Disentangled representation learning with physics-informed variational autoencoder for structural health monitoring

Koune, Ioannis; Cicirello, Alice

DOI
10.58286/29862

Publication date
2024

Document Version
Final published version

Published in
e-Journal of Nondestructive Testing

Citation (APA)
Koune, I., & Cicirello, A. (2024). Disentangled representation learning with physics-informed variational autoencoder for structural health monitoring. *e-Journal of Nondestructive Testing*, 1-11. https://doi.org/10.58286/29862

eJNDT e-Journal of Nondestructive Testing - ISSN 1435-4934 - www.ndt.net

EWSHM 2024
11th European Workshop on Structural Health Monitoring

Disentangled representation learning with physics-informed variational autoencoder for structural health monitoring

Ioannis KOUNE[1], Alice CICIRELLO[2]
[1] Technical University of Delft, Delft, The Netherlands, i.c.koune@tudelft.nl
[2] University of Cambridge, Cambridge, United Kingdom, ac685@cam.ac.uk

Abstract. Manual inspection and assessment of structures on a large scale is labour intensive and often infeasible, while data-driven machine learning techniques can fail to identify relevant failure mechanisms and suffer from poor generalization to previously unseen conditions, particularly when limited information is available. We propose a physics-informed variational autoencoder formulation for disentangled representation learning of confounding sources in the measurements with the aim of computing the posterior distribution of latent parameters of a physics-based model and predicting the response of a structure when limited measurements are available. The latent space of the autoencoder is augmented with a set of physics-based latent variables that are interpretable and allow for domain knowledge in the form of prior distributions and physics-based models to be included in the autoencoder formulation. To prevent the data-driven components of the model from overriding the known physics, a regularization term is included in the training objective that imposes constraints on the latent space and the generative model prediction. The feasibility of the proposed approach is evaluated on a synthetic case study.

Keywords: Generative models, variational autoencoders, structural health monitoring, physics-informed machine learning, disentangled representation learning.

Introduction

In the context of Structural Health Monitoring (SHM), three main tasks can typically be carried out: inferring the distribution of a set of latent variables (i.e. variables that are not directly observable) that best describe the parameters of a structure to identify its current health state; predicting the remaining useful life a structure, and predicting what the condition of a structure would have been if some rehabilitation measure were applied. Generative probabilistic models such as Normalizing Flows [1], Variational Autoencoders (VAE) [2] and Generative Adversarial Networks (GANs) [3] are a class of models, typically based on Neural Network (NN) and Deep Learning (DL) architectures, that can approximate the distribution of a given set of data and generate samples from the learned distribution. Generative probabilistic models have recently seen a broader use in performing SHM inferential and predictive tasks, and for constructing Digital Twins (DTs) of structures [4]. In the context of SHM, using these data-driven approaches might lead to poor generalization

Media and Publishing Partner
https://doi.org/10.58286/29862

performance and results that are inconsistent with known physics. Physics-enhanced machine learning approaches (also known as hybrid or grey-box models) overcome these limitations by utilizing previous knowledge in the form of informative prior distributions and physics-based models combined with measurements in order to reduce the required amount of data, improve accuracy and generalization performance and ensure the consistency of inferred quantities and model predictions with the known physics [5]. Importantly, incorporating physics-based terms in the predictions can yield models that are robust and explainable, representing physically meaningful and interpretable quantities. However, the same flexibility and learning capacity of NNs that enables them to accurately model physical processes from data can be problematic when combining them with physics-based models. Specifically, when attempting to minimize some discrepancy measure between measurements and predictions, the flexibility of overparametrized NNs can result in the data-driven components overriding the physics-based model, leading to inaccurate inference of latent variables and unrealistic and overconfident predictions [6].

In this work, we propose a VAE formulation for disentangled representation learning, enabling the identification and disentanglement of underlying confounding factors (including environmental and operational inputs, damage conditions, and other factors) hidden in the structural response measurements (observable data). This approach enables learning disentangled representations of the contributing factors in the measurements that are used to build an explainable, controllable and robust digital twin with enhanced generalisation performance. The approach is dependent on the availability of a physics-based model of the underlying structure, and labelled structural response measurements, environmental data, condition assessment data, and other relevant labelled data. This labelled training dataset is used to extract a low dimensional encoding of features in the structural response measurements that are informative about environmental and operational conditions, and damage in the structure. This is achieved by having the encoder, decoder and latent space of the VAE be semantically and functionally separated into a data-driven branch and physics-grounded branch, trained jointly in an end-to-end fashion. The VAE architecture and the training objective are further modified to prevent the data-driven branch from learning and compensating for the discrepancies between the physics-grounded branch predictions and the measured structural response.

1. Methods and Tools

Throughout this text, capital symbols denote random variables and lower case symbols denote deterministic quantities or realizations of random variables. Light symbols denote scalars while symbols in bold refer to vector or matrix quantities. Observable random variables or deterministic quantities that are available during testing will be denoted by X, labels that are only available during training by Y, and latent variables by Z. The subscripts p and d are used to denote components of the physics-grounded branch and data-driven branch respectively.

