Data-Centric Engineering (2025), 6: e50
doi:10.1017/dce.2025.10028

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Adversarial disentanglement by backpropagation with physics-informed variational autoencoder

Ioannis Christoforos Koune[1] and Alice Cicirello[2]

[1]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands
[2]Department of Engineering, University of Cambridge, Cambridge, UK
Corresponding author: Ioannis Christoforos Koune; Email: i.c.koune@tudelft.nl

Received: 05 July 2025; **Revised:** 21 September 2025; **Accepted:** 26 September 2025

Keywords: generative models; physics-informed machine learning; representation learning; structural health monitoring; variational autoencoders

Abstract

Inference and prediction under partial knowledge of a physical system is challenging, particularly when multiple confounding sources influence the measured response. Explicitly accounting for these influences in physics-based models is often infeasible due to epistemic uncertainty, cost, or time constraints, resulting in models that fail to accurately describe the behavior of the system. On the other hand, data-driven machine learning models such as variational autoencoders are not guaranteed to identify a parsimonious representation. As a result, they can suffer from poor generalization performance and reconstruction accuracy in the regime of limited and noisy data. We propose a physics-informed variational autoencoder architecture that combines the interpretability of physics-based models with the flexibility of data-driven models. To promote disentanglement of the known physics and confounding influences, the latent space is partitioned into physically meaningful variables that parametrize a physics-based model, and data-driven variables that capture variability in the domain and class of the physical system. The encoder is coupled with a decoder that integrates physics-based and data-driven components, and constrained by an adversarial training objective that prevents the data-driven components from overriding the known physics, ensuring that the physics-grounded latent variables remain interpretable. We demonstrate that the model is able to disentangle features of the input signal and separate the known physics from confounding influences using supervision in the form of class and domain observables. The model is evaluated on a series of synthetic case studies relevant to engineering structures, demonstrating the feasibility of the proposed approach.

Impact Statement

Models of complex physical systems, such as those encountered in structural health monitoring, typically fall under either the physics-based or data-driven paradigms. The former are often constrained by limited domain knowledge, while the latter can produce unrealistic predictions that are inconsistent with the known physical laws that govern the system. Hybrid approaches that integrate both physics-based and data-driven components face a trade-off between interpretability and flexibility. In variational autoencoders, which is the main focus of this paper, flexible data-driven components in the decoder can override the known physics, resulting in poor performance and loss of the physical meaning of the latent variables. This work contributes to the integration

This research article was awarded Open Materials badge for transparent practices. See the Data Availability Statement for details.

of domain knowledge with machine learning for hybrid modeling of engineering systems, by proposing an approach that aims to preserve the interpretability of physically meaningful latent variables while accounting for confounding influences in a data-driven manner.

1. Introduction

The aim of this work is to propose and evaluate an approach for learning disentangled representations of the underlying generative factors that characterize the behavior of an engineering system, of particular relevance for the monitoring of civil and mechanical structures. The proposed approach aims to identify and attribute variability observed in *response measurements* obtained from an engineering system to variability stemming from the *modeled physics*, *domain*, and *class* influences. We define the domain as the environmental and operational conditions that a system is exposed to, as well as other properties of the system that may not be directly specified in the model of the known physics. The class is defined as the characteristics of a structure related to the existence and extent of damage and degradation. Generally, we assume that domain information is relatively cheap and easy to collect, compared to class information. Such situations often arise when investigation by experts, costly equipment or elaborate experimental procedures are required to obtain measurements of the class variables. It is important to note that, although we view this problem from the perspective of civil and mechanical structural engineering systems, the approach described in this work can be adapted to other settings.

Our objective is to accurately infer a posterior distribution over physically meaningful latent variables, to reconstruct the structural response, quantify the associated uncertainty, and predict the damage and degradation condition of the system in previously unseen conditions. This is achieved using a limited number of noisy measurements of the structural response, domain and class variables. Due to the influence of the domain, class, and other unknown confounding factors, this will generally be an ill-posed inverse problem that requires learning a *disentangled representation* (Bengio et al., 2014) of the different generative factors. This task is further complicated by the limitations of physics-based models, which often represent structures under idealized nominal conditions and disregard the influence of environmental and operational variability, damage, and degradation. Most computational models of physical systems will contain simplifications and approximations due to lack of knowledge about certain aspects of the underlying physical process and to ensure computational tractability. Reducing this epistemic uncertainty is often infeasible due to cost or time constraints. As a result, only a partial description of the physical system is available in practical applications.

Generative probabilistic models such as variational autoencoders (VAE) (Kingma and Welling, 2022), normalizing flows (Rezende and Mohamed, 2016), and generative adversarial networks (GANs) (Goodfellow et al., 2014), are a class of models that employ deep learning architectures to approximate the distribution of a given set of data and generate samples from the learned distribution. Generative probabilistic models have recently seen broader use in structural health monitoring (SHM), and for constructing digital twins of structures (Bacsa et al., 2025; Coraça et al., 2023; Mao and Wang, 2021; Tsialiamanis et al., 2021). We propose a VAE architecture for approximating the joint distribution between the structural response and a set of physics-grounded latent variables, while accounting and correcting for the confounding influence of the domain and class of the structure, by leveraging observed domain and class variables. To achieve this, the VAE components are split into physics-based and data-driven branches, trained simultaneously in an end-to-end fashion. The data-driven branches are tasked with extracting features of the response that are informative about the domain and class variables, encoding them into the corresponding latent space, and using the latent code to augment the physics-based model predictions. Formulating the VAE as a combination of physics-based and data-driven components is not a straightforward task. The flexibility and learning capacity of feed-forward neural networks (NNs) that enables them to accurately model physical processes from data can be problematic when combining them with physics-based models, as the flexible NN components tend to override the known physics (Takeishi and Kalousis, 2021), resulting in inaccurate inference and overconfident or unrealistic predictions. To

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

address this issue we propose an adversarial training objective that encourages an interpretable and parsimonious representation of the physical system by constraining the data-driven components of the VAE. Once trained, the model can be used to simultaneously perform inference over physically meaningful latent variables for new measurements as they become available, and generate samples from the predictive distribution of the response. Given a set of response measurements, the trained model can also be used to predict the corresponding domain and class variables. The proposed approach aims to:

- Constrain the data-driven components of the model to avoid overriding the known physics and ground a subset of the latent variables to physically meaningful and interpretable quantities;
- Promote the learning of disentangled representations of the physics, domain and class generative factors, that are maximally informative about their corresponding modality while being minimally informative about other modalities.
- Infer unknown non-linear relationships between features in the response measurements and additional domain and class observables that can not be directly included in the physics-based model.
- Improve uncertainty quantification by preventing the data-driven components of the decoder from compensating for all discrepancies between the physics-based model prediction and the measured response.

To achieve these goals we investigate disentangled and invariant representation learning as a tool for regularizing machine learning components in VAE and properly utilizing the known physics, specified in terms of a nominal physics-based model. Additionally, we qualitatively and quantitatively evaluate the accuracy of the predictions and the complexity of the learned representation. The proposed model is assessed on three synthetic case studies and compared with fully data-driven approaches in a damage identification task.

2. Background

This section aims to clarify the terminology and notation used throughout this text, summarize the necessary background, and illustrate the challenges that the proposed approach aims to address. In what follows, bold capital and lower case symbols denote matrix and vector quantities respectively. Light symbols denote scalars. Latent variables that are not directly observed and must be inferred from data are denoted as z, while ϕ, θ and ψ denote encoder, decoder and auxiliary regressor/classifier parameters, respectively. The symbols x, c and y denote the response, domain, and class observables, jointly referred to as the *modalities* of a given physical system. When used as a subscript these symbols denote quantities that belong to a particular modality. As an example, $\boldsymbol{z}_{\mathrm{y}}$ denotes a set of latent variables that encode information about the class of a physical system. Throughout this text, $\mathcal{N}$ denotes the univariate or multivariate normal distribution parametrized by the mean and a scalar variance or matrix covariance respectively, and $\mathcal{U}$ denotes the uniform distribution parametrized by the lower and upper bound. The expectation of a function $f(\cdot)$ over a distribution $p(\cdot)$ is denoted as $\mathbb{E}_{p(\cdot)}[f(\cdot)]$. Finally, a distinction is made between the underlying generative factors s that determine the characteristics of the observed data, and the latent variables z, i.e. the learned representation of the generative factors.

2.1. Problem setting

Suppose that a nominal physics-based model and a dataset $\mathcal{D} = \{(\boldsymbol{x}_{\mathrm{i}}, \boldsymbol{y}_{\mathrm{i}}, \boldsymbol{c}_{\mathrm{i}})\}_{\mathrm{i}=1}^{N}$, composed of N triplets of response measurements $\boldsymbol{x}_{\mathrm{i}}$, domain variables $\boldsymbol{c}_{\mathrm{i}}$ and class variables $\boldsymbol{y}_{\mathrm{i}}$ are available for a given system under investigation. In structural and mechanical engineering applications, the response measurements will often be displacements, strains or accelerations, measured under operating conditions, that describe the static or dynamic performance of the system. The domain variables $\boldsymbol{c}_i$ can be measurements of environmental and operational parameters, such as the location, temperature, humidity or other properties of a structure or sensor. The class variables $\boldsymbol{y}_i$ describe properties of the system that are cumbersome to

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

obtain, such as assessments of the health condition of one or more structural components performed by experts or extracted from inspection reports. It is assumed that $\boldsymbol{y}$ is a quantity of interest to be predicted for new incoming observations of $\boldsymbol{x}$ and $\boldsymbol{c}$. Our goal is to simultaneously perform reconstruction of $\boldsymbol{x}$ and regression or classification on $\boldsymbol{y}$, and furthermore to utilize the observed domain and class variables to account for the impact of the domain and class influences on the measured response.

To highlight the intended application setting, three examples are presented in Figure 1 consisting of a beam, an oscillator, and population of bridges. In each example the available physics-based model fails to account for domain and class influences that are present in the measured response signals. In all three examples, the domain and class observables provide valuable information about the system that is necessary for accurately inferring a distribution over parameters of the physics-based model and reconstructing the response of the physical system. Furthermore, the class observables are significantly harder to measure and will not be available for future experiments. The oscillator and bridge examples also include unknown confounding influences, for which neither observations nor a physical description are available. The three examples are summarized below:

(a) **Beam:** In this example, the available physics-based model of a beam assumes simply supported boundary conditions and a point load acting on an unknown position. Noisy measurements of the displacement field are obtained from a set of sensors, equally spaced along the length of the beam. The domain influence is introduced as a dependence of the rotational stiffness of the right support on the temperature, and the class influence is taken as damage causing a reduction in the vertical stiffness of the right support that varies between experiments.

(b) **Oscillator:** Noisy displacement time-series are obtained from multiple experiments, where a mass-spring-dashpot system is deflected from the equilibrium position and released to perform a harmonic oscillation. The impact of damping on the motion is neglected in the physics-based model and must be inferred from additional observations of properties of the medium. The spring stiffness is taken to depend on the ambient temperature, and there is unknown variability in the initial displaced position. The damping, temperature and variability in initial position are taken as class, domain and unknown confounding influences, respectively.

(c) **Population of bridges:** A vehicle is used to excite the response of a large set of bridges belonging to a homogeneous population, with uncertain vertical stiffness of the supports and varying position of the central pier. Each bridge is monitored by a point strain gauge that yields an influence line for the moving load. The strain gauge measurements are supplemented by qualitative assessments of the condition of the deck, obtained during inspections performed by experts and considered as class variables. The variability

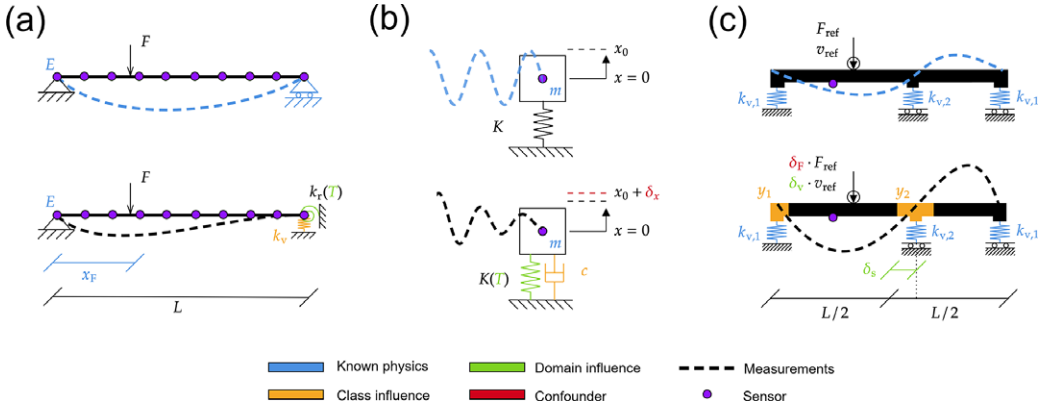


Figure 1. *Illustrative examples of the problem setting: a) Beam, b) Oscillator, and c) One member of a population of bridges. The objective is to learn components of the measured response (bottom row) that are not explicitly included in the nominal physics-based model (top row) using observations of related quantities.*

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

in the vehicle velocity and the existence of deterioration in the deck have an influence on the measured strain but are deemed too complicated to model, while the position of the pier is a known domain parameter and can be included in the physics-based model. A variability in the vehicle load is considered as an unknown confounding influence.

2.2. Epistemic uncertainty in Bayesian model updating

Uncertainty in the modeling of physical systems can generally be classified as either aleatoric or epistemic. Aleatoric uncertainty is the component of the uncertainty due to the inherent randomness of a physical process that can not be reduced, while epistemic uncertainty stems from lack of knowledge regarding the physical process (Kiureghian and Ditlevsen, 2009). Epistemic uncertainty is always present to some degree in practical applications, either due to lack of knowledge or due to simplifications and approximations used to make the evaluation of the physics-based model computationally tractable. The reader is referred to Kamariotis et al., 2024 for an extended overview on classification and treatment of uncertainties in SHM applications.

Suppose that an analytical or numerical physics-based model of a structure, defined as a function $f(\boldsymbol{z}_\mathrm{x})$ is available. The response measurements $\boldsymbol{x}$ can then be expressed as $\boldsymbol{x} = f(\boldsymbol{z}_\mathrm{x}) + \epsilon$, where ϵ is a realization of a random variable quantifying the discrepancy between $f(\boldsymbol{z}_\mathrm{x})$ and $\boldsymbol{x}$ due to the combined influence of aleatoric and epistemic uncertainties. In the Bayesian model updating framework, the measured response of a physical system is used to update the prior knowledge, expressed in terms of a prior distribution $p(\boldsymbol{z}_\mathrm{x})$ over physically meaningful latent variables $\boldsymbol{z}_\mathrm{x}$. This is achieved by approximating the data generating process (i.e. the real-world process that generated the observations) as a combination of a deterministic physics-based model and a probabilistic model (Kennedy and O'Hagan, 2001), where the latter accounts for the combined influence of epistemic and aleatoric uncertainty. In this work, it is assumed that the epistemic uncertainty stems from the confounding influence of the domain and class of a structure, and our inability (e.g. due to cost or time constraints) to account for these influences in the form and parameters of the physics-based model.

2.3. The variational autoencoder

In practical applications, the available domain knowledge is often not sufficient to guarantee that the coupled probabilistic-physical model is an accurate description of the data generating process, limiting the applicability of physics-based modeling. To remedy the lack of domain knowledge, data-driven models based on machine learning techniques have emerged as an alternative to physics-based models, where the unknown physical process is learned from measurements using flexible parametrized approximations. VAE (Kingma and Welling, 2019, 2022) are a popular data-driven approach for learning a joint distribution of data and the latent variables that are assumed to have generated the data using amortized variational inference (VI) (Blei et al., 2017). In VAE, the per-datapoint posterior distribution is approximated using a parametrized family of distributions, where the optimal parameters are obtained by minimizing the Kullback–Leibler divergence (KLD) between the true and approximate posteriors. The VAE is composed of an encoder network $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and a decoder network $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, parametrized by ϕ and θ, respectively, where $\boldsymbol{z}$ denotes latent variables that can not be observed directly and must be inferred from measurements. The encoder is typically implemented as a feed-forward NN that maps the inputs $\boldsymbol{x}$ to a conditional density over latent variables $\boldsymbol{z}$. The decoder network $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ works in the opposite direction by approximating the density of $\boldsymbol{x}$ conditioned on $\boldsymbol{z}$. The training process for VAE consists of simultaneously optimizing the parameters of the decoder that reconstructs the observations given samples of the latent variables, and the encoder that maps inputs to a posterior distribution over these latent variables. Optimization is performed by maximizing a lower bound on the marginal likelihood of the data known as the Evidence Lower BOund (ELBO), denoted as $\mathcal{L}_\mathrm{VAE}$ in Equation (2.1). Sampling $z \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and evaluating the decoder yields samples from the learned distribution of the data, which in the context of civil and mechanical structural systems can be used for downstream tasks such as remaining useful life assessment.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p_{\boldsymbol{\theta}}(\boldsymbol{z})\right) \\ &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right) \\ &\leq \log p_{\boldsymbol{\theta}}(\boldsymbol{x})\end{aligned} \tag{2.1}$$

While data-driven approaches might excel in accurately predicting the response of a physical system for a given set of input parameters when sufficient training data is available, the resulting models are typically black boxes that lack interpretability and yield no useful insights about the underlying physical process that generated the measurements. In cases where both the domain knowledge and the available data are limited, purely physics-based or data-driven approaches become infeasible, necessitating a compromise between the two extremes. Physics-enhanced machine learning (PEML) encompasses a wide range of approaches that combine machine learning with domain knowledge (Cicirello, 2024; Cross et al., 2022; Haywood-AIexander et al., 2024; von Rueden et al., 2023). In this paradigm, the available domain knowledge can be supplemented with data, resulting in more accurate and interpretable models than would be possible with either domain knowledge or data alone. PEML approaches have the potential to reduce the required amount of data, improve accuracy and generalization performance and ensure that model predictions are consistent with the known physics. Importantly, incorporating the known physics can yield interpretable representations of physically meaningful quantities, and models that are robust and explainable.

