Systems & Control Transactions

**Original Research Article**
*Peer Reviewed Conference Proceeding*
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

# Bayesian Optimization Framework for Agrochemical Formulation Design

**Yipei Zhao[a] Tong Liu[b], Robin Wesley[c] and Joan Cordiner[a]**

[a] School of Chemical, Materials and Biological Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom
[b] School of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom
[c] Syngenta, Jealott's Hill International Research Centre, Bracknell RG42 6EY, United Kingdom
* Corresponding Author: j.cordiner@sheffield.ac.uk.

## ABSTRACT

Manufacturing kinetically stable products remains a challenge in the agrochemical industry. Current agrochemical formulation design relies on semi-empirical and trial-and-error methods. The inconsistency is caused by the lack of a mechanistic understanding of the formulation, making the design a black-box optimisation problem. In addition, validating the ground truth of the high-dimensional design space is expensive, driving chemists to explore possible solutions using data-driven methods. We proposed a Bayesian optimisation framework employing a Gaussian process as the surrogate model to intelligently guide the screening of the design space. The uniqueness of our framework is the application to the classification task to increase the number of hits of stable formulation recipes. The framework was tested on a provided industry dataset with a focus on emulsifiable concentrates. The performance reached a comparable accuracy with only ~25% of the data being sampled and hit more stable formulations than a Monte Carlo search. Our framework accelerates the discovery of stable recipes and guides formulation screening.

**Keywords**: Agrochemical Formulation, Bayesian Optimisation, Gaussian Processes, Space-Filling Designs, Machine Learning

## INTRODUCTION

Developing stable and effective agrochemical formulations – specifically Emulsifiable Concentrates (EC) remains a challenging bottleneck in product development. The formulation design space is high-dimensional and is characterised by limited mechanistic understanding of the complex chemical interactions between active ingredients, surfactants, and solvents when diluted in water [1]. Conventionally, these complexities are addressed through iterative trial-and-error campaigns [2]. However, these methods are inefficient for exploring vast compositional design spaces and often fail to identify optimal formulations within tight industrial timelines. Moreover, the identification of subsequent candidate formulations is non-trivial. As empirical models often fail to generalise, researchers are forced into unguided experimental loops that waste materials and stall development.

Modelling this black-box function presents common constraints in chemical engineering: the gradient information ($\nabla f$) of the chemical interaction is usually unknown; thus, gradient descent methods cannot be applied [3]. Another challenge is that the cost of evaluations (experimental validation) is also expensive in labour and time to verify the function sufficiently.

To address these limitations, we present a data-driven, closed-loop Bayesian optimisation (BO) framework that systematically accelerates the formulation development process. Unlike the traditional design of experiments (DoE), our approach employs BO to actively learn the composition-stability relationship. By treating the formulation problem as a classification task (homogeneous vs. inhomogeneous), the framework iteratively proposes the next set of experimental candidates to maximise the probability of identifying a homogeneous recipe.

## METHODOLOGY

Mathematically, the challenge can be formulated into $\mathrm{x}^{*} \in \arg\max_{\mathrm{x}\in\mathrm{X}} \mathrm{f}(\mathrm{x})$ where the design space is represented by $\mathcal{X} \subset R^{d}$, and the black-box function is

represented by matching the input $x$ with output $f(x)$. Modelling the black-box function $f$ is usually expensive in the chemical engineering field, due to the following constraints:

- **No analytical form.** The underlying physics/chemical reactions between materials are difficult to evaluate and express in closed-form equations.
- **Derivative-free.** The gradient information $\nabla f(x)$ is unknown. The gradient descent approach can't be applied.
- **Expensive evaluations.** Conducting experiments to verify the function can be costly.

## Bayesian Optimisation

BO is a sample-efficient, iterative strategy designed for the global optimisation of expensive, black-box objective functions. Unlike Random Search or traditional DoE, which typically rely on static, pre-determined sampling plans, BO is inherently adaptive. It intelligently guides subsequent experiments by balancing the model's explored region (exploitation) with the unexplored region (exploration) [4]. This allows the algorithm to focus the search on the most promising areas of the design space, achieving the optimal solution with significantly fewer experimental runs than conventional methods.

