Marrying Causal Representation Learning with Dynamical Systems for Science

Anonymous Authors[1]

Abstract

Causal representation learning promises to extend causal models to hidden causal variables from raw entangled measurements. However, most progress has focused on proving identifiability results in different settings, and we are not aware of any successful real-world application. At the same time, the field of dynamical systems benefited from deep learning and scaled to countless applications but does not allow parameter identification. In this paper, we draw a clear connection between the two and their key assumptions, allowing us to apply identifiable methods developed in causal representation learning to dynamical systems. At the same time, we can leverage scalable differentiable solvers developed for differential equations to build models that are both identifiable and practical. Overall, we learn explicitly controllable models that isolate the trajectory-specific parameters for further downstream tasks such as out-of-distribution classification or treatment effect estimation. We experiment with a wind simulator with partially known factors of variation. We also apply the resulting model to real-world climate data and successfully answer downstream causal questions in line with existing literature on climate change.

1. Introduction

Causal representation learning (CRL) (Schölkopf et al., 2021) focuses on *provably* retrieving high-level latent variables from low-level data. Recently, there have been many casual representation learning works compiling, in various settings, different theoretical identifiability results for these latent variables (Brehmer et al., 2022; Kivva et al., 2022; Lachapelle et al., 2024; Lippe et al., 2022a;b; Squires et al., 2023; Sturma et al., 2024; Varici et al., 2023; Von Kügelgen et al., 2021; von Kügelgen et al., 2024; Xu et al., 2024; Zhang et al., 2024). The main open challenge that remains for this line of work is the broad applicability to real-world data. Following earlier works in disentangled representations (see (Locatello et al., 2019) for a summary of data sets), existing approaches have largely focused on visual data . This is challenging for various reasons. Most notably, it is unclear what the causal variables should be in computer vision problems and what would be interesting or relevant causal questions. The current standard is to test algorithms on synthetic data sets with "made-up" latent causal graphs, e.g., with the object class of a rendered 3d shape causing its position, hue, and rotation (Von Kügelgen et al., 2021).

In parallel, the field of machine learning for science (Mjolsness and DeCoste, 2001; Raghu and Schmidt, 2020) shows promising results on various real-world time series data collected from some underlying dynamical systems. Some of these works primarily focus on time-series forecasting, i.e., building a neural emulator that mimics the behavior of the given times series data (Chen et al., 2018; 2021; Kidger et al., 2021); while others try to additionally learn an explicit ordinary differential equation simultaneously (Brunton et al., 2016a;b; d'Ascoli et al., 2024; d'Ascoli et al., 2022; Kaheman et al., 2020; Schröder and Macke, 2023). However, to the best of our knowledge, none of these methods provide explicit identifiability analysis indicating whether the discovered equation recovers the ground truth underlying governing process given time series observations; or even whether the learned representation relates to the underlying steering parameters. At the same time, many scientific questions are inherently causal, in the sense that physical laws govern the measurements of all the natural data we can record, e.g., across different environments and experimental settings. Identifying such an underlying physical process can boost scientific understanding and reasoning in numerous fields; for example, in climate science, one could conduct sensitivity analysis of *layer thickness* parameter on atmosphere motion more efficiently, given a neural emulator that identifies the *layer thickness* in its latent space. However, whether mechanistic models can be practically identified from data is so far unclear (Schölkopf et al., 2021, Table 1).

This paper aims to identify the underlying *time-invariant* physical parameters from real-world time series, such as the previously mentioned *layer thickness* parameter, while

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at ICML 2024 AI for Science workshop. Do not distribute.

still preserving the ability to forecast efficiently. Thus, we connect the two seemingly faraway communities, causal representation learning and machine learning for dynamical systems, by phrasing parameter estimation problems in dynamical systems as a latent variable identification problem in CRL. The benefits are two folds: (1) we can import all identifiability theories for free from causal representation learning works, extending discovery methods with additional identifiability analysis and, e.g., multiview training constructs; (2) we showcase that the scalable mechanistic neural networks (Pervez et al., 2024) recently developed for dynamical systems can be directly employed with causal representation learning, thus providing a scalable implementation for both identifying and forecasting real-world dynamical systems.

Starting by comparing the common assumptions in the field of parameter estimation in dynamical systems and causal representation learning, we carefully justify our proposal to translate any parameter estimation problem into a latent variable identification problem; we differentiate three types of identifiability: *full identifiability*, *partial identifiability* and *non-identifiability*. We describe concrete scenarios in dynamical systems where each kind of identifiability can be theoretically guaranteed and restate *exemplary* identifiability theorems from the causal representation learning literature with slight adaptation towards the dynamical system setup. We provide a step-by-step recipe for reformulating a parameter estimation problem into a causal representation learning problem and discuss the challenges and pitfalls in practice. Lastly, we successfully evaluate our parameter identification framework on various *simulated* and *real-world* climate data. We highlight the following contributions:

- We establish the connection between causal representation learning and parameter estimation for differential equations by pinpointing the alignment of common assumptions between two communities and providing hands-on guidance on how to rephrase the parameter estimation problem as a latent variable identification problem in causal representation learning.

- We equip discovery methods with provably identifiable parameter estimation approaches from the causal representation learning literature and their specific training constructs. This enables us to maintain both the theoretical results from the latter and the scalability of the former.

- We successfully apply causal representation learning approaches to simulated and real-world climate data, demonstrating identifiability via domain-specific downstream causal tasks (OOD classification and treatment-effect estimation), pushing one step further on the applicability of causal representation for real-world problems.

Remark on the novelty of the paper: Our main contribution is establishing a connection between the dynamical systems and causal representation learning fields. As such, we do not introduce a new method per se. Meanwhile, this connection allows us to introduce CRL training constructs in methods that otherwise would not have any identification guarantees. Further, it provides the first avenue for causal representation learning applications on real-world data. These are both major challenges in the respective communities, and we hope this paper will serve as a building block for cross-pollination.

2. Parameter Estimation in Dynamical Systems

We consider dynamical systems in the form of

$$\dot{\mathbf{x}}(t) = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t)) \qquad \mathbf{x}(0) = \mathbf{x}_0,\ \boldsymbol{\theta} \sim p_{\boldsymbol{\theta}},\ t \in [0, t_{\max}] \tag{1}$$

where $\mathbf{x}(t) \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the state of a system at time t, $f_{\boldsymbol{\theta}} \in \mathcal{C}^1(\mathcal{X}, \mathcal{X})$ is some smooth differentiable vector field representing the constraints that define the system's evolution, characterized by a set of physical parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_N$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^N$ is an open, simply connected real space associated with the probability density $p_{\boldsymbol{\theta}}$. Formally, $f_{\boldsymbol{\theta}}$ can be considered as a functional mapped from $\boldsymbol{\theta}$ through $M : \boldsymbol{\Theta} \to \mathcal{C}^1(\mathcal{X}, \mathcal{X})$. In our setup, we consider ***time-invariant***, ***trajectory-specific*** parameters $\boldsymbol{\theta}$ that remain constant for the whole time span $[0, t_{\max}]$, but variable for different trajectories. For instance, consider a robot arm interacting with multiple objects of different mass; a parameter $\boldsymbol{\theta}$ could be the object's masses $m \in \mathbb{R}_+$ in Newton's second law $\ddot{x}(t) = \mathcal{F}(t)/m$, with $\mathcal{F}(t)$ denote the force applied at time t. Depending on the object the robot arm interacts with, m can take different values, following the prior distribution $p_{\boldsymbol{\theta}}$. $\mathbf{x}(0) = \mathbf{x}_0 \in \mathcal{X}$ denotes the initial value of the system. Note that higher-order ordinary differential equations can always be rephrased as a first-order ODE. For example, a ν-th order ODE in the following form:

$$x^{(\nu)}(t) = f = (x(t), x^{(1)}(t), \dots, x^{(\nu-1)}(t), \boldsymbol{\theta}),$$

can be written as $\dot{\mathbf{x}}(t) = f_{\boldsymbol{\theta}}(\mathbf{x}(t))$, where $\mathbf{x}(t) = (x(t), x^{(1)}(t), \dots, x^{(\nu-1)}(t)) \in \mathbb{R}^{\nu \cdot d}$ denotes state vector constructed by concatenating the derivatives. Formally, the solution of such a dynamical system can be obtained by integrating the vector field over time: $\mathbf{x}(t) = \int_0^t f(\mathbf{x}(\tau), \boldsymbol{\theta}) d\tau$.

What do we mean by "parameters"? The parameters $\boldsymbol{\theta}$ that we consider can be both explicit and implicit. When the functional form of the ODE is given, like Newton's second law, the set of parameters is defined explicitly and uniquely. For real-world physical processes where the functional form of the state evolution is unknown, such as the sea-surface temperature change, we can consider *latitude-related* features as parameters. Overall, we use *parameters* to generally refer to any *time-invariant*, *trajectory-specific* components of the underlying dynamical system.

Assumption 2.1 (Existence and uniqueness)**.** For every $\mathbf{x}_0 \in \mathcal{X}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists a unique continuous solution $\mathbf{x}_{\boldsymbol{\theta}} : [0, t_{\max}] \to \mathcal{X}$ satisfying the ODE (eq. (1)) for all $t \in [0, t_{\max}]$ (Ince, 1956; Lindelöf, 1894).

Assumption 2.2 (Structural identifiability)**.** An ODE (eq. (1)) is *structurally* identifiable in the sense that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, $\mathbf{x}_{\boldsymbol{\theta}_1}(t) = \mathbf{x}_{\boldsymbol{\theta}_2}(t)\, \forall t \in [0, t_{\max}]$ holds if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ (Bellman and Åström, 1970; Walter et al., 1997; Wieland et al., 2021).

Remark 2.1. Asm. 2.2 implies that it is *in principle* possible to identify the parameter $\boldsymbol{\theta}$ from a trajectory $\mathbf{x}_{\boldsymbol{\theta}}$ (Miao et al., 2011). Since this work focuses on providing concrete algorithms that guarantee parameter identifiability *given infinite number of samples*, the structural identifiability assumption is essential as a theoretical ground for further algorithmic analysis. It is noteworthy that a non-structurally identifiable system can become identifiable by reparamatization. For example, linear ODE $\dot{\mathbf{x}}(t) = ab\mathbf{x}(t)$ with parameters $a, b \in \mathbb{R}^2$ is structurally non-identifiable as a, b are commutative. But if we define $c := ab$ as the overall growth rate of the linear system, then c is structurally identifiable.