2.4. Challenges in combining physics-based and data-driven components in VAE

A straightforward approach to account for epistemic uncertainty in a data-driven manner would be to approximate the measured response $\boldsymbol{x}$ as the sum of the physics-based model $f(\boldsymbol{z}_{\mathrm{x}})$ and a trainable NN-based function $g_{\boldsymbol{\theta}}(\cdot)$, where the latter corrects the discrepancies between the physics-based model predictions and measurements. It is assumed that the gradients of the physics-based model with respect to the inputs can be evaluated efficiently to obtain a computationally tractable optimization problem. This type of hybrid model is referred to as a residual model. Parametrizing the data driven component of the residual model as $g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\mathrm{x}})$ is not feasible when it is required that $\boldsymbol{z}_{\mathrm{x}}$ is interpretable: The resulting hybrid generative model has a posterior distribution $p_{\theta}(\boldsymbol{z}_{\mathrm{x}}|\boldsymbol{x})$, where the latent variables $\boldsymbol{z}_{\mathrm{x}}$ are the input to a coupled physics-based and data-driven model, and thus no longer physically meaningful. Instead, the latent space can be partitioned as $(\boldsymbol{z}_{\mathrm{x}},\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}}) \in \boldsymbol{z}$ and the data driven component parametrized as $g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}})$, where $\boldsymbol{z}_{\mathrm{c}}$ and $\boldsymbol{z}_{\mathrm{y}}$ are physically meaningless latent variables intended to capture variability in the measured response due to the influence of the domain and the class. Assuming that the remaining aleatory uncertainties (e.g., caused by measurement noise) are independent of the signal being measured and can be sufficiently modeled as independent and identically distributed (i.i.d.) samples of Gaussian white noise with standard deviation σ_{x}, the response measurements can be expressed as:

$$\boldsymbol{x} = f(\boldsymbol{z}_{\mathrm{x}}) + g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}}) + \boldsymbol{\epsilon}_{\mathrm{x}}, \tag{2.2}$$

where $\boldsymbol{\epsilon}_{\mathrm{x}} \sim \mathcal{N}(\boldsymbol{0},\sigma_{\mathrm{x}}^2\boldsymbol{I})$ and $\boldsymbol{I}$ is the identity matrix. Substituting $\hat{\boldsymbol{x}}_{\mathrm{p}} = f(\boldsymbol{z}_{\mathrm{x}})$ and $\hat{\boldsymbol{x}}_{\mathrm{d}} = g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}})$ for clarity, the resulting generative model is defined as:

$$p_{\theta}(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{x}},\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}}) := \mathcal{N}(\hat{\boldsymbol{x}}_{\mathrm{p}} + \hat{\boldsymbol{x}}_{\mathrm{d}},\sigma_{\mathrm{x}}^2\boldsymbol{I}) \tag{2.3}$$

The observables $\boldsymbol{c}$ and $\boldsymbol{y}$ can be used to ensure that the latent variables $\boldsymbol{z}_{\mathrm{c}}$ and $\boldsymbol{z}_{\mathrm{y}}$ encode information about the domain and class of the structure by simultaneously training auxiliary tasks $r_{\mathrm{c}}(\boldsymbol{c}|\boldsymbol{z}_{\mathrm{c}})$ and $r_{\mathrm{y}}(\boldsymbol{y}|\boldsymbol{z}_{\mathrm{y}})$. The resulting hybrid generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{x}},\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}})$ can be coupled with variational posteriors $q_{\phi_{\mathrm{x}}}(\boldsymbol{z}_{\mathrm{x}}|\boldsymbol{x})q_{\phi_{\mathrm{c}}}(\boldsymbol{z}_{\mathrm{c}}|\boldsymbol{x})q_{\phi_{\mathrm{y}}}(\boldsymbol{z}_{\mathrm{y}}|\boldsymbol{x})$ to yield an architecture similar to VAE. This is the architecture derived in Section 3, without the additional constraints. It should be noted that the assumption of an additive structure for the discrepancy term $g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\mathrm{c}},\boldsymbol{z}_{\mathrm{y}})$ and the uncertainty term $\boldsymbol{\epsilon}_{\mathrm{x}}$ presented in Equation (2.2) is suitable for many physical systems and is commonly employed in hybrid models (Cross et al., 2022). We use it without loss of generality with the aim of promoting clarity and interpretability. Depending on

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

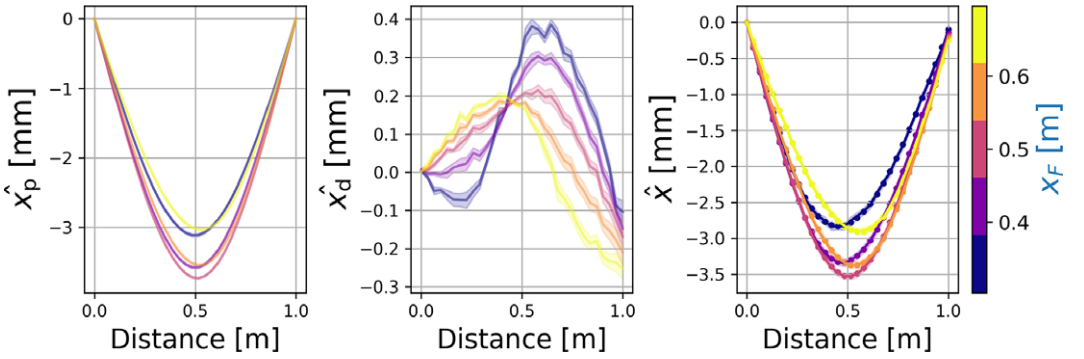


Figure 2. *Demonstration of the data-driven component of the decoder $g_\theta(\mathbf{z}_c, \mathbf{z}_y)$ overriding the physics-based model $f(\mathbf{z}_x)$. The effect of varying the position of the load x_F should be described by the known physics, but is instead captured by the data-driven components.*

domain knowledge regarding the problem at hand, a multiplicative or other form can also be specified. The implications of the additivity assumption are further discussed in Sections 3.2.3 and 6.2.

Neither the additive structure of the hybrid physics-based and data-driven model, nor the specified parametrization of the residual term $g_\theta(\boldsymbol{z}_c, \boldsymbol{z}_y)$ ensure that the known physics will be utilized by the model, or that $\boldsymbol{z}_x$ will be physically meaningful. Without further constraints, the model can learn combinations of arbitrary predictions from the physics-based and data-driven components $f(\boldsymbol{z}_x)$ and $g_\theta(\boldsymbol{z}_c, \boldsymbol{z}_y)$ that sum to an accurate prediction. To see why, it is sufficient to consider the form of the objective given in Equation (2.1). Both the encoder and decoder are aligned in the task of maximizing the reconstruction term $\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$, and the data-driven components of the model will account for discrepancies between the physics-based model prediction and measurements up to some level of noise. This results in an entangled representation, where the data-driven components override the known physics and the physics-grounded latent variables loose their physical meaning.

This issue is illustrated in Figure 2 using the beam case study shown in Figure 1(a). Further details of the case study are provided in Section 5.1. In this example, a VAE trained on a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i, \boldsymbol{y}_i)\}_{i=1}^N$ is evaluated on a new set of input measurements $\boldsymbol{x}$, generated from the ground truth data generating process by linearly varying the position of the load x_F. It can be seen that the effect of this variation on the measured response is largely captured by the data-driven component of the decoder, which overrides the known physics, despite the fact that the physics-based model includes the load position as an input parameter. The extent to which the data-driven components override the known physics can be inconsistent and hard to predict, and will depend on the neural network architectures and the physics of the problem at hand.

The results presented in Figure 2 are obtained under the assumption of a factorized variational posterior. Ideally, a single encoder with shared parameters $q_\phi(\boldsymbol{z}_x, \boldsymbol{z}_c, \boldsymbol{z}_y|\boldsymbol{x})$ would be used for the three subsets of the latent space $\boldsymbol{z}_x, \boldsymbol{z}_c$ and $\boldsymbol{z}_y$. However, using a shared encoder leads to further degradation of the performance as noted by Ilse et al., 2020. This example demonstrates how the interaction between the physics-based and data-driven components, which determines the resulting learned representation, depends on the capacity and flexibility of the individual machine learning components. Standard VAE offer no mechanism to control this interaction, and are therefore unable to guarantee that the physics will be utilized correctly in the presence of flexible NN-based decoder components.

3. Proposed approach

To address the issues presented in Section 2.4, we propose an approach that takes advantage of the domain and class observables to constrain the approximate posterior distribution. Our approach ensures that each

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

subset of the latent variables only encodes information that is relevant to the corresponding modality. This constraint in turn limits the amount of information available to the data-driven components of the decoder, preventing them from correcting every discrepancy between the physics-based model and measurements, and from overriding the known physics. This is achieved by imposing a latent bottleneck structure to the model, combined with an adversarial training objective. A detailed description of the model architecture and derivation of the training objective are provided in Section 3.1, followed by a brief discussion in Section 3.2. A method for quantitatively assessing the information encoded in subsets of the latent variables is presented in Section 3.3.

3.1. Detailed description of the model

It is assumed that three generative factors, the underlying physics of the structure, the domain, and the class, contribute to the measured response $\boldsymbol{x}$. Conversely, the latent variables are partitioned into subsets $(\boldsymbol{z}_x, \boldsymbol{z}_c, \boldsymbol{z}_y) \in \boldsymbol{z}$. It is emphasized that the separation of the latent variables is only semantic and used for clarity. In practice, they can be the output of a single encoder with shared parameters. The latent variables are the input to the hybrid probabilistic decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, wherein a NN-based function $g_{\boldsymbol{\theta}}(\boldsymbol{z}_c, \boldsymbol{z}_y)$ accounts for discrepancies between the measured response $\boldsymbol{x}$ and physics-based model prediction $f(\boldsymbol{z}_x)$. To ensure that information relevant to the domain and class is encoded in the corresponding subsets $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$, we utilize two auxiliary decoders $r_{\psi_c}(\boldsymbol{c}|\boldsymbol{z}_c)$ and $r_{\psi_y}(\boldsymbol{y}|\boldsymbol{z}_y)$. The latent variables $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$ are assigned conditional prior distributions $p_{\theta_c}(\boldsymbol{z}_c|\boldsymbol{c})$ and $p_{\theta_y}(\boldsymbol{z}_y|\boldsymbol{y})$ respectively, while the physics-grounded latent variables $\boldsymbol{z}_x$ are assigned a distribution $p(\boldsymbol{z}_x)$ based on the available prior knowledge. A schematic illustration of the architecture is provided in Figure 3(a).

To minimize the reconstruction error, the encoder tends to maximize the information in the posterior distribution over $\boldsymbol{z}$ that can be used to predict $\boldsymbol{x}$, $\boldsymbol{c}$ and $\boldsymbol{y}$, subject to the regularization imposed by the prior distribution. For $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$, this information includes features from the input signal $\boldsymbol{x}$ that are predictive of $\boldsymbol{c}$ and $\boldsymbol{y}$, but also irrelevant features that are only predictive of $\boldsymbol{x}$. These features can include systematic errors stemming from partial knowledge of the physics, and the influence of unknown confounding factors in the measurements. This *superfluous information* (Federici et al., 2020), i.e. information in $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$ that is not predictive of $\boldsymbol{c}$ and $\boldsymbol{y}$, can enable the data-driven component of the decoder to override the known physics. Motivated by this observation we aim to simultaneously maximize the information in $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$ that is predictive of $\boldsymbol{c}$ and $\boldsymbol{y}$, while minimizing the information that is predictive of $\boldsymbol{x}$. This trade-off can be formalized in terms of the mutual information (MI), a measure of the dependence between two random variables (Cover and Thomas, 2006). Denoting the MI between $\boldsymbol{x}$ and $\boldsymbol{z}$ for an encoder parametrized by $\boldsymbol{\phi}$

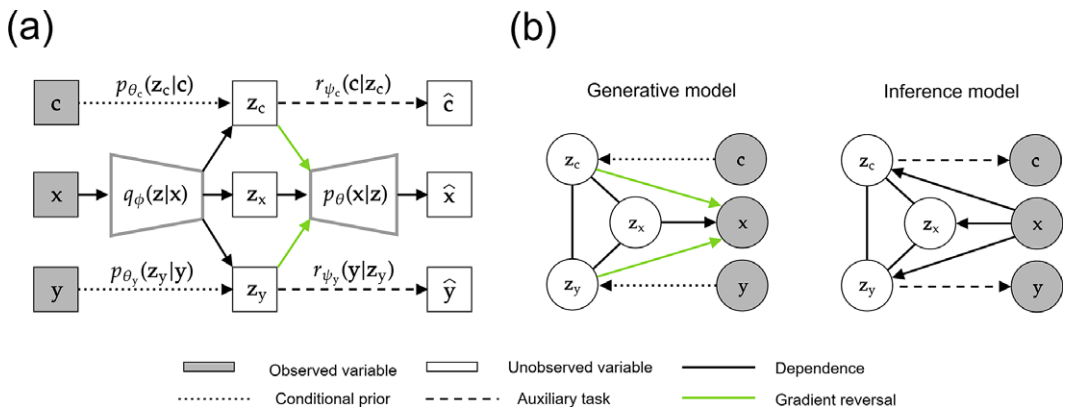


Figure 3. *a) Schematic diagram illustrating the components of the model and the encoder-decoder architecture, and b) Detailed structure of the dependencies in the generative and inference models.*

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

as $I_\phi(\boldsymbol{x};\boldsymbol{z})$, and introducing the trade-off parameters λ_c, λ_y, we define the following relaxed Lagrangian objectives:

$$\begin{aligned}\mathcal{L}_y(\boldsymbol{\phi};\lambda_y) &= I_\phi(\boldsymbol{y};\boldsymbol{z}_y) - \lambda_y I_\phi(\boldsymbol{x};\boldsymbol{z}_y)\\ \mathcal{L}_c(\boldsymbol{\phi};\lambda_c) &= I_\phi(\boldsymbol{c};\boldsymbol{z}_c) - \lambda_c I_\phi(\boldsymbol{x};\boldsymbol{z}_c)\end{aligned} \tag{3.1}$$

The quantities described in Equation (3.1) are optimized indirectly through a latent bottleneck structure (Alemi et al., 2019; Fischer, 2020; Moyer et al., 2018; Tishby et al., 2000) combined with adversarial training, as described in the following informal sketch. Note that in the inference model shown in Figure 3(b), the latent variables $\boldsymbol{z}_y$ do not depend directly on $\boldsymbol{y}$ (and analogously for $\boldsymbol{c}$). This is the conditional independence assumption typically used in the Information Bottleneck framework. Intuitively, the encoder is forced to distill the relevant information in $\boldsymbol{x}$ that is necessary for reconstructing $\boldsymbol{c}$ and $\boldsymbol{y}$ into the latent variables $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$. This results in the maximization of the $I_\phi(\boldsymbol{c};\boldsymbol{z}_c)$ and $I_\phi(\boldsymbol{y};\boldsymbol{z}_y)$ terms in Equation (3.1) during training. The additional requirement of minimizing $I_\phi(\boldsymbol{x};\boldsymbol{z}_c)$ and $I_\phi(\boldsymbol{x};\boldsymbol{z}_y)$ can be satisfied by introducing a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) at the input of the data-driven decoder component $g_\theta(\boldsymbol{z}_c,\boldsymbol{z}_y)$. During optimization, the gradient signal propagated backwards from $g_\theta(\boldsymbol{z}_c,\boldsymbol{z}_y)$ to the encoder is scaled by $-\lambda$, while the forward pass remains unchanged. Therefore, the GRL can be thought of as a pseudo function $R_\lambda(\boldsymbol{z})$ such that $R_\lambda(\boldsymbol{z}) = \boldsymbol{z}$ and $\frac{dR_\lambda}{d\boldsymbol{z}} = -\lambda \boldsymbol{I}$. Positive values of λ correspond to adversarial training. Conversely, negative values make the training "collaborative". The absolute value of λ determines the strength of the adversarial or collaborative objective, with larger values corresponding to a stronger regularization effect. By turning the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y)$ into an adversary, the GRL penalizes information in $\boldsymbol{z}_c$ and $\boldsymbol{z}_y$ that contributes to the reconstruction of $\boldsymbol{x}$, biasing the encoder towards representations that are minimally informative about $\boldsymbol{x}$.

The full structure of the model is shown in Figure 3(b). The variational lower bound can be obtained by considering the marginal likelihood over observed variables as shown in Equation (3.2).