1.1 Bayesian inference

The posterior distribution of latent variables $\boldsymbol{Z}$ given data $\boldsymbol{X}$ is [7]:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})/p(\boldsymbol{X}) \tag{1}$$

When physical knowledge in the form of an analytical or numerical (e.g. finite element) physics-based model is available, the likelihood function $p(\boldsymbol{X}|\boldsymbol{Z})$ can be expressed as a combination of a physics-based model and a probabilistic model. Prior knowledge on the latent variables (e.g. from previous experiments, literature or expert judgement) can be included in the prior distribution $p(\boldsymbol{Z})$. The marginal likelihood (i.e. the evidence) term in the denominator is often a high dimensional integral, and can quantify the relative strength of different models.

1.2 Variational Autoencoder

In Variational Inference (VI), the posterior distribution is approximated using a prescribed parametrized family of distributions. The optimal parameters are obtained by minimizing the KL-divergence between the true and approximate posteriors [8]. A VAE [2] can be utilized to perform approximate VI in order to learn the joint distribution $p(\boldsymbol{X}, \boldsymbol{Z})$. The VAE is composed of an *encoder* network $q_{\phi}(\boldsymbol{Z}|\boldsymbol{X})$ and a decoder network $p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{Z})$, parametrized by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ respectively. 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 X conditioned on Z. Sampling $\boldsymbol{z}{\sim}q_{\phi}(\boldsymbol{Z}|\boldsymbol{X})$ and evaluating the decoder yields samples from the learned distribution of the data, which in the context of SHM can be used for downstream tasks such as damage detection and remaining useful life assessment. Optimization is performed by maximizing a lower bound on the marginal likelihood of the data known as the Evidence LOwer Bound (ELBO) denoted here by $\mathcal{L}$.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathcal{L}_{\mathrm{VAE}}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\mathrm{q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})) \quad (2)$$

1.3 Conditional Variational Autoencoder

Conditional generative models are a class of generative models aimed at performing representation learning and structured prediction on (typically) high-dimensional outputs $\boldsymbol{Y}$ conditioned on a set of input variables $\boldsymbol{X}$. In the Conditional VAE (CVAE) architecture [9] the outputs are passed to the encoder that maps the input data to the latent space, conditioning the latent distribution on both $\boldsymbol{X}$ and $\boldsymbol{Y}$. Similarly, the decoder mapping from $\boldsymbol{Z}$ to $\boldsymbol{Y}$ is conditioned on $\boldsymbol{X}$**,** resulting in the following objective**:**

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathcal{L}_{\mathrm{CVAE}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\mathrm{q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) \quad (3)$$

2. Background and objectives

2.1 Problem description

We aim to construct a probabilistic generative model for a structure that, given a set of observations of the structural response, environmental parameters and damage, can be used to infer the posterior distribution of a set of latent variables of the physics-based model and learning some additional latent parameters of a data-driven part to generate samples of the future structural response and predict the damage condition of the structure. In practical applications, there is often a significant discrepancy between the structural response predicted by physics-based models and the measured response of the structure. Several different sources of uncertainty may contribute to this discrepancy, with the most significant

ones often being [10, 11]: (i) measurement uncertainty stemming from the presence of noise and sensor error in the measurements; (ii) parameter uncertainty, i.e. uncertainty in the values of the parameters of the physics-based model; (iii) model form uncertainty, i.e. uncertainty from simplifications and approximations in the physics-based model of a structure; (iv) uncertainty due to the influence of environmental conditions that are not considered in the physics-based model (unmeasurable signals that affect the output); (v) uncertainty caused by degradation and damage in the structure. If the physics-based model is a good approximation of the real behaviour of the physical system in its nominal condition, the model prediction error can indicate the type of damage affecting the system or the mechanism through which an environmental effect influences the response. In what follows, it is assumed that the damage conditions and environmental effects that are expected to affect the structure will have an influence in the measured response.

We seek to approximate the structural response as the sum of a physics-based model $f_{\mathrm{p}}(\boldsymbol{z}_{\mathrm{p}};\boldsymbol{\theta}_{\mathrm{p}})$, and a trainable parametrized function $f_{\mathrm{d}}(\boldsymbol{z}_{\mathrm{d}};\boldsymbol{\theta}_{\mathrm{d}})$ that captures the contribution of the environmental conditions and damage to the measured structural response. Furthermore, we assume that a physics-based model with some uncertain parameters (latent variables) for which prior knowledge is available. A set of measurements is obtained for a single structure, yielding N samples of the matrix $\boldsymbol{D} = \{\boldsymbol{X}_{\mathrm{r}}, \boldsymbol{X}_{\mathrm{e}}, \boldsymbol{Y}_{\mathrm{c}}\}$, where $\boldsymbol{X}_{\mathrm{r}} = (\boldsymbol{x}_{\mathrm{r},1}, \boldsymbol{x}_{\mathrm{r},2} \dots \boldsymbol{x}_{\mathrm{r,N}})$ and respectively for $\boldsymbol{X}_{\mathrm{e}}$ and $\boldsymbol{Y}_{\mathrm{c}}$. The vector $\boldsymbol{x}_{\mathrm{r,i}} \in \mathbb{R}^{d_{\mathrm{r}}}$ denotes a single measurement of the structural response from d_{r} sensors, while $\boldsymbol{x}_{\mathrm{e,i}} \in \mathbb{R}^{d_{\mathrm{e}}}$ is a vector of the environmental conditions, and $\boldsymbol{y}_{\mathrm{c,i}} \in \mathbb{R}^{d_c}$ is a vector of condition assessment measurements quantifying the existence and severity of damage in the structure (e.g. from inspections). The dimensionalities of the response, environmental and damage measurements are denoted as d_{r}, d_{e} and d_{c} respectively. In a real-world application, the damage condition data $\boldsymbol{Y}_{\mathrm{c}}$ would typically be obtained from inspections and would be available during training, but not during test time. In contrast, measurements of the structural response and environmental conditions are typically obtained from sensor networks and are therefore available both during training and testing.

