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

[2] Department of Engineering, University of Cambridge

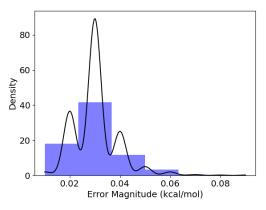[3] Department of Mathematics University of Cambridge
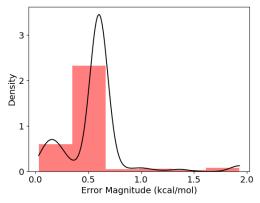
E-mail: `rrg27@cam.ac.uk`

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 consistently 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 (GP) surrogate model in conjunction with two straightforward adaptations of existing acquisition functions. First, we extend the augmented expected improvement (AEI) 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 and yield 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

Keywords: Bayesian Optimisation, Gaussian Processes, Heteroscedasticity

1. Introduction

Bayesian optimisation is proving to be a highly effective search methodology in areas such as drug discovery [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], materials discovery [11, 12, 13, 14, 15, 16, 17, 18, 19, 20], chemical reaction optimisation [21, 22, 23], robotics [24], sensor placement [25], tissue engineering [26] and genetics [27]. 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) [28]. 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 suprisingly 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 [29].
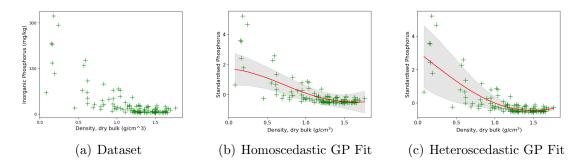

(a) Density plot of computational errors   (b) Density plot of experimental errors

Figure 1. (a) The density histogram of computational errors (kcal/mol) for the FreeSolv hydration energy dataset ([30]). 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.

In Figure 1 we illustrate real-world sources of heteroscedasticity using the FreeSolv dataset of [30]. The consequences of misrepresenting heteroscedastic noise as being homoscedastic, i.e. constant across the input domain, are illustrated using a second dataset [31] 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 [32] the cost of misrepresenting noise during the screening process could

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

Figure 2. Comparison of homoscedastic and heteroscedastic GP fits to the soil phosphorus fraction dataset [31].

lead to a substantial loss of time in material fabrication [33]. In this paper we present a heteroscedastic Bayesian optimisation algorithm capable of both representing and minimising aleatoric noise in its suggestions. Our contributions are:

(1) The introduction of a novel combination of surrogate model and acquisition function designed to minimise heteroscedastic aleatoric uncertainty.

(2) 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.

(3) The provision of an open-source implementation available at https://github.com/ Ryan-Rhys/Heteroscedastic-BO.

2. Background

2.1. Bayesian Optimisation

Bayesian optimisation solves the global optimisation problem defined as

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \tag{1}$$

where $\mathbf{x}^*$ is the global optimiser of a black-box function $f : \mathcal{X} \to \mathcal{Y}$. $\mathcal{X}$ 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:

(i) Black-Box Objective: We do not have the analytic form of f. We can however evaluate f pointwise anywhere in the design space $\mathcal{X}$.

(ii) Expensive Evaluations: Choosing an input location $\mathbf{x}$ and evaluating $f(\mathbf{x})$ takes a very long time.

(iii) 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 Gaussian processes (GPs) are the most popular choice of surrogate model because of their ability to represent posterior uncertainty without resorting to approximate Bayesian inference.

2.2. Gaussian Processes

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

Definition 1. A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Gaussian processes 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

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2}$$

and a covariance function

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x} - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}')))]. \tag{3}$$

The process is written as follows

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big). \tag{4}$$

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

$$k_{\text{SQE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \cdot \exp\Big(\frac{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\Big) \tag{5}$$

where σ_f^2 is the signal amplitude hyperparameter (vertical lengthscale) and ℓ is the (horizontal) lengthscale hyperparameter. We use the squared exponential kernel in all experiments. For a more detailed introduction to Gaussian processes the reader is referred to [34].

3. 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.