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}) &= \mathbb{E}_{q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y},\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y)}{q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})}\right]\\ &\leq \log p_\theta(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y})\end{aligned} \tag{3.2}$$

Rearranging the terms in Equation (3.2), noting that the generative model factorizes as $p(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y},\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y) = p_\theta(\boldsymbol{x}|\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y)p(\boldsymbol{z}_x)p_{\theta_c}(\boldsymbol{z}_c|\boldsymbol{c})p_{\theta_y}(\boldsymbol{z}_y|\boldsymbol{y})p(\boldsymbol{c})p(\boldsymbol{y})$, yields the following expression for the lower bound:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}) &= \mathbb{E}_{q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y)\right]\\ &\quad - D_{KL}\left(q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})\,\|\,p(\boldsymbol{z}_x)p_{\theta_c}(\boldsymbol{z}_c|\boldsymbol{c})p_{\theta_y}(\boldsymbol{z}_y|\boldsymbol{y})\right)\end{aligned} \tag{3.3}$$

Including the auxiliary tasks and additional regularization hyperparameters commonly used in representation learning, we rewrite the loss function as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi};\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}) &= \mathbb{E}_{q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})}\left[\alpha_x \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y) + \alpha_c \log r_{\psi_c}(\boldsymbol{c}|\boldsymbol{z}_c) + \alpha_y \log r_{\psi_y}(\boldsymbol{y}|\boldsymbol{z}_y)\right]\\ &\quad - \beta D_{KL}\left(q_\phi(\boldsymbol{z}_x,\boldsymbol{z}_c,\boldsymbol{z}_y|\boldsymbol{x})\,\|\,p(\boldsymbol{z}_x)p_{\theta_c}(\boldsymbol{z}_c|\boldsymbol{c})p_{\theta_y}(\boldsymbol{z}_y|\boldsymbol{y})\right)\end{aligned} \tag{3.4}$$

The loss function described in Equation (3.4)) includes additional regularization hyperparameters that can be used to balance the contribution of different terms. These are included for completeness, and are not used in the experiments described in Section 5. A scaling factor $\beta > 0$ on the KLD is commonly included in the ELBO as a means of adjusting the strength of the regularization imposed by the KLD term, and to control the capacity of the probabilistic encoder. It is often beneficial to begin training with $\beta = 0$ and gradually increase it to $\beta = 1$ using an annealing scheme such as the one proposed by Bowman et al., 2016. Annealing β can prevent the model from getting stuck in local minima of the KLD, and the posterior distribution from degenerating to the prior distribution. Conversely, setting $\beta > 1$ can promote

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

unsupervised disentanglement (Higgins et al., 2016). The impact of β is extensively discussed in the relevant literature, provided in Section 4. Additionally, the log-likelihood function $\log p_\theta\left(\boldsymbol{x}|\boldsymbol{z}_\mathrm{x},\boldsymbol{z}_\mathrm{c},\boldsymbol{z}_\mathrm{y}\right)$, and auxiliary decoders $\log r_{\psi_\mathrm{c}}\left(\boldsymbol{c}|\boldsymbol{z}_\mathrm{c}\right)$ and $\log r_{\psi_\mathrm{y}}\left(\boldsymbol{y}|\boldsymbol{z}_\mathrm{y}\right)$ are assigned weights α_x, α_c and α_y respectively to allow for balancing the relative strength of these terms (see e.g. (Ilse et al., 2020; Joy et al., 2022; Sun et al., 2022). It is important to note that using values other than unity for $\alpha_\mathrm{x},\alpha_\mathrm{c},\alpha_\mathrm{y}$ and β can have a significant impact on the interpretation of the ELBO and the inferred posterior distribution. Details of the implementation can be found in Appendix A.

3.2. Discussion of the approach

The latent bottleneck structure and GRL have several important implications for inference, conditional generation, and uncertainty quantification. These are discussed here, and demonstrated through the case studies presented in Section 5.

3.2.1. Interpretability

The main objective of the approach is to ensure that the known physics are properly utilized, which we interpret as variability in the generative factors being preferentially captured by the physics-grounded subset of the latent variables. As a result, the posterior distribution of the physics-grounded latent variables will be influenced by the domain and class contributions to the measured response, allowing for domain and class influences to be interpreted in terms of their effect on the physics-grounded latent variables. Therefore, the posterior over physics-grounded latent variables might not necessarily be accurate in the sense of a point estimate of a physical quantity obtained from the posterior being close to the underlying "true value". This is also in part due to the data-driven component of the decoder, which can yield a constant but not necessarily zero prediction when domain or class influences are not present in the measured response.

3.2.2. Conditional generation

The use of conditional prior networks and separate branches for the domain and class modalities makes it possible to perform data imputation and conditional generation. When conditioning on domain or class variables, the accuracy of the generated response will depend on the degree to which the corresponding influence is accounted for by the known physics. If the influence is primarily accounted for by the physics, the predicted response might become insensitive to changes in the domain or class latent variables. In this case, more accurate conditional generation might be possible by fixing the values of the physics-grounded latent variables based on domain knowledge. Another implication of the architecture is that only measurements of the response are needed to evaluate the model. Throughout this work, the model is evaluated only on response measurements, without using the domain observables. This did not result in any noticeable difference in accuracy, compared to using the domain observables.

3.2.3. Uncertainty quantification

The uncertainty associated with the predicted response stems from the approximate posterior distribution and the probabilistic decoder, and represents the combined influence of aleatoric and epistemic uncertainties. Without the GRL, and given sufficient data, the model would compensate for systematic discrepancies between the physics-based model and response measurements in a data-driven manner. In this case, the uncertainty in the reconstructed response would only represent aleatoric uncertainty. In contrast, if no data-driven component is used in the decoder, the uncertainty would also include epistemic uncertainty due to domain and class influences that are not included in the physics-based model. In our approach, part of this epistemic uncertainty is accounted for in a data-driven manner. Therefore, the proposed approach is expected to yield uncertainty bounds somewhere in-between these two extremes. We emphasize that the estimated uncertainty does not include the uncertainty over model parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. Finally, it is important to consider that the additivity assumption in Equation (2.3) is unlikely to

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

hold for many physical systems. The model is expected to perform sub-optimally in such cases, resulting in inaccurate uncertainty estimates. None of the case studies presented in 5 satisfy the additivity assumption, demonstrating that the model can still be feasibly applied in such cases.

3.2.4. Formulation of the latent space

The choice of a continuous latent representation for the domain and class variables provides a number of advantages over directly representing the variables themselves, i.e. using the same number and type (e.g., categorical) of latent variables as the domain and class variables. The mapping to a low-dimensional continuous latent space enables the model to deal with high-dimensional domain and class variables, and can improve generalization by promoting the encoding of richer representations of the domain and class (Joy et al., 2022). From an implementation perspective, it is convenient if the decoder inputs are not dependent on the type and dimensionality of the domain and class variables. Furthermore, for discrete and categorical domain and class variables, the latent space enables the model to make a continuous approximation by interpolating over the continuous latent space. Broadly speaking, the continuous latent space allows for more flexibility in the representation of the domain and class variables. Finally, the lack of an independence assumption facilitates the use of a single probabilistic encoder, potentially reducing the amount of trainable parameters in the model and allowing for more complex and expressive encoder formulations. In our experiments we did not observe a decrease in performance when using a single encoder for all latent variables, compared to using separate encoders for each subset of the latent space, when combined with adversarial training.

3.3. Description of the quantitative assessment approach

Quantitatively assessing disentanglement is a challenging problem and several metrics have been proposed (Chen et al., 2018; Higgins et al., 2016; Kim and Mnih, 2018). This difficulty can be partially attributed to the lack of a consistent definition of disentanglement (Locatello et al., 2019). In practice, the degree of disentanglement achieved by a model is often evaluated based on subjective expectations stemming from domain knowledge (Vowels et al., 2019). Our proposed approach aims to achieve "one-way" disentanglement and is conceptually more akin to techniques that temper the influence of misspecified model components (Carmona and Nicholls, 2020; Yu et al., 2022). Intuitively, variations in generative factors that are described by the known physics should not affect the data-driven subsets of the latent variables $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$, while variations in generative factors not included in the known physics should still be preferentially captured by the physics-grounded subset of the latent variables. The degree to which this is achieved can be evaluated by comparing the amount of information captured by each subset $\boldsymbol{z}_\mathrm{x}$, $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$ about a specified generative factor. When a subset of the latent variables is informative about a generative factor, it should be possible to train a regressor to predict the value of the generative factor from samples drawn from this subset of the latent variables. Based on this, we propose the following procedure to assess the amount of information about a given generative factor that is encoded in a subset of the latent variables for the trained model:

1. Draw two sets of samples of generative factors $\left\{\left(\boldsymbol{s}_\mathrm{x}^{(i)},\boldsymbol{s}_\mathrm{c}^{(i)},\boldsymbol{s}_\mathrm{y}^{(i)}\right)\right\}_{i=1}^{N_\mathrm{train}}$ and $\left\{\left(\boldsymbol{s}'^{(i)}_\mathrm{x},\boldsymbol{s}'^{(i)}_\mathrm{c},\boldsymbol{s}'^{(i)}_\mathrm{y}\right)\right\}_{i=1}^{N_\mathrm{test}}$ from $p_\mathrm{gt}\left(\boldsymbol{s}_\mathrm{x},\boldsymbol{s}_\mathrm{c},\boldsymbol{s}_\mathrm{y}\right)$ and generate two datasets $D=\{\boldsymbol{x}_i\}_{i=1}^{N_\mathrm{train}}$ and $D'=\left\{\boldsymbol{x}'_i\right\}_{i=1}^{N_\mathrm{test}}$ of response measurements from the ground truth generative process.
2. Draw a single sample from each of the approximate posterior distributions $\boldsymbol{z}_i\sim q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)$ and $\boldsymbol{z}'_i\sim q_\phi\left(\boldsymbol{z}'_i|\boldsymbol{x}'_i\right)$ for each $\boldsymbol{x}_i\in D$ and $\boldsymbol{x}'_i\in D'$ respectively, using the trained model. This yields two sets of samples from the latent variables $\left\{\left(\boldsymbol{z}_\mathrm{x}^{(i)},\boldsymbol{z}_\mathrm{c}^{(i)},\boldsymbol{z}_\mathrm{y}^{(i)}\right)\right\}_{i=1}^{N_\mathrm{train}}$, and $\left\{\left(\boldsymbol{z}'^{(i)}_\mathrm{x},\boldsymbol{z}'^{(i)}_\mathrm{c},\boldsymbol{z}'^{(i)}_\mathrm{y}\right)\right\}_{i=1}^{N_\mathrm{test}}$.
3. Train a regressor to predict the value of each set of generative factors $\left\{s_j^{(i)}\right\}_{i=1}^{N_\mathrm{train}}$ from each subset $\left\{\boldsymbol{z}_\mathrm{x}^{(i)}\right\}_{i=1}^{N_\mathrm{train}}$, $\left\{\boldsymbol{z}_\mathrm{c}^{(i)}\right\}_{i=1}^{N_\mathrm{train}}$ and $\left\{\boldsymbol{z}_\mathrm{y}^{(i)}\right\}_{i=1}^{N_\mathrm{train}}$, for $j=1,\ldots,N_f$, where N_f is the number of generative factors. This process yields $3\times N_\mathrm{f}$ regressors.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

4. Compute the R^2 value between each subset $\left\{z'^{(i)}_{\mathrm{x}}\right\}_{i=1}^{N_{\mathrm{test}}}$, $\left\{z'^{(i)}_{\mathrm{c}}\right\}_{i=1}^{N_{\mathrm{test}}}$ and $\left\{z'^{(i)}_{\mathrm{y}}\right\}_{i=1}^{N_{\mathrm{test}}}$ and each set of generative factors $\left\{s'^{(i)}_{j}\right\}_{i=1}^{N_{\mathrm{test}}}$ using the corresponding trained regression model.

This procedure yields N_f sets of pair-wise R^2 values $\left\{R^2_{z_{\mathrm{x}}\to s_j}, R^2_{z_{\mathrm{c}}\to s_j}, R^2_{z_{\mathrm{y}}\to s_j}\right\}_{j=1}^{N_f}$ with each of the N_f sets corresponding to a single generative factor. A more informative subset of the latent variables should yield a more accurate regressor than an uninformative subset, and therefore also a higher R^2 value. It is emphasized that the metric described here is only intended to be a surrogate quantity for the amount of information encoded in each subset of the latent variables, and not a metric of disentanglement. Furthermore, this metric requires access to the ground truth distribution and data generating process, and is therefore not generally applicable.

4. Previous work

Deep generative models such as VAE describe a mapping between a high-dimensional data manifold, and a low dimensional latent representation. Generative factors in the data are not generally controlled by individual dimensions of the latent variables, nor are they amenable to human interpretation or semantically meaningful. Disentangled representation learning is aimed at learning representations where perturbations of individual dimensions of the latent space correspond to interpretable perturbations of the data (Esmaeili et al., 2019). Approaches for disentangled representation learning can be broadly classified as either unsupervised, where disentanglement is achieved through the use of additional regularization terms on the ELBO, or supervised, semi-supervised, and weakly supervised methods that utilize additional observables or other information. For a comprehensive review of representation learning and of the different approaches, focusing on VAE, the reader is referred to Bengio et al., 2014 and Tschannen et al., 2018 respectively.

Unsupervised disentanglement necessarily relies on inductive bias and implicit supervision (Locatello et al., 2019). A common approach is to adjust the relative importance of the KLD and reconstruction error terms. In the β-VAE architecture (Higgins et al., 2016), the KLD is scaled by a factor $\beta \geq 1$ that determines how much the approximate posterior is penalized for deviating from the prior distribution, limiting the capacity of the latent distribution and encouraging the latent variables to be factorized, at the expense of reconstruction quality (Burgess et al., 2018). Other approaches involve weighting the importance of the *total correlation* term (Watanabe, 1960), a component of the KLD that quantifies and penalizes dependence between the dimensions of the aggregated posterior distribution (i.e. the posterior distribution marginalized over the entire dataset). These approaches avoid the high computational cost associated with estimating the total correlation by utilizing stochastic approximations based on mini-batches (Chen et al., 2018) and adversarial density-ratio estimation (Kim and Mnih, 2018). The *InfoVAE* approach proposed by Zhao et al., 2018 involves scaling specific terms in the ELBO, coupled with an additional term that promotes maximization of the MI between the inputs and latent variables. Other approaches for unsupervised disentanglement have been proposed in the literature, such as enforcing independence between and within groups of latent variables (Esmaeili et al., 2019), extending the standard architecture with adversarial components (Larsen et al., 2016) and additional decoders (Ding et al., 2020) and using sparsity inducing priors (Tonolini et al., 2020). Several works utilize additional observables in a fully supervised (Debbagh, 2023; Hadad et al., 2018; Sun et al., 2022, or semi-supervised (Louizos et al., 2017) manner, combined with inductive biases in the form of structured models (N et al., 2017) and penalties on dependence (Lopez et al., 2018) to promote disentanglement or invariance to nuisance factors. It has also been shown (Achille and Soatto, 2017) that disentanglement is closely related to the information bottleneck theory introduced by Tishby et al., 2000, and later extended to the variational setting by Alemi et al., 2019. Finally, the more general notion of decomposition, that admits disentanglement as a special case, was introduced by Mathieu et al., 2019.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

It has been hypothesized that representation learning approaches can be particularly useful in domain adaptation and transfer learning tasks due to their ability to capture underlying generative factors in data that are shared between tasks (Bengio et al., 2014). The gradient reversal approach utilized in this work was originally proposed to tackle domain adaptation for image classification (Ganin and Lempitsky, 2015; Ganin et al., 2016). Similar approaches have been extended to the setting of multi-view and multi-modal learning (Aguerri and Zaidi, 2019; Federici et al., 2020; Hwang et al., 2020; Mondal et al., 2023). Recent work has also explored the application of adversarial domain adaptation techniques to structural damage identification (Wang and Xia, 2022). Finally, it is important to note that our proposed architecture is similar to the domain invariant variational autoencoder (DIVA) proposed by Ilse et al., 2020, from which we adopt part of our terminology. DIVA is targeted towards invariant representation learning and domain generalization in a purely data-driven setting and does not utilize adversarial training, instead explicitly imposing an independence assumption between subsets of the latent variables.

VI offers a balance between accuracy and computational tractability. Combined with the inherent regularization of the Bayesian framework, these properties are particularly advantageous in the modeling of physical systems (Glyn-Davies et al., 2025). As a result, incorporating physical knowledge in VAE has received significant attention, and various physics-informed formulations have been proposed depending on the task of interest. Notably, Walker et al., 2024 present an approach for utilizing known physics to discover shared information in multi-modal data. The UQ-VAE (Goh et al., 2022) combines known governing equations and prior distributions over parameters of interest with paired input-output measurements to achieve computationally efficient uncertainty quantification for systems described by partial differential equations. Formulations of VAE that take advantage of known governing equations have also been proposed for solving forward and inverse problems in stochastic differential equations (Shin and Choi, 2023; Zhong and Meidani, 2023). In the context of surrogate modeling, Rixner and Koutsourelakis, 2021 introduce the notion of virtual observables as a means of encoding physical knowledge into probabilistic generative models. Despite these advances, the issue of balancing physics-based and data-driven components in VAE has received relatively limited attention. This issue is addressed by Takeishi and Kalousis, 2021 for systems described by ordinary differential equations. A similar setting is investigated in Linial et al., 2021 and Yildiz et al., 2019, with the latter introducing a regularized objective to ensure consistency of the latent space with the known physics.

5. Synthetic case studies

Three synthetic case studies of different complexity, illustrated in Figure 1, are discussed in detail throughout this section. The case study objectives, the definition of the physics-based models, the procedure used to generate the synthetic data, and details of the model implementation and visualization are provided below. For the purposes of reproducibility, the code needed to replicate the examples is made available on GitHub (Koune and Cicirello 2025). Additional information regarding the architecture, variable transformations, data, optimization, and visualization is provided in Appendix A.