2.2 Objective

A straightforward approach to infer the posterior distribution of the latent variables conditioned on the data and generate samples of the learned structural response would be to incorporate the physics-based model in the likelihood function, assign a prior distribution over $\boldsymbol{Z}$, and train a VAE using measurements of the structural response. Although feasible, this approach has a number of disadvantages: (i) the flexibility of the data-driven components can result in them dominating the predicted response; (ii) the physics-based latent variables $\boldsymbol{Z}_p$ lose their physical interpretation and the inferred posterior may be inaccurate; (iii) the components of the model prediction are not interpretable; (iv) poor performance in generalization to unseen conditions, particularly when training data is sparse. To address these issues we seek to learn interpretable and disentangled representations of the various factors of variation in the data, separating the influences for which domain knowledge (in the form of physics-based models and informative prior distributions) is available from those that can't be modelled directly and must be learned from the data. This is achieved by imposing constraints on the data-driven components of the VAE, such that they only approximate components of the discrepancy between the measurements and physics-based model predictions that are independent of the known physics and can be attributed to additional observations of the environmental and damage conditions of the structure.

2.3 Previous Work

A common approach for disentangled representation learning involves adjusting the relative importance of the KL-divergence and reconstruction error terms in the ELBO (Eq. 2). In the β-VAE architecture [12], the KL-divergence is scaled by a factor $\beta \geq 1$ that increases how much the approximate posterior is penalized for differing from the prior, effectively limiting the capacity of the latent distribution and forcing independence between the latent variables at the expense of reconstruction quality. An alternative decomposition of the KL-divergence term highlights the importance of the *total correlation* (TC) [13], a component of the KL-divergence that penalizes dependence between the latent variables marginalized over the data and can be estimated stochastically [14] or approximated with a discriminator network [15, 3]. Alternative approaches to promote disentanglement in VAEs involve extending the standard architecture with components such as discriminators [16] and additional decoders [17, 18], utilizing adversarial excitation and inhibition mechanisms [19], and separating components of the latent space [20]. Modifications to the standard VAE architecture are often coupled with modifications to the optimization objective, in order to adjust the mutual information between the input and the latent variables leveraging the information bottleneck theory [21, 22]. For a review on representation learning, the reader is referred to [23]. Limited guidance is available on incorporating physics-based models in VAE, and particularly on the impact of combining DL components and physics-based models on the interpretability and generalizability of the predictions. The issue of balancing physics-based and DL components is highlighted is highlighted in [6], and addressed using a regularized training objective based on the posterior predictive check [7].

3. Approach

The proposed VAE architecture is schematically illustrated in Fig. 1. Two feature extractors implemented as feed-forward NNs are utilized to reduce the dimensionality of the response measurements, which are subsequently provided as input to two sets of NNs that output the mean μ, standard deviation σ and a lower triangular matrix $\boldsymbol{L}$ of multivariate Normal distributions implementing the physics-based and data-driven latent spaces. The standard deviation and lower triangular factor are the **LDL** decomposition terms of the covariance matrix $\boldsymbol{\Sigma}$ of a multivariate Normal distribution, i.e. $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$. The decoder is composed of a physics-based model combined with multiple feed-forward NNs. The mean of the predicted response is obtained as a sum of the outputs of the physics-based model $f^{\mathrm{r}}_{\theta_{\mathrm{p}}}$ and a machine learning component $NN^{\mathrm{r}}_{\theta_{\mathrm{d}}}$. The environmental and damage condition predictions are obtained from $NN^{\mathrm{e}}_{\theta_{\mathrm{d}}}$ and $NN^{\mathrm{c}}_{\theta_{\mathrm{d}}}$ respectively. Crucially, a gradient reversal layer [24] is utilized to prevent the data-driven part of the encoder from learning features related to the reconstruction of the input data and the discrepancies between measurements and physics-based model predictions, instead forcing it to retain only a low-dimensional representation of the features that are informative about the environmental and damage condition observations. The motivation for this formulation of the objective is the assumption that the environmental conditions and damage modes may be identified by features of the response that can be extracted from measurement data and encoded into a low-dimensional representation in the latent space.

