HiBBO: HiPPO-based Space Consistency for High-dimensional Bayesian Optimisation

Anonymous authors
Paper under double-blind review

Abstract

Bayesian Optimisation (BO) is a powerful tool for optimising expensive black-box functions, but its effectiveness diminishes in high-dimensional spaces due to sparse data and poor surrogate model scalability. While Variational Autoencoder (VAE)-based approaches address this by learning low-dimensional latent representations, the reconstruction-based objective function often introduces a functional distribution mismatch between the latent space and the original space, leading to suboptimal optimisation performance. In this paper, we first analyse the reason why reconstruction-only loss may lead to distribution mismatch and then propose HiBBO, a novel BO framework that introduces space consistency into the latent space construction in VAE using HiPPO, which is a method for long-term sequence modelling, to reduce the functional distribution mismatch between the latent space and the original space. Experiments on high-dimensional benchmark tasks demonstrate that HiBBO outperforms existing VAE-BO methods in convergence speed and solution quality. Our work bridges the gap between high-dimensional sequence representation learning and efficient Bayesian Optimisation, enabling broader applications in neural architecture search, materials science, and beyond.

1 Introduction

Bayesian Optimisation (BO) (Brown et al., 2024; Hvarfner et al., 2024) is a powerful sequential strategy for optimising expensive-to-evaluate black-box functions. By leveraging probabilistic surrogate models (typical Gaussian processes (GP) (Williams & Rasmussen, 1995), Bayesian neural networks (Li et al., 2024), or variational Bayesian last layers (VBLLs) (Brunzema et al., 2025)) and acquisition functions to balance exploration and exploitation, BO efficiently guides the search for global optima with minimal function evaluations. This approach is particularly crucial in scenarios where each evaluation is costly or time-consuming, such as hyperparameter tuning in machine learning (Wang et al., 2024), experimental design in materials science (Frazier & Wang, 2015), LLMs in-context optimisation (Agarwal et al., 2025), or controller optimisation in robotics (Yuan et al., 2019). Its data-efficient nature and ability to handle noisy, non-convex objectives have made BO a cornerstone in automated decision-making systems, enabling optimisation in complex, real-world problems where gradient-based methods fail or are infeasible.

While Bayesian optimisation excels in low-dimensional spaces, its performance degrades sharply as dimensionality increases (Ziomek & Ammar, 2023; Moriconi et al., 2020; Lee et al., 2025b). This "curse of dimensionality" arises because: (1) the surrogate model's accuracy diminishes exponentially with more dimensions, as data becomes sparse relative to the search space volume; (2) acquisition functions struggle to balance exploration-exploitation in high-dimensional manifolds, often converging to suboptimal solutions; and (3) computational costs of inference and optimisation scale poorly. These limitations hinder applications in modern problems like neural architecture search (Kandasamy et al., 2018) or chemical compound design (Tripp et al., 2020), where data spaces often exceed hundreds or thousands of dimensions.

Variational Autoencoder (VAE) (Kingma & Welling, 2022)-based approaches have emerged as a promising solution to high-dimensional BO by learning compact, low-dimensional latent representations of the input space (Tripp et al., 2020; Ramchandran et al., 2025). These methods first train a VAE to encode high-dimensional data (e.g., molecular structures or neural architectures) into a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

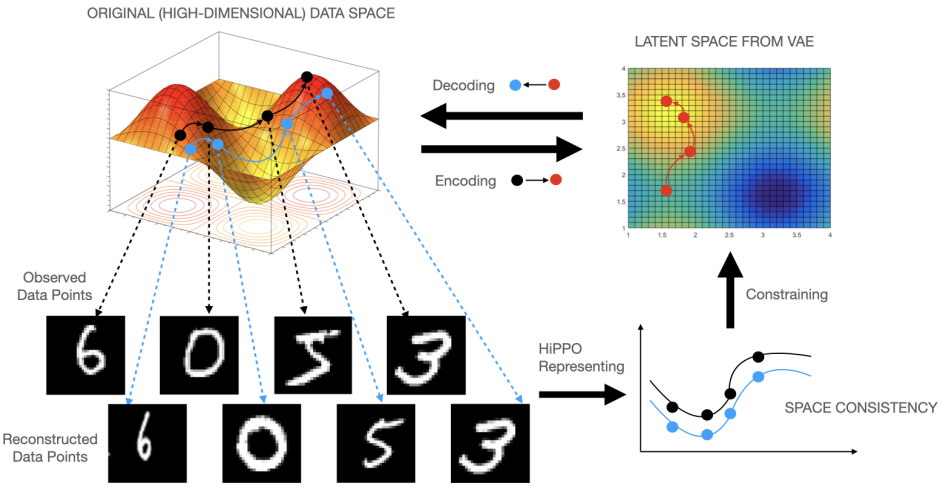


Figure 1: The illustration of our idea that is to increase the space consistency between the latent and original spaces by reducing the HiPPO representations of the reconstructed data sequence and the original data sequence.

continuous latent space, where traditional BO is then applied. Latent BO can also be built with other generative models, including normalising flows (Lee et al., 2025b), but invertible flow-based mappings generally preserve dimensionality because the latent and input spaces must have the same dimension. By contrast, a VAE can both support BO in a latent space and construct a latent dimension that is much smaller than the original input dimension. This dimensionality reduction is the main reason VAE-based BO remains attractive when the original search space is very high-dimensional. By exploiting the VAE's ability to capture meaningful structure and constraints in the data, this approach effectively reduces the optimisation problem's dimensionality while preserving critical features. The quality of BO solutions heavily depends on the VAE's representational capacity, which may discard important information during dimension reduction. To preserve the critical features, a number of strategies were proposed, such as metric learning to enhance the discriminative capability of latent representations (Grosnit et al., 2021), GP prior to model the relationship between data features (Ramchandran et al., 2025), sample reweighting to prioritise the space surrounding the data samples with extreme values (Tripp et al., 2020), and so on. However, one key challenge remains: there is a mismatch between functional distributions (GPs) in the latent space and the original data space. Such a mismatch would make the latent GP not a sufficiently good surrogate model for the following optimisation and the next evaluation point selection.

In this paper, we first analyse the underlying reason for the distribution mismatch in VAE-based latent BO: the commonly used reconstruction loss mainly enforces pointwise similarity, but does not directly preserve the kernel distance/relationship between data points, which is crucial for the GP surrogate used in BO. To address this issue, we propose HiBBO, a HiPPO-based space-consistency strategy that reduces this mismatch by constraining the HiPPO (High-order Polynomial Projection Operators) representations (Gu et al., 2020; 2023) of the original and reconstructed data sequences during latent-space construction. As illustrated in Fig. 1, the key idea is to improve consistency between the latent and original spaces by reducing the discrepancy between their HiPPO representations. Our theoretical study shows that the discrepancy in pairwise kernel statistics can be controlled through the discrepancy between HiPPO representations, together with the approximation error introduced by HiPPO truncation. This gives control of an average pairwise kernel statistic; preservation of the full kernel matrix or GP distribution requires additional assumptions. Moreover, under suitable conditions, incorporating the HiPPO regularisation term can lead to a smaller average kernel-statistic discrepancy than using reconstruction loss alone. These findings provide theoretical support for using HiPPO-based constraints to enhance consistency between the learned latent space and the original data space, thereby improving BO performance.

Our contributions can be summarised as follows:

- We identify the functional distribution mismatch in VAE-based Bayesian optimisation and show that standard reconstruction loss does not explicitly preserve the kernel geometry required by the GP surrogate.
- We propose HiBBO, a HiPPO-based space-consistency method that preserves the kernel distance/relationship between original and reconstructed data during latent-space inference.
- We provide theoretical and empirical validation: Lemma 1 bounds the kernel-statistic discrepancy using HiPPO coefficient discrepancy, Proposition 1 shows that HiPPO regularisation can improve over pure reconstruction under suitable conditions, and experiments demonstrate superior BO performance on benchmark tasks.

2 BACKGROUND: VAE-BASED BO

Given an unknown objective function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the input space, BO aims to find:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

where $f(x)$ is computationally expensive to evaluate and has no analytical form. The symbols used throughout the paper are summarised in Table 1 (see Appendix). The process consists of three steps:

- **Probabilistic Surrogate Model**: A probabilistic model, typically a GP, approximates $f(x)$. The GP provides a posterior distribution over $f(x)$ based on observed data $D_n = \{(x_i, y_i)\}_{i=1}^n$, where $y_i = f(x_i) + \epsilon_i$ and ϵ_i represents noise.
- **Acquisition Function**: An acquisition function $\alpha(x; D_n)$ guides the selection of the next query point x_{n+1}. Popular options include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB). The next point is selected as: $x_{n+1} = \arg\max_{x \in \mathcal{X}} \alpha(x; D_n)$.
- **Iterative Process**: The algorithm continuously updates the surrogate model with new observations (x_{n+1}, y_{n+1}) as it searches for the global optimum.

