MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation

To cite this article: Ryan-Rhys Griffiths *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 015004

View the article online for updates and enhancements.

You may also like

- Calibrated uncertainty estimation for interpretable proton computed tomography image correction using Bayesian deep learning
Yusuke Nomura, Sodai Tanaka, Jeff Wang et al.
- Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks
Jonas Busk, Peter Bjørn Jørgensen, Arghya Bhowmik et al.
- A statistical primer on classical period-finding techniques in astronomy
Naomi Giertych, Ahmed Shaban, Pragya Haravu et al.

This content was downloaded from IP address 208.52.96.42 on 17/01/2025 at 23:10

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 015004 https://doi.org/10.1088/2632-2153/ac298c

MACHINE LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
8 January 2021

REVISED
23 August 2021

ACCEPTED FOR PUBLICATION
23 September 2021

PUBLISHED
24 November 2021

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation

Ryan-Rhys Griffiths[1,*], **Alexander A Aldrick**[1], **Miguel Garcia-Ortegon**[2,3], **Vidhi Lalchand**[2] and **Alpha A Lee**[1,*]

[1] Department of Physics, University of Cambridge, Cambridge, United Kingdom
[2] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[3] Department of Mathematics, University of Cambridge, Cambridge, United Kingdom
[*] Authors to whom any correspondence should be addressed.

E-mail: rrg27@cam.ac.uk **and** aal44@cam.ac.uk

Keywords: Bayesian optimisation, Gaussian processes, heteroscedasticity

Abstract

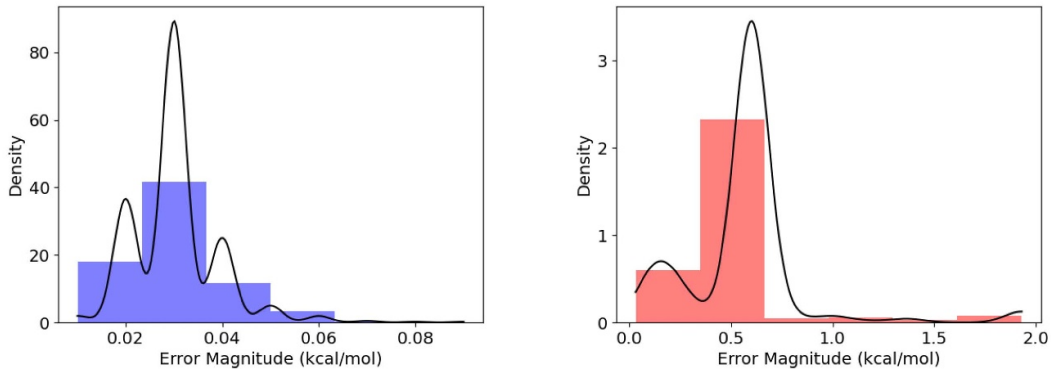
Bayesian optimisation is a sample-efficient search methodology that holds great promise for accelerating drug and materials discovery programs. A frequently-overlooked modelling consideration in Bayesian optimisation strategies however, is the representation of heteroscedastic aleatoric uncertainty. In many practical applications it is desirable to identify inputs with low aleatoric noise, an example of which might be a material composition which displays robust properties in response to a noisy fabrication process. In this paper, we propose a heteroscedastic Bayesian optimisation scheme capable of representing and minimising aleatoric noise across the input space. Our scheme employs a heteroscedastic Gaussian process surrogate model in conjunction with two straightforward adaptations of existing acquisition functions. First, we extend the augmented expected improvement heuristic to the heteroscedastic setting and second, we introduce the aleatoric noise-penalised expected improvement (ANPEI) heuristic. Both methodologies are capable of penalising aleatoric noise in the suggestions. In particular, the ANPEI acquisition yields improved performance relative to homoscedastic Bayesian optimisation and random sampling on toy problems as well as on two real-world scientific datasets. Code is available at: https://github.com/Ryan-Rhys/Heteroscedastic-BO

1. Introduction

Bayesian optimisation is proving to be a highly effective search methodology in areas such as drug discovery [1–3], materials discovery [4–6], chemical reaction optimisation [7–9], robotics [10], sensor placement [11], tissue engineering [12] and genetics [13]. Heteroscedastic aleatoric noise however, is rarely accounted for in these settings despite being an important consideration for real-world applications. Aleatoric uncertainty refers to uncertainty inherent in the observations (measurement noise) [14]. In contrast, epistemic uncertainty corresponds to model uncertainty and may be explained away given sufficient data. Heteroscedastic aleatoric noise refers to aleatoric noise which varies across the input domain and is a prevalent feature of many scientific datasets; perhaps surprisingly not only experimental datasets, but also datasets where properties are predicted computationally. One such source of heteroscedasticity in the computational case might be situations in which the accuracy of first-principles calculations deteriorate as a function of the chemical complexity of the molecule being studied [15].

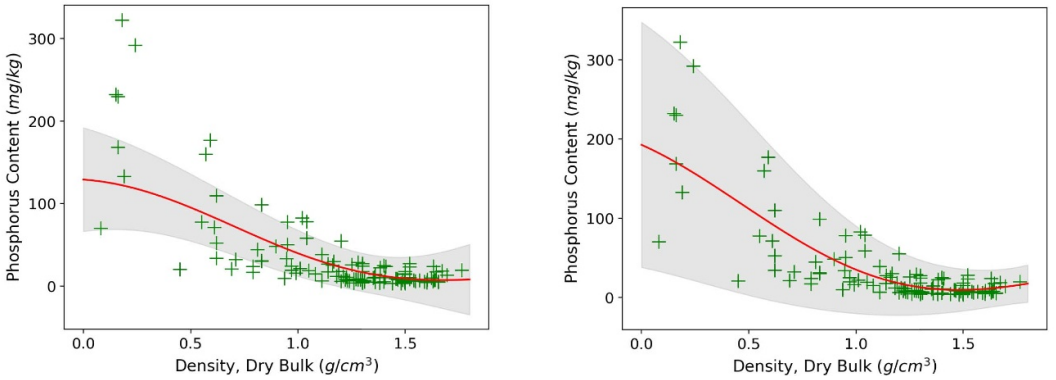
In figure 1 we illustrate real-world sources of heteroscedasticity using the FreeSolv dataset of [16]. The consequences of misrepresenting heteroscedastic noise as being homoscedastic, i.e. constant across the input domain, are illustrated using a second dataset [17] in figure 2. The homoscedastic model can underestimate noise in certain regions of the input space which in turn could induce a Bayesian optimisation scheme to suggest values possessing large aleatoric noise. In an application such as high-throughput virtual screening [18] the cost of misrepresenting noise during the screening process could lead to a substantial loss of time in

© 2021 The Author(s). Published by IOP Publishing Ltd



(a) Density plot of computational errors (b) Density plot of experimental errors

Figure 1. (a) The density histogram of computational errors (kcal/mol) for the FreeSolv hydration energy dataset ([16]). The computational errors in the hydration free energy arise from systematic errors in the force field used in alchemical free energy calculations based on classical molecular dynamics (MD) simulations. (b) A similar density histogram for the experimental errors where the source of uncertainty stems from the instrumentation used to obtain the measurement. The histograms are overlaid with kernel density estimates.



(a) Homoscedastic GP Fit (b) Heteroscedastic GP Fit

Figure 2. Comparison of homoscedastic and heteroscedastic GP fits to the soil phosphorus fraction dataset [17].

material fabrication [19]. In this paper we present a heteroscedastic Bayesian optimisation algorithm capable of both representing and minimising aleatoric noise in its suggestions. Our contributions are:

(a) The introduction of a novel combination of surrogate model and acquisition function designed to minimise heteroscedastic aleatoric uncertainty.
(b) A demonstration of our scheme's ability to outperform naive schemes based on homoscedastic Bayesian optimisation and random sampling on toy problems as well as two real-world scientific datasets.
(c) The provision of an open-source implementation.

The paper is structured as follows: section 2 introduces related work on heteroscedastic Bayesian optimisation. Section 3 provides background on Bayesian optimisation and homoscedastic Gaussian process (GP) surrogate models. Section 4 provides background on the heteroscedastic GP surrogate model used in this work and introduces the novel heteroscedastic augmented expected improvement (HAEI) and aleatoric noise-penalised expected improvement (ANPEI) acquisitions functions. Section 5 considers experiments on synthetic and scientific datasets possessing heteroscedastic noise where the goal is to be robust to, i.e. minimise, aleatoric noise in the suggestions. Section 6 presents an ablation study on noiseless tasks as well as tasks with homoscedastic and heteroscedastic noise in order to determine whether there is a detrimental effect to using a heteroscedastic surrogate when the noise properties of the problem are *a priori* unknown. Section 7 concludes with some limitations of the approach presented as well as fruitful sources for future work.

2. Related work

The most similar work to our own is that of [20] where experiments are reported on a heteroscedastic Branin-Hoo toy function using the variational heteroscedastic GP approach of [21]. This work defines and optimises a robustness index, making a compelling case for penalisation of aleatoric noise in real-world Bayesian optimisation problems. A modification to expected improvement (EI), expected risk improvement is introduced in [22] and is applied to problems in robotics where robustness to aleatoric noise is desirable. In this framework however, the relative weights of performance and robustness cannot be tuned [20]. References [23, 24] implement heteroscedastic Bayesian optimisation but do not introduce an acquisition function that penalises aleatoric noise [25, 26] consider the related problem of safe Bayesian optimisation through implementing constraints in parameter space. In this instance, the goal of the algorithm is to enforce a performance threshold for each evaluation of the black-box function. Recently, the winners of the 2020 NeurIPS Black-Box Optimisation Competition applied non-linear output transformations in their solution to tackle heteroscedasticity. The authors however are not interested in explicitly penalising aleatoric noise in this case. In terms of acquisition functions, [27, 28] propose principled approaches to handling aleatoric noise in the homoscedastic setting that could be extended to the heteroscedastic setting. Our primary focus in this work however, is to highlight that heteroscedasticity in the surrogate model is beneficial and so an examination of a subset of acquisition functions is sufficient for this purpose. We take the opportunity here to note earlier unpublished workshop versions of this paper which consider the same problem [29, 30].

3. Background

3.1. Bayesian optimisation

Bayesian optimisation [31, 32, 35] solves the global optimisation problem defined as

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

where $\mathbf{x}^*$ is the global optimiser of a black-box function $f\colon \mathcal{X} \to \mathcal{Y}$. $\mathcal{X}$ is the design space and is typically a compact subset of $\mathbb{R}^d$. What makes this optimisation problem practically relevant in applications are the following properties:

(a) Black-Box Objective: We do not have the analytic form of f. We can however evaluate f pointwise anywhere in the design space $\mathcal{X}$.
(b) Expensive Evaluations: Choosing an input location $\mathbf{x}$ and evaluating $f(\mathbf{x})$ takes a very long time.
(c) Noise: The evaluation of a given $\mathbf{x}$ is a noisy process. In addition, this noise may vary across $\mathcal{X}$, making the underlying process heteroscedastic.

We have a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$ consisting of observations of the black-box function f and fit a probabilistic surrogate model to these datapoints. We then leverage the predictive mean as well as the uncertainty estimates of the surrogate model to guide the acquisition of the next data point $\boldsymbol{x}_{n+1}$ according to a heuristic known as an acquisition function. In Bayesian optimisation, exact GPs are the most popular choice of surrogate model because of their ability to represent posterior uncertainty without resorting to approximate Bayesian inference.

3.2. Gaussian processes

In the terminology of stochastic processes we may formally define a GP as follows:

Definition 1. A GP [33] is a collection of random variables, any finite number of which have a joint Gaussian distribution.

GPs can be used to set a prior over functions in Bayesian modelling applications. In this setting, the random variables consist of function values $f(\mathbf{x})$ at different locations $\mathbf{x}$ within the design space. The GP is characterised by a mean function

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$

and a covariance function

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x} - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))].$$

The process is written as follows

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big).$$

In our experiments, the prior mean function will be set to the empirical mean of the data. The covariance function or kernel computes the pairwise covariance between two random variables (function values). The covariance between a pair of output values $f(\mathbf{x})$ and $f(\mathbf{x}')$ is a function of an input pair $\mathbf{x}$ and $\mathbf{x}'$. As such, the kernel encodes smoothness assumptions about the latent function being modelled. The most widely-utilised kernel is the squared exponential (SE) kernel

$$k_{\mathrm{SQE}}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma_f^2 \cdot \exp\Big(\frac{-\|\boldsymbol{x} - \boldsymbol{x'}\|^2}{2\ell^2}\Big) \tag{1}$$

where σ_f^2 is the signal amplitude hyperparameter (vertical lengthscale) and ℓ is the (horizontal) lengthscale hyperparameter. Although equation (1) is written with a single lengthscale shared across dimensions, for multidimensional input spaces we optimise a lengthscale per dimension. For consistency, we use the SE kernel in all experiments reported in the main paper. In appendix C we compare the performance of different kernels on a set of synthetic optimisation functions. For a more detailed introduction to GPs the reader is referred to [33].