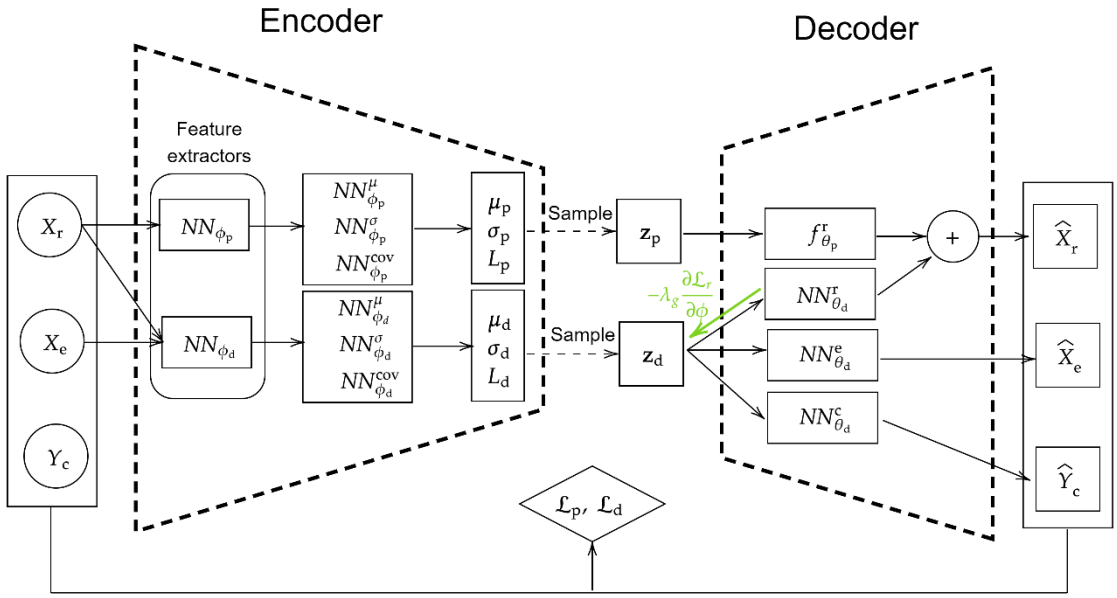

Fig. 1. Schematic overview of the proposed VAE architecture. Symbols with hat denote VAE predictions.

We aim to derive an objective that encourages the data-driven subset of the latent variables $\boldsymbol{Z}_\mathrm{d}$ to capture a low dimensional representation of the features in the response measurements $\boldsymbol{X}_\mathrm{r}$ that are informative about the environmental conditions or the existence of damage in the structure. This is achieved by including the mutual information between the physics-based and data-driven latent variables as an additional regularization term. We define $\bar{q}(\boldsymbol{z}) \coloneqq q(\boldsymbol{z}_\mathrm{p})q(\boldsymbol{z}_\mathrm{d})$ and write the optimization objective as the sum of the VAE losses, and the mutual information between $q(\boldsymbol{z})$ and $\bar{q}(\boldsymbol{z})$:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_\mathrm{p}\left(\boldsymbol{x}_\mathrm{r}; \boldsymbol{\theta}_\mathrm{p}, \boldsymbol{\phi}_\mathrm{p}\right) + \mathcal{L}_\mathrm{d}(\boldsymbol{x}_\mathrm{r}, \boldsymbol{x}_\mathrm{e}, \boldsymbol{y}_\mathrm{c}; \boldsymbol{\theta}_d, \boldsymbol{\phi}_d) + \lambda \cdot \mathrm{KL}(q(\boldsymbol{z})||\bar{q}(\boldsymbol{z})) \qquad (5)$$

The physics-based and data-driven losses $\mathcal{L}_\mathrm{p}$ and $\mathcal{L}_\mathrm{d}$ are obtained by deriving the VAE (Eq. 2) and CVAE (Eq. 3) objectives for the physics-based and data-driven components of the model respectively. The term $\mathrm{KL}(q(\boldsymbol{z})||\bar{q}(\boldsymbol{z}))$ penalizes the mutual information between the approximate posterior distributions of the physics-based and data-driven latent variables marginalized over the data, and is computed using an importance sampling approach [14]. The hyperparameter $\lambda \geq 0$ balances the quality of the reconstruction and the complexity of the learned encoding, with larger values resulting in a more severe penalty at the cost of reduced reconstruction accuracy. The result of optimizing the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ of the VAE using the objective given in Eq. 5 is that the features of $\boldsymbol{X}_\mathrm{r}$ that are informative about $\boldsymbol{X}_\mathrm{e}$ and $\boldsymbol{Y}_\mathrm{c}$ are encoded in the distribution of the latent variables from which the decoder reconstructs the corresponding outputs $\boldsymbol{X}_\mathrm{e}$ and $\boldsymbol{Y}_\mathrm{c}$.

