PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Review

Cite this article: Glyn-Davies A, Vadeboncoeur A, Akyildiz OD, Kazlauskaite I, Girolami M. 2025 A primer on variational inference for physics-informed deep generative modelling. *Phil. Trans. R. Soc. A* **383**: 20240324.
https://doi.org/10.1098/rsta.2024.0324

Received: 10 September 2024
Accepted: 24 March 2025

One contribution of 13 to a theme issue 'Generative modelling meets Bayesian inference: a new paradigm for inverse problems'.

Subject Areas:
computational mathematics, statistics

Keywords:
deep learning, physics-informed, variational inference, generative model, PDE

Author for correspondence:
Arnaud Vadeboncoeur
e-mail: av537@cam.ac.uk

THE ROYAL SOCIETY PUBLISHING

A primer on variational inference for physics-informed deep generative modelling

Alex Glyn-Davies[1], Arnaud Vadeboncoeur[1], O. Deniz Akyildiz[2], Ieva Kazlauskaite[3] and Mark Girolami[1]

[1]Department of Engineering, University of Cambridge, Cambridge, Cambridgeshire, UK
[2]Department of Mathematics, Imperial College London, London, UK
[3]Department of Statistics, The London School of Economics and Political Science, London, UK

AV, 0000-0003-4124-6763

Variational inference (VI) is a computationally efficient and scalable methodology for approximate Bayesian inference. It strikes a balance between accuracy of uncertainty quantification and practical tractability. It excels at generative modelling and inversion tasks due to its built-in Bayesian regularization and flexibility, essential qualities for physics-related problems. For such problems, the underlying physical model determines the dependence between variables of interest, which in turn will require a tailored derivation for the central VI learning objective. Furthermore, in many physical inference applications, this structure has rich meaning and is essential for accurately capturing the dynamics of interest. In this paper, we provide an accessible and thorough technical introduction to VI for forward and inverse problems, guiding the reader through standard derivations of the VI framework and how it can best be realized through deep learning. We then review and unify recent literature exemplifying the flexibility allowed by VI. This paper is designed for a general scientific audience looking to solve physics-based problems with an emphasis on uncertainty quantification.

This article is part of the theme issue 'Generative modelling meets Bayesian inference: a new paradigm for inverse problems'.

1. Introduction

This paper serves as a tutorial and review of methodologies for inference related to physical problems using variational inference (VI). We introduce basic concepts and the mathematical formulations pertaining to the most relevant and important tools in the field. We first consider the modelling of physical systems with partial differential equations (PDEs). We then present an overview of inverse problems through optimization and Bayesian perspectives and provide a detailed derivation of VI. Equipped with this knowledge, we then review salient methods in the literature for solving physical inference problems with forward model and weighted residual method (WRM)-based VI.

Forward problems in physical modelling refer to the computation, simulation or estimation of the solution to a mathematical physics problem. These can come in a variety of forms such as agent-based models [1], data-driven models [2], differential equations [3] and any number of combinations thereof. In this work, we focus on models which describe mechanistic understanding through differential equations. Broadly speaking, these models describe the change in certain quantities of interest, such as heat, velocity and electric potential, with respect to time or space. As such, these models are intrinsically linked to the setting in which they are considered, that is to say, initial conditions, boundary conditions, geometry and other physical quantities. If multiple forward problems must be solved for different sets of parameters, classical numerical solvers can be computationally intractable. These multi-query problems often arise in contexts of uncertainty quantification (UQ) through methods such as Monte Carlo sampling, Taylor expansion and perturbation methods. Surrogate models may alleviate this computational burden [4]. A classical example of surrogate models for forward problems is Gaussian processes (GPs), which have inherent UQ capabilities [5]. Many learning models have been recently developed for surrogate modelling of PDEs with functional inputs such as deep operator networks (DeepONet) and Fourier Neural Operators (FNO) [6,7]; however, these models do not have built-in UQ capabilities like [8,9].

Inverse problems, on the other hand, aim to recover model parameters that gave rise to a set of observations, i.e. inverting the forward problem. Classic application fields include computed tomography [10], cosmology [11] and geophysics [12]. When observations are noisy or sparse, the inverse problem is typically *ill-posed*, meaning that many different model parameter values could have provided the same observations. Then, inverse problems require a form of *regularization* on the model parameters to provide unique solutions [13]. Point-estimate-based inversion generally does not seek UQ [14], while Bayesian methods recover distributions over parameters [15].

VI is a statistical framework that strikes a practical balance between computational costs and accuracy of UQ [16,17]. It relies on the optimization of a statistical objective to provide uncertainty estimates in inference tasks [18]. There is a large variety of VI schemes with different advantages and limitations [19]. One of the most discernible advantages of constructing VI-based inference schemes is to allow one to circumvent expensive Markov Chain Monte Carlo (MCMC) sampling of intractable probability distributions, which often arise in the statistical treatment of uncertainty relating to nonlinear models. As these nonlinear models are essential for capturing the physical structure of many scientific problems, VI methods have great potential in making UQ for sciences computationally feasible. Furthermore, VI allows practitioners to construct computationally efficient frameworks with built-in conditional dependence structures reflecting the nature of the inferential task at hand [20–22]. This conditional dependence structure will often be represented as a Bayesian graphical model [16,23]. The ability to strictly enforce intricate dependencies between quantities of interest—such as in physics problems—is precisely what gives rise to the wide variety of methods explored in this paper.

We structure the rest of the paper as follows: §2 introduces the relevant mathematical background; forward problems are described in §2(a); optimization and Bayesian inference for inverse problems are covered in §2(b); VI methods are presented in §2(c). Section 3 reviews applications of

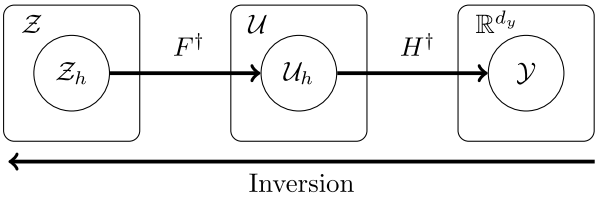

Figure 1. A depiction of the three spaces of inferential interest: the observation space $\mathcal{Y}$, the discretized solution space $\mathcal{U}_h$ and the discretized parameter space $\mathcal{Z}_h$. More specifically, we have an observation $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, a solution $u_h \in \mathcal{U}_h \subset \mathcal{U}$ and a parameter $z_h \in \mathcal{Z}_h \subset \mathcal{Z}$.

these methods to physics-based generative modelling tasks found in the literature. Applications are split into forward-model-based approaches in §3(a) and residual-based learning in §3(b).

2. Physics and inference

In this section, we introduce and elaborate on the core concepts and tools required to build variational inference schemes for the physical sciences. In figure 1, we show a depiction of the mathematical spaces that describe the three main quantities of inferential interest: parameter, solution and observation, which we denote as $z \in \mathcal{Z}, u \in \mathcal{U}, y \in \mathcal{Y}$, respectively. In the following sections, we denote the finite-dimensional representations of the parameter and solution as $z_h \in \mathcal{Z}_h, u_h \in \mathcal{U}_h$, respectively, where the subscript h is a parameter describing the degree of discretization. This means we only consider spaces of solutions and parameters that are finite-dimensional; hence, they have already been discretized. Rigorous mathematical treatment of inference schemes over functions, which are infinite-dimensional, is of great value but beyond the scope of this paper [15,24].

(a) Forward problems

We describe a generic forward model through a numerical scheme that relates the discretized physical setup, $z_h \in \mathcal{Z}_h$ to the realization of the physical process across time and space, which we call the solution and is denoted as $u_h \in \mathcal{U}_h$. The forward model is a mapping from a particular setup to the solution associated with that setup described as $F^\dagger : \mathcal{Z}_h \to \mathcal{U}_h$. The use of "†" refers to the near-exact numerical realization of the differential equations of interest, and we will see later how this might be approximated by a parametrized—less expensive to evaluate—surrogate model.

To discuss PDEs in more detail, we choose a canonical example, the Poisson problem. It describes a variety of steady-state diffusive physical systems, such as heat, electric potential and groundwater flow. A function is said to be a solution to this problem if it satisfies, for some physical domain Ω,

$$\nabla \cdot (z(x)\nabla u(x)) = f(x), \quad \text{for } x \in \Omega, \tag{2.1a}$$

$$u(x) = 0, \quad \text{for } x \in \partial\Omega, \tag{2.1b}$$

where $\partial\Omega$ denotes the boundary of Ω. The problem stated in this form is not amenable to numerical computation as u is currently an infinite-dimensional object and it must be discretized. How we represent this function $u \in \mathcal{U}$ and how it relates to (2.1) is given by the particular numerical scheme in use.