3.1. The Most Likely Heteroscedastic Gaussian Process

We adopt the most likely heteroscedastic Gaussian process (MLHGP) approach of [35], 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

$$\mu^* = E[\mathbf{t}^*] = K^*(K + R)^{-1}\mathbf{t} \tag{6}$$

and covariance matrix

$$\Sigma^* = \mathrm{var}[\mathbf{t}^*] = K^{**} + R^* - K^*(K + R)^{-1}K^{*T}, \tag{7}$$

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 most likely heteroscedastic GP algorithm [35] executes the following steps:

(i) Estimate a homoscedastic GP, G_1 on the dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}_{i=1}^n$

(ii) Given G_1, we estimate the empirical noise levels for the training data using $z_i = \log(\mathrm{var}[t_i, G_1(\mathbf{x}_i, \mathcal{D})]) = 0.5\,(t_i - \mathbb{E}[\mathbf{x}])^2$ forming a new dataset $\mathcal{D}' = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$

(iii) Estimate a second GP, G_2 on $\mathcal{D}'$.

(iv) Estimate a combined GP, G_3 on $\mathcal{D}$ using G_2 to predict the logarithmic noise levels r_i.

(v) 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.

3.2. Bayesian Optimisation with Aleatoric Noise Penalisation

Our heteroscedastic Bayesian optimisation problem may be framed as

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \chi} h(\boldsymbol{x}), \tag{8}$$

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

$$h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + (1 - \alpha)g(\boldsymbol{x}). \tag{9}$$

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 by a domain expert that trades off the weight of the principal objective relative to the noise objective.

3.3. Heteroscedastic Acquisition Functions

We investigate extensions of the expected improvement [36] 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

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big] = \int_{-\infty}^{\infty} (\eta - t)_+ \, p(t \,|\, \boldsymbol{x}) \, dt \tag{10}$$

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 expected improvement with plug-in [37], the plug-in value being the GP predictive mean [38].

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

$$\mathrm{AEI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big]\left(1 - \frac{\sigma_n}{\sqrt{\mathrm{var}[t] + \sigma_n^2}}\right), \tag{11}$$

of [39] where σ_n is the fixed aleatoric noise level. AEI was introduced 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:

$$\text{HAEI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big]\left(1 - \frac{\gamma\sqrt{r(\boldsymbol{x})}}{\sqrt{\text{var}[t] + \gamma^2 r(\boldsymbol{x})}}\right), \tag{12}$$

where $r(\boldsymbol{x})$ is the predicted aleatoric uncertainty at input $\boldsymbol{x}$ under the MLHGP and 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 $k = \frac{\text{var}[t]}{r(\boldsymbol{x})}$ 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 12 by $\sqrt{r(\boldsymbol{x})}$ yields

$$\text{HAEI}(\boldsymbol{x}) = \text{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k + \gamma^2}}\right). \tag{13}$$

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

$$\lim_{k \to \infty} \text{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k + \gamma^2}}\right) = \text{EI}(\boldsymbol{x}), \tag{14}$$

recovers the expected improvement acquisition.

$\square$

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 13 yields

$$\lim_{k \to 0} \text{EI}(\boldsymbol{x})\left(1 - \frac{\gamma}{\sqrt{k + \gamma^2}}\right) = 0. \tag{15}$$

$\square$

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

Letting $S(k) = 1 - \frac{\gamma}{k+\gamma^2}$ such that $\text{HAEI} = \text{EI}(\boldsymbol{x})S(k)$, consider the Taylor expansion of $S(k)$ around $k = 0$,

$$S(k) = 1 + S'(0)k + \frac{S''(0)}{2!}k^2 + \frac{S'''(0)}{3!}k^3 + \dots, \tag{16}$$

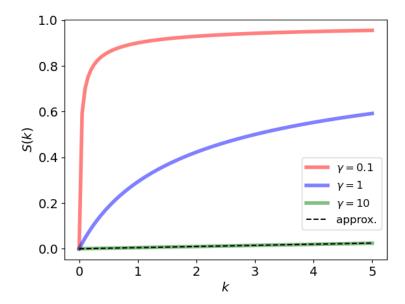Dropping terms of $O(k^2)$ and higher we obtain

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 $\gamma = 10$.

$$S(k) \approx 1 - \frac{\gamma^2}{2k}. \tag{17}$$

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 aleatoric noise-penalised expected improvement (ANPEI) and denote it

$$\text{ANPEI} = \beta \text{EI}(\boldsymbol{x}) - (1 - \beta)\sqrt{r(\boldsymbol{x})}, \tag{18}$$

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.

