On the Identifiability of Hybrid Deep Generative Models: Meta-Learning as a Solution

Yubo Ye *
Zhejiang University
22230131@zju.edu.cn

Maryam Toloubidokhti
Rochester Institute of Technology
mt6129@rit.edu

Sumeet Vadhavkar
Rochester Institute of Technology
sv6234@rit.edu

Xiajun Jiang
Rochester Institute of Technology
xj7056@rit.edu

Huafeng Liu ✉
Zhejiang University
liuhf@zju.edu.cn

Linwei Wang
Rochester Institute of Technology
linwei.wang@rit.edu

Abstract

The interest in leveraging physics-based inductive bias in deep learning has resulted in recent development of *hybrid deep generative models (hybrid-DGMs)* that integrates known physics-based mathematical expressions in neural generative models. To identify these hybrid-DGMs requires inferring parameters of the physics-based component along with their neural component. The identifiability of these hybrid-DGMs, however, has not yet been theoretically probed or established. How does the existing theory of the un-identifiability of general DGMs apply to hybrid-DGMs? What may be an effective approach to consutrct a hybrid-DGM with theoretically-proven identifiability? This paper provides the first theoretical probe into the identifiability of hybrid-DGMs, and present meta-learning as a novel solution to construct identifiable hybrid-DGMs. On synthetic and real-data benchmarks, we provide strong empirical evidence for the un-identifiability of existing hybrid-DGMs using unconditional priors, and strong identifiability results of the presented meta-formulations of hybrid-DGMs.

1 Introduction

There has been increasing interest in integrating mathematical expressions of known physics with neural-network functions for hybrid, or gray-box, modeling [1, 2, 3, 4, 5]. The incorporation of physics-based inductive bias has the potential to improve the generalizability and interpretability of deep learning [4], while the expressive and flexible neural functions offer the opportunity to fill the gap in our prior knowledge of physics (or *expert models*) [4, 5]. Recent progress in particular has started to see principled developments of hybrid deep generative models (DGM) where the data-generation process is described by a hybridization of physics-based and neural components. Both of these two components are then identified from data via, for instance, amortized variational inference [2, 4]. These hybrid-DGMs have shown benefits of intepretability, generalizability, and out-of-distribution (OoD) robustness, especially when learning to model physics systems from observational data [2, 4]. An important premise of these successes is that the hybrid expression of the physical systems can be accurately identified from the observed data: for example, if we can identify the parameters

---

*Work was done as a visiting exchange student at Rochester Institute of Technology.

and/or physics governing observed pendulum movements, we could then use the identified generative models/parameters for predicting the future trajectory of the pendulum [2].

However, substantial research based on the theory of nonlinear independent component analysis (nonlinear ICA) have established that a naive DGM or latent variable model is un-identifiable [6]. Active research has also reported various ways to construct an identifiable latent variable model, primarily via constructing a conditionally independent generative model [7, 8, 9, 6]. Most of the progress has been established for the identifiability of latent processes underlying time-series data, leveraging the temporal dependency of the latent variables [8] or observed domain index for non-stationary segments [7] to construct the conditional generative model. For static latent variables, the construction of conditionally independent generative model has mainly relied on the introduction of auxiliary observations such as class or domain labels [9].

How does the hybrid expression of a DGM impact its (un)-identifiability? Furthermore, since hybrid-DGMs are often intended to model physics systems with unknown parameters to be identified from data, what may be an effective approach to construct a hybrid-DGM with theoretically-proven identifiability? Unfortunately, despite the increasing interest in hybrid-DGMs, their identifiability remains unexplored. This identifiability however is critical for the intended use of a hybrid DGM, *e.g.,* to predict beyond the observations used to identify the model, or to be robust in out-of-distribution (OOD) generalizations.

Recent efforts in hybrid-DGMs indeed have increasingly noted the challenges in properly learning the two components within the hybrid models, especially in ensuring a non-trivial solution to the physics-based component due to the presence of a highly-expressive neural component [1, 2]. Various strategies have been presented, such as regularizing the expressiveness of the neural component [2], trajectory-based and adaptive optimization to allow automatic complexity adjustment for the neural component [1], and expert-augmentation leveraging the physics-based component to improve generalization to distributional shifts [4]. No existing works, however, have probed or established the identifiability of hybrid-DGMs, theoretically or empirically.

In this work, we present the first theoretical probe of the identifiability of hybrid-DGMs. Importantly, forgoing the need of *observed* auxiliary variable to condition the generative model, we present meta-learning as a novel solution to establish the identifiability of hybrid-DGMs. To this end, we present a learn-to-identify formulation for meta-learning hybrid-DGMs, and theoretically establish the identifiability of these meta-hybrid-DGMs via conditional independence of the DGM given few-shot context samples. We further provide empirical identifiability results of these meta-hybrid-DGMs in a variety of synthetic and real datasets generated from physics systems, in contrast to existing non-meta hybrid-DGMs where identifiability cannot be established without observed auxiliary variables. Finally, we note that while this work focuses on hybrid-DGMs, the presented meta-learning formulation provides a general new condition to construct identifiable DGMs beyond the hybrid setting.

2   Related Works

Hybrid-DGMs: A number of hybrid models have emerged in various domains to combine physics-based functions with neural networks, to compensate for unknown components in the known physics. While earlier approaches [10, 11, 12] focused on using neural networks to learn the residual between the measurements and those obtained from physics-based simulation, tighter integrations have been presented by, *e.g.,* NeuralSim [13] and universal differential equations (UDEs) [14], to include neural networks as different components within a physics-based function. Most of these earlier works are focused on hybrid modeling in specific application domains.

More recently, several hybrid-DGMs have been presented [2, 1, 4, 3], with a focus on properly learning the physics component within the hybrid-DGM so it is not overpowered by the expressive neural component. In physics-integrated hybrid VAE [2], this was achieved by a regularized learning method that controls the expressiveness of the neural component while preserving the semantics of the physics-based component. Similarly, regularization of the expressiveness of the neural component is considered in [1]. In [4], expert augmentation is proposed to improve the OOD robustness of these hybrid-DGMs by fine-tuning on synthetic OOD samples, generated by sampling outside the training distribution of the physics-based component within the trained hybrid model.

7715

Despite these efforts in better learning the two separate components within hybrid-DGMs, the identifiability of hybrid-DGMs has not been theoretically probed nor established. Empirically, no existing works have evaluated these hybrid-DGMs from the lens of identifiability metrics. Theoretically, no existing works have examined the conditions for constructing identifiable hybrid-DGMs. This work will take the first step to bridge these gaps, establish the un-identifiabilty of naive unconditional hybrid-DGMs, and present meta-learning as a novel solution to construct identifiable hybrid-DGMs.

Nonlinear ICA and Identifiability of DGMs: With nonlinear ICA, strong identifiability results have been established for unsupervised latent variable models [7, 8, 9, 6, 15, 16, 17, 18]. Fundamentally, construction of identifiable generative models is achieved by defining independence structure when conditioned on observed auxiliary variables [9]. Much progress has been made in identifying latent dynamic processes generating observed time series, including leveraging conditional independence of the DGM given time segment index temporal dependence [8], sparse temporal encoding [15], and even unobserved inferred domain labels for non-stationary time series [16, 17, 18]. For static latent variables, identifiability is mainly established via the introduction of auxiliary observations such as class or domain labels [9]. In specific, iVAE [6] established the identifiability of DGMs assuming an exponential families conditional distributions given observed auxiliary variables.

The reliance on observed conditioning variable, however, limits the applicability of these identifiability results to hybrid-DGMs: as hybrid-DGMs are intended for modeling physics system with unknown and often continuous parameters, it is not clear what *observed* auxiliary label may be available to construct an identifiable DGM. This work will forego this assumption of observed auxiliary variable and present meta-learning as a novel condition to construct identifiable DGMs. While presented for hybrid-DGMs, this connection between meta-learning and the identfiability of DGMs is a novel contribution to general non-hybrid DGMs that has not been constructed in the existing literature.

3   Problem Formulation: Unidentifiablity of Hybrid-DGMs

Let $\mathbf{x} \in \mathbb{R}^n$ be the observed random variables and $\mathbf{z} \in \mathbb{R}^d$ be the latent variables generating $\mathbf{x}$, with the following generation process:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}), \qquad p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})d\mathbf{z} \tag{1}$$

where $p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$ represents the prior distribution of independent generative factors z_i, $p(\mathbf{x}|\mathbf{z})$ is the likelihood function which can be defined based on the mixing function $\mathbf{x} = \mathcal{F}(\mathbf{z})$.

Construction of Hybrid-DGMs: When constructing a generative model for Equation (1), if we assume a known mathematical expression explaining the mechanism of the mixing function as f_{P}, we will have a physics-based expression of this generation process, with unknown parameters of f_{P} as physics-based latent variables $\mathbf{z}_{\mathrm{P}}$. If we have no prior knowledge, we can describe the mixing function with a neural network f_{N} with abstract representation of the latent variables as $\mathbf{z}_{\mathrm{N}}$ – the latter is the foundation of many successful DGMs including variational autoencoders (VAEs) [19].