The VAE-based dimensionality reduction approach, $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ and $d_z \ll d_x$, for high-dimensional BO can be summarised as: train a VAE based on current observations, i.e., $\{x_i\}$, to learn an encoder distribution $q_\phi(z|x)$ and deterministic decoder mean $d_\theta(z)$; define the decoded objective $h_\theta(z) := f(d_\theta(z))$ and use a latent GP surrogate $g \sim \mathrm{GP}(m_z, k_z)$ fitted to encoded observations; select $z_{n+1} \in \arg\max_{z \in \mathcal{Z}} \alpha(z; D_n^z)$, where $D_n^z = \{(e_\phi(x_i), y_i)\}_{i=1}^n$ and $e_\phi(x) = \mathbb{E}_{q_\phi(z|x)}[z]$; decode $x_{n+1} = d_\theta(z_{n+1})$; evaluate the objective, augment the observations, and repeat.

3 OUR METHOD

In this section, we first identify the potential functional distribution mismatch between the reconstructed and original spaces in Section 3.1. Then, in Section 3.2, we propose a HiPPO-based solution to mitigate this mismatch. Finally, in Section 3.3, we introduce a new BO algorithm that incorporates this approach.

3.1 DISTRIBUTION MISMATCH BETWEEN RECONSTRUCTED AND ORIGINAL SPACE

A latent GP can be compared with an original-space GP through a specified encoder map, or through a decoder inverse when that inverse exists. However, there is a distribution mismatch between reconstructed and original functions, which presents a fundamental challenge when applying VAEs to Bayesian optimisation, because it determines the extent to which the latent space preserves the information of the original space and also the extent to which we can trust the found optimum in the latent space.

The standard VAE loss function combines reconstruction error with latent space regularisation:

$$\mathcal{L}_{\mathrm{VAE}}(\theta, \phi; x) = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\|x - d_\theta(z)\|_2 \right] + \beta_{\mathrm{KL}} \mathcal{KL}[q_\phi(z|x) \| p_0(z)] \tag{1}$$

where $p_0(z)$ is a predefined prior and β_{KL} is the KL weight. While this formulation ensures approximate reconstruction and continuous latent representations, it does not guarantee preservation

of the relationships between $\{x\}$. When we fit a Gaussian process $g(z) \sim GP(m_z, k_z)$ in the latent space using observations $\{(z_i, f(x_i))\}$, an inverse-based comparison is available in the idealised case where the deterministic decoder d_θ is injective on the latent domain. On its image $\mathcal{M}_\theta = d_\theta(\mathcal{Z})$, the inverse then defines:

$$f^{\text{VAE}}(x) = g \circ d_\theta^{-1}(x) \sim GP(m_z(d_\theta^{-1}(x)), k_z(d_\theta^{-1}(x), d_\theta^{-1}(x'))).$$

In an ideal scenario where d_θ is perfectly invertible and deterministic, this would exactly match the GP $f(x) \sim GP(m_x(x), k_x(x, x'))$ fitted in the original space, provided the mean and kernel functions satisfy:

$$m_z \circ d_\theta^{-1}(x) = m_x(x), \quad k_z(d_\theta^{-1}(x), d_\theta^{-1}(x')) = k_x(x, x').$$

For a general VAE, a single-valued decoder inverse need not exist, so the inverse-based equations are only an idealised comparison. Recent studies have explored various methods to enforce invertibility. For example, inversion-based methods Chu et al. (2024) seek to find the best-aligned latent vector for a given input; while this approach improves invertibility, it cannot guarantee it. Conversely, normalising flow-based methods Lee et al. (2025a) can guarantee an invertible mapping between the latent vector and the input, but they generally do not reduce dimension. Consequently, the dimensions of the latent vector and the input space must match, even when dealing with high-dimensional inputs. This paper therefore concentrates on VAE-based methods, whose non-invertible decoder is precisely what enables a smaller latent BO space. Using this deterministic encoder map, pointwise mean and kernel discrepancies can be defined as:

$$\Delta^{\text{mean}}(x) = |m_z(e_\phi(x)) - m_x(x)|, \quad \Delta^{\text{kernel}}(x, x') = |k_z(e_\phi(x), e_\phi(x')) - k_x(x, x')|.$$

The general reconstruction loss in VAE (i.e., the first term of Eq. (1)) does not by itself guarantee control of either discrepancy without additional assumptions linking the encoder, decoder, and GP mean and kernel functions. Therefore, even when reconstruction error is minimised, the geometric relationships between points may not be preserved in the latent space, potentially leaving the surrogate model with mismatched data correlations. This can cause the acquisition function to either miss promising regions or exploit false optima. It has unfortunately been ignored by the existing VAE-based BO literature.

A key challenge in addressing this mismatch is that while we can minimise reconstruction error during VAE training, we cannot directly optimise the kernel discrepancy since the true kernel k_x is unknown, so fitting a GP in the original high-dimensional space $\mathcal{X}$ is typically infeasible. This motivates our following approach to indirectly reduce this kernel discrepancy, which helps preserve the critical relationships between data points during the latent space construction. By maintaining better consistency between the original and latent space representations, we aim to reduce both $\Delta^{\text{mean}}(x)$ and $\Delta^{\text{kernel}}(x, x')$, leading to more reliable optimisation performance.

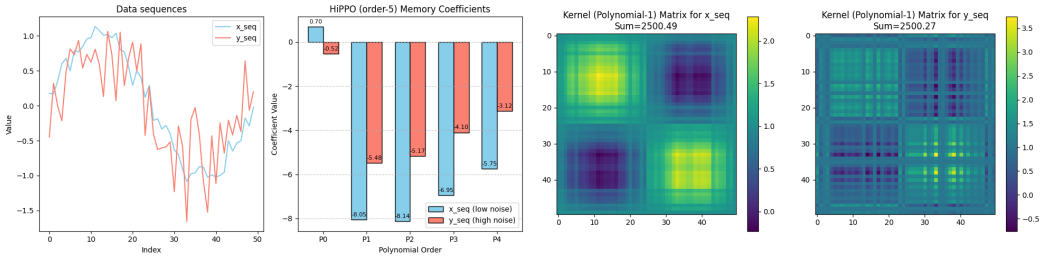
3.2 Control a kernel statistic via HiPPO memory representations

Our idea is to encourage preservation of kernel relationships[1] indirectly with the help of the HiPPO (Gu et al., 2020; 2023), which is a framework for online memory representation in continuous and discrete-time sequences. The key idea behind HiPPO is to project an input signal onto a basis of orthogonal polynomials (e.g., Legendre, Chebyshev) in a way that optimally summarises the history of the signal in a finite-dimensional state.

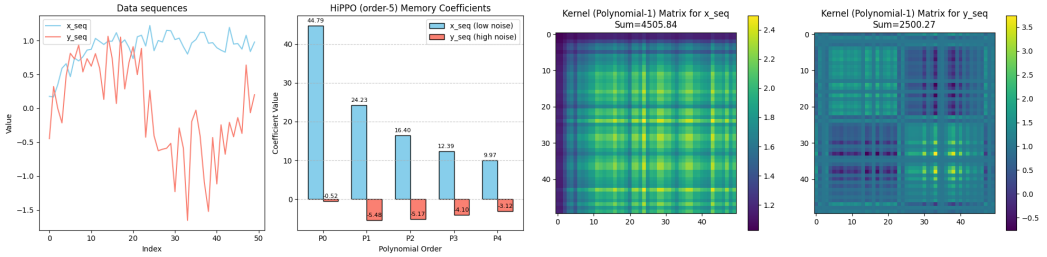
Given (sequentially obtained) data $\{x(t)\}$, HiPPO wants to maintain a memory matrix $M_t(x)$ that summarizes its history up to time t by projecting $\{x(t)\}$ onto a basis of orthogonal polynomials $\{\mathcal{P}_{r,t}\}_{r=0}^{\rho-1}$ (e.g., Legendre, Chebyshev, Laguerre) under the time-dependent weight density ω_t, i.e., $\int_{-\infty}^{t} \mathcal{P}_{n,t}(\tau)\mathcal{P}_{n',t}(\tau)\omega_t(\tau)\,\mathrm{d}\tau = \delta_{nn'}$, where ρ is the number of retained coefficients (the memory order), so the highest retained polynomial degree is $\rho - 1$. Given this projection, the original data can be approximated as:

$$x(\tau) \approx \sum_{n=0}^{\rho-1} c_{n,t}\mathcal{P}_{n,t}(\tau)$$

[1]Throughout this paper, we use the terms kernel distance, correlation, relationship, and the geometry of observations interchangeably.



(a) The left subfigure shows two sequences of data points (i.e., *x_seq* and *y_seq*) that have roughly similar correlations due to similar functional trends (the kernel matrix between data points is also visualised in the third and fourth subfigures). As shown in the second subfigure, their evaluated HiPPO representations are close.



(b) The left subfigure shows two sequences of data points (i.e., *x_seq* and *y_seq*) that have dissimilar correlations due to different functional trends (the kernel matrix between data points is also visualised in the third and fourth subfigures). As shown in the second subfigure, their evaluated HiPPO representations are far from each other.

