Adaptive RKHS Fourier Features for Compositional Gaussian Process Models

Xinxing Shi[1,2] **Thomas Baldwin-McDonald**[1] **Mauricio A. Álvarez**[1]

[1] Department of Computer Science
The University of Manchester, UK
[2]xinxing.shi@postgrad.manchester.ac.uk

Abstract

Deep Gaussian Processes (DGPs) leverage a compositional structure to model non-stationary processes. DGPs typically rely on local inducing point approximations across intermediate GP layers. Recent advances in DGP inference have shown that incorporating global Fourier features from the Reproducing Kernel Hilbert Space (RKHS) can enhance the DGPs' capability to capture complex non-stationary patterns. This paper extends the use of these features to compositional GPs involving linear transformations. In particular, we introduce Ordinary Differential Equation(ODE)–based RKHS Fourier features that allow for adaptive amplitude and phase modulation through convolution operations. This convolutional formulation relates our work to recently proposed deep latent force models, a multi-layer structure designed for modelling nonlinear dynamical systems. By embedding these adjustable RKHS Fourier features within a doubly stochastic variational inference framework, our model exhibits improved predictive performance across various regression tasks.

1 INTRODUCTION

Gaussian Processes (GPs) provide a principled Bayesian framework for function approximation, making them particularly useful in many applications requiring uncertainty calibration (Rasmussen and Williams, 2006), such as Bayesian optimisation (Snoek et al., 2012) and time-series analysis (Roberts et al., 2013). Despite offering reasonable uncertainty estimation, shallow GPs often struggle to model complex, non-stationary processes present in practical applications. To overcome this limitation, Deep Gaussian Processes (DGPs) employ a compositional architecture by stacking multiple GP layers, thereby enhancing representational power while preserving the model's intrinsic capability to quantify uncertainty (Damianou and Lawrence, 2013). However, the conventional variational formulation of DGPs heavily depends on local inducing point approximations across GP layers (Titsias, 2009; Salimbeni and Deisenroth, 2017), which hinder the model from capturing the global structures commonly found in real-world scenarios.

Incorporating *Fourier features* into GP models has shown promise in addressing this challenge in GP inference due to the periodic nature of these features. A line of research uses Random Fourier Features (RFFs) (Rahimi and Recht, 2007) of stationary kernels to convert the original (deep) GPs into Bayesian networks in weight space (Lázaro-Gredilla et al., 2010; Gal and Turner, 2015; Cutajar et al., 2017). Building on this concept within a sparse variational GP framework, recent advancements in inter-domain GPs (Lázaro-Gredilla and Figueiras-Vidal, 2009a; Van der Wilk et al., 2020) directly approximate the posterior of the original GPs by introducing *fixed* Variational Fourier Features (VFFs) through process projection onto a Reproducing Kernel Hilbert Space (RKHS)(Hensman et al., 2018; Rudner et al., 2020).

VFFs are derived by projecting GPs onto a different domain. The original GP posterior that these VFFs attempt to approximate remains within the same functional space as the original GP. In this setting, the VFFs produce a set of static basis functions determined by a fixed set of frequency values. To enhance these features and introduce greater flexibility, we pro-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

pose a generalised approach that incorporates Fourier features into inter-domain GPs through linear transformations, such as convolution operations.

In this paper, we focus on a type of GP characterised as the output of a convolution operation between a smoothing kernel and a latent GP. An example of this construction is the Latent Force Model (LFM) (Alvarez et al., 2009), in which the smoothing kernel corresponds to the Green's function associated with an Ordinary Differential Equation (ODE). By incorporating RKHS Fourier features into this framework, we derive adaptive global features inspired by the ODE, allowing for the optimisation of amplitudes and phases. We name the obtained features *Variational Fourier Response Features* (VFRFs) since they are derived from the output of a linear system. To enhance the capability of our model, we further use these adaptive features in a compositional GP model that stacks multiple LFMs, also known as Deep LFM (DLFM) (McDonald and Álvarez, 2021). This hierarchical structure facilitates more precise and robust modelling of complex, non-stationary data. Our experimental results on both synthetic and real-world data demonstrate that incorporating these ODE-inspired RKHS Fourier features improves upon the standard practice of using VFFs.

2 BACKGROUND

This section reviews concepts and preliminaries relevant to this work and establishes the notation used throughout the subsequent discussions.

2.1 Sparse Variational Gaussian Process

A GP $f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot'))$ places probability measures on a function space $\{f : \mathbb{R}^D \to \mathbb{R}\}$ (Rasmussen and Williams, 2006). Its behaviour is characterised by the mean function $m : \mathbb{R}^D \to \mathbb{R}$ and the covariance function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The evaluation of the function $f(\cdot)$ at an input of interest $\mathbf{x}$ is a random variable denoted as $f(\mathbf{x}) \in \mathbb{R}$. Given a dataset of inputs $\mathbf{X} = [\mathbf{x}_n]_{n=1}^N \in \mathbb{R}^{N \times D}$ and the corresponding measurements $\mathbf{y} = [y_n]_{n=1}^N \in \mathbb{R}^N$, we assume $\mathbf{y}$ is observed from a noise-corrupted GP: $y_n = f(\mathbf{x}_n) + \epsilon, \epsilon \sim \mathcal{N}(\epsilon \mid 0, \varepsilon^2)$, where ε^2 is the noise variance. The exact inference for the posterior distribution $p(f \mid \mathbf{y})$ suffers from $\mathcal{O}(N^3)$ time complexity and is limited to Gaussian likelihoods.

Sparse Variational Gaussian Processes (SVGPs) (Titsias, 2009; Hensman et al., 2013, 2015) provide a scalable inference framework by introducing a small set of $M(\ll N)$ inducing points $\mathbf{Z} = [\mathbf{z}_m]_m^M \in \mathbb{R}^{M \times D}$ and the corresponding inducing variables $\mathbf{u} = [u(\mathbf{z}_m)]_{m=1}^M \in \mathbb{R}^M$ from the GP prior, i.e., $p(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{0}, \mathbf{K}_{\mathbf{ZZ}})$. A variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}, \mathbf{S})$ is employed to approximate the posterior process $q(f(\mathbf{x})) = \int p(f(\mathbf{x}) \mid \mathbf{u}) q(\mathbf{u}) \, \mathrm{d}\mathbf{u} = \mathcal{N}(f \mid \tilde{m}, \tilde{\Sigma})$, where

$$\begin{aligned} \tilde{\mu}(\mathbf{x}) &= m(\mathbf{x}) + k_{\mathbf{xZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{m}, \\ \tilde{\Sigma}(\mathbf{x}, \mathbf{x}') &= k_{\mathbf{xx}'} + \mathbf{k}_{\mathbf{xZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} (\mathbf{S} - \mathbf{K}_{\mathbf{ZZ}}) \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{k}_{\mathbf{Zx}'}. \end{aligned} \tag{1}$$

SVGPs learn the optimal placement of the inducing points and the variational distribution by maximising an Evidence Lower BOund (ELBO) of $\log p(\mathbf{y} \mid \mathbf{X})$.

2.2 Variational Fourier Features

Inter-domain GPs (Lázaro-Gredilla and Figueiras-Vidal, 2009b; Álvarez et al., 2010; Van der Wilk et al., 2020) extend the domain of inducing variables by integrating the GP f with a deterministic inducing function g:

$$u(\mathbf{z}) = \int_{\mathbb{R}^D} g(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad \mathbf{z} \in \mathbb{R}^{D'}. \tag{2}$$

This formulation allows for a redefinition of inducing variables $\mathbf{u} = u(\mathbf{Z})$, which still share the GP prior, albeit with alternative expressions of $\mathbf{k}_{\mathbf{xZ}}$ and $\mathbf{K}_{\mathbf{ZZ}}$ used in (1). By choosing various functions for g, inter-domain GPs facilitate the construction of vector-valued basis functions $k(\cdot, \mathbf{Z})$ for more informative feature extraction while maintaining the standard SVGP framework.

VFFs (Hensman et al., 2018) define each inter-domain inducing variable u_m of $\mathbf{u}$ by projecting the original GP f onto a Fourier basis: $u_m = \langle \phi_m, f \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the Matérn RKHS inner product on an interval $[a, b]$, and ϕ_m is the m-th entry of a truncated Fourier basis

$$\begin{aligned} \boldsymbol{\phi}(x) = &[1, \cos(z_1(x-a)), \cdots, \cos(z_M(x-a)), \\ &\sin(z_1(x-a)), \cdots, \sin(z_M(x-a))], \end{aligned} \tag{3}$$

with x a scalar input. In this setting, $\mathbf{z} = [z_m]_{m=1}^M$ are M *inducing frequencies*, analogous to inducing points in the SVGP context. The projection of f onto this basis results in sinusoidal terms in the cross-covariance, i.e., $\mathrm{Cov}[f(\cdot), u_m] = \phi_m(\cdot)$ due to the reproducing property of the RKHS.

2.3 Latent Force Model

An LFM (Alvarez et al., 2009) is a GP model integrating differential equations to model dynamic physical systems probabilistically. Based on prior physical knowledge, a single-output LFM assumes the system's output $f(t)$ is influenced by Q *latent forces* $\{u_q(t)\}_{q=1}^Q$

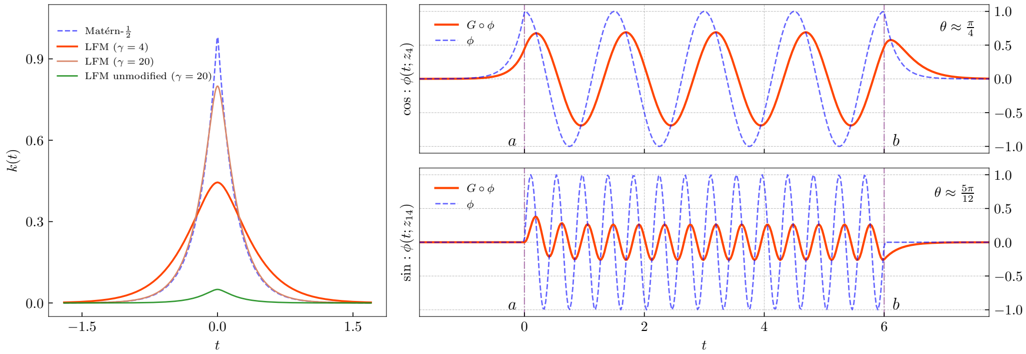


Figure 1: Covariance functions of LFMs (left) and Variational Fourier Response Features (VFRFs) (right). The latent force $u(t)$ uses a Matérn-$\frac{1}{2}$ kernel with length-scale $l = 0.2$ (left dashed). **Left**: The centred kernel of the input latent force (dashed) and the output process $f(t)$ of LFMs with different ODE parameters γ (solid). Unlike the LFM kernel induced by (4) (green), the modified LFM kernel from (5) can revert to the input Matérn-$\frac{1}{2}$ kernel if increasing γ (red to brown). **Right**: VFRFs ($G \circ \phi$, red solid) and VFFs (ϕ, blue dashed) with different inducing frequencies: $z_m = \frac{8\pi}{b-a}$ (upper) and $\frac{28\pi}{b-a}$ (lower). The upper panel depicts the cosine basis with a phase delay $\theta \approx \frac{\pi}{4}$ to the VFF, while the lower panel displays the sine basis with a phase delay $\theta \approx \frac{5\pi}{12}$.

through differential equations. Commonly, a first-order LFM uses the following form of ODE (Guarnizo and Álvarez, 2018)

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} + \gamma f(t) = \sum_{q=1}^{Q} S_q u_q(t), \tag{4}$$

where $\gamma > 0$ is a decay parameter, and $S_q \in \mathbb{R}$ is a sensitivity parameter. The solution for the output $f(t)$ takes the form of weighted convolution integrals $f(t) = \sum_{q=1}^{Q} S_q \int_0^t G(t-\tau)u_q(\tau)\,\mathrm{d}\tau$, where $G(\cdot)$ denotes the Green's function associated with the ODE.