Problem setting. Given an observed trajectory $\mathbf{x} := (\mathbf{x}_{\boldsymbol{\theta}}(t_0), \dots, \mathbf{x}_{\boldsymbol{\theta}}(t_T)) \in \mathcal{X}^T$ over the discretized time grid $\mathcal{T} := (t_0, \dots, t_T)$, our goal is to investigate the identifiability of structurally identifiable parameters by formulating concrete conditions under which the parameter $\boldsymbol{\theta}$ is (i) fully identifiable, (ii) partially identifiable, or (iii) non-identifiable *from the observational data*. We establish the identifiability theory for dynamical systems by converting classical parameter estimation problems (Bellman and Åström, 1970) into a latent variable identification problem in causal representation learning (Schölkopf et al., 2021). For both (i) and (ii), we empirically showcase that existing CRL algorithms with slight adaptation can successfully (*partially*) identify the underlying physical parameters.

3. Identifiability of Dynamical Systems

This section provides different types of theoretical statements on the identifiability of the underlying *time-invariant*, *trajectory-specific* physical parameters $\boldsymbol{\theta}$, depending on whether the functional form of $f_{\boldsymbol{\theta}}$ is known or not. We show that the parameters from an ODE with a known functional form can be *fully identified* while parameters from unknown ODEs are in general *non-identifiable*. However, by incorporating some weak form of supervision, such as multiple similar trajectories generated from certain overlapping parameters (Daunhawer et al., 2023; Locatello et al., 2020; Von Kügelgen et al., 2021; Yao et al., 2024), parameters from an unknown ODE can also be *partially identified*. Detailed proofs of the theoretical statements are provided in App. A.

3.1. Identifiability of dynamical systems with known functional form

We begin with the identifiability analysis of the physical parameters of an ODE with **known** functional form. Many real-world data we record are governed by known physical laws. For example, the bacteria growth in microbiology could be modeled with a simple logistic equation under certain conditions, where the parameter of interest in this case would be the *growth rate* $r \in \mathbb{R}_+$ and *maximum capacity* $K \in \mathbb{R}_+$. Identifying such parameters would be helpful for downstream analysis. To this end, we introduce the definition of *full identifiability* of a physical parameter vector $\boldsymbol{\theta}$.

Definition 3.1 (Full identifiability)**.** A parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is fully identified if the estimator $\hat{\boldsymbol{\theta}}$ converges to the ground truth parameter $\boldsymbol{\theta}$ almost surely.

Definition 3.2 (ODE solver)**.** An ODE solver $F : \boldsymbol{\Theta} \to \mathcal{X}^T$ computes the solution $\mathbf{x}$ of the ODE $f_{\boldsymbol{\theta}} = M(\boldsymbol{\theta})$ (eq. (1)) over a discrete time grid $T = (t_1, \dots, t_T)$.

Corollary 3.1 (Full identifiability with known functional form)**.** *Consider a trajectory $\mathbf{x} \in \mathcal{X}^T$ generated from a ODE $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$ satisfying Asms. 2.1 and 2.2, let $\hat{\boldsymbol{\theta}}$ be an estimator minimizing the following objective:*

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \left\| F(\hat{\boldsymbol{\theta}}) - \mathbf{x} \right\|_2^2 \tag{2}$$

*then the parameter $\boldsymbol{\theta}$ is **fully-identified** (Defn. 3.1) by the estimator $\hat{\boldsymbol{\theta}}$.*

Remark 3.1. The estimator $\hat{\theta}$ of eq. (2) is considered as some learnable parameters that can be directly optimized. If we have multiple trajectories $\mathbf{x}$ generated from different realizations of $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}$, we can also amortize the prediction $\hat{\boldsymbol{\theta}}$ using a smooth encoder $g : \mathcal{X}^T \to \boldsymbol{\Theta}$. In this case, the loss above can be rewritten as: $\mathcal{L}(g) = \mathbb{E}_{\mathbf{x},t}[||F(g(\mathbf{x})) - \mathbf{x}(t)\|_2^2]$, then the optimal encoder $g^* \in \operatorname{argmin} \mathcal{L}(g)$ can generalize to unseen trajectories $\mathbf{x}$ that follow the same class of physical law f and fully identify their trajectory-specific parameters $\boldsymbol{\theta}$.

Remark 3.2. In Cor. 3.1, we consider an ideal setup glossing over several practical challenges: (i) Although closed-form solution of $\boldsymbol{\theta}^*$ is provided by linear least squares when f is linear in $\boldsymbol{\theta}$ (see App. A.1 for details), finding the global optimum $\boldsymbol{\theta}^*$ in the nonlinear case using gradient descent is challenging in practice, both computationally and, despite the guarantee of theoretical full identifiability, it ignores non-convexity. (ii) Since the functional form $f_{\boldsymbol{\theta}}$ is known, we assume that the ODE solver is *exact* in the sense that the generated solution of the ground truth parameter $F(\boldsymbol{\theta})$ perfectly aligns with the observation $\mathbf{x}$, i.e., $\mathcal{L}(\boldsymbol{\theta}) = 0$.

However, in practice, numerical solvers preserve certain approximation errors (Lötstedt and Petzold, 1986). Although recent advances propose neural network-based ODE solvers (Chen et al., 2018) to alleviate this issue, end-to-end training that involves solving an ODE in the forward pass is not trivial. Most of the differentiable ODE solvers (Chen, 2018; Chen et al., 2018; 2021) solve the ODE autoregressively; thus, the time dimension cannot be parallelized in the GPU. To tackle this problem, Pervez et al. (2024) provided a highly efficient ODE solver that can be utilized in our framework. A more extensive discussion about different types of neural network-based solvers is provided in § 5.

Discussion. Many works on machine learning for dynamical system identification follow the principle presented in Cor. 3.1, and most of them solely differ concerning the architecture they choose for the ODE solver. For example, SINDy-like ODE discovery methods (Brunton et al., 2016a;b; Kaheman et al., 2020; Kaptanoglu et al., 2021; Pervez et al., 2024) approximate the ground truth vector field f using a linear weighted sum over a set of library functions and learn the linear coefficients by sparse regression. For any ODE f that is linear in $\boldsymbol{\theta}$, i.e., the ground truth vector field is in the form of $f_{\boldsymbol{\theta}}(\mathbf{x}, t) = \sum_{i=1}^{m} \theta_i \phi_i(\mathbf{x})$ for a set of known base functions $\{\phi_i\}_{i\in[m]}$, SINDy-like approaches can fully identify the parameters by imposing some sparsity constraint. Another line of work, gradient matching (Wenk et al., 2019), estimates the parameters probabilistically by modeling the vector field $f_{\boldsymbol{\theta}}$ using a Gaussian Process (GP). The modeled solution $\mathbf{x}(t)$ is thus also a GP since GP is closed under integrals (a linear operator). Given the functional form of $f_{\boldsymbol{\theta}}$, the model aims to match the estimated gradient $\dot{\mathbf{x}}$ and the evaluated vector field $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$ by maximizing the likelihood, which is equivalent to minimizing the least-squares loss (eq. (2)) under Gaussianity assumptions. Hence, the gradient matching approaches can *theoretically* identify the underlying parameters under Cor. 3.1. Formal statements and proofs for both SINDy-like and gradient matching approaches are provided in App. A. *Note that most ODE discovery approaches (Brunton et al., 2016a;b; Kaheman et al., 2020; Kaptanoglu et al., 2021; Pervez et al., 2024; Wenk et al., 2019) refrain from making identifiability statements and explicitly states it is unknown which settings yield identifiability.*

3.2. Identifiability of dynamical systems without known functional form

In traditional dynamical systems, identifiability analysis usually assumes the functional form of the ODE is known (Miao et al., 2011); however, for most real-world time series data, the functional form of underlying physical laws remains uncovered. Machine learning-based approaches for dynamical systems work in a black-box manner and can clone the behavior of an unknown system (Chen et al., 2018; 2021; Norcliffe et al., 2020), but understanding and identifiability guarantees of the learned parameters are so far missing. Since most of the physical processes are inherently steered by a few underlying *time-invaraint* parameters, identifying these parameters can be helpful in answering downstream scientific questions. For example, identifying climate zone-related parameters from sea surface temperature data could improve understanding of climate change because the impact of climate change significantly differs in polar and tropical regions. Hence, we aim to provide identifiability analysis for the underlying parameters of an unknown dynamical system by converting the classical parameter estimation problem of dynamical systems into a latent variable identification problem in causal representation learning. We start by listing the common assumptions in CRL and comparing the ground assumptions between these two fields.

Assumption 3.1 (Determinism)**.** The data generation process is deterministic in the sense that observation $\mathbf{x}$ is generated from some latent vector $\boldsymbol{\theta}$ using a deterministic solver F (Defn. 3.2).

Assumption 3.2 (Injectivity)**.** For each observation $\mathbf{x}$, there is only one corresponding latent vector $\boldsymbol{\theta}$, i.e., the ODE solve function F (Defn. 3.2) is injective in $\boldsymbol{\theta}$.

Assumption 3.3 (Continuity and full support)**.** $p_{\boldsymbol{\theta}}$ is smooth and continuous on $\boldsymbol{\Theta}$ with $p_{\boldsymbol{\theta}} > 0$ a.e.

Assumption justification. We observe strong alignment between the ground assumptions in CRL and system identification (Tab. 2) that justifies our idea of employing causal representation learning methods in parameter estimation problems for dynamical systems: (1) Asm. 2.1 implies that given a fixed initial value $\mathbf{x}_0 \in \mathcal{X}$, there exists a unique solution $\mathbf{x}(t)$, $t \in [0, t_{\max}]$ for any $f_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In other words, parameter domain $\boldsymbol{\Theta}$ is fully supported (Asm. 3.3), and these ODE solving processes from $F(\boldsymbol{\theta})$ (Defn. 3.2) are deterministic, which aligns with the standard Asm. 3.1 in CRL. Since the ODE solution $F(\boldsymbol{\theta})$ (§ 2) is continuous by definition, the continuity assumption from CRL (Asm. 3.3) is also fulfilled. (2) Asm. 2.2 emphasizes that each trajectory $\mathbf{x}$ can only be uniquely generated from one parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which means the generating process F (Defn. 3.2) is injective in $\boldsymbol{\theta}$ (Asm. 3.2).

