluding possible sources of entanglement from existing[4]. Assumption 5 restricts the causal complexity of the learnt representation, thereby reducing the admissible representation space. Assumption 6 enforces information diversity: if two parameters were to always affect the system in the exact same way, then no amount of data can tell them apart, preventing degenerate cases where the number of parameters is not representative of the true dimension of variation of the system.

Assumption 4 (Path Connectedness) *We assume the state variable domain $\mathcal{X}$ is path-connected, such that there exists no disjoint decomposition of the state space separated by boundaries that are fundamentally unreachable or untraversable by the system's trajectories.*

$$\forall \mathcal{X}_A, \mathcal{X}_B \subset \mathcal{X} \text{ s.t. } \mathcal{X}_A \cup \mathcal{X}_B = \mathcal{X}, \mathcal{X}_A \cap \mathcal{X}_B = \emptyset, \exists (\boldsymbol{x}_0, \boldsymbol{\theta}, t) \text{ s.t. } \boldsymbol{x}_0 \in \mathcal{X}_A \wedge \boldsymbol{x}_t \in \mathcal{X}_B \tag{25}$$

Assumption 5 (Sufficient Sparsity) *We assume that $\hat{\mathcal{S}}$ learns a representation and a system model that is at most as sparse as the underlying system, which in practise is enforced by assuming $\hat{\mathcal{S}}$ is a minimiser of $\|\hat{\mathcal{G}}\|_0$. This implies that $\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$.*

Assumption 6 (Sufficient Variability of the Jacobian (Lachapelle et al., 2022b)) *There must exist at least one $\boldsymbol{\theta} \in \Theta$ such that there exists a set of states $\{\boldsymbol{x}_p\}_{p=1}^{\|\mathcal{G}\|_0}, \boldsymbol{x} \in \mathcal{X}$ such that*

$$\textit{span}\left\{\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}_p, \boldsymbol{\theta})\right\}_{p=1}^{\|\mathcal{G}\|_0} = \mathbb{R}^{\mathcal{G}} \tag{26}$$

The real graph subset is defined in definition 11. This common formulation (Lachapelle et al., 2022b; Hyvarinen et al., 2019) requires that the effect of each causal edge in the system graph be independently observable from the state space.

4. For an intuition on path connectivity, refer to Appendix C

B.4. Definition and Proofs of Supporting Lemmas

The proof of Theorem 1 is built from several smaller supporting lemmas defined here. The final proof combining these lemmas is stated in section B.5. Lemma 1 combines assumption 4

Lemma 1 (Entanglement Map Independence) *If a system satisfies path connectivity (assumption 4), the entanglement mapping is independent of the starting state:* $\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0) = \boldsymbol{v}(\boldsymbol{\theta})$.

Proof The trajectory decoder function is composed of one-step Markovian predictions from a dynamical model, $\boldsymbol{f}(\cdot)$.

$$\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{\theta}) = [\boldsymbol{x}_0, \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_0), (\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{f}_{\boldsymbol{\theta}})(\boldsymbol{x}_0), ..., \boldsymbol{f}_{\boldsymbol{\theta}}^T(\boldsymbol{x}_0)]. \tag{27}$$

The proof follows two steps. First, we show that the entanglement mapping must be invariant to states that can originate from the same trajectory, and second, we show that it must be invariant between any states that could have originated from a chain of trajectories. For a trajectory $\boldsymbol{\tau} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T)$ generated by a fixed ground truth parameter $\boldsymbol{\theta}$. Under autoregressive rollouts, the model reuses the same inferred latent parameter $\hat{\boldsymbol{\theta}}$ to decode the dynamics at every step along the trajectory. Hence, for the rollout to remain consistent with the underlying system (as is guaranteed by the uniqueness of each trajectory in assumption 1), the entanglement mapping, $\boldsymbol{v}(\cdot)$, cannot depend on which state along the same trajectory is used for inference. If the rollout were to be reinitialised from the middle of some trajectory, $\tilde{\boldsymbol{x}}_0 = \boldsymbol{x}_t$, $\tilde{\boldsymbol{\tau}} = \boldsymbol{h}(\tilde{\boldsymbol{x}}_0, \boldsymbol{\theta})$, while keeping $\boldsymbol{\theta}$ unchanged, the inferred latent parameter must agree with the original $\hat{\boldsymbol{h}}^{-1}(\boldsymbol{\tau}) = \hat{\boldsymbol{h}}^{-1}(\tilde{\boldsymbol{\tau}})$. Otherwise, the model would assign different latent parameters to the same physical system depending on the temporal offset, thereby violating the Markovian assumption of the transition model.

Consequently, this implies $\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0) = \boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_1) = v(\boldsymbol{\theta}, \boldsymbol{x}_i) \forall \boldsymbol{\theta}, \boldsymbol{x}_i \in \boldsymbol{\tau}$. The corollary of this is that the entanglement mapping must be invariant to states within the same trajectory; therefore, defining a trajectory subset of the state space:

$$\mathcal{X}_{\boldsymbol{\tau}}(\boldsymbol{x}_0, \boldsymbol{\theta}) = \left\{ \tilde{\boldsymbol{x}} \mid \exists \boldsymbol{x}' \in \mathcal{X}_{\boldsymbol{\tau}} \text{ s.t. } \tilde{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{\theta}) \text{ and } \boldsymbol{x}_0 \in \mathcal{X}_{\boldsymbol{\tau}} \right\}. \tag{28}$$

Set $\mathcal{X}_{\boldsymbol{\tau}}$ represents the set of all positions that could have led to or have been produced from $\boldsymbol{x}_0$ given a particular parametrisation of a system. In the geometric deep learning literature, this concept is referred to as manifold invariance (Bronstein et al., 2021; Gerken et al., 2023). Furthermore, as the entanglement map is a function of state, then instantaneously evaluated at a specific state, $\boldsymbol{x}'$, the representation becomes independent of state. The invariance manifold can consequently be extended by considering local stationarity of the representation at fixed states. For any two trajectories, $\boldsymbol{\tau}_a = \boldsymbol{h}(\boldsymbol{x}_a, \boldsymbol{\theta}_a)$, $\boldsymbol{\tau}_b = \boldsymbol{h}(\boldsymbol{x}_b, \boldsymbol{\theta}_b)$, such that at some point in their trajectories, both contain a state $\boldsymbol{x}_c \in \boldsymbol{\tau}_a \cap \boldsymbol{\tau}_b$.

From the definition of $\mathcal{X}_{\boldsymbol{\tau}}$ and the consequential state invariance of the entanglement mapping,

$$\boldsymbol{v}(\boldsymbol{x}_a, \boldsymbol{\theta}) = \boldsymbol{v}(\boldsymbol{x}_c, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{x}_a, \boldsymbol{x}_c \in \mathcal{X}_{\boldsymbol{\tau}}(\boldsymbol{x}_a, \boldsymbol{\theta}_a) \tag{29}$$

$$\boldsymbol{v}(\boldsymbol{x}_b, \boldsymbol{\theta}) = \boldsymbol{v}(\boldsymbol{x}_c, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{x}_b, \boldsymbol{x}_c \in \mathcal{X}_{\boldsymbol{\tau}}(\boldsymbol{x}_b, \boldsymbol{\theta}_b) \tag{30}$$

$$\boldsymbol{x}_c \in \boldsymbol{\tau}_a \cap \boldsymbol{\tau}_b \implies \mathcal{X}_{\boldsymbol{\tau}}(\boldsymbol{x}_a, \boldsymbol{\theta}_a) \cap \mathcal{X}_{\boldsymbol{\tau}}(\boldsymbol{x}_b, \boldsymbol{\theta}_b) \neq \emptyset \tag{31}$$

Through the transitive property, whenever two trajectory manifolds overlap in state variable space, they must share the same inferred representation, implying that $\boldsymbol{v}$ is invariant on $\mathcal{X}_{\boldsymbol{\tau}_a} \cup \mathcal{X}_{\boldsymbol{\tau}_b}$ (and, by transitive closure, on any union of manifolds connected through overlaps).

$$\mathcal{X}_P(\boldsymbol{x}_0) = \left\{\tilde{\boldsymbol{x}} \mid \exists \boldsymbol{x}' \in \mathcal{X}_{\boldsymbol{\tau}}, \boldsymbol{\theta} \text{ s.t. } \tilde{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{\theta}) \text{ and } \boldsymbol{x}_0 \in \mathcal{X}_{\boldsymbol{\tau}}\right\}. \tag{32}$$

The set $\mathcal{X}_P$ defines a partition of the state space. The set $\mathcal{X}_P$ lifts $\mathcal{X}_{\boldsymbol{\tau}}$; the latter is defined as the set of states that are trajectory-connected to a specific starting point and dynamical parameter. The former extends $\mathcal{X}_{\boldsymbol{\tau}}$ to include every dynamical system that is either dynamically or trajectory-connected to a point through any arrangement. A corollary of this is that the set $\mathcal{X}_P \subseteq \mathcal{X}$ defines a partition of the state space for which $\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}') = \boldsymbol{v}(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}), \forall \boldsymbol{x}', \tilde{\boldsymbol{x}} \in \mathcal{X}_P$.