We look at the discretization of solution fields and PDE operators through the lens of the weighted residual method (WRM) [3], which encompasses most spatial discretization schemes such as finite element (FE), spectral methods, finite difference and physics-informed neural networks (PINNs). The advantage of taking this perspective on numerical discretization for machine learning (ML) is that inference schemes can be constructed independently of the *particular* WRM

method in use; hence, these can be swapped out with ease. To write out the WRMs, we first specify the residual function

$$R(u, z, f, x) = \nabla \cdot (z(x)\nabla u(x)) - f(x). \tag{2.2}$$

Choosing a set of weight functions $\{v_i\}_{i=1}^{d_r}$ with $v_i \in \mathcal{V}$ we can test the residual

$$r_i = \int_\Omega v_i(x) R(u, z, f, x)\, \mathrm{d}x = \int_\Omega v_i(x) \left(\nabla \cdot (z(x)\nabla u(x)) - f(x)\right)\, \mathrm{d}x. \tag{2.3}$$

Collecting $\mathbf{r} = \{r_i\}_{i=1}^{d_r}$ discretizes the action of the differential operator on the solution u. One can then use integration by parts on (2.3) if the test functions are differentiable to obtain the *weak form* of the Poisson equation,

$$r_i = \int_{\partial\Omega} v_i(x)(z(x)\nabla u(x)) \cdot \hat{n}(x)\, \mathrm{d}x - \int_\Omega \nabla v_i(x) \cdot (z(x)\nabla u(x))\, \mathrm{d}x - \int_\Omega f(x)\, \mathrm{d}x. \tag{2.4}$$

Various other Galerkin-type methods can be designed by varying the choice of test and trial functions. By choosing $v_i = \phi_i$ (implying $v \in \mathcal{V}^h = \mathcal{U}^h$ and $\mathcal{V}_h = \mathrm{span}\{\phi_i\}_{i=1}^{N_u}$ and for this problem choosing ϕ_i to be hat functions), we obtain a Bubnov–Galerkin method [25]. Working with such weak forms has notable advantages, mainly it reduces the differentiability requirements on the trial function as a derivative order is passed over to the test function. Linear approximants can be represented with the following basis function expansion $u_h(x) = \sum_{i=1}^{N_u} [\mathbf{u}]_i\, \phi_i(x)$, where $u_h \in \mathcal{U}_h$, $\mathbf{u} \in \mathbb{U}$ are the coefficients, and ϕ_i are the basis functions. When constructing inference schemes, we can now use $\mathbf{u}$ in lieu of u_h. Similarly, we can replace $z \in \mathcal{Z}$—which in this particular example is a function—with a finite-dimensional discretization $z_h \in \mathcal{Z}_h$ which in turn can be expressed with an expansion as $z_h(x) = \sum_{i=1}^{N_z} [\mathbf{z}]_i\, \psi_i(x)$ and summarized as $\mathbf{z} \in \mathbb{Z}$. We denote the chosen mapping from coefficients $\mathbf{z}, \mathbf{u}$ to interpolants z_h, u_h as $\pi_z(\mathbf{z}) = z_h$, $\pi_u(\mathbf{u}) = u_h$, respectively. Residuals like these can be efficiently computed in a GPU-efficient manner using array-shifting [26] or convolutions [27]. We note that a variety of variational formulations such as the Ritz method or energy functionals are amenable to equivalent residual formulations as in (2.3) [28].

PINNs are neural network-based methods for approximating the solution to differential equations [29]. Many of these methods can be obtained by taking u_h to be a nonlinear approximant as a neural network. A typical form is $u_h(x) = T_L \circ \ldots \circ T_0(x)$ where $T_i(x) = \boldsymbol{\sigma}_i(W_i\, x + b_i)$ where $\boldsymbol{\sigma}_i, W_i, b_i$ are the layers' activation function, weight matrix and bias vector, respectively, and choosing $v_i(x) = \delta(x_i - x)$ where δ is the Dirac delta function and x_i are collocation points. For these PDE solvers, the solution representation for inference is $\mathbf{u} = \{u_h(x_i)\}_{i=1}^{N_u}$. It is to be noted that when using this kind of approach, we no longer make use of the weak form. Neural network approximants may still be used with the variational form [30]. For further reading on this topic, we refer readers to [31–33].

The treatment of boundary conditions depends on the specific WRM method in use; FE-based methods typically use boundary-respecting meshes and the weak form naturally includes other boundary conditions; PINN-style methods can either include an additive boundary loss term to the residual or enforce certain types of boundary conditions through certain manipulations of u_h [34]. To numerically solve the PDE means to find u_h such that the residual vector $\mathbf{r} \approx 0$, within a pre-defined tolerance. In the case of the FE method for linear PDEs, a system of sparse linear equations can be set up, which can be directly solved using linear solvers, but the residual formulation may still be implemented as is often done in the case of PINNs.

(b) Inverse problems

Inversion methods map elements of $\mathcal{Y}$ to points or distributions in $\mathcal{U}$ or $\mathcal{Z}$[1]. That is, we either wish to recover the full solution from observations, or the parameters from observations. We find it appropriate to separate the full mapping between parameter-to-observation, denoted $G^\dagger$, into

the mapping from parameter-to-solution, $F^\dagger$ (forward model), and the mapping from solution-to-observation, $H^\dagger : \mathcal{U}_h \to \mathcal{Y}$ (observation model). Here, the "†" denotes the 'true mapping' to distinguish from settings where we might try and learn this map. The full parameter-to-observation map can be written as $G^\dagger(z_h) = (H^\dagger \circ F^\dagger)(z_h)$, the composition of the forward and observation maps.

(i) Point estimate inversion

If one is not interested in recovering uncertainty over model parameters given some data, point estimate inversion may be used. Inversion schemes rely on the combination of a data-fit term and a regularization term. As most inverse problems of interest are ill-posed, the quality of the estimated quantities from applying inversion schemes is tied to the quality of the regularization imposed. A classic approach to the regularization of inverse problems is the Tikhonov approach [13–15]

$$\mathbf{z}^\star = \arg\min_{\mathbf{z}\in\mathcal{Z}} \quad \frac{1}{2}\|\mathbf{y} - (H^\dagger \circ F^\dagger \circ \pi_z)(\mathbf{z})\|^2 + \frac{\beta}{2}\|\pi_z(\mathbf{z})\|^2, \tag{2.5}$$

where $F^\dagger$ is the forward model and β controls the strength of the bias towards z_h estimates that are small in the chosen norm. We note other forms of regularization are possible, such as total variation [35], sparsity promoting ℓ_1 regularization [36] and regularizing operators [12]. Alternative perspectives on inverse problems for physical systems use the regularization term to impose physical knowledge. These methods estimate the parameter of interest as

$$\mathbf{z}^\star = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \min_{\mathbf{u}\in\mathcal{U}} \quad \|\mathbf{y} - (H^\dagger \circ \pi_u)(\mathbf{u})\|^2 + \beta\|\mathbf{r}(\pi_u(\mathbf{u}); \pi_z(\mathbf{z}))\|^2, \tag{2.6}$$

where β now controls the trade-off between the data-fit and the physics regularization. In practice, the parameter β is often manually tuned. Taking u_h as the output of a PINNs and the WRM used for computing $\mathbf{r} \in \mathbb{R}^{d_r}$ to be a collocation method where the test functions are Diracs recovers a PINN-style parameter inversion method. We note one can choose u_h to be an FE expansion with a weak form result computation. An interesting development of these methods is to formulate the combined objectives in terms of a bilevel optimization problem [37], which eliminates the need to balance the physics residual with the data-fit term.

(ii) Bayesian inverse problems

Recovering a point estimate of the solution may be insufficient for many applications. Bayesian inverse problems (BIPs) provide an alternative approach through the probabilistic framework of Bayes' theorem that offers a unifying framework, UQ and some theoretical insights into the posterior consistency of the recovered solution. Bayes' theorem, given as

$$p(\mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{y})}, \quad \text{where} \quad p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}, \tag{2.7}$$

allows one to derive the full posterior distribution over the model parameters $\mathbf{z}$ given the observed data $\mathbf{y}$. This approach combines the likelihood $p(\mathbf{y}|\mathbf{z})$, derived from the data-generating model, and the prior distribution $p(\mathbf{z})$ as the regularizer, offering a direct parallel to the point-estimate-based approach. The model evidence, $p(\mathbf{y})$, also known as the marginal likelihood, which appears in (2.7), is often intractable. Hence the need for methods that do not require normalized probability densities such as MCMC or Bayesian VI. Note that the point estimate recovered using the optimization approach is typically the maximum a posteriori (MAP) estimate (as in equation (2.5)) where additive zero-mean Gaussian noise on the observations leads to a Gaussian likelihood. For typical physical systems, the mapping from parameter to observation can be expressed as $G = (H^\dagger \circ F^\dagger \circ \pi_z)$. We consider a set of observations that arise as independent and identically distributed (i.i.d.)

$$\mathbf{y} = G(\mathbf{z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \Gamma), \tag{2.8}$$

where Γ is the symmetric positive-definite noise covariance. The observation model (2.8) results in a Gaussian likelihood $p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}; G(\mathbf{z}), \Gamma)$.