4. Related Work

The most similar work to our own is that of [40] where experiments are reported on a heteroscedastic Branin-Hoo toy function using the variational heteroscedastic GP approach of [41]. 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 EI, expected risk improvement is introduced in [42] 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 [40]. [43, 44] implement heteroscedastic Bayesian optimisation but don't introduce an acquisition function that penalises aleatoric noise. [45, 46] 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 and as such. 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, [47, 48] 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.

5. Experiments on Robustness to Aleatoric Uncertainty

5.1. Implementation

Experiments were run using a custom Numpy implementation of Gaussian process regression and most likely heteroscedastic Gaussian process regression. All code to reproduce the experiments is available at https://github.com/Ryan-Rhys/Heteroscedastic-BO. The squared exponential 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 most likely heteroscedastic GP [35] 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 10 iterations and the sample size required to construct the variance estimator producing the auxiliary dataset was 100.

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 [49], taking the best of 3 random restarts. The objective function is

$$h(x) = \alpha f(x) - (1 - \alpha)g(x) \tag{19}$$

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

$$h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + (1 - \alpha)g(\boldsymbol{x}) \tag{20}$$

The sin wave and Branin-Hoo experiments are initialised with 25 data points 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, 0.5, $\frac{1}{6}$ and 0.5 for the sin, Branin-Hoo, soil and FreeSolv experiments respectively while β is set to 0.5, 0.5, $\frac{1}{6}$ and 0.5 for the sin, Branin-Hoo, soil and FreeSolv experiments. γ is set to 1, 1, 5 and 1 for the sin, Branin-Hoo, soil and FreeSolv experiments. We run 5 acquisition functions in all experiments: random sampling, homoscedastic EI, AEI, HAEI and ANPEI. Homoscedastic EI is included as a baseline to demonstrate the difference that consideration of aleatoric noise yields in the optimisation of the objective. AEI is included to demonstrate the difference that consideration of heteroscedastic aleatoric noise yields and random sampling is included as a baseline as it is known to be competitive with Bayesian optimisation in noisy settings.
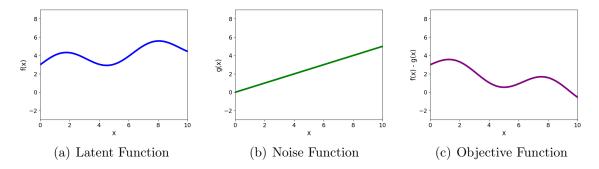
5.2. Heteroscedastic Sin Wave Function

The objective function has the form

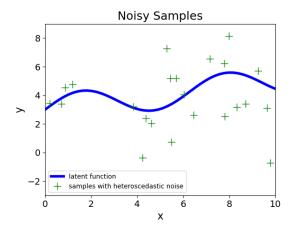$$h(x) = f(x) - g(x) \tag{21}$$

where $f(x) = \sin(x) + 0.2(x) + 3$ and $g(x) = 0.5(x)$. In this instance α from Equation 19 has a setting of 0.5 but we omit it explicitly as the objectives have equal weight. Over the course of the experiment samples

$$y_i = f(x_i) + g(x_i)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{22}$$

are observed. The problem setup is depicted in Figure 4 and Figure 5. The Bayesian optimisation problem is constructed such that the first maximum in 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 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 $h(x)$.

(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. Samples $y_i = f(x_i) + g(x_i)\epsilon$ from the heteroscedastic sin wave function.