Figure 2: Empirical demonstration of the capability of HiPPO representation in expressing the correlation between data points.

where $\tau \leq t$ and $c_{n,t}$ are the time-varying coefficients defined as

$$c_{n,t} = \int_{-\infty}^{t} x(\tau)\mathcal{P}_{n,t}(\tau)\omega_t(\tau)\mathrm{d}\tau.$$

At any time t, $M_t(x) = [c_{0,t}, c_{1,t}, \ldots, c_{\rho-1,t}]^\top \in \mathbb{R}^{\rho \times d_x}$ serves as the HiPPO memory of $\{x(\tau) : \tau \leq t\}$, which summarizes the projected trajectory history. Similarly, we can also build the corresponding HiPPO memory for the reconstructed observations $\{\bar{x}(t)\}$ from VAE mapping: $M_t(\bar{x})$, which encodes the geometric shape of $\{\bar{x}(t)\}$ from the latent space. Here x denotes an entire vector-valued trajectory, whereas x_i denotes one input vector. We propose the following final-time regularizer:

$$\|M_t(x) - M_t(\bar{x})\|_F^2. \tag{2}$$

Fig. 2 gives an intuitive example of why HiPPO memory is useful for our purpose. When two sequences have similar functional trends, their pairwise kernel matrices and HiPPO representations are both close; when the trends differ, both the kernel matrices and HiPPO representations diverge. The figure is intended only as a visual diagnostic, and the detailed data-generation settings are deferred to Appendix A.4. Next, we give our theoretical study of this regularizer and why it works.

Lemma 1. *Let $x, \bar{x} : (-\infty, t] \to \Omega \subset \mathbb{R}^{d_x}$, where Ω is compact. Let $M_t(x) = [c_{0,t}, \ldots, c_{\rho-1,t}]^\top$, $M_t(\bar{x}) = [\bar{c}_{0,t}, \ldots, \bar{c}_{\rho-1,t}]^\top$ be order-ρ HiPPO coefficient matrices, with $c_{n,t}, \bar{c}_{n,t} \in \mathbb{R}^{d_x}$. Define the order-$\rho$ HiPPO truncations $x_\rho(\tau_i) := \sum_{n=0}^{\rho-1} c_{n,t}\mathcal{P}_{n,t}(\tau_i)$, $\bar{x}_\rho(\tau_i) := \sum_{n=0}^{\rho-1} \bar{c}_{n,t}\mathcal{P}_{n,t}(\tau_i)$. For sampled times $\tau_1, \ldots, \tau_N \leq t$, define the average pairwise kernel statistic of a trajectory x by*

$$D_k(x) := \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n=1}^{N} k(x(\tau_m), x(\tau_n)).$$

Assume $k : U \times U \to \mathbb{R}$ is continuously differentiable, where $U \subseteq \mathbb{R}^{d_x}$ is a convex domain containing Ω and all sampled truncated values $x_\rho(\tau_i), \bar{x}_\rho(\tau_i)$. Suppose there exists $G_k <$

∞ *such that* $\|\nabla_1 k(a,b)\|_2 + \|\nabla_2 k(a,b)\|_2 \leq G_k$ *for all* $a, b \in U$. *Also assume* $E_\rho := \max_{1\leq i\leq N} \left(\sum_{n=0}^{\rho-1} |\mathcal{P}_{n,t}(\tau_i)|^2\right)^{1/2} < \infty$, *and that the HiPPO truncation residuals satisfy*

$$\max_i \|x(\tau_i) - x_\rho(\tau_i)\|_2 \leq \varepsilon_\rho, \qquad \max_i \|\bar{x}(\tau_i) - \bar{x}_\rho(\tau_i)\|_2 \leq \varepsilon_\rho.$$

Then

$$|D_k(x) - D_k(\bar{x})| \leq G_k E_\rho \|M_t(x) - M_t(\bar{x})\|_F + 2G_k\varepsilon_\rho.$$

Proof. Please see the Appendix for details. □

Let us compare this bound with the pure reconstruction-based bound. Under the above Lipschitz continuity assumption on the kernel function, we have

$$|D_k(x) - D_k(\bar{x})| \leq G_k \|x - \bar{x}\|_N.$$

The HiPPO-based bound is tighter than the reconstruction-based bound whenever

$$E_\rho \|M_t(x) - M_t(\bar{x})\|_F + 2\varepsilon_\rho < 2\|x - \bar{x}\|_N.$$

This condition characterises the regime in which HiPPO-based control provides a sharper estimate than reconstruction-based control. It holds when the discrepancy between the original and reconstructed trajectories is primarily concentrated in components outside the order-ρ HiPPO subspace, such as high-frequency residuals or local pointwise noise, while the HiPPO-visible low-order structure is well preserved. In this regime, the pointwise reconstruction error overestimates the error relevant to the pairwise kernel statistic. Consequently, a HiPPO regularizer can provide more targeted control of the kernel geometry than reconstruction loss alone.

The residual $r_\rho = x - x_\rho$ represents the component of the trajectory outside the order-ρ HiPPO subspace. Since this subspace is spanned by low-order orthogonal polynomials, r_ρ captures higher-order temporal variation, including local fluctuations, high-frequency oscillations, and pointwise noise. These components may contribute substantially to the pointwise reconstruction error, but they need not be important for the pairwise kernel statistic D_k, especially when k is smooth and D_k averages over all pairs.

Such residual components can be safely neglected when their norm is small relative to the kernel lengthscale, or when they behave as incoherent local perturbations whose effects are diluted by pairwise averaging. In this regime, reconstruction loss may over-penalise kernel-irrelevant residuals, whereas the HiPPO loss focuses on the low-order structural components that determine the global geometry. For example, suppose $x(\tau) = \sin(\tau) + 0.05\sin(100\tau)$. If ρ is moderate, the HiPPO coefficients mostly capture the slow component $\sin(\tau)$, while the fast oscillation $0.05\sin(100\tau)$ lies mostly outside the low-order subspace. This fast oscillation can increase pointwise reconstruction error, but may not strongly affect the global pairwise kernel statistic.

Proposition 1. *Let* $x : (-\infty, t] \to \Omega \subset \mathbb{R}^{d_x}$ *be the original trajectory, where* Ω *is compact, and let* $\mathcal{H}$ *be a compact set of trajectories attainable by the decoder. Let* $R(\bar{x}) := \|x - \bar{x}\|_N^2$, $H_\rho(\bar{x}) := \|M_t(x) - M_t(\bar{x})\|_F^2$, *and* $\bar{x}_{\mathrm{rec}} \in \arg\min_{\bar{x}\in\mathcal{H}} R(\bar{x})$ *be a reconstruction-only solution, and define its pairwise kernel-statistic error as* $K_{\mathrm{rec}} := |D_k(x) - D_k(\bar{x}_{\mathrm{rec}})|$. *Assume* $K_{\mathrm{rec}} > 0$ *and suppose the assumptions of Lemma 1 hold uniformly for* x *and all* $\bar{x} \in \mathcal{H}$. *That is, for some constants* $G_k, E_\rho, \varepsilon_\rho > 0$, $|D_k(x) - D_k(\bar{x})| \leq G_k E_\rho \|M_t(x) - M_t(\bar{x})\|_F + 2G_k\varepsilon_\rho$ *for every* $\bar{x} \in \mathcal{H}$. *Let* $H_\rho^\star := \min_{\bar{x}\in\mathcal{H}} H_\rho(\bar{x})$ *and assume the strict gap condition* $G_k E_\rho \sqrt{H_\rho^\star} + 2G_k\varepsilon_\rho < K_{\mathrm{rec}}$. *Then there exists* $\lambda_0 > 0$ *such that, for every* $\lambda \geq \lambda_0$, *every global minimizer*

$$\bar{x}_\lambda \in \arg\min_{\bar{x}\in\mathcal{H}} \{R(\bar{x}) + \lambda H_\rho(\bar{x})\}$$

satisfies

$$|D_k(x) - D_k(\bar{x}_\lambda)| < |D_k(x) - D_k(\bar{x}_{\mathrm{rec}})|.$$

Proof. Please see the Appendix for details. □

Proposition 1 shows that, under a suitable separation condition, adding the HiPPO regularisation term to the reconstruction objective can strictly improve the preservation of the pairwise kernel

statistic compared with reconstruction loss alone. The reconstruction-only solution $\bar{x}_{\mathrm{rec}}$ is optimal in the pointwise empirical norm, but this does not necessarily make it optimal for preserving the average pairwise kernel statistic D_k. In contrast, the HiPPO term directly controls the discrepancy between the low-order temporal representations $M_t(x)$ and $M_t(\bar{x})$, which are precisely the components that enter the kernel-statistic bound in Lemma 1. The strict gap condition states that the best attainable HiPPO approximation error, together with the truncation error ε_ρ, is small enough to yield a kernel-statistic error below that of the reconstruction-only solution. When this condition holds, choosing the regularisation weight λ sufficiently large forces any global minimiser of the combined objective to prioritise the HiPPO discrepancy strongly enough, and its induced kernel statistic is guaranteed to be closer to that of the original trajectory than the one from pure reconstruction.