4. Heteroscedastic Bayesian optimisation

We wish to perform Bayesian optimisation whilst minimising input-dependent aleatoric noise. In order to represent input-dependent aleatoric noise, a heteroscedastic surrogate model is required.

4.1. The most likely heteroscedastic Gaussian process

We adopt the most likely heteroscedastic Gaussian process (MLHGP) approach of [34], and for consistency, we use the same notation as the source work in our presentation. We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}_{i=1}^n$ in which the target values t_i have been generated according to $t_i = f(\mathbf{x}_i) + \epsilon_i$. We assume independent Gaussian noise terms $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ with variances given by $\sigma_i^2 = r(\mathbf{x}_i)$. In the heteroscedastic setting r is typically a non-constant function over the input domain $\mathbf{x}$. In order to perform Bayesian optimisation, we wish to model the predictive distribution $P(\mathbf{t}^* \mid \mathbf{x}_1^*, \ldots, \mathbf{x}_q^*)$ at the query points $\mathbf{x}_1^*, \ldots, \mathbf{x}_q^*$. Placing a GP prior on f and taking $r(\mathbf{x})$ as the assumed noise function, the predictive distribution is multivariate Gaussian $\mathcal{N}(\mu^*, \Sigma^*)$ with mean

$$\mu^* = E[\mathbf{t}^*] = K^*(K+R)^{-1}\mathbf{t}$$

and covariance matrix

$$\Sigma^* = \mathrm{var}[\mathbf{t}^*] = K^{**} + R^* - K^*(K+R)^{-1}K^{*T},$$

where $K \in \mathbb{R}^{n\times n}$, $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $K^* \in \mathbb{R}^{q\times n}$, $K_{ij}^* = k(\mathbf{x}_i^*, \mathbf{x}_j)$, $K^{**} \in \mathbb{R}^{q\times q}$, $K_{ij}^{**} = k(\mathbf{x}_i^*, \mathbf{x}_j^*)$, $\mathbf{t} = (t_1, t_2, \ldots, t_n)^T$, $R = \mathrm{diag}(\mathbf{r})$with $\mathbf{r} = (r(\mathbf{x}_1), r(\mathbf{x}_2), \ldots, r(\mathbf{x}_n))^T$, and $R^* = \mathrm{diag}(\mathbf{r}^*)$with $\mathbf{r}^* = (r(\mathbf{x}_1^*), r(\mathbf{x}_2^*), \ldots, r(\mathbf{x}_q^*))^T$.

The MLHGP algorithm [34] executes the following steps:

(a) Estimate a homoscedastic GP, G_1 on the dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}_{i=1}^n$.
(b) Given G_1, we estimate the empirical noise levels for the training data using $z_i = \log(\mathrm{var}[t_i, G_1(\mathbf{x}_i, \mathcal{D})])$ where $\mathrm{var}[t_i, G_1(\mathbf{x}_i, \mathcal{D})] \approx \frac{1}{s}\sum_j^s 0.5\,(t_i - t_i^j)^2$ with t_i^j a sample from the predictive distribution induced by the GP at $\mathbf{x}_i$, forming a new dataset $\mathcal{D}' = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$.
(c) Estimate a second GP, G_2 on $\mathcal{D}'$.
(d) Estimate a combined GP, G_3 on $\mathcal{D}$ using G_2 to predict the logarithmic noise levels r_i.
(e) If not converged, set G_3 to G_1 and repeat.

In essence, the defining characteristic of the MLHGP approach is that G_1 learns the latent function and G_2 learns the noise function.

4.2. Bayesian optimisation with aleatoric noise penalisation

Our heteroscedastic Bayesian optimisation problem may be framed as

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\chi} h(\boldsymbol{x}),$$

where the black-box objective h, to be minimised has the form

$$h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{x}),$$

where $f(\boldsymbol{x})$ is the black-box function of the principal objective i.e. the objective corresponding to classical Bayesian optimisation where noise is not optimised, and $g(\boldsymbol{x})$ is the latent heteroscedastic noise function which governs the magnitude of the noise at a given input location $\boldsymbol{x}$. α is a parameter chosen, for the purposes of evaluation, by a domain expert that trades off the weight of the principal objective relative to the noise objective. It is worth noting that α is a parameter that affects only the evaluation of an algorithm and not the execution. The evaluation criteria however, will dictate the optimal hyperparameters of the acquisition function.

4.3. Heteroscedastic acquisition functions

We investigate extensions of the EI [35] acquisition criterion, the form of which may be written in terms of the targets t and the incumbent best objective function value, η, found so far as

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big] = \int_{-\infty}^{\infty} (\eta - t)_+ p(t|\boldsymbol{x})\, dt$$

where $p(t|\boldsymbol{x})$ is the posterior predictive marginal density of the objective function evaluated at $\boldsymbol{x}$. $(\eta - t)_+ \equiv \max(0, \eta - t)$ is the improvement over the incumbent best objective function value η. Evaluations of the objective are noisy in all of the problems we consider and so we use EI with plug-in [36], the plug-in value being the GP predictive mean [37].

We propose two extensions to the EI criterion. The first is an extension of the augmented expected improvement (AEI) criterion

$$\mathrm{AEI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big]\left(1 - \frac{\sigma_n}{\sqrt{\mathrm{var}[t] + \sigma_n^2}}\right),$$

of [38] where σ_n is the fixed aleatoric noise level. AEI is introduced as a heuristic for the optimisation of noisy functions. EI is recovered in the case that $\sigma_n^2 = 0$ and in the case that $\sigma_n^2 > 0$ AEI operates as a rescaling of the EI acquisition function, penalising test locations where the GP predictive variance is small relative to the fixed noise level σ_n^2. We extend AEI to the heteroscedastic setting by exchanging the fixed aleatoric noise level with the input-dependent one:

$$\mathrm{HAEI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big]\left(1 - \frac{\gamma\sqrt{r(\boldsymbol{x})}}{\sqrt{\mathrm{var}[t] + \gamma^2 r(\boldsymbol{x})}}\right), \tag{2}$$

where $r(\boldsymbol{x})$ is the predicted aleatoric uncertainty at input $\boldsymbol{x}$ under the MLHGP and $\mathrm{var}[t]$ is the predictive variance of the MLHGP at input $\boldsymbol{x}$. γ in this instance is defined to be a positive penalty parameter for regions with high aleatoric noise.

Proposition 1 (Limit of Large Epistemic Uncertainty). *The HAEI acquisition function reduces to EI when the ratio of epistemic uncertainty to aleatoric uncertainty is much greater than γ^2.*

Proof. Let $k = \frac{\mathrm{var}[t]}{r(\boldsymbol{x})}$ denote the ratio of epistemic to aleatoric uncertainty at an arbitrary input location $\boldsymbol{x}$. Dividing the numerator and the denominator of the second term in the second factor of equation (2) by $\sqrt{r(\boldsymbol{x})}$ yields

$$\mathrm{HAEI}(\boldsymbol{x}) = \mathrm{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k + \gamma^2}}\right). \tag{3}$$

Taking the limit analytically as k tends to infinity and assuming finite γ

$$\lim_{k\to\infty} \mathrm{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k + \gamma^2}}\right) = \mathrm{EI}(\boldsymbol{x}),$$

recovers the EI acquisition.

□

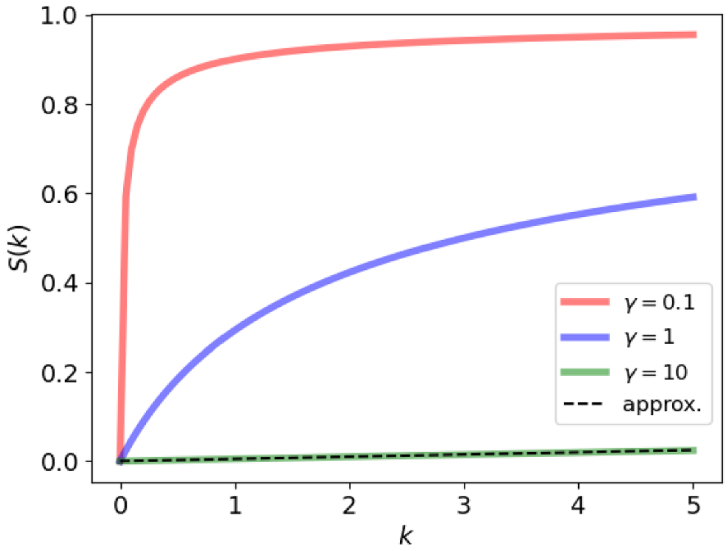


Figure 3. The HAEI scaling factor $S(k)$, now written as a function of k for different values of γ. When k, the ratio of epistemic to aleatoric uncertainty is small, the scaling factor goes to zero to reflect the penalty for regions of high aleatoric uncertainty. γ controls the decay rate of this penalty. Also shown is the linear approximation to the scaling factor for $\gamma = 10$.

Proposition 2 (Limit of Large Aleatoric Uncertainty). *The HAEI acquisition function goes to zero as the ratio of epistemic uncertainty to aleatoric uncertainty goes to zero.*

Proof. Taking the limit as k tends to zero in equation (3) yields

$$\lim_{k\to 0} \mathrm{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k+\gamma^2}}\right) = 0.$$

□

Remark. In the limit of large aleatoric uncertainty there is an approximation that is linear in k for the HAEI scaling factor.

Letting $S(k) = 1 - \frac{\gamma}{\sqrt{k+\gamma^2}}$ such that $\mathrm{HAEI} = \mathrm{EI}(\boldsymbol{x})S(k)$, consider the MacLaurin expansion of $S(k)$,

$$S(k) = S(0) + S'(0)k + \frac{S''(0)}{2!}k^2 + \frac{S'''(0)}{3!}k^3 + \ldots.$$

Dropping terms of $O(k^2)$ and higher we obtain

$$S(k) \approx \frac{k}{2\gamma^2}.$$

This approximation may be used when k is small relative to γ and could provide guidance in setting the γ parameter if prior knowledge about k and the desired trade-off between the principal and noise objectives is available. In figure 3 we provide insight into the effect that different values of γ will have on the scaling factor $S(k)$.

In addition to HAEI, we propose a simple modification to EI that explicitly penalises regions of the input space with large aleatoric noise. We call this acquisition function ANPEI and denote it

$$\mathrm{ANPEI} = \beta\mathrm{EI}(\boldsymbol{x}) - (1-\beta)\sqrt{r(\boldsymbol{x})}, \tag{4}$$

where β is a scalarisation constant. In the multiobjective optimisation setting a particular value of β will correspond to a point on the Pareto frontier. We showcase the advantages of both HAEI and ANPEI acquisition functions in conjunction with the MLHGP surrogate model in section 5.

5. Experiments on robustness to aleatoric uncertainty

5.1. Implementation

Experiments were run using a custom NumPy [39] implementation of GP regression and MLHGP regression. All code to reproduce the experiments is available at https://github.com/Ryan-Rhys/Heteroscedastic-BO. The SE kernel was chosen as the covariance function for both the homoscedastic GP as well as G_1 and G_2 of the MLHGP. Across all datasets, the lengthscales, ℓ, of the homoscedastic GP were initialised to 1.0 for each input dimension. The signal amplitude σ_f^2 was initialised to a value of 1.0. The lengthscale, ℓ, of G_2 of the MLHGP [34] was initialised to 1.0, the initial noise level of G_2 was set to 1.0. The EM-like procedure required to train the MLHGP was run for ten iterations and the sample size required to construct the variance estimator producing the auxiliary dataset was 100. All standard error confidence bands are computed using 50 independent random seed initialisations. Hyperparameter values, including the noise level of the homoscedastic GP, were obtained by optimising the marginal likelihood using the scipy implementation of the L-BFGS-B optimiser [40], taking the best of 20 random restarts. The objective function is

$$h(x) = \alpha f(x) - (1-\alpha)g(x)$$

for the one-dimensional sin wave experiment which is a maximisation problem and as such has a subtractive penalty for regions of large noise. For the remaining experiments, which are minimisation problems, the objective is

$$h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{x}). \tag{5}$$

The sin wave and Branin-Hoo tasks are initialised with 25 and 100 data points respectively drawn uniformly at random within the bounds of the design space. The soil and FreeSolv experiments are initialised with 36 and 129 data points respectively drawn uniformly at random from the datasets. α is set to 0.5 for all experiments while β is set to 0.5, $\frac{1}{11}$, 0.5 and 0.5 for the sin, Branin-Hoo, soil and FreeSolv experiments. γ is set to 1, 500, 1 and 1 for the sin, Branin-Hoo, soil and FreeSolv experiments. We run five acquisition functions in all experiments: random sampling, homoscedastic EI, AEI, HAEI and ANPEI. Homoscedastic EI is included as a baseline to demonstrate the difference consideration of aleatoric noise yields in the optimisation of the objective. AEI is included to demonstrate the difference consideration of heteroscedastic aleatoric noise yields and random sampling is included as a baseline as it is known to be highly competitive with Bayesian optimisation in noisy settings.