4. Case Study

It is assumed that the true behaviour (i.e. the ground-truth representation) of a structure can be accurately modelled as a two-dimensional Euler-Bernoulli beam with constant length $L = 1.0$ m, a point load $F = 5.0$ N, pinned left support, and rotational and vertical translation spring boundary conditions in the right support. It is assumed that we only possess partial knowledge (i.e. a reduced model) of the true physics. The ground-truth and reduced models are illustrated in Fig. 2 (left). The Young's modulus E and the position of the point load x_F are considered as uncertain latent variables of the physics-based models, such that $\boldsymbol{z}_p =$

$\{E, x_{\mathrm{F}}\}$. Additionally, the rotational stiffness of the right support is considered fully dependent on the temperature T through the non-linear function $k_{\mathrm{r}} = \exp\left(\frac{10}{1+e^{-T}} + 8.0\right)$, shown in Fig. 2 (right). It is assumed that the temperature is an observed environmental parameter such that $X_{\mathrm{e}} = \{T\}$, and that the vertical spring log-stiffness $\log k_{\mathrm{v}}$ represents damage in the structure, i.e. $Y_{\mathrm{c}} = \{\log k_{\mathrm{v}}\}$. Therefore, measurements of $\log k_{\mathrm{v}}$ will be available during training but not during test-time. The structure is assumed to be equipped with d_{r} displacement sensors, equally spaced along the length of the beam.

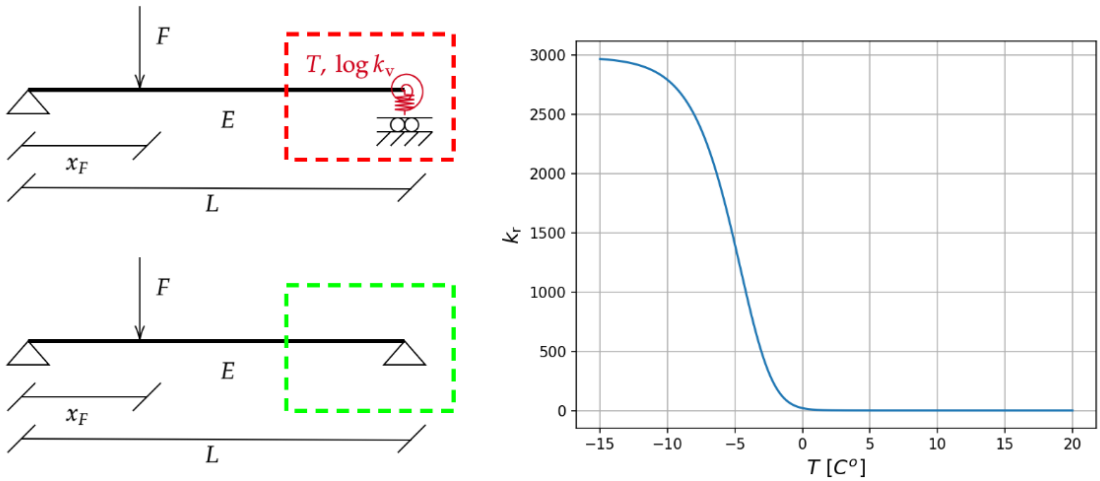

Fig. 2. Two-dimensional Euler-Bernoulli beam with point force. Ground truth model (top left) and reduced model (bottom left) representing our partial knowledge of the physics. The relationship between temperature and k_{r} is shown on the right.

The prior distributions over E and x_{F}, as well as the ground truth distributions of all the latent variables used to generate the training data are summarized in Table 1. The temperature and vertical spring log-stiffness are only included in the (unknown to us) true physical model and not in the partial model, and therefore no physically meaningful prior distribution can be specified. The dataset used to train the VAE is composed of $N = 1024$ measurements of the beam displacement, obtained by drawing N samples from the ground truth distribution of the latent variables and evaluating the true physical model on the sampled points, yielding $\boldsymbol{X}_{\mathrm{r}} = (\boldsymbol{x}_{\mathrm{r},1}, \boldsymbol{x}_{\mathrm{r},2} \ldots \boldsymbol{x}_{\mathrm{r,N}})$ where each element $\boldsymbol{x}_{\mathrm{r,i}}$ is a vector of length $d_{\mathrm{r}} = 32$. The measurements are subsequently contaminated with Gaussian white noise with standard deviation of 0.05 m. The environmental and damage condition measurements $\boldsymbol{X}_{\mathrm{e}}$ and $\boldsymbol{Y}_{\mathrm{c}}$ are taken as the sampled temperature and vertical log-stiffness values respectively.

Table 1. Prior distributions of the latent variables and ground truth distributions used to generate training data.

LATENT VARIABLE	PRIOR DISTRIBUTION	GROUND TRUTH
E [Pa]	$\mathcal{N}(4.0, 0.5)$	$\mathcal{N}(4.0, 0.35)$
x_{F} [m]	$\mathcal{N}(0.5, 0.1)$	$\mathcal{N}(0.5, 0.07)$
T [C^o]	-	$\mathcal{N}(8.0, 5.6)$
$\log k_{\mathrm{v}}$ [N/m]	-	$\mathcal{N}(-4.0, 2.1)$

For the physics-grounded branch, the feature extractor and encoder are formulated as feed-forward NNs while the decoder is the partial physics-based model. For the data-driven branch, the feature extractor, encoder and decoders are all formulated as feed-forward NNs. Details of all components are provided in Table 2.