If the latent parameter is surjective (definition 4) with respect to the dynamical model, then trivially $\mathcal{X}_P$ must span the whole domain $\mathcal{X}$ and global manifold invariance follows. If $\mathcal{X} = \mathcal{X}_P$, then the entanglement mapping must be invariant to the starting state, $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{v}(\boldsymbol{\theta})$. ■

Topological assumptions upon the data-generating distribution in representation learning are commonplace. Literature usually enforces stronger assumptions about the bijectivity of the latent space rather than a more relaxed connectivity requirement (Lachapelle et al., 2024; Fu et al., 2025), although Schug et al. (2024) implements a discrete analogue to our path connectivity. Similar concepts and methods have been used more generally to study emergent structures in latent representational space (Gu and Borovykh, 2023; Li et al., 2025a). Lachapelle et al. (2023) applies a similar path-connectedness assumption on the mapping between observational and representational space without the constraint of connectedness through a dynamical model.

The independence of the entanglement mapping allows proof techniques from Lachapelle et al. (2024) to be imported and applied to analyse the constraints upon the entanglement mapping.

Lemma 2 (Causal Inclusion) *Under assumption 6 (Sufficient Variability), there must exist a permutation matrix, $\boldsymbol{P}$, such that every edge contained in the ground truth graph, $\mathcal{G}$, must also be contained within the permuted version of the learnt graph $\hat{\mathcal{G}}$.*[5]

$$\exists \boldsymbol{P} \text{ s.t. } \quad \boldsymbol{A}_{\mathcal{G}} \subseteq \boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P}. \tag{33}$$

Proof We begin from the observational equivalence assumption (assumption 2) but consider the inverse of $\boldsymbol{v}(\cdot)$ which must exist as $\boldsymbol{v}(\cdot)$ is a diffeomorphism.

$$\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{v}^{-1}(\hat{\boldsymbol{\theta}})) = \hat{\boldsymbol{f}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}), \tag{34}$$

$$\nabla_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{f}} = \nabla_{\boldsymbol{\theta}} \boldsymbol{f} \nabla_{\hat{\boldsymbol{\theta}}} \boldsymbol{v}^{-1} \quad \forall \boldsymbol{x}_t,\ \hat{\boldsymbol{\theta}}, \tag{35}$$

$$[\nabla_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{f}}]_{i,j} = \sum_k [\nabla_{\boldsymbol{\theta}} \boldsymbol{f}]_{i,k} [\nabla_{\hat{\boldsymbol{\theta}}} \boldsymbol{v}^{-1}]_{k,j}, \tag{36}$$

$$\nabla_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{f}}_{i,j} = (\nabla_{\boldsymbol{\theta}} \boldsymbol{f}_{i,\cdot})^T \nabla_{\hat{\boldsymbol{\theta}}} \boldsymbol{v}^{-1}{}_{\cdot,j}. \tag{37}$$

From here, we show that given the invertibility of $\boldsymbol{v}^{-1}$ and following assumption 6, every edge contained in $\mathcal{G}$ must also be contained within some permutation of the learnt graph $\hat{\mathcal{G}}$. First, consider, rather than $\boldsymbol{v}^{-1}$, a permutation $\boldsymbol{v}^{-1}$. As $\boldsymbol{v}^{-1}$ is invertible, there exists a permutation thereof such that there are no zeros on the diagonal. This fact follows simply from the *Inverse Function Theorem* (Lee, 2012). The *Inverse Function Theorem* proves that the Jacobian of the entanglement mapping

5. Note all products and sums between adjacency matrices follow graph arithmetic (definition 9).

must be of full row rank. Consequently, it must be possible to rearrange the entanglement graph such that there are no zeros on the diagonal. The permuted composition will be denoted as $\nabla_{\hat{\boldsymbol{\theta}}}\boldsymbol{v}^{-1}\boldsymbol{P} = \boldsymbol{C}$. Formally, $\boldsymbol{C}$ is a function of $\hat{\boldsymbol{\theta}}$, but the inputs of $\boldsymbol{C}$, $\boldsymbol{f}$, and $\hat{\boldsymbol{f}}$ are omitted for notational clarity.

$$[\nabla_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{f}}\boldsymbol{P}]_{i,j} = \sum_k [\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,k}[\nabla_{\hat{\boldsymbol{\theta}}}\boldsymbol{v}^{-1}\boldsymbol{P}]_{k,j}, \tag{38}$$

$$= \sum_k [\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,k}[\boldsymbol{C}]_{k,j}, \tag{39}$$

$$= [\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,j}[\boldsymbol{C}]_{j,j} + \sum_{k\neq j} [\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,k}[\boldsymbol{C}]_{k,j}. \tag{40}$$

From the inverse function theorem, $\boldsymbol{C}_{j,j} \neq 0 \forall j$. This implies that if $[\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,j} \neq 0$ then $[\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,j}[\boldsymbol{C}]_{j,j} \neq 0$. This is not sufficient to imply that edge (i,j) is in $\boldsymbol{A}_{\hat{\mathcal{G}}}\boldsymbol{P}$ since the domains $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{\theta} \in \Theta$ may constrain $\nabla_{\boldsymbol{\theta}}\boldsymbol{f}_{i,\cdot}$ to the subspace of $\boldsymbol{C}_{\cdot,j}$. To combat this, we introduce assumption 6. The sufficient variability assumption ensures that there exists at least one input for which the Jacobian of the ground truth system is not in the subspace of the permuted inverse entanglement map. Formally:

$$\forall (i,j) \in \{i,j \mid [\nabla_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{f}}\boldsymbol{P}]_{i,j} \neq 0\}, \quad \exists(\boldsymbol{\theta}, \boldsymbol{x}_t) \text{ s.t } \sum_k [\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,k}[\boldsymbol{C}]_{k,j} \neq 0. \tag{41}$$

There must exist a permutation $\boldsymbol{P}$ such that there exists any point in the input space where $[\nabla_{\boldsymbol{\theta}}\boldsymbol{f}]_{i,j} \neq 0$ necessarily implies $[\nabla_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{f}}\boldsymbol{P}]_{i,j} \neq 0$. Consequently every edge in $\boldsymbol{A}_{\mathcal{G}}$ must exist in $\boldsymbol{A}_{\hat{\mathcal{G}}}\boldsymbol{P}$. This is not necessarily true in the other direction; not every edge in $\hat{\mathcal{G}}$ must exist in a permuted form of $\mathcal{G}$.[6] Formally,

$$\exists \boldsymbol{P} \text{ s.t.} \quad \boldsymbol{A}_{\mathcal{G}} \subseteq \boldsymbol{A}_{\hat{\mathcal{G}}}\boldsymbol{P}. \tag{42}$$

■

Lemma 3 (Graph Equivalence) *Under assumptions 5 and 6, the learnt graph $\mathcal{G}$ is equivalent to the ground truth graph up to permutation.*

$$\exists \boldsymbol{P} \textit{ s.t.} \quad \boldsymbol{A}_{\hat{\mathcal{G}}}\boldsymbol{P} \equiv \boldsymbol{A}_{\mathcal{G}}. \tag{43}$$

Proof Trivially true. From lemma 2 it is established that there exists a permutation such that every edge contained in $\mathcal{G}$ must also be contained within the permuted $\hat{\mathcal{G}}$. Assumption 5 states that the learnt representation's graph must contain at most as many edges as the underlying graph. Therefore, they must be the same. ■

6. Toy example to showcase this is that $\hat{\mathcal{G}}$ can be fully connected, and $\mathcal{G}$ can be sparse. Since only the sufficient influence of the true dynamics model is constrained and not that of the learnt model.