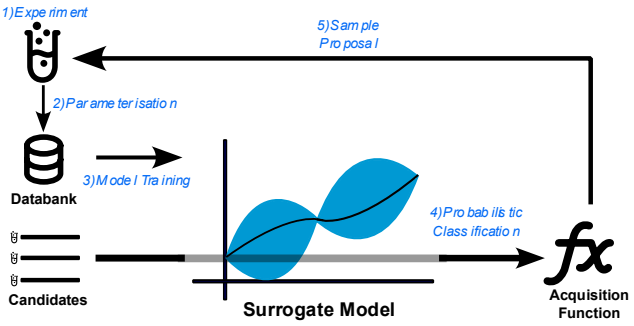



**Figure 1.** The iterative cycle of BO.

The BO procedure is an iterative strategy designed to optimise computationally expensive black-box functions. The workflow is characterised by the following stages:

(1) **Initial Sampling:** An initial set of experimental observations is collected.

(2) **Data Parameterisation:** The raw experimental inputs and corresponding outputs are transformed into a structured database for mathematical modelling.

(3) **Surrogate Model Construction:** A probabilistic surrogate model—most commonly a Gaussian Process (GP)—is trained on the available dataset to approximate the underlying objective function.

(4) **Acquisition Function Evaluation:** An acquisition function is employed to balance the exploration-exploitation trade-off. It utilises the surrogate model's mean predictions and associated predictive uncertainty to identify the most promising candidate for the next iteration.

(5) **Iterative Optimisation:** The suggested candidate is experimentally validated, the resulting data is appended to the training set, and the model is updated. This cycle repeats until a predefined termination criterion is reached.

The BO contains two core components: a surrogate model and an acquisition function. The surrogate model acts as the inference tool of the optimisation process, updating the probabilistic model given sampled data $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{t}$. The acquisition function is the decision maker, selecting the next point $x_{t+1}$ to sample by balancing exploitation (sampling regions known to be more promising) and exploration (sampling regions with higher uncertainty) using a predefined hyperparameter.

A **Gaussian process** is usually considered as the gold standard surrogate model as they provide reliable uncertainty estimates, analytic tractability and sample efficiency [4]. The Gaussian process is defined by a mean function $m(x)$ and a covariance function (also known as the kernel) $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{1}$$

both mean functions $m(x)$ and covariance function $k(x, x')$ are the prior beliefs of the objective function. They are critical in training as they assume the smoothness and periodicity, before any data is observed by the model. The mean function is usually considered to be $m(x) = 0$, while the selection of the covariance function is trickier. The Matérn kernel family offers a good balance between smoothness and noise-handling. The Matérn kernel $k_{a,h}: \mathcal{X}: \mathcal{X}$ is defined by:

$$k_{a,h}(x, x') = \frac{1}{2^{\alpha-1}\,\Gamma(\alpha)} \left( \frac{\sqrt{2\alpha}||x - x'||}{h} \right)^{\alpha} K_{\alpha} \left( \frac{\sqrt{2\alpha}||x - x'||}{h} \right)^{\alpha}, \tag{2}$$

Where $\Gamma$ is the gamma function, and $K_a$ is the modified Bessel function of the second kind of order $\alpha$ [5]. More details of the Gaussian process can be found in [6].

The other core component of BO is the acquisition function. In this work, we employed the Upper Confidence Bound (UCB) as the acquisition function. The UCB is defined as:

$$\alpha_{UCB}(x) = \mu(x) + \kappa \cdot \sigma(x), \tag{3}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the prediction from the surrogate model, respectively,

and $\kappa$ is a predefined hyperparameter to balance exploitation and exploration. Compared to other acquisition functions, UCB is mathematically straightforward, intuitive, and based solely on the prediction and its uncertainty.

## Space-Filling Designs

Space-filling designs aim to encourage the diversity included in the dataset, resulting in not only saving computational calculations but also leading to a more sophisticated model in interpolation/extrapolation by providing a better coverage or representation of the design space [7], [8]. In this work, we considered 3 space-filling design methods with random sampling.