3.3 HiBBO: BO with HiPPO-based space consistency

To reduce the distribution mismatch between the latent space and data space, we propose to use the following new loss function for VAE training:

$$\begin{aligned}\mathcal{L}_{\mathrm{HiBBO}}(\theta,\phi;X) = \sum_{x_i\in X} \mathbb{E}_{z_i\sim q_\phi(z|x_i)}\left[\|x_i - d_\theta(z_i)\|_2\right] + \beta_{\mathrm{KL}} \sum_{x_i\in X} \mathcal{KL}[q_\phi(z|x_i)\|p_0(z)] \\ + \lambda_H \mathbb{E}_{z_{1:T}\sim\prod_{i=1}^T q_\phi(z_i|x_i)}\left[\|C_T - \bar{C}_T\|_F^2\right],\end{aligned} \tag{3}$$

where $X = (x_1, \ldots, x_n)$ is ordered by observation time and $T = n$. For training, the discrete index i corresponds to the ith input. Define $M_i = M_i(x)$ and $\bar{M}_i = M_i(\bar{x})$, and concatenate the states horizontally as $C_T = [M_1, \ldots, M_T] \in \mathbb{R}^{\rho\times(Td_x)}$ and $\bar{C}_T = [\bar{M}_1, \ldots, \bar{M}_T]$. Thus the penalty equals $\sum_{i=1}^T \|M_i - \bar{M}_i\|_F^2$. The weight $\lambda_H > 0$ controls this penalty. We select HiPPO-LegS and leave other bases to future work.

As BO observes more expensive function values (i.e., as $\{x(t)\}$ expands), the HiPPO memory must be updated to incorporate the new observations. Fortunately, such an update is efficient in HiPPO because it is proven that the memory matrix M_t follows a linear ordinary differential equation (ODE):

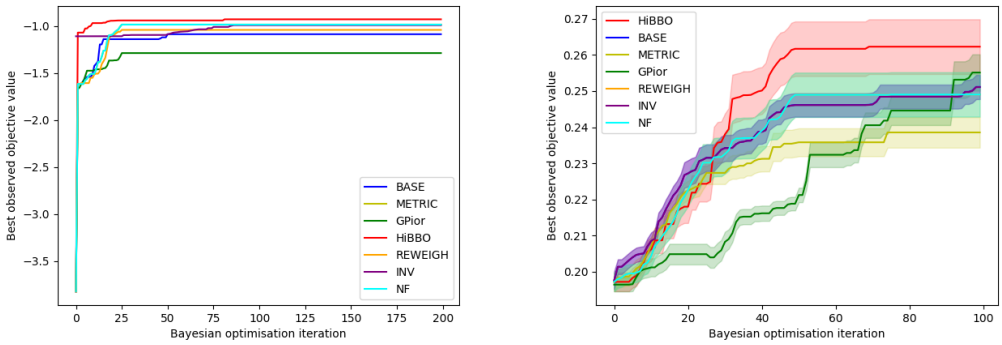
$$\frac{\mathrm{d}M_t}{\mathrm{d}t} = A_t M_t + B_t x(t)^\top \tag{4}$$

where $A_t \in \mathbb{R}^{\rho\times\rho}$ and $B_t \in \mathbb{R}^{\rho\times 1}$ depend on the chosen polynomial basis $\{\mathcal{P}_{r,t}\}_{r=0}^{\rho-1}$. Such an ODE formulation allows efficient online updates, making it useful for continuing observations from BO. Algorithm 1 in the Appendix summarises our approach.

4 Experiments

We evaluate the efficacy of our proposed method, **HiBBO**, across four high-dimensional optimisation problems: a standard Bayesian optimisation benchmark function, an MNIST-based synthetic problem, a 2D shape optimisation problem, and a chemical design application. These problems are selected to demonstrate the versatility of HiBBO in handling both synthetic and real-world scenarios while operating effectively in high-dimensional spaces.

For comparison, we consider several VAE-based Bayesian optimisation methods as baselines. **BASE** serves as the simplest baseline, employing a standard VAE without any additional constraints on the latent space. **METRIC** (Grosnit et al., 2021) improves upon this by integrating a triplet loss to enhance the discriminative power of the learned latent representations. In contrast, **GPior** (Ramchandran et al., 2025) replaces the conventional Gaussian prior with a Gaussian Process prior during VAE training, allowing the model to incorporate additional information directly into the latent structure. **REWEIGH** (Tripp et al., 2020) modifies the VAE's training objective by reweighting data points to prioritise those with higher performance values. **INV** (Chu et al., 2024) proposes an inversion strategy to alleviate the 'misalignment problem'. Finally, **NF** (Lee et al., 2025b) constructs a latent space via normalising flows instead of a VAE. To ensure a fair and unbiased comparison, the VAE-based methods share the same VAE architecture for each problem; NF uses its own flow architecture, and training procedures, including optimisation steps, retraining schedules, BO budget, and hyperparameter settings, are kept consistent across experiments.



(a) Ackley (b) MNIST

Figure 3: Results on standard function optimisation and MNIST-based synthetic problem

4.1 Standard function optimisation

We begin our experimental evaluation with a high-dimensional optimisation benchmark using the Ackley function, implemented through the **BoTorch** framework[2]. The Ackley function represents a challenging multimodal optimisation problem in d_x-dimensional space, typically evaluated on the hypercube $[-32.768, 32.768]^{d_x}$. For our experiments, we set $d_x = 1000$ to create an especially demanding high-dimensional optimisation scenario. This function features numerous local minima but has a single global minimum of 0 at the origin, providing a clear optimisation target for evaluating algorithm performance. To express this task in the maximisation convention, define $f(x) = -a(x)$, where a is the Ackley function; reported raw Ackley values $a(x)$ are minimised. The dimension of the latent space from VAE is set 10, representing a 100-fold dimensionality reduction from the original 1000-dimensional space. More details of the setup can be found in the Appendix. As shown in Fig. 3a, HiBBO demonstrates superior performance compared to baseline methods. Notably, it achieves solutions closest to the global minimum, while also converging faster in terms of iteration count. Since this problem is relatively easy, the differences among various methods are small.

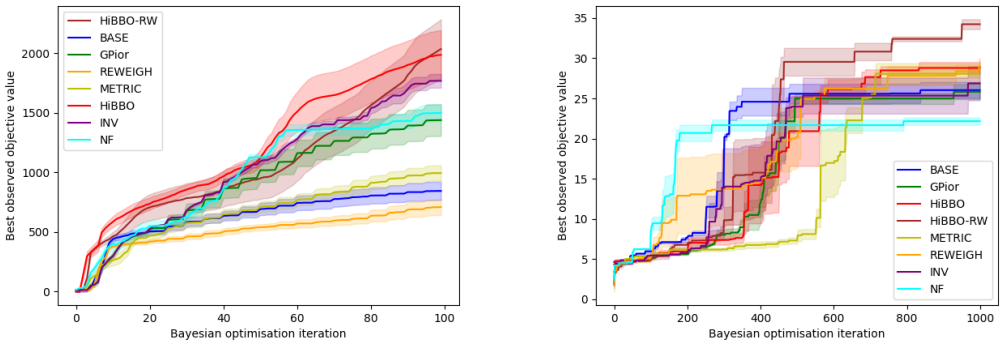
4.2 MNIST-based synthetic problem

The second benchmark problem adopts the official high-dimensional optimisation example from BoTorch, where the black-box objective function maps MNIST digit images to scalar scores through a carefully designed evaluation pipeline. In this task, the original 784-dimensional search space (28×28 pixels) represents handwritten digits from the MNIST dataset, with the optimisation target being the discovery of images maximising the utility assigned to digit '3'. Let $\pi_v(x)$ be the pre-trained classifier probability for digit $v \in \{0, \ldots, 9\}$, and define $w_v = \exp[-2(v-3)^2]$. The digit-3 utility is $s(x) = \sum_{v=0}^{9} \pi_v(x) w_v$, whose maximum is 1 when $\pi_3(x) = 1$. An equivalent minimization objective is $L_{\text{digit}}(x) = 1 - s(x)$, whose minimum is 0. The BO maximisation convention uses $f(x) = s(x)$. Fig. 3b compares the methods over four seeds. The complete experimental setup, including CNN architecture details and optimisation hyperparameters, is in the Appendix.

4.3 Shape optimisation

Another illustrative task, originally proposed by (Tripp et al., 2020), involves optimising for the shape with the largest total area within the space of 64×64 binary images—that is, maximising the number of pixels with a value of 1. The underlying generative model is a variational autoencoder (VAE) with a simple convolutional architecture (details provided in the Appendix), and its latent space is two-dimensional, i.e., $\mathbb{R}^2$. Given the low-dimensional nature of the latent space, we perform direct optimisation by enumerating a uniform grid over the range $[-3, 3]^2$ instead of GP, following (Tripp et al., 2020). The results are presented in Fig. 4a (5 seeds). As shown, our proposed method,

[2]https://botorch.org/



(a) Shape (b) Chem

Figure 4: Results on shape optimisation and chemical design

HiBBO, achieves the best performance compared to baseline approaches. When combined with the reweighting strategy (**HiBBO-RW**), the method still attains a strong final objective value, though slightly lower than that of HiBBO alone.