Lemma 4 (The Inverse Entanglement Graph is Right Graph Consistent) *Given all prior assumptions, the permuted Given lemma 3, the permuted inverse entanglement mapping must be right $\mathcal{G}$ consistent (definition 10). Which, in the form of adjacency matrix arithmetic, requires*

$$\exists \boldsymbol{P} \text{ s.t. } \quad \boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}. \tag{44}$$

Proof Beginning from the full form of equation 39,

$$\underbrace{\nabla_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{f}}}_{\subseteq \hat{\mathcal{G}} = \mathcal{G} \boldsymbol{P}^T} \boldsymbol{P} = \underbrace{\nabla_{\boldsymbol{\theta}} \boldsymbol{f}}_{\subseteq \mathcal{G}} \underbrace{\nabla_{\hat{\boldsymbol{\theta}}} \boldsymbol{v}^{-1}}_{\subseteq \mathcal{V}^{-1}} \boldsymbol{P}, \tag{45}$$

From Lemma 2, we know that there exists a point in the input space such that no cancellations occur in the product between the ground truth system Jacobian and the Jacobian of the inverse entanglement map (the Jacobian of the ground truth graph varies sufficiently and consequently it is not constrained to the null-space of the Jacobian of the entanglement mapping). The relation below can be rewritten in adjacency matrix arithmetic by considering the set of non-zero values each element of equation 45 can take. The set of non-zero values is given by the elements of the graph in each underbrace.

$$\boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P} \subseteq \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}. \tag{46}$$

The lack of cancellations allows the subset to be tightened to equality:

$$\boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}, \tag{47}$$

$$\boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}. \tag{48}$$

■

Lemma 5 (Graphical Criterion) *If $\mathcal{G}$ satisfies the following graphical criterion:*

$$\forall i: \quad \cap_{a \in Ch(i)} Pa(a) = \{i\}. \tag{49}$$

Then the only valid right consistent graph on $\mathcal{G}$, $\mathcal{R}: \ \boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{R}}$ is the identity.

Proof We will now formalise the allowed structures on $\mathcal{R}$, given $\mathcal{G}$. Consider a generic 3x3 example, where a $\star$ symbol denotes that any possible value could be taken.

$$\underbrace{\begin{bmatrix} \star & \star & \star \\ \star & \star & \star \\ \star & \star & \star \end{bmatrix}}_{\mathcal{G}} = \underbrace{\begin{bmatrix} \star & \star & \star \\ \star & \star & \star \\ \star & \star & \star \end{bmatrix}}_{\mathcal{G}} \underbrace{\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}}_{\mathcal{R}}.$$

In adjacency matrix arithmetic, cancellations of edges are impossible; consequently, the product can be written out in Boolean logic form as:

$$\mathcal{G}_{ij} = 0 \implies (\mathcal{G}_{i1} \wedge R_{1j} = 0) \vee (\mathcal{G}_{i2} \wedge R_{2j} = 0) \vee ... \vee (\mathcal{G}_{ik} \wedge R_{kj}) = 0. \tag{50}$$

By iterating over all appearances of R_{ij}, it can be shown through boolean algebra that R_{ij} must be 0 if and only if:

$$\bigvee_k (\mathcal{G}_{kj} = 0 \wedge \mathcal{G}_{ik} \neq 0) = 1. \tag{51}$$

Further rearrangement shows that the set of nonzero elements in each row of the entanglement mapping is given by $\mathcal{R}_{i,:}$:

$$\cap_{a \in \text{Ch}(i)} \text{Pa}(a). \tag{52}$$

Where $\text{Ch}(\cdot)$ and $\text{Pa}(\cdot)$ are the children and parents of a node in each graph, respectively.

Therefore, the set of permissible elements in each row of $\boldsymbol{A}_{\mathcal{R}}$ is restricted to the diagonal (equivalently the elements of the identity), by iterating over row elements in the graphical criterion.

$$\forall i: \quad \cap_{a \in \text{Ch}(i)} \text{Pa}(a) = \{i\}. \tag{53}$$

■

B.5. Proof of Theorem 1

Theorem 1 (Disentanglement of Latent System Parameters) *Let $\mathcal{S}$, and $\hat{\mathcal{S}}$ correspond to two systems satisfying assumptions 1, 2, and 3. Further assume that,*

1. *The system, $\mathcal{S}$, is path connected (assumption 4).*
2. *$\hat{\mathcal{S}}$ satisfies $\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$ (assumption 5).*
3. *The Jacobian of the transition model, $\nabla_{\boldsymbol{\theta}} \boldsymbol{f}$,* varies sufficiently *(assumption 6).*

Then the dynamical representations of $\mathcal{S}$, and $\hat{\mathcal{S}}$ are equivalent up to permutation and element-wise diffeomorphism if $\mathcal{G}$ satisfies the following graphical criterion:

$$\forall i: \quad \cap_{a \in Ch_{\mathcal{G}}(i)} Pa_{\mathcal{G}}(a) = \{i\}, \tag{54}$$

where $Ch_{\mathcal{G}}(j)$, and $Pa_{\mathcal{G}}(j)$ denotes the set of children and parent nodes of a node j in a graph G.

Proof

Step 1. We define an entanglement map $\boldsymbol{v}: \Theta \times \mathcal{X} \to \hat{\Theta}$ such that $\hat{\boldsymbol{\theta}} = \boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x})$. This mapping describes how the learned parameters relate to the ground truth at any given state. The dependencies in this mapping are captured by an entanglement graph $\mathcal{V}$, where an edge exists if a ground-truth parameter influences a learned latent dimension:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0) = (\hat{\boldsymbol{h}}^{-1} \circ \boldsymbol{h})(\boldsymbol{\theta}, \boldsymbol{x}_0). \tag{55}$$

Step 2. By assumption 4, any two states in the domain are connected through some sequence of trajectories. Lemma 1 uses the assumption to show that the entanglement map must be invariant to the starting state:

$$\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{x}_0) = \boldsymbol{v}(\boldsymbol{\theta}) \quad \forall \boldsymbol{x}_0, \boldsymbol{\theta} \tag{56}$$

Step 3. We establish the conditions under which the entanglement map $\boldsymbol{v}$ implies representational disentanglement. By definition, $\boldsymbol{v} = \hat{\boldsymbol{h}}^{-1} \circ \boldsymbol{h}$. From Assumption 1, $\boldsymbol{h}(\cdot)$ must be injective, and from Assumption 2, $\hat{\boldsymbol{h}}(\cdot)$ must be surjective; together with the requirement that both models are defined over the same image, this ensures that $\boldsymbol{v}$ is a bijection. Furthermore, Assumption 3 requires that the transition model (and consequently $\boldsymbol{h}$) and its first derivative are continuous, rendering $\boldsymbol{v}$ a diffeomorphism. If the entanglement graph is shown to be a permutation of the identity, $\boldsymbol{A}_{\mathcal{V}} = \boldsymbol{P}$, it implies that each learned parameter $\hat{\theta}_i$ is a function of exactly one unique ground-truth parameter θ_j. Because the requirements of bijectivity and differentiability must hold, such a structural constraint necessitates that $\boldsymbol{v}$ is an element-wise diffeomorphism, satisfying the criteria for disentanglement (Definition 7). By the same logic, if the inverse entanglement graph is a permutation of the identity ($\boldsymbol{A}_{\mathcal{V}^{-1}} = \boldsymbol{P}$), then each ground-truth parameter is mapped to a single learned latent, which similarly forces an element-wise diffeomorphic relationship between the two representations.

Step 4. From observational equivalence (assumption 2), we relate the Jacobians of the ground truth and learnt transition models via the inverse entanglement mapping,

$$\nabla_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{f}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{v}^{-1}(\hat{\boldsymbol{\theta}})) \nabla_{\hat{\boldsymbol{\theta}}} \boldsymbol{v}^{-1}(\hat{\boldsymbol{\theta}}). \tag{57}$$

Step 5. We use lemma 2 to show that given sufficient variability of the Jacobian of the transition model (assumption 6), there must exist a point in the input space such that every non-zero entry in the Jacobian of the ground truth transition model must also be non-zero up to permutation in the learnt transition model. In terms of the structural graphs, this implies that:

$$\exists \boldsymbol{P} \text{ s.t. } \quad \boldsymbol{A}_{\mathcal{G}} \subseteq \boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P}. \tag{58}$$

Step 6. Additionally, introducing assumption 5 (Sparsity), which posits that the learned graph's cardinality is bounded by the ground truth ($\|\hat{\mathcal{G}}\|_0 \leq \|\mathcal{G}\|_0$). By lemma 3, since $\boldsymbol{A}_{\mathcal{G}} \subseteq \boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P}$ and the number of edges in $\hat{\mathcal{G}}$ cannot exceed those in $\mathcal{G}$, the two graphs must be equivalent up to permutation: $\boldsymbol{A}_{\hat{\mathcal{G}}} \boldsymbol{P} \equiv \boldsymbol{A}_{\mathcal{G}}$.

Step 7. Using the graph equivalence, we return to the Jacobian formulation from step 4. The sufficient-variability assumptions require that the transition model not be constrained to the nullspace of the Jacobian of the inverse entanglement mapping. Consequently, lemma 4 shows that the inverse entanglement mapping must be right consistent with the ground truth graph:

$$\boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}. \tag{59}$$

Step 8. The requirement of right consistency means that the ground truth graph, $\mathcal{G}$, constrains the allowed structure upon $\boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P}$. Lemma 5 proves that if $\mathcal{G}$ satisfies the graphical criterion:

$$\forall i: \quad \cap_{a \in \mathrm{Ch}(i)} \mathrm{Pa}(a) = \{i\}. \tag{60}$$

Then the only matrix $\boldsymbol{A}_{\mathcal{R}}$ that satisfies the consistency equation, $\boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}} \boldsymbol{A}_{\mathcal{R}}$ is the identity matrix. Therefore $\boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P} = \boldsymbol{I}$.