5.2. Heteroscedastic sin wave function

The objective function has the form

$$h(x) = f(x) - g(x)$$

where $f(x) = \sin(x) + 0.2(x) + 3$ and $g(x) = 0.5(x)$. In this instance α from section 5.1 has a setting of 0.5 but we omit it explicitly as the objectives have equal weight. Over the course of the experiment samples

$$y_i = f(x_i) + g(x_i)\epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$$

are observed. The problem setup is depicted in figures 4 and 5. The Bayesian optimisation problem is constructed such that the first maximum in figure 4(a) is to be preferred as samples from this region of the input space will have low aleatoric noise. The black-box objective in figure 4(c) illustrates this trade-off. In figure 6 we compare the performance of all surrogate model/acquisition function combinations. We observe the low aleatoric noise-seeking behaviour of HAEI and ANPEI on $g(x)$ as well as their ability to optimise the composite objective $h(x)$.

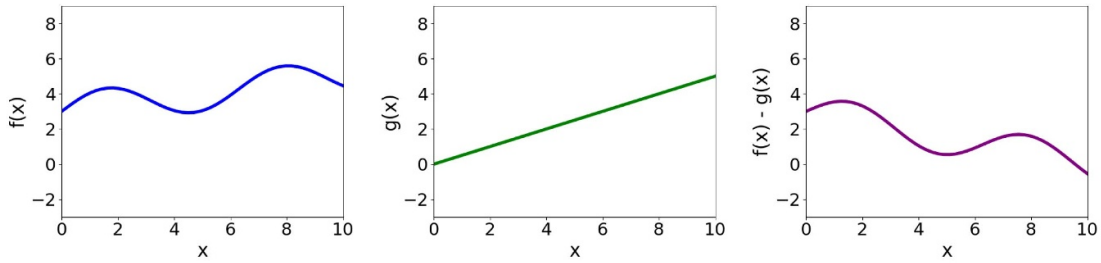
5.3. Heteroscedastic Branin-Hoo function

In the second experiment we consider the objective

$$h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$$

with an additive penalty because the task is a minimisation problem and an α setting of 0.5 for equal-weight objectives.

$$f(\boldsymbol{x}) = \frac{1}{51.95}\left[\left(\bar{x}_2 - \frac{5.1\,\bar{x}_1^2}{4\,\pi^2} + \frac{5\,\bar{x}_1}{\pi} - 6\right)^2 + \left(10 - \frac{10}{8\,\pi}\right)\cos(\bar{x}_1) - 44.81\right] \tag{6}$$



(a) Latent Function (b) Noise Function (c) Objective Function

Figure 4. Illustrative toy problem. The latent function in (a) is corrupted with heteroscedastic Gaussian noise according to the function in (b) where $g(x)$ is a constant multiplier of a sample from a standard Gaussian. The combined objective is given in (c) and is obtained by subtracting the noise function from the latent function.

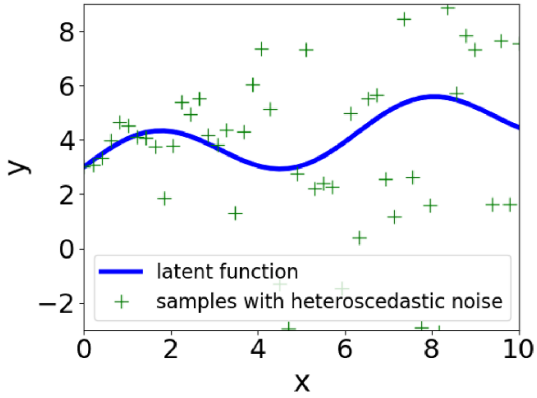
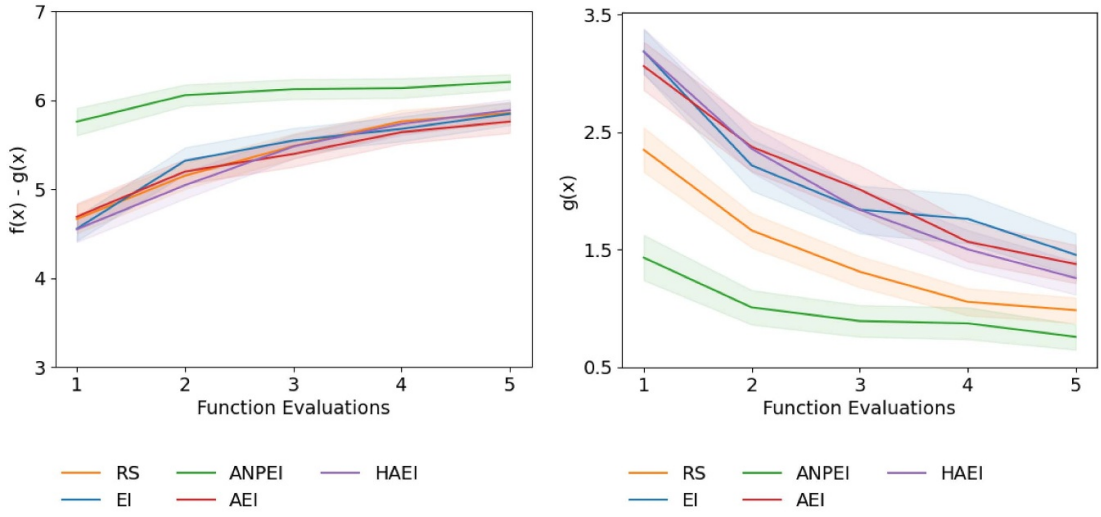
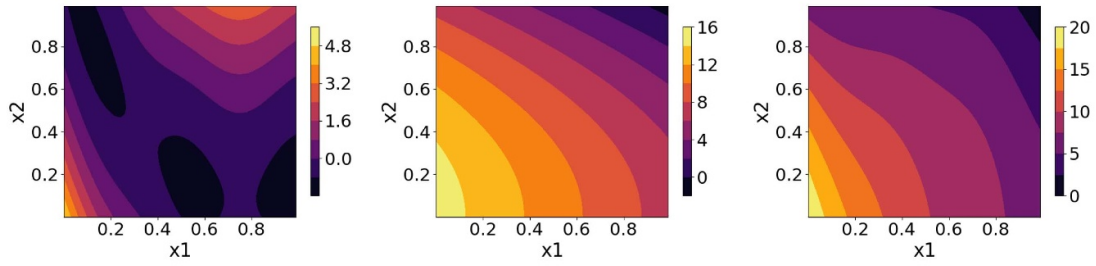


Figure 5. Noisy samples $y_i = f(x_i) + g(x_i)\epsilon$ from the heteroscedastic sin wave function.



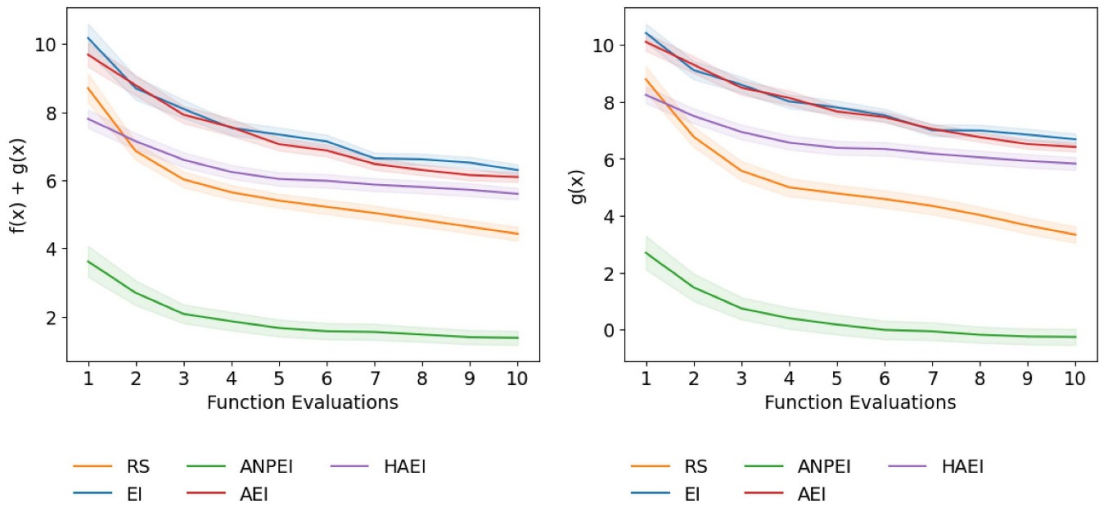
(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure 6. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the sin wave problem. (a) Shows the optimisation of $h(x) = f(x) - g(x)$ (higher is better) whereas (b) shows the values $g(x)$ obtained over the course of the optimisation of $h(x)$. This latter plot demonstrates the propensity of ANPEI to seek low aleatoric noise solutions.



(a) Latent Function (b) Non-linear Noise Function (c) Objective Function

Figure 7. Branin-Hoo Optimisation Problem. The latent function in (a) is corrupted by heteroscedastic Gaussian noise function according to the function in (b). The combined objective function is given in (c) and is obtained by summing the functions in (a) and (b). The sum is required to penalise regions of large aleatoric noise because the objective is being minimised.



(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure 8. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the Branin-Hoo problem. (a) Shows the optimisation of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ (lower is better) whereas (b) shows the values $g(\boldsymbol{x})$ obtained over the course of the optimisation of $h(\boldsymbol{x})$.

with $\bar{x}_1 = 15x_1 - 5$, $\bar{x}_2 = 15x_2$ and $\boldsymbol{x} = (x_1, x_2)$ is the standardised Branin-Hoo function introduced in [36]. The noise function $g(\boldsymbol{x})$ is in this instance

$$g(\boldsymbol{x}) = 15 - 8x_1 + 8x_2^2. \tag{7}$$

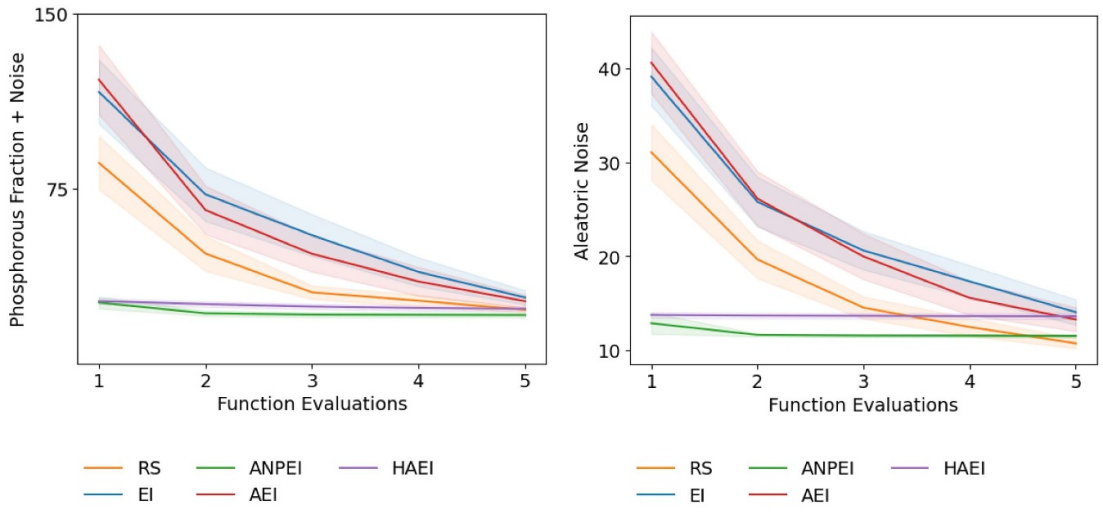
Samples are again generated according to

$$y_i = f(\boldsymbol{x}_i) + g(\boldsymbol{x}_i)\epsilon, \quad \epsilon \sim \mathcal{N}(0,1).$$

The problem setup is shown in figure 7 and the performance of all surrogate model/acquisition function pairs is depicted in figure 8. The gulf in performance between the heteroscedastic and homoscedastic surrogate models is more pronounced in this case because the noise function is more severe relative to the sin wave problem.

5.4. Soil phosphorus fraction optimisation

In this experiment we consider the optimisation of the phosphorus fraction of soil. Soil phosphorus is an essential nutrient for plant growth and is widely used as a fertiliser in agriculture. While the amount of arable land worldwide is declining, global population is expanding concomitantly with food demand. As such, understanding the availability of plant nutrients that increase crop yield is a topic worthy of attention. To this end, [17] have curated a dataset on soil phosphorus, relating phosphorus content to variables such as soil particle size, total nitrogen, organic carbon and bulk density. We choose to study the relationship between



(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure 9. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the soil phosphorus fraction optimisation problem. (a) Shows the optimisation of $h(x) = f(x) + g(x)$ (lower is better) where x is the dry bulk density of the soil. (b) Shows the values $g(x)$ obtained over the course of the optimisation of $h(x)$.

bulk soil density and the phosphorus fraction, the goal being to minimise the phosphorus content of soil subject to heteroscedastic noise. In lieu of performing a formal test for heteroscedasticity, we provide evidence that there is heteroscedasticity in the dataset by comparing the fits of a homoscedastic GP and the MLHGP in figure 2 and provide a predictive performance comparison based on negative log predictive density values in appendix A.