5.3. Heteroscedastic Branin-Hoo Function

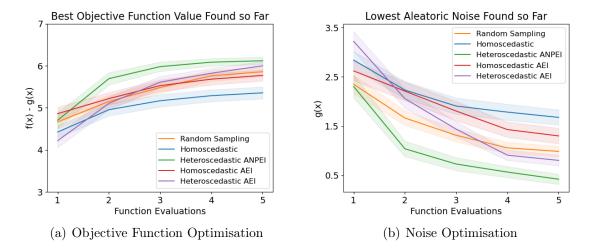In the second experiment we consider the objective

$$h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{23}$$

with an additive penalty because the task is a maximisation problem and an α setting of 0.5 for equal-weight objectives.
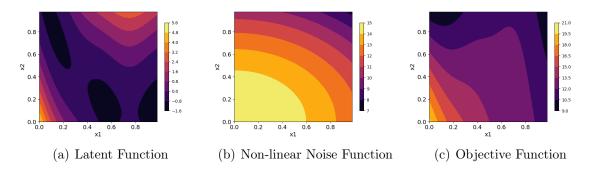
$$f(\boldsymbol{x}) = \frac{1}{51.95} \left[\left(\bar{x}_2 - \frac{5.1\bar{x}_1^2}{4\pi^2} + \frac{5\bar{x}_1}{\pi} - 6 \right)^2 + \left(10 - \frac{10}{8\pi} \right) \cos\left(\bar{x}_1 \right) - 44.81 \right] \tag{24}$$

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 [37]. The noise function $g(\boldsymbol{x})$ is in this instance

$$g(\boldsymbol{x}) = 15 - 2.8x_1^2 - 4.8x_2^2. \tag{25}$$

(a) Objective Function Optimisation   (b) Noise Optimisation

Figure 6. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the sin wave problem. Standard error bands are computed using 50 random initialisations. (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 and HAEI (labelled as Heteroscedastic AEI) to seek low aleatoric noise solutions.

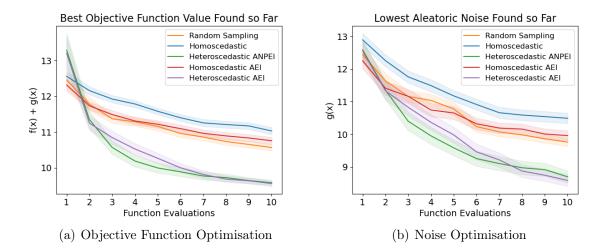(a) Latent Function   (b) Non-linear Noise Function   (c) Objective Function

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.

Samples are again generated according to

$$y_i = f(\boldsymbol{x}_i) + g(\boldsymbol{x}_i)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{26}$$

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.

(a) Objective Function Optimisation (b) Noise Optimisation

Figure 8. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the Branin-Hoo problem. Standard error bands are computed using 50 random initialisations. (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 $h(\boldsymbol{x})$.

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, [31] 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 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 most likely heteroscedastic GP 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:

(1) Fit a homoscedastic GP to the full dataset.

(2) At each point x_i, compute the corresponding squared error $s_i^2 = (y_i - \mu(x_i))^2$.

(3) 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.
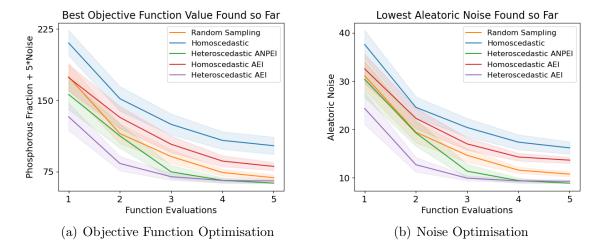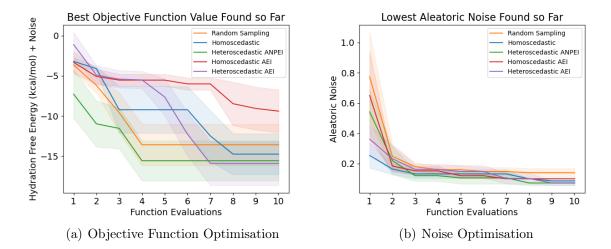
(a) Objective Function Optimisation (b) Noise Optimisation

Figure 9. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the soil phosphorus fraction optimisation problem. Standard error bands are computed using 50 random initialisations. (a) shows the optimisation of $h(x) = \frac{1}{6}f(x) + \frac{5}{6}g(x)$ (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 $h(x)$. The β parameter of ANPEI is set to $\frac{1}{6}$ and the γ parameter of HAEI is set to 5 in order to encourage low-noise seeking behaviour.