Step 9. Since $\boldsymbol{A}_{\mathcal{V}^{-1}} \boldsymbol{P} = \boldsymbol{I}$, it follows that the entanglement graph is a permutation of the identity. This forces the mapping $\boldsymbol{v}(\boldsymbol{\theta})$ to be an element-wise diffeomorphism, meaning the system parameters have been disentangled. This concludes the proof. ■

B.6. Extension to Local Causal Graphs

Many dynamical systems exhibit state-dependent causal structures, where only a subset of the edges in the global graph are active at a given state. Let the state space admit a finite cover by non-zero measure partitions $\{\mathcal{X}_l\}_{l=1}^{b}$, such that for $\boldsymbol{x}_t \in \mathcal{X}_l$, the one-step transition $\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta})$ is Markov not only with respect to the global DAG, but also the local DAG $\mathcal{G}_l \subseteq \mathcal{G}$ associated with that partition. The partition must have non-zero measure so the sufficient variability assumption can be applied. In practice, this means that when dealing with continuous systems, instantaneous events such as collisions can not be factored into discrete local graph partitions. However, when operating in discrete time, collisions occupy a non-zero measure.

$$\left[\boldsymbol{A}_{\mathcal{G}_l}\right]_{i,j} = 0 \iff \frac{\partial[\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta})]_i}{\partial \theta_j} = 0 \quad \forall \boldsymbol{x}_t \in \mathcal{X}_l, \theta. \tag{61}$$

We further assume:

- Assumptions 1, and 2 hold globally. This is required to prevent trivial solutions.
- Assumption 4 must also hold globally. In practice, this imposes the additional constraint that the system must be able to navigate between state space partitions. Otherwise, a separate representation might be learnt for each partition. Global path connectedness ensures the same representation is used for each partition and allows theoretical reasoning about the reuse of representations across partitions.
- Assumption 5-6 must hold *within* each partition $\mathcal{X}_l$. The former allows local representations to be sparser than global representations, whilst the latter ensures that the effect of each edge in a local causal graph is measurable.

B.6.1. POINTWISE CONSISTENCY OF THE ENTANGLEMENT MAPPING.

Recall the global *right-consistency* relation from lemma 4, which forces the inverse entanglement map to respect the system graph in adjacency arithmetic: $\boldsymbol{A}_{\mathcal{G}} = \boldsymbol{A}_{\mathcal{G}}\boldsymbol{A}_{\mathcal{V}^{-1}}\boldsymbol{P}$, obtained from the Jacobian factorisation, $\nabla_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{f}} = \nabla_{\boldsymbol{\theta}}\boldsymbol{f}\nabla_{\hat{\boldsymbol{\theta}}}\boldsymbol{v}^{-1}$, and $\nabla_{\boldsymbol{\theta}}\boldsymbol{f} = \nabla_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{f}}\nabla_{\boldsymbol{\theta}}\boldsymbol{v}$. Building the same calculation, restricted to a partition $\mathcal{X}_l$, gives, pointwise in $\boldsymbol{x}_t \in \mathcal{X}_l$:

$$\boldsymbol{A}_{\mathcal{G}_l} = \boldsymbol{A}_{\mathcal{G}_l}\boldsymbol{A}_{\mathcal{V}^{-1}}\boldsymbol{P}. \tag{62}$$

Intuitively, whenever an edge is active in $\mathcal{G}_l$, the corresponding influence must be representable through the inverse entanglement mapping, $\boldsymbol{v}^{-1}$, at that state.

B.6.2. ALLOWED STRUCTURE OF V UNDER LOCAL GRAPHS.

In the global case, Boolean adjacency arithmetic shows that the set of indices allowed to be non-zero in row i of $\mathcal{V}^{-1}$ is:

$$\text{supp}(\mathcal{V}^{-1}_{i,:}) \subseteq \bigcap_{a \in \text{Ch}(i)} \text{Pa}(a). \tag{63}$$

Which yields the global graph criterion for disentanglement $\bigcap_{a \in \mathrm{Ch}(i)} \mathrm{Pa}(a) = \{i\} \forall i$. Note that the same applies to each and every partition:

$$\mathrm{supp}(\mathcal{V}^{-1}_{i,:}) \subseteq \bigcap_{a \in \mathrm{Ch}_{\mathcal{G}_l}(i)} \mathrm{Pa}_{\mathcal{G}_l}(a) \quad \forall l. \tag{64}$$

Since the representation is reused globally and throughout partitions, the allowed supports must satisfy **all** local constraints simultaneously. Hence, the allowed support upon the entanglement graph is given as:

$$\mathrm{supp}(\mathcal{V}^{-1}_{i,:}) \subseteq \bigcap_{l=1}^{b} \bigcap_{a \in \mathrm{Ch}_{\mathcal{G}_l}(i)} \mathrm{Pa}_{gG_l}(a) \tag{65}$$

Therefore, the local-graph graphical criterion guaranteeing disentanglement is:

$$\forall i : \bigcap_{l=1}^{b} \bigcap_{a \in \mathrm{Ch}_{\mathcal{G}_l}(i)} \mathrm{Pa}_{gG_l}(a) = \{i\}. \tag{66}$$

B.6.3. COROLLARIES AND EDGE CASES.

- **Monotonicity**. Adding more local graphs can only shrink the admissible set for each row of the entanglement graph, so incorporating additional partitions can never worsen disentanglement and may only help it.
- **Sufficient Subgraph**. If full latent parameter disentanglement holds in *any* one local graph, the representation must be entangled globally.
- **Insufficient Subgraph**. If full latent parameter disentanglement holds in *none* of the individual sub-graphs, or the global graph, disentanglement throughout the intersections is still feasible given the right structure.
- **Overspecification**. The number of dynamical parameters that can be disentangled can be significantly larger than the number of state parameters considered. Theroretically in the limit it can be infinite as long as all assumptions are still met, and the full set of dynamical parameters is inferable from a trajectory.
- **Nodes without children**. If $\mathrm{Ch}_{\mathcal{G}_l} = \emptyset$ in some sub-graph, the local constraint does not restrict row i. When evaluating the intersection, these should be ignored. In practice, it is easiest to replace such a case with the set of all nodes, as this best aligns with the allowed elements of the entanglement mapping. Suppose a node has no children in any subset of the global causal graph. In that case, the invertibility of dynamics is unmet, and the latents are not deterministically encodable from a trajectory. The theory does not apply in cases where the number of learnt parameters exceeds the minimal set of parameters required to describe a system.

Appendix C. An Example of Path Connectedness

Example.

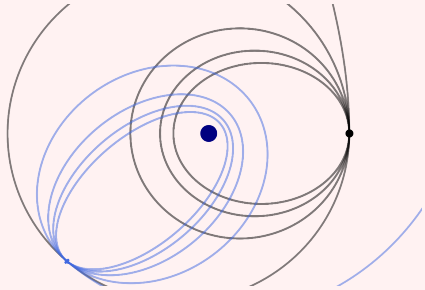


Figure 6: Illustration of how partition sets are defined using a planetary orbit example. Consider this system where the dynamical factor of variation is the (central) sun's mass. From an initial starting position and velocity, a large set of trajectories can unfold depending on the sun's mass. Throughout every point in black, the learnt representation must be independent of the system state. Each black line represents one trajectory that could have originated from the black point by varying the dynamical parameter. This process can be repeated iteratively at any point along any of the black lines to expand the set. In the figure, we choose the blue point from the set of black trajectories and, from the same starting position and velocity, change the dynamical parameter again. The result is a set of states, which can be expanded to cover a partition of the state-space.

Path connectivity uncovers trajectory-invariant symmetries in dynamical families. This means that the usefulness of a representation to our theorem is highly sensitive to the available state representation. Trajectory invariant symmetries can be challenging to spot. Consider the example from figure 6. It is evident that, given a broad enough set of allowable starting conditions and sun masses, the entire position space can be covered by one singular partition, as there exist trajectories between any two positions. However, if the state space were to include the planet's velocity, then two partitions would exist in the system. One for clockwise circular motion and one for counter-clockwise circular motion, since no dynamical parameter or state can transfer the system from one partition to another. Therefore, we expect any dynamical model trained on this system to learn up to two state-dependent distinct representations, one for clockwise circular motion and one for counter-clockwise circular motion. This also means that invariant or redundant system information cannot be part of the state space. For example, imagine a dynamical model predicting the behaviour of cubes on a table, except sometimes the cube is red and sometimes the cube is blue. If this property were encoded into the state space, it would be invariant along a trajectory, and any dynamical parameter can be a function of the cube colour.