(c) Variational inference

At its core, VI poses statistical inference as an optimization problem by minimizing a data-informed *regularized loss* over a *variational family* of distributions. Abstractly, we seek

$$q^\star(\mathbf{z}) \in \arg\min_{q \in \mathcal{Q}(\mathcal{Z})} J(q(\mathbf{z}); \mathbf{y}), \tag{2.9}$$

where $\mathcal{Q}(\mathcal{Z}) \subseteq \mathcal{P}(\mathcal{Z})$ is the variational family—a subset of all possible probability measures on $\mathcal{Z}$. To realize this approach, we typically choose $\mathcal{Q}(\mathcal{Z})$ to have a parametric form with parameters ϕ. The variational approximation $q_\phi(\mathbf{z})$ (with ϕ being the mean and covariance for Gaussian approximations, for example) is then parametrized by ϕ and loss is minimized with respect to ϕ. In some cases, closed forms of the updates on ϕ can be derived, but in many modern applications, one resorts to gradient descent schemes. The choice of loss function $J(\cdot\,; \mathbf{y})$ is crucial and determines the object recovered by the method. We next discuss two pertinent concepts: Bayesian VI and probabilistic generative models.

(i) Bayesian variational inference

Bayesian VI is the optimization formulation of the Bayes' theorem. It performs inference with a principled balance between data-fit and prior knowledge and recovers a probability distribution over model parameters. The loss function for Bayesian VI is based on the Kullback–Leibler (KL) divergence

$$D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})) = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right], \tag{2.10}$$

given absolute continuity between q and p, meaning q assigns zero probability to sets for which p also assigns zero probability. The KL divergence quantifies the difference between two probability distributions. Bayesian VI aims to minimize the KL divergence between the true posterior $p(\mathbf{z}|\mathbf{y})$, and the variational approximation $q_\phi(\mathbf{z})$, parametrized by ϕ. To derive the objective function, we write out the KL divergence, before applying Bayes' theorem and simplifying

$$\begin{aligned} D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z}|\mathbf{y})) &= \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})}\right] = \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log \frac{p(\mathbf{y})q_\phi(\mathbf{z})}{p(\mathbf{y}|\mathbf{z})p(\mathbf{z})}\right], \\ &= \log p(\mathbf{y}) - \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p(\mathbf{y}|\mathbf{z})\right] + \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{z})}{p(\mathbf{z})}\right]. \end{aligned} \tag{2.11}$$

As $p(\mathbf{y})$ does not depend on the variational approximation $q_\phi(\mathbf{z})$ [38], minimizing $D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z}|\mathbf{y}))$ is equivalent to minimizing

$$J(\phi; \mathbf{y}) := \mathbb{E}_{q_\phi(\mathbf{z})}\left[-\log p(\mathbf{y}|\mathbf{z})\right] + D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z})). \tag{2.12}$$

In this form $J(\phi; \mathbf{y})$ avoids the expensive computation of the model evidence $p(\mathbf{y})$ and is directly minimizing the KL divergence between the variational approximation and the Bayesian posterior. Seeking $\phi^\star = \arg\min_\phi J(\phi; \mathbf{y})$, yields a Bayesian VI approximation to the posterior. In practice, the expectations in (2.12) are approximated via Monte Carlo using samples $\mathbf{z}^{(s)} \sim q(\mathbf{z})$, $s = 1, \ldots, S$ [39].

(ii) Probabilistic generative models

Probabilistic generative models are defined by a joint distribution $p_\vartheta(\mathbf{z}, \mathbf{y})$, parametrized by ϑ which are to be estimated from the observed data. In order to learn the generative model,

these parameters are typically estimated via maximization of the Bayesian model evidence, $p_\theta(\mathbf{y}) = \int p_\theta(\mathbf{z}, \mathbf{y})\mathrm{d}\mathbf{z}$, which now depends on θ. Methods in variational inference, such as Variational Autoencoders (VAEs) [40], will often combine estimation of generative model parameters with the variational approximation of the posterior $q_\phi(\mathbf{z})$, where, in general, the exact posterior $p_\theta(\mathbf{z}|\mathbf{y}) = p_\theta(\mathbf{z}, \mathbf{y})/\int p_\theta(\mathbf{z}, \mathbf{y})\mathrm{d}\mathbf{z}$ cannot be evaluated due to the intractable normalization constant arising from the complex generative model structure. In such cases, the joint estimation of parameters $\{\phi, \theta\}$ is required. Taking the prior $p(\mathbf{z})$ as fixed and the likelihood $p_\theta(\mathbf{y}|\mathbf{z})$ as the parametrized model, we can rearrange (2.11) to obtain an expression for the log-marginal likelihood,

$$\log p_\theta(\mathbf{y}) = D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p_\theta(\mathbf{z}|\mathbf{y})) + \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{y}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z})), \tag{2.13}$$

which is intractable due to the evaluation of the posterior in the first right-hand term, but can be bounded from below due to the non-negativity of the KL

$$\log p_\theta(\mathbf{y}) \geq \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{y}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z})) := \mathcal{L}(\phi, \theta; \mathbf{y}). \tag{2.14}$$

Here, $\mathcal{L}$ is known as the evidence lower bound (ELBO), and in practice is maximized via gradient-based stochastic optimization schemes, using Monte Carlo to estimate expectations. For optimization, the objective is defined in terms of both ϕ, θ as the negative ELBO, $J(\phi, \theta; \mathbf{y}) := -\mathcal{L}(\phi, \theta; \mathbf{y})$, where optimal parameters minimize this objective $\phi^\star, \theta^\star = \arg\min_{\phi,\theta} J(\phi, \theta; \mathbf{y})$.

We note that the ELBO is often derived via Jensen's inequality (see e.g. [41]), which applies to concave transformations of expectations, and for the natural log reads $\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$ [20], and is applied for (2.18) below

$$\log p_\theta(\mathbf{y}) = \log\left(\int p_\theta(\mathbf{z}, \mathbf{y})\mathrm{d}\mathbf{z}\right) = \log\left(\int \frac{p_\theta(\mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z})} q_\phi(\mathbf{z})\mathrm{d}\mathbf{z}\right) \tag{2.15}$$

$$= \log\left(\mathbb{E}_{q_\phi(\mathbf{z})}\left[\frac{p_\theta(\mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z})}\right]\right) \geq \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log \frac{p_\theta(\mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z})}\right] \tag{2.16}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{y}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z})) = \mathcal{L}(\mathbf{y}; \phi, \theta). \tag{2.17}$$

It is important to note that since the KL term dropped from (2.13) depends on θ, $\mathcal{L}$ is a *lower bound*, whereas in (2.11) the objective is directly minimizing the posterior KL without approximation (as $\log p(\mathbf{y})$ does not depend on θ).

The ELBO is used for unsupervised learning in VAEs, which are probabilistic generative models defined by an encoder and decoder. The encoder is a conditional distribution $q_\phi(\mathbf{z}|\mathbf{y})$ which, intuitively, *encodes* a data point $\mathbf{y}$ into the latent space $\mathcal{Z}$ by returning a probability distribution over it (rather than a fixed embedding). Similarly, the probabilistic decoder $p_\theta(\mathbf{y}|\mathbf{z})$ is a probability measure for fixed $\mathbf{z}$ and θ, meaning that the decoder returns a distribution over the data $\mathbf{y}$ given the latent vector $\mathbf{z}$. The latent space is typically low-dimensional, forcing the model to learn parsimonious representations of the data, and is regularized by a (often simple) prior distribution, e.g. $p(\mathbf{z}) = \mathcal{N}(0, \mathrm{I})$. Both q_ϕ and p_θ, in general, are parametrized with neural networks. For a dataset $\mathcal{D} = \{\mathbf{y}^{(n)}\}_{n=1}^N$, and assuming i.i.d. observations such that the log likelihood decomposes as $\log p_\theta(\mathbf{y}^{(1:N)}) = \sum_{n=1}^N \log p_\theta(\mathbf{y}^{(n)})$, we can write the log marginal likelihood as

$$\log p_\theta(\mathbf{y}^{(1:N)}) \geq \sum_{n=1}^N \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y}^{(n)})}\left[\log p_\theta(\mathbf{y}^{(n)}|\mathbf{z})\right]}_{\text{reconstruction error}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{y}^{(n)})||p(\mathbf{z}))}_{\text{regularisation}} =: \sum_{n=1}^N \mathcal{L}(\mathbf{y}^{(n)}; \theta, \phi). \tag{2.18}$$