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 [50, 30] over 10 iterations of data collection. 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 [51] 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 and as such the heteroscedastic modelling problem is more difficult, leading to only very marginal gains in obtaining low noise solutions.

(a) Objective Function Optimisation (b) Noise Optimisation

Figure 10. Comparison of heteroscedastic and homoscedastic Bayesian optimisation on the FreeSolv hydration free energy optimisation problem. Standard error bands are computed using 5 random initialisations. (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 $h(\boldsymbol{x})$.

5.6. Experiment Summary

The experiments provide strong evidence that modelling heteroscedasticity in Bayesian optimisation is a more flexible approach to assuming homoscedastic noise. Both heteroscedastic surrogate model/acquisition function pairs outperform homoscedastic baselines and random sampling across all problems.

6. 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. Nonetheless, there are a number of limitations of the current approach which inhibit its application to high-dimensional, real-world datasets and act as fruitful sources for future work:

(1) **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 [52, 53, 54, 55, 56, 57, 58] and have demonstrated success in modelling applications [59, 60, 61, 62]. Of particular interest for real-world problems are scalable heteroscedastic GPs [63, 64] which could circumvent the computationally-intensive bottleneck of fitting multiple exact GPs as a subroutine of the MLHGP Bayesian optimisation procedure.

(2) **Advances in Surrogate Model Machinery**: Advances in areas such as efficient sampling of GPs [65] are liable to yield improvements to sampled-based acquisition functions such as Thompson sampling [66] while fully Bayesian approaches to hyperparameter estimation for sparse GPs [67] are liable to yield improvements in model fitting procedures.

(3) **Scalable Bayesian Optimisation:** Scalable Bayesian optimisation can also be enabled via dimensionality reduction techniques [68, 69]. Such approaches, when combined with efficient libraries [70, 71] could facilitate heteroscedastic Bayesian optimisation in high-dimensional settings.

(4) **Acquisition Function Optimisation:** Recent work on acquisition function optimisation [72, 73] has the potential to yield gains in empirical performance.

(5) **Data Transformation:** Input-warping [74] and output transformations [75] have recently shown success towards addressing heteroscedastic datasets.

(6) **Approaches for Molecular Bayesian Optimisation:** In relation to molecules, the use of tailored GP kernels such as Tanimoto kernels [76, 77] and more expressive dimensionality reduction techniques [78] could lead to performance gains and enhanced scalability respectively.

(7) **Exploration in the Noise Objective:** Incorporating exploration in the noise objective in the multi-objective setting as in [42].

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 [79] 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 [80]. 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.

7. Acknowledgements

References

[1] 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.

[2] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11(2):577–586, 2020.

[3] Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. In *International Conference on Artificial Intelligence and Statistics*, pages 3393–3403. PMLR, 2020.

[4] Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. *arXiv preprint arXiv:1809.02032*, 2018.

[5] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Karam MJ Thomas, Simon Blackburn, Connor W Coley, Jian Tang, Sarath Chandar, and Bengio Yoshua. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *37th International Conference on Machine Learning*, volume 119, pages 3668–3679, 2020.

[6] David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *arXiv preprint arXiv:2012.07127*, 2020.

[7] Samuel Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. Optimizing molecules using efficient queries from property evaluations. *arXiv preprint arXiv:2011.01921*, 2020.

[8] Omar Mahmood and José Miguel Hernández-Lobato. A cold approach to generating optimal samples. *arXiv preprint arXiv:1905.09885*, 2019.

[9] Shali Jiang, Gustavo Malkomes, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search with applications in drug and materials discovery. *arXiv preprint arXiv:1811.08871*, 2018.

[10] 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, 2020.