When understanding complex systems in many domains, recent years have seen an increasing interest in the middle-ground of the above two scenarios: some prior knowledge in the form of a mathematical expression f_{P} of physics exists about the data observed, yet often inexact with unknown gaps to the actual data-generating mechanisms. This motivated recent developments of hybrid-DGMs where the mixing function $\mathbf{x} = \mathcal{F}(\mathbf{z})$ is described as a combination of a physics-based and neural component, f_{P} and f_{N}, respectively, with corresponding physics-based and abstract latent variables $\mathbf{z}_{\mathrm{P}}$ and $\mathbf{z}_{\mathrm{N}}$.

As a concrete example, suppose that we observe the time-series of the angular position φ of a damped forced pendulum system, $\mathbf{x} = [\varphi_0, \varphi_1, \ldots, \varphi_T]$ where $\varphi_i = \varphi(i\Delta t)$, governed by:

$$\frac{d^2\varphi(t)}{dt^2} + \omega^2 \sin\varphi(t) + \xi\frac{d\varphi(t)}{dt} - A\cos(2\pi\phi t) = 0 \tag{2}$$

where the first two terms describe the physics of an *ideal* pendulum system, the third term the damping effect, and the last term the external force. Parameters $\omega, \xi, A,$ and ϕ specify different systems governed by the same physics, representing the independent latent generative factors.

Suppose that we choose to leverage our prior knowledge about the *ideal* pendulum physics $f_{\mathrm{P}}(\mathbf{x}; \mathbf{z}_{\mathrm{P}}) = \ddot{\varphi} + \mathbf{z}_{\mathrm{P}}^2 \sin\varphi$ with unknown parameter ω. To bridge its potential gap with actual

data-generating physics, we then choose to complete it with a neural-network expression of an ODE $f_{N_\theta}(\mathbf{x}; \mathbf{z}_N)$, parameterized by θ and with an abstract latent variable $\mathbf{z}_N$: while abstract, we hope f_{N_θ} to absorb the effect of the last two terms in the data-generating physics (Equation (2) with $\mathbf{z}_N$ representing its remaining data-generating parameters ξ, A, and ϕ. This gives rise to a hybrid mixing function $\mathbf{x} = \mathcal{F}[f_P, f_{N_\theta}; \mathbf{z}_P, \mathbf{z}_N] = \text{ODEsolve}[f_P(\mathbf{x}; \mathbf{z}_P) + f_{N_\theta}(\mathbf{x}; \mathbf{z}_N) = 0]$.

Assuming Gaussian observation noises, we obtain the likelihood of the Hybrid-DGMs as:
$$p_\theta(\mathbf{x}|\mathbf{z}_P, \mathbf{z}_N) = \mathcal{N}(\mathbf{x}|\mathcal{F}[f_P, f_{N_\theta}; \mathbf{z}_P, \mathbf{z}_N], \Sigma_x) \tag{3}$$
with prior distribution for the latent variables assumed to be Gaussian as:
$$p(\mathbf{z}_P) := \mathcal{N}(\mathbf{z}_P|\mu_P, \sigma_P^2 \mathbf{I}), \qquad p(\mathbf{z}_N) := \mathcal{N}(\mathbf{z}_N|0, \mathbf{I}) \tag{4}$$
where the mean μ_P and variance σ_P^2 for $p(\mathbf{z}_P)$ can be defined by prior knowledge due to its physical meanings. Note that $\mathbf{z}_P$ will be directly interpretable and physically meaningful as they will be semantically grounded to the parameters of the physics model f_P; $\mathbf{z}_N$, in comparison, will be abstract but need to absorb effect of the varying parameters underlying the missing physics f_N.

The data-generating process of the hybrid-DGM with parameter θ can thus be defined as:
$$p_\theta(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N) = p_\theta(\mathbf{x}|\mathbf{z}_P, \mathbf{z}_N)p(\mathbf{z}_P)p(\mathbf{z}_N), \qquad p_\theta(\mathbf{x}) = \int \int p_\theta(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N)d\mathbf{z}_P d\mathbf{z}_N \tag{5}$$

While described in the context of a conrete example, Equations (3)-(5) represent a general expression of hybrid-DGMs where the hybrid-mixing function $\mathbf{x} = \mathcal{F}[f_P, f_{N_\theta}; \mathbf{z}_P; \mathbf{z}_N]$ can be designed in various forms (and not limited to the hybrid-ODE described in the example above).

Unidentifiability of Hybrid-DGMs: Now assume we have access to data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ generated by $p_{\theta^*}(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N)$, where θ^* is the true but unknown data-generating parameter. For an identifiable DGM, our goal is to learn θ such that the likelihood $p_\theta^*(\mathbf{x}|\mathbf{z}_P, \mathbf{z}_N)$, the priors $p(\mathbf{z}_P)$ and $p(\mathbf{z}_N)$, and the posteriors $p_\theta^*(\mathbf{z}_P|\mathbf{x})$ and $p_\theta^*(\mathbf{z}_N|\mathbf{x})$ are all correctly recovered.

In practice, because we only observe $\mathbf{x}$ without access to the latent variables $\mathbf{z}_P$ or $\mathbf{z}_N$, we optimize θ to match the marginal density of $\mathbf{x}$, *e.g.,* via amortized variational inference to maximize the evidence lower bound (ELBO) of $p_\theta(\mathbf{x})$ in VAEs, such that:
$$p_\theta(\mathbf{x}) \approx p_{\theta^*}(\mathbf{x}) \tag{6}$$

Unfortunately, even if the above optimization is done perfectly (*i.e.,* $p_\theta(\mathbf{x}) = p_{\theta^*}(\mathbf{x})$), there is no guarantee that $p_\theta^*(\mathbf{x}|\mathbf{z}_P, \mathbf{z}_N)$, $p_\theta^*(\mathbf{z}_N|\mathbf{x})$, or $p_{\theta^*}(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N)$ are correctly identified. As shown in [6], for these densities to be recovered based on matching $p_\theta(\mathbf{x})$, the DGM needs to satisfy:
$$\forall(\theta, \tilde{\theta}) : p_\theta(\mathbf{x}) = p_{\tilde{\theta}}(\mathbf{x}) \Rightarrow \theta = \tilde{\theta} \tag{7}$$

In another word, the matching of $p_\theta^*(\mathbf{x})$ needs to imply a unique solution of θ^* to be obtained. In reality, however, it has been shown in [6] that DGMs with unconditional prior are unidentifiable. We refer the readers to the important theorems and the proof in [6, 9, 20] for more details, but to what extent is the hybrid-DGM affected by this theory of identifiability?

At a concept level, let us first consider the neural component ($\mathbf{z}_N$ and f_{N_θ}) within the hybrid-DGM. For $\mathbf{z}_N$ of any distributions, there are always nonlinear transformations that change its values but not its distributions [6]. In a highly expressive neural network, this transformation can be learned in the likelihood $p_\theta(\mathbf{x}|\mathbf{z}_N)$. This will give the same fit of $p_\theta(\mathbf{x})$, but non-unique solutions of $p_\theta(\mathbf{x}|\mathbf{z}_N)$ and $p_\theta(\mathbf{z}_N|\mathbf{x})$, resulting in non-identifiable neural component within the hybrid-DGM.

For the physics-based component ($\mathbf{z}_P$ and f_{P_θ}), this theory does not hold because the pre-specified form of f_P can not trivially accommodate any transformations applied to $\mathbf{z}_P$ without affecting the fitting of observations. In practice, however, because both components jointly contribute to $p_\theta(\mathbf{x}|\mathbf{z}_P, \mathbf{z}_N)$ as defined in Equation (3), the optimization of $\mathbf{z}_P$ and f_{P_θ} maybe overpowered by the overly-expressive neural component, rendering trivial solutions to $\mathbf{z}_P$ and f_{P_θ}.

In other words, identification of hybrid-DGMs is inflicted by two fundamental challenges: 1) the theoretical un-identifiability of its neural component, and 2) this un-identifiable but highly expressive component overpowering the identification of the otherwise identifiable physics-based component. While existing works [2, 1, 3] have focused on the second challenge, we address the problem from its fundamentals at the un-identifiability of $\mathbf{z}_N$ and f_{N_θ}: we introduce a novel condition to construct identifiable hybrid-DGMs, and establish their identifiability both theoretically and empirically.

7717

4 Learn-to-Identify Hybrid-DGMs via Meta-Learning

Based on nonlinear ICA theory [9], an identifiable DGM can be constructed by conditionally independent latent variables $p(\mathbf{z}|\mathbf{u})$: in the literature, the conditioning variable $\mathbf{u}$ is often assumed to be additionally-observed side information such as class labels or domain indexes [6]. Here, we show that meta-learning formulations offer an opportunity to construct an identifiable DGM, with $\mathbf{u}$ taking forms of k-shot context samples sharing the same data-generation process as the query samples.