For large datasets, one often uses a *mini-batch*, $B \subseteq \mathcal{D}$, of the dataset per gradient step, giving an approximate minimization objective $J(\theta, \phi; \mathbf{y}^{(1:N)}) := -\frac{N}{|B|}\sum_{n\in B}\mathcal{L}(\mathbf{y}^{(n)}; \theta, \phi)$. As we approximate this lower bound stochastically through Monte Carlo, our objective is a 'doubly-stochastic' approximation to the true ELBO, which is found to improve learning [40]. If we now choose $q_\phi(\mathbf{z}|\mathbf{y}^{(n)}) = \mathcal{N}(\mathbf{z}; m_\phi(\mathbf{y}^{(n)}), C_\phi(\mathbf{y}^{(n)}))$ and $p_\theta(\mathbf{y}^{(n)}|\mathbf{z}) = \mathcal{N}(\mathbf{y}^{(n)}; G_\theta(\mathbf{z}), C_\eta)$ with $m_\phi(\cdot)$, $C_\phi(\cdot)$, $G_\theta(\cdot)$, being

neural networks, we obtain the classic VAE. The choice of prior distribution affects the latent regularization and is typically chosen as a standard Gaussian, $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; 0, \mathrm{I})$.

A practical consideration when training VAEs is the computation of the loss function's gradient with respect to the VI parameters $\nabla_\phi J(\theta, \phi; \mathbf{y}^{(1:N)})$, which requires gradient backpropagation through the Monte Carlo sampled latent variables $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}|\mathbf{y}^{(n)})$. In order to facilitate the gradient backpropagation, practitioners employ the so-called 'reparameterization-trick' [40], which defines the latent random variable as a *differentiable* transformation of the variational parameters, and a noise random variable, $\epsilon \sim p(\epsilon)$. For the Gaussian variational posterior above, this can be done by first sampling $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \mathrm{I})$, then transforming these to samples from the variational posterior as $\mathbf{z}^{(i)} = m_\phi(\mathbf{y}^{(n)}) + L_\phi(\mathbf{y}^{(n)}) \odot \boldsymbol{\epsilon}$, where $L_\phi(\mathbf{y}^{(n)})$ is the Cholesky factor of $C_\phi(\mathbf{y}^{(n)}) = L_\phi(\mathbf{y}^{(n)}) L_\phi(\mathbf{y}^{(n)})^\top$.

Constructing more expressive variational approximations can be achieved through normalizing flows [42,43]. A complicated distribution is modelled as a series of invertible transformations of a simple reference distribution, e.g. $p(\mathbf{w}) = \mathcal{N}(0, \mathrm{I})$. More explicitly, $\mathbf{w}^{(i)} \sim p(\mathbf{w})$, $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z})$, where $\mathbf{z}^{(i)} = f_\phi(\mathbf{w}^{(i)})$. The density for $q_\phi(\mathbf{z})$ is computed through the change of variable formula $q_\phi(\mathbf{z}) = p(f_\phi^{-1}(\mathbf{z})) \det |\partial_{\mathbf{z}} f_\phi^{-1}(\mathbf{z})|$. Conditional normalizing flows extended the normalizing flow method to learn conditional densities, i.e. $q_\phi(\mathbf{z}|\mathbf{y})$ similar to the encoder for a VAE. Normalizing flows have the benefit over VAEs of being *invertible* transformations, but as a result are constrained to having the same latent dimension as that of the data, so do not benefit from dimensionality reduction.

3. Physics-informed generative models

We now delve into salient works taken from the literature that best exemplify the flexibility and versatility of VI for physics. In what follows, we cast the central VI objective of selected works in a notation consistent with the previously presented material. This should be interpreted as a paraphrasing of the methods in the referenced works to help the reader best understand their differences and similarities. Particular implementation details such as precise residual computations or variational forms will vary.

(a) Forward-model-based learning

In this section, we describe inverse problem methodologies that embed the forward model into the probabilistic generative model. It is assumed the forward model (while still potentially expensive) can be evaluated for a given input $\mathbf{z}$—outputting a corresponding $\mathbf{y}$—and the dataset is a collection of these physical model input–output pairs, $\mathcal{D} = \{\mathbf{z}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^N$. For a probabilistic generative model, this amounts to sampling from the joint distribution $p(\mathbf{z}, \mathbf{y}) \propto p(\mathbf{z})p(\mathbf{y}|\mathbf{z})$. In this setting, the likelihood describes a probabilistic forward map, as determined by the true forward model $G^\dagger(\cdot)$ and an assumed noise model, e.g. (2.8). The central goal of these methodologies is to learn a variational approximation $q_\phi(\mathbf{z}|\mathbf{y})$, that once trained, provides a calibrated posterior estimate over parameters for a previously unseen data point.

(i) Supervised VAEs for calibrated posteriors

This class of models is for *supervised* learning problems—meaning we have access to input–output pairs. This allows for the use of the *forward* KL, $D_{\mathrm{KL}}(p(\mathbf{z}|\mathbf{y})||q_\phi(\mathbf{z}|\mathbf{y}))$ in the objective, as opposed to the mode-seeking reverse KL. The estimation of the mean-seeking forward KL requires an expectation with respect to the true posterior, which is unavailable to us. However, the average over the data distribution can be computed using samples from the joint distribution $p(\mathbf{z}, \mathbf{y})$ via

$$\mathbb{E}_{p(\mathbf{y})}\left[D_{\mathrm{KL}}(p(\mathbf{z}|\mathbf{y})||q_\phi(\mathbf{z}|\mathbf{y}))\right] = \mathbb{E}_{p(\mathbf{z},\mathbf{y})}\left[-\log q_\phi(\mathbf{z}|\mathbf{y})\right]. \tag{3.1}$$

This approach is used in [44] to learn an amortized variational approximation with sampled input–output pairs, computed via the true forward model by pushing prior samples $\mathbf{z}^{(n)} \sim p(\mathbf{z})$

through the forward model and sampling $\mathbf{y}^{(n)} \sim \mathcal{N}(G^\dagger(\mathbf{z}^{(n)}), \sigma^2\mathrm{I})$. A conditional normalizing flow provides the variational approximation $q_\phi(\mathbf{z}|\mathbf{y}) = \mathcal{N}(f_\phi^{-1}(\mathbf{z};\mathbf{y}); 0, \mathrm{I}) \det|\partial_{\mathbf{z}} f_\phi^{-1}(\mathbf{z};\mathbf{y})|$, mapping data to the latent space and acting as a surrogate. The forward KL averaged over the data distribution, and (3.1) is the objective to learn the conditional normalizing flow as

$$\phi^\star = \arg\min_\phi J(\phi;\mathbf{y}), \quad J(\phi;\mathbf{y}) = \mathbb{E}_{p(\mathbf{z},\mathbf{y})}\left[\frac{1}{2}\|f_\phi^{-1}(\mathbf{z};\mathbf{y})\|_2^2 - \log\det|\partial_{\mathbf{z}} f_\phi^{-1}(\mathbf{z};\mathbf{y})|\right]. \tag{3.2}$$

The posterior given an unseen data point is then computed by sampling $\mathbf{w} \sim \mathcal{N}(0, \mathrm{I})$ and pushing through the trained conditional normalizing flow $f_{\phi^\star}(\mathbf{w};\mathbf{y})$ which approximately samples from $p(\mathbf{z}|\mathbf{y})$.