[11] Jialin Song, Yury S. Tokpanov, Yuxin Chen, Dagny Fleischman, Kate T. Fountaine, Harry A. Atwater, and Yisong Yue. Optimizing Photonic Nanostructures via Multi-fidelity Gaussian Processes. *arXiv:1811.07707*, 2018.

[12] Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity Bayesian optimization with Gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3158–3167, 2019.

[13] Henry C Herbol, Weici Hu, Peter Frazier, Paulette Clancy, and Matthias Poloczek. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. *npj Computational Materials*, 4(1):51, 2018.

[14] Florian Hase, Loic M Roch, Christoph Kreisbeck, and Alan Aspuru-Guzik. Phoenics: A Bayesian optimizer for chemistry. *ACS Central Science*, 4(9):1134–1145, 2018.

[15] Florian Hase, Loic M Roch, and Alan Aspuru-Guzik. Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry. *arXiv preprint arXiv:2003.12127*, 2020.

[16] Martha M Flores-Leonar, Luis M Mejia-Mendoza, Andres Aguilar-Granda, Benjamin Sanchez-Lengeling, Hermann Tribukait, Carlos Amador-Bedolla, and Alan Aspuru-Guzik. Materials Acceleration Platforms: On the way to autonomous experimentation. *Current Opinion in Green and Sustainable Chemistry*, page 100370, 2020.

[17] Florian Häse, Matteo Aldeghi, Riley J Hickman, Loïc M Roch, Melodie Christensen, Elena Liles, Jason E Hein, and Alán Aspuru-Guzik. Olympus: a benchmarking framework for noisy optimization and experiment planning. *arXiv preprint arXiv:2010.04153*, 2020.

[18] Kevin Maik Jablonka, Giriprasad Melpatti Jothiappan, Shefang Wang, Berend Smit, and Brian Yoo. Bias free multiobjective active learning for materials design and discovery. *ChemRxiv*, 2020.

[19] Kei Terayama, Masato Sumita, Ryo Tamura, Daniel T Payne, Mandeep K Chahal, Shinsuke Ishihara, and Koji Tsuda. Pushing property limits in materials discovery via boundless objective-free exploration. *Chemical Science*, 2020.

[20] Melodie Christensen, Lars Yunker, Folarin Adedeji, Florian Hase, Loic Roch, Tobias Gensch, Gabriel dos Passos Gomes, Tara Zepel, Matthew Sigman, Alan Aspuru-Guzik, and Jason Hein. Data-science driven autonomous process optimization. *ChemRxiv*, 11 2020.

[21] Kobi Felton, Daniel Wigh, and Alexei Lapkin. Multi-task Bayesian optimization of chemical reactions. *ChemRxiv*, 2020.

[22] Kobi Felton, Jan Rittig, and Alexei Lapkin. Summit: Benchmarking machine learning methods for reaction optimisation. *ChemRxiv*, 2020.

[23] Chonghuan Zhang, Yehia Amar, Liwei Cao, and Alexei A. Lapkin. Solvent selection for Mitsunobu reaction driven by an active learning surrogate model. *Organic Process Research & Development*, 24(12):2864–2873, 2020. doi: 10.1021/acs.oprd.0c00376.

[24] Roberto Calandra, André Seyfarth, Jan Peters, and Marc Peter Deisenroth. Bayesian optimization for learning gaits under uncertainty. *Annals of Mathematics and Artificial Intelligence*, 76(1-2):5–23, 2016.

[25] James Grant, Alexis Boukouvalas, Ryan-Rhys Griffiths, David Leslie, Sattar Vakili, and Enrique Munoz De Cote. Adaptive sensor placement for continuous spaces.

In *Proceedings of the 36th International Conference on Machine Learning*, pages 2385–2393, 2019.

[26] Simon Olofsson, Mohammad Mehrian, Roberto Calandra, Liesbet Geris, Marc Peter Deisenroth, and Ruth Misener. Bayesian multiobjective optimisation with mixed analytical and black-box functions: Application to tissue engineering. *IEEE Transactions on Biomedical Engineering*, 66(3):727–739, 2018.