**Latin Hypercube Sampling** (LHS) is a space-filling design strategy that improves upon simple random sampling by ensuring one-dimensional uniformity. For each of the $k$ input variables, the range is divided into $n$ equally spaced intervals, with exactly one observation per interval. Mathematically, a design point $d_{ij}$ (representing the $i$-th run and $j$-th factor) is calculated using the formula $d_{ij} = \frac{l_{ij}+(n-1)/2+u_{ij}}{n}$, while $l_{ij}$ is a value from a permutation of $n$ levels (specifically taken as $-(n-1)/2$ to $(n-1)/2$) which determines the specific interval, and $u_{ij}$ is a random number between 0 and 1 that provides a random position within that interval. This stratification ensures that the variance of the LHS estimator is lower than that of simple random sampling, specifically by subtracting the variance contributions of the main effects. As one of the most popular space-filling design strategies, LHS offers a reduction of the variance of the sample mean compared to random sampling and provides better coverage of the range of each input [9].

**Sobol Sequence** is a quasi-random number generator which offers lower discrepancy compared to a random number generator. A pseudo-random number generator can cause the data points to cluster together, leaving a large area unexplored. Unlike random sampling, where the discrepancy decreases probabilistically, Sobol sequences are constructed using radical inversion in base-2 to ensure the discrepancy decreases, resulting in a faster conversion rate [10].

Neither the LHS nor the Sobol sequence directly select samples from the existing database; instead, they generate a list of theoretical coordinates to sample. To sample from the existence database, samples are selected by comparing the distance between each data point to match the theoretical points proposed by both generators with the same dimension. Euclidean distance $d(p,q) = ||p-q||^2 = \sum_{i=1}^{m}(p_i - q_i)^2$ is used to measure the distance between data points pairwise.

**Maximin distance design** is more intuitive. By starting with a random sample, the algorithm iteratively selected samples that maximise the minimum Euclidean distance (can be substituted by other distance measures) from the sampled data points. This algorithm ensures that no data point will be close to other data points pairwise, thus ensuring a maximum spread of the samples.

# CASE STUDY

To evaluate our framework, we utilise an experimental dataset provided by Syngenta. The data was generated using a customised automated robotic platform specifically designed for high-throughput formulation design and screening. Our goal is to validate the framework using an experimental dataset by sampling only a proportion of the data and testing the model with unseen data points. In the context of this case study, our optimisation objective is to maximise the probability of identifying a kinetically stable (homogeneous) formulation using BO. The underlying objective is to optimise for the most stable formulation composition. The optimisation variable defining our design space is the continuous weight percentages of the raw materials comprising the formulation product. Consequently, unguided formulation screening is time and cost-intensive; thus, we employed BO to accelerate the screening process.

## Formulation Composition and Design Space

The dataset focuses on EC formulations. These formulations consist of a mixture of active ingredients (AI), solvents, and emulsifiers. To ensure a robust evaluation across a diverse chemical space, a total of 25 raw materials were tested:

- Active ingredients: Five unique AI and control formulations containing no active ingredients.
- Solvent Systems: Five unique solvent systems, comprising blends of two or three solvents (from 11 unique solvents) at various ratios.
- Emulsifier Systems: A primary anionic emulsifier along with six different co-emulsifier systems, each a blend of two emulsifiers selected from a pool of nine non-ionic emulsifiers.

## Data Acquisition

Following preparation, the EC formulations were diluted in water at a 1:100 ratio to simulate practical manufacturing conditions. Kinetic (short-term) homogeneity was assessed 24 hours after dilution using a high-throughput image module.

To maximise information gain and capture subtle stability characteristics, the photography module utilised a multi-exposure technique. Each final data point is a concatenated image with a resolution of 9328 ×2550 pixels. These high-resolution images are an aggregation of 13 separate captures taken at varying exposures and

| Category | Components | Count | Percentage/Details |
|---|---|---|---|
| **Dataset Statistics** | Total Samples ($N$) | 811 | 100% |
| | Homogeneous | 341 | 42.05% |
| | Inhomogeneous | 470 | 57.95% |
| **Formulation Components** | Active Ingredients (AI) | 5 | Plus AI-free controls |
| | Solvent Systems | 5 | 11 unique solvents |
| | Emulsifier Systems | 6 | 9 unique emulsifiers |
| **Total Unique Materials** | | **25** | |

**Table 1.** A detailed description of the dataset and label distributions.

laser intensities.

## Data Labelling & Distribution

The labels of all the formulations are evaluated via an automated image analysis tool designed to classify formulations based solely on visual characteristics. This tool achieved a binary classification accuracy of 94.83%, providing a highly reliable and scalable method for processing large image batches.