In [45], the decoder of a VAE is replaced by the known physical forward model, which acts to physically regularize the problem. Data is assumed to be observed under some known noise model $\mathbf{y} \sim \mathcal{N}(G^\dagger(\mathbf{z}) + m_\epsilon, C_\epsilon)$, which can include a bias through the mean m_ϵ. Input–output pairs are used to learn an amortized variational posterior with mean $m_\phi(\cdot)$, and covariance square root $C_\phi^{1/2}(\cdot)$ parametrized by neural networks, yielding $q_\phi(\mathbf{z}|\mathbf{y}) = \mathcal{N}(m_\phi(\mathbf{y}), C_\phi(\mathbf{y}))$. The Jensen–Shannon divergence, which is parametrized by $\alpha \in [0,1]$, interpolates between the forward ($\alpha = 0$) and reverse ($\alpha = 1$) KL. The form of this divergence between $q := q(\mathbf{z})$ and $p := p(\mathbf{z})$ is

$$\mathrm{JS}_\alpha(q||p) = \alpha D_{\mathrm{KL}}(q||(1-\alpha)q + \alpha p) + (1-\alpha)D_{\mathrm{KL}}(p||(1-\alpha)q + \alpha p). \tag{3.3}$$

A weighted Jensen–Shannon divergence is incorporated into their variational objective alongside the standard reverse KL as

$$\phi^\star = \arg\min_\phi J(\phi;\alpha,\mathbf{y}), \quad J(\phi;\alpha,\mathbf{y}) = \frac{1}{\alpha}\mathrm{JS}_\alpha(q_\phi(\mathbf{z}|\mathbf{y})||p(\mathbf{z}|\mathbf{y})) + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{y})||p(\mathbf{z}|\mathbf{y})), \tag{3.4}$$

where the parameter α allows for a trade-off between data-fit and regularization, said to help regularize the problem, preventing either extremely low or high values of posterior variance. For expensive forward models, the exact forward model can be replaced by a surrogate decoder $p_\theta(\mathbf{y}|\mathbf{z}) = \mathcal{N}(G_\theta(\mathbf{z}), \sigma^2\mathrm{I})$, $G_\theta H^\dagger \circ F_\theta \circ \pi_z$ and the encoder and decoder parameters are learned simultaneously.

(ii) Dynamical latent spaces

Embedding dynamical structure into the latent space of a VAE has been considered to model time-indexed data $\mathbf{y}_{1:N} = \{\mathbf{y}_n\}_{n=1}^N$. In [46], a probabilistic forward model drives the latent solution, and an auxiliary variable, $\mathbf{x}_n$ is introduced as the *pseudo-observable*, representing the observations of the latent Gaussian state-space model. This yields the likelihoods $p(\mathbf{x}_n|\mathbf{u}_n) = \mathcal{N}(\tilde{H}(\mathbf{u}_n), \sigma_{\mathbf{x}}^2\mathrm{I})$ and $p(\mathbf{u}_n|\mathbf{u}_{n-1}) = \mathcal{N}(\Psi^\dagger(\mathbf{u}_{n-1};\mathbf{z}), \sigma_{\mathbf{u}}^2\mathrm{I})$, where $\tilde{H}$ is the *known* pseudo-observation operator, and $\Psi^\dagger$ is the one-step evolution operator of the latent dynamical system, which depends on parameters $\mathbf{z}$. The generative model learns to reconstruct data from the pseudo-observable with a probabilistic decoder, $p_\theta(\mathbf{y}_n|\mathbf{x}_n) = \mathcal{N}(H_\theta(\mathbf{x}_n), \sigma^2\mathrm{I})$, where the true mapping is approximated $H^\dagger \approx H_\theta \circ \tilde{H}$. The variational posterior is factorized as

$$q(\mathbf{u}_{1:N}, \mathbf{x}_{1:N}, \mathbf{z}|\mathbf{y}_{1:N}) \propto p(\mathbf{u}_{1:N}|\mathbf{x}_{1:N}) q_\vartheta(\mathbf{z}) \prod_n q_\phi(\mathbf{x}_n|\mathbf{y}_n), \tag{3.5}$$

which uses an amortized encoder $q_\phi(\mathbf{x}_n|\mathbf{y}_n)$, variational approximation $q_\vartheta(\mathbf{z})$ and exact posterior $p(\mathbf{u}_{1:N}|\mathbf{x}_{1:N})$. We obtain the desired parameters $(\theta^\star, \phi^\star, \vartheta^\star)$ by maximizing the ELBO

$$\begin{aligned} J(\theta,\phi,\vartheta;\mathbf{y}_{1:N}) = \sum_n \mathbb{E}_{q_\phi(\mathbf{x}_n|\mathbf{y}_n)}\left[\log\frac{p_\theta(\mathbf{y}_n|\mathbf{x}_n)}{q_\phi(\mathbf{x}_n|\mathbf{y}_n)}\right] \\ + \mathbb{E}_{q_\phi(\mathbf{x}_n|\mathbf{y}_n)q_\vartheta(\mathbf{z})}\left[\log p(\mathbf{x}_{1:N}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\vartheta(\mathbf{z})||p(\mathbf{z})). \end{aligned} \tag{3.6}$$

The term $\log p(\mathbf{x}_{1:N}|\mathbf{z})$ is computed using Kalman filtering. Similarly, dynamical latent structure is imposed in [47] by constraining the latent embeddings to non-Euclidean manifolds, improving the robustness to noise and improving interpretability of latent dynamics.

(iii) Deep generative priors for regularization

When the parameter space is high-dimensional, regularizing the inverse problem is essential. Furthermore, if direct observations of the parameters are available, a possible method of regularization is through the use of a deep generative prior (DGP) over the parameter space. By introducing a lower-dimensional auxiliary latent variable $\mathbf{w}$, a generative model $p_\theta(\mathbf{w}, \mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{w})p(\mathbf{w})$ can be trained to approximately generate samples from the prior $p(\mathbf{z})$, where the likelihood is constructed as a probabilistic decoder, e.g. $p(\mathbf{z}|\mathbf{w}) = \mathcal{N}(f_\theta(\mathbf{w}), \sigma^2\mathrm{I})$, with learnable generator function f_θ. Including the DGP in the inverse problem acts as a form of regularization when optimization is performed over the low-dimensional $\mathbf{w}$ rather than the high-dimensional $\mathbf{z}$. Typically, VAEs are suitable here [48] because of the in-built dimensionality reduction, and once trained, the decoder can produce samples from the DGP via $\mathbf{z}^{(i)} = f_{\theta\star}(\mathbf{w}^{(i)})$, with $\mathbf{w}^{(i)} \sim p(\mathbf{w})$ (here f_θ need not be invertible). The auxiliary prior can be set arbitrarily, most simply as a standard multivariate Gaussian.

For solving the inverse problem, in [49], a point-estimate-based inversion viewpoint is taken, where the optimization is performed w.r.t. auxiliary variables, which are pushed through the trained generator and then the forward model to obtain the data-misfit loss

$$J(\mathbf{w}; \mathbf{y}, \theta^\star) = \|G^\dagger \circ f_{\theta\star}(\mathbf{w}) - \mathbf{y}\|^2 + \beta(\|\mathbf{w}\| - \mu_\chi)^2, \tag{3.7}$$

where the constant μ_χ in the regularization term preferences $\mathbf{w}$ lie on a ring centred at the origin. The resulting parameter estimate is found by pushing the optimal $\mathbf{w}^\star = \arg\min_{\mathbf{w}} J(\mathbf{w}; \mathbf{y}, \theta^\star)$ through the generator, giving $\mathbf{z}^\star = f_{\theta\star}(\mathbf{w}^\star)$.

One might consider learning probabilistic priors for inversion through the use of normalizing flows. In [50], the authors trained a normalizing flow to learn a prior in an embedded space—where the embedding itself is learned with a VAE or generative adversarial network (GAN).

In [51], a simple DGP is trained for sampling $p(\mathbf{z})$, which is included in a Bayesian VI problem where the auxiliary posterior $p(\mathbf{w}|\mathbf{y})$ is approximated by the VI approximation $q_\phi(\mathbf{w})$. The objective is

$$\phi^\star = \arg\min_\phi J(\phi; \mathbf{y}, \theta^\star),\ J(\phi; \mathbf{y}, \theta^\star) = \mathbb{E}_{q_\phi(\mathbf{w})}\left[-\log p(\mathbf{y}|\mathbf{w})\right] + \mathrm{KL}(q_\phi(\mathbf{w})|p(\mathbf{w})), \tag{3.8}$$

where the likelihood $p(\mathbf{y}|\mathbf{w}) := p(\mathbf{y}|\mathbf{z} = f_{\theta\star}(\mathbf{w}))$ is determined by the forward model, $\mathbf{y} = G^\dagger \circ f_{\theta\star}(\mathbf{w}) + \boldsymbol{\epsilon}$. Posterior samples can then be readily obtained by sampling from this variational posterior and pushing through the generator, $\mathbf{z}^{(i)} = f_{\theta\star}(\mathbf{w}^{(i)})$, with $\mathbf{w}^{(i)} \sim q_{\phi\star}(\mathbf{w})$.

(b) Residual-based learning

The objective of VI-based deep surrogate modelling is to predict solutions of PDEs using deep learning models that output uncertainty about their predictions. Such surrogates are of great use for solving inverse problems as they can replace computationally expensive numerical forward models while quantifying the error of their approximations, which can be incorporated into inversion schemes [52].