Construction of Identifiable Meta-Hybrid-DGMs: Given is a set of k context samples, $\mathcal{D}^s$, sharing the same data-generation process as query samples $\mathbf{x}$, we define the generation of $\mathbf{x}$ as:

$$p_\phi(\mathbf{x}, \mathbf{z}_\mathrm{P}, \mathbf{z}_\mathrm{N}|\mathcal{D}^s) = p_\theta(\mathbf{x}|\mathbf{z}_\mathrm{P}, \mathbf{z}_\mathrm{N})p(\mathbf{z}_\mathrm{P})p_\zeta(\mathbf{z}_\mathrm{N}|\mathcal{D}^s), \quad p_\phi(\mathbf{x}|\mathcal{D}^s) = \int \int p_\phi(\mathbf{x}, \mathbf{z}_\mathrm{P}, \mathbf{z}_\mathrm{N}|\mathcal{D}^s)d\mathbf{z}_\mathrm{P}d\mathbf{z}_\mathrm{N}$$

(8)

where $\phi = \{\theta, \zeta\}$, $p_\theta(\mathbf{x}|\mathbf{z}_\mathrm{P}, \mathbf{z}_\mathrm{N})$ is defined in Equation (3), and the conditional prior $p_\zeta(\mathbf{z}_N|\mathcal{D}^s)$ is assumed to be factorized exponential family distributions [21]:

$$p_\zeta(\mathbf{z}_N|\mathcal{D}^s) = p_{\mathbf{T}, \boldsymbol{\lambda}_\zeta}(\mathbf{z}_N|\mathcal{D}^s) = \prod_{i=1}^d \frac{Q_i(z_{\mathrm{N},i})}{Z_i(\mathcal{D}^s)} \exp\left[\mathbf{T}_i(z_{\mathrm{N},i})^\top \boldsymbol{\lambda}_{\zeta,i}(\mathcal{D}^s)\right]$$

(9)

where Q_i is the base measure, $Z_i(\mathcal{D}^s)$ is the normalizing constant, $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,e})$ are the sufficient statistics that are fixed (not estimated). $\boldsymbol{\lambda}_i^\zeta(\mathcal{D}^s) = (\lambda_{i,1}^\zeta(\mathcal{D}^s), \ldots, \lambda_{i,e}^\zeta(\mathcal{D}^s))$ are the corresponding parameters conditioning on $\mathcal{D}^s$. In this paper, we realize this conditioning through a composite of a neural network $h_\zeta(\mathbf{x}^s)$, parameterized by ζ, for individual context samples $\mathbf{x}^s \in \mathcal{D}^s$, and an averaging function across all context samples:

$$\boldsymbol{\lambda}_\zeta(\mathcal{D}_s) = \frac{1}{|\mathcal{D}^s|} \sum_{\mathbf{x}^s \in \mathcal{D}^s} h_\zeta(\mathbf{x}^s)$$

(10)

Note that exponential families have universal approximation capabilities, thus this commonly-used assumption is not very restrictive [20].

Amortized Variational Inference: To enable inference over the hybird-DGM in Equation (8), we approximate the posterior density $p(\mathbf{z}_\mathrm{P}|\mathbf{x})$ as $q_\eta(\mathbf{z}_\mathrm{P}|\mathbf{x})$ and $p_\zeta(\mathbf{z}_\mathrm{N}|\mathbf{x}, \mathcal{D}^s)$ as $q_\zeta(\mathbf{z}_\mathrm{N}|\mathbf{x} \cup \mathcal{D}^s)$, the latter realized with the same network as defined in (9) but with the additional input $\mathbf{x}$ in addition to $\mathcal{D}^s$.

Formally, we cast the variational inference into a meta-learning formulation. Consider a dataset $\mathcal{D}$ with M similar but distinct data-generation process: $\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^M$. For each $\mathcal{D}_m$, we consider disjoint context samples $\mathcal{D}_m^s = \{\mathbf{x}^{s,1}, \mathbf{x}^{s,2}, \ldots, \mathbf{x}^{s,k}\}$ and query samples $\mathcal{D}_m^q = \{\mathbf{x}^{q,1}, \mathbf{x}^{q,2}, \ldots, \mathbf{x}^{q,l}\}$, where $k \ll l$. Instead of maximizing the marginal likelihood of $\mathbf{x}$ for all $\mathbf{x} \in \mathcal{D}$, we formulate a meta-objective to learn to maximize the marginal likelihood $p(\mathbf{x}^q|\mathcal{D}_m^s)$ of all query samples $\mathbf{x}^q \in \mathcal{D}_m^q$ when conditioned on support set $\mathcal{D}_m^s$, for all underlying data-generation processes $m \in \{1, 2, \ldots, M\}$:

$$\sum_{m=1}^M \sum_{\mathbf{x}^q \in \mathcal{D}_m^q} \log p(\mathbf{x}^q|\mathcal{D}_m^s) \geq \sum_{m=1}^M \sum_{\mathbf{x}^q \in \mathcal{D}_m^q} \{\mathbb{E}_{q_\eta(\mathbf{z}_\mathrm{P}|\mathbf{x}^q), q_\zeta(\mathbf{z}_\mathrm{N}|\mathbf{x}^q \cup \mathcal{D}_m^s)}[\log p_\theta(\mathbf{x}^q|\mathbf{z}_\mathrm{P}, \mathbf{z}_\mathrm{N})]$$
$$- \mathrm{KL}(q_\eta(\mathbf{z}_\mathrm{P}|\mathbf{x}^q)||p(\mathbf{z}_\mathrm{P})) - \mathrm{KL}(q_\zeta(\mathbf{z}_\mathrm{N}|\mathbf{x}^q \cup \mathcal{D}_m^s)||p_\zeta(\mathbf{z}_\mathrm{N}|\mathcal{D}_m^s))\}$$

(11)

where the first term represents likelihood of the meta-hybrid-DGM on query sample $\mathbf{x}^q \in \mathcal{D}_m^q$, and the last two terms represent Kullback–Leibler divergences between the prior and posterior densities of $\mathbf{z}_\mathrm{P}$ and $\mathbf{z}_\mathrm{N}$, respectively, where densities of $\mathbf{z}_\mathrm{N}$ are conditioned on context samples $\mathcal{D}_m^s$. The maximization of Equation (11) is performed over $\{\eta, \zeta, \theta\}$ in an episodic training where, in each episode, the division of $\mathcal{D}_m^s$ and $\mathcal{D}_m^q$ is shuffled within each $\mathcal{D}_m$. The likelihood is calcuated by the reparameterization trick [19], while the two KL-terms are calculated analytically.

Note that in Equation (11), the meta-hybrid-DGM is formulated on the neural component within the hybrid-DGM to establish its identifiablity. While the identifiability of the physics-based component is not hinged on this construction, in practice, it is possible to extend the conditional prior (Equation

(9)) to $\mathbf{z}_P$ to obtain $p_\zeta(\mathbf{z}_P, \mathbf{z}_N | \mathcal{D}_m^s)$, leading to an alternative meta-formulation:

$$\sum_{m=1}^{M} \sum_{\mathbf{x}^q \in \mathcal{D}_m^q} \log p(\mathbf{x}^q | \mathcal{D}_m^s) \geq \sum_{m=1}^{M} \sum_{\mathbf{x}^q \in \mathcal{D}_m^q} \{ \mathbb{E}_{q_\zeta(\mathbf{z}_P, \mathbf{z}_N | \mathbf{x}^q \cup \mathcal{D}_m^s)} [\log p_\theta(\mathbf{x}^q | \mathbf{z}_P, \mathbf{z}_N)]$$
$$- \mathrm{KL}(q_\zeta(\mathbf{z}_P, \mathbf{z}_N | \mathbf{x}^q \cup \mathcal{D}_m^s) || p_\zeta(\mathbf{z}_P, \mathbf{z}_N | \mathcal{D}_m^s)) \} \tag{12}$$

We will empirically compare the two formulations above, and demonstrate how the identifiability of $\mathbf{z}_P$ and $\mathbf{z}_N$ is each affected by the proposed meta-construction of the hybrid-DGMs.

5 Identifiablility Theory for Meta-Hybrid-DGMs

We consider the meta-formulation a novel condition to ensure the identifiability of unsupervised DGMs. Here, we show that, built on nonlinear ICA and the theory of identifiability for conditional-VAEs established in [6], the presented meta-hybrid-DGMs are identifiable.

For readability, below we collect the formulations of the hybrid-DGM from previous sections:

$$p_\phi(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N | \mathcal{D}^s) = p_\theta(\mathbf{x} | \mathbf{z}_P, \mathbf{z}_N) p(\mathbf{z}_P) p_{\mathbf{T}, \boldsymbol{\lambda}_\zeta}(\mathbf{z}_N | \mathcal{D}^s) \tag{13}$$

$$p_\theta(\mathbf{x} | \mathbf{z}_P, \mathbf{z}_N) = p_\varepsilon(\mathbf{x} - \mathcal{F}[f_P, f_{N_\theta}; \mathbf{z}_P, \mathbf{z}_N]) \tag{14}$$

$$p_{\mathbf{T}, \boldsymbol{\lambda}_\zeta}(\mathbf{z}_N | \mathcal{D}^s) = \prod_{i=1}^{d} \frac{Q_i(z_{N,i})}{Z_i(\mathcal{D}^s)} \exp \left[\mathbf{T}_i(z_{N,i})^\top \boldsymbol{\lambda}_{\zeta,i}(\mathcal{D}^s) \right] \tag{15}$$

where $\phi = (\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$ are model parameters and $\boldsymbol{\lambda}_\zeta$ is defined in Equation (10). Equation (14) gives a more general definition than Equation (3) where ε is an independent noise variable.