Case study objectives

Each case study addresses a different set of challenges. The beam case study demonstrates that the proposed approach preferentially utilizes the known physics, yielding an interpretable and parsimonious representation of the physical system. The oscillator case study highlights how the adversarial training can prevent the model from learning arbitrary components of the response in a data-driven manner, and investigates the impact of the GRL hyperparameter. Finally, the bridge case study demonstrates the feasibility of using the model for damage detection in a more complex synthetic case, and compares the performance to that of existing data-driven approaches. It is noted that the case studies are only meant as didactic examples, intended to elaborate the issues with combining physics-based and data-driven components in VAE, provide intuition about the interaction between these components, and demonstrate the behavior of the model. Therefore, emphasis is placed on clarity rather than realism.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

Physics-based models

Three separate physics-based models are considered for every case study: A high-fidelity *simulator*, a *full* model, and a *nominal* model. The simulator is an accurate but generally computationally expensive model of the physical system, typically in the form of a finite element (FE) model, used to train the full and nominal models for each case study. The full model is a computationally efficient surrogate model of the ground truth data generating process: one or more structures with varying physical characteristics, subject to operational and environmental conditions, damage and degradation. To produce the training dataset for the full model, the simulator is evaluated on a set of generative factors $\left\{\left(\boldsymbol{s}_{\mathrm{x}}^{(i)},\boldsymbol{s}_{\mathrm{c}}^{(i)},\boldsymbol{s}_{\mathrm{y}}^{(i)}\right)\right\}_{i=1}^{N_{\mathrm{full}}}$, sampled uniformly and independently from prescribed ranges of values. The ranges are chosen to provide sufficient coverage over the support of the corresponding ground truth distribution $p_{\mathrm{gt}}\left(\boldsymbol{s}_{\mathrm{x}},\boldsymbol{s}_{\mathrm{c}},\boldsymbol{s}_{\mathrm{y}}\right)$. The full model is then obtained by fitting a NN-based surrogate to the dataset composed of N_{full} input-output pairs $D_{\mathrm{full}}=\left\{\left(\boldsymbol{s}_{\mathrm{x}}^{(i)},\boldsymbol{s}_{\mathrm{c}}^{(i)},\boldsymbol{s}_{\mathrm{y}}^{(i)},\boldsymbol{x}^{(i)}\right)\right\}_{i=1}^{N_{\mathrm{full}}}$ obtained from the simulator. Using a NN as the forward model for the data generating process enables the efficient visualization of the latent space and the reconstructions generated by the VAE for different inputs, simplifies the generation of test data to evaluate the performance of the VAE, and makes it possible to account for randomness in the hyperparameter initialization and data generation by averaging results over multiple runs with i.i.d. datasets. The nominal model corresponds to the available incomplete representation of the physics of the system under investigation. When an analytical expression describing the partially known physics is available, this is used as the nominal model. Alternatively, the nominal model is built by training a NN-based surrogate on a limited dataset $D_{\mathrm{nom}}=\left\{\left(\boldsymbol{s}_{\mathrm{x}}^{(i)},\boldsymbol{x}^{(i)}\right)\right\}_{i=1}^{N_{\mathrm{nom}}}$, obtained by evaluating the simulator only on the physics-based subset of the generative factors $\left\{\boldsymbol{s}_{\mathrm{x}}^{(i)}\right\}_{i=1}^{N_{\mathrm{nom}}}$, while $\boldsymbol{s}_{\mathrm{c}}$ and $\boldsymbol{s}_{\mathrm{y}}$ are set to a constant reference value corresponding to the nominal condition of the structure.

Synthetic data generation

The VAE is trained and validated on a dataset composed of $N_{\mathrm{total}}=N_{\mathrm{train}}+N_{\mathrm{val}}$ triplets of observables $\mathcal{D}=\{(\boldsymbol{x}_i,\boldsymbol{c}_i,\boldsymbol{y}_i)\}_{i=1}^{N_{\mathrm{total}}}$. This dataset is generated by first drawing samples of the generative factors from the ground truth distribution $\left(\boldsymbol{s}_{\mathrm{x}}^{(i)},\boldsymbol{s}_{\mathrm{c}}^{(i)},\boldsymbol{s}_{\mathrm{y}}^{(i)}\right)\sim p_{\mathrm{gt}}\left(\boldsymbol{s}_{\mathrm{x}},\boldsymbol{s}_{\mathrm{c}},\boldsymbol{s}_{\mathrm{y}}\right)$ for $i=1,\ldots,N_{\mathrm{total}}$, applying a set of deterministic transformations, and subsequently adding i.i.d. samples of zero-mean Gaussian white noise $\boldsymbol{\epsilon}_{\mathrm{x}}\sim\mathcal{N}\left(\boldsymbol{0},\sigma_{\mathrm{x}}^2\boldsymbol{I}\right)$, $\boldsymbol{\epsilon}_{\mathrm{c}}\sim\mathcal{N}\left(\boldsymbol{0},\sigma_{\mathrm{c}}^2\boldsymbol{I}\right)$ and $\boldsymbol{\epsilon}_{\mathrm{y}}\sim\mathcal{N}\left(\boldsymbol{0},\sigma_{\mathrm{y}}^2\boldsymbol{I}\right)$. Denoting the full model as $h_{\mathrm{x}}(\cdot)$, the response observables are obtained as $\boldsymbol{x}_i=h_{\mathrm{x}}\left(\boldsymbol{s}_{\mathrm{x}}^{(i)},\boldsymbol{s}_{\mathrm{c}}^{(i)},\boldsymbol{s}_{\mathrm{y}}^{(i)}\right)+\boldsymbol{\epsilon}_{\mathrm{x}}$. The domain and class observables are obtained as $\boldsymbol{c}_i=h_{\mathrm{c}}\left(\boldsymbol{s}_{\mathrm{c}}^{(i)}\right)+\boldsymbol{\epsilon}_{\mathrm{c}}$ and $\boldsymbol{y}_i=h_{\mathrm{y}}\left(\boldsymbol{s}_{\mathrm{y}}^{(i)}\right)+\boldsymbol{\epsilon}_{\mathrm{y}}$ respectively. This procedure is illustrated in Figure 4.

Implementation details

For all the case studies presented in this section, h_{c} and h_{y} are taken as the identity function for simplicity. Furthermore we use $N_{\mathrm{train}}=1024$ and $N_{\mathrm{val}}=512$, and consider no other regularization except for the GRL, i.e. $\beta=\alpha_{\mathrm{x}}=\alpha_{\mathrm{c}}=\alpha_{\mathrm{y}}=1.0$. Unless stated otherwise, the number of the domain and class latent variables are taken to be twice the number of domain and class generative factors. The intention behind this choice is to avoid biasing the model towards a disentangled representation by matching the number of latent variables to the ground truth generative factors, ensuring that any disentanglement in the learned representation is not a consequence of limited latent space capacity. To enable the formulation of problems with bounded latent variables and to ensure a stable optimization procedure, the physics-grounded latent variables are obtained through a sequence of invertible transformations applied to the encoder output, mapping samples from an unbounded base latent space to the target latent space. All physics-grounded

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

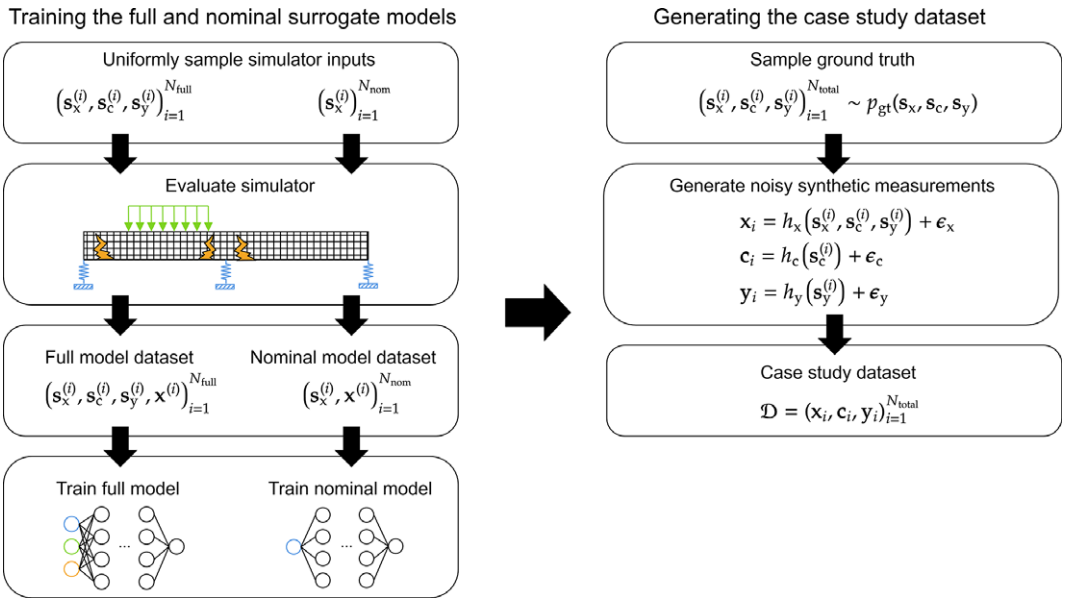


Figure 4. *Illustration of the procedure used to obtain the full and nominal physics-based models (left), and to generate the datasets used in the case studies (right).*

latent variables are constrained to lie within ranges that ensure consistency with the underlying physics of each system.

Visualization

A particularly useful tool for assessing the learned representation is to "traverse" the latent space and the space of reconstructions of the VAE (Higgins et al., 2016; Kim and Mnih, 2018). This can be achieved by generating synthetic data while interpolating over a specified generative factor, and setting the remaining generative factors to a constant reference value. The VAE is then evaluated on the generated data, yielding samples from the latent space, realizations of the reconstructed input $\hat{\boldsymbol{x}}$, and the mean physics-based and data-driven components $\hat{\boldsymbol{x}}_\mathrm{p}$ and $\hat{\boldsymbol{x}}_\mathrm{d}$. Each generative factor is linearly interpolated within the 1$^{\mathrm{st}}$ and 99$^{\mathrm{th}}$ percentiles of the corresponding ground truth distribution.

5.1. Beam case study

5.1.1. Case study description

The case study consists of a beam with fixed length $L = 1.0$ m and a point load with magnitude $F = 1.0$ N acting on an uncertain position x_F along the length of the beam. The material is linear elastic with uncertain Young's modulus E, Poisson ratio $\nu = 0.3$, area moment of inertia $I = 2 \cdot 10^{-6}$ m^4 and cross-sectional area $A = 2.4 \cdot 10^{-3}$ m^2. The rotational stiffness of the right-hand side support is temperature dependent, with the dependence modeled as an increase in the rotational stiffness of the support at lower temperatures. The relationship between the temperature and the support rotational stiffness is formulated as $\log k_\mathrm{r} = 8 - \frac{10}{1+e^{-T/2}}$. The beam is subject to variability in the vertical stiffness of the right-hand side support, e.g. due to damage or a deficiency of the support, simulated as a translational spring boundary condition with stiffness k_v. This quantity can span several orders of magnitude, and therefore we parametrize the model using $\log k_\mathrm{v}$ instead. The beam is equipped with $d_\mathrm{x} = 32$ sensors measuring the vertical displacement, equally spaced along the length as shown in Figure 1(a).

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

Table 1. *Summary of generative factors and the corresponding ground truth and prior distributions*

Variable	Unit	Type	Prior distribution	Ground truth	Reference value
E	MPa	Physical	$\mathcal{N}(4.0,1.0)$	$\mathcal{U}(2.5,4.5)$	3.0
x_{F}	m	Physical	$\mathcal{N}(0.5,0.04)$	$\mathcal{U}(0.3,0.7)$	0.5
$\log k_{\mathrm{v}}$	N/m	Class	-	$\mathcal{U}(6.0,8.0)$	8.0
T	C^o	Domain	-	$\mathcal{U}(-11.0,5.0)$	5.0

The Young's modulus E and the position of the point load x_{F} are considered as uncertain latent variables, such that $\boldsymbol{z}_{\mathrm{x}} = (E, x_{\mathrm{F}})$. It is assumed that the temperature is an observed domain variable such that $\boldsymbol{c} = (T)$, and that the vertical spring log-stiffness $\log k_{\mathrm{v}}$ is taken as a class variable representing damage in the structure, such that $\boldsymbol{y} = (\log k_{\mathrm{v}})$. Since the class variable $\boldsymbol{y}$ represents damage in the structure it will not be quantitatively measurable. In a realistic scenario, observations of the condition of the support on a qualitative scale (e.g. from 0 representing no damage to 5 denoting a fully damaged support) might be available. In this example we simplistically consider $\boldsymbol{y}$ as the ground truth value of $\log k_{\mathrm{v}}$ with some added noise. The variable symbols, units, types, as well as the prior distributions over the physics-grounded latent variables and the ground truth distributions of the generative factors used to generate the training data are summarized in Table 1. We additionally provide a reference value which is used to produce the figures as discussed in Section 5. To ensure physical consistency and to avoid numerical issues, the Young's modulus is truncated below a small positive value, and the load position x_{F} is restricted to the range $(0,1)$.

A partial description of the physics is available, in the form of an analytical expression for the vertical deflection of a simply supported Euler-Bernoulli beam with a point load acting at x_{F}:

$$w(x) = \begin{cases} \dfrac{Pbx\left(L^2 - b^2 - x^2\right)}{6LEI}, & 0 \le x \le x_{\mathrm{F}} \\ \dfrac{Pbx\left(L^2 - b^2 - x^2\right)}{6LEI} + \dfrac{P(x - x_{\mathrm{F}})^3}{6EI}, & x_{\mathrm{F}} < x \le L \end{cases} \tag{5.1}$$

where $b = L - x_{\mathrm{F}}$, and the non-bold x refers to the position along the beam. This nominal model represents the beam in the undamaged condition at a reference temperature, and is directly incorporated in the physics-based branch of the VAE decoder.

Following the procedure described in Section 5, the full model (trained on input-output pairs from an FE-based simulator) is used to produce synthetic data by first drawing samples of the input parameters from the ground truth distribution, and subsequently contaminating the resulting model predictions with zero-mean Gaussian white noise with standard deviation $\sigma_{\mathrm{x}} = 0.02$ m. The dataset used to train the VAE is composed of N_{train} measurements of the beam displacement $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{N_{\mathrm{train}}}$ where each element $\boldsymbol{x}_i$ is a vector of length $d_{\mathrm{x}} = 32$. The domain and class observables $\boldsymbol{c} = \{\boldsymbol{c}_i\}_{i=1}^{N_{\mathrm{train}}}$ and $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^{N_{\mathrm{train}}}$ are obtained as the ground truth values used to generate the dataset, with the addition of i.i.d. samples of Gaussian white noise with standard deviations of $\sigma_c = \sigma_y = 0.02$, respectively. The dimensionality of the domain and class latent variables is taken as $d_{z_{\mathrm{c}}} = d_{z_{\mathrm{y}}} = 2$.

5.1.2. Qualitative assessment of disentanglement

After training, the disentanglement between physics-grounded and data-driven components is qualitatively assessed by examining the latent space and samples of the reconstructed response of the beam. The predicted physics-based $\hat{\boldsymbol{x}}_{\mathrm{p}}$ and data-driven $\hat{\boldsymbol{x}}_{\mathrm{d}}$ components, as well as the combined prediction $\hat{\boldsymbol{x}}$ are shown in Figure 5. It can be observed that the data-driven component of the reconstruction $\hat{\boldsymbol{x}}_{\mathrm{d}}$ (middle

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

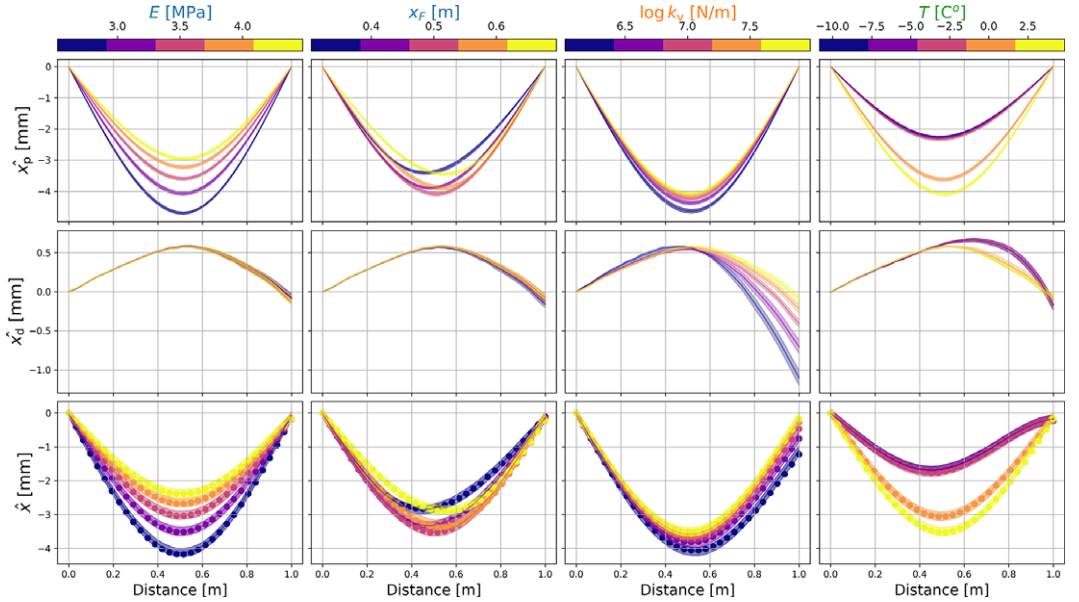
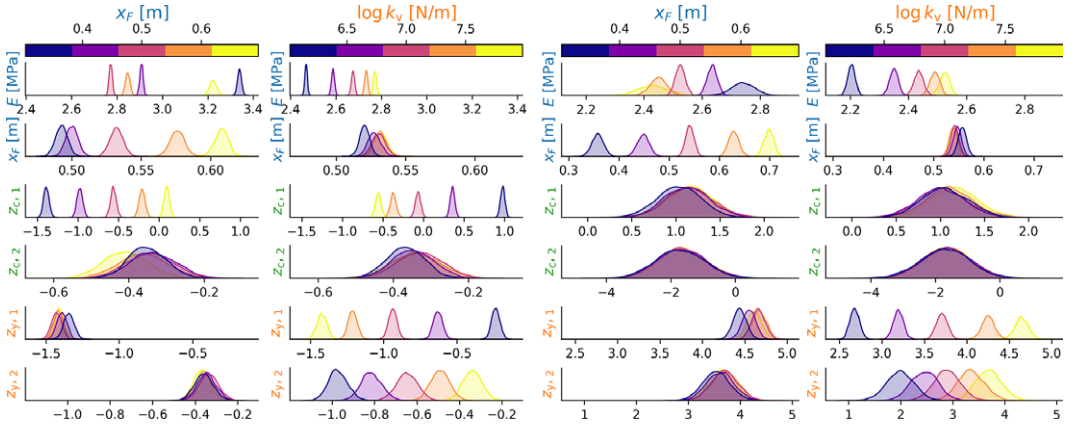


Figure 5. *Mean prediction and* $\pm 2\sigma$ *uncertainty bounds for the physics-based* $\hat{\mathbf{x}}_\mathrm{p}$ *and data-driven* $\hat{\mathbf{x}}_\mathrm{d}$ *components, and combined prediction* $\hat{\mathbf{x}}$ *while traversing the generative factors. The input response measurements are denoted as dots in the bottom row.*

row) is invariant to changes in E and x_F contributing only a constant deformed shape to the total predicted response. On the other hand, the physics-based model captures variability in both the physics-grounded generative factors $\boldsymbol{s}_\mathrm{x} = (E, x_\mathrm{F})$, but also domain and class generative factors $\boldsymbol{s}_\mathrm{c} = (T)$ and $\boldsymbol{s}_\mathrm{y} = (\log k_\mathrm{v})$. In contrast to the behavior of the unconstrained VAE, presented in Section 2.4, here the model preferentially utilizes the known physics. Only the variability in the measured response due to $\log k_\mathrm{v}$ and T that can not be captured by the physics-based model is accounted for by the data-driven part of the decoder, indicating that the model can disentangle components of the response that can be attributed to the known physics from those that cannot. A key aspect of the adversarial training is the degree to which it allows interaction between the physics-based and data-driven components of the prediction. In this case study, the additional displacement of the beam due to the reduced vertical stiffness of the right-hand side support will also depend on the load position x_F. More positive values of λ tend to prevent the model from capturing this interaction, whereas more negative values enable it but may result in the data-driven components overriding the known physics.