Next, we reformulate the parameter estimation problem in the language of causal representation learning. We first cast the generative process of the dynamical system $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$ as a latent variable model by considering the underlying physical parameters $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}$ as a set of *latent variables*. Given a trajectory $\mathbf{x}$ generated by a set of underlying factors $\boldsymbol{\theta}$ based on the vector field $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$, we consider the observed trajectory as some *unknown nonlinear* mixing of the underlying $\boldsymbol{\theta}$, with the mixing process specified by individual vector field $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$. This interpretation of observations aligns with the standard setup of causal representation learning; for instance, high-dimensional images are usually generated from some lower-dimensional latent generating factors through an unknown nonlinear process. Thus, estimating the parameters of unknown dynamical systems becomes equivalent to inferring the underlying generating factors in causal representation learning.

After transforming the parameter estimation into a latent variable identification problem in CRL, we can directly invoke the identifiability theory from the literature. Based on Locatello et al. (2019, Theorem 1.), we conclude that the underlying parameters from an unknown system are in general ***non-identifiable***. Nevertheless, several works proposed different weakly supervised learning strategies that can *partially identify* the latent variables (Ahuja et al., 2022; Brehmer et al., 2022; Daunhawer et al., 2023; Locatello et al., 2020; Von Kügelgen et al., 2021; Yao et al., 2024). To this end, we define partial identifiability in the context of dynamical systems by slightly adapting the definition of block-identifiability proposed by Von Kügelgen et al. (2021):

Definition 3.3 (Partial identifiability)**.** A partition $\boldsymbol{\theta}_S := (\boldsymbol{\theta}_i)_{i \in S}$ with $S \subseteq [N]$ of parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is partially identified by an encoder $g : \mathcal{X}^T \to \boldsymbol{\Theta}$ if the estimator $\hat{\boldsymbol{\theta}}_S := g(\mathbf{x})_S$ contains all and only information about the ground truth partition $\boldsymbol{\theta}_S$, i.e. $\hat{\boldsymbol{\theta}}_S = h(\boldsymbol{\theta}_S)$ for some invertible mapping $h : \boldsymbol{\Theta}_S \to \boldsymbol{\Theta}_S$ where $\boldsymbol{\Theta}_S := \times_{i \in S} \boldsymbol{\Theta}_i$.

Note that the inferred partition $\hat{\boldsymbol{\theta}}_S$ can be a set of *entangled* latent variables rather than a single one. In the multivariate case, one can consider the $\hat{\boldsymbol{\theta}}_S$ as a bijective mixture of the ground truth parameter $\boldsymbol{\theta}_S$.

Corollary 3.2 (Identifiability without known functional form)**.** *Assume a dynamical system f satisfying Asms. 2.1 and 2.2, a pair of trajectories $\mathbf{x}, \tilde{\mathbf{x}}$ generated from the same system f but specified by different parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$, respectively. Assume a partition of parameters $\boldsymbol{\theta}_S$ with $S \subseteq [N]$ is shared across the pair of parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$. Let $g : \mathcal{X}^T \to \Theta$ be some smooth encoder and $\hat{F} : \boldsymbol{\Theta} \to \mathcal{X}^T$ be some left-invertible smooth solver that minimizes the following objective:*

$$\mathcal{L}(g, \hat{F}) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}}} \underbrace{\|g(\mathbf{x})_S - g(\tilde{\mathbf{x}})_S\|_2^2}_{\textit{Alignment}} + \underbrace{\left\|\hat{F}(g(\mathbf{x})) - \mathbf{x}\right\|_2^2 + \left\|\hat{F}(g(\tilde{\mathbf{x}})) - \tilde{\mathbf{x}}\right\|_2^2}_{\textit{Sufficiency}}, \tag{3}$$

then the shared partition $\boldsymbol{\theta}_S$ is partially identified *(Defn. 3.3) by g in the statistical setting.*

Discussion. We remark that an implicit ODE solver $\hat{F}$ is introduced in eq. (3) because the functional form $f_{\boldsymbol{\theta}}$ is unknown. Intuitively, Cor. 3.2 provides partial identifiability results for the shared partition of parameters between two trajectories. We can consider the trajectories to be different simulation experiments but with certain sharing conditions, such as two wind simulations that share the same *layer thickness* parameter. This partial identifiability statement is mainly concluded from the theory in the multiview CRL literature (Ahuja et al., 2022; Brehmer et al., 2022; Daunhawer et al., 2023; Locatello et al., 2020; Schölkopf et al., 2021; Von Kügelgen et al., 2021; Yao et al., 2024). Note that this corollary is *one exemplary demonstration* of achieving partial identifiability in dynamical systems. Many identifiability results from the causal representation works can be reformulated similarly by replacing their decoder with a differentiable ODE solver $\hat{F}$. The high-level idea of multiview CRL is to identify the shared part between different views by enforcing alignment on the shared coordinates while preserving a sufficient information representation. *Alignment* can be obtained by either minimizing the L_2 loss between the encoding from different views on the shared coordinates (Daunhawer et al., 2023; Von Kügelgen et al., 2021; Yao et al., 2024) or maximizing the correlation on the shared dimensions correspondingly (Lyu and Fu, 2022; Lyu et al., 2021); *Sufficiency* of the learned representation is often prompted by maximizing the entropy (Daunhawer et al., 2023; Von Kügelgen et al., 2021; Yao et al., 2024; Zimmermann et al., 2021) or minimizing the reconstruction error (Ahuja et al., 2022; Brehmer et al., 2022; Locatello et al., 2020; Schölkopf et al., 2021). Other types of causal representation learning works will be further discussed in § 5.

4. CRL-construct of Identifiable Neural Emulators for Dynamical Systems

This section provides a step-by-step construct of a neural emulator that can (1) identify the *time-invariant, trajectory-specific* physical parameters from some unknown dynamical systems if the identifiability conditions are met and (2) efficiently forecast future time steps. Identifiability can be guaranteed by employing causal representation learning approaches (§ 3) while forecasting ability can be obtained by using an efficient mechanistic solver (Pervez et al., 2024) as a decoder. For the sake of simplicity, we term these identifiable neural emulators as *identifiers*. We remark that the general architecture remains consistent for most CRL approaches, while the learning object differs slightly in *latent regularization*, which is specified by individual identifiability algorithms. Intuitively, the *latent regularization* can be interpreted as an additional constraint put on the learned encodings imposed by the setting-specific assumptions, such as the *alignment* term in multiview CRL (Cor. 3.2). In the following, we demonstrate building an *identifier* in the multiview setting from scratch and showcase how it can be easily generalized to other CRL approaches with slight adaptation.

Architecture. Since the parameters of interest are *time-invariant* and *trajectory-specific* (§ 2), we input the whole trajectory $\mathbf{x} = (\mathbf{x}(t_1), \dots, \mathbf{x}(t_T))$ to a smooth encoder $g : \mathcal{X}^T \to \mathbf{\Theta}$, as shown in Fig. 5. Then, we decode the trajectory $\hat{\mathbf{x}}$ from estimated parameter vector $\hat{\boldsymbol{\theta}} := g(\mathbf{x})$ using a mechanistic solver (Pervez et al., 2024). The high-level idea of mechanistic neural networks is to approximate the underlying dynamical system using a set of explicit ODEs $\mathcal{U}_{\hat{\boldsymbol{\theta}}} : C(\boldsymbol{\alpha}, \hat{\boldsymbol{\theta}}) = 0$ with learnable coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{d_{\boldsymbol{\alpha}}}$. The explicit ODE family $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$ can then be interpreted as a constrained optimization problem and can thus be solved using a *neural relaxed linear programming solver* (Pervez et al., 2024, Sec 3.1).

In more detail, the original design of MNN predicts the coefficients from the input trajectory $\mathbf{x}$ using an MNN encoder g_{mnn}; however, as we enforce the estimated parameter $\boldsymbol{\theta}$ to preserve *sufficient* information of the entire trajectory $\mathbf{x}$, we instead predict the coefficients $\boldsymbol{\alpha}$ from the estimated parameter $\hat{\boldsymbol{\theta}}$ with the encoder $g_{\text{mnn}} : \mathbf{\Theta} \to \mathbb{R}^{d_{\boldsymbol{\alpha}}}$. Formally, the coefficients $\boldsymbol{\alpha}$ are computed as $\boldsymbol{\alpha} = g_{\text{mnn}}(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}} = g(\mathbf{x})$. The resulting ODE family $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$ provides a broad variability of ODE parametrizations. A detailed formulation of $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$ at t (Pervez et al., 2024, eq. (3)) is given by

$$\underbrace{\textstyle\sum_{i=0}^{l} c_i(t;\hat{\boldsymbol{\theta}})u^{(i)}}_{\text{linear terms}} + \underbrace{\textstyle\sum_{j=0}^{r} \phi_k(t;\hat{\boldsymbol{\theta}})g_k(t,\{u^{(j)}\})}_{\text{nonlinear terms}} = b(t;\hat{\boldsymbol{\theta}}), \tag{4}$$

where $u^{(i)}$ is $i-$th order approximations of the ground truth state $\mathbf{x}$. Like in any ODE solving in practice, solving eq. (4) requires discretization of the continuous coefficients in time (e.g., $c_i(t;\hat{\boldsymbol{\theta}})$). Discretizing the ODE representation $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$:

$$\begin{aligned} \sum_{i=0}^{l} c_{i,t}u_t^{(i)} + \sum_{j=0}^{r} \phi_{k,t}g_k(\{u_t^{(j)}\}) &= b_t \\ s.t. \quad (u_{t_1}, u'_{t_1}, \dots) &= \omega, \end{aligned} \tag{5}$$

where ω denotes the initial state vector of the ODE representation $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$. To this end, we present the explicit definition of the learnable coefficients $\boldsymbol{\alpha} := (c_{i,t}, \phi_{k,t}, b_t, s_t, \omega)$ with $t \in \mathcal{T}, i \in [l], k \in [r]$, which is a concatenation of linear coefficients $c_{i,t}$, nonlinear coefficients $\phi_{i,k}$, adaptive step sizes s_t and initial values ω. Note that we dropped the $\hat{\boldsymbol{\theta}}$ in the notation for simplicity, but all of these coefficients $\boldsymbol{\alpha}$ are predicted from $\hat{\boldsymbol{\theta}}$, as described previously. At last, MNN converts ODE solving into a constrained optimization problem by representing the $\mathcal{U}_{\hat{\boldsymbol{\theta}}}$ using a set of constraints, including ODE equation constraints, initial value constraints, and smoothness constraints (Pervez et al., 2024, Sec 3.1.1). This optimization problem is then solved by *neural relaxed linear programming* solver (Pervez et al., 2024, Sec 3.1) in a time-parallel fashion, thus making the overall mechanistic solver scalable and GPU-friendly.