4.4 Chemical design

This experiment builds upon the molecular optimisation task introduced by Tripp et al. (2020); Ramchandran et al. (2025). We adopt the standardised benchmark from Gómez-Bombarelli et al. (2018), which involves synthesising molecules with the highest penalised water-octanol partition coefficient (logP), using the ZINC250k dataset (Irwin et al., 2012) as the starting point. To ensure valid chemical structures, we use a junction tree VAE (Jin et al., 2018) to generate the latent space. Optimisation is performed using the expected improvement acquisition function. For evaluating M_t and $\bar{M}_t$, we use vector representations concatenated from *tree_vecs* and *mol_vecs* obtained from the junction tree VAE encoder, rather than the original SMILES strings. All models are trained and evaluated under consistent settings, following the original codebase[3]. Additional implementation details are provided in the Appendix. Fig. 4b (5 seeds) shows the optimisation results. Our method, HiBBO, achieves excellent final objective values, and the HiBBO-RW variant stands out, attaining the highest logP score with the lowest query budget among all compared methods. This reinforces the benefit of incorporating reweighting, particularly in complex design tasks.

5 Conclusion

In this work, we introduced HiBBO, a novel Bayesian Optimisation (BO) framework designed to address the persistent challenge of functional distribution mismatch in VAE-based high-dimensional optimisation. By incorporating HiPPO-based memory representations into the latent space construction, HiBBO encourages consistency between original and reconstructed sequences. Our analysis bounds an average pairwise kernel statistic; it does not by itself establish preservation of the complete GP mean or covariance. Our theoretical analysis and empirical results demonstrate that this added space consistency significantly enhances the fidelity of the surrogate model in the latent space, leading to more reliable and efficient optimisation. The proposed HiPPO-based regularisation is lightweight, generalisable, and compatible with existing VAE architectures, making it a practical enhancement for a wide range of BO applications.

Reproducibility and AI Use Statement

The implementation and experimental details are provided in the Appendix, and the code is included in the supplementary materials. In this work, we used generative AI tools for assistant coding and

[3]https://github.com/cambridge-mlg/weighted-retraining

paper proofreading. We have not used generative AI tools for idea proposal. We have reviewed all AI-assisted work. We checked LLM-generated code and proofreading results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Dhruv Agarwal, Manoj Ghuhan Arivazhagan, Rajarshi Das, Sandesh Swamy, Sopan Khosla, and Rashmi Gangadharaiah. Searching for optimal solutions with llms via bayesian optimization. In *International Conference on Learning Representations*, 2025.

Rika Antonova, Akshara Rai, Tianyu Li, and Danica Kragic. Bayesian optimization in variational latent spaces with dynamic compression. In *Conference on Robot Learning*, pp. 456–465. PMLR, 2020.

Theodore Brown, Alexandru Cioba, and Ilija Bogunovic. Sample-efficient bayesian optimisation using known invariances. *Advances in Neural Information Processing Systems*, 37:47931–47965, 2024.

Paul Brunzema, Mikkel Jordahn, John Willes, Sebastian Trimpe, Jasper Snoek, and James Harrison. Bayesian optimization via continual variational last layer training. In *International Conference on Learning Representations*, 2025.

Jaewon Chu, Jinyoung Park, Seunghun Lee, and Hyunwoo Kim. Inversion-based latent bayesian optimization. *Advances in Neural Information Processing Systems*, 37:68258–682, 2024.

Peter I Frazier and Jialei Wang. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pp. 45–75. Springer, 2015.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Antoine Grosnit, Rasul Tutunov, Alexandre Max Maraval, Ryan-Rhys Griffiths, Alexander I Cowen-Rivers, Lin Yang, Lin Zhu, Wenlong Lyu, Zhitang Chen, Jun Wang, et al. High-dimensional bayesian optimisation with variational autoencoders and deep metric learning. *arXiv preprint arXiv:2106.03609*, 2021.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in Neural Information Processing Systems*, 33: 1474–1487, 2020.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your hippo: State space models with generalized orthogonal basis projections. In *International Conference on Learning Representations*, 2023.

Carl Hvarfner, Frank Hutter, and Luigi Nardi. A general framework for user-guided bayesian optimization. In *International Conference on Learning Representations*, 2024.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52 (7):1757–1768, 2012.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pp. 2323–2332. PMLR, 2018.

Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Neural architecture search with bayesian optimisation and optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arxiv.org/abs/1312.6114*, 2022.

Seunghun Lee, Jinyoung Park, Jaewon Chu, Minseo Yoon, and Hyunwoo Kim. Latent bayesian optimisation via autoregressive normalizing flows. In *International Conference on Learning Representations*, volume 2025, pp. 89789–89811, 2025a.

Seunghun Lee, Jinyoung Park, Jaewon Chu, Minseo Yoon, and Hyunwoo J Kim. Latent bayesian optimization via autoregressive normalizing flows. In *International Conference on Learning Representations*, 2025b.

Yucen Lily Li, Tim GJ Rudner, and Andrew Gordon Wilson. A study of bayesian neural network surrogates for bayesian optimization. In *International Conference on Learning Representations*, 2024.

Riccardo Moriconi, Marc Peter Deisenroth, and KS Sesh Kumar. High-dimensional bayesian optimization using low-dimensional feature spaces. *Machine Learning*, 109(9):1925–1943, 2020.

Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in vae latent space using decoder uncertainty. *Advances in Neural Information Processing Systems*, 34:802–814, 2021.

Siddharth Ramchandran, Manuel Haussmann, and Harri Lähdesmäki. High-dimensional bayesian optimisation with gaussian process prior variational autoencoders. In *International Conference on Learning Representations*, 2025.

Zhichao Tian, Yang Yang, Sui Zhou, Tian Zhou, Ke Deng, Chunlin Ji, Yejun He, and Jun S Liu. High-dimensional bayesian optimization for metamaterial design. *Materials Genome Engineering Advances*, 2(4):e79, 2024.

Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *Advances in Neural Information Processing Systems*, 33:11259–11272, 2020.

Zi Wang, George E Dahl, Kevin Swersky, Chansoo Lee, Zachary Nado, Justin Gilmer, Jasper Snoek, and Zoubin Ghahramani. Pre-trained gaussian processes for bayesian optimization. *Journal of Machine Learning Research*, 25(212):1–83, 2024.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8, 1995.

Kai Yuan, Iordanis Chatzinikolaidis, and Zhibin Li. Bayesian optimization for whole-body control of high-degree-of-freedom robots through reduction of dimensionality. *IEEE Robotics and Automation Letters*, 4(3):2268–2275, 2019.

Yimeng Zeng, Hunter Elliott, Phillip Maffettone, Peyton Greenside, Osbert Bastani, and Jacob R Gardner. Antibody design with constrained bayesian optimization. In *ICLR Workshop on Generative and Experimental Perspectives for Biomolecular Design*. GEM workshop, ICLR 2024, 2024.

Juliusz Krzysztof Ziomek and Haitham Bou Ammar. Are random decompositions all we need in high dimensional bayesian optimisation? In *International Conference on Machine Learning*, pp. 43347–43368. PMLR, 2023.

A Appendix

A.1 Notation Table

The notation used throughout the paper is collected below.

Table 1: Notation used throughout the paper.