Table 2. Summary of the formulation of the VAE components. Numbers in brackets denote the width of the hidden layers in the NNs.

BRANCH	FEATURE EXTRATOR	ENCODER	DECODER
Physics-based	$NN^{f}_{\phi_\mathrm{p}}$ [256, 64]	$NN^{\mu}_{\phi_\mathrm{p}}$ [128] $NN^{\sigma}_{\phi_\mathrm{p}}$ [128] $NN^{cov}_{\phi_\mathrm{p}}$ [128]	$f^{r}_{\theta_\mathrm{p}}$ [-]
Data-driven	$NN^{f}_{\phi_\mathrm{d}}$ [256, 64]	$NN^{\mu}_{\phi_\mathrm{d}}$ [128] $NN^{\sigma}_{\phi_\mathrm{d}}$ [128] $NN^{cov}_{\phi_\mathrm{d}}$ [128]	$NN^{r}_{\theta_\mathrm{d}}$ [256, 64] $NN^{e}_{\theta_\mathrm{d}}$ [128] $NN^{c}_{\theta_\mathrm{d}}$ [128]

Optimization is performed using the Adam algorithm [25] with minibatch gradient estimation [2]. The model is trained for 25.000 iterations, and the objective and gradients are estimated using 32 Monte Carlo samples and a batch size of 64. A value of $\lambda = 2$ is used for training as it was found to result in a good compromise between disentanglement and reconstruction accuracy. The coefficient of the gradient reversal layer is taken as $\lambda_\mathrm{g} = 0$, preventing the backwards flow of gradient information from the data-driven decoder. The decoders are formulated as Normal distributions with mean equal to the output of the corresponding NN. For the displacement decoder, the standard deviation is included in the vector $\boldsymbol{\theta}$ and jointly optimized with the other hyperparameters, while for the environment and damage decoders it is set to $\sigma_\mathrm{e} = \sigma_\mathrm{d} = 0.05\ C^o$ and N/m respectively. Both feature extractor NNs are set to output eight features, and the number of dimensions of the data-driven latent space is set to two. The VAE is intentionally overparametrized by specifying a larger number of features in the output of the feature extractors than necessary, in order to demonstrate that prior knowledge of the number of failure mechanisms, environmental influences and physical processes affecting the response of a structure is not necessary.

After training, the disentanglement between physics-grounded and data-driven components is qualitatively assessed by examining the latent space and predictions. Each variable is linearly interpolated within the interval $[\mu_\mathrm{gt} - 3\sigma_\mathrm{gt}, \mu_\mathrm{gt} + 3\sigma_\mathrm{gt}]$ (where μ_gt and σ_gt are the ground truth mean and standard deviation corresponding to the variable as shown in Table 1), while the other variables are held constant at their ground truth mean value. The resulting sets of input variables are then used to generate synthetic displacement and temperature data, which are in turn given as input to the VAE. The resulting predictions of the physics-based and data-driven components, as well as the combined prediction are shown in Fig. 3. It can be observed that the data-driven component of the prediction is invariant to changes in E and x_F contributing only a constant deformed shape to the total predicted response, and that additionally, the physics-based component of the prediction is relatively insensitive to varying $\log k_\mathrm{v}$ and T values. However, it can also be seen that the predictions obtained from the two branches of the VAE are neither fully disentangled, nor do they perfectly capture the true structural response. The trade-off between these two objectives can be adjusted using the λ parameter (Eq. 5).