(i) Data-free inference

For the work in [27], the authors model the PDE solution $\mathbf{u}$ given a parameter $\mathbf{z}$ probabilistically through a residual $\mathbf{r}(u_h, z_h)$ with

$$p_\beta(\mathbf{u}|\mathbf{z}) \propto \exp\left(-\beta\, \|\mathbf{r}(\pi_u(\mathbf{u}), \pi_z(\mathbf{z}))\|_2^2\right), \tag{3.9}$$

where the exact formulation of the residual $\mathbf{r}(u_h, z_h)$ can vary, but its purpose remains the same; $\mathbf{r} = 0$ when u_h satisfies the PDE system for parameters z_h. We then seek the parameters

$$\phi^\star = \arg\min_\phi D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z})p(\mathbf{z})||p_\beta(\mathbf{u}|\mathbf{z})p(\mathbf{z})), \tag{3.10}$$

where β controls the intensity of the physics constraint and is selected such that the surrogate model $q_\phi(\mathbf{u}|\mathbf{z})$ provides calibrated uncertainty estimates given a dataset $\mathcal{D} = \{\mathbf{u}^{(n)}, \mathbf{z}^{(n)}\}_{n=1}^N$ of solution–parameter pairs. In their work, the authors make use of a normalizing flow to model the forward problem $q_\phi(\mathbf{u}|\mathbf{z})$. This variational construction learns a probabilistic forward model.

In [53,54], different variational frameworks are proposed which allow for the learning of both forward and inverse probabilistic maps. The construction is posed through a parametrized probabilistic model $p_\theta(\hat{\mathbf{r}}, \mathbf{u}, \mathbf{z}) = p(\hat{\mathbf{r}}|\mathbf{u}, \mathbf{z})p_\theta(\mathbf{z}|\mathbf{u})p(\mathbf{u})$ and a variational approximation $q_\phi(\mathbf{u}, \mathbf{z}) = q_\phi(\mathbf{u}|\mathbf{z})q(\mathbf{z})$. Here, $\hat{\mathbf{r}}$ represents a zero-valued *virtual observable* [55] posed as

$$\hat{\mathbf{r}} = \mathbf{r}(\pi_u(\mathbf{u}), \pi_z(\mathbf{z})) + \epsilon_r, \quad \epsilon_r \sim \mathcal{N}(0, \sigma_r^2 \mathrm{I}). \tag{3.11}$$

We note that other virtual noise models may be considered, leading to different residual likelihoods [56]. The factorization of the joint variational approximation $q_\phi(\mathbf{u}, \mathbf{z})$ and the model $p_\theta(\mathbf{u}, \mathbf{z}|\hat{\mathbf{r}})$ is chosen such that

$$\phi^\star, \theta^\star = \arg\max_{\phi,\theta} J(\phi, \theta), \quad J(\phi, \theta) = \mathbb{E}_{q_\phi(\mathbf{u}|\mathbf{z})p(\mathbf{z})} \log \frac{p(\hat{\mathbf{r}} = 0|\mathbf{u}, \mathbf{z})p_\theta(\mathbf{z}|\mathbf{u})p(\mathbf{u})}{q_\phi(\mathbf{u}|\mathbf{z})p(\mathbf{z})}, \tag{3.12}$$

learns mapping for forward UQ ($q_\phi(\mathbf{u}|\mathbf{z})$) and inversion ($p_\theta(\mathbf{z}|\mathbf{u})$). It is a lower bound on the log marginal probability of $\hat{\mathbf{r}}$. In the same spirit as (3.9) (with $\beta = 1/2\sigma_r^2$), the distribution over the residual is posed as $p(\hat{\mathbf{r}} = 0|\mathbf{u}, \mathbf{z}) \propto \exp(-\frac{1}{2\sigma_r^2}\|\mathbf{r}(\pi_u(\mathbf{u}), \pi_z(\mathbf{z}))\|_2^2)$. These frameworks construct variational uncertainty quantifying surrogates in the data-free regime.

(ii) Small data regime

In some settings, one may have access to small datasets alongside knowledge of the form of the underlying physics. Methods for constructing probabilistic forward surrogates may pose their likelihood as a product measure between a virtually observed residual $\hat{\mathbf{r}}$ and data $\mathbf{y}$ as in [57]. Using this approach, one can combine (possibly high fidelity) data with fast to evaluate physics residuals in the likelihood

$$p(\hat{\mathbf{r}}, \mathbf{y}|\mathbf{u}, \mathbf{z}) = p(\hat{\mathbf{r}} = 0|\mathbf{u}, \mathbf{z})p(\mathbf{y}|\mathbf{u}, \mathbf{z}), \tag{3.13}$$

where the balance between data and physics residual is given by the estimated variance of the data noise and chosen virtual observational noise of the residual. A Bayesian VI objective can be written using (2.12) to obtain an approximate posterior over the solution $\mathbf{u}$ and parameters $\mathbf{z}$ as

$$\phi^\star = \arg\min_\phi J(\phi), \quad J(\phi) = D_{\mathrm{KL}}(q_\phi(\mathbf{u}, \mathbf{z})||p(\mathbf{u}, \mathbf{z}|\mathbf{y}, \hat{\mathbf{r}})). \tag{3.14}$$

Here $q_\phi(\mathbf{u}, \mathbf{z})$ is factorized independently as $q_\phi(\mathbf{u})q_\phi(\mathbf{z})$—called the mean field approximation [58]—and the dependence between the parameter and solution to the PDE is captured in the likelihood through the virtual observable $\hat{\mathbf{r}} = 0$. Similar in objective is [59], where a joint variational approximation $q_\phi(\mathbf{u}, \mathbf{z})$ is used to approximate the Bayesian posterior $p(\mathbf{u}, \mathbf{z}|\mathbf{y})$, factorizing $q_\phi(\mathbf{u}, \mathbf{z}) = q_\phi(\mathbf{u}|\mathbf{z})q_\phi(\mathbf{z})$ where the likelihood $q_\phi(\mathbf{u}|\mathbf{z}) = \mathcal{N}(\mathbf{u}; F_\phi(\mathbf{z}), \epsilon^2 C(\mathbf{z}))$ captures the forward map. Furthermore, [59] uses the information from the physics problem through the stiffness matrix to inform the covariance $C(\mathbf{z})$. The parameter ϵ controls the strength of the physics constraint in the likelihood, and in the limit $\epsilon \to 0$, the following problem is recovered

$$\theta^\star, \phi^\star = \arg\min_{\theta,\phi} \mathbb{E}_{q_\phi(\mathbf{z})}\left[-\log p(\mathbf{y}|\mathbf{u} = F_\theta(\mathbf{z}))\right] + D_{\mathrm{KL}}(q_\phi(\mathbf{z})||p(\mathbf{z})), \tag{3.15a}$$

$$\text{s.t. } \|\mathbf{r}(\pi_u(F_\theta(\mathbf{z})), \pi_z(\mathbf{z}))\|_2^2 = 0. \tag{3.15b}$$

Notice in this interpretation, the learning of $F_\theta(\mathbf{z})$ is part of the *probabilistic model* not the *variational approximation*, hence changing F_ϕ for F_θ. This constrained optimization view is in effect similar to having access to the forward model $F^\dagger$. In [60], a deterministic forward surrogate $F_\theta \approx F^\dagger$ is

learned by minimizing $\|(F_\theta - F^\dagger)\circ\pi_z(\mathbf{z})\|_2^2$ in conjunction with a normalizing flow that probabilistically solves the inverse problem. We note that for many of these inversion methods, amortization could be used to learn a mapping to the posteriors given data from varying physical systems. Relevant to the aforementioned methods, the work in [61] uses VI to synthesize information for coarse-grained models in the small data regime. This model is also used to learn efficient latent representations of structured high-dimensional feature spaces, arising in problems in porous media [62]. Further methods propose VI surrogate models in the small data regime for related applications [63].

Methods for handling stochastic PDEs have also been developed to solve forward and inverse problems when the solution, parameters and source terms are described by random fields. These fields may only be sparsely observed over a number of sensor locations. The variational autoencoder approaches in [64,65] encode observations to auxiliary random variables, which capture the stochastic behaviour of the PDEs, with physics-informed losses constructed from PDE residual terms. Aside from VAEs, other VI variants include physics-informed generative adversarial networks (PI-GAN) [66], and normalizing field flows (NFF) [67] use physics-informed flows and are agnostic to sensor/observation location.