Notation	Meaning
$\mathcal{X} \subseteq \mathbb{R}^{d_x}, \mathcal{Z} \subseteq \mathbb{R}^{d_z}$	Input and latent domains; d_x, d_z are their dimensions.
$\mathcal{Y} \subseteq \mathbb{R}$	Scalar output domain.
$x, x_i, \bar{x}_i, z, z_i$	Input vector, indexed input, reconstructed input, latent vector, and indexed latent sample.
$x(\tau), \bar{x}(\tau)$	Original and reconstructed vector-valued trajectories; bare $x, \bar{x}$ also denote trajectories in the theory.
f, g, h_θ	Black-box objective, latent GP surrogate, and decoded objective $h_\theta = f \circ d_\theta$.
y_i, ϵ_i	Observed response $y_i = f(x_i) + \epsilon_i$ and observation noise; $\epsilon_i = 0$ in noiseless tasks.
D_n, D_n^z, X, n	Observed pairs, encoded observed pairs, ordered input sequence, and current number of observations.
$x^\star, x_{\text{best}}, i^\star$	True global maximiser, best observed point, and its observation index.
$x_{n+1}, z_{n+1}, \hat{x}, \hat{z}, \hat{y}$	Next-query input and latent point; candidate input, latent point, and observed response in the algorithm.
$\alpha(z; D_n^z), \eta$	Acquisition function conditional on encoded observations, and threshold triggering earlier VAE retraining.
m_x, m_z, k_x, k_z	Original-space and latent-space GP mean functions and covariance kernels.
$q_\phi(z \mid x), \phi$	Variational encoder distribution and its parameters.
$p_\theta(x \mid z), d_\theta(z), \theta$	Decoder likelihood, deterministic decoder mean, and decoder parameters.
$p_0(z), e_\phi(x)$	Latent prior and encoder mean $e_\phi(x) = \mathbb{E}_{q_\phi}[z]$.
$\mathcal{M}_\theta, d_\theta^{-1}$	Decoder image and its inverse, used only under the stated injective-decoder idealisation.
$f^{\text{VAE}}, f^{\text{enc}}$	Ideal inverse-composed surrogate $g \circ d_\theta^{-1}$ and encoder-composed surrogate $g \circ e_\phi$.
$\Delta^{\text{mean}}(x), \Delta^{\text{kernel}}(x, x')$	Pointwise mean and covariance discrepancies under the specified encoder map.
t, τ, τ_i, N, T	Continuous terminal time, integration time, sampled times, number of theory samples, and final discrete training index ($T = n$).
$\rho, \mathcal{P}_{r,t}, \omega_t, \delta_{rr'}$	Number of retained coefficients, time-dependent orthonormal polynomial, weight density, and Kronecker delta.
$c_{r,t}, \bar{c}_{r,t}$	Original and reconstructed coefficient vectors in $\mathbb{R}^{d_x}$ for degree r.
$M_t(x), M_t(\bar{x}), M_i, \bar{M}_i$	Single-time coefficient matrices in $\mathbb{R}^{\rho \times d_x}$; discrete states use index i.
$C_T, \bar{C}_T$	Horizontal concatenations of all discrete states, in $\mathbb{R}^{\rho \times (T d_x)}$.
A_t, B_t	HiPPO state matrix in $\mathbb{R}^{\rho \times \rho}$ and input matrix in $\mathbb{R}^{\rho \times 1}$.
$x_\rho, \bar{x}_\rho, r_\rho, \bar{r}_\rho$	Truncated trajectories and residuals $r_\rho = x - x_\rho$, $\bar{r}_\rho = \bar{x} - \bar{x}_\rho$.
Ω, U	Compact range of original/reconstructed trajectories and convex domain containing all kernel-evaluation paths.
$k, D_k(x)$	Kernel used in the theory and average pairwise kernel statistic of the sampled trajectory.
$G_k, E_\rho, \varepsilon_\rho$	Uniform sum-of-gradient bound, maximum sampled basis-vector norm, and uniform truncation-error bound.
$\nabla_1 k, \nabla_2 k$	Gradients with respect to the first and second kernel arguments.
$\|\cdot\|_2, \|\cdot\|_F, \|\cdot\|_N$	Euclidean norm, Frobenius norm, and empirical RMS trajectory norm.
$\mathcal{H}, R, H_\rho$	Compact attainable trajectory class, squared RMS reconstruction loss, and squared terminal-state discrepancy.
$\bar{x}_{\text{rec}}, K_{\text{rec}}$	Reconstruction-only minimizer and its absolute average kernel-statistic error.

Continued on next page

Table 1 (continued)

Notation	Meaning
$H_\rho^\star, \bar{x}^\star$	Minimum attainable terminal-state discrepancy and a trajectory attaining it.
$\bar{x}_\lambda, \lambda, \lambda_0$	Minimiser of the simplified combined objective, its theoretical weight, and a sufficient weight threshold.
$F(C), C, \bar{C}, \Delta C, \Delta c_j, \Delta x_\rho$	Proof-local kernel statistic as a function of coefficients, generic coefficient matrices, and coefficient/trajectory perturbations.
$\mathcal{L}_{\mathrm{VAE}}, \mathcal{L}_{\mathrm{HiBBO}}, \lambda_H$	VAE loss, HiBBO training loss, and all-state HiPPO penalty weight.
$\mathcal{KL}, \beta_{\mathrm{KL}}, \mathbb{E}, \mathrm{GP}$	Kullback–Leibler divergence, its loss weight, expectation, and Gaussian process.
$\mathcal{B}, \nu, b, \ell$	Additional-query budget, VAE update interval, completed additional queries, and inner-loop index.
$a(x), \pi_v(x), w_v, v$	Raw Ackley objective, digit-class probability, digit utility weight, and digit index $v \in \{0, \ldots, 9\}$.
$s(x), L_{\mathrm{digit}}(x)$	Digit-3 utility and equivalent loss $L_{\mathrm{digit}} = 1 - s$.
$z_{\mathrm{tree}}, z_{\mathrm{graph}}$	JT-VAE latent components for tree structure and graph connectivity.

A.2 Pseudocode for the algorithm

Algorithm 1 HiBBO

Input: additional-query budget $\mathcal{B}$, update interval ν, ordered initial inputs X and stored responses, HiPPO weight λ_H, acquisition threshold η
Output: best observed point x_{best}
Set additional-query count $b = 0$.
while $b < \mathcal{B}$ **do**
  for each VAE training epoch **do**
    Set $T = |X|$, $M_0 = \bar{M}_0 = 0$ and $C_T = \bar{C}_T = \emptyset$.
    for $i = 1$ **to** T **do**
      Sample $z_i \sim q_\phi(z \mid x_i)$; set $\bar{x}_i = d_\theta(z_i)$.
      Update M_i from (M_{i-1}, x_i) and $\bar{M}_i$ from $(\bar{M}_{i-1}, \bar{x}_i)$ using the bilinear discretization of Eq. (4).
      Append M_i to C_T and $\bar{M}_i$ to $\bar{C}_T$ horizontally.
    end for
    Update (θ, ϕ) using a Monte Carlo estimate of Eq. (3).
  end for
  for $\ell = 1$ **to** $\min(\nu, \mathcal{B} - b)$ **do**
    Re-encode current inputs with e_ϕ and fit the GP to D_n^z.
    Find $\hat{z} \in \arg\max_{z \in \mathcal{Z}} \alpha(z; D_n^z)$.
    Decode $\hat{x} = d_\theta(\hat{z})$, observe $\hat{y}$, and append $(\hat{x}, \hat{y})$ to the data; set $b = b + 1$.
    if $\alpha(\hat{z}; D_n^z) < \eta$ **then**
      End this inner loop and update the VAE before the next query.
    end if
  end for
end while
Return $x_{\mathrm{best}} = x_{i^\star}$, where $i^\star \in \arg\max_i y_i$ (for noiseless observations, $y_i = f(x_i)$).

The budget $\mathcal{B}$ counts additional objective evaluations after initialisation, and frequency ν denotes the number of BO steps performed before updating the VAE model. There are mainly two procedures in Algorithm 1. The first procedure updates the VAE under the HiPPO-based space consistency constraint. We first reconstruct the data X with the VAE, then construct the HiPPO memory matrices C_T and $\bar{C}_T$, and finally use the loss function in Eq. (3) to guide the optimisation of the VAE encoder and decoder. Compared with classical VAE training, this new solution is expected to reduce the distribution mismatch between the latent space and data space. The second procedure performs BO in the latent space, which is common to all VAE-based BO.

A.3 Proof of Lemma 1 and Proposition 1

Proof of Lemma 1 For a generic coefficient matrix $C = [c_0, \dots, c_{\rho-1}]^\top$, define

$$F(C) := \frac{1}{N^2} \sum_{m,n=1}^{N} k\left(\sum_{j=0}^{\rho-1} c_j \mathcal{P}_{j,t}(\tau_m), \sum_{j=0}^{\rho-1} c_j \mathcal{P}_{j,t}(\tau_n)\right).$$

Then

$$D_k(x_\rho) = F(M_t(x)), \qquad D_k(\bar{x}_\rho) = F(M_t(\bar{x})).$$

For any coefficient perturbation ΔC,

$$\Delta x_\rho(\tau_i) = \sum_{j=0}^{\rho-1} \Delta c_j \mathcal{P}_{j,t}(\tau_i),$$

and by Cauchy–Schwarz,

$$\|\Delta x_\rho(\tau_i)\|_2 \le E_\rho \|\Delta C\|_F.$$

Using the mean value theorem and the gradient bound on k, this gives

$$|F(C) - F(\bar{C})| \le G_k E_\rho \|C - \bar{C}\|_F.$$

Therefore,

$$|D_k(x_\rho) - D_k(\bar{x}_\rho)| \le G_k E_\rho \|M_t(x) - M_t(\bar{x})\|_F.$$

For the residual terms, again using the mean value theorem,

$$|D_k(x) - D_k(x_\rho)| \le G_k \varepsilon_\rho, \qquad |D_k(\bar{x}) - D_k(\bar{x}_\rho)| \le G_k \varepsilon_\rho.$$

Hence

$$\begin{aligned} |D_k(x) - D_k(\bar{x})| &\le |D_k(x) - D_k(x_\rho)| + |D_k(x_\rho) - D_k(\bar{x}_\rho)| + |D_k(\bar{x}_\rho) - D_k(\bar{x})| \\ &\le G_k E_\rho \|M_t(x) - M_t(\bar{x})\|_F + 2G_k \varepsilon_\rho. \end{aligned}$$

This proves the lemma.