Learning objective and latent regularizers. Depending on whether the functional form of the underlying dynamical system is known or not, the proposed neural emulator can be trained using the losses given in Cor. 3.1 or Cor. 3.2, respectively. When the functional form is unknown, we employ CRL approaches to *partially* identify the physical parameters. We remark that the causal representation learning schemes mainly differ in the latent regularizers, specified by the assumptions and settings. Therefore, we provide a more extensive summary of different causal representation learning approaches and their corresponding latent regularizer in Tab. 6.

5. Related Work

Multi-environment CRL. Another important line of work in causal representation learning focuses on the multi-environment setup, where the data are collected from multiple different environments and thus *non-identically distributed*. One common way to collect multi-environmental data is to perform *single node interventions* (Ahuja et al., 2023; Buchholz et al., 2024; Squires et al., 2023; Varici et al., 2023; Von Kügelgen et al., 2021; Zhang et al., 2024). Identifiability proofs were provided for different settings, varying from types of mixing functions, causal models and interventions. Squires et al. (2023) considers linear Gaussian model and linear mixing functions, showing identifiability under both hard and soft interventions; Ahuja et al. (2023) considers a more general causal model with bounded support, together with finite degree polynomial mixing function, and provides identifiability proof for *do* and hard interventions. Buchholz et al. (2024) extends Squires et al. (2023) to general nonlinear mixing functions and linear Gaussian latent model. Zhang et al. (2024) show identifiability guarantee for a nonlinear causal model with polynomial mixing functions under soft interventions. Jin et al. (2023) considers linear mixing function with nonlinear model or linear non-Gaussian model under soft interventions. Overall, given the fruitful literature in multi-environment causal representation learning, we believe applying multi-environments methods to build identifiable neural emulators (§ 4) would be an exciting future avenue.

ODE discovery. The ultimate goal of ODE discovery is to learn a human-interpretable equation for an unknown system, given discretized observations generated from this system. Recently, many machine learning frameworks have been used for ODE discovery, such as sparse linear regression (Brunton et al., 2016a;b; Kaheman et al., 2020; Rudy et al., 2017), symbolic regression (Becker et al., 2023; d'Ascoli et al., 2024; d'Ascoli et al., 2022), simulation-based inference (Cranmer et al., 2020; Schröder and Macke, 2023). Becker et al. (2023); d'Ascoli et al. (2022) exploit transformer-based approaches to dynamical symbolic regression for univariate ODEs, which is extended by d'Ascoli et al. (2024) to multivariate case. Schröder and Macke (2023) employs *simulation-based variational* *inference* to jointly learn the operators (like addition or multiplication) and the coefficients. However, this approach typically runs simulations inside the training loop, which could introduce a tremendous computational bottleneck when the simulator is inefficient. On the contrary, our approach works offline with pre-collected data, avoiding simulating on the fly. Although ODE discovery methods can provide symbolic equations for data from an unknown trajectory, the inferred equation does not have to align with the ground truth. In other words, theoretical identifiability guarantees for these methods are still missing.

Identifiability of dynamical systems. Identifiability of dynamical systems has been studied on a *case-by-case* basis in traditional system identification literature (Åström and Eykhoff, 1971; Miao et al., 2011; Villaverde et al., 2016). Liang and Wu (2008) studied ODE identifiability under measurement error. Scholl et al. (2023) investigated the identifiability of ODE discovery with non-parametric assumption, but only for univariate cases. More recently, several works have advanced in identifiability analysis of *linear* ODEs from a *single* trajectory (Duan et al., 2020; Qiu et al., 2022; Stanhope et al., 2014). Overall, current theoretical results cannot conclude whether an unknown nonlinear ODE can be identified from observational data. Hence, in our work, we do not aim to identify the whole equation of the dynamical systems but instead focus on identifying the time-invariant parameters.

6. Experiments

This section provides experiments and results on both simulated and real-world climate data. In both cases, the true functional form of the underlying physical process is unknown, so we employ the multiview CRL approach together with mechanistic neural networks to build our identifiable neural emulator (termed as *mechanistic identifier*), following the steps in § 4. We compare *mechanistic identifier* with three baselines: (1) *Ada-GVAE* (Locatello et al., 2020), a traditional multiview model that uses a vanilla decoder instead of a mechanistic solver. (2) *Time-invariant MNN*, proposed by (Pervez et al., 2024). We choose this variant of MNN as our baseline for a fair comparison. (3) *Contrastive identifier*, a contrastive loss-based CRL approach without a decoder (Daunhawer et al., 2023; Von Kügelgen et al., 2021; Yao et al., 2024). We train *mechanistic identifier* using eq. (3) and other baselines following the steps given in the original papers. After training, we evaluate these methods on their identifiability and long-term forecasting capability.

6.1. Wind simulation

Experimental setup. Our experiment considers longitudinal and latitudinal wind velocities (also termed u, v wind components) from the global wind simulation data generated by various *layer-thickness* parameters. To train the multiview approaches, we generate a tuple of three views: After sampling the first view $\mathbf{x}^1$ randomly throughout the whole training set, we sample another trajectory $\mathbf{x}^2$ from a different location which shares the same simulation condition as the first one, compared to the first view, the third view $\mathbf{x}^3$ is then sampled from another simulation but at the same location. Overall, $\mathbf{x}^1, \mathbf{x}^2$ share the global simulation conditions like the *layer thickness* parameter while $\mathbf{x}^1, \mathbf{x}^3$ only share the local features. All three views share global atmosphere-related features that are not specified as simulation conditions. More details about the data generation process and training pipeline are provided in App. B.2.

Parameter identification. In this experiment, we use the learned representation to classify the ground-truth labels generated by discretizing the generating factor *layer thickness*, and report the accuracy in Fig. 1. In more detail, we use `latent dim`=12 for all models and split the learned encodings into three partitions S_1, S_2, S_3, with four dimensions each. Then, we individually predict the ground truth *layer thickness* labels from each partition. According to the previously mentioned view-generating process, the *layer thickness* parameter should be encoded in S_1 for both *contrastive* and *mechanistic identifiers*. This hypothesis is verified by Fig. 1 since both *contrastive* and *mechanistic identifiers* show a high accuracy of `acc`≈1 in the first partition S_1 and low accuracy in other partitions. On the contrary, *Ada-GVAE* and *TI-MNN* performed significantly worse with an average acc. of 60% everywhere. Overall, Fig. 1 shows both the necessity of explicit time modeling using MNN solver (compared to *Ada-GVAE*) and identifiability power of multiview CRL (compared to *TI-MNN*).

6.2. Real-world sea surface temperature

Experimental setup. We evaluate the models on sea surface temperature dataset *SST-V2* (Huang et al., 2021). For the multiview training, we generate a pair trajectories from a small neighbor region ($\pm 5°$) along the ***same latitude***. We believe these pairs share certain climate properties as the locations from the same latitude share *roughly* the amount of direct sunlight which will directly affect the sea surface temperature. Further infromation about the dataset and training procedure is provided in App. B.2.

Time series forecasting. We chunk the time series into slices of 4 years in training while keeping last four years as out-of-distribution forecasting task. To predict the last chunk, we input data from 2015 to 2018 to get the learned representation $\hat{\boldsymbol{\theta}}$. Since we assume $\hat{\boldsymbol{\theta}}$ to be *time-inavriant*, we decode $\hat{\boldsymbol{\theta}}$ together with 10 initial steps of 2019 to predict the last chunk. Note that *contrastive identifier* is excluded from this task as it does not have a decoder. As shown in Tab. 1, the forecasting performance of *mechanistic Identifier* surpasses *Ada-GVAE* by a great margin, showcasing the

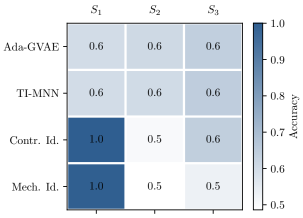

Figure 1: **Prediction accuracy** on *layer thickness* parameter on wind simulation data, evaluated on encoding partitions S_1, S_2, S_3.

Table 1: **Performance evaluation on the SST-V2 data on various types of tasks**. Results averaged over three random seeds with standard deviation, provided as (m ± std).

	SST V2		
	Acc.(ID)(↑)	**Acc.**(OOD)(↑)	**Forecast. error**(↓)
Ada-GVAE	0.468 ± 0.001	0.467 ± 0.000	0.043 ± 0.044
TI-MNN	0.697 ± 0.049	0.668 ± 0.074	0.024 ± 0.016
Contr. Identifier	$\mathbf{0.904} \pm 0.011$	$\mathbf{0.861} \pm 0.022$	✗
Mech. Identifier	$\mathbf{0.902} \pm 0.005$	0.824 ± 0.016	$\mathbf{0.007} \pm 0.003$

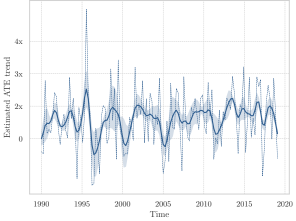

Figure 2: **Causal effect estimation** on SST-V2 from 1990 to 2023, with climate zone as treatment and zonal average temp. as outcome.

superiority of integrating scalable mechanistic solvers in real-world time series datasets. At the same time, *TI-MNN* performed worse and unstably despite the MNN component, verifying the need of the additional information bottleneck (parameter encoder g) and the multiview learning scheme.

Climate-zone classification. Since there is no ground truth latitude-related parameters available, we design a downstream classification task that verifies our learned representation encodes the latitude-related information. The goal of the task is to predict the climate zone *(tropical, temperate, polar)* from the learned *shared* representation because the latitude uniquely defines climate zones. We evaluated the methods in both *in-distribution* (ID) and *out-of-distribution* (OOD) setup for all baselines. In the OOD setting, we input data from longitude $10°$ to longitude $360°$ when training the classifier while keeping the first 10 degree as our out-of-distribution test data. Tab. 1 show that both *contrastive* and *mechanistic identifiers* perform decently, supporting the applicability of identifiable multiview CRL algorithms in dynamical systems. Overall, the performance of multiview CRL-based approaches (*contrastive and mechanistic identifiers*) far exceeds *Ada-GVAE* and *TI-MNN*, again showcasing the superiority of the combination of causal representation learning and mechanistic solvers.