[27] Henry Moss, David Leslie, Daniel Beck, Javier Gonzalez, and Paul Rayson. Boss: Bayesian optimization over string spaces. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[29] Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A. Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. *ChemRxiv*, 2018.

[30] Guilherme Duarte Ramos Matos, Daisy Y Kyu, Hannes H Loeffler, John D Chodera, Michael R Shirts, and David L Mobley. Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the freesolv database. *Journal of Chemical & Engineering Data*, 62(5):1559–1569, 2017.

[31] Enqing Hou, Xiang Tan, Marijke Heenan, and Dazhi Wen. A global dataset of plant available and unavailable phosphorus in natural soils derived by Hedley method. *Scientific Data*, 5(1):180166, 2018.

[32] Edward O Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annual Review of Materials Research*, 45:195–216, 2015.

[33] José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *International Conference on Machine Learning*, pages 1470–1479, 2017.

[34] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[35] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic Gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, pages 393–400, 2007.

[36] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[37] Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-

based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

[38] Emmanuel Vazquez, Julien Villemonteix, Maryan Sidorkiewicz, and Eric Walter. Global optimization based on noisy evaluations: an empirical study of two statistical approaches. In *Journal of Physics: Conference Series*, volume 135, page 012100. IOP Publishing, 2008.

[39] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34(3):441–466, 2006.

[40] Roberto Calandra. *Bayesian Modeling for Optimization and Control in Robotics*. PhD thesis, Technische Universität Darmstadt, 2017.

[41] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic Gaussian process regression. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 841–848. Omnipress, 2011.

[42] Scott R Kuindersma, Roderic A Grupen, and Andrew G Barto. Variable risk control via stochastic optimization. *The International Journal of Robotics Research*, 32(7): 806–825, 2013.

[43] John-Alexander M Assael, Ziyu Wang, Bobak Shahriari, and Nando de Freitas. Heteroscedastic treed Bayesian optimisation. *arXiv preprint arXiv:1410.7172*, 2014.

[44] Ryo Ariizumi, Matthew Tesch, Howie Choset, and Fumitoshi Matsuno. Expensive multiobjective optimization for robotics with consideration of heteroscedastic noise. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2230–2235, 2014.

[45] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning*, pages 997–1005, 2015.

[46] Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*, 2016.

[47] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[48] Benjamin Letham, Brian Karrer, Guilherme Ottoni, Eytan Bakshy, et al. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14 (2):495–519, 2019.

[49] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[50] David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and

calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28(7):711–720, 2014.

[51] Greg Landrum. Rdkit: Open-source cheminformatics. URL http://www.rdkit.org.

[52] Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic Gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 489–496, 2005.

[53] Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

[54] Ibrahim Almosallam. *Heteroscedastic Gaussian processes for uncertain and incomplete data*. PhD thesis, University of Oxford, 2017.

[55] Luis Muñoz-González, Miguel Lázaro-Gredilla, and Aníbal R Figueiras-Vidal. Heteroscedastic Gaussian process regression using expectation propagation. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2011.

[56] Zilong Wang and Marianthi Ierapetritou. A novel surrogate-based optimization method for black-box simulation with heteroscedastic noise. *Industrial & Engineering Chemistry Research*, 56(38):10720–10732, 2017.

[57] Chunyi Wang and Radford M Neal. Gaussian process regression with heteroscedastic or non-Gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

[58] Qiu-Hu Zhang and Yi-Qing Ni. Improved most likely heteroscedastic Gaussian process regression via Bayesian residual moment estimator. *IEEE Transactions on Signal Processing*, 68:3450–3460, 2020.

[59] Filipe Rodrigues and Francisco C Pereira. Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data. *Transportation Research Part C: Emerging Technologies*, 95:636–651, 2018.

[60] Lucie Tabor, James-A Goulet, Jean-Philippe Charron, and Clelia Desmettre. Probabilistic modeling of heteroscedastic laboratory experiments using Gaussian process regression. *Journal of Engineering Mechanics*, 144(6):04018038, 2018.