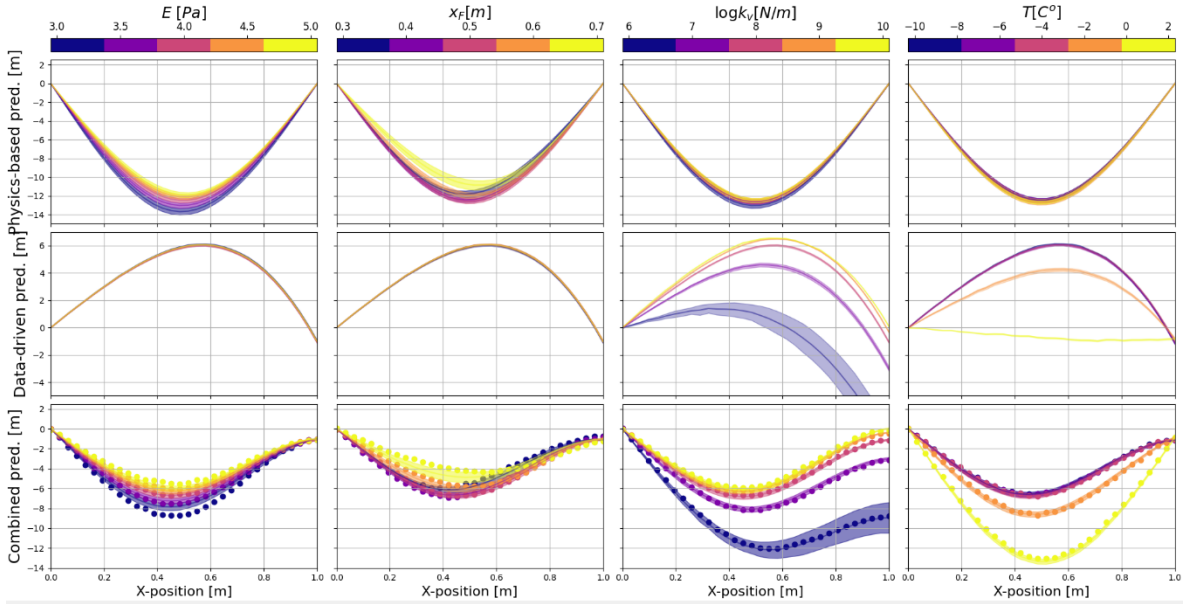

Fig. 3. Comparison of VAE predictions for a single input measurement per value of the latent variables. Mean and std. dev. of the physics-based (top), data-driven (middle), and combined prediction (bottom). Dots denote the input measurements.

Further insights on the impact of the additional regularization term and gradient reversal layer on the performance of the VAE and the disentanglement of the physics and data terms can be obtained by examining the behaviour of the latent space with respect to the varying input data, shown in Fig. 4. It can be seen that, although correlation between E and x_{F} is present as evidenced by the curved shape of the "paths" of samples for varying input, as well as the correlation that is present in the posterior of each individual input, only limited correlation is observed between the two subsets of laten variables $\{E, x_{\mathrm{F}}\}$ and $\{z_{\mathrm{d},1}, z_{\mathrm{d},2}\}$. The limited correlation between the two sets of latent variables indicates that the modified TC objective, in addition to the structure of the VAE itself, promotes independence between the physics-grounded and data-driven components of the latent space.

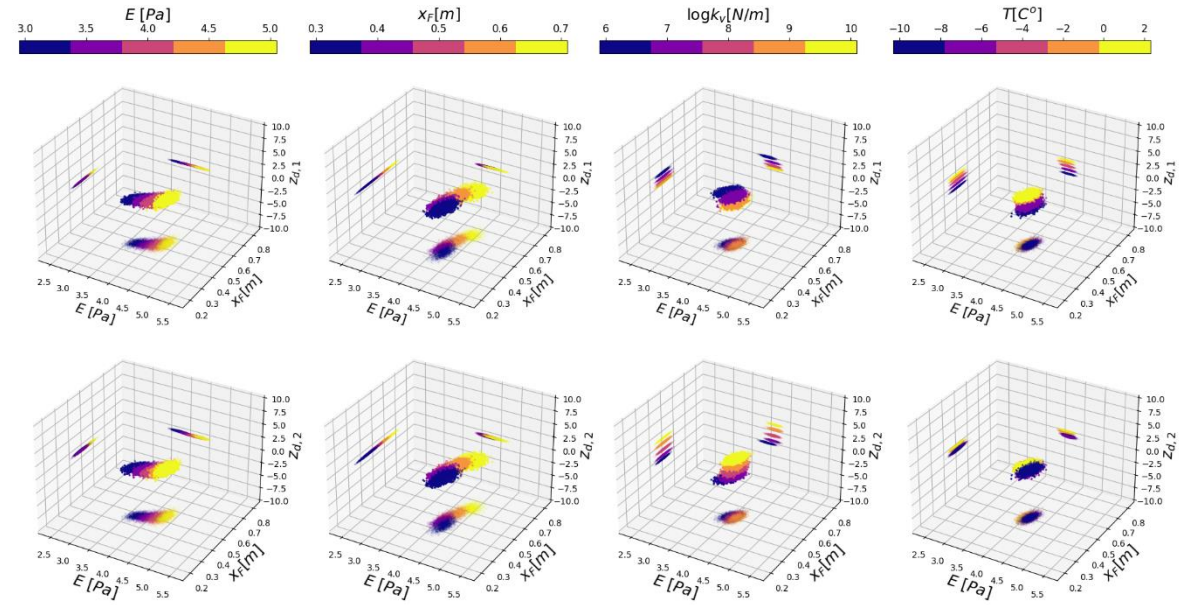

Fig. 4. Samples from the latent space for a single input measurement per value of the latent variables. The vertical axis are the first (top row) and second (bottom row) data-driven latent variables $z_{d,1}, z_{d,2}$.

It is noted that the objective of this case study is not to provide a comprehensive assessment of the performance of the proposed architecture compared to a purely data-driven approach,

but rather to highlight the benefits of disentangling the physics-based and data-driven components of the model and demonstrate the feasibility of the proposed architecture. Therefore, the results shown in this section are only qualitative. A thorough, quantitative comparison with other data-driven and physics-informed approaches is left for future work.