Appendix D. Discussion

D.1. Comparison with the Recent Works Lachapelle et al. (2024) and Yao et al. (2024a)

The proof presented in Appendix B builds upon the proposed concepts by Yao et al. (2024a) and the proof technique of Lachapelle et al. (2024). Lachapelle et al. (2024) introduces *Mechanism Sparsity* a principle for nonlinear ICA that achieves disentanglement by assuming latent factors depend only on a small subset of past factors or auxiliary variables. They derive an identifiability theorem showing that latent variables can be recovered up to permutation and element-wise diffeomorphism if the ground-truth causal structure satisfies a graphical criterion. To implement this in practice, they propose a method that simultaneously learns latent representations and their underlying causal graph structure via L_1 regularisation. We adapt the framework of Lachapelle et al. (2024) to a novel setting, necessitating three key theoretical extensions: the incorporation of auxiliary variable representation learning, the introduction of state-dependent local-causal graphs, and the application of a path-connectedness assumption.

Learning the Representation of Auxiliary Variables. Prior works, such as (Lachapelle et al., 2024; Hyvarinen et al., 2019), are primarily focused on learning state representations while assuming auxiliary variables are directly observed. We show that in the framework of mechanism sparsity, these results can be extended to learning representations of auxiliary variables, which are of particular importance in settings of modelling dynamical systems, as dynamical parameters can be seen as auxiliary variables.

Path Connectedness Assumption. Switching to auxiliary representational learning does introduce additional challenges not addressed by Lachapelle et al. (2024) and not discussed in Yao et al. (2024a). The representation of parameters can become entangled with the representations of state. We formalise path-connectedness, an assumption that resolves this issue. To our knowledge, we are the first to implement such an assumption in the context of a mechanism sparsity identifiability theorem.

Local Graphs. Finally, we introduce the notion of local graphs, which are state-dependent subgraphs of a system's global graph. To the best of our knowledge, ours is the first work to establish a representational disentanglement identifiability theorem incorporating local graphs. Critically, we show that incorporating subgraphs must be equiperformant with global graph methods or improve upon them. We anticipate that the shift from global to local intersections is a generalizable insight applicable to other graphical criteria within the mechanism sparsity framework, although we leave the formal proof of this extension to future work.

Yao et al. (2024a) propose an isomorphism between Causal Representation Learning (CRL) and dynamical systems, arguing that integrating the two fields equips system identification with theoretical guarantees and improves the real-world applicability of CRL. Empirically, they rely on *Mechanistic Neural Networks* (Pervez et al., 2024) to model the parameters of an Ordinary Differential Equation (ODE) via differentiable solvers. This reliance on ODEs fundamentally distinguishes their domain from ours. While ODEs model continuous vector fields, our theorem allows for object-centric or factored representations that are not bound by global continuity. Consequently, our method is inherently capable of capturing sparse (both spatially and temporally), discontinuous interactions, such as collisions, that lie outside the standard expressivity of continuous ODE formulations.

Regarding identifiability, Yao et al. (2024a) presents two key propositions. The first, applicable when the functional form is known (see their Corollary 3.1), aligns with the framework of sparse identification methods such as SINDy (Brunton et al., 2016). By assuming a fixed dictionary of terms, this approach maps predicted parameters to pre-defined equation components; thus, identifiability arises from the imposed structural priors rather than through representation learning of latent variables.
Their second proposition (Corollary 3.2) addresses the case of unknown functional forms and relies on *multi-view* assumptions. They show that optimising an alignment loss across paired trajectories partially identifies shared parameters. In contrast, our work does not consider multi-view data or paired counterfactuals. Instead, we prove that *mechanism sparsity* is sufficient in some settings to disentangle parameters from raw trajectories.

D.2. Discussion on the Sufficient Influence Assumption

The sufficient variability assumption (alternatively termed sufficient influence, assumption 6) emerges as a theoretical necessity. In this section, we demonstrate that this condition possesses a clear, intuitive interpretation within our framework. We support this claim through illustrative examples and a comparative analysis of how analogous constraints are utilised in the broader causal representation learning literature (section D.2.1). The assumption is restated below for convenience.

Assumption 6 (Sufficient Variability of the Jacobian (Lachapelle et al., 2022b)) *There must exist at least one $\boldsymbol{\theta} \in \Theta$ such that there exists a set of states $\{\boldsymbol{x}_p\}_{p=1}^{\|\mathcal{G}\|_0}, \boldsymbol{x} \in \mathcal{X}$ such that*

$$\textit{span}\left\{\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}_p, \boldsymbol{\theta})\right\}_{p=1}^{\|\mathcal{G}\|_0} = \mathbb{R}^{\mathcal{G}} \tag{67}$$

The real graph subset is defined in definition 11.

In essence, the assumption requires that every causal edge exerts a unique influence that is separable from the influence of every other edge. Practically, this requires that the graph $\mathcal{G}$ and our model of the ground-truth environment $\boldsymbol{f}$ are irreducible. Consequently, the assumption is mild and additionally imposes that the assumed cardinality of the parameter space is correct.

In the setting of physical parameters, irreducibility can be expressed as a lack of redundancy. For example, given a system under the influence of two springs with strengths θ_0, and θ_1, as per equation 68.

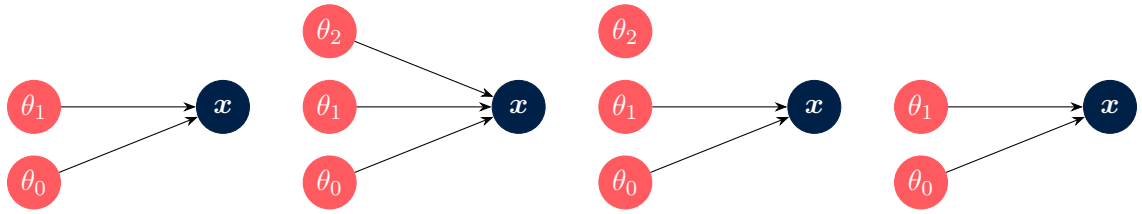
Example A. Redundant Causal Edge

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}) = -(\theta_0 + \theta_1)\boldsymbol{x}_t \tag{68}$$

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f} = \begin{bmatrix} -\boldsymbol{x}_t & -\boldsymbol{x}_t \end{bmatrix} \tag{69}$$

Equation 69 expresses the corresponding Jacobian of the transition model. As per our definition of causal graphs (non-zero Jacobian at any point), the graph is fully connected. The graph is depicted in figure 7A.

From the Jacobian, it is evident that the assumption can never be satisfied with this system. Intuitively, because both springs have the same effect upon the system, their combined influence can be aggregated into one parameter. Another implication is that, given any observed trajectory, the



(A) (B) (C) (D)

Figure 7: The causal DAGs pertaining to examples A,B,C and D.

ground-truth parameters are not uniquely recoverable. Satisfying the sufficient variability assumption by itself does not guarantee disambiguation of the influence of parameters given any specific trajectory, but it does guarantee that there exist states where the influence can be disambiguated.

Example B. Redundant Causal Edge

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}) = \frac{\theta_0}{\theta_1}\boldsymbol{x}_t + \theta_2 \tag{70}$$

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f} = \begin{bmatrix} \frac{1}{\theta_1}\boldsymbol{x}_t & -\frac{\theta_0}{{\theta_1}^2}\boldsymbol{x}_t & 1 \end{bmatrix} \tag{71}$$

Example B showcases the same effect in a slightly more complex setting. Many physical effects are functions of the ratios of natural variables. For example, elastic collisions are a function of the ratio of the masses of the colliding objects. In the same setting, using two parameters introduces linearly dependent columns, which violate the assumption.

Example C. Non-Redundant Causal Edge

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{\theta}) = \theta_0\boldsymbol{x}_t + \theta_1 \tag{72}$$

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f} = \begin{bmatrix} \boldsymbol{x}_t & 1 & 0 \end{bmatrix} \tag{73}$$

In contrast, if the redundant parameter is replaced with a single parameter, the first two columns become linearly separable. Note that in this new example, the Jacobian with respect to θ_2 is zero. Consequently, the parameter no longer exerts causal influence on $\boldsymbol{x}$ and the assumption no longer requires that the column is linearly independent.

D.2.1. SUFFICIENT INFLUENCE ASSUMPTION IN RELATED WORKS

A notion of sufficient influence (or variability) is a recurring requirement across causal representation learning (CRL) methods, as it is fundamental for resolving ambiguity between causal influences. While this assumption is intuitive in physical systems, where parameters exert identifiable effects on system states,its interpretation becomes less direct in more abstract representation learning settings.