Definition 1:   *Let $\sim$ be an equivalence relation on the parameter space Φ. We say the DGM in 1 is $\sim$-identifiable if $p_\phi(\mathbf{x}) = p_{\tilde{\phi}}(\mathbf{x}) \Rightarrow \phi \sim \tilde{\phi}$*

Definition 2:   *Let $\sim$ be the equivalence relation on Φ defined as follows:*

$$(\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta) \sim (\tilde{\theta}, \tilde{\mathbf{T}}, \boldsymbol{\lambda}_{\tilde{\zeta}}) \Leftrightarrow \exists \mathbf{A}, \mathbf{c} : \mathbf{T}(\mathcal{F}_\theta^{-1}(\mathbf{x})) = \mathbf{A}\tilde{\mathbf{T}}(\mathcal{F}_{\tilde{\theta}}^{-1}(\mathbf{x})) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{X} \tag{16}$$

where $\mathbf{A}$ is a $de \times de$ matrix and $\mathbf{c}$ is a vector of dimension de, d being the dimension of $\mathbf{z}_N$ and e the dimensionality of the sufficient statistics $\mathbf{T}_i$. If $\mathbf{A}$ is invertible, we denote this relation by $\sim_A$.

Definition 2 establishes a specific equivalence relation that allows to recover the sufficient statistics of the generative model up to a linear matrix multiplication.

Theorem 1:   *Assume we observe data sampled from $p_\theta(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N | \mathcal{D}^s)$ as defined in Equations (13)-(15) with parameters $\phi = (\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$. Assume the following holds:*

(i) The set $\{\mathbf{x} \in \mathcal{X} | \varphi_\varepsilon(\mathbf{x}) = 0\}$ has measure zero, where φ_ε is the characteristic function of the density p_ε defined in $p_\varepsilon(\mathbf{x} - \mathcal{F}[f_P, f_{N_\theta}; \mathbf{z}_P, \mathbf{z}_N])$.

(ii) The hybrid mixing function $\mathcal{F}_\theta$ is injective.

(iii) The sufficient function $T_{i,j}$ are differentiable almost everywhere, and linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.

(iv) There exist $de + 1$ distinct context sets $\mathcal{D}^{s,0}, \mathcal{D}^{s,1}, \ldots, \mathcal{D}^{s,de}$ such that the $de \times de$ matrix $\mathbf{L}$ defined as follows is invertible:

$$\mathbf{L} = (\boldsymbol{\lambda}(\mathcal{D}^{s,1}) - \boldsymbol{\lambda}(\mathcal{D}^{s,0}), \ldots, \boldsymbol{\lambda}(\mathcal{D}^{s,de}) - \boldsymbol{\lambda}(\mathcal{D}^{s,0})) \tag{17}$$

Then the parameters $\phi = (\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$ are $\sim_A$-identifiable.

The proof for Theorem 1 directly builds on that presented in [6]. Condition (iv) in Theorem 1 establishes that, if the generative factors in the neural component of the hybrid-DGM is d-dimensional each specified with e-dimensional sufficient statistics, we will need observations from a minimum of $de+1$ distinct and independent generation processes in order to identify the hybrid-DGM. Considering the pendulum-system example discussed in Section 3 and Equation (2), since the neural component is generated by three independent system parameters $\xi, A,$ and ϕ ($d = 3$) and assuming Gaussian statistics ($e = 2$), the hybrid-DGMs will be identifiable if we have observations generated from more than $3 \times 2 + 1 = 7$ distinct combination of parameter values for $\omega, \xi, A,$ and ϕ in the true data-generation process (Equation (2)). We will empirically verify this condition in Section 6.

Table 1: Quantitative identifiability metrics for the presented meta-hybrid-VAE in comparison to physics-integrated hybrid-VAE [2] and APHYNITY [1], including MSE and MCC metrics on the latent variables, as well as MSE of generated $\mathbf{x}$ during reconstruction of observed samples (Rec) and prediction of unobserved samples (Pre).

	Forced Damped Pendulum			Advection-Diffusion System		
	APHYNITY	Hybrid-VAE	Meta-Hybrid-VAE	APHYNITY	Hybrid-VAE	Meta-Hybrid-VAE
MSE of $\mathbf{z}_\mathrm{P}$ ↓	6.96(0.01)e-2	4.14(0.29)e-2	**1.59(0.07)e-2**	1.9(0.3)e-2	7.12(5.32)e-4	**1.34(0.75)e-4**
MCC ↑	0.79(0.02)	0.59(0.03)	**0.99(0.00)**	0.98(0.01)	0.94(0.00)	**0.99(0.00)**
MSE of $\mathbf{x}$ (Rec) ↓	5.2(0.01)e-2	**2.66(0.03)e-2**	2.91(0.06)e-2	6.22(0.41)e-3	4.90(0.75)e-3	**2.93(0.22)e-3**
MSE of $\mathbf{x}$ (Pre) ↓	2.37(0.67)	1.74(0.20)	**6.85(1.11)e-2**	8.25(0.58)e1	8.53(5.72)e-2	**5.63(1.34)e-3**
	Double Pendulum			Real Double Pendulum		
	APHYNITY	Hybrid-VAE	Meta-Hybrid-VAE	APHYNITY	Hybrid-VAE	Meta-Hybrid-VAE
MSE of $\mathbf{z}_\mathrm{P}$ ↓	4.58(0.04)e-1	3.97(0.06)e-1	**3.85(0.19)e-1**	/	/	/
MCC ↑	0.50(0.00)	0.51(0.00)	**0.98(0.00)**	/	/	/
MSE of $\mathbf{x}$ (Rec) ↓	4.88(0.00)e-2	4.05(0.25)e-2	**3.83(0.06)e-2**	3.78(0.52)e-2	**2.20(0.14)e-3**	2.67(0.34)e-2
MSE of $\mathbf{x}$ (Pre) ↓	1.29(0.23)e1	5.52(0.06)	**2.80(0.48)e-1**	5.16(0.69)	1.87(0.13)	**2.88(0.34)e-2**

6 Experiments and Results

Data: We considered three simulated and one real-world benchmarks for hybrid-DGMs, including three simulated physics systems of forced damped pendulum [2], advection-diffusion system [2] and double pendulum [4], and one real-world system double pendulum [4]. For each of the three simulated physics systems, we randomly sampled the initial states and parameters of the governing function to generate observations. We refer to the governing functions used for the generation of data as *full physics* functions, and design *partial physics* functions for represent our imperfect prior knowledge about the observed data, reflecting a variety of potentially additive and multiplicative errors – these *partial physics* functions are used as the f_P in the hybrid models. In data generation, we varied the parameters in the components both present and absent in the prior physics. More details on each dataset can be found in Appendix A.

Baselines: We considered 1) physics-integrated hybrid VAE [2] and 2) APHYNITY [1], both using regularization to control the expressiveness of the neural component. Note that while the original formulation of APHYNITY does not have an inference network to identify $\mathbf{z}_\mathrm{P}$ or $\mathbf{z}_\mathrm{N}$ (making it intended for data generated from a single parameter setting), we added an inference network (that shares a similar architecture with the other two models) for fairness of comparison.

Metrics: With a focus on identifiability, we focus on the following metrics: 1) for physics-based latent variable $\mathbf{z}_\mathrm{P}$, we consider mean squared error (MSE) owing to its physical meaning; 2) for abstract latent variable $\mathbf{z}_\mathrm{N}$ modeled in the neural component, we consider the mean correlation coefficient (MCC) between the true data-generating parameters and $\mathbf{z}_\mathrm{N}$ sampled from its learned posterior density. The calculation of MCC follows standard practice in the identifiability literature [6, 9]: to evaluate $\sim_A$-identifiablity, we calculate the weak MCC with the method described in [22], where a high MCC provides evidence of successful identification. We also evaluate a more strong identifiable relation $\sim_P$ with strong MCC and the details can be found in Appendix E.