Average treatment effect estimation. We further investigate the effect of climate zone on average temperature along one specific latitude through *average treatment effect* (ATE) estimation. Formally, we consider the latitudinal average temperature as outcome Y, two climate zones (*tropical* $(T = 0)$, *polar*$(T = 1)$) as binary treatments, and the predicted latitude-specific features as unobserved mediators. Formally, ATE is defined as: $\text{ATE} := \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)]$. Since ATE cannot be computed directly (Holland, 1986), we estimate it using the popular *AIPW* estimator (Robins et al., 1994). Fig. 2 illustrates the estimated ATE change ratio from 1990 to 2020, computed by ${}^{ATE(year)-ATE(1990)}/_{ATE(1990)}$. We observe that the recent ATE ratio has risen to 2x compared to the year 1990, which surprisingly aligns with the fact that the Arctic Ocean recently became at least twice as warm as before (Rantanen et al., 2022).

7. Limitations and Conclusion

In this paper, we build a bridge between causal representation learning and dynamical system identification. By virtue of this connection, we successfully equipped existing mechanistic models (focusing on (Pervez et al., 2024) in practice for scalability reasons) with identification guarantees. Our analysis covers a large number of papers, including (Brunton et al., 2016a;b; Kaheman et al., 2020; Kaptanoglu et al., 2021; Pervez et al., 2024; Wenk et al., 2019) explicitly refraining from making identifiability statements. At the same time, our work demonstrated that causal representation learning training constructs are ready to be applied in the real world, and the connection with dynamical systems offers untapped potential due to its relevance in the sciences. This was an overwhelmingly acknowledged limitation of the causal representation learning field (Ahuja et al., 2023; Buchholz et al., 2024; Daunhawer et al., 2023; Locatello et al., 2020; Squires et al., 2023; Varici et al., 2023; Von Kügelgen et al., 2021; Yao et al., 2024). Having clearly demonstrated the mutual benefit of this connection, we hope that future work will scale up identifiable mechanistic models and apply them to even more complex dynamical systems and real scientific questions. Nevertheless, this paper has several technical limitations that could be addressed in future work. First of all, the proposed theory explicitly requires *determinism* as one of the key assumptions (Asm. 3.1), which directly excludes another important type of differential equation: Stochastic Differential Equations. Second, we assume we directly observe the state $\mathbf{x}$ without considering measurement noise. Although the empirical results were promising on real-world noisy data (§ 6.2), we believe explicitly modeling measurement noise would elevate the theory. Finally, our identifiability analysis focuses on the infinite data regime, which is unrealistic in real-world scenarios.

References

Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100 (1):90–93, 1974. 14

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022. 5, 18

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023. 6, 8

Karl Johan Åström and Peter Eykhoff. System identification—a survey. *Automatica*, 7(2):123–162, 1971. 7

Sören Becker, Michal Klein, Alexander Neitz, Giambattista Parascandolo, and Niki Kilbertus. Predicting ordinary differential equations with transformers. In *International Conference on Machine Learning*, pages 1978–2002. PMLR, 2023. 6

Ror Bellman and Karl Johan Åström. On structural identifiability. *Mathematical biosciences*, 7(3-4):329–339, 1970. 3

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35: 38319–38331, 2022. 1, 5

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, 2016a. 1, 4, 6, 8, 13

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Sparse identification of nonlinear dynamics with control (sindyc). *IFAC-PapersOnLine*, 49(18):710–715, 2016b. 1, 4, 6, 8, 13

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, 2024. 6, 8

Ricky T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq. 4

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018. 1, 4, 15

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Learning neural event functions for ordinary differential equations. *International Conference on Learning Representations*, 2021. 1, 4, 15

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. 6

Stéphane d'Ascoli, Sören Becker, Philippe Schwaller, Alexander Mathis, and Niki Kilbertus. ODEFormer: Symbolic regression of dynamical systems with transformers. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 6

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023. 3, 5, 7, 8, 13, 18

Xiaoyu Duan, JE Rubin, and David Swigon. Identification of affine dynamical systems from a single trajectory. *Inverse Problems*, 36(8):085004, 2020. 7

Stéphane d'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, and Francois Charton. Deep symbolic regression for recurrence prediction. In *International Conference on Machine Learning*, pages 4520–4536. PMLR, 2022. 1, 6

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 8

Boyin Huang, Chunying Liu, Viva Banzon, Eric Freeman, Garrett Graham, Bill Hankins, Tom Smith, and Huai-Min Zhang. Improvements of the daily optimum interpolation sea surface temperature (doisst) version 2.1. *Journal of Climate*, 34(8):2923–2939, 2021. 7, 15

Edward L Ince. *Ordinary differential equations*. Courier Corporation, 1956. 3

Songyao Jin, Feng Xie, Guangyi Chen, Biwei Huang, Zhengming Chen, Xinshuai Dong, and Kun Zhang. Structural estimation of partially observed linear non-gaussian acyclic model: A practical approach with identifiability. In *The Twelfth International Conference on Learning Representations*, 2023. 6

Kadierdan Kaheman, J Nathan Kutz, and Steven L Brunton. Sindy-pi: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A*, 476(2242):20200279, 2020. 1, 4, 6, 8

Alan A Kaptanoglu, Jared L Callaham, Aleksandr Aravkin, Christopher J Hansen, and Steven L Brunton. Promoting global stability in data-driven models of quadratic nonlinear dynamics. *Physical Review Fluids*, 6(9):094401, 2021. 4, 8

Patrick Kidger, Ricky T. Q. Chen, and Terry J. Lyons. "hey, that's not an ode": Faster ode adjoints via seminorms. *International Conference on Machine Learning*, 2021. 1, 15

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15687–15701. Curran Associates, Inc., 2022. 1

Milan Klöwer and the SpeedyWeather.jl Contributors. Speedyweather.jl, 2023. 14

Sébastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. *arXiv preprint arXiv:2207.07732*, 2022. 17, 18

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. In *International Conference on Machine Learning*, pages 18171–18206. PMLR, 2023. 17, 18

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, 2024. 1

Hua Liang and Hulin Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008. 7

Ernest Lindelöf. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 116(3):454–457, 1894. 3

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2022a. 1

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, 2022b. 1

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 1, 5, 17

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020. 3, 5, 7, 8, 14, 18

Per Lötstedt and Linda Petzold. Numerical solution of nonlinear differential equations with algebraic constraints. i. convergence results for backward differentiation formulas. *Mathematics of computation*, 46(174):491–516, 1986. 4

Peter Y Lu, Joan Ariño Bernad, and Marin Soljačić. Discovering sparse interpretable dynamics from partial observations. *Communications Physics*, 5(1):206, 2022. 13, 17

Qi Lyu and Xiao Fu. On finite-sample identifiability of contrastive learning-based nonlinear independent component analysis. In *International Conference on Machine Learning*, pages 14582–14600. PMLR, 2022. 5

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. *arXiv preprint arXiv:2106.07115*, 2021. 5

Hongyu Miao, Xiaohua Xia, Alan S Perelson, and Hulin Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM review*, 53(1):3–39, 2011. 3, 4, 7

Eric Mjolsness and Dennis DeCoste. Machine learning for science: state of the art and future prospects. *science*, 293 (5537):2051–2055, 2001. 1

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. 17, 18

Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. On second order behaviour in augmented neural odes. *Advances in neural information processing systems*, 33:5911–5921, 2020. 4

Adeel Pervez, Francesco Locatello, and Efstratios Gavves. Mechanistic neural networks for scientific machine learning. *International Conference on Machine Learning*, 2024. 2, 4, 5, 6, 7, 8, 14, 15

Xing Qiu, Tao Xu, Babak Soltanalizadeh, and Hulin Wu. Identifiability analysis of linear ordinary differential equation systems with a single trajectory. *Applied Mathematics and Computation*, 430:127260, 2022. 7

Maithra Raghu and Eric Schmidt. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*, 2020. 1

Mika Rantanen, Alexey Yu Karpechko, Antti Lipponen, Kalle Nordling, Otto Hyvärinen, Kimmo Ruosteenoja, Timo Vihma, and Ari Laaksonen. The arctic has warmed nearly four times faster than the globe since 1979. *Communications earth & environment*, 3(1):168, 2022. 8

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994. 8

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. 6

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. 1, 3, 5

Philipp Scholl, Aras Bacho, Holger Boche, and Gitta Kutyniok. The uniqueness problem of physical law learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 7

Cornelius Schröder and Jakob H Macke. Simultaneous identification of models and parameters of scientific simulators. *arXiv preprint arXiv:2305.15174*, 2023. 1, 6

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, volume 202, pages 32540–32560. PMLR, 2023. 1, 6, 8

Shelby Stanhope, Jonathan E Rubin, and David Swigon. Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory. *SIAM Journal on Applied Dynamical Systems*, 13(4):1792–1815, 2014. 7

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

Francesco Tonolini, Bjørn Sand Jensen, and Roderick Murray-Smith. Variational sparse coding. In *Uncertainty in Artificial Intelligence*, pages 690–700. PMLR, 2020. 18

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Score-based causal representation learning from interventions: Nonparametric identifiability. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. 1, 6, 8

Alejandro F Villaverde, Antonio Barreiro, and Antonis Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLoS computational biology*, 12(10):e1005153, 2016. 7

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021. 1, 3, 5, 6, 7, 8, 13, 17, 18

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, 2024. 1

Eric Walter, Luc Pronzato, and John Norton. *Identification of parametric models from experimental data*, volume 1. Springer, 1997. 3

Philippe Wenk, Alkis Gotovos, Stefan Bauer, Nico S Gorbach, Andreas Krause, and Joachim M Buhmann. Fast gaussian process based gradient matching for parameter identification in systems of nonlinear odes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1351–1360. PMLR, 2019. 4, 8, 13

Franz-Georg Wieland, Adrian L Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. *Current Opinion in Systems Biology*, 25:60–69, 2021. 3

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *International Conference on Machine Learning*, 2024. 1, 17, 18

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 5, 7, 8, 13, 14, 17, 18

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 6

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. 5

A. Proofs

A.1. Proofs for full identifiability

Corollary 3.1 (Full identifiability with known functional form)**.** *Consider a trajectory $\mathbf{x} \in \mathcal{X}^T$ generated from a ODE $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$ satisfying Asms. 2.1 and 2.2, let $\hat{\boldsymbol{\theta}}$ be an estimator minimizing the following objective:*

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \left\| F(\hat{\boldsymbol{\theta}}) - \mathbf{x} \right\|_2^2 \tag{2}$$

*then the parameter $\boldsymbol{\theta}$ is **fully-identified** (Defn. 3.1) by the estimator $\hat{\boldsymbol{\theta}}$.*

Proof. We begin by showing the global minimum of $\mathcal{L}(\hat{\boldsymbol{\theta}})$ exists and equals zero. Then, we show by contradiction that any estimators $\hat{\boldsymbol{\theta}}$ that obtains this global minimum has to equal the ground truth parameters $\boldsymbol{\theta}$.