Latent forces are presumed to follow GP priors, $u_q(t) \sim \mathcal{GP}(0, k_q(t,t'))$, leading to a covariance function for the outputs $k_f(t,t') = \sum_{q=1}^{Q} S_q^2 \iint G(t-\tau)k_q(\tau,\tau')G(t'-\tau')\,\mathrm{d}\tau\,\mathrm{d}\tau'$. For some types of covariance functions $k_q(t,t')$, e.g., the radial basis function (RBF), $k_f(t,t')$ can either be computed explicitly (Lawrence et al., 2006) or approximated by using convolved RFFs (Guarnizo and Álvarez, 2018; Rahimi and Recht, 2007). By plugging the physics-informed kernels into the GP posterior, LFMs embed domain-specific knowledge into the learning process and can utilise the sparse approximation techniques in GP inference.

3 METHODOLOGY

This section describes integrating RKHS Fourier features into compositional GPs, with a specific focus on LFMs within the SVGP framework. We start by adapting the conventional ODE used in LFMs to incorporate VFFs as a special instance of our model (Section 3.1). Details on VFRFs are provided in Section 3.2. We then extend our model from a single-layer to a hierarchical structure in Section 3.3.

3.1 LFMs with Modified ODEs

In this work, we focus on a dynamical system modelled by a potential first-order ODE without loss of generality

$$\beta\frac{\mathrm{d}f(t)}{\mathrm{d}t} + \alpha f(t) = u(t), \tag{5}$$

where α, β are positive coefficients and $u(t) \sim \mathcal{GP}(0, k(t,t'))$ represents an unknown latent force with a Matérn kernel with half-integer order. The Green's function of (5) is $G(t) = \frac{1}{\beta}\exp(-\frac{\alpha}{\beta}t) = \frac{1}{\beta}\exp(-\gamma t)$. We introduce $\gamma = \frac{\alpha}{\beta}$ to remain consistent with the decay parameter in (4). Unlike the conventional formulation (4), which involves a weighted sum of multiple latent forces, the modified ODE (5) simplifies it to a single process $u(t)$. It can be trivially decomposed into distinct latent forces if necessary. Moreover, we will further show that, by introducing coefficients α and β in our model, the output process f can revert to a GP with VFFs as $\beta \to 0^+$. This formulation enables practitioners to apply the proposed approach in scenarios where prior knowledge of the system is limited and there is no prior knowledge indicating if the dynamics encoded in the kernel accurately reflect the observed data.

A solution $f(t)$ can be expressed as a convolution integral

$$f(t) = \int_{-\infty}^{t} G(t-\tau)u(\tau)\,\mathrm{d}\tau = G \circ u, \tag{6}$$

where the integral's lower limit is extended to negative infinity to maintain variance near the origin, though it can be adjusted based on data range or prior knowledge in practice. The covariance function of the output process $f(\cdot)$ is derived by applying the convolution operator to the kernel k's arguments, respectively:

$$\begin{aligned}&\mathrm{Cov}[f(t), f(t')] = \\ &\int_{-\infty}^{t'}\int_{-\infty}^{t} G(t-\tau)k(\tau,\tau')G(t'-\tau')\,\mathrm{d}\tau\,\mathrm{d}\tau'.\end{aligned} \tag{7}$$

The covariance function (also called LFM kernel in this paper) can be calculated analytically if k is a Matérn kernel with half-integer orders. We give the closed-form covariances in Table 3 of Appendix C.

Model interpretation The Green's function $G(\cdot)$, determined by the system's dynamics, serves as a signal filter. It effectively acts as a low-pass filter described by the ODE (5), with γ representing the "cutoff frequency". Mathematically, the convolution operator $G(\cdot)$ of the modified ODE will behave like the Dirac delta function in (6) as $\alpha = 1, \beta \to 0^+$ (i.e.,$\gamma \to +\infty$), causing $f(t)$ to closely replicate $u(t)$. Fig. 1, left, illustrates this behaviour. We use the Matérn-$\frac{1}{2}$ kernel for the covariance $k(t,t')$ of the latent force. The figure shows this covariance function (dashed) and two LFMs covariance functions (solid) $k_f(t,t')$ with different γ values. The LFM kernel reverts to the latent force kernel as γ increases. However, the conventional LFM kernel without the ODE modification, i.e., (4) (Guarnizo and Álvarez, 2018) will get flattened since the corresponding Green's function $\exp(-\gamma t)$ does not effectively mimic a valid Dirac delta function.

3.2 Variational Fourier Response Features

Building upon the modified ODE described by (5), we introduce a spectral approximation for the LFMs within the inter-domain GP framework. The latent force u is initially projected onto the Fourier basis entries ϕ_m as defined in (3), yielding its spectral representations

$$v_m = \langle \phi_m, u\rangle_{\mathcal{H}}, \quad m = 0, 1, \ldots, 2M. \tag{8}$$

The projected inducing variables v_m are collected as $\mathbf{v} = [v_m]_{m=0}^{2M} \in \mathbb{R}^{2M+1}$.

By the closure of GPs under linear operations, the output f and the projection $\mathbf{v}$ share a joint augmented GP prior. The covariance matrix of inducing variables $\mathrm{Cov}[\mathbf{v},\mathbf{v}]$ has a low-rank-plus-diagonal structure if inducing frequencies $\mathbf{z} = [z_m]_{m=1}^{M}$ are harmonic on $[a,b]$, i.e., $z_m = \frac{2\pi m}{b-a}$, facilitating faster posterior computation (Hensman et al., 2018). For a given input t, the cross-covariance of the output process f and the inducing variable v_m is computed as

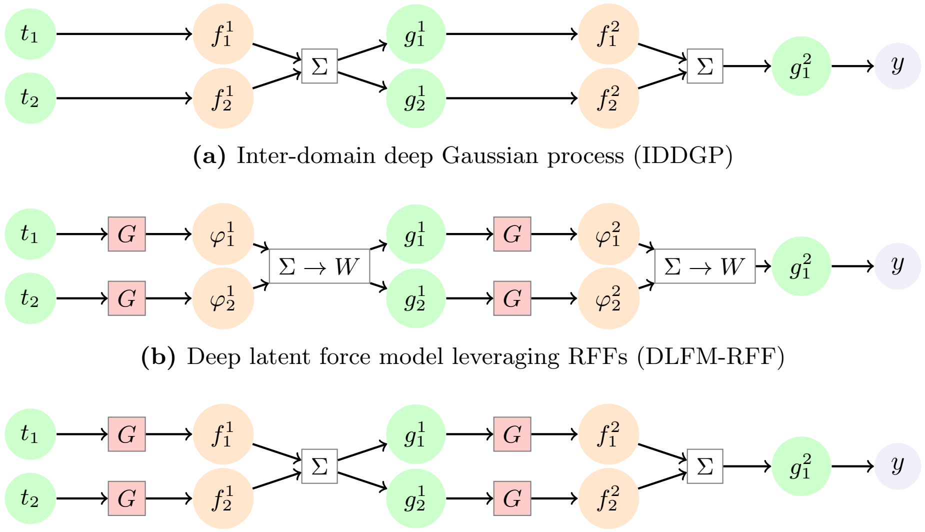
$$\mathrm{Cov}[f(t), v_m] = \int_{-\infty}^{t} G(t-\tau)\cdot\langle k(\tau,\cdot), \phi_m(\cdot)\rangle_{\mathcal{H}}\,\mathrm{d}\tau, \tag{9}$$

where we take advantage of the linearity and calculate the expectation over $u(\cdot)$. The reproducing property of Matérn RKHS ensures that the inner product $\langle k(\tau,\cdot), \phi_m(\cdot)\rangle_{\mathcal{H}}$ results in well-defined sinusoidal functions within the interval $\tau \in [a,b]$. Therefore, we can derive the RKHS Fourier features for LFMs on $t \in [a,b]$ as follows:

$$\begin{aligned}\mathrm{Cov}[f(t), v_m] &= \int_{-\infty}^{t} G(t-\tau)\phi_m(\tau)\,\mathrm{d}\tau \\ &= \begin{cases}\frac{\cos(z_i(t-a)+\theta)}{\beta\sqrt{z_i^2+\gamma^2}} + \xi_i & i = 0,\ldots,M, \\ \frac{\sin(z_i(t-a)+\theta)}{\beta\sqrt{z_i^2+\gamma^2}} + \xi_i & i = M+1,\ldots,2M,\end{cases}\end{aligned} \tag{10}$$

where $z_0 = 0$, cutoff frequency $\gamma = \frac{\alpha}{\beta}$, phase shift $\theta = -\arctan(\frac{z_i}{\gamma})$, and ξ_i represents an exponential decay term. Since the integration variable τ ranges from negative infinity and the inner product $\langle k(\tau,\cdot), \phi(\cdot,z)\rangle_{\mathcal{H}}$ has different expressions beyond $\tau \in [a,b]$, the covariance of f and v_m for $t \in \mathbb{R}$ emerges as a continuous piece-wise function (see Fig. 1 right). The detailed derivation and complete expressions of $\mathrm{Cov}[f(t), v(z)]$ for Matérn-$\frac{1}{2}/\frac{3}{2}/\frac{5}{2}$ are given in Appendix D and illustrated in Fig. 8 and 9.