In addition to metrics on $\mathbf{z}_\mathrm{P}$ and $\mathbf{z}_\mathrm{N}$, we further introduce MSE metrics on the generated $\mathbf{x}$ in two distinct scenarios: 1) *reconstruction MSE* measures how well an observed $\mathbf{x}$ is reconstructed from the identified $\mathbf{z}_\mathrm{P}$ and $\mathbf{z}_\mathrm{N}$, and 2) *prediction MSE* measures how well the identified $\mathbf{z}_\mathrm{P}$ and $\mathbf{z}_\mathrm{N}$ can be used to generate outside the observed $\mathbf{x}$ used to identify them, *e.g.,* for predicting over long time domains, or predicting for a different sample that comes from the same generation process but with different initial conditions. We use these two metrics to showcase that a good reconstruction does not guarantee a good identification, while the ability to use the identified latent variables to predict for different samples (other than observed) may be a surrogate for identifiaiblity.
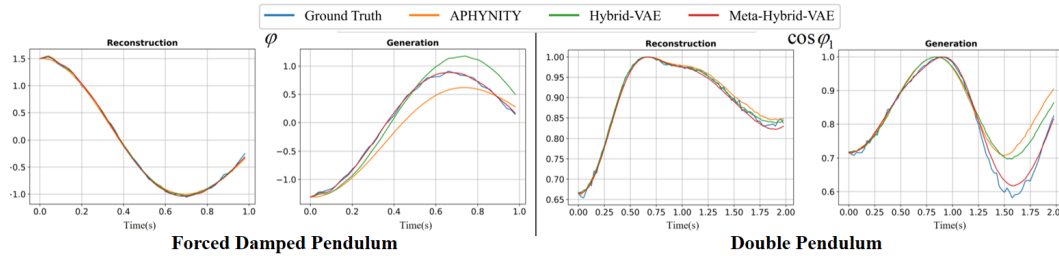
Figure 1: Visual results on synthetic data for reconstruction and prediction performances.

6.1 Results on Synthetic Data

Reconstruction *vs*. Prediction Performance: Table 1 summarizes quantitative identifiability metrics on the presented meta-hybrid-VAE in comparison to its baselines, where the prediction performance of all models is mentioned by their ability – once identified – to predict for a sample sharing the same data-generating parameters but different initial conditions. As shown, meta-hybrid-VAE significantly improved the identifiability of the abstract latent variable z_N (the highest MCC close to 1) and via which, moderately improved the accuracy of the physics-based latent variable z_P (the lowest MSE). Very importantly, while both baseline models identified z_P with a reasonable accuracy, their MCC values were significantly lower – providing strong empirical support for the un-identifiability of the neural component of the hybrid-DGM as theorized in this paper.

Also importantly, all hybrid models achieved comparable reconstruction MSE for observed x's, where the meta-hybrid-VAE did not necessarily produce the lowest MSE. The ability of the identified hybrid-DGM to predict for different samples other than observed, however, varied significantly: while the meta-hybrid-VAE was able to deliver a MSE comparable to the reconstruction task, the two baseline models saw a deterioration of MSE by two magnitudes. Fig. 1 provides visual examples to demonstrate this performance difference between *reconsruction* and *prediction* tasks, stressing the importance to 1) look beyond reconstruction performance for evaluating hybrid-DGMs, and 2) consider prediction of unobserved samples as a potential surrogate for identifiability measures when ground-truth of latent variables is not available (real data settings in Section 6.2).

Interestingly, among the three datasets, the advection-diffusion system appeared to be relatively simple to identify as the baselines did not exhibit as significant a degradation of performance in the more difficult identifiability metrics (MCC and prediction-MSE). This is potentially because the missing component from the prior physics has a small effect. In comparison, the MCC and prediction-MSE of the baseline models were significantly poorer in double pendulum, suggesting a difficult missing physics to be identified. Regardless, meta-hybrid-VAE was able to obtain significant margins of improvements compared to baselines across all datasets.

Predictions over Longer Time Domains: Fig. 2 shows the prediction performance of the three models when predicting the trajectory of pendulum movement beyond the time domain used to identify the model, with a visual sample of forced damped pendulum system. As shown, while all models' performance deteriorated as they predicted outside the time domain used in training, the presented meta-hybrid-VAE demonstrated significantly smaller drop in performance and slower rate of deterioration as the prediction horizon increased. This provides strong evidence regarding the importance of the identifiability, even for the neural component, of a hybrid model for the purpose of predictive tasks.

OOD Performance: Figure 3 (left) summarizes the performance of the three models in settings where either the physics-based component z_P or the neural component z_N was outside the training distribution for forced damped pendulum system. Results demonstrated that, meta-hybrid-VAE obtained the strongest performance in OOD settings for both the identification of the physics and neural components. Details of the OOD parameters can be found in the Appendix A.1.

Empirical Verification of the Condition for Identifiability: Theorem 1-iv) stipulates that the identifiability of the hybrid-DGM depends on the number of independent conditional densities observed, in relation to the dimension of the latent factors to be identified. To experimentally
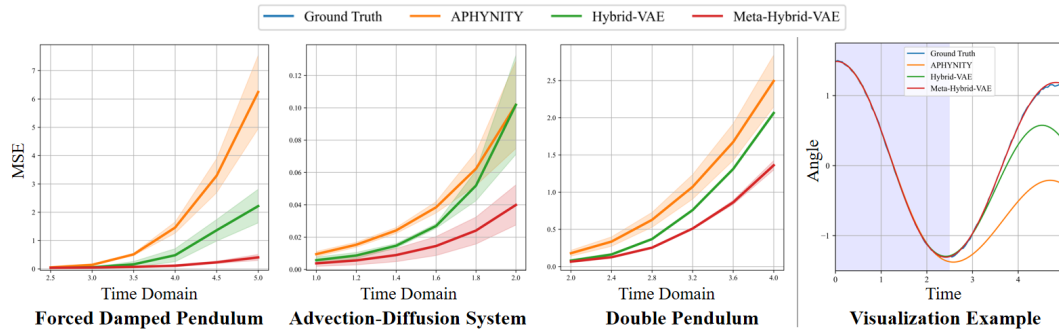
Figure 2: (Left) The performance of predictions over longer time domains; (Right) A visual example of forced damped pendulum system.The dark part is the training time domain, and the light part is the time domain outside training (longer time domain).
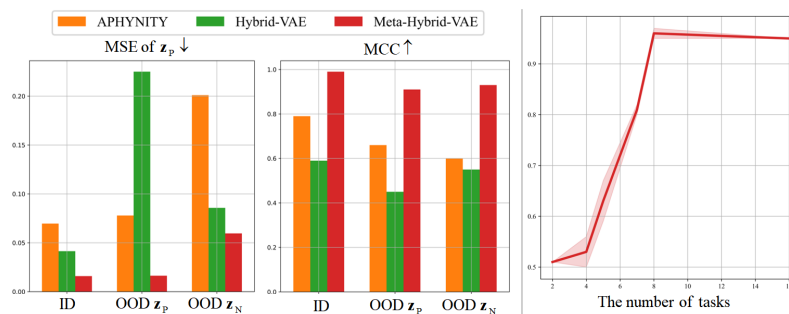

Figure 3: (left) Results of OOD; (right) Results of the condition for identifiability.

verify this, we evaluated the identifiability metric MCC of the hybrid-DGM when learning over data generated from an increasing number of distinct parameter values governing the pendulum system (Equation (2)), *i.e.,* number of tasks in meta-learning. The results in Figure 3 (right) showed that the MCC metric exceeded 0.85 when 7 tasks appeared in the training data and stablized when 8 or more tasks appeared. This agreed with the theorem that, for the pendulum system that had a 3-dimensional parameter vector to be identified in the neural component of the hybrid model with a Gaussian assumption of 2-dimensional sufficient statistics, a minimum of 3*2+1 = 7 distinct "tasks" s needed to identify the system.

Additional Alation Analyses of Meta-Hybrid-VAE: In Appendix C.1, we provided ablation results on the pendulum system to demonstrate that the identifiability results of the presented meta-hybrid-VAE were minimally affected by the number of parameters to be identified, as long as the theoretical condition for identifiability is met. In Appendix C.2, we provided further ablation studies to show that the physics-based component did not suffer from the type of un-identifiabilty discussed in this paper as the neural component, although the meta-formulation did moderately improve the accuracy of the estimation of the physics-based parameters.

6.2 Real Data of Double Pendulum

Experimental settings: We used the dataset of a double pendulum introduced by [4], which contains 21 videos of the pendulum. Each run lasts approximately 40 seconds and is recorded at 400Hz. We extractd the position of the pendulum limbs from each frame with elementary computer-vision tools. We divided each video into observations of $\mathbf{x}$'s to consist of 20 temporal frames with a sampling frequency of 100 Hz. Because there is no clear indication of which samples belong to the same data-generation process, we use 7 samples preceding the current query sample as the context samples. We generated 10000, 3400, and 3400 training, validation, and test samples, respectively.