Step 1. We show that the global minimum zero can be obtained for $\mathcal{L}(\hat{\boldsymbol{\theta}})$. Consider the ground truth parameter $\boldsymbol{\theta} \in \Theta$, then by definition of the ODE solver F (Defn. 3.2), we have:

$$\mathcal{L}(\boldsymbol{\theta}) = \| F(\boldsymbol{\theta}) - \mathbf{x} \|_2^2 = \| \mathbf{x} - \mathbf{x} \|_2^2 = 0. \tag{6}$$

Step 2. Suppose for a contraction that there exists a $\boldsymbol{\theta}^* \in \Theta$ that minimizes the loss eq. (2) but differs from the ground truth parameters $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$. This implies:

$$\mathcal{L}(\boldsymbol{\theta}^*) = \| F(\boldsymbol{\theta}^*) - \mathbf{x} \|_2^2 = 0 \tag{7}$$

Note that $\mathcal{L}(\boldsymbol{\theta}^*)$ can be rewritten as:

$$\mathcal{L}(\boldsymbol{\theta}^*) = \sum_{k=1}^{T} \| F(\boldsymbol{\theta}^*)_{t_k} - \mathbf{x}(t_k) \|_2^2 = 0 \tag{8}$$

To make sure the sum is zero, each individual term has to be zero, that is $F(\boldsymbol{\theta}^*)_{t_k} = \mathbf{x}(t_k), \forall t \in \{t_1, \ldots, t_T\}$. According to the uniqueness assumption of the ODE (Asm. 2.1), this implies $\boldsymbol{\theta}^* = \boldsymbol{\theta}$, which leads to a contradiction.

Thus, we have shown that minimizing eq. (2) will yield the ground truth parameter $\boldsymbol{\theta}$. In other words, any estimator $\hat{\boldsymbol{\theta}}$ that minimizes eq. (2) fully identifies $\boldsymbol{\theta}$. □

Full identifiability with closed form solution when $f_{\boldsymbol{\theta}}$ is linear in $\boldsymbol{\theta}$. We show that a closed-form solution can be obtained through linear least squares when the vector field $f_{\boldsymbol{\theta}}$ is linear in $\boldsymbol{\theta}$ and if we observe a *first-order* trajectory. A *first-order* trajectory means the first-order derivatives are included in the state-space vector. This statement is formalized as follows:

Observation A.1. Given a first-order trajectory $(\mathbf{x}, \dot{\mathbf{x}}) = (\mathbf{x}(t), \dot{\mathbf{x}}(t))_{t \in \mathcal{T}}$ generated from a dynamical system $f_{\boldsymbol{\theta}}(\mathbf{x}(t))$ satisfying Asms. 2.1 and 2.2. In particular, this ODE $f_{\boldsymbol{\theta}}$ can be written as a weighted sum of a set of base functions $\{\phi_1, \ldots, \phi_m\}$, i.e., $f_{\boldsymbol{\theta}}$ is linear in $\boldsymbol{\theta}$:

$$f_{\boldsymbol{\theta}}(\mathbf{x}(t)) = \textstyle\sum_{i=1}^m \theta_i \phi_i(\mathbf{x}). \tag{9}$$

Define $\Phi_{\mathbf{x}} := [\phi_i(\mathbf{x}(t))]_{i \in [m], t \in \mathcal{T}} \in \mathbb{R}^{m \times T}$, then the global optimum of the loss eq. (2) is given by

$$\boldsymbol{\theta}^* = \left(\Phi_{\mathbf{x}}^\intercal \Phi_{\mathbf{x}}\right)^{-1} \phi_{\mathbf{x}} \dot{\mathbf{x}} \tag{10}$$

As a direct implication, SINDy-like approaches (Brunton et al., 2016a;b; Lu et al., 2022) and gradient matching (Wenk et al., 2019) can fully identify the underlying physical parameters $\boldsymbol{\theta}$ even with a closed-form solution if the underlying vector field $f_{\boldsymbol{\theta}}$ is can be represented as a sparse weighted sum of the given base functions $\{\phi_i\}_{i \in [m]}$.

A.2. Proofs for partial identifiability

Corollary 3.2 (Identifiability without known functional form)**.** *Assume a dynamical system f satisfying Asms. 2.1 and 2.2, a pair of trajectories $\mathbf{x}, \tilde{\mathbf{x}}$ generated from the same system f but specified by different parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$, respectively. Assume a partition of parameters $\boldsymbol{\theta}_S$ with $S \subseteq [N]$ is shared across the pair of parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$. Let $g : \mathcal{X}^T \to \Theta$ be some smooth encoder and $\hat{F} : \boldsymbol{\Theta} \to \mathcal{X}^T$ be some left-invertible smooth solver that minimizes the following objective:*

$$\mathcal{L}(g, \hat{F}) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}}} \underbrace{\| g(\mathbf{x})_S - g(\tilde{\mathbf{x}})_S \|_2^2}_{\textit{Alignment}} + \underbrace{\left\| \hat{F}(g(\mathbf{x})) - \mathbf{x} \right\|_2^2 + \left\| \hat{F}(g(\tilde{\mathbf{x}})) - \tilde{\mathbf{x}} \right\|_2^2}_{\textit{Sufficiency}}, \tag{3}$$

then the shared partition $\boldsymbol{\theta}_S$ is partially identified *(Defn. 3.3) by g in the statistical setting.*

Proof. This proof can be directly adapted from the proofs with by Daunhawer et al. (2023); Von Kügelgen et al. (2021); Yao et al. (2024) with slight modification. So we briefly summarize the **Step 1.** and **Step 2.** that are imported from previous work and focus on the modification (**Step 3.**).

Step 1. We show that the loss function eq. (3) is lower bounded by zero and construct optimal encoder $g^* : \mathcal{X}^T \to \boldsymbol{\Theta}$ that reach this lower bound. Define $g^* : \mathcal{X}^T \to \boldsymbol{\Theta} := F^{-1}$ as the inverse of the ground truth data generating process, i.e., for all trajectories $\mathbf{x} = F(\boldsymbol{\theta})$ that generated from parameter $\boldsymbol{\theta}$, it holds:

$$g^*(\mathbf{x}) = \boldsymbol{\theta} \tag{11}$$

Thus, we have shown that the global minimum *zero* exists and can be obtained by the inverse mixing function $F^{-1} : \mathcal{X}^T \to \boldsymbol{\Theta}$ (Defn. 3.2).

Table 2: **Comparing typical assumptions** of parameter estimation for dynamical systems and latent variable identification in causal representation learning. We justify that the common assumptions in both fields are aligned, providing theoretical ground for applying identifiable CRL methods to learning-based parameter estimation approaches in dynamical systems.

param. estimation		**CRL**		**Explanation**
ref	*assumption*	*assumption*	*ref*	
2.1	*existence & uniqueness*	*determ. gen.*	3.1	Both 2.1 and 3.1 implies deterministic generative process.
		$supp(\boldsymbol{\theta}) = \boldsymbol{\Theta}$	3.3	2.1 implies 3.3 as $\mathbf{x}_{\boldsymbol{\theta}}$ uniquely exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
2.2	*structural identifiability*	*injectivity*	3.2	2.2 implies 3.2 of the solution $\mathbf{x}_{\boldsymbol{\theta}}$.

Step 2. We show that any optimal encoders g that minimizes eq. (3) must have the alignment equal zero, in other words, it has to satisfy the ***invariance*** condition, which is formalized as

$$g(\mathbf{x})_S = g(\tilde{\mathbf{x}}) \qquad a.s. \tag{12}$$

Following Yao et al. (2024, Lemma D.3), we conclude that both $g(\mathbf{x})_S$ and $g(\tilde{\mathbf{x}})_S$ can only depend on information about the shared partition about the ground truth parameter $\boldsymbol{\theta}_S$. In other words,

$$g(\mathbf{x})_S = g(\tilde{\mathbf{x}})_S = h(\boldsymbol{\theta}_S) \tag{13}$$

for some smooth $h : \Theta_S \to \Theta_S$.

Step 3. At last, we show that h is invertible. Note that any optimal encoders g that minimizes eq. (3) must have zero reconstruction error on both $\mathbf{x}$ and $\tilde{\mathbf{x}}$. Taking $\mathbf{x}$ as an example, we have

$$\mathbb{E} \left\| \hat{F}(g(\mathbf{x})) - \mathbf{x} \right\|_2^2 = 0 \tag{14}$$

which implies

$$\hat{F}(g(\mathbf{x})) = \mathbf{x} \qquad a.s. \tag{15}$$

If two continuous functions $\hat{F}(g(\mathbf{x}))$ and $\mathbf{x}$ equals *almost* everywhere on $\boldsymbol{\Theta}$, then they are equal everywhere on $\boldsymbol{\Theta}$, which implies:

$$\hat{F}(g(\mathbf{x})) = \mathbf{x} \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{16}$$

Substituting $\mathbf{x}$ with the ground truth generating process F:

$$\hat{F}(g(\mathbf{x})) = F(\boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{17}$$

applying the left inverse of $\hat{F}$, we have:

$$\hat{F}^{-1} \circ \hat{F}(g(\mathbf{x})) = \hat{F}^{-1} \circ F(\boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{18}$$

i.e.,

$$g(\mathbf{x}) = \hat{F}^{-1} \circ F(\boldsymbol{\theta}) = \hat{F}^{-1} \circ F(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{\bar{S}}) \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{19}$$

Define $h^* := \hat{F}^{-1} \circ F$, note that h^* is bijective as a composition of bijections. Imposing the ***invariance*** constraint, we have $g(\mathbf{x})_S = h^*(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{\bar{S}})_S$. Since $g(\mathbf{x})_S$ cannot depend on $\boldsymbol{\theta}_{\bar{S}}$, we have $g(\mathbf{x})_S = h^*_S(\boldsymbol{\theta}_S)$ with $h := h^*_S : \boldsymbol{\Theta}_S \to \boldsymbol{\Theta}_S$.

Thus we have shown that $g(\mathbf{x})_S$ *partially* identifies $\boldsymbol{\theta}_S$.