In the temporal CRL setting of Lachapelle et al. (2024), the sufficient variability assumption is formalized as a requirement that the family of functions influencing each variable be linearly independent. In their framework, these functions correspond to components of the transition function, which closely parallels our setting, in which identifiability depends on the distinguishability of parameter-induced effects on state transitions. Linear dependence among these functions introduces redundancy, preventing parameters from being uniquely identified. Recent work on hierarchical temporal CRL (Li et al., 2025b), employs a similar modified sufficient variability assumption.

Similar principles appear outside of temporal CRL. In iVAE, (Khemakhem et al., 2020), identifiability is achieved by leveraging an observed auxiliary variable that modulates a latent distribution. There, sufficient influence is enforced by requiring independence between the sufficient statistics and the natural parameters of an assumed exponential family. As in our setting, this condition rules out redundant or indistinguishable latent effects that would otherwise break identifiability.

More broadly, related assumptions are commonplace across a wide range of CRL approaches. In interventional CRL, for example, methods typically require interventions to induce sufficiently distinct changes in the latent variables. The work of von Kügelgen et al. (2023) formally assumes that interventions is sufficiently distinguishable such that only the correct causal graph can explain a set of observations. CITRIS (Lippe et al., 2022) requires the independence of intervention targets, preventing scenarios in which it is ambiguous which intervention produced a given outcome. Although expressed differently, these conditions serve a common purpose, preventing ambiguity in causal influence.

Appendix E. Dataset Details

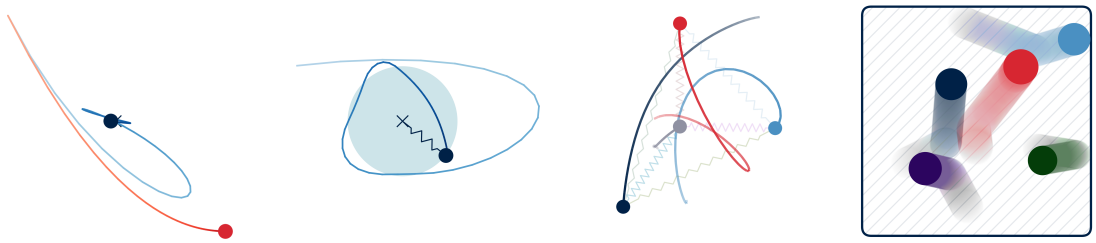


Figure 8: Graphical representation of the four evaluation environments, in order from left to right. *Left:* Dual Particle. *Center-Left:* Local Particle. *Center-Right:* Springs. *Right:* Bounce.

DUAL PARTICLE

Consider a point-mass that is connected to the origin by a spring (controlled by dynamical parameter, k) and under the influence of separate x-axis and y-axis damping (parameters c_x, c_y). The state information is taken as a four-dimensional vector containing position and velocity information. As per the graphical criterion derived in section 2, the dynamical factors are not disentangle-able using the system's causal graph. To achieve disentanglement, we modify the system to perform two counterfactual tasks: general prediction, and then counterfactually what would have happened if the spring were disabled. The resultant DAG is visualised in figure 2 and satisfies the graphical criterion globally. This environment also showcases an interesting edge case in the theory. The path-connectedness condition is not technically satisfied. Consider the environment as visualised in Figure 8. The red particle represents the counterfactual trajectory with no spring connected. The entire state-space can be partitioned into four, depending upon the sign of the x and y velocity of the non-spring-connected particle. The path-connectedness assumption is violated, as there exists no position in state space or no (positive) damping parameters that would change the sign of the velocity components of the red particle. The assumption being violated is not catastrophic in this scenario, as it only affects the damping parameters, which the blue particle must also reuse, and those are well-behaved relative to the path-connectedness assumption. The result is a representation which is surjective, simply many-to-one. Within the set of experiments, this instantiates itself as a V or inverted-V learnt representation. In practice, we also regularise the state-to-state causal DAG, which isolates the dynamical behaviour of each particle, and forces the learnt representation to be diffeomorphism again.

LOCAL PARTICLE

The graphical criterion can also be met locally, with state-dependent local causal graphs. The local particle environment implements the same physics as the multi-task particle environment, but disables the spring within a ball around the origin. Consequently, locally within the ball the causal graph is the same as the counterfactual graph of the multi-task environment, and outside the same as the primary task. Interestingly, **neither** local graph is sufficient to disentangle all parameters by itself, but when used in conjunction they satisfy the local graphical criterion.

SPRINGS

The springs environment contains four-point particles connected by six springs, one for every possible pairing. The causal connections are quantised at the per-object level, meaning each token input to SPARTAN corresponds to an object, not a state dimension. The spring constants are distributed via the absolute value of a standard normal variable, but there is only a seventy per cent probability of a given spring being active at all. The global causal graph for the environment is consequently heavily populated, and can be seen in figure 2. This environment is an example in which there are more dynamic factors of variation than direct observables, yet the theory guarantees disentanglement with respect to the global graph. Variations of this environment, usually with static spring constants, are commonplace across works in causal discovery and representation learning (Löwe et al., 2022; Kipf et al., 2018; Battaglia et al., 2016; Li et al., 2020).

BOUNCE

The bounce environment consists of five round balls constrained in a square box, observable through object-centric representations. Each ball can collide elastically with another ball and with the wall. The characteristics of each collision are governed by the mass of each object, which is the dynamical factor of variation, sampled from a non-zero mean normal distribution. For ease of observation, the radius of each ball is also proportional to its mass, creating a dynamic environment that is sensitive to its mass. As with the spring environment, variations of this environment are prevalent testing grounds for works within compositional world models, causal discovery, and representation learning (Lei et al., 2025; Battaglia et al., 2016; Chang et al., 2017; Jiang et al., 2024), and were initially introduced in video form by Van Steenkiste et al. (2018). The environment is much more difficult to learn as the causal graphs occur discontinuously only upon collision events. The global graph for this environment is fully connected and provides zero disentanglement guarantees; however, as visualised in figure 2, there exists a combinatorial cardinality of local causal graphs. Only a small subset of them is required to prove the theory guarantees disentanglement.

Appendix F. Extended Results

F.1. MCC Metric for Disentanglement

The MCC disentanglement metric is computed by encoding all trajectories in a validation dataset into the learnt parameters, $\hat{\boldsymbol{\theta}}$, and fitting a small MLP onto every combination of marginalised predictions and marginalised ground truth parameters, $\theta_i \approx \text{MLP}_{ij}(\hat{\theta}_j)$. The predictions from this collection of MLPs are used to form a matrix of non-linear correlation coefficients, $R^2 \in \mathbb{R}^{I \times J}$. The MCC metric is calculated as $\text{MCC} = \frac{1}{I}\sum_i^I \max_j(R^2_{i,j})$. The MCC metric is permutation, scale, and shift invariant in representations and accounts for any non-linearities. It does not determine whether a representation is a diffeomorphism, but it does enforce surjectivity. Consequently, an MCC of ≈ 1 indicates that there exists at least one element in the learnt representation that perfectly specifies, functionally up to continuous, and differentiable surjection, every element in the ground-truth representation. The use of MCC to measure disentanglement is commonplace in causal representation learning literature (Lachapelle et al., 2022a; Yao et al., 2022; Lachapelle and Lacoste-Julien, 2022; Li et al., 2019, 2024; Gresele et al., 2021; Khemakhem et al., 2020).

The metric is an approximation as exact values depend upon the size and convergence of the MLP used. If not enough data points are used to calculate the mapping, overfitting might also skew

the results. Consequently, we opt to use a one-hidden-layer MLP with a hidden dimension of 32 and 5,000 sampled training points, as well as cross-validation with a 10% and 90% split to prevent overfitting and ensure that the model converges quickly.

F.2. Examples of Learnt Representations

To illustrate the learnt representations, the empirical entanglement mapping can be plotted. We encode over a validation dataset of trajectories all trajectories into pairs of latent vectors and ground-truth dynamical factor vectors. Then in a grid by plotting the marginals we can represent the multidimensional mapping in a way that visually shows disentanglement. On the y axis we depict the ground truth parameters and on the x axis the learnt parameters. Surjective disentanglement (as measured by MCC) is evident when specifying one learnt parameter allows for a precise specification of a ground truth factor of variation. Each diagram is a representative sample from one of the trials in the main results.