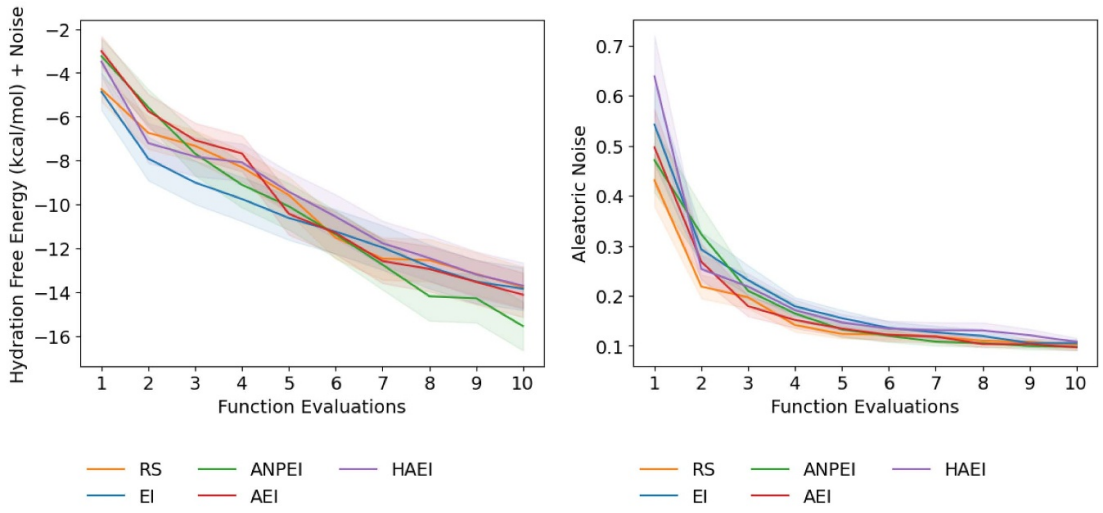
In this problem, we do not have access to a continuous-valued black-box function or a ground truth noise function. As such, the surrogate models were initialised with a subset of the data and the query locations selected by Bayesian optimisation were mapped to the closest datapoints in the heldout data. The following kernel smoothing procedure was used to generate pseudo ground-truth noise values:

(a) Fit a homoscedastic GP to the full dataset.
(b) At each point x_i, compute the corresponding squared error $s_i^2 = (y_i - \mu(x_i))^2$.
(c) Estimate variances by computing a moving average of the squared errors, where the relative weight of each s_i^2 was assigned with a Gaussian kernel.

The performances of heteroscedastic and homoscedastic Bayesian optimisation are compared in figure 9. Given that regions of low phosphorus fraction coincide with regions of small aleatoric noise, we apply an α value of $\frac{1}{6}$ to the composite objective $h(x)$ to admit a finer granularity for distinguishing between degrees of low aleatoric noise in the solutions.

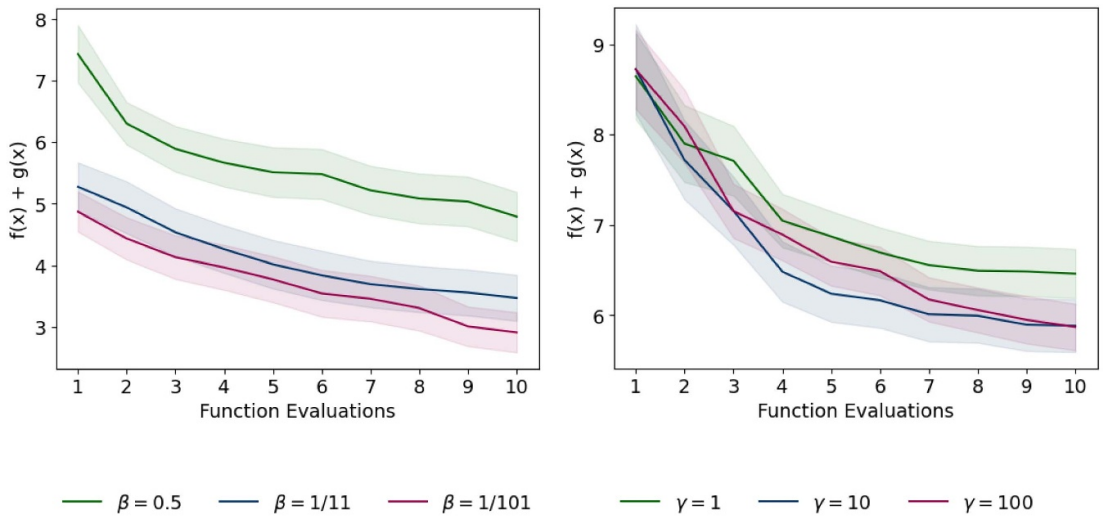
5.5. Molecular hydration free energy optimisation

We perform a retrospective virtual screening experiment with the aim of identifying molecules with favourable hydration free energy, a property important in determining the binding affinity of a drug candidate. Experiments were performed with an initialisation of 129 out of the 642 molecules in the FreeSolv dataset [16, 41] over ten iterations of data collection. Unlike the soil phosphorus fraction dataset, ground truth measurement error (aleatoric noise $g(\mathbf{x})$) values are available for the FreeSolv dataset. The remaining 513 molecules were reserved as a heldout set where at each iteration of data collection one of the heldout molecules was selected. Chemical fragments computed using RDKit [42] were used as the molecular representation based on the fact that these global features, unlike local Morgan fingerprints, act as good predictors of the hydration free energy. The fragment features were projected down to 14 components using principal component analysis, retaining more than 90% of the variance on average across random trials. The results are shown in figure 10. Compared to previous experiments, the noise is smaller in this instance relative to the magnitude of the hydration free energy (Signal-to-noise ratio of approximately 10) and as such the heteroscedastic modelling problem is more difficult, leading to only very marginal gains in obtaining low noise solutions. While ANPEI obtains the lowest objective function value over the Bayesian optimisation trace, the results are unlikely to be statistically significant according to the standard error bands.



(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure 10. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the FreeSolv hydration free energy optimisation problem. (a) Shows the optimisation of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ (lower is better) where $\boldsymbol{x}$ is the fragment set of molecular descriptors, $f(\boldsymbol{x})$ is the hydration free energy and $g(\boldsymbol{x})$ is the aleatoric noise. (b) Shows the values $g(\boldsymbol{x})$ obtained over the course of the optimisation of $h(\boldsymbol{x})$.



(a) ANPEI (b) HAEI

Figure 11. Performance of ANPEI and HAEI plotted for different values of the β and γ hyperparameters respectively. *Smaller* values of β encourage avoidance of regions of high aleatoric noise whilst *larger* values of γ encourage avoidance of regions of high aleatoric noise.

5.6. Heteroscedastic acquisition function hyperparameters

The β hyperparameter of ANPEI in equation (4) and the γ hyperparameter of HAEI in equation (3) are designed to modulate the avoidance of aleatoric noise in the acquisitions. In figure 11 we offer some intuition as to the effect of various settings of β and γ by examining the heteroscedastic Branin-Hoo function introduced in section 5.3. The results demonstrate that the performance of the algorithms is strongly dependent on the setting of the β hyperparameter for ANPEI whereas γ is less influential on the performance of HAEI. It is worth noting in figure 11(b) that if too large a value of γ is chosen the principal objective $f(\mathbf{x})$ may be compromised through overly aggressive avoidance of aleatoric noise. In practice choosing the value of β in line with the value of the evaluation criterion parameter α in equation (5) is likely to be a sensible approach i.e. if the noise objective is more important relative to the principal objective by a factor of 10 then the value of β should be $\frac{1}{11}$.

5.7. Robustness experiments summary

The experiments of this section provide strong evidence that modelling heteroscedasticity in Bayesian optimisation is a useful approach for problems in which there is a strong degree of aleatoric noise present. The ANPEI acquisition tends to outperform HAEI on the majority of the tasks where there is a small degree of aleatoric noise whilst the acquisitions are more evenly matched when the extent of the aleatoric noise is high. The outstanding questions for these methods however, is how well they perform on tasks where heteroscedastic noise is not present. Such a situation may easily arise for real-world problems where the noise properties of the tasks are a prior unknown and as such, it is important to ascertain whether there is a deleterious effect on performance in noiseless and homoscedastic noise settings.

6. Ablation study on noiseless, homoscedastic noise and heteroscedastic noise tasks

In this section we perform an ablation study where components of the ablation constitute different noise properties. We examine the noiseless case as a base task before adding first a homoscedastic noise component and second, a heteroscedastic noise component. Additionally, we examine the effect of the size of the initialisation grid on performance in the heteroscedastic noise tasks.

6.1. Ablation

The ablation study makes use of three synthetic optimisation functions: The Branin-Hoo function, the Hosaki function and the Goldstein-Price function. The form of the Branin-Hoo function is the same standardised Branin-Hoo function introduced in equation (6) with heteroscedastic noise function given in equation (7). The Hosaki function, defined on the domain $x_1, x_2 \in [0, 5]$, is

$$\text{Hosaki}(x_1, x_2) = \left(1 - 8x_1 + 7{x_1}^2 - \frac{7}{3}{x_1}^3 + \frac{1}{4}{x_1}^4\right){x_2}^2 \exp(-x_2).$$

To facilitate the GP fit, the Hosaki function is subsequently standardised by its mean (0.817) and standard deviation (0.573). The noise function is

$$g_{\text{Hosaki}}(x_1, x_2) = 50 \cdot \frac{1}{(x_1 - 3.5)^2 + 2.5} \cdot \frac{1}{(x_2 - 2)^2 + 2.5}. \tag{8}$$

The logarithmic Goldstein-Price function [36] is

$$\text{G-P}(x_1, x_2) = \frac{1}{2.427}\left[\log\left([1 + (\bar{x}_1 + \bar{x}_2 + 1)^2(19 - 14\bar{x}_1 + 3\bar{x}_1^2 - 14\bar{x}_2 + 6\bar{x}_1\bar{x}_2 + 3\bar{x}_2^2)] \right. \tag{9}$$

$$\left. [30 + (2\bar{x}_1 - 3\bar{x}_2)^2(18 - 32\bar{x}_1 + 12\bar{x}_1^2 + 48\bar{x}_2 - 36\bar{x}_1\bar{x}_2 + 27\bar{x}_2^2)]\right) - 8.693\right] \tag{10}$$

where $\bar{x}_1 = 4x_1 - 2$ and $\bar{x}_2 = 4x_2 - 2$. The Goldstein-Price noise function is

$$g_{\text{G-P}}(x_1, x_2) = \frac{3}{2} \cdot \frac{1}{(x_1 - 0.5)^2 + 0.2} \cdot \frac{1}{(x_2 - 0.3)^2 + 0.3}. \tag{11}$$

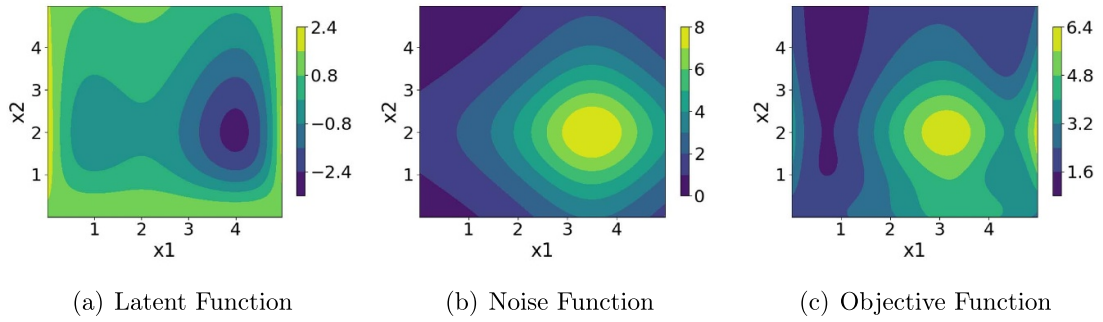
For clarity, only the results of the Hosaki function are presented in the main paper with the Branin-Hoo and Goldstein-Price results presented in B. The Hosaki function is visualised in figure 12. The value of β for ANPEI is set to 0.5 and the value of γ is set to 500 for all Hosaki function experiments.

6.1.1. Noiseless case

In this case, the synthetic functions do not possess any observation noise and the optimisation function corresponds to the situation in figure 12(a). Nine points sampled uniformly at random are used for initialisation and the results are displayed in figure 13. As expected, all Bayesian optimisation methods outperform random search in the noiseless case. In this example it is unclear as to whether heteroscedastic Bayesian optimisation methods are detrimental as HAEI performs best whereas ANPEI performs worst.