Proof of Proposition 1. Since $\mathcal{H}$ is compact and R, H_ρ are continuous, minimizers exist. Choose

$$\bar{x}^\star \in \arg\min_{\bar{x} \in \mathcal{H}} H_\rho(\bar{x}),$$

so that

$$H_\rho(\bar{x}^\star) = H_\rho^\star.$$

Let $\bar{x}_\lambda$ be a global minimizer of

$$R(\bar{x}) + \lambda H_\rho(\bar{x}).$$

By optimality,

$$R(\bar{x}_\lambda) + \lambda H_\rho(\bar{x}_\lambda) \le R(\bar{x}^\star) + \lambda H_\rho^\star.$$

Since $R(\bar{x}_\lambda) \ge 0$, we have

$$H_\rho(\bar{x}_\lambda) \le H_\rho^\star + \frac{R(\bar{x}^\star)}{\lambda}. \tag{A.1}$$

By Lemma 1,

$$|D_k(x) - D_k(\bar{x}_\lambda)| \le G_k E_\rho \|M_t(x) - M_t(\bar{x}_\lambda)\|_F + 2G_k \varepsilon_\rho.$$

Using the definition of H_ρ,

$$\|M_t(x) - M_t(\bar{x}_\lambda)\|_F = \sqrt{H_\rho(\bar{x}_\lambda)}.$$

Therefore,

$$|D_k(x) - D_k(\bar{x}_\lambda)| \le G_k E_\rho \sqrt{H_\rho(\bar{x}_\lambda)} + 2G_k \varepsilon_\rho.$$

Combining with Eq. (A.1),

$$|D_k(x) - D_k(\bar{x}_\lambda)| \le G_k E_\rho \sqrt{H_\rho^\star + \frac{R(\bar{x}^\star)}{\lambda}} + 2G_k\varepsilon_\rho. \tag{A.2}$$

By the strict gap condition,

$$G_k E_\rho \sqrt{H_\rho^\star} + 2G_k\varepsilon_\rho < K_{\mathrm{rec}}.$$

Hence there exists $\delta > 0$ such that

$$G_k E_\rho \sqrt{H_\rho^\star + \delta} + 2G_k\varepsilon_\rho < K_{\mathrm{rec}}.$$

Choose $\lambda_0 > 0$ such that

$$\frac{R(\bar{x}^\star)}{\lambda_0} \le \delta.$$

Then for every $\lambda \ge \lambda_0$, Eq. (A.2) gives

$$|D_k(x) - D_k(\bar{x}_\lambda)| \le G_k E_\rho \sqrt{H_\rho^\star + \delta} + 2G_k\varepsilon_\rho < K_{\mathrm{rec}}.$$

Since

$$K_{\mathrm{rec}} = |D_k(x) - D_k(\bar{x}_{\mathrm{rec}})|,$$

we conclude that

$$|D_k(x) - D_k(\bar{x}_\lambda)| < |D_k(x) - D_k(\bar{x}_{\mathrm{rec}})|.$$

This proves the claim. □

A.4 Details for the visualisation in Figure 2

In Fig. 2(a), *x_seq* contains 50 points from a *sin()* function with additive standard-normal noise scaled by 0.1, while *y_seq* contains 50 points from a *sin()* function with additive standard-normal noise scaled by 0.5. In Fig. 2(b), *x_seq* is generated from the same noisy *sin()* function, while *y_seq* is generated from a *tanh()* function with additive standard-normal noise scaled by 0.5. The correlation of data points in the two right-hand panels of each row is calculated using a polynomial kernel with order 1, while the HiPPO representation uses order 5. More examples are given in Fig. 5.

A.5 Related Works

The literature on VAE-based high-dimensional Bayesian Optimisation (BO) explores methods to efficiently optimise expensive black-box functions in high-dimensional spaces by leveraging VAEs for dimensionality reduction. Traditional BO struggles in high dimensions due to the curse of dimensionality, but VAEs provide a solution by learning a compressed, structured latent space that retains the essential features of the input data. One of the earliest and most influential applications of VAE-based BO was in automatic chemical design, where Tripp et al. (2020) used a VAE to encode molecular structures into a continuous latent space, enabling gradient-based optimisation of molecular properties. Building on this, Grosnit et al. (2021) introduced a method that combines VAEs with deep metric learning to structure the latent space using label information from the black-box function. This approach enhances the smoothness and informativeness of the latent space, improving the performance of the GP surrogate model used in BO. Their method achieved state-of-the-art results on benchmarks like penalised logP for molecule generation, requiring significantly less labelled data than previous methods.

Recent studies have also explored structured priors, such as GP (Ramchandran et al., 2025)), for the latent variable of VAE to capture the relationship between data and their external features. Meanwhile, Notin et al. (2021) proposed leveraging the epistemic uncertainty of the decoder to guide the optimisation process to improve its robustness and sample validity. Applications of VAE-based BO extend beyond chemistry. In protein design, Zeng et al. (2024) utilised VAE-BO to efficiently generate diverse, high-affinity antibody candidates, which were subsequently validated through in vitro synthesis. Similarly, in materials science, Tian et al. (2024) demonstrated the use of VAE-BO

for efficient design of electromagnetic metamaterials by reducing the high-dimensional microstructure space into a compact latent space. VAE-BO was also used in robotics (Antonova et al., 2020) to enable ultra-data-efficient controller tuning by learning low-dimensional latent representations of simulated trajectories, eliminating the need for expert-designed features and reducing exploration in undesirable state spaces.

In summary, VAE-based BO is a powerful framework for optimising complex, high-dimensional functions. However, its success depends critically on the quality of the latent space, for which this work offers a new strategy. This paper focuses its method design on VAE-based BO and includes NF as an additional latent-BO baseline. Other latent generative models, such as normalising flows, provide useful invertible mappings, but their dimensionality-preserving structure is different from the VAE setting considered here. Our goal is therefore not to compare all possible latent BO frameworks, but to improve the kernel-geometry preservation of VAE-based dimensionality reduction.

A.6 Experiment details

For REWEIGH, we use the direct weighting method (instead of using weighted sampling for batches), which means the weights are directly multiplied by the loss errors of the corresponding data points. The weights are rank-based. For the GPior, we only use the target function values to evaluate the GP prior kernel values without any additional information. For INV, the original paper includes a trust-region component aside from the inversion idea, but we exclude it in this paper to make it fair to other methods because such a trust-region component can be attached to any VAE-based method.

Table 2: Hyperparameters or setup for Sections 4.1 and 4.2

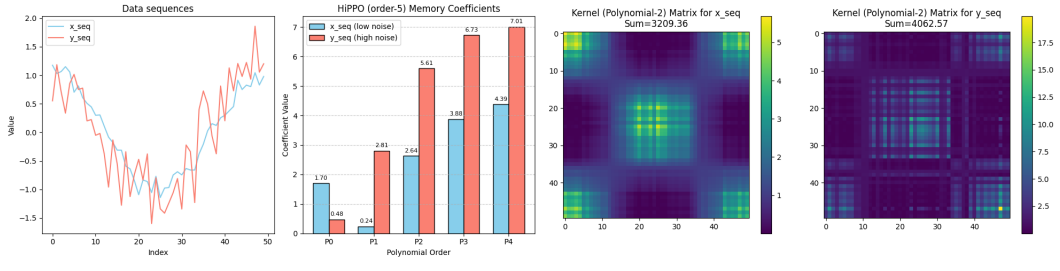
hyperparameter	**value**
VAE architecture	INPUT-FC(50)-FC(50)-FC(50)-FC(50)-FC(50)-(MU, SIGMA)
seed	123
the epoch number for BO	100
the initial data samples used for the first VAE pretraining	10
the epoch number VAE update	10
the acquisition function	UpperConfidenceBound
ode discrete method for HiPPO	'bilinear'
HiPPO dimension	50

Table 3: Hyperparameters or setup for Section 4.3

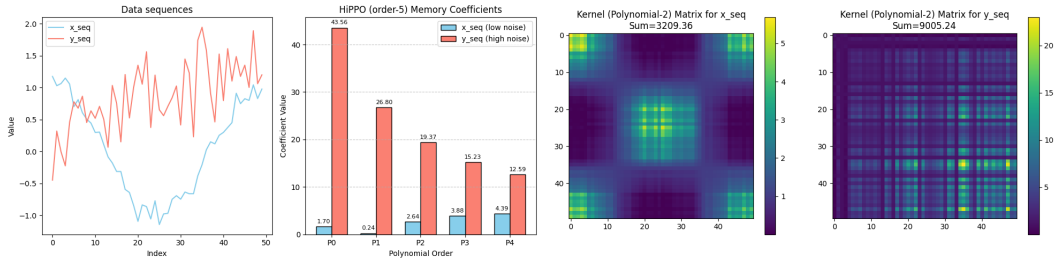
hyperparameter	**value**
VAE encoder	CONV(1,4)-CONV(4,8)-CONV(8,8)-CONV(8,16)-CONV(16,16) x 7-FC(256-32)-FC(32,2*latent-dim)
VAE decoder	FC(latent-dim,32)-FC(32, 256)-CONV(16,32)-CONV(32,32) x 2-CONV(32,16)-CONV(16,16) x 3-CONV(16-8)-CONV(8-8) x 2-CONV(8,1)
seed	1
query budget	500
VAE update frequency	5
ode discrete method for HiPPO	'bilinear'
HiPPO dimension	50