As summarised in Table 1, the final dataset consists of 811 samples. The distribution is relatively balanced, with 341 samples (42.05%) labelled as homogeneous and 470 samples (57.95%) labelled as inhomogeneous.

The model's input is compositional data, consisting of the weight percentages of each unique material. The input has been transformed using the centred log ratio (CLR) transformation and a standard scaler to make the model more sophisticated. The CLR transformation is used to handle the feature of compositional data (all data points have a constant sum) and preserves the mean distances [11]. The dataset was partitioned into two subsets using a 70:30 ratio. Seventy per cent of the data constituted the sampling pool, from which all initial and subsequent samples were drawn. The remaining 30% functioned as a held-out test set to evaluate model performance.

## Optimisation Strategy

Since we are working on a classification rather than a regression model, we employed a different approach than Gaussian process regression. The construction and training of Gaussian process models are carried out using the Python library GPflow [12]. A Gaussian process classifier (GPC) can be slightly different to a regression task. A latent function must be applied to convert the output to the desired likelihood, which is a Bernoulli likelihood for a binary classification problem.

We have chosen Matérn kernel with $\alpha = 5/2$ (also known as the Matérn 5/2 kernel) as it provides a smoother result than other common hyperparameters. The GP models' other parameters, such as length scale and variance will be optimised by the GPflow library. The hyperparameter $\kappa$ of UCB was chosen to be 1.5 to balance exploitation and exploration.

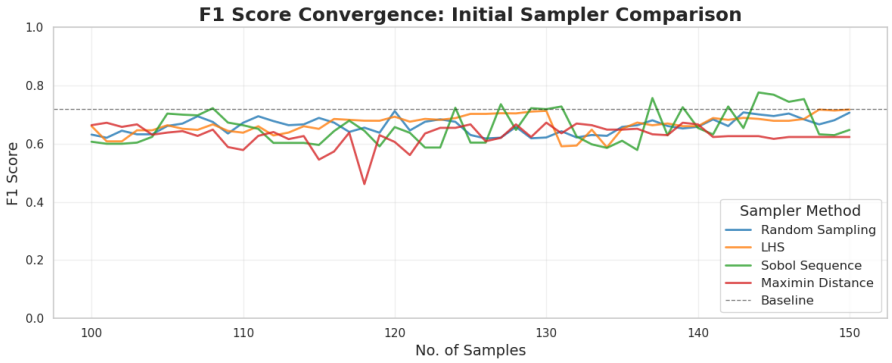
## RESULT

We applied our framework to the provided dataset with an initial sample size of 100 and ran 50 iterations. As illustrated in Figure 2(a), the diverse sampling strategies fluctuate in performance but generally demonstrate the ability to achieve high predictive accuracy with significantly reduced data requirements. While the baseline model required 500 random samples to achieve an F1 score of 71.9%, the BO framework achieved comparable F1 scores (ranging between 0.6 and 0.8) with only 100 to 150 samples. Among the space-filling designs, the Sobol Sequence demonstrated strong potential, peaking above the baseline in the final iterations. This framework reduced the experimental burden by approximately 70% compared to the baseline random sampling required to achieve the same F1 score.

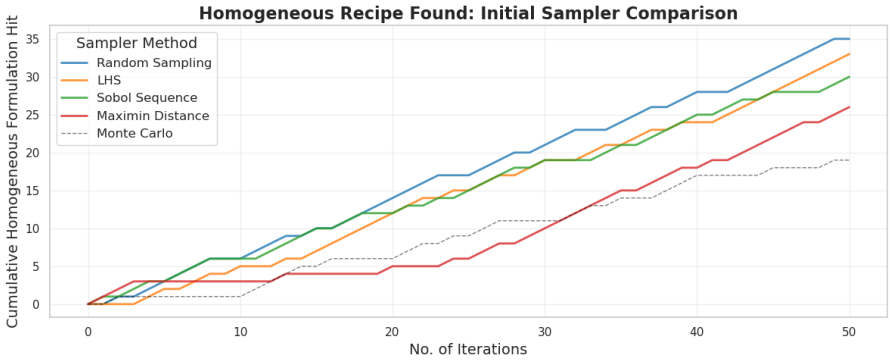
This underperformance may be attributed to the tendency of Maximin distance designs to prioritise the boundaries of the design space. Consequently, the central regions remain undersampled, leading to poorer generalisation in those areas compared to Sobol or Random initialisation.