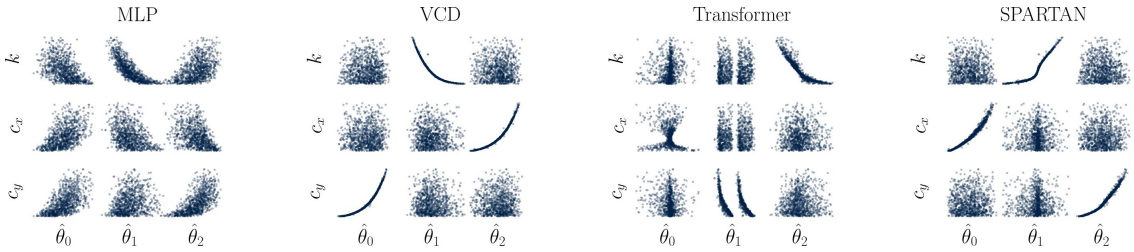


Figure 9: Representative samples of representations learnt on the Dual Particle environment. Each subplot plots the marginal between one ground-truth factor of variation on the vertical axis against a learnt parameter dimension on the horizontal axis. Increasingly sharp one-to-one mappings show disentanglement. The MLP learns a representation where parameters are correlated with the factors of variation but not disentangled. The Transformer learns a non-bijective distribution in which only the spring constant, k, exhibits clear diffeomorphic disentanglement. The multi-form nature of the Transformers distribution indicates entanglement with the state. Both SPARTAN and VCD show clear, sharp, and crisp disentanglement.

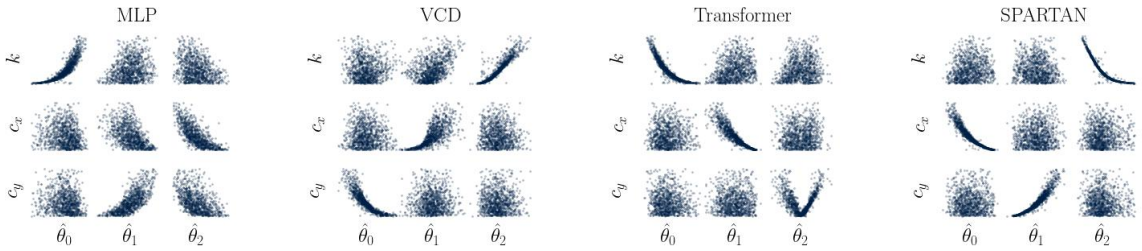


Figure 10: Representative samples of representations learnt on the Local Particle Environment. Each subplot plots the marginal between one ground-truth factor of variation on the vertical axis against a learnt parameter dimension on the horizontal axis. Increasingly sharp one-to-one mappings show disentanglement. Both the MLP and VCD baselines learn representations that are correlated with but not fully disentangled. In this instance, the global graphical criterion predicts partial disentanglement; the spring constant k is guaranteed to be disentangled, but not the damping constants. The VCD model strongly disentangles k and c_y, but c_x remains entangled. SPARTAN learns a diffeomorphism mapping as predicted by the theory, whilst a Transformer learns a surjective mapping. Both perform well on the MCC metric.

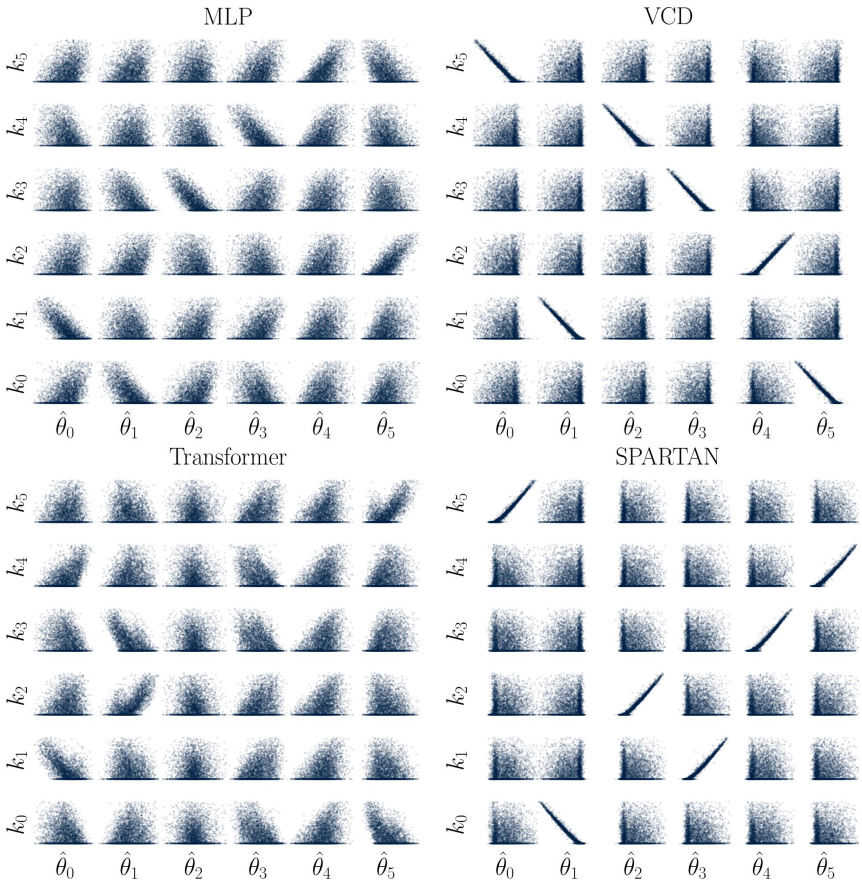


Figure 11: Representative samples of representations learnt in the Springs environment. Each subplot plots the marginal between one ground-truth factor of variation on the vertical axis against a learnt parameter dimension on the horizontal axis. Increasingly sharp diagonals in each subplot showcase disentanglement. VCD and SPARTAN both learn highly disentangled representations with evident structure that is missing from the MLP and Transformer baselines.

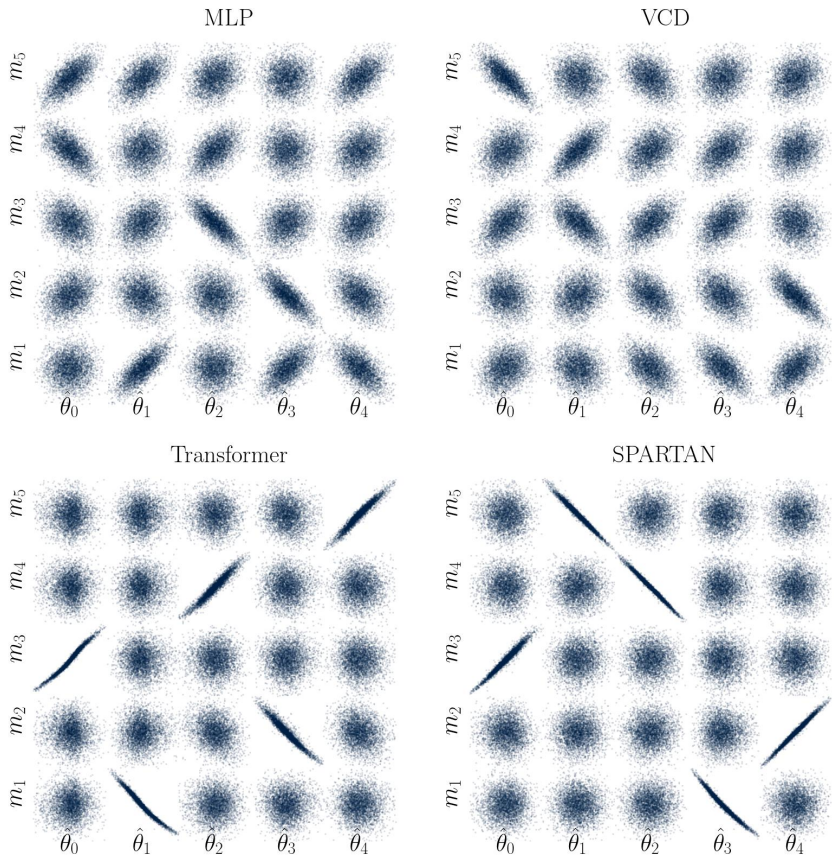


Figure 12: Representative samples of representations learnt in the Bounce environment. Each subplot plots the marginal between one ground-truth factor of variation on the vertical axis against a learnt parameter dimension on the horizontal axis. Increasingly sharp diagonals indicate disentanglement. As this environment requires adherence to local-causal graphs for disentanglement, both the MLP and VCD baselines fail to disentangle. Both SPARTAN and a Transformer selectively attend to parameters, and in this instance, the Transformer's softer bias is sufficient for both to disentangle strongly.

F.3. Examples of Learnt Causal Graphs

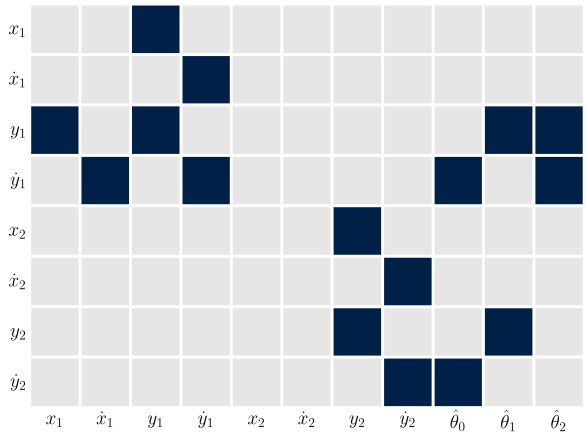


Figure 13: Representative *global* causal graph learnt by the VCD architecture on the dual particle environment. State space tokenisation is used. The parameter influence graph (referred to as $\hat{\mathcal{G}}$ in the theory) is represented by the final three columns. The sparse representation satisfies the global graph criterion for disentanglement. Dark blue squares represent an edge probability of approximately 1, light grey squares an edge probability of approximately 0, and colours in between represent values in between.