The derived inter-domain features (10) reflect the filtering effect of the system, i.e., how the ODE adaptively transforms the frequency components of latent forces to the output through amplitude attenuation and phase shift. By analogy with the frequency response of linear systems, we name the obtained Fourier features from RKHS as *Variational Fourier Response Features* (VFRFs). Fig. 1, right, depicts the VFRFs of the LFM from the left subplot ($\gamma = 4$, solid red) and the VFFs of the corresponding latent force (dash blue). They show that the VFRFs are learnable basis functions that adjust both the amplitude and the phase according to the input frequencies and the ODE parameters. Particularly, the system will allow nearly all frequency components of the input process to pass through as $\gamma \to +\infty$. Under this condition, the VFRFs converge to the VFFs. We would like to emphasise here that the features derived from (9) can apply to more general inter-domain GPs with other linear transformations $G(\cdot)$, not just limited to LFMs. Moreover, for stable dynamical systems governed by higher-order ODEs within the LFM framework, the derivation of the VFRFs described above can be readily extended by using corresponding Green's functions.



(a) Inter-domain deep Gaussian process (IDDGP)

(b) Deep latent force model leveraging RFFs (DLFM-RFF)

(c) Our deep latent force model with VFRF (DLFM-VFRF)

Figure 2: A conceptual illustration of how our model (2c) differs from the IDDGP (2a) and the DLFM-RFF (2b). Compared to (2a), our model additionally applies convolution operators G from the ODEs to each input dimension: $f(t) = \int G(t-\tau)u(\tau)\,\mathrm{d}\tau$, where $G(\cdot)$ represents the Green's function and $u(\cdot)$ is a GP prior with Matérn kernels. Compared to (2b) using RFFs $\varphi(\cdot)$ for low-rank covariance matrix approximation and making inference over weights W, our model uses Fourier features derived from applying linear transformations to GPs and make inference in an inter-domain way. For a high-level comparison with other models, see Fig. 7.

3.3 Deep LFMs with VFRFs

DLFMs extend the concept of shallow LFMs by stacking them in a non-parametric cascade, similar to DGPs. This hierarchical structure allows DLFMs to model the non-stationarities present in nonlinear dynamical systems. In this section, we detail the construction of a hierarchical composition of L LFMs within the framework of variational DGPs, each governed by the modified ODE and enhanced with VFRFs for variational approximation of the posterior. Fig. 2 gives a conceptual illustration of how our proposed DLFM differs from a DGP. We leverage the layer-wise Monte Carlo technique in doubly stochastic variational inference (Salimbeni and Deisenroth, 2017) to allow functional samples to propagate through the compositional architecture efficiently.

The first layer of a DLFM processes a D^0-dimensional input $\mathbf{t} = [t_d]_{d=1}^{D^0}$ and outputs a D^1-dimensional independent process $\boldsymbol{g}^1(\mathbf{t}) = [g_r^1(\mathbf{t})]_{r=1}^{D^1}$ (the superscripts indicate the layer index). To extend the application of VFRFs to multidimensional inputs, we follow Hensman et al. (2018) to employ additive LFM kernels for each output dimension, i.e., each output dimension $g_r^1(\mathbf{t})$ is modelled as $g_r^1(\mathbf{t}) = \sum_{d=1}^{D^0} f_d^1(t_d)$, where $\{f_d^1\}_{d=1}^{D^0}$ are LFMs with ODE-induced covariance functions. In this work, we assume the LFMs f_d^1 are independent, but this assumption can be relaxed by allowing them to share the same latent forces, which can lead to more complex kernels for the outputs.

Following the construction of a single-layer LFM, each $g_r^1(\mathbf{t})$ is equipped with a set of M inducing frequencies $\mathbf{Z}^0 \in \mathbb{R}^{M \times D^0}$ and corresponding inducing variables $\mathbf{V}^1 = [v_{m,d}^1]_{m=1,d=1}^{2M+1,D^0}$. These variables are created by the RKHS projection $v_{m,d}^1 = \langle u_d^1, \phi_m \rangle_{\mathcal{H}}$. Therefore, the covariance functions necessary for sparse variational inference are given by

$$\begin{aligned}\mathrm{Cov}[g_r^1(\mathbf{t}), v_{m,d}^1] &= \int_{-\infty}^{t_d} G_d^1(t_d-\tau)\phi_m(\tau)\,\mathrm{d}\tau,\\ \mathrm{Cov}[v_{m,d}^1, v_{m',d'}^1] &= 0 \quad (d \neq d').\end{aligned}$$

Assuming a variational distribution $q(\mathbf{V}^1)$, the approximate posterior $q(\boldsymbol{g}^1 \mid \mathbf{t})$ of the first layer is derived by substituting $k(\mathbf{x}, \mathbf{Z})$ and $\mathbf{K_{ZZ}}$ in (1) with the expressions of VFRFs. Samples from this approximate posterior are drawn using the re-parameterisation trick (Kingma et al., 2015).

Given a training dataset with inputs $[\mathbf{t}_i]_{i=1}^N$ and targets $[\mathbf{y}_i]_{i=1}^N$, the training of DLFMs involves maximising the average ELBO over a mini-batch $\mathcal{B}$:

$$\begin{aligned}\mathrm{ELBO} = &\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}_{q(\mathbf{g}_i^L|\mathbf{t}_i)} \log\left[p(\mathbf{y}_i \mid \mathbf{g}_i^L)\right]\\ &- \frac{1}{N}\sum_{l=1}^{L} \mathrm{KL}\left[q(\mathbf{V}^l) \mid\mid p(\mathbf{V}^l)\right],\end{aligned}$$

where $\mathbf{V}^l$ are the inducing variables of the l-th layer, and the collection $\{\mathbf{g}^l\}_{l=1}^L$ denotes the output random variables at hidden layers. The output of each layer serves as the input for the subsequent layer, creating a chain of dependencies where the posterior of each layer is computed based on the propagated samples $\hat{\mathbf{g}}_i^l \sim q(\mathbf{g}_i^l \mid \hat{\mathbf{g}}_i^{l-1})$. The predictive distribution $q(\mathbf{y}_*) = \int p(\mathbf{y}_* \mid \mathbf{g}_*^L) q(\mathbf{g}_*^L)\, \mathrm{d}\mathbf{g}_*^L$ at test location $\mathbf{t}_*$ follows a similar layer-wise procedure, where $q(\mathbf{g}_*^L)$ is a Gaussian mixture of S hidden-layer samples: $q(\mathbf{g}_*^L) \approx \frac{1}{S}\sum_{s=1}^S q(\mathbf{g}_*^L \mid \hat{\mathbf{g}}_*^{(s)^{L-1}})$.

4 RELATED WORK

LFMs present a physically-inspired approach to combining data-driven modelling with differential equations (Alvarez et al., 2009). Álvarez et al. (2010) further proposed variational inducing functions to handle non-smooth latent processes within convolved GPs (Alvarez and Lawrence, 2011). Our model builds upon LFMs and DGPs (Salimbeni and Deisenroth, 2017). Recently, various approximate inference methods have been explored for DGP-based models, which are generally categorised into variational inference techniques (Salimbeni and Deisenroth, 2017; Salimbeni et al., 2019; Lindinger et al., 2020) and Monte Carlo approaches(Havasi et al., 2018).

As outlined in Section 1, RFFs (Rahimi and Recht, 2007) and VFFs (Hensman et al., 2018) have recently been incorporated into GP models. RFFs were used in shallow LFMs models to approximate covariance matrices (Guarnizo and Álvarez, 2018) and expanded to a deeper architecture (McDonald and Álvarez, 2021, 2023). VFFs were once integrated with harmonizable mixture kernels in shallow GP models (Shen et al., 2019). While related to these studies, our approach primarily uses features similar to VFFs within the scope of inter-domain GPs (Lázaro-Gredilla and Figueiras-Vidal, 2009a; Van der Wilk et al., 2020). Unlike RFF-based DGP models, which often modify the original covariance functions by introducing a fully parametric variational distribution over random frequencies, our model preserves the integrity of the original kernel forms and approximates the DGP posterior directly. Another closely related work is Inter-Domain DGPs (IDDGPs) (Rudner et al., 2020), which employ fixed VFFs without ODEs. We provide an illustrative plot of IDDGP in Fig. 2a. In contrast, our model extends the compositional inter-domain GPs by integrating ODEs to provide trainable, physics-informed RKHS Fourier features. Fig. 7 in Appendix A illustrates a high-level comparison of our work with related studies.

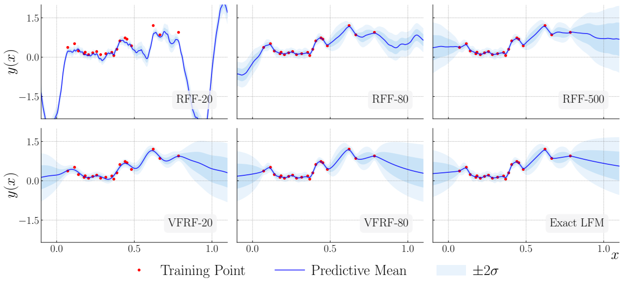


Figure 3: Illustrative example of Matérn-$\frac{1}{2}$ LFM posteriors with VFRFs / RFFs. The model's feature is indicated at the lower right. **Top row**: predictive posteriors of 20, 80, and 500 RFFs. **Bottom row**: predictive posteriors of 20 and 80 inducing frequencies and an exact LFM. Noisy observations are marked with red dots, posterior predictive means with blue lines, and uncertainty (one or two standard deviations) with varying shades of blue. In this example, VFRFs show a better approximation to the true posterior, whereas RFFs indicate variance underestimation with fewer features.

5 EXPERIMENTS

This section presents experiments designed to evaluate our model using VFRFs. We begin by examining the approximation quality of shallow LFMs with VFRFs and RFFs on synthetic data. We then evaluate our model on a highly non-stationary speech signal dataset and benchmark regression tasks, comparing it to various baselines in both cases.[1]

5.1 Synthetic Datasets

We first evaluate the shallow LFMs and DLFMs using the proposed VFRFs on two synthetic datasets, respectively.

5.1.1 Posterior Approximation for Shallow LFMs

VFRFs and RFFs both leverage Fourier features to facilitate approximate inference in LFMs. In Fig. 3, we compare the approximation quality of VFRFs and RFFs in a regression task using models with a Matérn-$\frac{1}{2}$ kernel. The kernel's parameters and the noise variance are initially optimised by maximising the marginal likelihood of an exact LFM and then fixed across all models. The frequencies of RFFs are sampled from the corresponding Cauchy distribution of the kernel (detailed in Appendix C.2).