We adopted Equation (19) as the physics-component of the hybrid-DGM, using the known lengths of the two arms ($L_1 = 91mm$ and $L_2 = 70mm$) and assuming $\tilde{m} = 1$. The total energy of the double
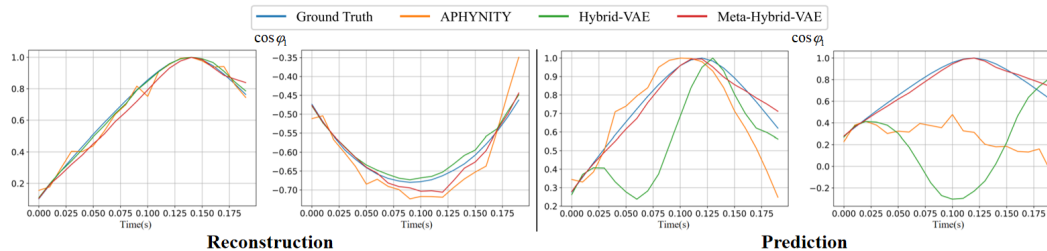
Figure 4: Examples of reconstruction and generation results on real double pendulum data.

pendulum decreases over time in all videos, which is not explained by this physics model. Nor does the physics model consider potential vibrations or errors in extracting the arms' positions, which is expected to be learned by the neural component of the hybrid-DGM with the dimension of $\mathbf{z}_N = 4$.

Results: As ground-truth generative factors are not available in the real data, we relied on prediction MSE as a surrogate for the identifiability metric as informed by results in Sections A.1-A.3. While all models delivered comparable reconstruction MSE, meta-hybrid-VAE was able to achieve a prediction MSE at a level similar to reconstruction while the two baselines experienced a deterioration by 2-3 magnitudes. Visual examples are provided in Fig. 4. This strong performance in real data provides solid evidence for the presented meta-formulation to establish identifiability of hybrid-DGMs.

7   Conclusions & Discussion

In this paper, we probe the un-identifiability of hybrid-DGMs with unconditioned priors, and present meta-learning as a novel solution to construct identifiabile hybrid-DGMs – both results were supported by strong theoretical and empirical evidence. Moving forward, this work can be improved as follows.

Comprehensive ablation studies of the meta-hybrid-VAE are needed to examine the effect of several key hyperparameters, such as the size of k-shot context set and the dimension of $\mathbf{z}_N$. The latter was set to the number of true generative factors in synthetic experiments, and its effect on model performance needs to be further examined. Note that this setting was identical across baselines, thus it had no effect on the significant margins of improvements seen by meta-hybrid-VAE compared to baselines. Future ablation studies should also delve into the benefit of hybrid-DGMs in comparison to purely physics-based and purely neural DGMs, which we considered out of scope of the current study that focused on establishing the identifiability of hybrid-DGMs both theoretically and empirically.

Regarding the generality of the meta-formulation as a solution to identifiable DGMs beyond the hybrid formulation, Appendix D presented an initial investigation on the performance of the presented meta-formulation on a general non-hybrid VAE on a synthetic dataset of non-stationary Gaussian time-series, in comparison to the identifiable-VAE [6] constructed using known class labels to condition the generative model.

Future works could investigate this effect in a broader context. Finally, while the meta-formulation foregoes explicitly observed auxiliary variable, it does assume the ability to pair context and query samples from the same data generation process. This is a common assumption in meta-learning and, in scenarios where this knowledge is not evident, simply pairing by preceding temporal samples was shown to be effective in our real-data experiments. In scenarios where no information is available for pairing, the presented meta-solution will not be applicable.

Acknowledgments and Disclosure of Funding

References

[1] Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

[2] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34:14809–14821, 2021.

[3] Ziming Liu, Yuanqi Du, Yunyue Chen, and Max Tegmark. Physics-augmented learning: A new paradigm beyond physics-informed learning. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[4] Antoine Wehenkel, Jens Behrmann, Hsiang Hsu, Guillermo Sapiro, Gilles Louppe, and Jörn-Henrik Jacobsen. Robust hybrid learning with expert augmentation. *Transactions on Machine Learning Research*, 2023.

[5] Zhaozhi Qian, W. R. Zame, L. M. Fleuren, Elbers P, and M. van der Schaar. Integrating expert ODEs into neural ODEs: Pharmacology and disease progression. In *NeurIPS*, 2021.

[6] 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.

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

[8] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[9] 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.

[10] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

[11] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 36(4):1307–1319, 2020.

[12] Florian Golemo, Adrien Ali Taiga, Aaron Courville, and Pierre-Yves Oudeyer. Sim-to-real transfer with neural-augmented robot simulation. In *Conference on Robot Learning*, pages 817–828. PMLR, 2018.

[13] Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9474–9481. IEEE, 2021.

[14] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

[15] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *ICLR*, 2021.

[16] 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 *NeurIPS*, 2023.

[17] Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. In *NeurIPS*, 2021.

[18] H Halva and A Hyvarinen. Hidden markov nolinear ica: Unsupervised learning from nonstationary time series. In *UAI*, 2020.

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

[20] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyv, Revant Kumar, et al. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

[21] Christopher M Biship. Pattern recognition and machine learning (information science and statistics). *Springer New York*, 2007.

[22] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020.

A   Details on Experimental Data Generation

A.1   Forced Damped Pendulum

We generated data from the *full physics* function in Equation (2), with observed $\mathbf{x} = [\varphi_0, \ldots, \varphi_i, \ldots, \varphi_T], \varphi_i = \varphi(i\Delta t), T = 49, \Delta t = 0.05$. We randomly sampled the initial condition $\varphi_0 \in (-1.57, 1.57)$ (with $\dot{\varphi}_0 = 0$ fixed). Physics parameters $[\omega, \xi, A, \phi]$ to be identified were varied as follows: $\omega \in [0.785, 3.14]$, $\xi \in [0.0, 0.8]$, $A \in [0, 40]$ and $\phi \in [3.14, 6.28]$. We generated total 60000 samples and separated them into a training, validation, and test sets with 4,0000, 10000, and 1,0000 samples, respectively. In all hybrid-DGMs, we considered the first two terms n in Equation (2) as the *partial physics*, thus $\mathbf{z}_P = \omega$ and $\mathbf{z}_N$ with a dimension of 3 is expected to recover $[\xi, A, \phi]$. This setting induces an additive error in the prior physics. In OOD setting, we set OOD $\mathbf{z}_P$ as $\omega \in [3.14, 3.5]$ and OOD $\mathbf{z}_N$ as $\xi \in [0.8, 1.0]$, while keeping the rest of the parameters unchanged.

A.2   Advection-diffusion System

We generate data from the following *full physics* function:

$$\frac{\partial T}{\partial t} - a\frac{\partial^2 T}{\partial s^2} + b\frac{\partial T}{\partial s} = 0 \tag{18}$$

where s is the spatial dimension. The solution $T(s,t)$ was observed as $\mathbf{x} := [\mathbf{T}_0, \ldots, \mathbf{T}_T]$, where $\mathbf{T}_j := [T(0,t_j), \ldots, T(s_{max}, t_j)]^T$ at $t_j := j\Delta t$ is uniformly distributed in space with a dimension of 20. We set $T = 49, \Delta t = 0.02$, the boundary condition as $T(0,t) = T(s_{max}, t) = 0$, and the initial condition as $T(s, 0) = c\sin(\pi s/s_{max})$ with $c \in [0.5, 1.5]$. Physics parameters to be identified $[a, b]$ were sampled as: $a \in [0.01, 0.1]$ and $b \in [0.01, 0.1]$. We generated 20000, 5000, and 5000 training, validation, and test samples, respectively. In all hybrid-DGMs, we set *partial physics* as the first two terms of Equation (18), thus $z_p = a$ and z_N with a dimension of 1 is expected to recover b. This represents an additive error in the prior physics.

A.3   Double Pendulum

We generated data from the following *full physics* function:

$$\frac{d}{dt}\begin{bmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{bmatrix} = \begin{bmatrix} \frac{G\sin\varphi_2\cos(\varphi_1-\varphi_2)-\sin(\varphi_1-\varphi_2)(L_1\dot{\varphi}_1^2\cos(\varphi_1-\varphi_2)+L_2\dot{\varphi}_2^2)-(\tilde{m}+1)G\sin\varphi_1}{L_1(\tilde{m}+\sin(\varphi_1-\varphi_2)^2)} \\ \frac{(\tilde{m}+1)(L_1\dot{\varphi}_1^2\sin(\varphi_1-\varphi_2)-G\sin\varphi_2+G\sin\varphi_1\cos(\varphi_1-\varphi_2))+L_2\dot{\varphi}_2^2\cos(\varphi_1-\varphi_2)\sin(\varphi_1-\varphi_2)}{L_2(\tilde{m}+\sin(\varphi_1-\varphi_2)^2)} \end{bmatrix} \tag{19}$$

where $\tilde{m} = m_1/m_2$. The observed $\mathbf{x} := [(\varphi_{1,0}, \varphi_{2,0}), \ldots, (\varphi_{1,j}, \varphi_{2,j}) \ldots, (\varphi_{1,T}, \varphi_{2,T})]$, where $(\varphi_{1,j}, \varphi_{2,j})$ is the value of a solution at $t_j := j\Delta t$ with $T = 79, \Delta t = 0.025, L_1 = L_2 = 1.0$. We randomly sampled the initial condition $\varphi_{1,0}, \varphi_{2,0} \in (-1.57, 1.57)$ (with $\dot{\varphi}_{1,0} = \dot{\varphi}_{2,0} = 0$). Physical parameters $[G, \tilde{m}]$ to be identified were sampled as follows: $G \in [5.0, 15.0]$ and $\tilde{m} \in [0.5, 1.5]$. We generated 20000, 5000, and 5000 training, validation, and test samples, respectively. In all hybrid-DGMs, we set *partial physics* to be Equation (19) with $\tilde{m} = 1$, thus $z_p = G$ and z_N with a dimension of 1 is expected to recover $\tilde{m}$. This induces a multiplicative error in the prior physics.