Figure 2(b) details the efficiency of the framework in locating stable (homogeneous) formulations. The effectiveness of the BO approach is evident when compared to the Monte Carlo baseline (grey dashed line), which represents a pure random search without model guidance. The figure shows that the Gaussian model using random sampling achieves the highest hit rate (35 found), while demonstrating no superiority over LHS (33 Maximin distance (26 found). Whereas all Gaussian models found more homogeneous samples than a pure random search (19 found).

The result demonstrates the advantages of our framework, as all four different initial samplers captured more homogeneous formulations than the Monte Carlo method. This highlights the limitations of traditional trial-and-error approaches, which suffer from inconsistency and the inability to intelligently guide the search toward promising compositional regions.




**(a)**




**(b)**

**Figure 2. (a)** The change of the F1 score over iterations. The grey dashed line is the baseline model trained using a Gaussian process with random sampling with 500 sampled data points, which leads to an F1 score of 71.9%. **(b)** The cumulative number of homogeneous formulations identified over 50 iterations using different initial sampling strategies compared to a Monte Carlo baseline.

## CONCLUSIONS AND FUTURE WORK

This study presented a data-driven, closed-loop BO framework designed to address the inefficiencies in traditional trial-and-error practice for agrochemical formulation design. By treating the formulation design of EC as a binary classification task (homogeneous vs. found), the Sobol sequence (30 found), and outperforming inhomogeneous, we successfully demonstrated the framework's ability to navigate a high-dimensional design space with limited mechanistic understanding using Gaussian process models.

The next step will be applying the framework to a larger dataset to validate its performance against a dataset with higher dimensions and more samples. In addition, the idea can be applied to a multi-class classification task, but perhaps with slightly different approaches.

## DIGITAL SUPPLEMENTARY MATERIAL

Due to commercial confidentiality, specific compositional data cannot be provided. However, we have included representative images captured by the camera module of the automated formulation system: https://github.com/yipeizhao/ESCAPE-material.

## ACKNOWLEDGEMENTS

This is an iCASE awarded project, sponsored by Syngenta and the UK Engineering and Physical Sciences Research Council (EPSRC) under the grand code EP/Y528808/1.

## REFERENCES

1. Tadros TF, Ed., *Emulsion formation and stability*. in Topics in colloid and interface science. Wiley-VCH (2013). https://doi.org/10.1002/9783527647941
2. Montgomery DC, Design and analysis of experiments, Ninth edition. John Wiley & Sons, Inc. (2017).
3. Jones DR, Schonlau M. Efficient Global Optimization of Expensive Black-Box Functions'.
4. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning

Algorithms. arXiv preprint (2012). arXiv:1206.2944.
5. Kanagawa M, Hennig P, Sejdinovic D, Sriperumbudur BK. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. arXiv preprint (2018). arXiv:1807.02582.
6. Rasmussen CE, Williams CKI, *Gaussian processes for machine learning*, 3. print. In: Adaptive computation and machine learning. MIT Press (2008).
7. Joseph VR. Space-filling designs for computer experiments: a review. Quality Engineering 28:28-35 (2016). https://doi.org/10.1080/08982112.2015.1100447
8. Huang S. Surrogates: gaussian process modeling, design, and optimization for the applied sciences. Journal of Quality Technology 53:440-441 (2020). https://doi.org/10.1080/00224065.2020.1764416
9. Lin CD, Tang B. Latin Hypercubes and Space-filling Designs. arXiv preprint (2012) *arXiv*: arXiv:2203.06334.
10. Michel ES. Population Ecology in Practice. John Wiley & Sons, inc. (2020). ISBN: 9780470674147.
11. Galletti A, Maratea A. Numerical stability analysis of the centered log-ratio transformation. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) :713-716 (2016). https://doi.org/10.1109/sitis.2016.119
12. de G. Matthews AG, van der Wilk M, Nickson T, Fujii K, Boukouvalas A, Léon-Villagrá P, Ghahramani Z, Hensmen J. GPflow: A Gaussian process library using TensorFlow, *J Machine Learning Res* 18:1–6 (2017)

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See https://creativecommons.org/licenses/by-sa/4.0/