Fig. 3 shows that the model using 20 VFRFs has

[1] Our code is publicly available in the repository: `https://github.com/shixinxing/LFM-VFF`

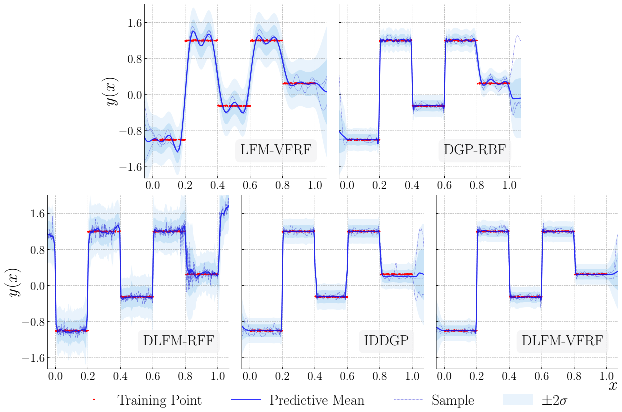


Figure 4: Posterior predictive distribution comparison of different models on data points from a noisy multi-step function. The models and the features used are noted at the bottom right of each subplot. The dashed lines are samples from the predictive distributions. The experiment uses two layers for deep models and Matérn-$\frac{3}{2}$ kernels except for the DGP (upper left) and DLFM-RFF (lower left) that use RBF kernels. All models are trained with 20 inducing points/Fourier features per layer. The DLFM models with VFRFs perform best among the models.

already fitted the data points reasonably well. Increasing VFRFs to 80 fills in the details of the region with more observations, and the approximate predictive posterior is quite close to the exact one. The same number of RFFs yields a poorly fitted predictive mean and tends to underestimate the variance in different regions of the input space, which is a phenomenon known as *variance starvation* (Wang et al., 2018). From the top row of Fig. 3, we can observe that achieving a comparable approximation requires more RFFs than VFRFs due to the heavy-tailed spectral density of the Matérn-$\frac{1}{2}$ kernels used in RFFs.

5.1.2 Multi-step Function for Deep Structures

To further evaluate our DLFM-VFRF's performance against other models, we conduct tests on a synthetic multi-step function (as shown in Fig. 4). This task is challenging for shallow GP models due to the need to capture global structures in highly non-stationary data (Rudner et al., 2020).

Although equipped with VFRFs, our single-layer LFM (upper left) struggles to fit the non-stationarity with a stationary kernel. In contrast, the models with compositional layers exhibit better performance. The DLFM-RFF (lower left) (McDonald and Álvarez, 2021) generates high-frequency, wiggly posterior predictive samples, resulting in an easily overfitting model struggling to seize the slow-changing trend in the data. DGP-RBF (upper left), IDDGP (lower middle), and our DLFM-VFRF (lower right) offer smoother samples from the posterior distributions. Due to the VFRFs' flexibility to capture the global data structure, our DLFM-VFRF outperforms both the DGP-RBF with local inducing points and the IDDGP, especially inside $[0.8, 1]$. Our model provides a more accurate predictive mean throughout the steps and at abrupt step transitions and demonstrates narrower confidence intervals, indicating a better uncertainty calibration. In the plot, GP models based on function-space inference tend to revert to prior distributions outside the observed data range, displaying wide uncertainty bands. In contrast, the DLFM-RFF yields relatively more confident non-zero predictions in these areas.

To measure the performance quantitatively, we conducted additional experiments to train five independent copies of the IDDGP and our model. We summarise the Root Mean Square Error (RMSE) and Negative Marginal Log-Likelihood (NMLL) on the test points in the following Table 1. Additionally, the outputs of the intermediate layers for DGP, IDDGP, and our model are shown in Fig. 10 in Appendix E.

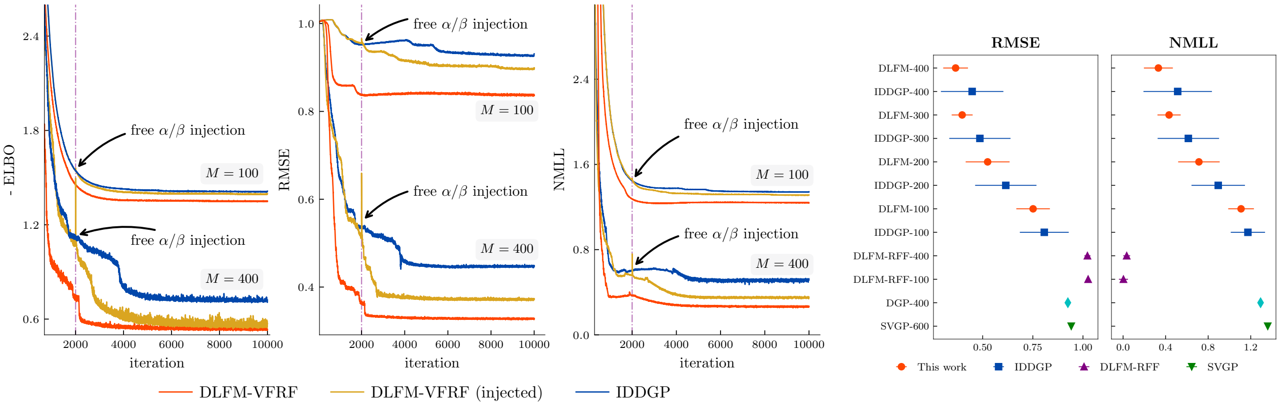
Table 1: Performance of IDDGP and our DLFM-VFRF on fitting the multi-step function over five runs (lower is better).

Model	RMSE	NMLL
IDDGP	0.107 ± 0.015	-1.051 ± 0.131
Ours	**0.095 ± 0.010**	**-1.304 ± 0.120**

5.2 TIMIT Speech Signals

We apply our model to a regression task on the TIMIT dataset, a speech recognition resource previously used by Rudner et al. (2020), to explore the capability of GP-based models in handling complex, non-stationary data. The dataset features rapidly changing audio waves, posing significant challenges for shallow GP models reliant on local approximation. Initially, we apply a moving average filter to smooth the audio waves and select the first 10,000 data points for our analysis, reserving 30% as test data. Our method uses the Matérn-$\frac{3}{2}$ kernels.

One of our goals is to evaluate how the performance of IDDGPs and DLFMs varies with the number of global Fourier features and the effect of the ODE parameters α and β on the learning process. Fig. 5a illustrates the progressions of performance metrics, e.g., test RMSE and NMLL using 100 and 400 inducing



(a) Learning progression of DLFM-VFRF and IDDGP

(b) Standardised RMSE and NMLL

Figure 5: **(a)** Learning progression of DLFMs and IDDGPs with M inducing frequencies on the TIMIT dataset, presented in negative ELBO, test average RMSE and NMLL. The DLFM in yellow maintains fixed $\beta = 10^{-6}$ throughout the first 2000 training iterations, after which α/β are allowed to vary. The DLFMs in red employ trainable ODE parameters from the start. The DLFM-VFRFs consistently outperform the IDDGPs; **(b)** Mean standardised RMSE and NMLL with the standard deviations (over 10 random seeds) for models employing varying numbers of inducing frequencies. The numbers following the hyphen in the y-axis labels indicate the number of inducing frequencies/points. A lower value (to the left) indicates better performance.

frequencies. The yellow lines of DLFMs align closely with the IDDGP's blue line during the first 2000 iterations, where DLFMs maintain a fixed small β value ($\beta = 10^{-6}, \alpha = 1$). We use this setting to illustrate how our DLFM-VFRF can replicate the original IDDGPs as expected when $\beta \to 0^+$. After 2000 iterations, we allow optimisation of α and β, leading to subsequent improvements in the testing metrics, and suggesting continuous learning with ODE-based Fourier features. Additionally, optimising all parameters from the beginning (red lines) yields the best results across various setups. Fig. 5b compares the performance of different models with all parameters optimised from the beginning. It is unsurprising that increasing the number of inducing frequencies typically results in better performance. The results reveal that while RFF-based DLFMs exhibit the lowest NMLL, they show the highest RMSE, reflecting a lack of precision on test data points. DLFMs equipped with VFRFs consistently surpass both the DGP with local inducing points and the IDDGPs in terms of both RMSE and NMLL, highlighting our model's enhanced ability to accurately capture the global structure and non-stationarity of the data.

Running Time Comparison Theoretically, the extra running time of our model compared to IDDGPs mainly lies in the more complex forward computation on covariance entries and the backward gradient update on the extra ODE parameters. We record the wall-clock running time (per iteration) of models with the number of inducing frequencies ranging from 100 to 400 in Table 2 below. The results are averaged over five runs, and we exclude the standard deviations as they are quite small.

Table 2: Wall-clock training time of IDDGPs and our DLFM-VFRFs with different numbers of inducing frequencies M.

Model/M	100	200	300	400
IDDGP-M	0.279s	0.633s	1.138s	1.891s
Ours-M	0.313s	0.686s	1.207s	1.977s

From Table 2, we observe that our model incurs slightly higher runtime overhead compared to IDDGP. Despite this, the runtime difference compared with IDDGP with VFFs remains acceptable even with 400 inducing features, especially given the improved flexibility and modelling capacity of our approach. The runtime gap might be reduced with some computational optimization techniques (e.g., JIT) implemented.

5.3 UCI Regression Benchmarks

To demonstrate the versatility and effectiveness of our model on domain-agnostic real-world data, we conduct evaluations on five diverse regression datasets from the UCI Machine Learning Repository (Dua and Graff, 2019). These datasets vary in size and feature dimensionality, allowing us to test the model's adaptability

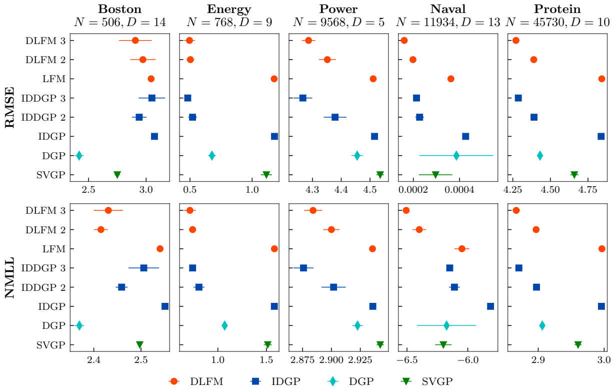


Figure 6: Regression test RMSE and NMLL results on UCI datasets, averaged over 10 random seeds. Lower values (to the left) indicate better performance. Model names include the number of layers.

across different scenarios (see Fig. 6). Consistent with standard practice (Salimbeni and Deisenroth, 2017), our regression tasks involve multivariate inputs and a univariate target. We reserve 10% of each dataset for testing, normalise the inputs to the range $[0, 3]$, and standardise the outputs based on the mean and standard deviation of the training set (these transformations are then reversed for evaluation). Following the setups in Rudner et al. (2020) and McDonald and Álvarez (2021), all models run with Matérn-$\frac{3}{2}$ kernels and employ 20 inducing points or frequencies. We employ three output dimensions per layer. We maintained the same experimental settings and initialisation across all tests. The figure illustrates that our models achieve comparable performance to the baselines. Notably, our models with two layers outperform IDDGP counterparts on the Energy, Power, and Naval datasets. We also observe that increasing the number of layers generally enhances the model's representational capacity, resulting in improved performance.