6.1.2. Homoscedastic noise case

In this case the functions are subject to homoscedastic noise of the form 25ϵ where epsilon is noise sampled from a standard Gaussian $\mathcal{N}(0, 1)$. The GP surrogates are again initialised with nine points. The results are



(a) Latent Function (b) Noise Function (c) Objective Function

Figure 12. (a) The latent Hosaki Function $f(\mathbf{x})$ together with (b) its heteroscedastic noise function $g(\mathbf{x})$ and (c) the objective function $f(\mathbf{x}) + g(\mathbf{x})$.

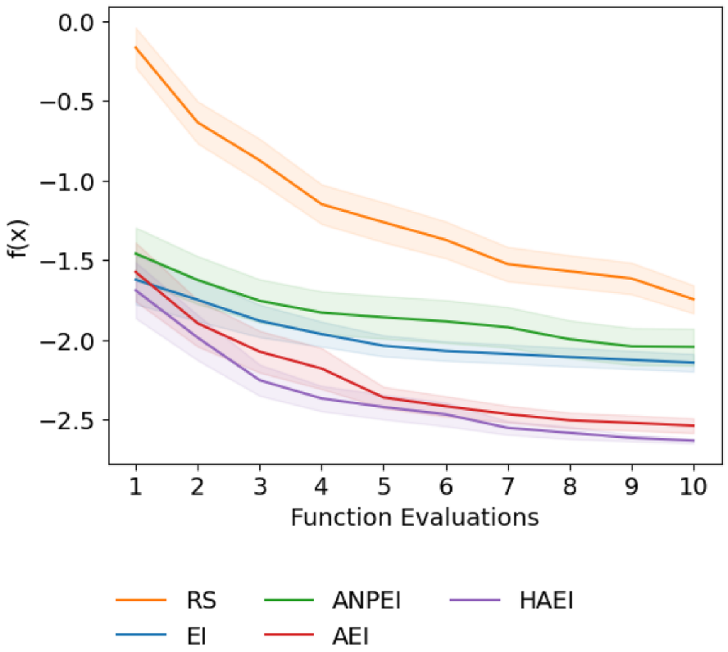


Figure 13. Hosaki function noiseless case. All Bayesian optimisation methods outperform random search. HAEI performs best and ANPEI performs worst.

displayed in figure 14. The Bayesian optimisation methods perform worse in the homoscedastic noise case relative to the noiseless case although the rank order of the methods mirrors that of the noiseless case.

6.1.3. Heteroscedastic noise

In the heteroscedastic noise case the Hosaki function is subject to the noise function given in equation (8) and visualised in figure 12. 144 points were used to initialise the GP surrogates. The results are shown in figure 15. In this instance, given that the extent of heteroscedastic noise is very strong (relative to the homoscedastic noise case), random search is highly competitive with the Bayesian optimisation methods. ANPEI however, is the best-performing algorithm. The large number of initialisation points chosen for this experiment reflects one limitation of the heteroscedastic surrogate approach; for the MLHGP to effectively learn a decomposition of the function into signal and noise components it needs access to more samples. As such, this merits an investigation into the effect of the number of samples on the performance of the heteroscedastic acquisitions.

6.2. Effect of initialisation set size

The effect of the size of the initialisation set on the heteroscedastic Branin-Hoo task is investigated in figure 16. The value of β used for ANPEI is $\frac{1}{11}$ and the value of γ used for HAEI is 500. The performance of the heteroscedastic acquisitions ANPEI and HAEI is observed to improve as the size of the initialisation set increases. In contrast, the homoscedastic methods EI and AEI do not improve on obtaining access to more samples as they are unable to model the heteroscedastic noise component of the task.

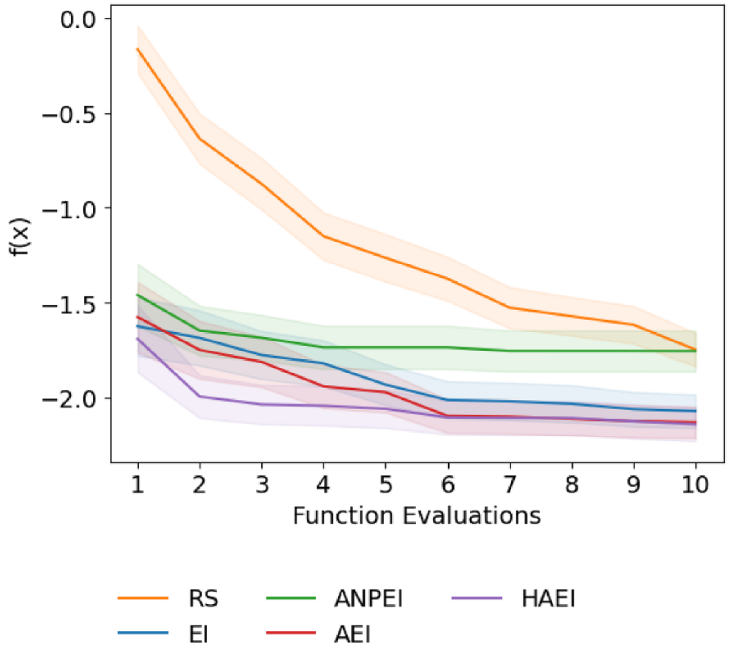
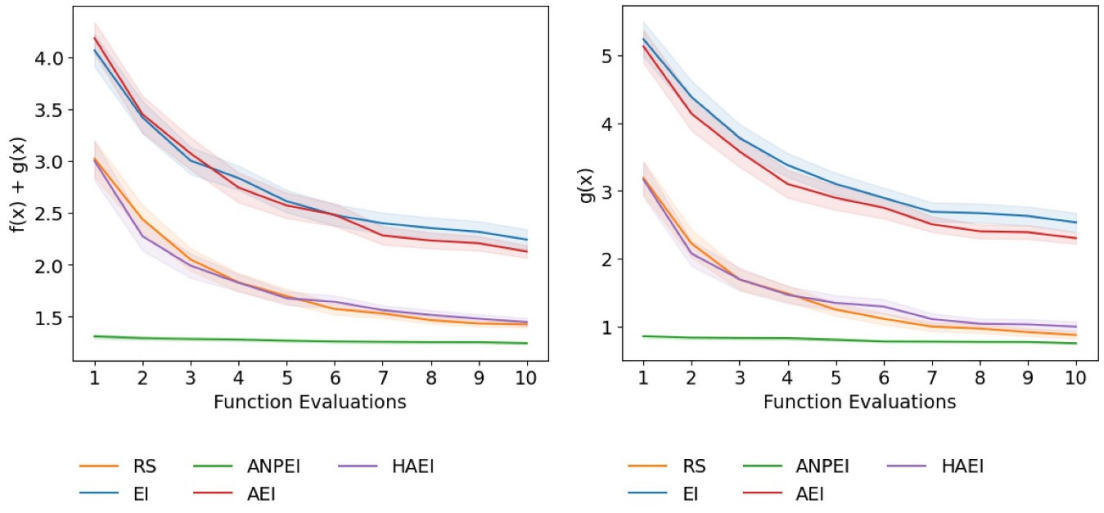


Figure 14. Hosaki function homoscedastic noise case. All Bayesian optimisation methods outperform random search with HAEI the best and ANPEI the worst.



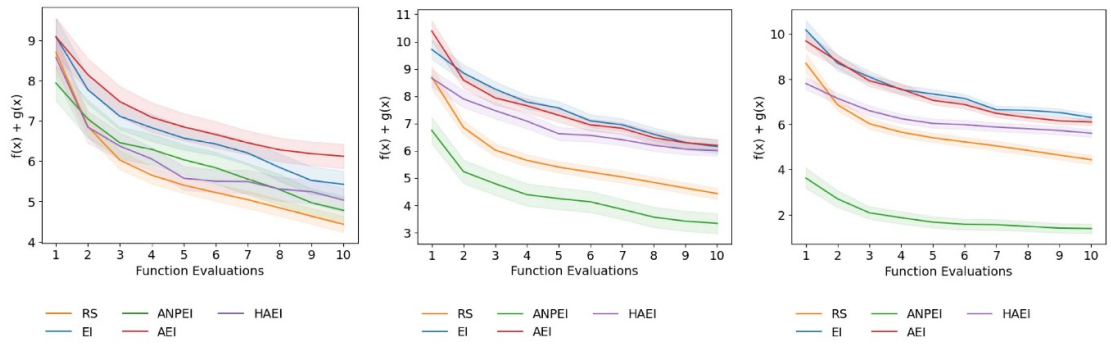
(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure 15. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the heteroscedastic 2D Hosaki function. (a) Shows the optimisation of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ (lower is better) where $g(\boldsymbol{x})$ is the aleatoric noise. (b) Shows the values $g(\boldsymbol{x})$ obtained over the course of the optimisation of $h(\boldsymbol{x})$.

6.3. Conclusions from ablation experiments

Synthesising the results from the additional ablation experiments in B some trends may be observed:

(a) All Bayesian optimisation methods outperform random search in the noiseless case and homoscedastic noise cases on aggregate across the three synthetic functions.
(b) On aggregate there is no significant difference between Bayesian optimisation methods in the noiseless or homoscedastic noise cases (HAEI marginally outperforms ANPEI on 2/3 noiseless tasks and 2/3 homoscedastic noise tasks).
(c) The heteroscedastic acquisitions ANPEI and HAEI perform competitively on the noiseless and homoscedastic noise tasks most likely because the MLHGP is capable of effecting nonstationary behaviour by 'fantasising' heteroscedastic noise. As such, the MLHGP surrogate may be achieving enhanced flexibility relative to the homoscedastic GP in this setting.



(a) 9 Points (b) 49 Points (c) 100 Points

Figure 16. The effect of the initialisation set size on the heteroscedastic Branin-Hoo function. The performance of heteroscedastic acquisitions ANPEI and HAEI increases as they are given access to more samples. An excess of samples do not help the homoscedastic Bayesian optimisation methods as they are unable to model the heteroscedastic noise component.

(d) The heteroscedastic acquisitions tend to outperform other Bayesian optimisation approaches on the heteroscedastic noise tasks although crucially this depends on the size of the initialisation set. In order to detect heteroscedastic noise the MLHGP surrogate needs access to more samples relative to the noiseless and homoscedastic cases.
(e) ANPEI outperforms HAEI.

In summary, the experiments would appear to show that there is no significant downside to employing a heteroscedastic surrogate and acquisition function on noiseless tasks or tasks with homoscedastic noise save for the increased training time for the model.

7. Conclusions

We have presented an approach for performing Bayesian optimisation with the explicit goal of minimising aleatoric noise in the suggestions. We posit that such an approach can prove useful for the natural sciences in the search for molecules and materials that are robust to experimental measurement noise. The synthetic function ablation study highlights no particular downside to the use of the MLHGP in conjunction with ANPEI or HAEI in cases where the noise structure of the problem is *a priori* unknown i.e the black-box optimisation problem is either noiseless or homoscedastic. Nonetheless, we anticipate that this type of approach may be particularly relevant for the experimental natural sciences where noiseless objectives or those with homoscedastic noise are highly uncommon. In terms of concrete recommendations on when to apply the algorithm, we foresee the best performance in situations where the user has access to a moderately-sized initialisation set in order to provide the MLHGP with enough samples to distinguish heteroscedastic noise from intrinsic function variability. There are a number of possible extensions to the current approach which may facilitate its application to high-dimensional datasets and act as fruitful sources for future work:

(a) **Surrogate Model:** One disadvantage of the MLHGP model is the lack of convergence guarantees for the EM-like procedure required for fitting. Various other forms of heteroscedastic GP exist [43–49] and have demonstrated success in modelling applications [50–53]. Of particular interest for real-world problems are scalable heteroscedastic GPs [54, 55] which could circumvent the computationally-intensive bottleneck of fitting multiple exact GPs as a subroutine of the MLHGP Bayesian optimisation procedure.
(b) **Advances in Surrogate Model Machinery**: Advances in areas such as efficient sampling of GPs [56] are liable to yield improvements to sampled-based acquisition functions such as Thompson sampling [57] while fully Bayesian approaches to hyperparameter estimation for sparse GPs [58] are liable to yield improvements in model fitting procedures.
(c) **Scalable Bayesian Optimisation:** Scalable Bayesian optimisation can also be enabled via dimensionality reduction techniques [59–61]. Such approaches, when combined with efficient libraries [62, 63] could facilitate heteroscedastic Bayesian optimisation in high-dimensional settings.
(d) **Acquisition Function Optimisation:** Recent developments in acquisition function optimisation including Monte Carlo reformulations [64, 65], compositional optimisers [65, 66] and tight relaxations [67] of common acquisition functions have the potential to yield gains in empirical performance.

(e) **Data Transformation:** Input-warping [68] and output transformations [69] have recently shown success towards addressing heteroscedastic datasets.
(f) **Approaches for Molecular Bayesian Optimisation:** In relation to molecules, the use of tailored GP kernels such as Tanimoto kernels [70, 71] and more expressive dimensionality reduction techniques [72] could lead to performance gains and enhanced scalability respectively.
(g) **Exploration in the Noise Objective:** Incorporating exploration in the noise objective in the multi-objective setting as in [22].