4. Discussion

This paper introduces the core concepts necessary for constructing VI schemes for solving physics-based forward and inverse problems. Furthermore, we review the literature that employs VI and deep learning in the context of physics, presenting the contributions under a unified notation. Our approach is intended to help readers better understand the similarities, differences and nuances among the various methodologies proposed in the field. A few limitations are to be kept in mind when applying and developing some of the mentioned works. As highlighted in [27], care must be taken in assessing the accuracy of UQ with VI, which remains an open practical [68] and theoretical challenge [69]. In applications, one should also assess the computational advantage of *training* any surrogate model versus directly making use of classical numerical schemes [70]. Software libraries are being developed to aid practitioners in the implementation of these schemes, e.g. [71]. Furthermore, the use of the KL divergence may not always be well-posed, particularly when dealing with functional objects such as in physics applications [72]. As such, beyond the Bayesian formulation of VI, promising areas of research consider other divergences [19] such as those based on the Wasserstein [73,74] and Sliced Wasserstein metrics [26,75] or Maximum Mean Discrepancy [65,76] as these do not have the same conditions on absolute continuity and are readily computable from random samples. Finally, many promising developments in solving physics-based inverse problems through deep learning and possibly variational inference focus on learning better priors [26,77–79] along with important earlier works in Earth sciences [48,80].

Data accessibility. This article has no additional data.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. A.G.-D.: conceptualization, writing—original draft, writing—review and editing; A.V.: conceptualization, writing—original draft, writing—review and editing; O.D.A.: conceptualization, writing—review and editing; I.K.: conceptualization, writing—original draft, writing—review and editing; M.G.: conceptualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. AGD is supported by Splunk Inc. [G106483] PhD scholarship funding. AV is supported through the EPSRC ROSEHIPS grant [EP/W005816/1]. MG is supported by a Royal Academy of Engineering Research Chair and EPSRC grants [EP/X037770/1, EP/Y028805/1, EP/W005816/1, EP/V056522/1, EP/V056441/1, EP/T000414/1 and EP/R034710/1].

References

1. Janssen MA, Ostrom E. 2006 Empirically based, agent-based models. *Ecol. Soc* **11**, 34. (doi:10.5751/es-01861-110237)
2. Kirchdoerfer T, Ortiz M. 2016 Data-driven computational mechanics. *Comput. Methods Appl. Mech. Eng*. **304**, 81–101. (doi:10.1016/j.cma.2016.02.001)
3. Finlayson BA. 2013 *The method of weighted residuals and variational principles*. Philadelphia, PA: SIAM.
4. Sullivan TJ. 2015 *Introduction to uncertainty quantification*. vol. 63. Cham, Switzerland: Springer.
5. Kennedy MC, O'Hagan A. 2001 Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B* **63**, 425–464. (doi:10.1111/1467-9868.00294)
6. Li Z, Kovachki NB, Azizzadenesheli K, liu B, Bhattacharya K, Stuart A, Anandkumar A. 2021 Fourier neural operator for parametric partial differential equations. In *Int. Conf. on Learning Representations*, Appleton, WI: ICLR.
7. Lu L, Jin P, Karniadakis GE. 2019 DeepONet: learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv*. (doi:10.48550/arXiv.1910.03193)
8. Lin G, Moya C, Zhang Z. 2023 B-deeponet: an enhanced Bayesian deepoNet for solving noisy parametric PDEs using accelerated replica exchange sgld. *J. Comput. Phys*. **473**, 111713. (doi:10.1016/j.jcp.2022.111713)
9. Psaros AF, Meng X, Zou Z, Guo L, Karniadakis GE. 2023 Uncertainty quantification in scientific machine learning: methods, metrics, and comparisons. *J. Comput. Phys*. **477**, 111902. (doi:10.1016/j.jcp.2022.111902)
10. Ramm AG, Katsevich AI. 2020 The radon transform and local tomography. Boca Raton, FL: CRC press.
11. Trotta R. 2008 Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemp. Phys*. **49**, 71–104. (doi:10.1080/00107510802066753)
12. Zhdanov MS. 2002 *Geophysical inverse theory and regularization problems*. vol. 36. Amsterdam, The Netherlands: Elsevier.
13. Benning M, Burger M. 2018 Modern regularization methods for inverse problems. *Acta Numer*. **27**, 1–111. (doi:10.1017/s0962492918000016)
14. Arridge S, Maass P, Öktem O, Schönlieb CB. 2019 Solving inverse problems using data-driven models. *Acta Numer*. **28**, 1–174. (doi:10.1017/s0962492919000059)
15. Stuart AM. 2010 Inverse problems: a Bayesian perspective. *Acta Numer*. **19**, 451–559. (doi:10.1017/s0962492910000061)
16. Bishop CM. *Pattern recognition and machine learning*. vol. 4. Cham, Switzerland: Springer.
17. McGrory CA, Titterington DM. 2007 Variational approximations in Bayesian model selection for finite mixture distributions. *Comput. Stat. Data Anal*. **51**, 5352–5367. (doi:10.1016/j.csda.2006.07.020)
18. Blei DM, Kucukelbir A, McAuliffe JD. 2017 Variational inference: a review for statisticians. *J. Am. Stat. Assoc*. **112**, 859–877. (doi:10.1080/01621459.2017.1285773)
19. Knoblauch J, Jewson J, Damoulas T. 2019 Generalized variational inference: three arguments for deriving new posteriors. *arXiv*. (doi:10.48550/arXiv.1904.02063)
20. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. 1999 An introduction to variational methods for graphical models. *Mach. Learn*. **37**, 183–233.
21. Pearl J. 2014 *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Elsevier.
22. Wainwright MJ, Jordan MI. 2007 Graphical models, exponential families, and variational inference. *Found. Trends® Mach. Learn*. **1**, 1–305. (doi:10.1561/2200000001)
23. Barber D. 2012 *Bayesian reasoning and machine learning*. Cambridge, UK: Cambridge University Press.
24. Giné E, Nickl R. 2021 *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press. (doi:10.1017/9781009022811)
25. Reddy JN. 1993 *An introduction to the finite element method*, p. 14, vol. 27. New York, NY, McGraw-Hill.

26. Akyildiz OD, Girolami M, Stuart AM, Vadeboncoeur A. 2024 Efficient prior calibration from indirect data. *arXiv*. (doi:10.48550/arXiv.2405.17955)
27. Zhu Y, Zabaras N, Koutsourelakis PS, Perdikaris P. 2019 Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys*. **394**, 56–81. (doi:10.1016/j.jcp.2019.05.024)
28. Leissa AW. 2005 The historical bases of the rayleigh and ritz methods. *J. Sound Vib*. **287**, 961–978. (doi:10.1016/j.jsv.2004.12.021)
29. Raissi M, Perdikaris P, Karniadakis GE. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys*. **378**, 686–707. (doi:10.1016/j.jcp.2018.10.045)
30. Kharazmi E, Zhang Z, Karniadakis GEM. 2021 hp-VPINNs: variational physics-informed neural networks with domain decomposition. *Comput. Methods Appl. Mech. Eng*. **374**, 113547. (doi:10.1016/j.cma.2020.113547)
31. Cai S, Mao Z, Wang Z, Yin M, Karniadakis GE. 2021 Physics-informed neural networks (PINNs) for fluid mechanics: a review. *Acta Mech. Sin*. **37**, 1727–1738. (doi:10.1007/s10409-021-01148-1)
32. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. 2021 Physics-informed machine learning. *Nat. Rev. Phys*. **3**, 422–440. (doi:10.1038/s42254-021-00314-5)
33. Lu L, Meng X, Mao Z, Karniadakis GE. 2021 Deepxde: a deep learning library for solving differential equations. *SIAM Rev*. **63**, 208–228. (doi:10.1137/19m1274067)
34. Sukumar N, Srivastava A. 2022 Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks. *Comput. Methods Appl. Mech. Eng*. **389**, 114333. (doi:10.1016/j.cma.2021.114333)
35. Chan T, Esedoglu S, Park F, Yip A. 2005 Recent developments in total variation image restoration. *Math. Model. Comput. Vis*. **17**, 17–31.
36. Tibshirani R. 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288. (doi:10.1111/j.2517-6161.1996.tb02080.x)
37. Holler G, Kunisch K, Barnard RC. 2018 A bilevel approach for parameter learning in inverse problems. *Inverse Probl*. **34**, 115012. (doi:10.1088/1361-6420/aade77)
38. Sanz-Alonso D, Stuart AM, Taeb A. 2023 *Inverse problems and data assimilation*. vol. 107. Cambridge, UK: Cambridge University Press.
39. Ranganath R, Gerrish S, Blei D. 2014 Black box variational inference. In *Artificial intelligence and statistics* (eds S Kaski, J Corander), pp. 814–822. PMLR.
40. Kingma DP, Welling M. 2013 Auto-encoding variational Bayes. *arXiv*. (doi:10.48550/arXiv.1312.6114)
41. Zhang C, Butepage J, Kjellstrom H, Mandt S. 2019 Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell*. **41**, 2018. (doi:10.1109/TPAMI.2018.2889774)
42. Dinh L, Sohl-Dickstein J, Bengio S. 2022 Density estimation using real NVP. In *Int. Conf. on Learning Representations*, Appleton, WI: ICLR.
43. Rezende D, Mohamed S. 2015 Variational inference with normalizing flows. In *Int. Conf. on machine learning*, Lille, France, pp. 1530–1538. PMLR.
44. Siahkoohi A, Rizzuti G, Orozco R, Herrmann FJ. 2023 Reliable amortized variational inference with physics-based latent distribution correction. *Geophysics* **88**, R297–R322. (doi:10.1190/geo2022-0472.1)
45. Goh H, Sheriffdeen S, Wittmer J, Bui-Thanh T. 2022 Solving Bayesian inverse problems via variational autoencoders. In *Proc. of the 2nd Mathematical and Scientific Machine Learning Conf*, pp. 386–425. Virtual: PMLR.
46. Glyn-Davies A, Duffin C, Deniz Akyildiz O, Girolami M. 2024 Φ-DVAE: physics-informed dynamical variational autoencoders for unstructured data assimilation. *J. Comput. Phys*. **515**, 113293. (doi:10.1016/j.jcp.2024.113293)
47. Lopez R, Atzberger PJ. 2020 Variational autoencoders for learning nonlinear dynamics of physical systems. *arXiv*. (doi:10.48550/arXiv.2012.03448)
48. Laloy E, Hérault R, Lee J, Jacques D, Linde N. 2017 Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour*. **110**, 387–405. (doi:10.1016/j.advwatres.2017.09.029)