6 CONCLUSION

In this work, we adapt VFFs to the latent force framework, which inherently involves convolution operators with Green's functions. This adaptation introduces flexibility in modelling dynamics while preserving computational traceability. By introducing trainable parameters in the Green's function, we provide a mechanism for dynamically adjusting the inter-domain features. We further employ the inter-domain Fourier features in hierarchical LFMs. Our empirical evaluations across various datasets demonstrate that our model extends inter-domain GPs with RKHS Fourier features and has enhanced their modelling capacity for non-stationary and global structures.

Limitations and Future Work The current experiments are only based on models from first-order ODEs. Besides, computing the piece-wise VFRFs at intermediate layers may result in extra computational costs. Future work will focus on developing a normalization method at intermediate layers to accelerate inference and on extending our model's use to other challenging machine-learning tasks requiring the integration of specific domain knowledge of higher-order ODEs. Extending DLFMs to incorporate other recently proposed Fourier features, such as those in Cheema and Rasmussen (2024), represents a promising direction.

Acknowledgements

We are grateful for the computational resources from UoM CSF3, UK, and appreciate the insightful discussions with Xiaoyu Jiang. X.S. was supported by the UoM-CSC Joint Scholarship.

References

Alvarez, M., Luengo, D., and Lawrence, N. D. (2009). Latent force models. In *Artificial Intelligence and Statistics*, pages 9–16. PMLR.

Álvarez, M., Luengo, D., Titsias, M., and Lawrence, N. D. (2010). Efficient multioutput gaussian processes through variational inducing kernels. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 25–32. JMLR Workshop and Conference Proceedings.

Alvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research*, 12:1459–1500.

Cheema, T. M. and Rasmussen, C. E. (2024). Integrated variational fourier features for fast spatial modelling with gaussian processes. *Transactions on Machine Learning Research*.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. (2017). Random feature expansions for deep gaussian processes. In *International Conference on Machine Learning*, pages 884–893. PMLR.

Damianou, A. and Lawrence, N. D. (2013). Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215. PMLR.

Dua, D. and Graff, C. (2019). UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Sciences.

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. (2016). Detecting periodicities with gaussian processes. *PeerJ Computer Science*, 2:e50.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. (2014). Avoiding pathologies in very deep networks. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 202–210, Reykjavik, Iceland. PMLR.

Gal, Y. and Turner, R. (2015). Improving the gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *International Conference on Machine Learning*, pages 655–664. PMLR.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*.

Guarnizo, C. and Álvarez, M. A. (2018). Fast kernel approximations for latent force models and convolved multiple-output gaussian processes. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 835–844. Association For Uncertainty in Artificial Intelligence (AUAI).

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. *Advances in Neural Information Processing Systems*, 31.

Hensman, J., Durrande, N., and Solin, A. (2018). Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA. AUAI Press.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 28.

Lawrence, N., Sanguinetti, G., and Rattray, M. (2006). Modelling transcriptional regulation using gaussian processes. *Advances in Neural Information Processing Systems*, 19.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. (2009a). Inter-domain gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. (2009b). Inter-domain gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.

Lindinger, J., Reeb, D., Lippert, C., and Rakitsch, B. (2020). Beyond the mean-field: Structured deep gaussian processes improve the predictive uncertainties. *Advances in Neural Information Processing Systems*, 33:8498–8509.

McDonald, T. and Álvarez, M. (2021). Compositional modeling of nonlinear dynamical systems with ode-based random features. *Advances in Neural Information Processing Systems*, 34:13809–13819.

McDonald, T. and Álvarez, M. (2023). Deep latent force models: Ode-based process convolutions for bayesian deep learning. *arXiv preprint arXiv:2311.14828*.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550.

Rudin, W. (2017). *Fourier analysis on groups*. Courier Dover Publications.

Rudner, T. G., Sejdinovic, D., and Gal, Y. (2020). Inter-domain deep gaussian processes. In *International Conference on Machine Learning*, pages 8286–8294. PMLR.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep gaussian processes. *Advances in Neural Information Processing Systems*, 30.

Salimbeni, H., Dutordoir, V., Hensman, J., and Deisenroth, M. (2019). Deep gaussian processes with importance-weighted variational inference. In *International Conference on Machine Learning*, pages 5589–5598. PMLR.

Shen, Z., Heinonen, M., and Kaski, S. (2019). Harmonizable mixture kernels with variational fourier features. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3273–3282. PMLR.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V., and Hensman, J. (2020). A framework for interdomain and multioutput gaussian processes. *arXiv preprint arXiv:2003.01115*.

Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. (2018). Batched large-scale bayesian optimization in high-dimensional spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 745–754. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Model Comparison

We give a high-level comparison of our approach with other related models in Fig. 7. Each point represents a corresponding model. The reference is attached to the blue tag. The comparison dimensions include whether the model has a multi-layer structure, whether it incorporates physics-informed modelling involving ODEs and convolutions, and which kind of feature it uses. The red points represent our work.

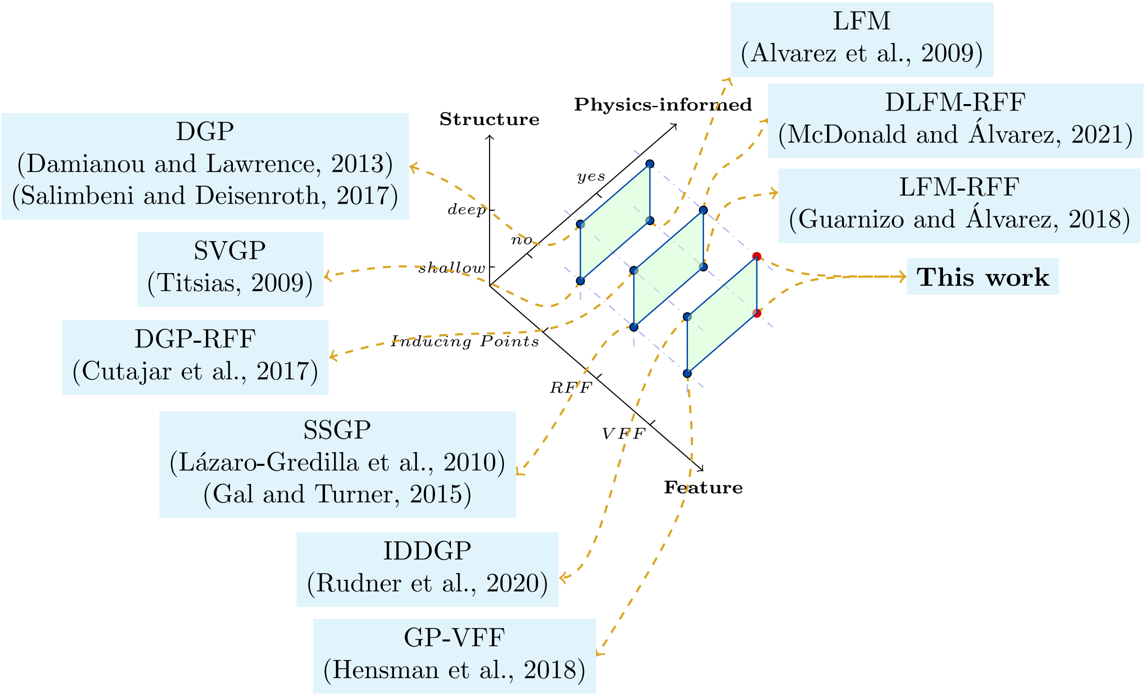


Figure 7: An illustration on a comparison of our model with recent related work. The comparison dimensions are the used feature, the structure depth, and whether incorporating physical dynamics.

B Variational Fourier Feature

Variational Fourier Features (VFFs, (Hensman et al., 2018)) are designed on a Matérn Reproducing Kernel Hilbert Space (RKHS). Specifically, Matérn-$\frac{1}{2}/\frac{3}{2}/\frac{5}{2}$ kernels with inputs $x, x' \in \mathbb{R}$ are of particular interest:

$$k_{1/2}(r) = \sigma^2 e^{-\lambda r},\ \lambda = \frac{1}{l}, \tag{11}$$

$$k_{3/2}(r) = \sigma^2 (1+\lambda r) e^{-\lambda r},\ \lambda = \frac{\sqrt{3}}{l}, \tag{12}$$

$$k_{5/2}(r) = \sigma^2 (1+\lambda r + \frac{1}{3}\lambda^2 r^2) e^{-\lambda r},\ \lambda = \frac{\sqrt{5}}{l}, \tag{13}$$

where $r = |x - x'|$, σ^2 is the kernel's output-scale (or variance), and l is the length-scale. We reiterate the closed-form RKHS inner products for Matérn-$\frac{1}{2}$ on $[a, b]$ here (for other Matérn kernels and more details see Durrande et al. (2016)) :

$$\langle g, h \rangle_{\mathcal{H}_{\frac{1}{2}}} = \frac{1}{2\lambda\sigma^2} \int_a^b (\lambda g(x) + g'(x))(\lambda h(x) + h'(x))\, \mathrm{d}x + \frac{1}{\sigma^2} g(a)h(a). \tag{14}$$

The explicit expressions of the Matérn RKHS not only allow us to verify the reproducing property when $t \in [a, b]$

$$\langle k(t, \cdot), h(\cdot) \rangle_{\mathcal{H}} = h(t),\ \forall h \in \mathcal{H}, t \in [a, b], \tag{15}$$

but also make it feasible to complete the cross-covariance if t is outside $[a,b]$. In Appx. D, we further utilise the conclusion of VFFs in Hensman et al. (2018) for $t \in \mathbb{R}$ to calculate our *Variational Fourier Response Features* (VFRFs).

Given a Matérn GP $f(t) \sim \mathcal{GP}(0, k(t,t'))$, the explicit RKHS inner product provides an alternative linear operator to construct an inter-domain GP by $u_m = \langle f, \phi_m \rangle_{\mathcal{H}}$, where $\phi_m, m = 0, \dots, 2M$ is from a set of truncated Fourier basis with harmonic *inducing frequencies*

$$\phi_0(\cdot) = 1,\ \phi_m(\cdot) = \cos(z_m(\cdot - a)),\ \phi_{M+m}(\cdot) = \sin(z_m(\cdot - a)),\ z_m = \frac{2\pi m}{b-a}. \tag{16}$$

$u(\cdot)$ is an *inter-domain* GP sharing a joint Gaussian prior with $f(\cdot)$. For $t \in [a,b]$, the covariances are

$$\mathrm{Cov}[f(t), u_m] = \langle k(t,\cdot), \phi_m(\cdot) \rangle_{\mathcal{H}} = \phi_m(t), \quad \mathrm{Cov}[u_i, u_j] = \langle \phi_i(\cdot), \phi_j(\cdot) \rangle_{\mathcal{H}}. \tag{17}$$

The VFFs approximate the posterior by replacing the covariance matrix in Sparse Variational GPs (SVGPs) appropriately.