Lastly, a further use-case of the machinery developed in this paper is obtained by turning the noise minimisation problem into a noise maximisation problem. As an example, in materials discovery, we may derive benefit from being antifragile [73] towards (i.e. derive benefit from) high aleatoric noise. In an application such as the search for performant perovskite solar cells, we are faced with an extremely large compositional space, with millions of potential candidates possessing high aleatoric noise for identical reproductions [74]. In this instance we may want to guide search towards a candidate possessing a high photoluminescence quantum efficiency with high aleatoric noise. If the cost of repeating material syntheses is small relative to the cost of the search, the large aleatoric noise will present opportunities to synthesise materials possessing efficiencies far in excess of their mean values.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

The authors would like to thank James T Wilson for discussion about the experimental setup and applications domains of heteroscedastic Bayesian optimisation, Henry Moss for useful discussions regarding Bayesian optimisation in the noisy setting and Luke Corcoran for discussions relating to the limiting behaviour of the HAEI acquisition. We would additionally like to thank the anonymous reviewers who were instrumental in improving the quality of the empirical analysis as well as the clarity of the manuscript. VL is funded by The Alan Turing Institute Doctoral Studentship under the EPSRC Grant EP/N510129/1.

Appendix A. Heteroscedasticity of the soil phosphorus fraction dataset

Table A1 is used to demonstrate the efficacy of modelling the soil phosphorus fraction dataset using a heteroscedastic GP. The heteroscedastic GP outperforms the homoscedastic GP on prediction based on the metric of negative log predictive density (NLPD)

$$\text{NLPD} = \frac{1}{n}\sum_{i=1}^{n} -\log p(t_i|\boldsymbol{x_i})$$

which penalises both over and under-confident predictions.

Appendix B. Additional ablation experiments

In this section we present the ablation results on noiseless, homoscedastic and heteroscedastic noise tasks in line with section 6 of the main paper.

B.1. Goldstein-Price function

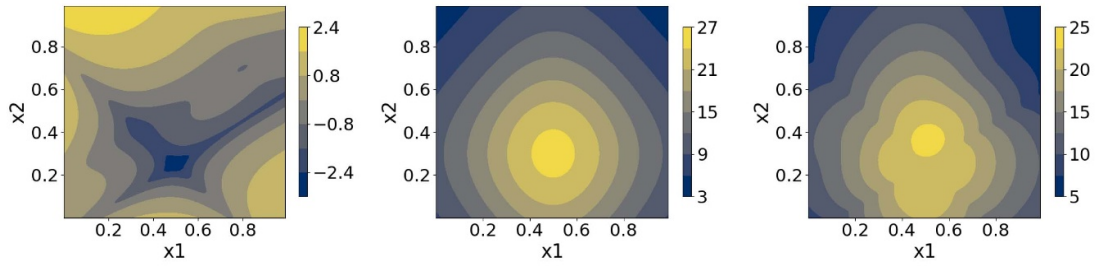
The form of the Goldstein-Price function is given in equation (9) with noise function in equation (11). The function is visualised in figure B1. Nine data points are used for initialisation in the noiseless and homoscedastic noise cases whereas 100 data points are used for initialisation in the heteroscedastic noise case. β is set to 0.5 for the noiseless and homoscedastic noise tasks and $\frac{1}{11}$ for the heteroscedastic noise task. γ is set to 500 for all experiments.

B.1.1. Noiseless case

The results of the noiseless case for Goldstein-Price are given in figure B2. All Bayesian optimisation methods outperform random search with ANPEI best and HAEI second best.

Table A1. Comparison of NLPD values on the soil phosphorus fraction dataset. Standard errors are reported for ten independent train/test splits. Lower scores are better.

Soil phosphorus fraction dataset	GP	Het GP
NLPD	1.35 ± 1.33	1.00 ± 0.95



(a) Latent Function (b) Noise Function (c) Objective Function

Figure B1. (a) The latent Goldstein-Price Function $f(\mathbf{x})$ together with (b) its heteroscedastic noise function $g(\mathbf{x})$ and (c) the objective function $f(\mathbf{x}) + g(\mathbf{x})$.

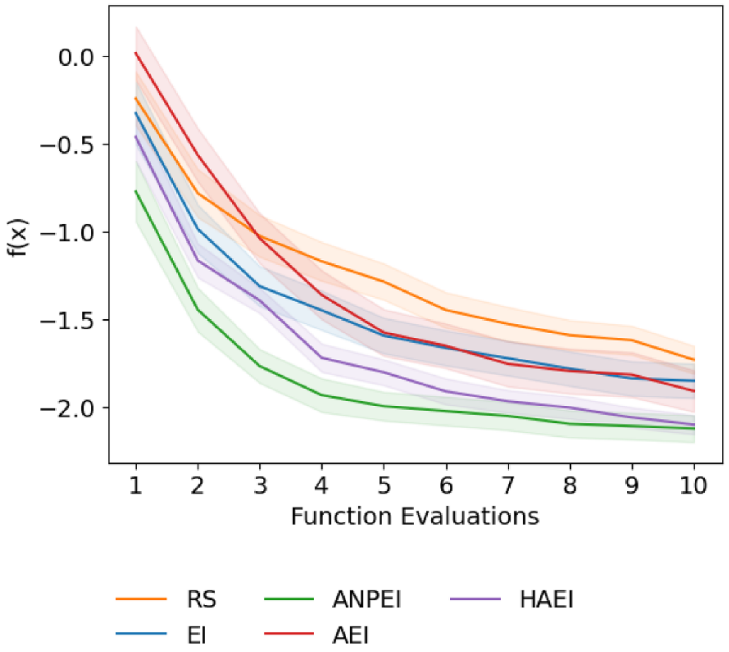


Figure B2. Goldstein-Price function noiseless case. All Bayesian optimisation methods outperform random search. ANPEI performs best and HAEI is runner-up.

B.1.2. Homoscedastic noise case

The results of the homoscedastic noise case for Goldstein-Price are shown in figure B3. In this instance HAEI performs best.

B.1.3. Heteroscedastic noise

The results of the heteroscedastic noise case for Goldstein-Price are shown in figure B4. ANPEI performs best whilst HAEI performs worse than random search.

B.2. Branin-Hoo function

The form of the Branin-Hoo function is given in equation (6) with noise function in equation (7). The function is visualised in figure B5, a figure from the main paper repeated here for clarity. Nine data points are used for initialisation in the noiseless and homoscedastic noise cases whereas 100 data points are used for initialisation in the heteroscedastic noise case. β is set to 0.5 and γ is set to 500 for all experiments.

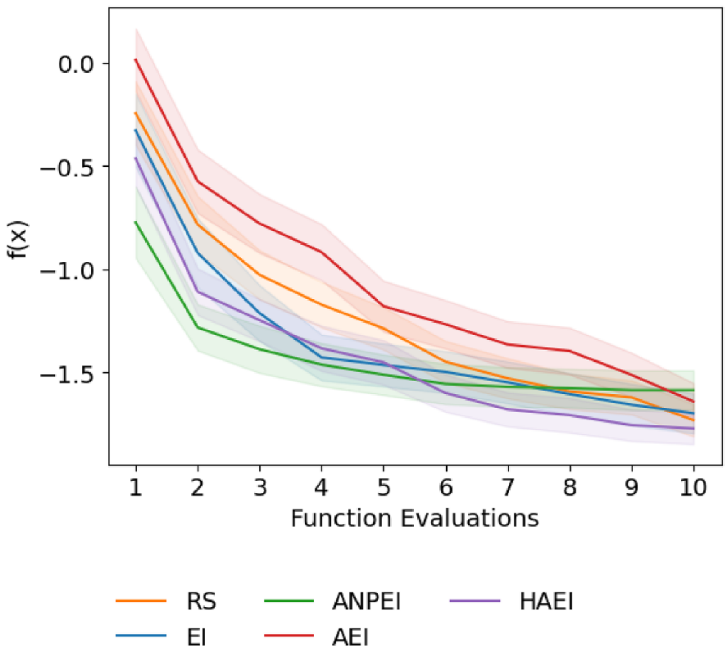
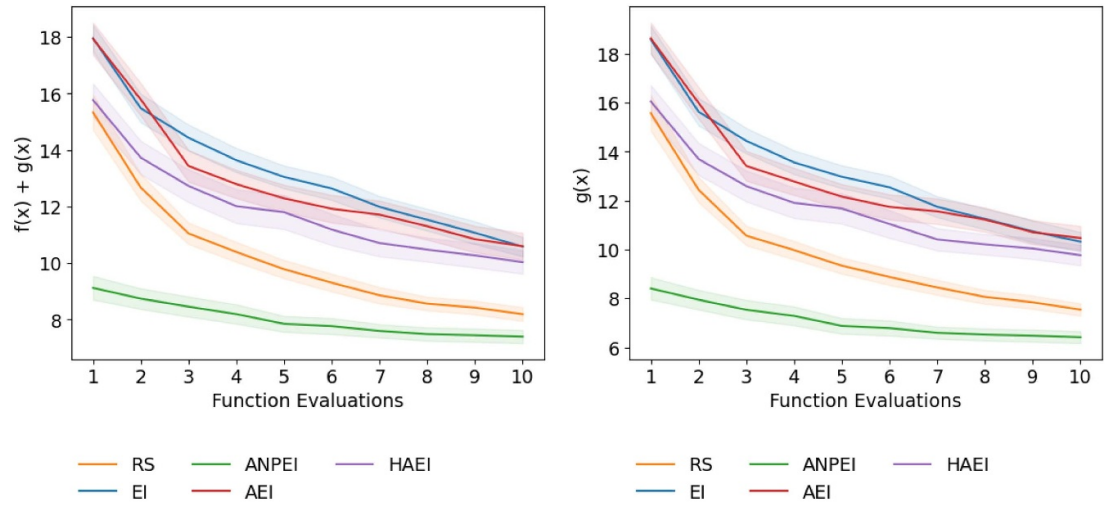
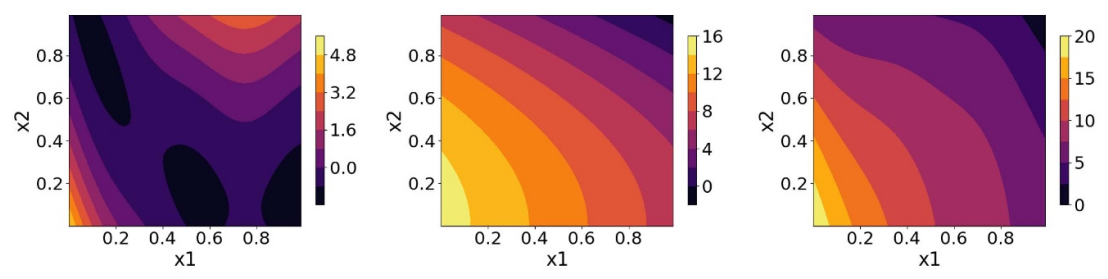


Figure B3. Goldstein-Price function homoscedastic noise case. HAEI performs best.



(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure B4. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the heteroscedastic 2D Goldstein-Price function. (a) Shows the optimisation of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ (lower is better) where $g(\boldsymbol{x})$ is the aleatoric noise. (b) Shows the values $g(\boldsymbol{x})$ obtained over the course of the optimisation of $h(\boldsymbol{x})$.



(a) Latent Function (b) Noise Function (c) Objective Function

Figure B5. Heteroscedastic Branin function.

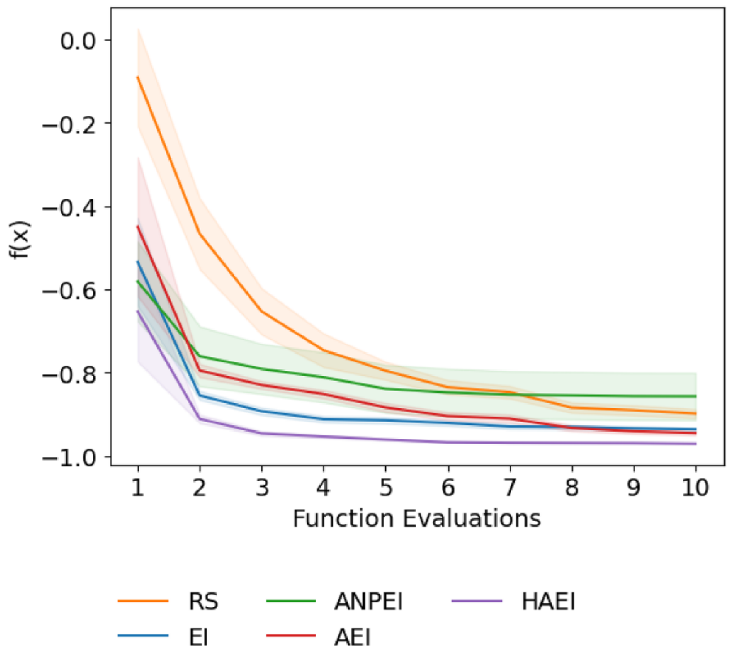


Figure B6. Branin-Hoo function noiseless case. HAEI performs best. ANPEI performs worst.