49. Lopez-Alvis J, Laloy E, Nguyen F, Hermans T. 2021 Deep generative models in inversion: the impact of the generator's nonlinearity and development of a new approach based on a variational autoencoder. *Comput. Geosci*. **152**, 104762. (doi:10.1016/j.cageo.2021.104762)
50. Levy S, Laloy E, Linde N. 2023 Variational Bayesian inference with complex geostatistical priors using inverse autoregressive flows. *Comput. Geosci*. **171**, 105263. (doi:10.1016/j.cageo.2022.105263)
51. Xia Y, Liao Q, Li J. 2023 VI-DGP: a variational inference method with deep generative prior for solving high-dimensional inverse problems. *J. Sci. Comput*. **97**, 16. (doi:10.1007/s10915-023-02328-w)
52. Cleary E, Garbuno-Inigo A, Lan S, Schneider T, Stuart AM. 2021 Calibrate, emulate, sample. *J. Comput. Phys*. **424**, 109716. (doi:10.1016/j.jcp.2020.109716)
53. Vadeboncoeur A, Akyildiz ÖD, Kazlauskaite I, Girolami M, Cirak F. 2023 Fully probabilistic deep models for forward and inverse problems in parametric PDEs. *J. Comput. Phys*. **491**, 112369. (doi:10.1016/j.jcp.2023.112369)
54. Vadeboncoeur A, Kazlauskaite I, Papandreou Y, Cirak F, Girolami M, Akyildiz OD. 2023 Random grid neural processes for parametric partial differential equations. In *Int. Conf. on Machine Learning*, pp. 34759–34778. Honolulu, HI: PMLR.
55. Rixner M, Koutsourelakis PS. 2021 A probabilistic generative model for semi-supervised training of coarse-grained surrogates and enforcing physical constraints through virtual observables. *J. Comput. Phys*. **434**, 110218. (doi:10.1016/j.jcp.2021.110218)
56. Chatzopoulos M, Koutsourelakis PS. 2024 Physics-aware neural implicit solvers for multiscale, parametric pdes with applications in heterogeneous media. *arXiv*. (doi:10.48550/arXiv.2405.19019)
57. Kaltenbach S, Koutsourelakis PS. 2020 Incorporating physical constraints in a deep probabilistic machine learning framework for coarse-graining dynamical systems. *J. Comput. Phys*. **419**, 109673. (doi:10.1016/j.jcp.2020.109673)
58. Parisi G, Shankar R. 1988 Statistical field theory. Redwood City, CA: Addison Wesley.
59. Tait DJ, Damoulas T. 2020 Variational autoencoding of PDE inverse problems. *arXiv*. (doi:10.48550/arXiv.2006.15641)
60. Wang Y, Liu F, Schiavazzi DE. 2022 Variational inference with NoFAS: normalizing flow with adaptive surrogate for computationally expensive models. *J. Comput. Phys*. **467**, 111454. (doi:10.1016/j.jcp.2022.111454)
61. Grigo C, Koutsourelakis PS. 2019 A physics-aware, probabilistic machine learning framework for coarse-graining high-dimensional systems in the small data regime. *J. Comput. Phys*. **397**, 108842. (doi:10.1016/j.jcp.2019.05.053)
62. Dasgupta A, Patel DV, Ray D, Johnson EA, Oberai AA. 2024 A dimension-reduced variational approach for solving physics-based inverse problems using generative adversarial network priors and normalizing flows. *Comput. Methods Appl. Mech. Eng*. **420**, 116682. (doi:10.1016/j.cma.2023.116682)
63. Rixner M, Koutsourelakis PS. 2022 Self-supervised optimization of random material microstructures in the small-data regime. *Npj Comput. Mater*. **8**, 46. (doi:10.1038/s41524-022-00718-6)
64. Shin H, Choi M. 2023 Physics-informed variational inference for uncertainty quantification of stochastic differential equations. *J. Comput. Phys*. **487**, 112183. (doi:10.1016/j.jcp.2023.112183)
65. Zhong W, Meidani H. 2023 PI-VAE: physics-informed variational auto-encoder for stochastic differential equations. *Comput. Methods Appl. Mech. Eng*. **403**, 115664. (doi:10.1016/j.cma.2022.115664)
66. Yang L, Zhang D, Karniadakis GE. 2020 Physics-informed generative adversarial networks for stochastic differential equations. *SIAM J. Sci. Comput*. **42**, A292–A317. (doi:10.1137/18m1225409)
67. Guo L, Wu H, Zhou T. 2022 Normalizing field flows: solving forward and inverse stochastic differential equations using physics-informed flow models. *J. Comput. Phys*. **461**, 111202. (doi:10.1016/j.jcp.2022.111202)
68. Povala J, Kazlauskaite I, Febrianto E, Cirak F, Girolami M. 2022 Variational Bayesian approximation of inverse problems using sparse precision matrices. *Comput. Methods Appl. Mech. Eng*. **393**, 114712. (doi:10.1016/j.cma.2022.114712)

69. Wang Y, Blei DM. 2019 Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc*. **114**, 1147–1161. (doi:10.1080/01621459.2018.1473776)
70. de Hoop MV, Huang DZ, Qian E, Stuart AM. 2022 The cost-accuracy trade-off in operator learning with neural networks. *arXiv* (doi:10.48550/arXiv.2203.13181)
71. Zou Z, Meng X, Psaros AF, Karniadakis GE. 2024 NeuralUQ: a comprehensive library for uncertainty quantification in neural differential equations and operators. *SIAM Rev*. **66**, 161–190. (doi:10.1137/22m1518189)
72. Bunker J, Girolami M, Lambley H, Stuart AM, Sullivan T. 2024 Autoencoders in function space. *arXiv*. (doi:10.48550/arXiv.2408.01362)
73. Ambrogioni L, Güçlü U, Güçlütürk Y, Hinne M, van Gerven MA, Maris E. 2018 Wasserstein variational inference. In *Advances in neural information processing systems*. Red Hook, NJ: Curran Associates.
74. Yao R, Yang Y. 2022 Mean field variational inference via Wasserstein gradient flow. *arXiv*. (doi:10.48550/arXiv.2207.08074)
75. Yi M, Liu S. 2023 Sliced Wasserstein variational inference. In *Asian Conf. on Machine Learning*, pp. 1213–1228. İstanbul, Turkey: PMLR.
76. Chérief-Abdellatif BE, Alquier P. 2020 MMD-Bayes: robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 1–21. Vancouver, Canada: PMLR.
77. Meng X, Yang L, Mao Z, del Águila Ferrandis J, Karniadakis GE. 2022 Learning functional priors and posteriors from data and physics. *J. Comput. Phys*. **457**, 111073. (doi:10.1016/j.jcp.2022.111073)
78. Patel DV, Oberai AA. 2021 GAN-based priors for quantifying uncertainty in supervised learning. *SIAM/ASA J. Uncertain. Quantif*. **9**, 1314–1343. (doi:10.1137/20m1354210)
79. Patel DV, Ray D, Oberai AA. 2022 Solution of physics-based Bayesian inverse problems with deep generative priors. *Comput. Methods Appl. Mech. Eng*. **400**, 115428. (doi:10.1016/j.cma.2022.115428)
80. Laloy E, Hérault R, Jacques D, Linde N. 2018 Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resour. Res*. **54**, 381–406. (doi:10.1002/2017wr022148)