C LFMs for First-Order Dynamical System

We recall in this work a dynamical system modelled by a first-order ODE

$$\beta \frac{\mathrm{d}f(t)}{\mathrm{d}t} + \alpha f(t) = u(t),\ u(t) \sim \mathcal{GP}(0, k(t,t')),\ \alpha, \beta > 0, \tag{18}$$

where u is an unobserved latent force with a Matérn kernel. The Green's function is $G(t) = \frac{1}{\beta} \exp(-\gamma t), \gamma = \frac{\alpha}{\beta}$. We take the solution

$$f(t) = \int_{-\infty}^{t} G(t-\tau) u(\tau)\, \mathrm{d}\tau = G \circ u, \tag{19}$$

where a convolutional operator G acting on u is represented as $f = G \circ u$. Conventional LFMs establish a GP over $f \sim \mathcal{GP}(0, G \circ k \circ G)$, where $k \circ G$ signifies G operating the second argument of the kernel. The LFM kernels are computed analytically in Lawrence et al. (2006); Alvarez et al. (2009), but their expressions are based on RBF kernels. In the subsequent part, we will present the closed-form covariance expressions $G \circ k \circ G$ for the Matérn-$\frac{1}{2}/\frac{3}{2}/\frac{5}{2}$ kernels, respectively. All analytical LFM covariance functions discussed in this work are summarised in Table 3 and illustrated in the left panel of Fig. 1, 8 and 9. Furthermore, we introduce the approximation of our LFM kernels using random Fourier features in Appx. C.2.

C.1 Analytical LFM Matérn Kernels

The LFM kernel of a Matérn-$\frac{1}{2}$ latent force when $t > t'$ is given by

$$\begin{aligned} G \circ k \circ G &= \int_{-\infty}^{t} \int_{-\infty}^{t'} \frac{1}{\beta} e^{-\gamma(t-\tau)} \cdot \sigma^2 e^{-\lambda|\tau - \tau'|} \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')}\, \mathrm{d}\tau\, \mathrm{d}\tau' \\ &= \frac{\sigma^2}{\beta^2} \int_{-\infty}^{t'} \int_{\tau'}^{t} e^{-\gamma(t-\tau) - \lambda(\tau - \tau') - \gamma(t'-\tau')}\, \mathrm{d}\tau\, \mathrm{d}\tau' \\ &\quad + \frac{\sigma^2}{\beta^2} \int_{-\infty}^{t'} \int_{-\infty}^{\tau'} e^{-\gamma(t-\tau) - \lambda(\tau' - \tau) - \gamma(t'-\tau')}\, \mathrm{d}\tau\, \mathrm{d}\tau' \\ &= \frac{\sigma^2}{\beta^2 \gamma (\gamma^2 - \lambda^2)} \left[\gamma e^{-\lambda(t-t')} - \lambda e^{-\gamma(t-t')} \right]. \end{aligned} \tag{20}$$

The derivation for $t < t'$ is similar. As a result, we obtain a stationary LFM kernel for $\forall t, t' \in \mathbb{R}$,

$$G \circ k \circ G = \begin{cases} \frac{\sigma^2}{\beta^2 \gamma (\gamma^2 - \lambda^2)} \left[\gamma e^{-\lambda|t-t'|} - \lambda e^{-\gamma|t-t'|} \right] & \text{if } \gamma \neq \lambda, \\ \frac{\sigma^2 (1 + \lambda|t-t'|)}{2\beta^2 \lambda^2} e^{-\lambda|t-t'|} & \text{if } \gamma = \lambda. \end{cases} \tag{21}$$

Likewise, the expressions of the other LFM kernels are present in Table 3. The expressions exhibit continuity but non-differentiability at the point where $\gamma = \lambda$.

Table 3: LFM kernels of Matérn-$\frac{1}{2}/\frac{3}{2}/\frac{5}{2}$ latent forces

Latent Force $k(r)$	LFM $G \circ k \circ G$ [1]
Matérn-$\frac{1}{2}$	$\frac{\sigma^2}{\beta^2\gamma(\gamma^2-\lambda^2)}\left[\gamma e^{-\lambda r} - \lambda e^{-\gamma r}\right]$
Matérn-$\frac{3}{2}$	$\frac{\sigma^2}{\beta^2}\left[\frac{\lambda r+1}{\gamma^2-\lambda^2} - \frac{2\lambda^2}{(\gamma^2-\lambda^2)^2}\right] e^{-\lambda r} + \frac{2\lambda^3\sigma^2}{\beta^2\gamma(\gamma^2-\lambda^2)^2} e^{-\gamma r}$
Matérn-$\frac{5}{2}$	$\frac{\sigma^2}{3\beta^2}\left[\frac{\lambda^2 r^2+3}{\gamma^2-\lambda^2} + \frac{\lambda(3\gamma^2-7\lambda^2)r}{(\gamma^2-\lambda^2)^2} + \frac{4\lambda^2(3\lambda^2-\gamma^2)}{(\gamma^2-\lambda^2)^3}\right] e^{-\lambda r} - \frac{8\lambda^5\sigma^2}{3\beta^2\gamma(\gamma^2-\lambda^2)^3} e^{-\gamma r}$

[1] r is the input distance $r = |t - t'|$. $\{\beta, \gamma\}$ and $\{\sigma^2, \lambda\}$ are parameters from the ODE and the Matérn kernel, respectively.

C.2 Random Fourier Approximation of Kernels

Stationary kernels can be approximated using random Fourier features (Rahimi and Recht, 2007) using Bochner's theorem (Rudin, 2017). In the illustrative experiment, we approximate the Matérn LFM covariance proposed in our work using random Fourier features (also termed *Random Fourier Response Features* (RFRFs) by (Guarnizo and Álvarez, 2018)). The random Fourier features of the LFM with a Matérn-ν ($\nu = \frac{1}{2}/\frac{3}{2}/\frac{5}{2}$) kernel is given by

$$\varphi(t;\omega) = \int_{-\infty}^{t} e^{j\omega\tau} \cdot \frac{1}{\beta} e^{-\gamma(t-\tau)}\, \mathrm{d}\tau = \frac{e^{j\omega t}}{\beta(\gamma + j\omega)},\ \omega = \frac{\omega'}{l},\ \omega' \sim t_{2\nu}(\omega'), \tag{22}$$

where $j^2 = -1$, $t_{2\nu}$ is a zero-mean Student's t-distribution with 2ν degrees of freedom, and l is the length-scale of the Matérn kernel. Therefore,

$$G \circ k \approx \frac{\sigma^2}{M} \sum_{m=1}^{M} \varphi(t;\omega_m) \cdot e^{-j\omega_m t'}, \tag{23}$$

$$G \circ k \circ G \approx \frac{\sigma^2}{M} \sum_{m=1}^{M} \varphi(t;\omega_m) \cdot \bar{\varphi}(t';\omega_m), \tag{24}$$

where σ^2 is Matérn kernel's variance, $\{\omega_m\}_{m=1}^{M}$ are M random Fourier frequencies sampled from the corresponding Student's t-distribution. $\bar{\varphi}$ denotes the complex conjugate of φ.

D Variational Fourier Response Features for LFMs

We represent the projection of the latent force u onto the truncated Fourier basis ϕ as

$$v(z) = P \circ u = \langle \phi(\cdot; z), u(\cdot) \rangle_{\mathcal{H}}, \tag{25}$$

where $v(z)$ is the projection process for the latent force, and z is the inducing frequency. For simplicity, the projection operator is denoted as P. Consequently, the output process f and the projection process v share a joint GP prior:

$$\begin{bmatrix} f \\ v \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} G \circ k \circ G & G \circ k \circ P \\ P \circ k \circ G & P \circ k \circ P \end{bmatrix}\right), \tag{26}$$

where the covariance terms are given by

$$\begin{aligned} \mathrm{Cov}[f(t), v(z)] &= \mathbb{E}\left[f(t)\langle u(\cdot), \phi(\cdot; z)\rangle_{\mathcal{H}}\right] = \langle \mathbb{E}[f(t)u(\cdot)], \phi(\cdot; z)\rangle_{\mathcal{H}} \\ &= \langle G \circ k(t, \cdot), \phi(\cdot; z)\rangle_{\mathcal{H}} = G \circ \langle k(t, \cdot), \phi(\cdot; z)\rangle_{\mathcal{H}} = G \circ k \circ P. \end{aligned} \tag{27}$$

$$\mathrm{Cov}[v(z_i), v(z_j)] = \langle \phi(\cdot; z_i), \phi(\cdot; z_j)\rangle_{\mathcal{H}} = P \circ k \circ P. \tag{28}$$

The rest of this section will specify the closed-form VFRF expressions of $P \circ k \circ G$. With the integration lower limit going to negative infinity, the input values outside the interval $[a, b]$ should be considered.

We detail the derivation of Matérn-$\frac{1}{2}$ (listed in Table 5 and 6) and directly give the results of Matérn-$\frac{3}{2}$ (see Table 8 and 9) and Matérn-$\frac{5}{2}$ (see Table 11 and 12). These expressions of VFRFs will revert to VFFs under certain conditions. The LFM kernel and the VFRFs for Matérn-$\frac{3}{2}$ and Matérn-$\frac{5}{2}$ with the same hyperparameters in the main text are depicted in Fig. 8 and Fig. 9, respectively.

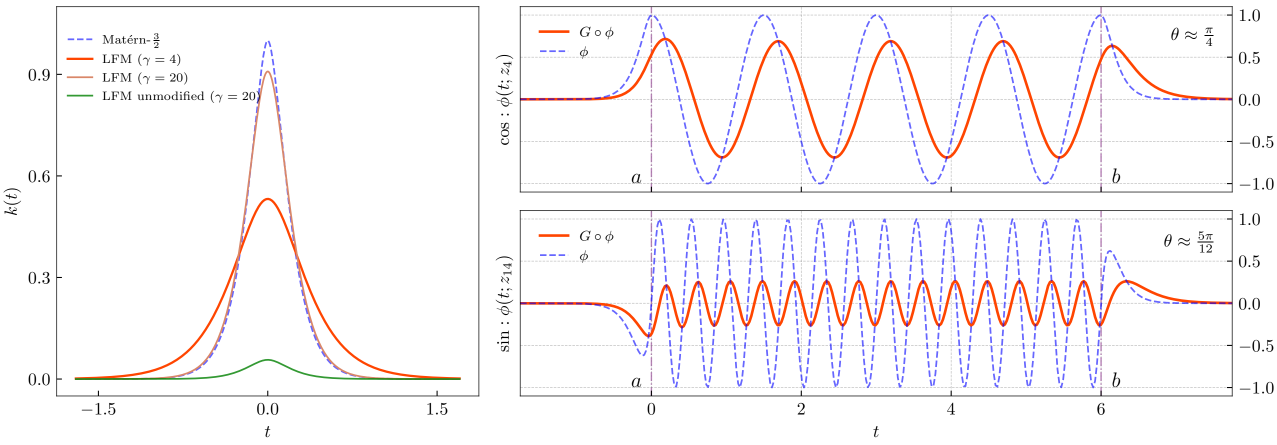


Figure 8: Covariance functions and VFRFs for LFMs with Matérn-$\frac{3}{2}$ kernel.