[61] TJ Rogers, P Gardner, N Dervilis, K Worden, AE Maguire, E Papatheou, and EJ Cross. Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression. *Renewable Energy*, 148:1124–1136, 2020.

[62] Qi-Ang Wang and Yi-Qing Ni. Measurement and forecasting of high-speed rail track slab deformation under uncertain shm data using variational heteroscedastic Gaussian process. *Sensors*, 19(15):3311, 2019.

[63] Wenjing Wang and Xi Chen. Distributed variational inference-based heteroscedastic Gaussian process metamodeling. In *2019 Winter Simulation Conference (WSC)*, pages 380–391. IEEE, 2019.

[64] Haitao Liu, Yew-Soon Ong, and Jianfei Cai. Large-scale heteroscedastic regression via Gaussian process. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[65] James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, 2020.

[66] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[67] Vidhi Lalchand and Carl Edward Rasmussen. Approximate inference for fully Bayesian Gaussian process regression. *arXiv preprint arXiv:1912.13440*, 2019.

[68] 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.

[69] Antonio Candelieri and Riccardo Perego. Dimensionality reduction methods to scale Bayesian optimization up. *Numerical Computations: Theory and Algorithms NUMTA 2019*, page 167, 2019.

[70] Maximilian Balandat, Brian Karrer, Daniel R Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: Programmable bayesian optimization in pytorch. *arXiv preprint arXiv:1910.06403*, 2019.

[71] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R. Collins, Jeff Schneider, Barnabas Poczos, and Eric P. Xing. Tuning hyperparameters without grad students: Scalable and robust Bayesian optimisation with dragonfly. *Journal of Machine Learning Research*, 21(81):1–27, 2020. URL http://jmlr.org/papers/v21/18-223.html.

[72] James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for Bayesian optimization. *Advances in Neural Information Processing Systems*, 31: 9884–9895, 2018.

[73] Antoine Grosnit, Alexander I Cowen-Rivers, Rasul Tutunov, Ryan-Rhys Griffiths, Jun Wang, and Haitham Bou-Ammar. Are we forgetting about compositional optimisers in Bayesian optimisation? *arXiv preprint arXiv:2012.08240*, 2020.

[74] Johannes Wiebe, Inês Cecílio, Jonathan Dunlop, and Ruth Misener. A robust approach to warped Gaussian process-constrained optimization. *arXiv preprint arXiv:2006.08222*, 2020.

[75] Alexander I Cowen-Rivers, Wenlong Lyu, Zhi Wang, Rasul Tutunov, Hao Jianye, Jun Wang, and Haitham Bou Ammar. Hebo: Heteroscedastic evolutionary Bayesian optimisation. *arXiv preprint arXiv:2012.03826*, 2020.

[76] Henry B Moss and Ryan-Rhys Griffiths. Gaussian process molecule property prediction with FlowMO. *arXiv preprint arXiv:2010.01118*, 2020.

[77] Aditya R Thawani, Ryan-Rhys Griffiths, Arian Jamasb, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander A Aldrick, and Alpha A Lee. The

photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry. *arXiv preprint arXiv:2008.03226*, 2020.

[78] Bingqing Cheng, Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bonan Zhu, Volker L Deringer, Noam Bernstein, Johannes T Margraf, Karsten Reuter, et al. Mapping materials and molecules. *Accounts of Chemical Research*, 53(9):1981–1991, 2020.

[79] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, New York, 1st ed edition, 2012. ISBN 978-1-4000-6782-4.

[80] Yuanyuan Zhou and Yixin Zhao. Chemical stability and instability of inorganic halide perovskites. *Energy & Environmental Science*, 12(5):1495–1511, May 2019. ISSN 1754-5706. doi: 10.1039/C8EE03559H.

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)

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

which penalises both over and under-confident predictions.

Table A1. Comparison of NLPD values on the soil phosphorus fraction dataset. Standard errors are reported for 10 independent train/test splits. Lower scores are better.

Soil Phosphorus Fraction Dataset	GP	Het GP
NLPD	1.35 ± 1.33	1.00 ± 0.95