To further highlight the impact of the GRL, the latent space traversals of the unconstrained model and the model trained adversarially are compared in Figure 6. Without adversarial training, the domain latent variables $\boldsymbol{z}_\mathrm{c}$ encode the variability in the load position x_F as shown in Figure 6(a), providing the data-driven decoder components with the information needed to reconstruct this component of the measured response and resulting in an entangled representation, as discussed in Section 2.4. In contrast, when $\lambda = 1/256$ the adversarial training results in a posterior distribution over $\boldsymbol{z}_\mathrm{c}$ that is invariant to changes in x_F. Instead, the variability is captured by the corresponding physics-grounded latent variable, as shown in Figure 6(b), indicating disentanglement of the physics-grounded and domain generative factors. The results shown previously suggest that the latent bottleneck architecture and GRL regularization result in a sparse and parsimonious representation of the physical system, and can yield domain and class latent variables that are invariant to changes in the underlying physics. The influence of the GRL hyperparameter λ is further investigated in the oscillator case study presented below.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press



(a) Marginal approximate posterior without adversarial training ($\lambda = -1$).

(b) Marginal approximate posterior with adversarial training ($\lambda = 1/256$).

Figure 6. *Visualizations of the VAE latent space during traversal of the generative factors x_F and $\log k_v$. Each column corresponds to variation of a single generative factor, and each row shows the marginal approximate posterior distribution of a single latent variable.*

5.2. Oscillator case study

5.2.1. Case study description

This example demonstrates how the adversarial training prevents the model from compensating for all discrepancies between the physics-based model predictions and measurements. Suppose that a mass-spring-dashpot system undergoes damped harmonic motion, starting from an initial displaced position x_0, with no external excitation. It is assumed that each experiment is performed under varying temperature T, which is taken as the domain variable. The temperature affects the spring stiffness through the relationship $k(T) = k_{\text{ref}} + \alpha_T(T_{\text{ref}} - T)$, with $T_{\text{ref}} = 20.0$ C^o and $\alpha_T = 0.01$. The reference spring stiffness at $T_{\text{ref}} = 20.0$ C^o is assumed known and equal to $k_{\text{ref}} = 1.0$ N/m. The mass m is considered unknown and treated as a physics-grounded latent variable to be inferred from data. The viscous damping coefficient c_d is taken to define the class of the system. Finally, it is assumed that the observations are subject to an unknown confounding influence in the form of small random perturbations of the initial displacement x_0. A summary of the generative factors, the prior distribution and the ground truth distribution is provided in Table 2.

The equation of motion describing the system can be written as:

$$m\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2} + c_d\frac{\mathrm{d}x(t)}{\mathrm{d}t} + k(T)x(t) = 0 \tag{5.2}$$

A partial description of the physics is available in the form of an analytical solution under the assumption that the initial displacement is $x_0 = 1.0$ m, and the initial velocity is $\dot{x}_0 = 0.0$ m/s for all

Table 2. *Summary of generative factors and the corresponding ground truth and prior distributions*

Variable	Unit	Type	Prior distribution	Ground truth	Reference value
m	kg	Physical	$\mathcal{U}(1.0,2.0)$	$\mathcal{U}(1.2,1.8)$	1.5
c_d	kg / s	Class	-	$\mathcal{U}(0.0,2.0)$	0.0
T	C^o	Domain	-	$\mathcal{U}(0.0,40.0)$	20.0
x_0	m	Unknown	-	$\mathcal{U}(0.9,1.1)$	1.0

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

experiments, and that there is no damping affecting the motion of the oscillator. Furthermore, it is assumed that the relationship between temperature and stiffness is not known, and the temperature effect is therefore not included in the nominal physics-based model. Under the assumptions described previously, the displacement of the oscillator at time t can be expressed as:

$$x(t) = \cos\left(\sqrt{\frac{k_{\text{ref}}}{m}}t\right) \tag{5.3}$$

Each triplet of observations is composed of a noisy displacement time series, and noisy measurements of the viscous damping coefficient c_{d} and temperature T, which are considered as class and domain variables respectively such that $\boldsymbol{c} = (T)$ and $\boldsymbol{y} = (c_{\text{d}})$. Training and validation datasets are generated by drawing samples from the ground truth distribution and generating the oscillator displacement time-series using the full model, trained on input-output pairs simulated using the equation of motion shown in Equation (5.2). Each of the measured time-series is a vector of 64 measurements, equally spaced within a time interval $t \in [0, 10]$ s. The synthetic response measurements are subsequently contaminated with i.i.d. realizations of Gaussian white noise with standard deviation $\sigma_{\text{x}} = 0.01$ m. The standard deviation of the measurement uncertainty of the domain and class observables are taken as $\sigma_{\text{c}} = 0.01$ and $\sigma_{\text{y}} = 0.01$ respectively. To ensure that the model has sufficient capacity to learn the unknown confounding influence if the adversarial training were not present, the dimensionality of the latent space is specified to be significantly larger than the number of ground truth generative factors. The domain and class latent space dimensions are taken as $d_{z_{\text{c}}} = d_{z_{\text{y}}} = 4$.

5.2.2. Model behavior in the presence of unknown confounders

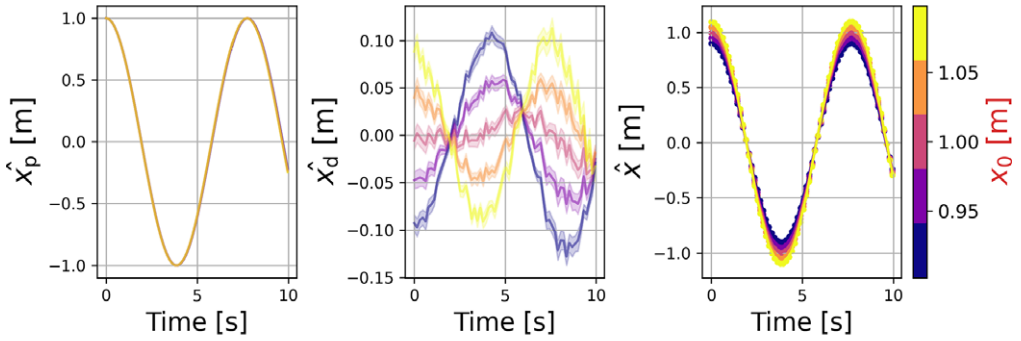
The proposed model with no adversarial training ($\lambda = -1$) is trained and evaluated on the synthetic example. The reconstructed response obtained from a traversal of the initial displacement x_0, shown in Figure 7a, highlights another issue that occurs when combining physics-based and data-driven components in VAE: Although the variability in x_0 can not be accounted for by the physics-based model, and there is no information in the domain or class variables regarding the value of x_0, the lack of regularization results in a model that is free to capture the components of the measured displacement stemming from the variability in the initial displacement x_0. Although in this case the effect is benign, in more complex physical systems it can result in the model learning unknown confounding influences in a non-interpretable black-box manner. When the GRL regularization is utilized (Figure 7b), the data-driven encoder is unable to capture the variability in x_0, depriving the data-driven decoder from the information needed to reconstruct this component of the input measurements.

The results shown in Figure 7 demonstrate how the unconstrained VAE will compensate for discrepancies between the physics-based model prediction and the measurements caused by unknown confounding influences. The reason why this can be detrimental for the learning task is illustrated in Figure 8a. The unconstrained VAE accounts for the learned confounding influence of x_0, which can lead to underestimation of the uncertainty over the latent variables and predictions. In contrast, when the model is trained with the adversarial objective, the encoder is prevented from learning a representation of x_0. The uncertainty stemming from the partial knowledge of the physics, including the unknown influence of the viscous damping and the variability in x_0, is more accurately accounted for in the reconstructed response, as shown in (Figure 8b). The additional uncertainty can also be attributed to the fact that the mass-spring-dashpot system does not satisfy the additivity assumption described in Section 2.4.

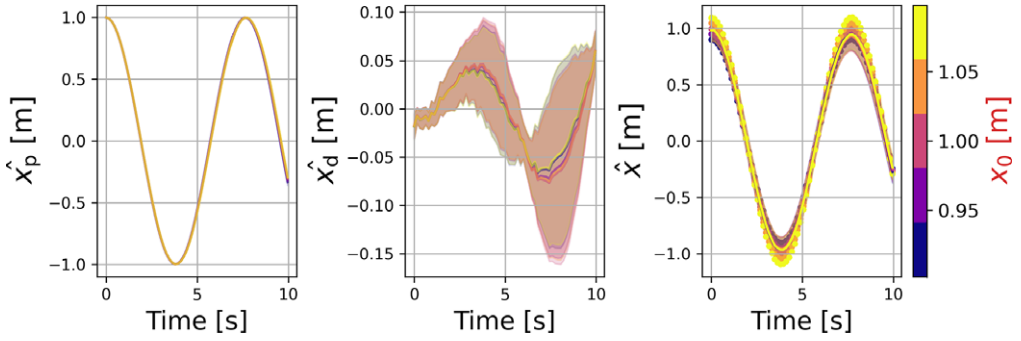
5.2.3. Quantitative assessment of disentanglement

The trade-off between invariance of the domain and class latent variables to non-domain or class influences and prediction accuracy can be adapted by tuning the GRL hyperparameter λ. A parameter study is performed to assess the impact of different choices for λ on the learned representation. The model is trained for varying values of $\lambda = \{-1, -1/10, -1/100, -1/1000, 0, 1/1000, 1/100, 1/10, 1\}$, and the metric described in Section 3.3 is computed for each trained model using linear regression. The training

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press



(a) Mean and $\pm 2\sigma$ uncertainty bounds of the reconstructed response without adversarial training ($\lambda = -1$).



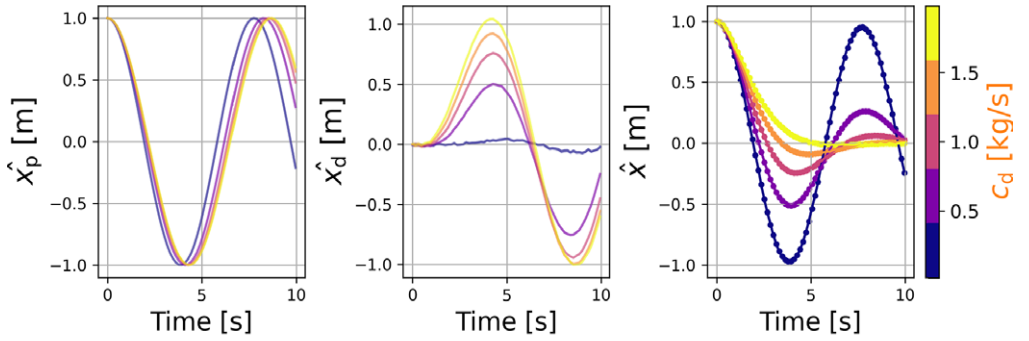
(b) Mean and $\pm 2\sigma$ uncertainty bounds of the reconstructed response with adversarial training ($\lambda = 1/128$).

Figure 7. *Physics-based model prediction* $\hat{\mathbf{x}}_\mathrm{p}$*, data-driven model prediction* $\hat{\mathbf{x}}_\mathrm{d}$*, and combined prediction* $\hat{\mathbf{x}}$ *for varying initial displacement* x_0*. With* $\lambda = -1.0$ *(top) the data-driven components in the VAE are free to account for the variability in the initial position. For* $\lambda = 1/128$ *(bottom) the model does not learn this component of the response.*

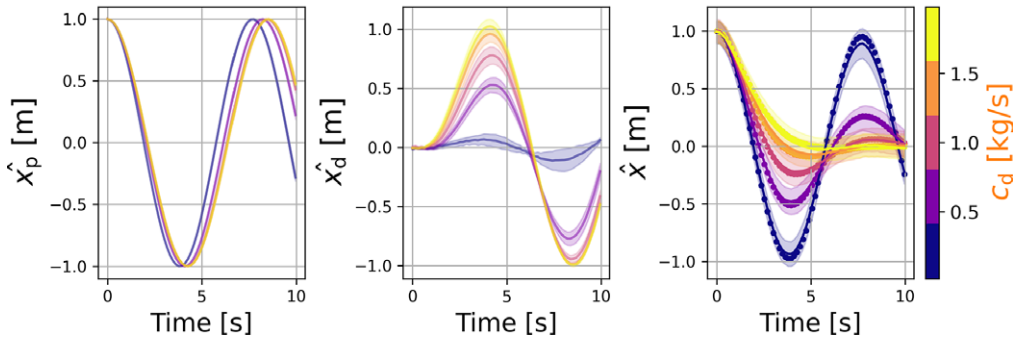
and testing datasets D and D' are composed of 2048 samples each. To account for the impact of randomness in the synthetic dataset, neural network parameter initialization, and training procedure, the results for each value of λ are averaged over multiple runs. The values of the metric for each generative factor and subset of the latent space, as a function of λ, are shown in Figure 9.

It can be seen that the sign of λ determines the nature of the training procedure, with positive values resulting in adversarial training. For negative values of λ the training becomes collaborative, in the sense that the encoder attempts to find approximate posterior distributions over $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$ that are jointly informative about their respective modality as well as the response measurements $\boldsymbol{x}$. The magnitude of λ determines the strength of the adversarial or collaborative training. To aid in the interpretation of the results, the behavior of the model is classified into four regimes. When λ approaches -1 from above, the training is *strongly collaborative*, and the tasks of minimizing the error in the reconstruction of $\boldsymbol{x}$ and the prediction of the domain and class variables $\boldsymbol{c}$ and $\boldsymbol{y}$ are jointly prioritized. This is reflected by the relatively high scores obtained by the subsets $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$ for the generative factors m and x_0. For $\lambda \to 0^-$, the auxiliary tasks are prioritized over the main task, and the amount of information about m and x_0 that is encoded in $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$ is limited. In this regime the behavior can be characterized as *weakly collaborative.* Conversely, for small positive values of λ the training becomes *weakly adversarial.* In this regime the encoder will seek latent codes over $\boldsymbol{z}_\mathrm{c}$ and $\boldsymbol{z}_\mathrm{y}$ that are uninformative about $\boldsymbol{x}$. Further increasing the GRL coefficient such that $\lambda \to 1$ yields a *strongly adversarial* model, and any information that can be used to reconstruct the domain and class variables is heavily weighted against the potential improvement in the reconstruction of $\boldsymbol{x}$. In this case the encoder fails to capture the variation in any of the generative factors.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press



(a) Mean and $\pm 2\sigma$ uncertainty bounds of the reconstructed response without adversarial training ($\lambda = -1$).



(b) Mean and $\pm 2\sigma$ uncertainty bounds of the reconstructed response with adversarial training ($\lambda = 1/128$).

Figure 8. *Physics-based model prediction $\hat{\mathbf{x}}_p$, data-driven model prediction $\hat{\mathbf{x}}_d$, and combined prediction $\hat{\mathbf{x}}$ for varying viscous damping coefficient c_d. The data-driven decoder components are prevented from fully accounting for the discrepancies between the physics-based model and measurements, resulting in wider uncertainty bounds for the proposed model.*

5.3. Bridge case study

5.3.1. Case study description

The final synthetic case study utilizes the two-span bridge benchmark presented in Tatsis and Chatzi, 2019, illustrated in Figure 1(c). Members of a homogeneous population (Bull et al., 2021) of bridges are subjected to controlled loading tests, where a vehicle with known mass and moving at a constant velocity is used to excite the bridge response. The response is obtained as a strain influence line, expressed in parts per thousand (‰), measured by a point strain gauge placed at a distance of 5.625 m from the start of the bridge, and at a height of 0.1 m from the bottom of the cross section. Each time-series is composed of 64 measurements, equally spaced in time $t \in [1,21]$ s, where $t = 0$ s is the moment the vehicle enters the bridge.