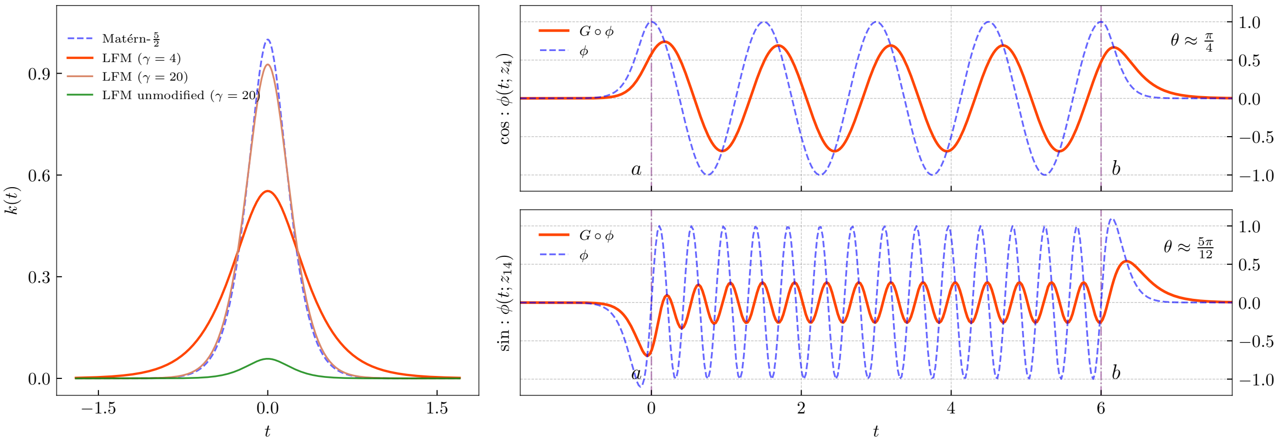


Figure 9: Covariance functions and VFRFs for LFMs with Matérn-$\frac{5}{2}$ kernel.

D.1 Matérn-$\frac{1}{2}$ Cross-covariance

We write the cross-covariance as

$$P \circ k \circ G = \langle \phi(\cdot; z), k(\cdot, \tau') \rangle_{\mathcal{H}} \circ G = \int_{-\infty}^{t'} h(z, \tau') \cdot G(t' - \tau') \, \mathrm{d}\tau'. \tag{29}$$

The location of τ determines the expression of $h(z, \tau') = \langle \phi(\cdot; z), k(\cdot, \tau') \rangle_{\mathcal{H}}$ (i.e., $P \circ k$) (Hensman et al., 2018), as summarized in the subsequent table (For Matérn-$\frac{3}{2}$/$\frac{5}{2}$, see Table 7/10). The columns of the tables indicate the VFFs for input τ' located inside/outside $[a, b]$. The cross-covariance can be derived by substituting $h(z, \tau')$ in (29). Table 5 and Table 6 collect the VFRFs for Matérn-$\frac{1}{2}$ LFMs with cosine/sine projection basis functions at different locations of t' in (29).

Table 4: VFFs $h(z, \tau')$ with Matérn-$\frac{1}{2}$ kernel

$\phi(\cdot; z)$	$\tau' < a$	$a \le \tau' \le b$	$\tau' > b$
$\cos(z(\cdot - a))$	$e^{-\lambda(a-\tau')}$	$\cos(z(\tau' - a))$	$e^{-\lambda(\tau'-b)}$
$\sin(z(\cdot - a))$	0	$\sin(z(\tau' - a))$	0

D.1.1 Cosine Features ($\phi(t; z) = \cos(z(t-a))$)

Case 1: $t' < a$,

$$P \circ k \circ G = \int_{-\infty}^{t'} e^{-\lambda(a-\tau')} \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')} \, \mathrm{d}\tau' = \frac{1}{\beta(\gamma+\lambda)} e^{-\lambda(a-t')}. \tag{30}$$

The covariance will converge to the VFF in Table 4 with a scaling coefficient α:

$$\lim_{\beta \to 0^+} P \circ k \circ G = \frac{1}{\alpha} e^{-\lambda(a-t')}. \tag{31}$$

Case 2: $a \leq t' \leq b$,

$$\begin{aligned} P \circ k \circ G &= \int_{-\infty}^{a} e^{-\lambda(a-\tau')} \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')} \, \mathrm{d}\tau' + \int_{a}^{t'} \cos(z(\tau'-a)) \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')} \, \mathrm{d}\tau' \\ &= \frac{\gamma \cos(z(t'-a)) + z \sin(z(t'-a))}{\beta(z^2+\gamma^2)} + \frac{(z^2-\gamma\lambda)}{\beta(\gamma+\lambda)(z^2+\gamma^2)} e^{-\gamma(t'-a)} \\ &= \frac{\cos(z(t'-a)+\theta)}{\beta\sqrt{z^2+\gamma^2}} + \xi_{\cos}, \end{aligned} \tag{32}$$

where $\theta = -\arctan(\frac{z}{\gamma})$, and ξ is a decay term. Particularly, the cross-covariance extends the VFFs of the latent force since

$$\lim_{\beta \to 0_+} P \circ k \circ G = \frac{1}{\alpha} \cos(z(t'-a)), \tag{33}$$

Case 3: $t' > b$,

$$P \circ k \circ G = \frac{e^{-\lambda(t'-b)}}{\beta(\gamma-\lambda)} + \frac{(z^2-\gamma\lambda)e^{-\gamma(t'-a)}}{\beta(\gamma+\lambda)(z^2+\gamma^2)} - \frac{(z^2+\gamma\lambda)e^{-\gamma(t'-b)}}{\beta(\gamma-\lambda)(z^2+\gamma^2)}, \tag{34}$$

which utilizes harmonic $z = \frac{2\pi m}{b-a}, m \in \mathbb{Z}_+$. The covariance will also return to a scaled term in Table 4 when $\beta \to 0_+$.

D.1.2 Sine Features ($\phi(t; z) = \sin(z(t-a))$)

Case 1: $t' < a, \; P \circ k \circ G = 0$.

Case 2: $a \leq t' \leq b$,

$$\begin{aligned} P \circ k \circ G &= \int_{a}^{t'} \sin(z(\tau'-a)) \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')} \, \mathrm{d}\tau' \\ &= \frac{-z\cos(z(t'-a)) + \gamma \sin(z(t'-a))}{\beta(z^2+\gamma^2)} + \frac{z}{\beta(z^2+\gamma^2)} e^{-\gamma(t'-a)} \\ &= \frac{\sin(z(t'-a)+\theta)}{\beta\sqrt{z^2+\gamma^2}} + \xi_{\sin}, \quad \theta = -\arctan(\frac{z}{\gamma}) \end{aligned} \tag{35}$$

Case 3: $t' > b$,

$$P \circ k \circ G = \int_{a}^{b} \sin(z(\tau'-a)) \cdot \frac{1}{\beta} e^{-\gamma(t'-\tau')} \, \mathrm{d}\tau' = \frac{ze^{-\gamma(t'-a)} - ze^{-\gamma(t'-b)}}{\beta(\gamma^2+z^2)}. \tag{36}$$

The VFRFs of the LFM with a Matérn-$\frac{1}{2}$ latent force are summarised in Table 5 and 6, where the absolute distances to the interval ends are denoted as $r_a = |t'-a|$ and $r_b = |t'-b|$ and the phase shift is $\theta = -\arctan(\frac{z}{\gamma})$. The features are continuous at $\gamma = \lambda$ when $t' > b$.

Table 5: Matérn-$\frac{1}{2}$ VFRFs on Fourier basis $\phi(x;z) = \cos(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (cosine part)
$t' < a$	$\frac{1}{\beta(\gamma+\lambda)} e^{-\lambda r_a}$
$a \le t' \le b$	$\frac{\cos(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} - \left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{1}{\beta(\gamma+\lambda)}\right] e^{-\gamma r_a}$
$t' > b$ $(\gamma \neq \lambda)$	$-\left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{1}{\beta(\gamma+\lambda)}\right] e^{-\gamma r_a} + \left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{1}{\beta(\gamma-\lambda)}\right] e^{-\gamma r_b} + \frac{1}{\beta(\gamma-\lambda)} e^{-\lambda r_b}$
$t' > b$ $(\gamma = \lambda)$	$-\left[\frac{\lambda}{\beta(z^2+\lambda^2)} - \frac{1}{2\beta\lambda}\right] e^{-\lambda r_a} + \left[\frac{\lambda}{\beta(z^2+\lambda^2)} + \frac{r_b}{\beta}\right] e^{-\lambda r_b}$

Table 6: Matérn-$\frac{1}{2}$ VFRFs on Fourier basis $\phi(x;z) = \sin(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (sine part)
$t' < a$	0
$a \le t' \le b$	$\frac{\sin(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} + \frac{z}{\beta(z^2+\gamma^2)} e^{-\gamma r_a}$
$t' > b$	$\frac{z}{\beta(z^2+\gamma^2)} e^{-\gamma r_a} - \frac{z}{\beta(z^2+\gamma^2)} e^{-\gamma r_b}$

D.2 Matérn-$\frac{3}{2}$/$\frac{5}{2}$ Cross-covariance

Based on Table 7 and 10, we give the VFRFs for Matérn-$\frac{3}{2}$ and Matérn-$\frac{5}{2}$ LFMs with $\theta = -\arctan(\frac{z}{\gamma})$ in Table 8,9,11 and 12. Also, the derived cross-covariance expressions can return to the scaled VFFs of the latent force.