□

B. Experimental results

General remarks. All models used in the experiments (§ 6) (*Ada-GVAE, TI-MNN, contrastive identifier, mechanistic identifier*) were built upon open-sourced code provided by the original works (Locatello et al., 2020; Pervez et al., 2024; Yao et al., 2024), under the MIT license. For *mechanistic identifiers*, we add a regularizer multiplier on the *alignment* constraint (Defn. 3.3), which is shown in Tabs. 3 and 5.

B.1. Wind simulation: `SpeedyWeather.jl`

We simulate global air motion using using the `ShallowWaterModel` from `speedy weather` Julia package (Klöwer and the SpeedyWeather.jl Contributors, 2023). We consider a *layer thickness* as the primary generating factor in `ShallowWaterModel` varying from 8e3[m] to 2e4[m], which is a reasonable range given by the climate science literature. Taking the minimal and maximal values, we simulate the wind in a binary fashion and obtain 9024 trajectories across the globe under different conditions. Each trajectory constitutes three output variables discretized on `ts`=121 time steps, on a 3D resolution grid of size: latitude `lat`=47; longitude `lon`=96; level `lev`=1. The three output variables represent *u wind component* (parallel to longitude), *v wind component* (parallel to latitude), and *relative vorticity*, respectively. An illustrative example of all three components is depicted in Fig. 3. further details about the simulation output are provided in Tab. 4. In particular, to train more efficiently, we pre-process the data using a *discrete cosine transform* (DCT) proposed by Ahmed et al. (1974) and only keep the first 50% frequencies. This is feasible as the original data possesses a certain periodic pattern, as shown in Fig. 4. For all baselines, we train the

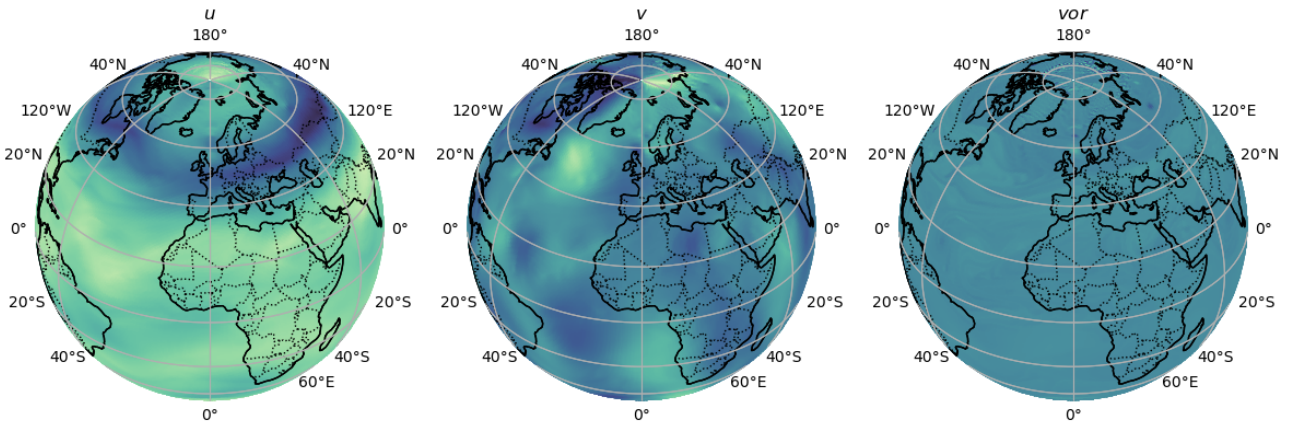

Figure 3: **Example of wind simulation**: *Left:* longitudinal wind velocity (u) [m/s]. *Middle*: latitudinal wind velocity (v)[m/s], *Right*: relative vorticity (vor) [1/s].

model till convergence. More training and test details for the tasks in § 6.1 are summarized in Tab. 3. To validate identifiability, we use `LogisticRegression` model from `scikit-learn` in its default setting to evaluate the classification accuracy in Fig. 1.

B.2. Sea surface temperature: SST-V2

The sea surface temperature data SST-V2 (Huang et al., 2021) contains the *weekly* sea surface temperature data from 1990 to 2023, on a resolution grid of 180 × 360 (latitudes × longitudes). An example input is depicted in Fig. 7. Each time series contains 1727 times steps. To generate multiple views that share specific climate properties, we sample two different trajectories from a small neighbor region ($\pm 5°$) along the ***same latitude***, as the latitude differs in the amount of direct sunlight thus directly affecting the sea surface temperature.

Fig. 5 gives an overview of *mechnistic identifier's* working pipeline and potential applicability in causal downstream task. For a fair comparison, we train all baselines till convergence following the setup summarized in Tab. 5. Similar to the wind simulation data, we pre-process the SST-V2 data using DCT and keep the first 25% frequencies. We only keep $^1/_4$ of the frequencies because sea surface temperature data is highly periodic due to seasonality patterns. Fig. 6 shows an example of predicted trajectories over three randomly sampled locations. As for the downstream classification task, we use `LogisticRegression` model from `scikit-learn` in its default setting to evaluate the classification accuracy in Tab. 1.

B.3. Experiments and compute

In this paper, we train four different models, each over three independent seeds. All 12 jobs ran with 24GB of RAM, 8 CPU cores, and a single node GPU, which is, in most cases, `NVIDIA GeForce RTX2080Ti`. Given different model sizes and convergence rates, the required amount of compute could vary slightly, despite the pre-fixed training epochs. Thus, we report an upper bound of the compute hours on `NVIDIA GeForce RTX2080Ti`. On average, all runs converge within 22 GPU hours. Therefore, the experimental results in this paper can be reproduced with 264 GPU hours.

C. Discussion

Why mechanistic neural networks (Pervez et al., 2024). As mentioned in § 4, the ODE solver F given in Cors. 3.1 and 3.2 can be interpreted as the decoder in a traditional representation learning regime; however, several challenges arise when integrating ODE solving in the training loop: First of all, the ODE solver must be differentiable to utilize the automatic differentiation implementation of the state-of-the-art deep learning frameworks; this obstacle has been tacked by the line of work termed *NeuralODE*, which models the ODE vector field using a neural network thus enable differentiability (Chen et al., 2018; 2021; Kidger et al., 2021). Nevertheless, most differentiable ODE solvers solve the ODE autoregressively and thus cannot be parallelized by the GPU very efficiently. Dealing with long-term trajectories (for example, weekly climate data during the last few decades) would be extremely computationally heavy. Therefore, we advocate for a time- and memory-efficient differentiable ODE solver: the mechanistic neural networks (Pervez et al., 2024).

Latent regularizers in CRL. The framework proposed in § 4 can be generalized to many causal representation learning works, by specifying the latent regularizes according to individual assumptions and settings. For example, in the multiview setting, the latent regularizer can be the L_2 *alignment* between the learned representations on the shared

Table 3: Training setup for wind simulation in § 6.1. Non-applicable fields are marked with ✗.

	Ada-GVAE	**TI-MNN**	**Cont. Identifier**	**Mech. Identifier**
Pre-process	DCT	DCT	DCT	DCT
Encoder	6-layer MLP	6-layer MLP	6-layer MLP	6-layer MLP
Decoder	6-layer MLP	6-layer MLP	✗	3 proj. × 6-layer MLP
Time dim	121	121	121	121
State dim	2	2	2	2
Hidden dim	1024	1024	1024	1024
Latent dim	12	12	12	12
Optimizer	Adam	Adam	Adam	Adam
Adam: learning rate	1e−5	1e−5	1e−5	1e−5
Adam: beta1	0.9	0.9	0.9	0.9
Adam: beta2	0.999	0.999	0.999	0.999
Adam: epsilon	1e−8	1e−8	1e−8	1e−8
Batch size	1128	1128	1128	1128
Temperature τ	✗	✗	0.1	✗
Alignment reg.	✗	✗	✗	10
# Initial values	10	10	✗	10
# Iterations	< 30,000	< 30,000	< 30,000	< 30,000
# Seeds	3	3	3	3

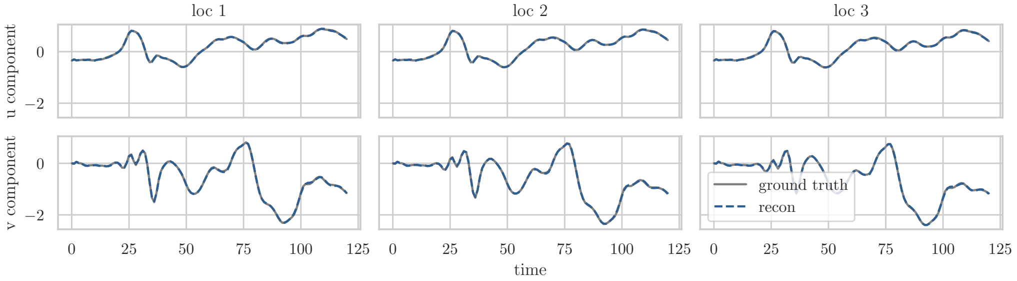

Figure 4: **Wind simulation**: *mechanistic identifier* reconstruction of highly irregular time series. The first half of the trajectory is provided as initial values, while the second half is predicted.

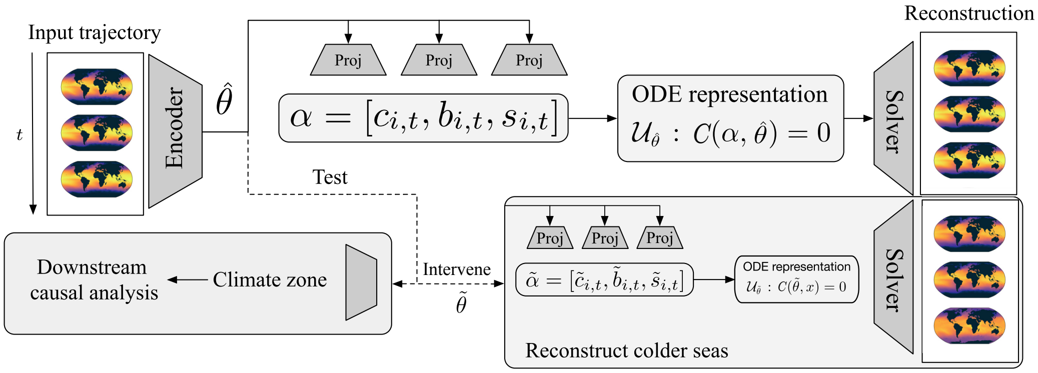
Input trajectory
t
Encoder
$\hat{\theta}$
Proj
Proj
Proj
$\alpha = [c_{i,t}, b_{i,t}, s_{i,t}]$
ODE representation
$\mathcal{U}_{\hat{\theta}} : C(\alpha, \hat{\theta}) = 0$
Solver
Reconstruction
Test
Downstream causal analysis
Climate zone
Intervene
$\tilde{\theta}$
Proj
Proj
Proj
$\tilde{\alpha} = [\tilde{c}_{i,t}, \tilde{b}_{i,t}, \tilde{s}_{i,t}]$
ODE representation
$\mathcal{U}_{\hat{\theta}} : C(\tilde{\theta}, x) = 0$
Solver
Reconstruct colder seas

Figure 5: Our ***mechanistic identifier*** learns the underlying physical parameters $\boldsymbol{\theta}$, providing a versatile neural emulator for downstream causal analysis.