The behavior of each bridge is partially determined by the unknown vertical stiffnesses of the supports $k_{v,1}$, $k_{v,2}$ and $k_{v,3}$, which are taken to vary between different bridges due to variability in the design, construction and soil conditions. The boundary conditions are known to be symmetric such that $k_{v,1} = k_{v,3}$. The base-10 logarithms of the vertical stiffnesses are considered as physics-grounded latent variables. In the horizontal direction, only the left support has a large stiffness, while the rest are unconstrained. It is assumed that the position of the central pier can vary between members of the population by up to ± 1.0 m from $L/2$. Furthermore, fluctuations from the prescribed reference vehicle velocity $v_{ref} = 1$ m/s were observed during the tests that can not be accounted for in the nominal physics-based model. These fluctuations are modeled as a multiplicative term δ_v such that $v = \delta_v \cdot v_{ref}$. Noisy measurements of the vehicle velocity, and the known pier offsets δ_s are included as domain variables. It is

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

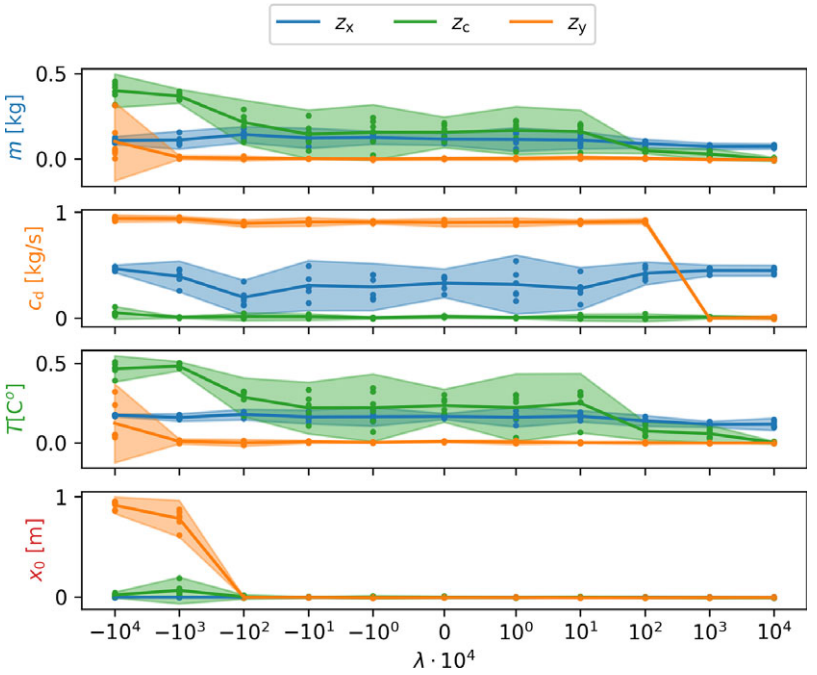


Figure 9. *R^2 value per subset of the latent variables and generative factor as a function of λ, averaged over 6 runs. The shaded intervals correspond to two standard deviations.*

assumed that the bridge decks are prone to deterioration in a region around the supports. During inspections performed by experts, each bridge is assigned scores $\boldsymbol{y} = (y_1, y_2)$, quantifying the deterioration of the structure near the left and middle supports respectively. Each y_i takes values between zero and one, where zero represents pristine condition and unity corresponds to severe damage of the cross-section at that position. These scores are considered as class observables. Finally, a small variability is considered in the vehicle load such that $F = \delta_F \cdot F_{\text{ref}}$, where $F_{\text{ref}} = 100$ kN is the reference load. This variability is caused by deviation in the transverse position of the vehicle, and is considered as an unknown confounding influence. The quantities involved in the case study along with their prior and ground truth distributions are summarized in Table 3.

Training data for the full and nominal physical models is generated using the FE model of Tatsis and Chatzi, 2019 as a simulator. The FE model is composed of quadrilateral isoparametric plane stress elements with 9 nodes each, arranged in a 200×6 grid. The length, width and height are assumed constant and equal to $L = 25.0$ m, $w = 0.1$ m and $h = 0.6$ m for all of the bridges. The material is linear elastic, with Young's modulus $E = 200$ GPa, density $\rho = 7850$ kg/m^3 and Poisson's ratio $\nu = 0.3$. All supports are modeled as linear springs in the vertical direction. The equations of motion are integrated from $t_0 = 0$ seconds to $t_1 = 25$ seconds with a timestep of $\mathrm{d}t = 0.00025$ seconds, using an implicit Newmark scheme with parameters $\gamma = 1/2$ and $\beta = 1/6$. The deterioration is modeled as a reduction of the cross section width, ranging from 0% (for $y_i = 0.0$) to 90% (for $y_i = 1.0$) in the region spanning $L/20$ around the corresponding support. It is noted that the deterioration is intentionally exaggerated to a large extent to ensure that it can be observed in the strain influence lines. The damaged regions are illustrated in Figure 1(c).

The nominal physics-based model is assumed to be a simplistic but representative model for the behavior of the bridges in the population under study in their nominal condition, obtained through the procedure described in Section 5. In addition to the physics-grounded latent variables, the nominal model also includes the parameter δ_s as an input, describing the offset of the pier relative to the center of the

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

Table 3. *Summary of physics-based, class and domain variables for the two-span bridge case study*

Variable	Unit	Type	Prior distribution	Ground truth	Reference value
$\log_{10} k_{\mathrm{v},1}$	N/m	Physical	$\mathcal{U}(9.0,12.0)$	$\mathcal{U}(9.5,11.5)$	11.5
$\log_{10} k_{\mathrm{v},2}$	N/m	Physical	$\mathcal{U}(9.0,12.0)$	$\mathcal{U}(9.5,11.5)$	11.5
y_1	-	Class	-	$\mathcal{U}(0.0,1.0)$	0.1
y_2	-	Class	-	$\mathcal{U}(0.0,1.0)$	0.1
δ_{v}	-	Domain	-	$\mathcal{U}(0.9,1.1)$	1.0
δ_{s}	m	Domain	-	$\mathcal{U}(-1.0,1.0)$	0.0
δ_{F}	-	Unknown	-	$\mathcal{U}(0.9,1.1)$	1.0

bridge in the longitudinal direction. Given the vertical stiffness of the abutments and support and the offset of the central pier, the nominal model returns a time-series of strains.

The response, domain and class observables are contaminated with i.i.d. samples of Gaussian white noise with standard deviations $\sigma_{\mathrm{x}} = \sigma_{\mathrm{c}} = \sigma_{\mathrm{y}} = 10^{-4}$. It is important to note that this case study is not intended to be a realistic representation of system identification and SHM for bridges, since it circumvents several important practical difficulties such as ensuring the consistency and alignment of data collected over long time-scales from a large number of structures. Furthermore, it is assumed for simplicity that the degradation condition of the bridges does not change significantly within the amount of time required to obtain the dataset.

5.3.2. Qualitative assessment of disentanglement

The model is trained with $\lambda = 1/1024$ and $d_{z_{\mathrm{c}}} = d_{z_{\mathrm{y}}} = 4$. The predictions generated by the model while traversing each of the generative factors are shown in Figure 10. It can be seen that the data-driven component of the decoder is prevented from capturing variability in the reconstructed response when varying $\log_{10} k_{\mathrm{v},1}$, $\log_{10} k_{\mathrm{v},2}$ and δ_{F}, but is able to contribute to the components caused by the variation of the domain and class generative factors. Furthermore, the figure illustrates that the unknown confounder δ_{F} can be partially accounted for by the physics-based model. This is in contrast to the oscillator example (Section 5.2) where the influence of the unknown confounder could not be accounted for by the known physics.

The previous conclusions are further supported by the traversal of the latent space, shown in Figure 11, which indicates that the domain and class subsets of the latent variables encode information that enables the auxiliary decoders to predict the domain and class labels, and the response decoder to correct the physics-based model prediction. It can be seen that the influence of the unknown confounder δ_{F} is partly captured as variability in the physics-based subset $\boldsymbol{z}_{\mathrm{x}}$, indicating that model form uncertainty is compensated by inferring an "effective" value of the physics-grounded latent variables. Figure 11 also suggests that the domain and class latent variables only capture variability in the corresponding generative factors, whereas the physics-grounded latent variables are always active, providing additional evidence for the claim that the adversarial objective induces a disentangled representation while prioritizing the use of the known physics. Importantly, Figures 10 and 11 illustrate that the adversarial training can feasibly constrain $g_\theta(\boldsymbol{z}_{\mathrm{c}}, \boldsymbol{z}_{\mathrm{y}})$, such that it only contributes to the prediction when justified by additional domain and class observables.

5.3.3. Application to damage identification

As discussed in Section 3, the model is trained in a fully supervised manner to simultaneously reconstruct the domain and class variables from the input measurements, making it possible to handle tasks such as damage detection, where predicting the class labels $\boldsymbol{y}$ from input measurements $\boldsymbol{x}$ is of interest. Given a trained model, the condition labels of a similar bridge can be predicted from response measurements. The

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

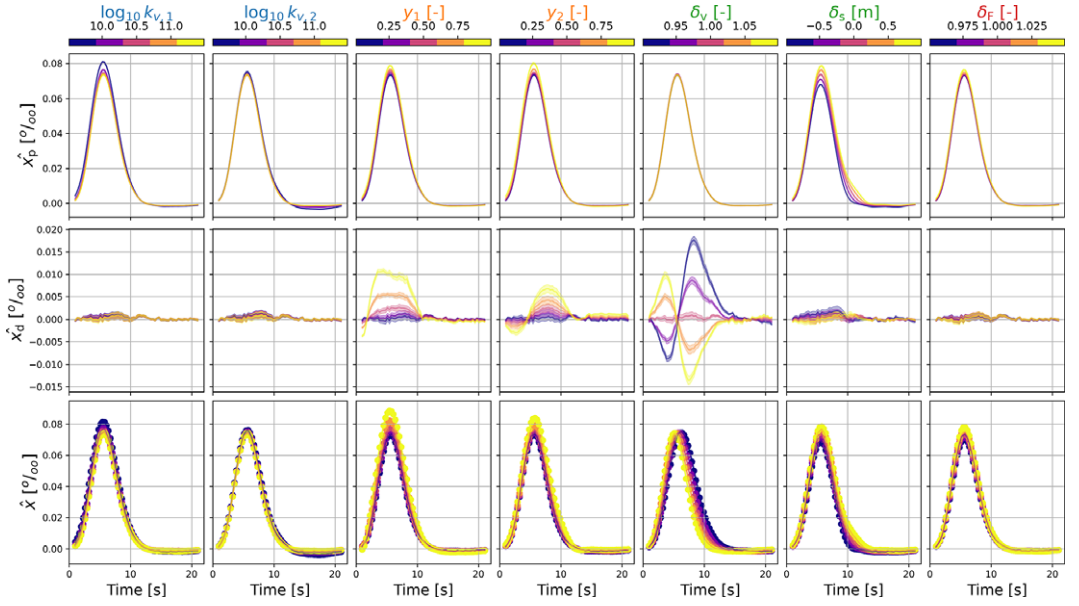


Figure 10. *Mean prediction and $\pm 2\sigma$ uncertainty bounds for the physics-based $\hat{\mathbf{x}}_p$ and data-driven $\hat{\mathbf{x}}_d$ components, and combined prediction $\hat{\mathbf{x}}$ while traversing the generative factors. The input response measurements are denoted as dots in the bottom row.*

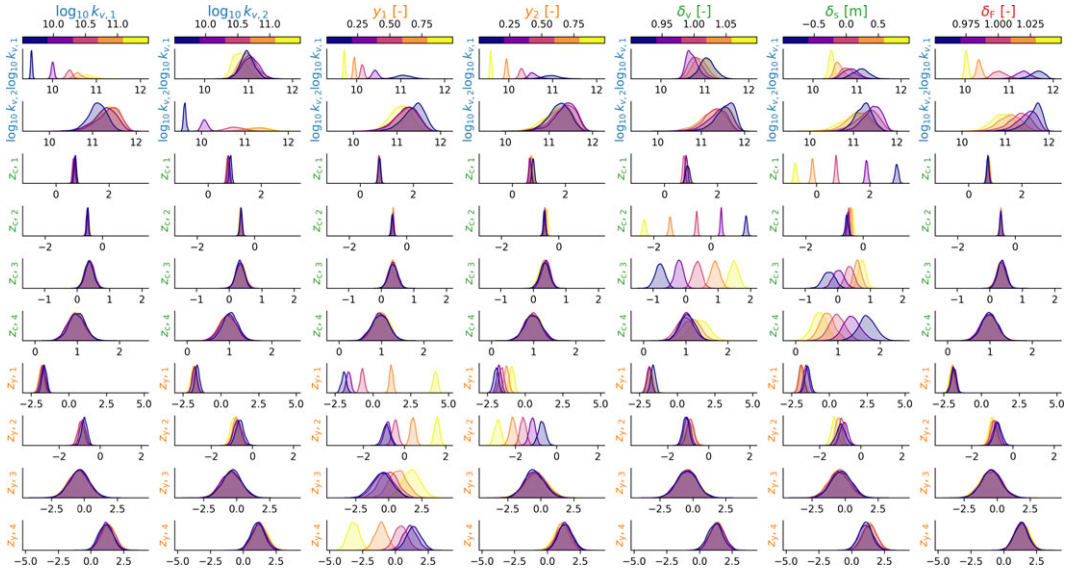


Figure 11. *Visualization of the VAE latent space during traversal of the generative factors. Each column corresponds to variation of a single generative factor, and each row shows the marginal approximate posterior distribution of a single latent variable.*

performance is evaluated in two different cases, illustrated in Figure 12, referred to as "interpolation" and "extrapolation," respectively. For each case, the space of physics-grounded generative factors is subdivided into four quarters. In the interpolation case, the model is trained on $N_{\text{train}} = 1024$ samples from three quarters and evaluated on $N_{\text{test}} = 512$ samples from the fourth. In the extrapolation case, the model is

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

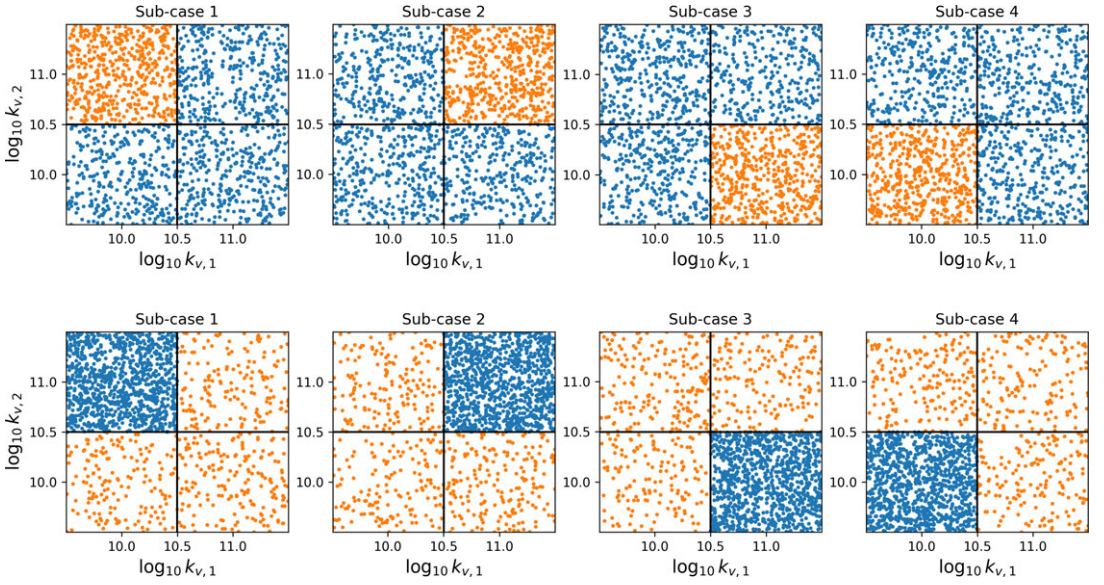


Figure 12. *Samples of physics-grounded generative factors used for creating the synthetic training set (blue) and test set (orange). Two cases are constructed in order to evaluate performance in interpolation (top) and extrapolation (bottom).*

trained on data from a single quarter and evaluated on the remaining three, using the same train and test set sizes. All other generative factors are sampled from the ground truth distributions presented in Table 3. To obtain a more comprehensive evaluation, each of the two cases is divided into four sub-cases, over which the results are averaged.