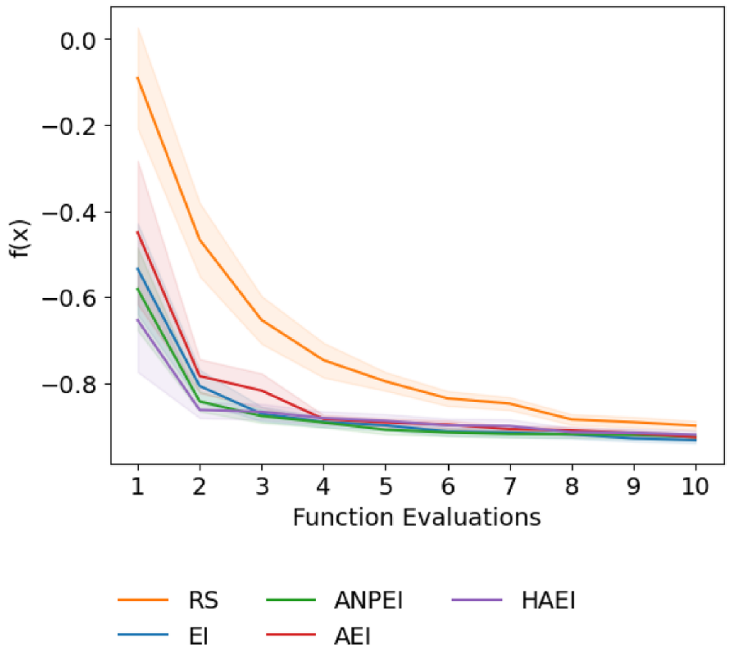


Figure B7. Branin-Hoo function homoscedastic noise case. All Bayesian optimisation methods outperform random search.

B.2.1. Noiseless case

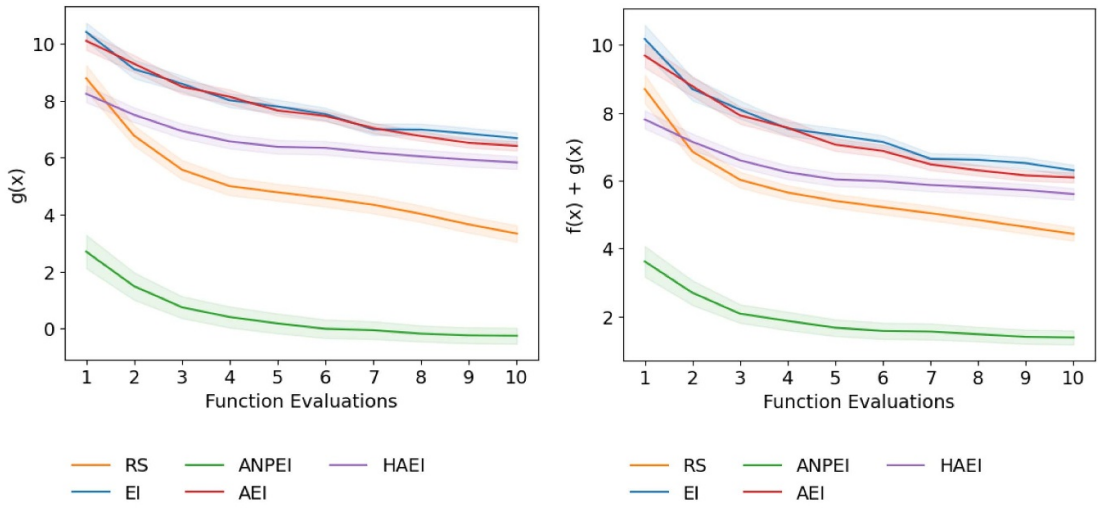
The results of the noiseless case for the Branin-Hoo function are given in figure B6. HAEI performs best in this case whereas ANPEI performs worst.

B.2.2. Homoscedastic noise case

The results of the homoscedastic noise case for the Branin-Hoo function are given in figure B7. All Bayesian optimisation methods outperform random search yet perform comparably against each other.

B.2.2.1. Heteroscedastic noise

The results of the heteroscedastic noise case for the Branin-Hoo function are shown in figure B8. ANPEI performs best whilst HAEI performs worse than random search.



(a) Best Objective Value Found so Far (b) Lowest Aleatoric Noise Found so Far

Figure B8. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the heteroscedastic 2D Branin function. (a) Shows the optimisation of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ (lower is better) where $g(\boldsymbol{x})$ is the aleatoric noise. (b) Shows the values $g(\boldsymbol{x})$ obtained over the course of the optimisation of $h(\boldsymbol{x})$.

Appendix C. Performance impact of the kernel choice

In this section we analyse the impact that the choice of GP kernel has on Bayesian optimisation performance. We select three kernels for this purpose: the SE kernel

$$k_{\text{SQE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \cdot \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right)$$

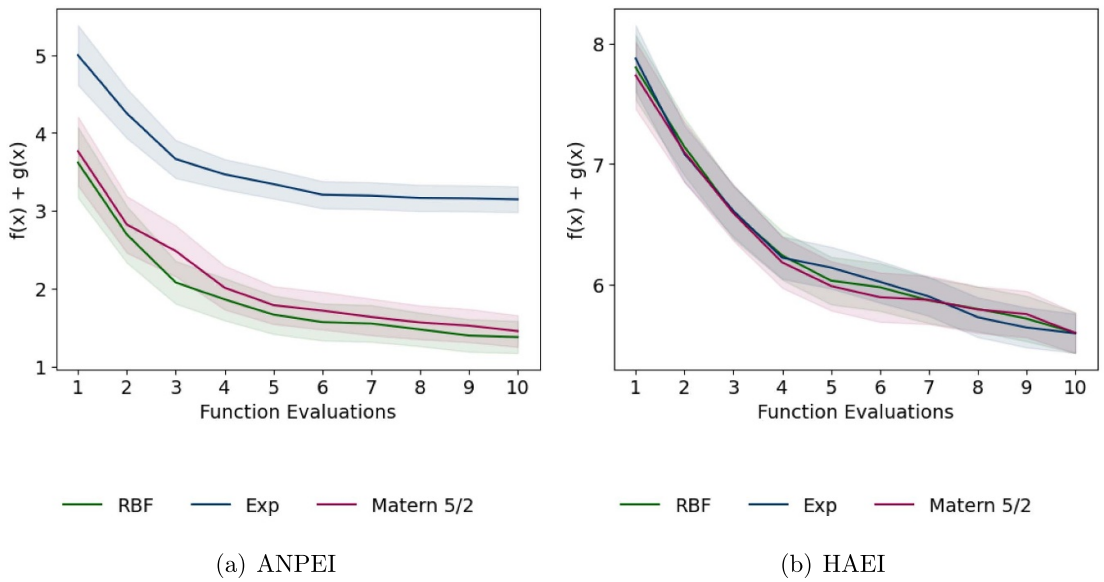
used for all experiments in the main paper, the exponential kernel

$$k_{\text{exp}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \cdot \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right),$$

a special instance of the Matérn kernel for values of $\nu = \frac{1}{2}$ [33] as well as the Matérn 5/2 kernel

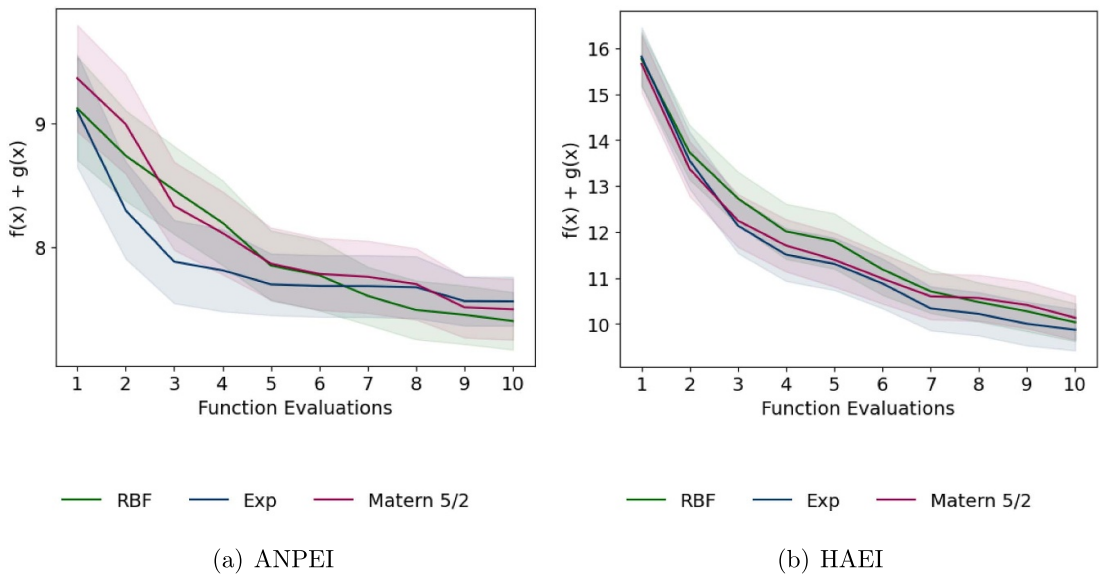
$$k_{\text{Matérn}(5/2)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \cdot \left(1 + \frac{\sqrt{5}\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell} + \frac{5\,\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{3\ell^2}\right) \cdot \exp\left(\frac{-\sqrt{5}\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right)$$

which is one of the most popular kernels for large scale empirical studies [64, 65]. It should be noted that while the equations are written assuming a single scalar lengthscale, in practice for the experiments in greater than 1D, each lengthscale is optimised per dimension under the marginal likelihood. For all experiments we choose the same kernel for both GPs of the MLHGP model i.e. the GP modelling the objective as well as the GP modelling the noise. 100 points are used for initialisation in the Branin-Hoo and Goldstein-Price functions and 144 points are used for the Hosaki function. β is set to 0.5 for the Branin-Hoo and Hosaki functions and $\frac{1}{11}$ for the Goldstein-Price function. γ is set to 500 for all experiments. The results are shown in figures C1–C3 for the Branin-Hoo function, Goldstein-Price function and Hosaki functions respectively. There is no significant difference in performance using each kernel save for the Branin-Hoo function where ANPEI underperforms using the somewhat rougher exponential kernel.



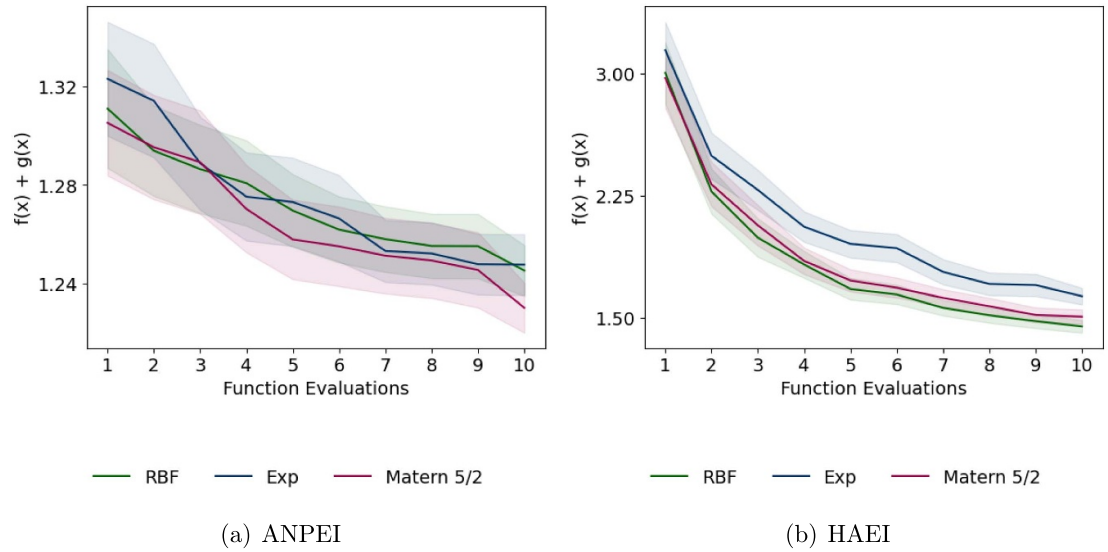
(a) ANPEI (b) HAEI

Figure C1. Branin-Hoo function kernel comparison.



(a) ANPEI (b) HAEI

Figure C2. Goldstein-Price function kernel comparison.



(a) ANPEI (b) HAEI

Figure C3. Hosaki Function kernel comparison.