B   The proof of Theorem 1

Theorem 1:   *Assume we observe data sampled from* $p_\theta(\mathbf{x}, \mathbf{z}_P, \mathbf{z}_N | \mathcal{D}^s)$ *as defined in Equations* (13)-(15) *with parameters* $\phi = (\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$. *Assume the following holds:*

(i) *The set* $\{\mathbf{x} \in \mathcal{X} | \varphi_\varepsilon(\mathbf{x}) = 0\}$ *has measure zero, where* φ_ε *is the characteristic function of the density* p_ε *defined in* $p_\varepsilon(\mathbf{x} - \mathcal{F}_\theta[f_P, f_{N_\theta}; \mathbf{z}_P, \mathbf{z}_N])$.

(ii) *The hybrid mixing function* $\mathcal{F}_\theta$ *is injective.*

(iii) *The sufficient function* $T_{i,j}$ *are differentiable almost everywhere, and linearly independent on any subset of* $\mathcal{X}$ *of measure greater than zero.*

(iv) *There exist* $de + 1$ *distinct context sets* $\mathcal{D}_0^s, \mathcal{D}_1^s, \ldots, \mathcal{D}_{de}^s$ *such that the* $de \times de$ *matrix* $\mathbf{L}$ *defined as follows is invertible:*

$$\mathbf{L} = (\boldsymbol{\lambda}(\mathcal{D}^{s,1}) - \boldsymbol{\lambda}(\mathcal{D}^{s,0}), \ldots, \boldsymbol{\lambda}(\mathcal{D}^{s,de}) - \boldsymbol{\lambda}(\mathcal{D}^{s,0})) \tag{20}$$

Then the parameters $\phi = (\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$ are $\sim_A$-identifiable.

Proof: This proof mainly refers to [6]. For simplicity, we denote $\mathcal{F}_\theta$ as $\mathbf{f}$, $\mathcal{F}_{\tilde{\theta}}$ as $\tilde{\mathbf{f}}$, $\boldsymbol{\lambda}_\zeta$ as $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_{\tilde{\zeta}}$ as $\tilde{\boldsymbol{\lambda}}$. Suppose we have two sets of parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ and $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ such that $p_{\mathbf{f},\mathbf{T},\boldsymbol{\lambda}}(\mathbf{x}|\mathcal{D}^s) = p_{\tilde{\mathbf{f}},\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{x}|\mathcal{D}^s)$ for all pairs $(\mathbf{x}, \mathcal{D}^s)$. Then:

$$\int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathcal{D}^s) p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} = \int_{\mathcal{Z}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{z}|\mathcal{D}^s) p_{\tilde{\mathbf{f}}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} \tag{21}$$

$$\int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathcal{D}^s) p_\varepsilon(\mathbf{x} - \mathbf{f}(\mathbf{z})) d\mathbf{z} = \int_{\mathcal{Z}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{z}|\mathcal{D}^s) p_\varepsilon(\mathbf{x} - \tilde{\mathbf{f}}(\mathbf{z})) d\mathbf{z} \tag{22}$$

$$\int_{\mathcal{X}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{f}^{-1}(\bar{\mathbf{x}})|\mathcal{D}^s) \mathrm{vol} J_{\mathbf{f}^{-1}}(\bar{\mathbf{x}}) p_\varepsilon(\mathbf{x} - \bar{\mathbf{x}}) d\bar{\mathbf{x}} = \int_{\mathcal{X}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\tilde{\mathbf{f}}^{-1}(\bar{\mathbf{x}})|\mathcal{D}^s) \mathrm{vol} J_{\tilde{\mathbf{f}}^{-1}}(\bar{\mathbf{x}}) p_\varepsilon(\mathbf{x} - \bar{\mathbf{x}}) d\bar{\mathbf{x}} \tag{23}$$

$$\int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathcal{D}^s}(\bar{\mathbf{x}}) p_\varepsilon(\mathbf{x} - \bar{\mathbf{x}}) d\bar{\mathbf{x}} = \int_{\mathbb{R}^d} \tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathcal{D}^s}(\bar{\mathbf{x}}) p_\varepsilon(\mathbf{x} - \bar{\mathbf{x}}) d\bar{\mathbf{x}} \tag{24}$$

$$(\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathcal{D}^s} * p_\varepsilon)(\mathbf{x}) = (\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathcal{D}^s} * p_\varepsilon)(\mathbf{x}) \tag{25}$$

$$F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathcal{D}^s}](\omega) \varphi_\varepsilon(\omega) = F[\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathcal{D}^s}](\omega) \varphi_\varepsilon(\omega) \tag{26}$$

$$F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathcal{D}^s}](\omega) = F[\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathcal{D}^s}](\omega) \tag{27}$$

$$\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathcal{D}^s}(\mathbf{x}) = \tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathcal{D}^s}(\mathbf{x}) \tag{28}$$

By taking the logarithm on both sides of equation (28) and replacing $p_{\mathbf{T},\boldsymbol{\lambda}}$ by its expression from (9), we get:

$$\begin{aligned}
&\log \mathrm{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) + \sum_{i=1}^n (\log Q_i(f_i^{-1}(\mathbf{x})) - \log Z_i(\mathcal{D}^s)) + \sum_{j=1}^k T_{i,j}(f_i^{-1}(\mathbf{x}) \lambda_{i,j}(\mathcal{D}^s)) = \\
&\log \mathrm{vol} J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{x}) + \sum_{i=1}^n (\log \tilde{Q}_i(\tilde{f}_i^{-1}(\mathbf{x})) - \log \tilde{Z}_i(\mathcal{D}^s)) + \sum_{j=1}^k \tilde{T}_{i,j}(\tilde{f}_i^{-1}(\mathbf{x}) \tilde{\lambda}_{i,j}(\mathcal{D}^s))
\end{aligned} \tag{29}$$

Let $\mathcal{D}_0^s, \mathcal{D}_1^s, \ldots, \mathcal{D}_{de}^s$ be the context sets provided by assumption (iv) of the Theorem, and define $\bar{\boldsymbol{\lambda}}(\mathcal{D}_l^s) = \boldsymbol{\lambda}(\mathcal{D}_l^s) - \boldsymbol{\lambda}(\mathcal{D}_0^s)$. We plug each of those $\mathcal{D}_l^s$ in (29) to obtain $de + 1$ such equations. We subtract the first equation for $\mathcal{D}_0^s$ from the remaining de equations to get for $l = 1, \ldots, de$:

$$\langle \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})), \bar{\boldsymbol{\lambda}}(\mathcal{D}_l^s) \rangle + \sum_i \log \frac{Z_i(\mathcal{D}_0^s)}{Z_i(\mathcal{D}_l^s)} = \langle \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})), \bar{\tilde{\boldsymbol{\lambda}}}(\mathcal{D}_l^s) \rangle + \sum_i \log \frac{\tilde{Z}_i(\mathcal{D}_0^s)}{\tilde{Z}_i(\mathcal{D}_l^s)} \tag{30}$$

Let L be the matrix defined in assumption (iv), and $\tilde{L}$ similarly defined for $\tilde{\boldsymbol{\lambda}}$. Define $b_l = \sum_i \log \frac{\tilde{Z}_i(\mathcal{D}_0^s) Z_i(\mathcal{D}_l^s)}{Z_i(\mathcal{D}_0^s) \tilde{Z}_i(\mathcal{D}_l^s)}$ and $\mathbf{b}$ the vector of all b_l for $l = 1, \ldots, de$. Experssing equation (30) for all points $\mathbf{u}_l$ in matrix form, we get:

$$L^T \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \tilde{L}^T \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{b} \tag{31}$$

We multiply both sides by the transpose of the inverse of L^T from the left to find:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{A} \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c} \tag{32}$$

where $\mathbf{A} = L^{-T} \tilde{L}$ and $\mathbf{c} = L^{-T} \mathbf{b}$.