Table 4: Hyperparameters or setup for Section 4.4

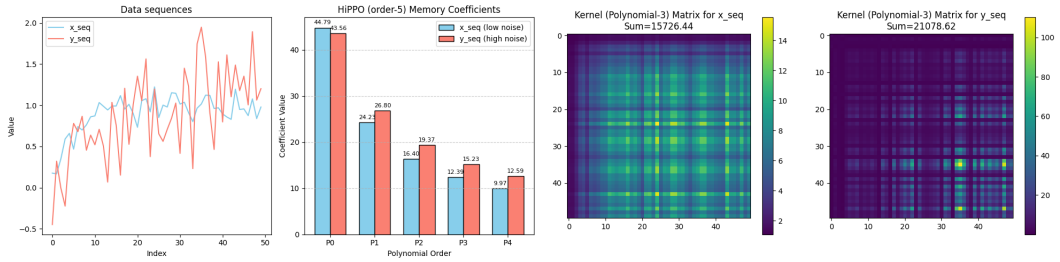
hyperparameter	**value**
VAE	JTVAE
seed	2
query budget	1000
VAE update frequency	50
ode discrete method for HiPPO	'bilinear'
HiPPO dimension	50



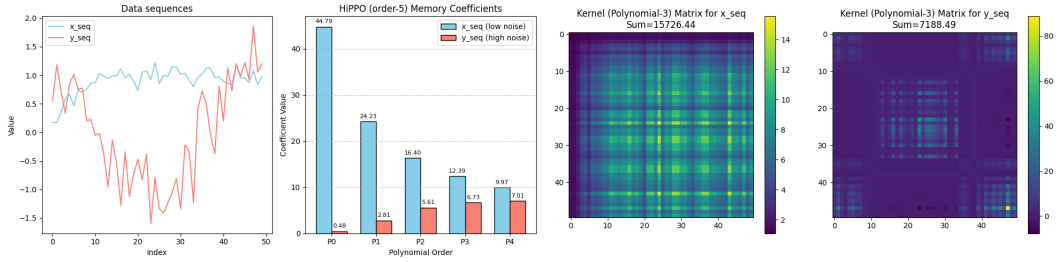
(a) Both *x_seq* and *y_seq* from the cos function



(b) *x_seq* from the cos function, and *y_seq* from tanh function



(c) Both *x_seq* and *y_seq* from the tanh function



(d) *x_seq* from the tanh function, and *y_seq* from cos function

Figure 5: Empirical demonstration of the capability of HiPPO representation in expressing the correlation between data points. The left subfigure shows two sequences of data points (i.e., *x_seq* and *y_seq*) that have roughly similar correlations due to similar functional trends (the kernel matrix between data points is also visualised in the third and fourth subfigures). As shown in the second subfigure, when the shapes or trends of two sequences are similar, their corresponding HiPPO representations are close; in contrast, their distance increases when the sequences differ.

In the molecule generation experiment, we employ the Junction Tree Variational Autoencoder (JT-VAE), which extends traditional VAEs to molecular graphs through specialised encoders and decoders. The encoder learns two distinct latent representations: one captures the tree structure and high-level cluster information, while the other encodes fine-grained connectivity details.

Molecular graph generation proceeds in two stages. First, the model constructs a tree-structured scaffold composed of chemical substructures. Then, it assembles these substructures into a complete molecule. The molecule is represented by a latent vector $z = [z_{\text{tree}}^\top, z_{\text{graph}}^\top]^\top$, where z_{tree} encodes

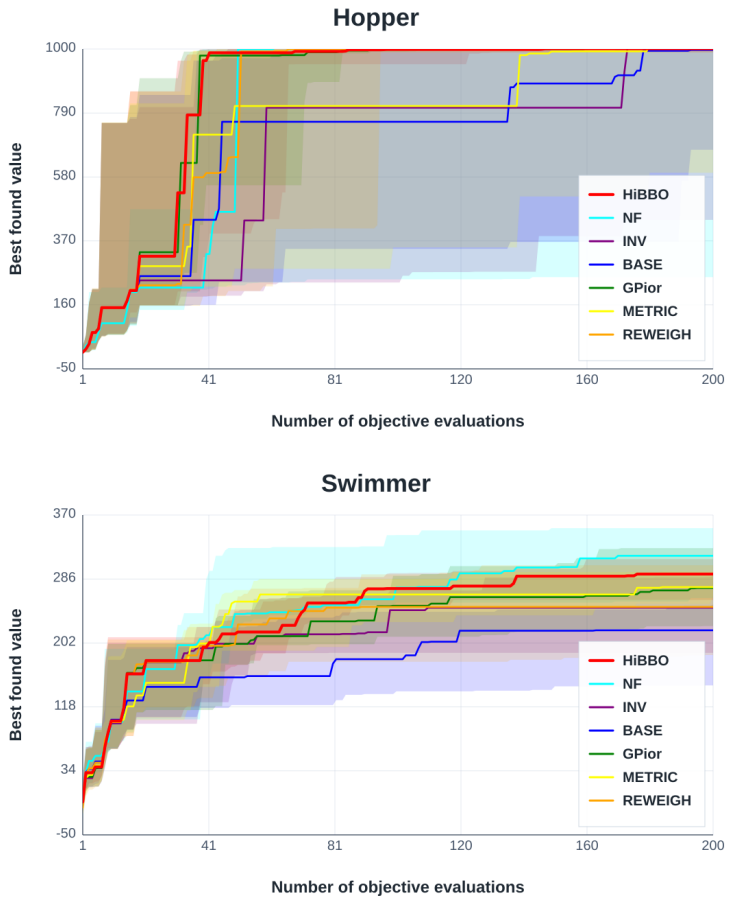


Figure 6: Results on *hopper* and *swimmer*

the tree structure and cluster-level information, and z_{graph} captures the detailed connectivity between clusters. The latent space has a total dimensionality of 56, with 28 dimensions each for z_{tree} and z_{graph}.

Decoding also occurs in two phases. The tree decoder reconstructs the junction tree, followed by the graph decoder, which predicts the fine-grained connections between clusters to form the full molecular graph. This component-wise generation approach avoids atom-by-atom assembly and prevents chemically invalid intermediates. Given the complexity of molecular structures and their limited computability, we utilise intermediate representations from JT-VAE: tree_vecs and moles_vecs, each with 450 dimensions. Together, they form the original data space with a total dimensionality of 900.

A.7 Further results on Hopper and Swimmer

We evaluate the compared Bayesian optimisation methods on two continuous-control MuJoCo benchmark tasks, *Hopper* and *Swimmer*. In both tasks, the objective is to maximise the best return found within a fixed evaluation budget. Each method is initialised from the same Sobol-style random design and then iteratively proposes new candidates using a latent-space Bayesian optimisation loop. Fig. 6 reports the best-found objective value as a function of the number of objective evaluations. Curves show the median across random seeds, while shaded regions indicate the interquartile range between the 25th and 75th percentiles. We compare HiBBO against NF, INV, BASE, GPior, METRIC, and REWEIGH.

On Hopper, most methods eventually reach a similar final performance range near 1000 return, indicating that the benchmark is solvable by several latent-space BO variants within the plotted budget.

Table 5: Important hyperparameters used for the Hopper and Swimmer trajectory plots.

Category	Hyperparameter	Value
BO budget	Plotted evaluations	200
Initialization	Initial design size	30
Summary statistic	Center line	Median
Uncertainty band	Percentiles	25th–75th
Acquisition optimization	Raw samples	512
Acquisition optimization	Restarts	4
Acquisition optimization	Max iterations	300
VAE	Latent dimension	8
VAE	Minimum latent dimension	2
VAE	Hidden dimension	64
VAE training	Epochs	10
VAE training	Batch size	32
VAE training	KL weight	10^{-4}
HiBBO	HiPPO memory dimension	10
HiBBO	HiPPO coefficient	1
HiBBO	ODE discretization	bilinear
Hopper	Input dimension	33
Swimmer	Input dimension	16

HiBBO is one of the fastest methods to reach the final performance. NF has a much wider interquartile range, suggesting less consistent behaviour across seeds. HiBBO is comparatively stable: its final interquartile range is approximately $[988.8, 1020.7]$, showing that the HiPPO-regularized latent model can reliably recover high-performing Hopper policies.

On Swimmer, the methods separate more clearly. HiBBO achieves a final median best value of approximately 292.4, outperforming INV, BASE, GPior, METRIC, and REWEIGH in median performance. NF obtains the highest median, around 316.5, but again with a relatively wide interquartile range. Compared with the standard VAE baseline and INV, HiBBO improves both final performance and robustness, suggesting that the HiPPO memory regularisation is especially useful in this lower-dimensional locomotion setting. Overall, the two plots show that HiBBO is consistently competitive: it matches the strongest group on Hopper and clearly outperforms most VAE-based baselines on Swimmer.