ORCID iD

Ryan-Rhys Griffiths https://orcid.org/0000-0003-3117-4559

References

[1] Gómez-Bombarelli R *et al* 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Cent. Sci.* **4** 268–76
[2] Griffiths R-R and Hernández-Lobato Je M 2020 Constrained Bayesian optimization for automatic chemical design using variational autoencoders *Chem. Sci.* **11** 577–86
[3] Hoffman S, Chenthamarakshan V, Wadhawan K, Chen P-Y and Das P 2020 Optimizing molecules using efficient queries from property evaluations (arXiv:2011.01921)
[4] Hase F, Roch L M and Aspuru-Guzik A 2020 Gryffin: an algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry (arXiv:2003.12127)
[5] Hase F, Aldeghi M, Hickman R, Roch L, Liles E, Christensen M, Hein J and Aspuru-Guzik Aan 2021 Olympus: a benchmarking framework for noisy optimization and experiment planning *Mach. Learn.: Sci. Technol.* **2** 035021
[6] Terayama K, Sumita M, Tamura R, Payne D T, Chahal M K, Ishihara S and Tsuda K 2020 Pushing property limits in materials discovery via boundless objective-free exploration *Chem. Sci.* **11** 5959–68
[7] Felton K, Wigh D and Lapkin A 2020 Multi-task Bayesian optimization of chemical reactions *ChemRxiv* (https://doi.org/10.26434/chemrxiv.13250216.v2)
[8] Felton K, Rittig J and Lapkin A 2020 Summit: benchmarking machine learning methods for reaction optimisation *Chemistry-Methods* **1** 116–22
[9] Zhang C, Amar Y, Cao L and Lapkin A A 2020 Solvent selection for Mitsunobu reaction driven by an active learning surrogate model *Org. Process Res. Develop.* **24** 2864–73
[10] Calandra R, Seyfarth Ae, Peters J and Deisenroth M P 2016 Bayesian optimization for learning gaits under uncertainty *Ann. Math. Artif. Intell.* **76** 5–23
[11] Grant J, Boukouvalas A, Griffiths R-R, Leslie D, Vakili S and De Cote E M 2019 Adaptive sensor placement for continuous spaces *Proc. 36th Int. Conf. on Machine Learning* pp 2385–93
[12] Olofsson S, Mehrian M, Calandra R, Geris L, Deisenroth M P and Misener R 2018 Bayesian multiobjective optimisation with mixed analytical and black-box functions: application to tissue engineering *IEEE Trans. Biomed. Eng.* **66** 727–39
[13] Moss H, Leslie D, Beck D, Gonzalez J and Rayson P 2020 BOSS: Bayesian optimization over string spaces *Advances in Neural Information Processing Systems* p 33 (available at: https://proceedings.neurips.cc//paper/2020/hash/b19aa25ff58940d974234b48391b9549-Abstract.html)
[14] Kendall A and Gal Y 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* pp 5574–84
[15] Griffiths R-R, Schwaller P and Lee A A 2018 Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design *ChemRxiv*
[16] Matos G D R, Kyu D Y, Loeffler H H, Chodera J D, Shirts M R and Mobley D L 2017 Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database *J. Chem. Eng. Data* **62** 1559–69
[17] Hou E, Tan X, Heenan M and Wen D 2018 A global dataset of plant available and unavailable phosphorus in natural soils derived by Hedley method *Sci. Data* **5** 180166
[18] Pyzer-Knapp E O, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J and Aspuru-Guzik Aan 2015 What is high-throughput virtual screening? A perspective from organic materials discovery *Ann. Rev. Mater. Res.* **45** 195–216
[19] Hernández-Lobato Je M, Requeima J, Pyzer-Knapp E O and Aspuru-Guzik Aan 2017 Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space *Int. Conf. on Machine Learning* pp 1470–9

[20] Calandra R 2017 Bayesian modeling for optimization and control in robotics *PhD Thesis* Technische Universität Darmstadt
[21] Lázaro-Gredilla M and Titsias M K 2011 Variational heteroscedastic Gaussian process regression *Proc. 28th Int. Conf. on Int. Conf. on Machine Learning* (Omnipress) pp 841–8
[22] Kuindersma S R, Grupen R A and Barto A G 2013 Variable risk control via stochastic optimization *Int. J. Robot. Res.* **32** 806–25
[23] Assael J-A M, Wang Z, Shahriari B and de Freitas N 2014 Heteroscedastic treed Bayesian optimisation (arXiv:1410.7172)
[24] Ariizumi R, Tesch M, Choset H and Matsuno F 2014 Expensive multiobjective optimization for robotics with consideration of heteroscedastic noise *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* pp 2230–5
[25] Sui Y, Gotovos A, Burdick J and Krause A 2015 Safe exploration for optimization with Gaussian processes *Int. Conf. on Machine Learning* pp 997–1005
[26] Berkenkamp F, Krause A and Schoellig A P 2016 Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics (arXiv:1602.04450)
[27] Frazier P, Powell W and Dayanik S 2009 The knowledge-gradient policy for correlated normal beliefs *INFORMS J. Comput.* **21** 599–613
[28] Letham B, Karrer B, Ottoni G and Bakshy E *et al* 2019 Constrained Bayesian optimization with noisy experiments *Bayesian Anal.* **14** 495–519
[29] Griffiths R-R, Aldrick A A, Miguel G-O and Lee A A 2019 Heteroscedastic Bayesian optimisation in scientific discovery *NeurIPS Workshop on Machine Learning and the Physical Sciences*
[30] Griffiths R-R, Garcia-Ortegon M, Aldrick A A and Lee A A 2019 Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation *NeurIPS Workshop on Safety and Robustness in Decision-Making*
[31] Kushner H J 1964 A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise *J. Basic Eng.* **86** 97+
[32] Tiesis V, Mockus J and Žilinskas A 1978 The application of Bayesian methods for seeking the extremum *Toward Global Optimization* ed I Dixon and G Szego (Amsterdam: Elsevier)
[33] Rasmussen C E and Williams C K I 2006 *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press)
[34] Kersting K, Plagemann C, Pfaff P and Burgard W 2007 Most likely heteroscedastic Gaussian process regression *Proc. 24th Int. Conf. on Machine Learning* pp 393–400
[35] Jones D R, Schonlau M and Welch W J 1998 Efficient global optimization of expensive black-box functions *J. Glob. Optim.* **13** 455–92
[36] Picheny V, Wagner T and Ginsbourger D 2013 A benchmark of kriging-based infill criteria for noisy optimization *Struct. Multidiscip. Optim.* **48** 607–26
[37] Vazquez E, Villemonteix J, Sidorkiewicz M and Walter E 2008 Global optimization based on noisy evaluations: an empirical study of two statistical approaches *J. Phys.: Conf. Ser.* **135** 012100 IOP Publishing
[38] Huang D, Allen T T, Notz W I and Zeng N 2006 Global optimization of stochastic black-box systems via sequential kriging meta-models *J. Glob. Optim.* **34** 441–66
[39] Harris C R *et al* 2020 Array programming with NumPy *Nature* **585** 357–62
[40] Zhu C, Byrd R H, Peihuang L and Nocedal J 1997 Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization *ACM Trans. Math. Softw. (TOMS)* **23** 550–60
[41] Mobley D L and Peter Guthrie J 2014 FreeSolv: a database of experimental and calculated hydration free energies, with input files *J. Comput.-Aided Mole. Des.* **28** 711–20
[42] Landrum G Rdkit: Open-source cheminformatics (available at: http://www.rdkit.org)
[43] Le Q V, Smola A J and Canu Sephane 2005 Heteroscedastic Gaussian process regression *Proc. 22nd Int. Conf. on Machine Learning* pp 489–96
[44] Binois M, Gramacy R B and Ludkovski M 2018 Practical heteroscedastic Gaussian process modeling for large simulation experiments *J. Computat. Graph. Stat.* **27** 808–21
[45] Almosallam I 2017 Heteroscedastic Gaussian processes for uncertain and incomplete data *PhD Thesis* University of Oxford
[46] Muñoz-González L, Lázaro-Gredilla M and Figueiras-Vidal Aibal R 2011 Heteroscedastic Gaussian process regression using expectation propagation *2011 IEEE Int. Workshop on Machine Learning for Signal Processing* (IEEE) pp 1–6
[47] Wang Z and Ierapetritou M 2017 A novel surrogate-based optimization method for black-box simulation with heteroscedastic noise *Ind. Eng. Chem. Res.* **56** 10720–32
[48] Wang C and Neal R M 2012 Gaussian process regression with heteroscedastic or non-Gaussian residuals (arXiv:1212.6246)
[49] Zhang Q-H and Yi-Qing N 2020 Improved most likely heteroscedastic Gaussian process regression via Bayesian residual moment estimator *IEEE Trans. Signal Process.* **68** 3450–60
[50] Rodrigues F and Pereira F C 2018 Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data *Transp. Res.* C **95** 636–51
[51] Tabor L, Goulet J-A, Charron J-P and Desmettre C 2018 Probabilistic modeling of heteroscedastic laboratory experiments using Gaussian process regression *J. Eng. Mech.* **144** 04018038
[52] Rogers T J, Gardner P, Dervilis N, Worden K, Maguire A E, Papatheou E and Cross E J 2020 Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression *Renew. Energy* **148** 1124–36
[53] Wang Q-A and Yi-Qing N 2019 Measurement and forecasting of high-speed rail track slab deformation under uncertain SHM data using variational heteroscedastic Gaussian process *Sensors* **19** 3311
[54] Wang W and Chen Xi 2019 Distributed variational inference-based heteroscedastic Gaussian process metamodeling *2019 Winter Conf. (WSC)* (IEEE) pp 380–91
[55] Liu H, Ong Y-S and Cai J 2020 Large-scale heteroscedastic regression via Gaussian process *IEEE Trans. Neural Netw. Learn. Syst.* **32** 708–21
[56] Wilson J T, Borovitskiy V, Terenin A, Mostowsky P and Deisenroth M P 2020 Efficiently sampling functions from Gaussian process posteriors *Int. Conf. on Machine Learning*
[57] Thompson W R 1933 On the likelihood that one unknown probability exceeds another in view of the evidence of two samples *Biometrika* **25** 285–94
[58] Lalchand V and Rasmussen C E 2019 Approximate inference for fully Bayesian Gaussian process regression (arXiv:1912.13440)
[59] Moriconi R, Deisenroth M P and Kumar K S S 2020 High-dimensional Bayesian optimization using low-dimensional feature spaces *Mach. Learn.* **109** 1925–43
[60] Candelieri A and Perego R 2019 Dimensionality reduction methods to scale Bayesian optimization up *Numerical Computations: Theory and Algorithms NUMTA 2019* 167

[61] Grosnit A *et al* 2021 High-dimensional Bayesian optimisation with variational autoencoders and deep metric learning (arXiv:2106.03609)
[62] Balandat M, Karrer B, Jiang D, Daulton S, Letham B, Wilson A G and Bakshy E 2020 BoTorch: a framework for efficient Monte-Carlo Bayesian optimization *Adv. Neural Inform. Process. Syst.* **33** 21524–38
[63] Kandasamy K, Vysyaraju K R, Neiswanger W, Biswajit Paria C R Collins J S, Poczos B and Xing E P 2020 Tuning hyperparameters without grad students: scalable and robust Bayesian optimisation with dragonfly *J. Mach. Learn. Res.* **21** 1–27
[64] Wilson J, Hutter F and Deisenroth M 2018 Maximizing acquisition functions for Bayesian optimization *Adv. Neural Inf. Process. Syst.* **31** 9884–95
[65] Grosnit A, Cowen-Rivers A I, Tutunov R, Griffiths R-R, Wang J and Bou-Ammar H 2020 Are we forgetting about compositional optimisers in Bayesian optimisation? (arXiv:2012.08240)
[66] Tutunov R, Minne Li, Cowen-Rivers A I, Wang J and Bou-Ammar H 2020 Compositional adam: an adaptive compositional solver (arXiv:2002.03755)
[67] Schweidtmann A M, Bongartz D, Grothe D, Kerkenhoff T, Lin X, Najman J and Mitsos A 2020 Global optimization of Gaussian processes (arXiv:2005.10902)
[68] Wiebe J, Cecílio Inês, Dunlop J and Misener R 2020 A robust approach to warped Gaussian process-constrained optimization (arXiv:2006.08222)
[69] Cowen-Rivers A I, Lyu W, Tutunov R, Wang Z, Grosnit A, Griffiths R R, Jianye H, Wang J and Ammar H B 2020 An empirical study of assumptions in Bayesian optimisation (arXiv:2012.03826)
[70] Moss H B and Griffiths R-R 2020 Gaussian process molecule property prediction with FlowMO (arXiv:2010.01118)
[71] Thawani A R, Griffiths R-R, Jamasb A, Bourached A, Jones P, McCorkindale W, Aldrick A A and Lee A A 2020 The photoswitch dataset: a molecular machine learning benchmark for the advancement of synthetic chemistry (arXiv:2008.03226)
[72] Cheng B *et al* 2020 Mapping materials and molecules *Acc. Chem. Res.* **53** 1981–91
[73] Taleb N N 2012 *Antifragile: Things That Gain From Disorder* 1st edn (New York: Random House)
[74] Zhou Y and Zhao Y 2019 Chemical stability and instability of inorganic halide perovskites *Energy Environ. Sci.* **12** 1495–511