Table 7: VFFs $h(z, \tau')$ with Matérn-$\frac{3}{2}$ kernel

$\phi(\cdot; z)$	$\tau' < a$	$a \le \tau' \le b$	$\tau' > b$
$\cos(z(\cdot - a))$	$(1 + \lambda(a - \tau'))e^{-\lambda(a-\tau')}$	$\cos(z(\tau' - a))$	$(1 + \lambda(\tau' - b))e^{-\lambda(\tau'-b)}$
$\sin(z(\cdot - a))$	$z(\tau' - a)e^{-\lambda(a-\tau')}$	$\sin(z(\tau' - a))$	$z(\tau' - b)e^{-\lambda(\tau'-b)}$

Table 12: Matérn-$\frac{5}{2}$ VFRFs on Fourier basis $\phi(x;z) = \sin(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (sine part)
$t' < a$	$-\frac{z}{\beta}\left[\frac{\lambda r_a^2}{(\gamma+\lambda)} + \frac{(\gamma+3\lambda)r_a}{(\gamma+\lambda)^2} + \frac{\gamma+3\lambda}{(\gamma+\lambda)^3}\right] e^{-\lambda r_a}$
$a \le t' \le b$	$\frac{\sin(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} + \frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{(\gamma+3\lambda)}{(\gamma+\lambda)^3}\right] e^{-\gamma r_a}$
$t' > b$ $(\gamma \neq \lambda)$	$\frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{(\gamma+3\lambda)}{(\gamma+\lambda)^3}\right] e^{-\gamma r_a} - \frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{(\gamma-3\lambda)}{(\gamma-\lambda)^3}\right] e^{-\gamma r_b}$ $+\frac{z}{\beta}\left[\frac{\lambda r_b^2}{(\gamma-\lambda)} + \frac{(\gamma-3\lambda)r_b}{(\gamma-\lambda)^2} - \frac{\gamma-3\lambda}{(\gamma-\lambda)^3}\right] e^{-\lambda r_b}$
$t' > b$ $(\gamma = \lambda)$	$\frac{z}{\beta}\left[\frac{1}{(z^2+\lambda^2)} - \frac{1}{2\lambda^2}\right] e^{-\lambda r_a} - \frac{z}{\beta}\left[\frac{1}{(z^2+\lambda^2)} - \frac{(2\lambda r_b+3)r_b^2}{6}\right] e^{-\lambda r_b}$

Table 8: Matérn-$\frac{3}{2}$ VFRFs on Fourier basis $\phi(x;z) = \cos(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (cosine part)
$t' < a$	$\left[\frac{\lambda r_a+1}{\beta(\gamma+\lambda)} + \frac{\lambda}{\beta(\gamma+\lambda)^2}\right] e^{-\lambda r_a}$
$a \leq t' \leq b$	$\frac{\cos(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} - \left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{\gamma+2\lambda}{\beta(\gamma+\lambda)^2}\right] e^{-\gamma r_a}$
$t' > b \ (\gamma \neq \lambda)$	$-\left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{\gamma+2\lambda}{\beta(\gamma+\lambda)^2}\right] e^{-\gamma r_a} + \left[\frac{\gamma}{\beta(z^2+\gamma^2)} - \frac{\gamma-2\lambda}{\beta(\gamma-\lambda)^2}\right] e^{-\gamma r_b}$ $+\left[\frac{\lambda r_b+1}{\beta(\gamma-\lambda)} - \frac{\lambda}{\beta(\gamma-\lambda)^2}\right] e^{-\lambda r_b}$
$t' > b \ (\gamma = \lambda)$	$-\left[\frac{\lambda}{\beta(z^2+\lambda^2)} - \frac{3}{4\beta\lambda}\right] e^{-\lambda r_a} + \left[\frac{\lambda}{\beta(z^2+\lambda^2)} + \frac{(\lambda r_b+2)r_b}{2\beta}\right] e^{-\lambda r_b}$

Table 9: Matérn-$\frac{3}{2}$ VFRFs on Fourier basis $\phi(x;z) = \sin(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (sine part)
$t' < a$	$-\frac{z}{\beta}\left[\frac{r_a}{\gamma+\lambda} + \frac{1}{(\gamma+\lambda)^2}\right] e^{-\lambda r_a}$
$a \leq t' \leq b$	$\frac{\sin(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} + \frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{1}{(\gamma+\lambda)^2}\right] e^{-\gamma r_a}$
$t' > b \ (\gamma \neq \lambda)$	$\frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{1}{(\gamma+\lambda)^2}\right] e^{-\gamma r_a} - \frac{z}{\beta}\left[\frac{1}{(z^2+\gamma^2)} - \frac{1}{(\gamma-\lambda)^2}\right] e^{-\gamma r_b}$ $+\frac{z}{\beta}\left[\frac{r_b}{\gamma-\lambda} - \frac{1}{(\gamma-\lambda)^2}\right] e^{-\lambda r_b}$
$t' > b \ (\gamma = \lambda)$	$\frac{z}{\beta}\left[\frac{1}{(z^2+\lambda^2)} - \frac{1}{4\lambda^2}\right] e^{-\lambda r_a} + \frac{z}{\beta}\left[-\frac{1}{(z^2+\lambda^2)} + \frac{r_b^2}{2}\right] e^{-\lambda r_b}$

Table 10: VFFs $h(z,\tau')$ with Matérn-$\frac{5}{2}$ kernel [1]

$\phi(\cdot, z)$	$\tau' < a$	$a \leq \tau' \leq b$	$\tau' > b$
$\cos(z(\cdot - a))$	$\left[1 + \lambda r - \frac{(z^2-\lambda^2)r^2}{2}\right] e^{-\lambda r}$	$\cos(z(\tau' - a))$	$\left[1 + \lambda r - \frac{(z^2-\lambda^2)r^2}{2}\right] e^{-\lambda r}$
$\sin(z(\cdot - a))$	$z(\tau' - a)(1 + \lambda r)e^{-\lambda r}$	$\sin(z(\tau' - a))$	$z(\tau' - b)(1 + \lambda r)e^{-\lambda r}$

[1] $r = \min\{|\tau' - a|, |\tau' - b|\}$.

Table 11: Matérn-$\frac{5}{2}$ VFRFs on Fourier basis $\phi(x;z) = \cos(z(x-a))$.

$t' \in \mathbb{R}$	LFM Fourier Feature $P \circ k \circ G$ (cosine part)
$t' < a$	$-\left[\frac{(z^2-\lambda^2)r_a^2}{2\beta(\gamma+\lambda)} + \frac{(z^2-\gamma\lambda-2\lambda^2)r_a}{\beta(\gamma+\lambda)^2} + \frac{z^2-\gamma^2-3\gamma\lambda-3\lambda^2}{\beta(\gamma+\lambda)^3}\right] e^{-\lambda r_a}$
$a \leq t' \leq b$	$\frac{\cos(zr_a+\theta)}{\beta\sqrt{z^2+\gamma^2}} - \left[\frac{z^2-\gamma^2-3\gamma\lambda-3\lambda^2}{\beta(\gamma+\lambda)^3} + \frac{\gamma}{\beta(z^2+\gamma^2)}\right] e^{-\gamma r_a}$
$t' > b \ (\gamma \neq \lambda)$	$-\left[\frac{z^2-\gamma^2-3\gamma\lambda-3\lambda^2}{\beta(\gamma+\lambda)^3} + \frac{\gamma}{\beta(z^2+\gamma^2)}\right] e^{-\gamma r_a} + \left[\frac{z^2-\gamma^2+3\gamma\lambda-3\lambda^2}{\beta(\gamma-\lambda)^3} + \frac{\gamma}{\beta(z^2+\gamma^2)}\right] e^{-\gamma r_b}$ $-\left[\frac{(z^2-\lambda^2)r_b^2}{2\beta(\gamma-\lambda)} - \frac{(z^2+\gamma\lambda-2\lambda^2)r_b}{\beta(\gamma-\lambda)^2} + \frac{z^2-\gamma^2+3\gamma\lambda-3\lambda^2}{\beta(\gamma-\lambda)^3}\right] e^{-\lambda r_b}$
$t' > b \ (\gamma = \lambda)$	$-\left[\frac{z^2-7\lambda^2}{8\beta\lambda^3} + \frac{\lambda}{\beta(z^2+\lambda^2)}\right] e^{-\lambda r_a} + \left[\frac{\lambda}{\beta(z^2+\lambda^2)} - \frac{[(z^2-\lambda^2)r_b^2-3\lambda r_b-6]r_b}{6\beta}\right] e^{-\lambda r_b}$

E Experimental Details

All models in the experimental section are implemented using GPyTorch (Gardner et al., 2018), trained by Adam Optimizer on an NVIDIA A100-SXM4 GPU (for TIMIT and UCI datasets) or Apple Macbook CPUs (for illustrative examples), with a learning rate of 0.01 and a batch size of 10,000. The models using doubly stochastic variational inference, e.g., IDDGPs, DLFM-VFRF, employ five samples for layer-wise sampling during training. We follow Salimbeni and Deisenroth (2017) to set up a linear mean function for all the inner layers and a zero-mean function for the outer layer to avoid pathological behaviour (Duvenaud et al., 2014). The weights of the linear mean function are fixed and determined by SVD if the input and output dimensions are not equal. The variational distributions over inducing variables are initialised to normal distributions with zero mean and variances identity for the outer layers and 10^{-5} for the inner layers. The inducing points are initialised with K-means. All models, including RFF-based models, used 100 Monte Carlo samples for evaluations on test data.

Unless specifically stated, the RKHS interval is set to $[a, b] = [-1, 4]$, and all input data are normalised to $[0, 3]$. We initialise our model with length-scale $l = 0.1$ for the TIMIT dataset and $l = 1$ for the UCI datasets, ODE parameters $\alpha = 1, \beta = 0.01$, kernel variance $\sigma^2 = 0.1$ and noise variance $\varepsilon^2 = 0.01$.

Intermediate Outputs for Synthetic Data We present here the posterior distributions of the DGP, the IDDGP and our model DLFM-VFRF on the synthetic data in Section 5.

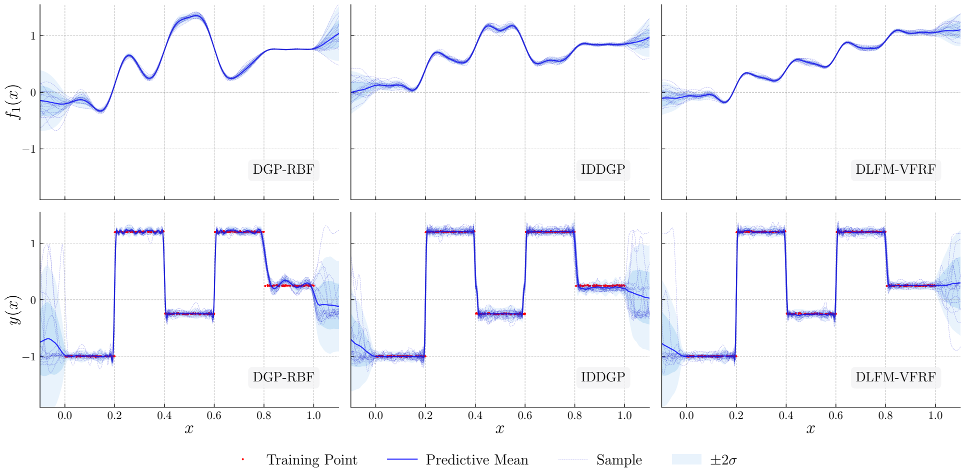


Figure 10: Comparison of posterior distributions of different compositional GP models on synthetic data. **Top row**: The output distributions of the intermediate layers. **Bottom row**: The posterior predictive distributions. Training points are marked with red dots, posterior means with blue lines, and uncertainty with varying shades of blue. Each panel depicts 20 samples from the posterior distribution. Although both the IDDGP and our model show better fitting to the multi-step function, they have very different intermediate posterior distributions.