Table 4: **Wind simulation**: output variables.

Output variable [unit]	Shape
Longitudinal wind velocity (u) [m/s]	(`ts`, `lev`, `lat`, `lon`)
Latitudinal wind velocity (v) [m/s]	(`ts`, `lev`, `lat`, `lon`)
Relative vorticity (vor) [1/s]	(`ts`, `lev`, `lat`, `lon`)

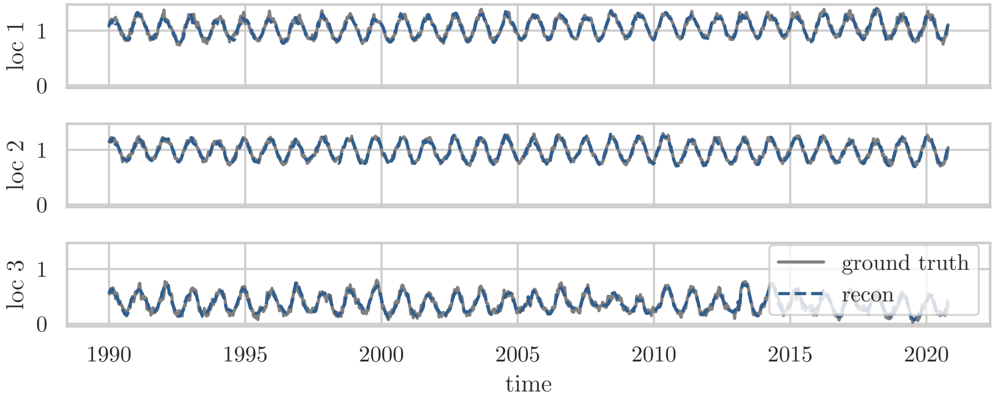

Figure 6: **SST-V2**: *mechanistic identifier* reconstruction over long-term time series. Results are produced by concatenating subsequently predicted chunks.

partition eq. (3), as it was assumed that the paired views are generated based on this overlapping set of latents (Locatello et al., 2019; Von Kügelgen et al., 2021; Yao et al., 2024); in sparse causal representation learning the underlying generative process assumes observations are generated from sparse latent variables; therefore, the proposed algorithms actively enforce some sparsity constraint on the learned representation (Lachapelle and Lacoste-Julien, 2022; Lachapelle et al., 2023; Moran et al., 2022; Xu et al., 2024), We provide a more extensive summary of different causal representation learning approaches and their corresponding latent regularizer in Tab. 6. By replacing the *alignment* term (Cor. 3.2) with the specific latent constraints, one can plug in many causal representation learning algorithms to construct an identifiable neural emulator using our framework.

Identifying time-varying parameters Time-varying parameters $\boldsymbol{\theta}(t)$ could also be potentially identified when they change sparsely in time. For example, a time-varying parameter $\boldsymbol{\theta}_k$ remains constant between (t_k, t_{k+1}). Then, the states in between $\mathbf{x}(t), \mathbf{x}(t+1), \dots, \mathbf{x}(t+k)$ can be considered as multiple views that share the same parameter $\boldsymbol{\theta}_k$. Following this perspective, the time-invariant parameters considered in the scope of this paper remain consistent through the whole timespan $(0, t_{\max}$, thus all discretized states $\mathbf{x}(t_1), \dots, \mathbf{x}(t_T)$ are views that share this parameter. This inductive bias is directly built into the architecture design by inputting the whole trajectory into the encoder instead of doing so step by step (where the time axis is considered as batch dimension). From another angle, the time-varying parameters $\boldsymbol{\theta}(t)$ could be interpreted as a *hidden* part of the state space vector $\mathbf{x}(t)$ without an explicitly defined differential equation, which gives rise to a partial observable setup. This direction has been studied in the context of sparse system identification without explicit identifiability analysis (Lu et al., 2022).

Model evaluation on real-world data. A great obstacle hindering causal representation learning scaling to real-world data is that no ground truth latent variables are available. Since the methods aim to *identify* the latent variables, it is hard to validate the identifiability theory without ground truth-generating factors. However, properly evaluating the CRL models on real-world data can be conducted by carefully designing causal downstream tasks, such as climate zone classification and ATE estimation shown in § 6.2. Overall, we believe by incorporating domain knowledge of the applied datasets, we can use CRL to answer important causal questions from individual fields, thus indirectly validating the identifiability.

Table 5: Training setup for sea surface temperature in § 6.2. Non-applicable fields are marked with ✗.

	Ada-GVAE	**TI-MNN**	**Cont. Identifier**	**Mech. Identifier**
Pre-process	DCT	DCT	DCT	DCT
Encoder	6-layer MLP	6-layer MLP	6-layer MLP	6-layer MLP
Decoder	6-layer MLP	6-layer MLP	✗	3 proj. × 6-layer MLP
Time dim	208	208	208	208
State dim	1	1	1	1
Hidden dim	1024	1024	1024	1024
Latent dim	20	20	20	20
Optimizer	Adam	Adam	Adam	Adam
Adam: learning rate	$1e{-5}$	$1e{-5}$	$1e{-5}$	$1e{-5}$
Adam: beta1	0.9	0.9	0.9	0.9
Adam: beta2	0.999	0.999	0.999	0.999
Adam: epsilon	$1e{-8}$	$1e{-8}$	$1e{-8}$	$1e{-8}$
Batch size	2160	2160	2160	2160
Temperature τ	✗	✗	0.1	✗
Alignment reg.	✗	✗	✗	10
# Initial values	10	10	✗	10
# Iterations	< 30,000	< 30,000	< 30,000	< 30,000
# Seeds	3	3	3	3

Table 6: A non-exhaustive summary of latent regularizers in recent CRL approaches.

Principle	**Assumption**	**Latent regularizer**	**References**
multiview	*part. shared* latents	$\lVert g(\mathbf{x})_S - g(\tilde{\mathbf{x}})_S \rVert_2^2$	Locatello et al. (2020); Von Kügelgen et al. (2021) Daunhawer et al. (2023); Yao et al. (2024)
		$\lVert g(\tilde{\mathbf{x}}) - g(\mathbf{x}) - \delta \rVert_2^2$	Ahuja et al. (2022)
sparsity	*sparse* causal graph	$\lVert g(\mathbf{x}) \rVert_1$	Lachapelle et al. (2023); Xu et al. (2024)
		Spike and Slab prior	Moran et al. (2022); Tonolini et al. (2020)
	temporal sparsity	KL $(q(z^t \mid x^t) \vert\vert \hat{p}(z^t \vert z^{<t}, a^{<t}))$	Lachapelle and Lacoste-Julien (2022)

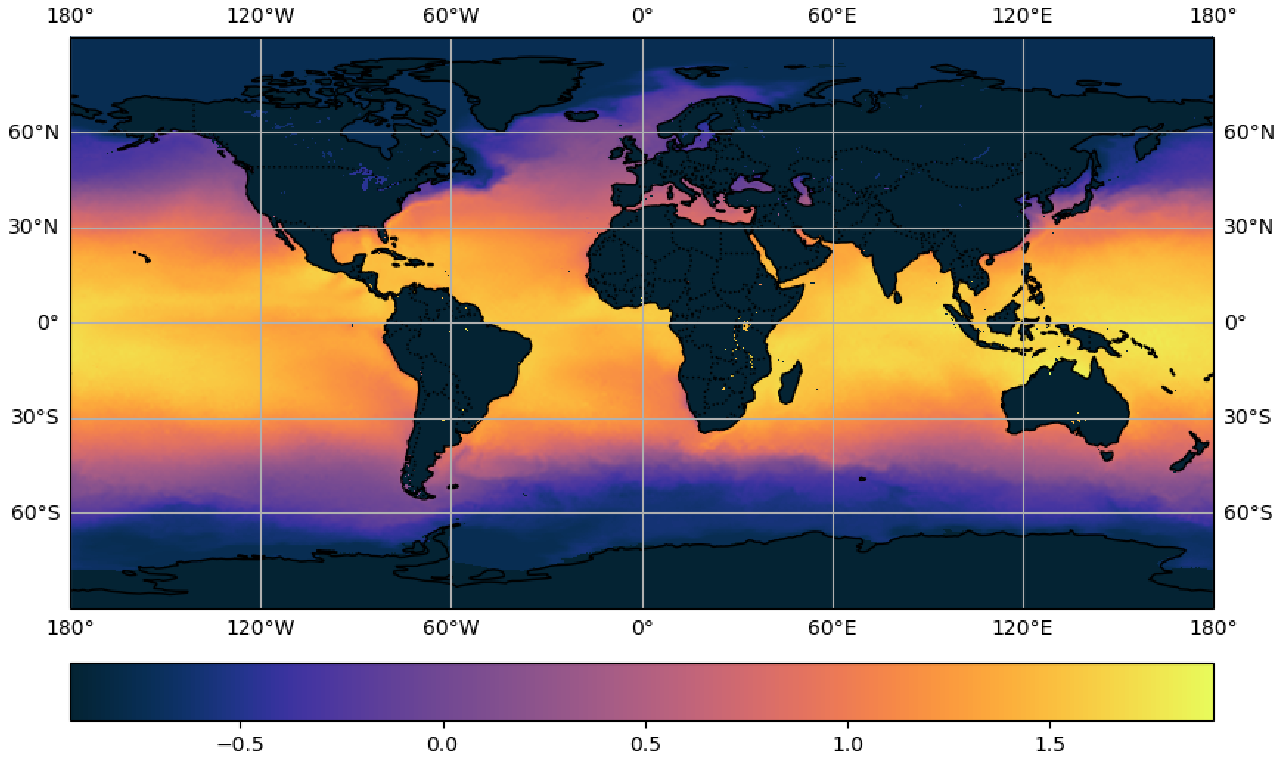

Figure 7: **Example of global sea surface temperature** in January, 1990.