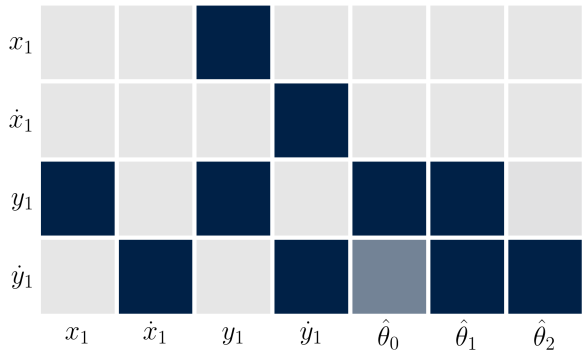


Figure 14: Representative *global* causal graph learnt by the VCD architecture on the local particle environment. State space tokenisation is used. The parameter influence graph (referred to as $\hat{\mathcal{G}}$ in the theory) is represented by the final three columns. In contrast to the local-graphs learnt by SPARTAN, the global graph displayed does not satisfy the disentanglement condition. Dark blue squares represent an edge probability of approximately 1, light grey squares an edge probability of approximately 0, and colours in between represent values in between.

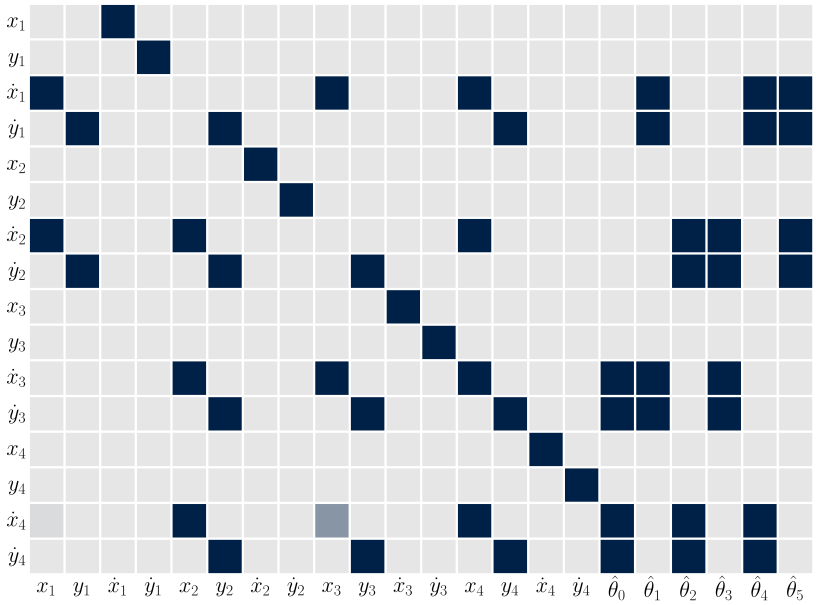


Figure 15: Representative *global* causal graph learnt by the VCD architecture on the springs environment. State space tokenisation is used. The parameter influence graph (referred to as $\hat{\mathcal{G}}$ in the theory) is represented by the final six columns. The sparse representation satisfies the global graph criterion for disentanglement. Dark blue squares represent an edge probability of approximately 1, light grey squares an edge probability of approximately 0, and colours in between represent values in between.

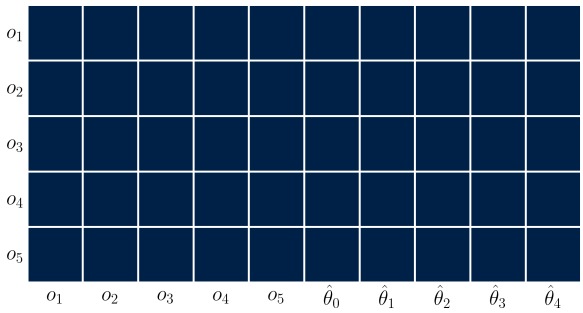


Figure 16: Representative *global* causal graph learnt by the VCD architecture on the bounce environment. Object-level tokenisation is used. The parameter influence graph (referred to as $\hat{\mathcal{G}}$ in the theory) is represented by the final five columns. The graph is fully connected, as any pair of objects and any set of parameters *can* interact at some point in the dataset, even though interactions are very sparse in practice. The *local* sparsity is taken into account by the SPARTAN baseline. Dark blue squares represent an edge probability of approximately 1, light grey squares an edge probability of approximately 0, and colours in between represent values in between.

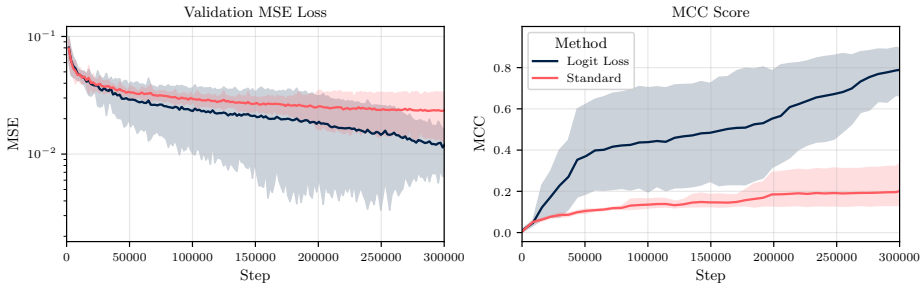


Figure 17: Comparison of the impact that the attention logit regularisation loss has on training a transformer in the bounce environment. Both lines consists of 8 trials completed with different random seeds, and the shaded region represents plus-minus one standard deviation.

F.4. Ablation of the Logit Regularisation

We demonstrate the necessity of attention logit regularisation through an ablation study within the bounce environment, comparing performance with and without across eight random seeds. Because the encoder is bootstrapped to the decoder, its capacity to encode dynamical behaviour emerges only after the decoder has adequately modelled the system and identified the relevant factors of variation. This results in a two stage information adoption process akin to the dual optimisation procedure used with the SPARTAN and VCD experiments. Without the attention logit regularisation, the decoder struggles to incorporate the new information. We conjecture that this is due to vanishing gradients in the softmax of the transformer. The effect is visualised in figure 17, which shows that the transformer can only disentangle in this setting when introducing the logit loss. We observed in experimental settings that undertuning or removing the logit loss hyperparameter, λ_{logit}, in the SPARTAN experimental runs results in the path loss plateauing during training, and consequently, the model does not sparsify sufficiently for representational disentanglement.

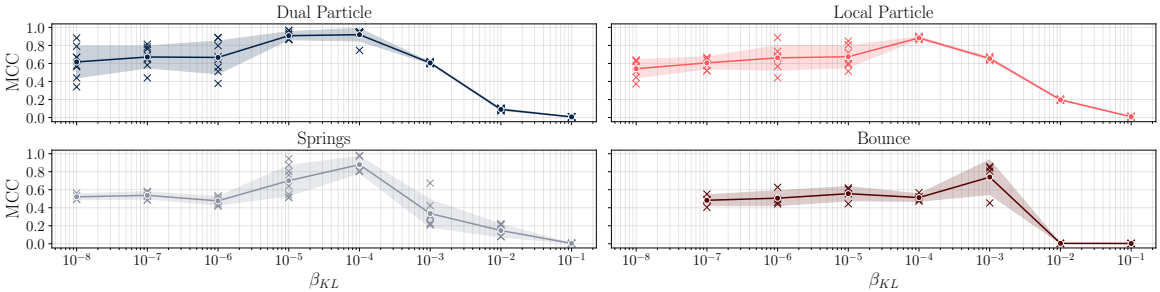


Figure 18: Comparison of the disentanglement achieved by tuning the weighting on the KL regularisation and using an MLP decoder. Each data point represents one trial over a random seed, and the lines represent the mean over all trials at that value. The shaded region covers plus-minus one standard deviation.

F.5. Disentanglement Induced by the ELBO Loss.

During all experiments, the weighting on the KL regularisation β_{KL} is set to 10^{-6}. This is to ensure that latent regularisation does not contribute to the disentanglement of factors, an issue widely studied in the non-linear ICA and representation learning literature (Burgess et al., 2018; Kim and Mnih, 2018; Khemakhem et al., 2020; Chen et al., 2018; Esmaeili et al., 2019; Mathieu et al., 2019). The impact of varying the KL regularisation parameter is shown in figure 18.