5. Conclusions and Future Work

A VAE architecture with two separate encoder and decoder branches, one physics-based and one data-driven, is proposed as a means to tackle SHM problems and create DT models for which domain knowledge can be expressed in terms of a physics-based model. The training objective is designed to promote disentangled learning, separating the physics-based components of the latent space (the uncertain parameters of the physics-based model for which a prior distribution is available) from the data-driven components of the latent space, as well as the physics-based predictions from the data-driven predictions, by incorporating an additional KL-divergence term scaled by a hyperparameter λ. Additionally, a gradient reversal layer is included before the data-driven decoder of the structural response in order to prevent the data-driven latent space from encoding features of the structural response that are not relevant to the reconstruction of the additional environmental condition measurements and damage condition labels. The approach is demonstrated using a synthetic case study, with the results indicating that the proposed architecture enables disentangled learning of the structural response components that correspond to the partially known physics from those stemming from environmental influences and the existence of damage in the structure. Follow-up work will be focused on investigating the impact of the λ and λ_g parameters on the accuracy of the inferred posterior and the quality of the disentanglement, and more generally on the accuracy of the approximate posterior obtained from the proposed model formulation. Determining the applicability of existing disentanglement metrics for evaluating the disentanglement of known physics from environmental and damage influences, or deriving a specialized metric for this task, are also interesting avenues for future work.

Acknowledgements

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).

References

[1] D. J. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows,* 2016.

[2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114,* 2013.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *Generative Adversarial Networks,* 2014.

[4] G. Tsialiamanis, D. J. Wagg, N. Dervilis and K. Worden, "On generative models as the basis for digital twins," *Data-Centric Engineering,* vol. 2, p. e11, 2021.

[5] E. J. Cross, S. J. Gibson, M. R. Jones, D. J. Pitchforth, S. Zhang and T. J. Rogers, "Physics-Informed Machine Learning for Structural Health Monitoring," in *Structural Health Monitoring Based on Data Science Techniques*, Springer International Publishing, 2022, pp. 347-367.

[6] N. Takeishi and A. Kalousis, "Physics-Integrated Variational Autoencoders for Robust and Interpretable Generative Modeling," in *Advances in Neural Information Processing Systems 34*, 2021.

[7] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari and D. Rubin, Bayesian Data Analysis (3rd ed.), Chapman and Hall/CRC, 2013.

[8] D. M. Blei, A. Kucukelbir and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association,* vol. 112, no. 518, pp. 859-877, 2017.

[9] K. Sohn, H. Lee and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in *Advances in Neural Information Processing Systems 28*, 2015.

[10] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 63, no. 3, pp. 425-464, 2001.

[11] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?," *Structural Safety,* vol. 31, no. 2, pp. 105-112, 2009.

[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed and A. Lerchner, *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,* 2017.

[13] S. Watanabe, "Information Theoretical Analysis of Multivariate Correlation," *IBM Journal of Research and Development,* vol. 4, no. 1, pp. 66-82, 1960.

[14] R. T. Q. Chen, X. Li, R. Grosse and D. Duvenaud, "Isolating Sources of Disentanglement in Variational Autoencoders," in *Advances in Neural Information Processing Systems 31*, 2018.

[15] H. Kim and A. Mnih, "Disentangling by Factorising," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[16] A. B. L. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of Machine Learning Research*, 2016.

[17] H. Sun, N. Pears and Y. Gu, "Information Bottlenecked Variational Autoencoder for Disentangled 3D Facial Expression Modelling," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[18] X. Hou, L. Shen, K. Sun and G. Qiu, "Deep Feature Consistent Variational Autoencoder," *arXiv,* vol. 1610.00291, 2024.

[19] Z. a. X. Y. Ding, W. Xu, G. Parmar, Y. Yang, M. Welling and Z. Tu, "Guided Variational Autoencoder for Disentanglement Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] A. Mondal, A. Sailopal, P. Singla and P. AP, " SSDMM-VAE: variational multi-modal disentangled representation learning," *Applied intelligence,* vol. 53, pp. 8467-8481, 2022.

[21] N. Tishby, F. C. Pereira and W. Bialek, "The information bottleneck method," *arXiv:physics/0004057,* 2000.

[22] A. A. Alemi, I. Fischer and J. V. Dillon, "Deep Variational Information Bottleneck," *ArXiv,* vol. abs/1612.00410, 2017.

[23] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798-1828, 2013.

[24] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[25] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization,* arXiv:1412.6980, 2017.

[26] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,* 2016.

[27] R. Ranganath, S. Gerrish and D. Blei, "Black box variational inference," *Artificial intelligence and statistics,* pp. 814-822, 2014.

[28] W. Zhong and H. Meidani, "PI-VAE: Physics-Informed Variational Auto-Encoder for stochastic differential equations," *Computer Methods in Applied Mechanics and Engineering,* vol. 403, p. 115664, 2023.

[29] X. L. a. H. Bolandi, M. Masmoudi, T. Salem, N. Lajnef and V. N. Boddeti, "Mechanics-Informed Autoencoder Enables Automated Detection and Localization of Unforeseen Structural Damage," *arXiv:2402.15492,* 2024.