The proposed disentangled physics-informed variational autoencoder (DPIVAE), using two different hyperparameter settings denoted as DPIVAE-A and DPIVAE-B, is compared with linear regression (LIN), Gaussian process regression (GPR) and a multi-layer perceptron (MLP). For DPIVAE-A the GRL is not utilized, i.e., $\lambda = -1$, and separate encoders are used for each subset of the latent variables. For DPIVAE-B the GRL hyperparameter is taken as $\lambda = 1/1024$. The GPR is implemented with a radial basis function kernel and additive Gaussian white noise. The MLP is formulated with two hidden layers, each with a width of 64 units and a rectified linear unit (ReLU) activation function. Results in terms of interpolation and extrapolation performance of each model is quantified in terms of the R^2 and mean squared error (MSE), shown in Table 4. These results are intended to highlight that the performance of the different models is comparable, and that the proposed model can feasibly be used to predict the class

Table 4. *Mean and standard deviation of* R^2 *and* MSE *for the task of predicting* $\mathbf{y}$*, averaged over* 6 *runs*

	Interpolation		Extrapolation	
Model	$R^2(\uparrow)$	MSE($\downarrow$)	$R^2(\uparrow)$	MSE($\downarrow$)
DPIVAE-A	0.905 ± 0.023	0.008 ± 0.002	0.676 ± 0.153	0.027 ± 0.013
DPIVAE-B	0.943 ± 0.012	0.005 ± 0.001	0.809 ± 0.090	0.016 ± 0.008
GPR	0.957 ± 0.022	0.004 ± 0.002	0.820 ± 0.112	0.015 ± 0.009
LIN	0.863 ± 0.005	0.011 ± 0.000	0.617 ± 0.285	0.032 ± 0.024
MLP	0.839 ± 0.035	0.013 ± 0.003	0.365 ± 0.397	0.053 ± 0.033

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

variables in a complex high-dimensional case study. Manual hyperparameter tuning is performed for the models involved in the comparison.

The proposed approach using adversarial training doesn't result in any improvement over existing approaches for the specific interpolation and extrapolation tasks in the present case study. This can be attributed to the adversarial training, which forces the encoder to weigh any information that is relevant to the prediction of $\boldsymbol{y}$ against the potential improvement it provides towards the reconstruction of $\boldsymbol{x}$. It can be seen that the GPR outperforms all other models while only using a fraction of the parameters, possibly due to the smoothness of the input influence line measurements. The results also indicate that the model performs better when using a single encoder combined with adversarial training in both interpolation and extrapolation. Despite this negative result, we speculate that the disentangled representation induced by the architecture, the invariance of the class latent variables to unknown confounding influences, and the incorporation of the known physics, might be beneficial in certain tasks. Future work will aim to investigate the factors that affect the performance of the proposed approach, and the conditions under which it can provide a benefit in class prediction tasks. This analysis indicates that the proposed model performs on par with other commonly used data-driven models, but with the added benefit of ensuring the proper use of the known physics and the additional interpretability of the physics-grounded latent space.

6. Discussion

6.1. Contributions and strengths

The results presented in Section 5 indicate that the proposed architecture and adversarial objective effectively constrain the posterior distribution over domain and class latent variables, and by extension, the flexibility of the data-driven decoder components. The constraint is controlled by an interpretable hyperparameter that determines the strength of the gradient reversal. This allows for the main and auxiliary decoders to be trained in a collaborative or adversarial manner. This hyperparameter effectively controls the relative importance of the physics-based and data-driven components, and can be used to encourage the model to preferentially utilize the known physics. When the training is adversarial, the domain and class latent spaces encode features of the response measurements that can be related to the observed domain or and class variables, and that cannot be accounted for by the known physics. Simultaneously, the data-driven components of the model are constrained to avoid overriding the physics-based model predictions. Because neither the domain or class observables are necessary during model evaluation, the proposed approach has the potential to reduce the need for cumbersome and expensive data collection methods, such as those involving elaborate experimental procedures or expert assessments.

6.2. Assumptions and limitations

It is important to consider the assumptions and limitations of the proposed approach. One of the main drawbacks of the model is the additivity assumption imposed on the physical, domain and class components of the response. It is expected that the model will perform sub-optimally when this assumption is violated. Furthermore, the accuracy of the inferred physics-grounded latent variables will depend on the relative contribution of the physics, domain and class influences to the measured response. Significant domain and class contributions to the response, or violating the additivity assumption of Equation (2.3), can lead to inaccurate inference of physics-grounded latent variables and large uncertainty in the predictions. Additionally, the model requires that multiple types of data are available, namely measurements of the structural response and information on domain and class. In SHM applications, this might necessitate data alignment procedures of response measurements, environmental conditions and damage level descriptions, and could potentially limit the immediate applicability of the proposed architecture. It is worth mentioning that the interaction between the encoder, decoders, the GRL, and the known physics can be unintuitive in some applications, limiting the applicability of the approach and potentially necessitating implicit supervision by a human expert.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

6.3. Practical considerations

Specifying an appropriate value of λ for a given learning problem is not straightforward. Schemes for scheduling or adaptively tuning the strength of the GRL during training have been proposed (Ganin and Lempitsky, 2015; Li et al., 2023; Qu et al., 2025), but have not been considered in this work. Instead, we focus on providing intuition and clarity regarding the influence of λ through the qualitative and quantitative results presented in Section 5. Furthermore, it is known that adversarial training can be unstable (Wiatrak et al., 2020). Throughout this work, occasional instability and overfitting were observed when using small datasets and large batch sizes. Depending on the case study and the value of the λ hyperparameter, oscillatory behavior may also occur. We found that these issues could be addressed by adjusting the λ hyperparameter, implementing early stopping based on the value of the ELBO on a held-out validation set, and reducing the batch size.

The dimensionality of the latent space is an important design parameter in VAE, and excessively small or large dimensionality can result in poor reconstruction quality (Doersch, 2021). Depending on the available computational budget and problem complexity, approaches for determining an appropriate dimensionality are often based on manual trial and error or grid search (Sejnova et al., 2024). More sophisticated approaches include dynamically adjusting the number of latent variables during optimization (De Boom C et al., 2021; Sejnova et al., 2024), automatic relevance determination (Saha et al., 2025) and multi-stage models (Dai and Wipf, 2019). A key advantage of VAE in engineering, physical, and scientific applications, is that domain knowledge can guide reasoning about the type and number of the dominant generative factors in the data, informing the design of the latent space. VAE are generally insensitive to over-specification of the latent space dimensionality, with superfluous dimensions becoming inactive and ignored by the decoder (Asperti, 2019; Yeung et al., 2017). Choosing the dimensionality of the domain and class latent space to be a multiple of the expected number of generative factors, based on domain knowledge, and subsequently refining this choice by monitoring the number of inactive dimensions after training can therefore be a viable approach.

It is not possible to provide a rule-of-thumb about the amount of data required for effective training. This would be dependent on the specific problem, noise levels, physics-based model, and accuracy of the domain and class information, and also on the particular architecture choices (e.g. the number and depth of layers used in the feed-forward NNs in the encoder and decoder). In some applications, the incorporation of the known physics might lead to a reduction in the data requirements. This has not been investigated in the current paper since it is believed that it would be strongly dependent on the case study chosen, rather than offering any general insight.

7. Conclusions

The present work contributes to the emerging applications of probabilistic generative models in engineering, by investigating disentangled and invariant representation learning as a tool for grounding VAE to the known physics. Specifically, a physics-enhanced machine learning strategy utilizing a VAE architecture is proposed, with the aim of learning a disentangled representation of physical, domain and class confounding influences that are present in the response measurements of physical systems. This is achieved by having the decoder and latent space of the VAE be semantically and functionally separated into data-driven and physics-grounded branches. An easy to implement regularization method based on the GRL is used to constrain the data-driven components, resulting in a model that preferentially utilizes the known physics. An interpretable and intuitive hyperparameter is used to specify the strength of GRL, and wether the model is trained in a collaborative or adversarial manner. Moreover, a strategy for quantifying the type and relative amount of information encoded in different sets of latent variables is proposed, yielding insights on the degree of disentanglement achieved by the model.

Three synthetic case studies involving a beam, an oscillator, and a population of bridges were investigated. In these cases, a nominal model representing the partially known physics was available or built from a simulator. For each case, noisy observations of the structural response and information on

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

domain (the environmental and operational conditions that a system is exposed to) and class (the characteristics of a structure related to the existence and extent of damage and degradation) are assumed available. It was shown that the proposed architecture promotes the learning of disentangled representations, and mitigates the issues that occur when including physics-based components in standard VAE. Furthermore, it was shown that the proposed approach is able to: (i) Preferentially utilize the known physics, resulting in an interpretable and physically meaningful posterior distribution over physics-grounded latent variables, (ii) Accurately reconstruct the structural response in the presence of domain and class influences that are not described by the known physics, and (iii) Predict the class variables associated with a structure under previously unseen conditions using noisy measurements of the structural response.

Although the results of the case studies do not indicate improvement in the prediction of class variables, compared to commonly used data-driven approaches, it is likely that the invariance of the learned domain and class representations with respect to unknown confounding influences can be advantageous for certain problems. Future work will aim to investigate this, as well as the performance of the approach in more complex tasks and in real-world problems. Other possible avenues for future work include the extension of the approach to the semi-supervised setting, the application to dynamical systems described by ordinary differential equations, and automating the tuning of the GRL hyperparameter.

Data availability statement. The code and data (generated via the synthetic use cases) needed to replicate the results shown in this paper can be accessed through the link: https://doi.org/10.5281/zenodo.15813028 (Koune and Cicirello, 2025).

Author contribution. Conceptualization: I.K.; A.C. Methodology: I.K; A.C. Data curation: I.K.; A.C. Data visualisation: I.K.; A.C. Writing original draft: I.K; A.C. All authors approved the final submitted draft.

Funding statement. This publication is part of the project LiveQuay: Live Insights for Bridges and Quay walls (project number NWA.1431.20.002) of the research programme NWA UrbiQuay which is (partly) funded by the Dutch Research Council (NWO). Open access funding provided by Delft University of Technology.

Competing interests. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Achille A and Soatto S (2017) *Information Dropout: Learning Optimal Representations Through Noisy Computation.* arXiv: 1611.01353 [stat.ML] . Available at https://arxiv.org/abs/1611.01353.

Aguerri IE and Zaidi A (2019) *Distributed Variational Representation Learning.* arXiv: 1807.04193 [stat.ML] . Available at https://arxiv.org/abs/1807.04193.

Alemi AA, **Fischer I**, **Dillon JV and Murphy K** (2019) *Deep Variational Information Bottleneck.* arXiv: 1612.00410 [cs.LG] . Available at https://arxiv.org/abs/1612.00410.

Asperti A (2019) *Sparsity in Variational Autoencoders.* arXiv: 1812.07238 [cs.LG] . Available at https://arxiv.org/abs/1812.07238.

Bacsa K, **Liu W**, **Abdallah I and Chatzi E** (2025) Structural dynamics feature learning using a supervised variational autoencoder. *Journal of Engineering Mechanics 151*(2), 04024106. https://doi.org/10.1061/JENMDT.EMENG-7635.

Bengio Y, **Courville A and Vincent P** (2014) *Representation Learning: A Review and New Perspectives.* arXiv: 1206.5538 [cs.LG] . Available at https://arxiv.org/abs/1206.5538.

Blei DM, **Kucukelbir A and McAuliffe JD** (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773.

Bowman SR, **Vilnis L**, **Vinyals O**, **Dai AM**, **Jozefowicz R and Bengio S** (2016) *Generating Sentences from a Continuous Space.* arXiv: 1511.06349 [cs.LG] . Available at https://arxiv.org/abs/1511.06349.

Bull L, **Gardner P**, **Gosliga J**, **Rogers T**, **Dervilis N**, **Cross E**, **Papatheou E**, **Maguire A**, **Campos C and Worden K** (2021) Foundations of population-based SHM, part I: Homogeneous populations and forms. *Mechanical Systems and Signal Processing 148*, 107141. https://doi.org/10.1016/j.ymssp.2020.107141.

Burgess CP, **Higgins I**, **Pal A**, **Matthey L**, **Watters N**, **Desjardins G and Lerchner A** (2018) *Understanding disentangling in β-VAE.* arXiv: 1804.03599 [stat.ML] . Available at https://arxiv.org/abs/1804.03599.

Carmona C and Nicholls G (2020) Semi-modular inference: Enhanced learning in multi-modular models by tempering the influence of components. In Chiappa S and Calandra R (eds), *Proceedings of the Twenty Third International Conference on*

Artificial Intelligence and Statistics. Vol *108*. Proceedings of Machine Learning Research. PMLR, pp. 4226–4235. Available at https://proceedings.mlr.press/v108/carmona20a.html.

Chen RTQ, **Li X**, **Grosse RB and Duvenaud DK** (2018) Isolating sources of disentanglement in Variational autoencoders. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.), *Advances in Neural Information Processing Systems*. Vol. *31*. Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper_files/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Cicirello A (2024) *Physics-Enhanced Machine Learning: a position paper for dynamical systems investigations.* arXiv: 2405.05987[cs.LG] . Available at https://arxiv.org/abs/2405.05987.

Coraça EM, **Ferreira JV and Nóbrega EG** (2023) An unsupervised structural health monitoring framework based on Variational autoencoders and hidden Markov models. *Reliability Engineering & System Safety 231*, 109025. https://doi.org/10.1016/j.ress.2022.109025.

Cover TM and Thomas JA (2006) *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, p. 0471241954

Cross EJ, **Gibson SJ**, **Jones MR**, **Pitchforth DJ**, **Zhang S and Rogers TJ** (2022) Physics-informed machine learning for structural health monitoring. In *Structural Health Monitoring Based on Data Science Techniques*, Cury A, Ribeiro D, Ubertini F and Todd MD (eds.), Cham: Springer International Publishing, 347–367. https://doi.org/10.1007/978-3-030-81716-9_17.

Dai B and Wipf D (2019) *Diagnosing and Enhancing VAE Models.* arXiv: 1903.05789[cs.LG] . Available at https://arxiv.org/abs/1903.05789.

De Boom C, **Wauthier S**, **Verbelen T and Dhoedt B** (2021) Dynamic narrowing of VAE bottlenecks using GECO and L0 regularization. *International Joint Conference on Neural Networks (IJCNN) 2021*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533671.

Debbagh M (2023) *Learning Structured Output Representations from Attributes using Deep Conditional Generative Models.* arXiv: 2305.00980[cs.CV] . Available at https://arxiv.org/abs/2305.00980.

Ding Z, **Xu Y**, **Xu W**, **Parmar G**, **Yang Y**, **Welling M and Tu Z** (2020) Guided Variational autoencoder for disentanglement learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Doersch C (2021) *Tutorial on Variational Autoencoders.* arXiv: 1606.05908[stat.ML] . Available at https://arxiv.org/abs/1606.05908.

Esmaeili B, **Wu H**, **Jain S**, **Bozkurt A**, **Siddharth N**, **Paige B**, **Brooks DH**, **Dy J and Meent JW van de** (2019) Structured disentangled representations. In Chaudhuri K and Sugiyama M (eds), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. *89*. Proceedings of Machine Learning Research. PMLR, pp. 2525–2534. Available at https://proceedings.mlr.press/v89/esmaeili19a.html.

Federici M, **Dutta A**, **Forré P**, **Kushman N and Akata Z** (2020) *Learning Robust Representations via Multi-View Information Bottleneck.* arXiv: 2002.07017[cs.LG] . Available at https://arxiv.org/abs/2002.07017.

Fischer I (2020) The conditional entropy bottleneck. *Entropy 22*(9), 999. https://doi.org/10.3390/e22090999.

Ganin Y and Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In Bach F and Blei D (eds.), *Proceedings of the 32nd International Conference on Machine Learning*. Vol. *37*. Lille, France: Proceedings of Machine Learning Research: PMLR, pp. 1180–1189. Available at https://proceedings.mlr.press/v37/ganin15.html.

Ganin Y, **Ustinova E**, **Ajakan H**, **Germain P**, **Larochelle H**, **Laviolette F**, **Marchand M and Lempitsky V** (2016) *Domain-Adversarial Training of Neural Networks.* arXiv: 1505.07818[stat.ML] . Available at https://arxiv.org/abs/1505.07818.

Glyn-Davies A, **Vadeboncoeur A**, **Akyildiz OD**, **Kazlauskaite I and Girolami M** (2025) A primer on variational inference for physics-informed deep generative modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 383*(2299), 20240324. https://doi.org/10.1098/rsta.2024.0324.

Goh H, **Sheriffdeen S**, **Wittmer J and Bui-Thanh T** (2022) Solving Bayesian inverse problems via Variational autoencoders. In Bruna J, Hesthaven J and Zdeborova L (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Vol. *145*. Proceedings of Machine Learning Research. PMLR, pp. 386–425. Available at https://proceedings.mlr.press/v145/goh22a.html.

Goodfellow IJ, **Pouget-Abadie J**, **Mirza M**, **Xu B**, **Warde-Farley D**, **Ozair S**, **Courville A and Bengio Y** (2014) *Generative Adversarial Networks.* arXiv: 1406.2661[stat.ML] . Available at https://arxiv.org/abs/1406.2661.

Hadad N, **Wolf L and Shahar M** (2018) A two-step disentanglement method. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haywood-AIexander M, **Liu W**, **Bacsa K**, **Lai Z and Chatzi E** (2024) Discussing the spectrum of physics-enhanced machine learning: A survey on structural mechanics applications. *Data-Centric Engineering 5*, e30. https://doi.org/10.1017/dce.2024.33.

Higgins I, **Matthey L**, **Pal A**, **Burgess CP**, **Glorot X**, **Botvinick MM**, **Mohamed S and Lerchner A** (2016) Beta-VAE: Learning basic visual concepts with a constrained Variational framework. *International Conference on Learning Representations*.

Hwang H, **Kim GH**, **Hong S and Kim KE** (2020) Variational interaction information maximization for cross-domain disentanglement. In Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds.), *Advances in Neural Information Processing Systems*. Vol *33*. Curran Associates, Inc., 22479–22491. Available at https://proceedings.neurips.cc/paper_files/paper/2020/file/fe663a72b27bdc613873fbbb512f6f67-Paper.pdf.

Ilse M, **Tomczak JM**, **Louizos C and Welling M** (2020) DIVA: Domain invariant Variational autoencoders. In Arbel T, Ben Ayed I, Bruijne M de, Descoteaux M, Lombaert H and Pal C (eds.), *Proceedings of the Third Conference on Medical Imaging with Deep*

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

Learning. Vol. *121*. Proceedings of Machine Learning Research. PMLR, pp. 322–348. Available at https://proceedings.mlr.press/v121/ilse20a.html.