Now by the definition of $\mathbf{T}$ and according to assumption (iii), its Jacobian exists and is an $de \times d$ matrix of rank d. This implies that the Jacobian of $\tilde{\mathbf{T}} \circ \tilde{\mathbf{f}}^{-1}$ exists and is of rank d and so is A. We distinguish two cases:
(1) If $e = 1$, then this means that A is invertible (becuase A is $d \times d$).
(2) If $e > 1$, define $\bar{\mathbf{x}} = \mathbf{f}^{-1}(\mathbf{x})$ and $\mathbf{T}_i(\bar{\mathbf{x}}_i) = (\mathbf{T}_{i,1}(\bar{\mathbf{x}}_i), \ldots, \mathbf{T}_{i,e}(\bar{\mathbf{x}}_i))$. According to Lemma 3 in [6], for each $i \in [1, \ldots, d]$ there exist e points $\bar{\mathbf{x}}_i^1, \ldots, \bar{\mathbf{x}}_i^e$ such that $(\mathbf{T}_i'(\bar{\mathbf{x}}_i^1), \ldots, \mathbf{T}_i'(\bar{\mathbf{x}}_i^e))$ are linearly independent. Collect those points into e vectors $(\bar{\mathbf{x}}^1, \ldots, \bar{\mathbf{x}}^e)$, and concatenate the e Jacobians $\mathbf{J}_\mathbf{T}(\bar{\mathbf{x}}^l)$

Table 2: The results of the different dimensions of unknown parameters

The number of parameters	MSE of $\mathbf{z}_P$ ↓	MCC ↑	MSE of $\mathbf{x}$ (Rec)	MSE of $\mathbf{x}$ (Pre)
2	1.22(0.08)e-2	0.97(0.01)	1.22(0.08)e-2	2.02(0.13)e-2
3	3.62(0.01)e-2	0.99(0.00)	1.29(0.05)e-2	5.01(0.00)e-2
4	1.59(0.07)e-2	0.99(0.00)	2.91(0.06)e-2	6.85(1.11)e-2

Table 3: Comparison results of Non-meta and Meta

Method	MSE of $\mathbf{z}_P$ ↓	MCC ↑	MSE of $\mathbf{x}$ (Rec)	MSE of $\mathbf{x}$ (Pre)
Non-Meta $\mathbf{z}_P$	9.17(1.33)e-3	0.99(0.00)	3.28(0.17)e-2	8.82(2.03)e-2
Meta $\mathbf{z}_P$	5.63(0.42)e-3	0.99(0.00)	3.18(0.19)e-2	3.37(0.00)e-2

evaluated at each of those vectors horizontally into the matrix $Q = (\mathbf{J_T}(\bar{\mathbf{x}}^1), \ldots, \mathbf{J_T}(\bar{\mathbf{x}}^e))$ (and similarly define $\tilde{Q}$ as the concatenation of the Jacobians of $\tilde{\mathbf{T}}(\mathbf{f}^{-1}(\tilde{\mathbf{f}}(\bar{\mathbf{x}})))$ evaluated at those points). Then the matrix Q is invertible (through a combination of Lemma 3 and the fact that each component of $\tilde{\mathbf{T}}$ is univariate). By differentiating equation (32) for each $\mathbf{x}^l$, we get (in matrix form): $Q = \mathbf{A}\tilde{Q}$. The invertibility of Q implies the invertibility of $\mathbf{A}$ and $\tilde{Q}$.

Hence, equation (32) and the invertibility of $\mathbf{A}$ mean that $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim_A (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$

C Alation Analyses of Meta-Hybrid-VAE

C.1 Effect of the Dimension of Unknown Parameters

Here, we provided ablation results on the pendulum system to demonstrate that the identifiability results of the presented meta-hybrid-VAE were minimally affected by the number of parameters to be identified, as long as the theoretical condition for identifiability is met. The results can be seen in Table 2.

C.2 Meta-component on cp

Here, we provided further ablation studies to show that the physics-based component did not suffer from the type of un-identifiabilty discussed in this paper as the neural component, although the meta-formulation did moderately improve the accuracy of the estimation of the physics-based parameters. The results can be seen in Table 3.

D Results on General Non-Hybrid VAEs

We compare our meta-formulation of the non-hybrid VAE (Meta-VAE) compared to the original VAE and identifiable VAE (iVAE) on a synthetic datasets about non-stationary Gaussian time-series in [6]. To evaluate the performance of the method, we compute the MCC between the original sources and the corresponding latent samples from the learned posterior. The results can be seen in Table 4.

Table 4: The results on General Non-Hybrid VAEs

	VAE	iVAE	Meta-VAE
MCC	0.67(0.04)	0.91(0.03)	0.88(0.00)

E Strong MCC results

We also evaluate proposed model on a more strong identifiable relation $\sim_P$ defined as follow:

Definition 3: *Let $\sim$ be the equivalence relation on Φ defined the same as 16:*

$$(\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta) \sim (\tilde{\theta}, \tilde{\mathbf{T}}, \boldsymbol{\lambda}_{\tilde{\zeta}}) \Leftrightarrow \exists \mathbf{A}, \mathbf{c} : \mathbf{T}(\mathcal{F}_\theta^{-1}(\mathbf{x})) = \mathbf{A}\tilde{\mathbf{T}}(\mathcal{F}_{\tilde{\theta}}^{-1}(\mathbf{x})) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{X} \quad (33)$$
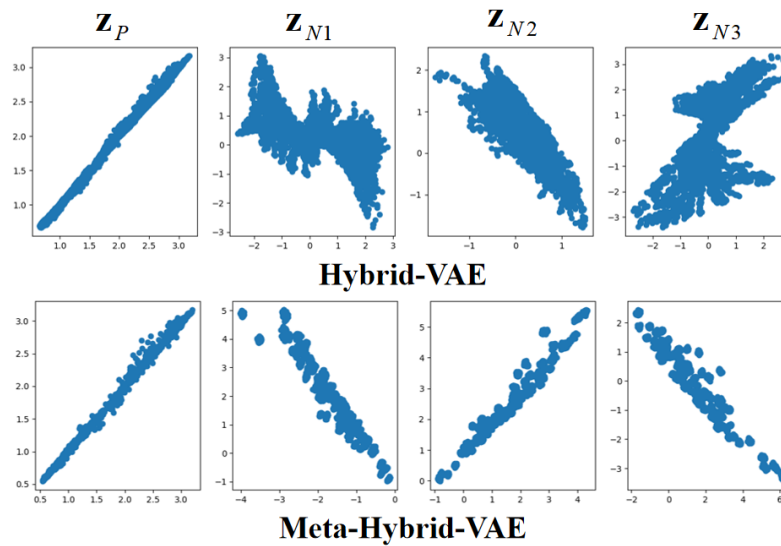
Figure 5: Visual samples of learned latent variables between different runs.

Table 5: The results of strong MCC

	Forced Damped Pendulum		Advection-Diffusion System		Double Pendulum	
	Strong MCC Ground Truth	Strong MCC Different Runs	Strong MCC Ground Truth	Strong MCC Different Runs	Strong MCC Ground Truth	Strong MCC Different Runs
APHYNITY	0.38(0.02)	0.81(0.08)	0.86(0.05)	0.83(0.04)	0.50(0.00)	0.73(0.00)
Hybrid-VAE	0.29(0.00)	0.80(0.01)	0.80(0.12)	0.79(0.11)	0.50(0.00)	0.72(0.00)
Meta-Hybrid-VAE	**0.55(0.02)**	**0.90(0.06)**	**0.97(0.03)**	**0.96(0.00)**	**0.96(0.00)**	**1.00(0.00)**

If $\mathbf{A}$ is a block permutation matrix, we denote this relation by $\sim_P$.

The theory according to $\sim_P$ is following:

Theorem 2: *($e \geq 2$) Assume the hypotheses of Theorem 1 hold and that $e \geq 2$. Further assume:*

(2.i) The sufficient statistics $T_{i,j}$ in Equation 15 are twice differentiable.

(2.ii) The mixing function $\mathcal{F}_\theta$ has all second order cross derivatives.

then the parameters $(\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$ are $\sim_P$-identifiable.

Theorem 3: *($e = 1$) Assume the hypotheses of Theorem 1 hold and that $e = 1$. Further assume:*

(3.i) The sufficient statistics $T_{i,1}$ not monotonic.

(3.ii) All partial derivatives of $\mathcal{F}_\theta$ are continuous.

then the parameters $(\theta, \mathbf{T}, \boldsymbol{\lambda}_\zeta)$ are $\sim_P$-identifiable.

The proof of Theorem 2 and 3 can be found in [6]. Then to evaluate the performance on $\sim_P$, we calculate the strong MCC between learned latent variables and ground truth variables (Strong MCC Ground Truth) and strong MCC between runs of the model (Strong MCC Different Runs) [6, 9, 22]. The results can be found in Table 5 and we provide a visualization example of forced damped pendulum system to show the relation of learned latent variables between two different runs.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our method and experiments have show our main claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have discussed it on Section 7.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: The proofs of the theorem in the main paper can be found in Appendix B .

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We have provided enough information to reproduce the results in Section 6 and Appendix A.

Guidelines:

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

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

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

Answer: [NA]

Justification: The data and code will be released soon.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: We provide experimental settings and details on Section 6.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: We provide the standard deviation of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

Answer: [No]

Justification:

Guidelines:

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: Yes

Guidelines:

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

10. **Broader Impacts**

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

Answer: [No]

Justification:

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [NA]

Justification:

Guidelines:

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

13. **New Assets**

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

Answer: [NA]

Justification:

Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

Answer:[NA]

Justification:

Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [NA]

Justification:

Guidelines:

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