Joy T, **Schmon SM**, **Torr PHS**, **Siddharth N and Rainforth T** (2022) *Capturing Label Characteristics in VAEs*. arXiv: 2006.10102 [cs.LG] . Available at https://arxiv.org/abs/2006.10102.

Kamariotis A, **Vlachas K**, **Ntertimanis V**, **Koune I**, **Cicirello A and Chatzi E** (2024) On the consistent classification and treatment of uncertainties in structural health monitoring applications. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg 11*(1), 011108. https://doi.org/10.1115/1.4067140.

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294.

Kim H and Mnih A (2018) Disentangling by factorising. In Dy J and Krause A (eds.), *Proceedings of the 35th International Conference on Machine Learning*. Vol *80*. Proceedings of Machine Learning Research. PMLR, pp. 2649–2658. Available at https://proceedings.mlr.press/v80/kim18b.html.

Kingma DP and Ba J (2017) *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG] . Available at https://arxiv.org/abs/1412.6980.

Kingma DP and Welling M (2019) An introduction to Variational autoencoders. *Foundations and Trends® in Machine Learning 12*(4), 307–392. https://doi.org/10.1561/2200000056.

Kingma DP and Welling M (2022) *Auto-Encoding Variational Bayes*. arXiv: 1312.6114 [stat.ML] . Available at https://arxiv.org/abs/1312.6114.

Kiureghian AD and Ditlevsen O (2009) Aleatory or epistemic? Does it matter? *Structural Safety 31*(2), 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.

Koune IC and Cicirello A (2025) *Replication Data for: Adversarial Disentanglement by Backpropagation with Physics-Informed Variational Autoencoder*. Available at https://doi.org/10.5281/zenodo.15813028.

Larsen ABL, **Sønderby SK**, **Larochelle H and Winther O** (2016) Autoencoding beyond pixels using a learned similarity metric. In Balcan MF and Weinberger KQ (eds.), *Proceedings of The 33rd International Conference on Machine Learning*. Vol. *48*. New York: Proceedings of Machine Learning Research. PMLR, pp. 1558–1566. Available at https://proceedings.mlr.press/v48/larsen16.html.

Li C, **Li Z**, **Sun J**, **Zhang Y**, **Jiang X and Zhang F** (2023) Dynamic weighted gradient reversal network for visible-infrared person re-identification. *ACM Transactions on Multimedia Computing Communications and Applications 20*(1), 1551–6857. https://doi.org/10.1145/3607535.

Linial O, **Ravid N**, **Eytan D and Shalit U** (2021) Generative ODE modeling with known unknowns. In *Proceedings of the Conference on Health, Inference, and Learning*. ACM CHIL '21. ACM. Available at https://doi.org/10.1145/3450439.3451866.

Locatello F, **Bauer S**, **Lucic M**, **Rätsch G**, **Gelly S**, **Schölkopf B and Bachem O** (2019) *Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations*. arXiv: 1811.12359 [cs.LG] . Available at https://arxiv.org/abs/1811.12359.

Lopez R, **Regier J**, **Jordan MI and Yosef N** (2018) *Information Constraints on Auto-Encoding Variational Bayes*. arXiv: 1805.08672 [cs.LG] . Available at https://arxiv.org/abs/1805.08672.

Louizos C, **Swersky K**, **Li Y**, **Welling M and Zemel R** (2017) *The Variational Fair Autoencoder*. arXiv: 1511.00830 [stat.ML] . Available at https://arxiv.org/abs/1511.00830.

Mao J, **Wang H and Spencer Jr BF** (2021) Toward data anomaly detection for automated structural health monitoring: Exploiting generative adversarial nets and autoencoders. *Structural Health Monitoring 20*(4), 1609–1626. eprint: https://doi.org/10.1177/1475921720924601.

Mathieu E, **Rainforth T**, **Siddharth N and Teh YW** (2019) *Disentangling Disentanglement in Variational Autoencoders*. arXiv: 1812.02833 [stat.ML] . Available at https://arxiv.org/abs/1812.02833.

Mondal AK, **Sailopal A**, **Singla P and AP** (2023) SSDMM-VAE: Variational multi-modal disentangled representation learning. *Applied Intelligence 53*, 8467–8481.

Moyer D, **Gao S**, **Brekelmans R**, **Galstyan A and Ver Steeg G** (2018) Invariant representations without adversarial training. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.), *Advances in Neural Information Processing Systems*. Vol. *31*. Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper_files/paper/2018/file/415185ea244ea2b2bedeb0449b926802-Paper.pdf.

N S, **Paige B**, **Meent JW van de**, **Desmaison A**, **Goodman N**, **Kohli P**, **Wood F and Torr P** (2017) Learning disentangled representations with semi-supervised deep generative models. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*. Vol. *30*. Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper_files/paper/2017/file/9cb9ed4f35cf7c2f295cc2bc6f732a84-Paper.pdf.

Qu L, **Weber C**, **Wang W**, **Jin J**, **Gao Y**, **Li T and Wermter S** (2025) Disentanglement of prosody representations via diffusion models and scheduled gradient reversal. *IEEE Transactions on Neural Networks and Learning Systems 36*(8), 15043–15054. https://doi.org/10.1109/TNNLS.2025.3534822.

Rezende DJ and Mohamed S (2016) *Variational Inference with Normalizing Flows*. arXiv: 1505.05770 [stat.ML] . Available at https://arxiv.org/abs/1505.05770.

Rixner M and Koutsourelakis PS (2021) A probabilistic generative model for semi-supervised training of coarse-grained surrogates and enforcing physical constraints through virtual observables. *Journal of Computational Physics 434*, 110218. https://doi.org/10.1016/j.jcp.2021.110218.

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

Saha S, **Joshi S and Whitaker R** (2025) ARD-VAE: A statistical formulation to find the relevant latent dimensions of Variational autoencoders. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025*, 889–898. https://doi.org/10.1109/WACV61041.2025.00096.

Sejnova G, **Vavrecka M and Stepanova K** (2024) Adaptive compression of the latent space in variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN2024*. Springer Nature Switzerland, pp. 89–101. Available at https://doi.org/10.1007/978-3-031-72332-2_7.

Shin H and Choi M (2023) Physics-informed variational inference for uncertainty quantification of stochastic differential equations. *Journal of Computational Physics 487*, 112183. https://doi.org/10.1016/j.jcp.2023.112183. https://www.sciencedirect.com/science/article/pii/S0021999123002784.

Sun H, **Pears N and Gu Y** (2022) Information bottlenecked Variational autoencoder for disentangled 3D facial expression modelling. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022*, 2334–2343. http://doi.org/10.1109/WACV51458.2022.00239.

Takeishi N and Kalousis A (2021) Physics-integrated Variational autoencoders for robust and interpretable generative Modeling. In Ranzato M, Beygelzimer A, Dauphin Y, Liang P and Vaughan JW (eds.), *Advances in Neural Information Processing Systems*. Vol. *34*. Curran Associates, Inc. pp. 14809–14821. Available at https://proceedings.neurips.cc/paper_files/paper/2021/file/7ca57a9f85a19a6e4b9a248c1daca185-Paper.pdf.

Tatsis K and Chatzi E (2019) A numerical benchmark for system identification under operational and environmental variability. In *8th IOMAC – International Operational Modal Analysis Conference.* pp. 101–106.

Tishby N, **Pereira FC and Bialek W** (2000) *The Information Bottleneck Method.* arXiv: physics/0004057[physics.data-an] . Available at https://arxiv.org/abs/physics/0004057.

Tonolini F, **Jensen BS and Murray-Smith R** (2020) Variational sparse coding. In Adams RP and Gogate V (eds), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. vol *115*. Proceedings of Machine Learning Research. PMLR, pp. 690–700. Available at https://proceedings.mlr.press/v115/tonolini20a.html.

Tschannen M, **Bachem O and Lucic M** (2018) *Recent Advances in Autoencoder-Based Representation Learning.* arXiv: 1812.05069[cs.LG] . Available at https://arxiv.org/abs/1812.05069.

Tsialiamanis G, **Wagg DJ**, **Dervilis N and Worden K** (2021) On generative models as the basis for digital twins. *Data-Centric Engineering 2*, e11. https://doi.org/10.1017/dce.2021.13.

von Rueden L, **Mayer S**, **Beckh K**, **Georgiev B**, **Giesselbach S**, **Heese R**, **Kirsch B**, **Pfrommer J**, **Pick A**, **Ramamurthy R**, **Walczak M**, **Garcke J**, **Bauckhage C and Schuecker J** (2023) Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering 35*(1), 614–633. https://doi.org/10.1109/TKDE.2021.3079836.

Vowels MJ, **Camgoz NC and Bowden R** (2019) *Gated Variational AutoEncoders: Incorporating Weak Supervision to Encourage Disentanglement.* arXiv: 1911.06443 [cs.CV] . Available at https://arxiv.org/abs/1911.06443.

Walker E, **Trask N**, **Martinez C**, **Lee K**, **Actor JA**, **Saha S**, **Shilt T**, **Vizoso D**, **Dingreville R and Boyce BL** (2024) *Unsupervised physics-informed disentanglement of multimodal data.* https://doi.org/10.3934/fods.2024019.

Wang X and Xia Y (2022) Knowledge transfer for structural damage detection through re-weighted adversarial domain adaptation. *Mechanical Systems and Signal Processing 172*, 108991. https://doi.org/10.1016/j.ymssp.2022.108991.

Watanabe S (1960) Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development 4*(1), 66–82. http://doi.org/10.1147/rd.41.0066.

Wiatrak M, **Albrecht SV and Nystrom A** (2020) *Stabilizing Generative Adversarial Networks: A Survey.* arXiv: 1910.00927 [cs.LG] . Available at https://arxiv.org/abs/1910.00927.

Yeung S, **Kannan A**, **Dauphin Y and Fei-Fei L** (2017) *Tackling Over-pruning in Variational Autoencoders.* arXiv: 1706.03643 [cs.LG] . Available at https://arxiv.org/abs/1706.03643.

Yildiz C, **Heinonen M and Lahdesmaki H** (2019) ODE2VAE: Deep generative second order ODEs with Bayesian neural networks. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E and Garnett R (eds.), *Advances in Neural Information Processing Systems*. Vol. *32*. Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper_files/paper/2019/file/99a401435dcb65c4008d3ad22c8cdad0-Paper.pdf.

Yu X, **Nott DJ and Smith MS** (2022) *Variational inference for cutting feedback in misspecified models.* arXiv: 2108.11066 [stat. ME] . Available at https://arxiv.org/abs/2108.11066.

Zhao S, **Song J and Ermon S** (2018) *InfoVAE: Information Maximizing Variational Autoencoders.* arXiv: 1706.02262[cs.LG] . Available at https://arxiv.org/abs/1706.02262.

Zhong W and Meidani H (2023) PI-VAE: Physics-informed Variational auto-encoder for stochastic differential equations. *Computer Methods in Applied **Mechanics** and Engineering 403*, 115664. https://doi.org/10.1016/j.cma.2022.115664.

A. Implementation details

Encoder formulation. The encoder reduces the dimensionality of the input measurements and maps them to vectors of mean values $\boldsymbol{\mu}_\phi(\boldsymbol{x})$, standard deviations $\boldsymbol{\sigma}_\phi(\boldsymbol{x})$ and a lower triangular matrix $\boldsymbol{L}'_\phi(\boldsymbol{x})$. The lower triangular factor $\boldsymbol{L}_\phi(\boldsymbol{x}) = \boldsymbol{L}'_\phi(x) + \boldsymbol{\sigma}_\phi(\boldsymbol{x})\boldsymbol{I}$ is the Cholesky decomposition factor of the covariance matrix $\Sigma_\phi(\boldsymbol{x})$, i.e. $\Sigma_\phi(\boldsymbol{x}) = \boldsymbol{L}_\phi(\boldsymbol{x})\boldsymbol{L}_\phi(\boldsymbol{x})^T$, such that the posterior distribution corresponding to each input is a multivariate Normal distribution. The encoder outputs the log of the standard deviations, which are

https://doi.org/10.1017/dce.2025.10028 Published online by Cambridge University Press

then exponentiated to avoid negative values. The reparametrization trick (Kingma and Welling, 2022) is exploited to define the latent variables $\boldsymbol{z}$ as a deterministic transformation of a noise variable $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$. This facilitates the computation of unbiased Monte Carlo gradient estimates of the objective with respect to the variational parameters, using automatic differentiation. With the exception of the introductory examples presented in Section 2.4, the encoder is everywhere formulated as a single feed-forward NN using a shallow architecture with a single hidden layer. The input and hidden layer widths are $[d_{\mathrm{x}}, 128]$, where d_{x} is the dimensionality of the input. The output layer is composed of three heads, corresponding to the mean, standard deviation and covariance outputs. The mean and standard deviation heads have output sizes of d_{z}, while the covariance head has an output size of d_{z}^2, where $d_{\mathrm{z}} = d_{\mathrm{z_x}} + d_{\mathrm{z_c}} + d_{\mathrm{z_y}}$, with $d_{\mathrm{z_i}}$ denoting the size of the i'th subset of the latent space. A ReLU activation function is applied on all layers except the final output layer. For the introductory examples, an independent encoder network is used for each subset of the latent variables. The hidden layer widths of each independent network are set to 64 units, and the output shapes are adjusted according to the dimensionality of the corresponding subset of the latent variables. To further ensure numerical stability, the outputs of all encoder NNs are clamped within ranges of values that are expected to be encountered for the case studies investigated in this work.

Decoders. The decoder of the response is formulated as a feed-forward NN with a single, 128-unit-wide hidden layer and a ReLU nonlinearity at the output of the hidden layer. The size of the input is $d_{\mathrm{z_c}} + d_{\mathrm{z_y}}$ and the output size is d_{x}. A gradient reversal layer is placed at the input of this network. For the structural response prediction, the standard deviation σ_{x} is included in the vector $\boldsymbol{\theta}$ and jointly optimized with the NN hyperparameters. The auxiliary networks are formulated with a single hidden layer with a width of 64 units and a ReLU nonlinearity between the input and the hidden layer. The auxiliary decoders are composed of two prediction heads, responsible for the mean prediction and standard deviation respectively. The input and output shapes are d_{z_i} and $2 \cdot d_i$, where i denotes the corresponding domain or class modality.

Conditional prior networks. The conditional prior distributions are formulated as factorized Gaussian distributions. The corresponding neural networks use a single hidden layer with a width of 64 units. The input and output shapes are also adjusted to d_{i} and d_{z_i} respectively, where the subscript $i \in \{x, c, y\}$ denotes the corresponding modality and subset of the latent space.

Latent variable transformation. To facilitate the application of the model to cases involving physics-grounded latent variables with bounded support, and to improve numerical stability, all parameters are transformed from an unbounded and normalized base latent space to the target latent space in which they are defined. This is achieved by applying a sequence of deterministic transformations to the samples and corresponding scaling of the log-densities. In the following, variables in the base space are denoted as u. The samples at the output of the encoder are first bounded by applying the logistic transform $u' = \frac{1}{1+e^{-u}}$, and subsequently scaled and shifted using an affine transform $z = u' \cdot (\mathrm{UB} - \mathrm{LB}) + \mathrm{LB}$ to bound the variables to their specified supports defined by the lower and upper bound LB and UB. The samples and densities can also be mapped from the target latent space to the base latent space by applying the corresponding inverse transforms in reverse order.

Optimization. Optimization is carried out using the Adam algorithm (Kingma and Ba, 2017) with minibatch gradient estimation (Kingma and Welling, 2022). The model is trained for up to 20,000 iterations with a batch size of 64. Early stopping is implemented by monitoring the value of the ELBO, evaluated on a held-out validation set with size $N_{\mathrm{val}} = 512$. The training is terminated if no improvement of the ELBO is observed over 2,000 iterations. Gradient and objective estimates are obtained using 16 Monte Carlo samples during training, 64 during validation, and 512 during evaluation, although in practice the training was found to be insensitive to the number of samples. All learning rates are set to 0.001, except for the learning rate of the standard deviation parameter for the response σ_{x}, which is set to 0.005. The α and β hyperparameters of the optimization objective are taken as $\beta = \alpha_{\mathrm{x}} = \alpha_{\mathrm{c}} = \alpha_{\mathrm{y}} = 1.0$ for all the experiments presented in this work.

Visualization. Figures illustrating the traversal of the latent space and the space of reconstructions are provided for each case study. Samples from the latent space and reconstructions are obtained as follows: Five linearly spaced values between the 1st and 99th percentile of the ground truth distribution are computed for each generative factor in turn, while the remaining generative factors are fixed to a constant value. For each combination of generative factors, 1000 realizations of response measurements are generated using the procedure described in Section 5. The model is evaluated on the response measurements, and a single sample is drawn from the approximate posterior distribution for each response measurement. The decoder is then evaluated on each sample from the posterior, yielding deterministic predictions $\hat{\boldsymbol{x}}_{\mathrm{p}}$ and $\hat{\boldsymbol{x}}_{\mathrm{d}}$ from the physics-based and data-driven components respectively. The combined prediction is sampled from $\mathcal{N}\left(\hat{\boldsymbol{x}}_{\mathrm{p}} + \hat{\boldsymbol{x}}_{\mathrm{d}}, \sigma_{\mathrm{x}}^2 \boldsymbol{I}\right)$. The visualizations of the latent space and reconstructions therefore also include the randomness in the data generating process, in addition to the randomness in the approximate posterior distribution and decoder.

Cite this article: Koune IC and Cicirello A (2025). Adversarial disentanglement by backpropagation with physics-informed variational autoencoder. *Data-Centric Engineering*, 6, e50. doi:10.1017/dce.2025.10028