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

Pareto Front Guided Sampling for Efficient Bioprocess Experimentation

Stricker Samuel[ab*], Lucas Francisco dos Santos[c], Claus Wirnsperger[d], Alessandro Butté[d], Antonio del Rio Chanona[ab], Mehmet Mercangöz[ab], Gonzalo Guillén Gosálbez[c]

[a] Imperial College London, Department of Chemical Engineering, London, United Kingdom
[b] Sargent Center for Process Systems Engineering, London, United Kingdom
[c] ETH Zurich, Department of Chemical- and Bioprocess Engineering, Zurich, Switzerland
[d] DataHow AG, Zurich, Switzerland
* Corresponding Author: sam.stricker25@imperial.ac.uk

ABSTRACT

This work presents Pareto Front Guided Sampling (PFGS), a model-guided Design of Experiments (DoE) strategy for bioprocess development that makes the exploration–exploitation trade-off explicit and integrates human expertise into experiment selection. Starting from an initial experimental design, PFGS fits a probabilistic surrogate and then proposes new experiments by solving a multi-objective design problem that simultaneously rewards (i) high predicted performance (posterior mean) and (ii) high information gain (posterior uncertainty). Rather than collapsing this trade-off into a single acquisition value, PFGS generates a Pareto set of candidate experiments, that reflect different balances between improvement-seeking and learning. To prevent wasted runs, an automated screening step is performed to remove candidates in (i) low predicted-mean regions unlikely to yield near-optimal performance and (ii) low-uncertainty regions already well explained by the surrogate, concentrating effort on promising yet under-explored areas and encouraging coverage of multiple near-optimal basins. For industrial application, PFGS was extended to batch selection by choosing a diverse subset of Pareto candidates to support parallel experimentation without substantial loss in sample efficiency, and a termination rule to stop experimentation once user-defined accuracy thresholds are met. PFGS is benchmarked against Latin Hypercube Sampling (LHS) and Bayesian Optimisation (BO) on two representative problems: (i) a bioprocess case study with shallow performance gradients, where naive exploitation is unreliable and non-adaptive designs are inefficient, and (ii) a multi-modal landscape, where methods that focus on a single basin can miss alternative optima. Across these settings, PFGS allocates experimental resources more effectively than non-adaptive baselines and, in the multi-modal case, uniquely identifies and characterises all stationary regions while maintaining strong overall performance. By combining Pareto-transparent decision support, automated filtering, batch practicality, and principled stopping, PFGS provides an interpretable and resource-efficient DoE methodology for biopharmaceutical process optimisation.

Keywords: Design of Experiments (DoE), Bayesian Optimization (BO), Bioprocesses, Optimization, Pareto Front

INTRODUCTION

The biopharmaceutical industry faces mounting pressure to optimize processes due to elevated development costs, stringent regulatory requirements, and the inherent complexity of bioprocesses. Digital tools, including hybrid modeling and advanced statistical methodologies, have become essential for reducing development timelines and ensuring consistent product quality. Design of experiments has emerged as a critical methodology for systematically identifying and optimizing the critical factors that influence bioprocess outcomes, thereby enabling researchers to efficiently navigate complex experimental domains [1].

Classical approaches, such as Full Factorial Design (FFD), remain widely utilized in the pharmaceutical

industry owing to their simplicity and effectiveness in systematically mapping investigation spaces [2]. However, these methodologies prove insufficient for the demands of modern biopharmaceutical development, which increasingly necessitate rapid iterative refinement and stringent cost management [1]. Classical DoE methods typically necessitate a large number of experiments, which introduces substantial costs and extended timelines. Furthermore, conventional DoE frameworks, including FFD, are fundamentally limited in their capacity to model highly nonlinear processes, as they are restricted to approximating response functions using first- or second-order polynomial equations. Although LHS provides improved coverage for complex response surfaces, it similarly lacks adaptive and iterative refinement capabilities.

In contrast, model-based DoE approaches based on BO are becoming increasingly relevant because they (i) accommodate model uncertainty as actionable information for decision-making, and (ii) enable iterative experimental design optimization through systematic incorporation of knowledge from preceding experiments [1]. Methods such as Batch Bayesian Optimization (BBO) facilitate simultaneous optimization of multiple parallel experiments. Despite their considerable potential, these approaches remain underutilized in industrial settings, primarily due to challenges associated with integrating complex algorithms into established workflows [3]. Moreover, these methods are primarily designed for optimization rather than comprehensive mapping of the optimal region, which is essential for developing robust models capable of generalizing across diverse operational conditions.

Problem Statement

This work addresses the development of a comprehensive Design of Experiments framework designed to efficiently guide experimentation in bioprocesses. The overarching objectives are:

- to optimize an objective function (e.g., maximize product concentration),
- to systematically explore the investigation space such that the underlying, often nonlinear, process behavior can be accurately modeled with minimal experimental burden, and
- to accommodate practical constraints inherent to industrial operations (e.g., parallel experimentation or discretization).
- to provide transparent, human-interactable results that facilitate domain expert engagement and enable systematic incorporation of prior knowledge into the experimental design process.

The primary goal is to efficiently acquire process knowledge in regions of the investigation space near optima. This necessitates balancing between exploitation of known promising regions and exploration of uncertain or under-sampled areas, while simultaneously accounting for process nonlinearity and operational constraints of real-world systems. Furthermore, the practical requirements for adherence to experimental timelines and parallelization of resources motivate the development of algorithms capable of proposing multiple diverse, high-performing experimental candidates concurrently.

A fundamental challenge lies in the inherent trade-off between exploitation and exploration. Optimization-focused sampling strategies tend to concentrate proposed experiments in regions of high performance, potentially overlooking broader system dynamics and behavior. Additionally, designs proposed by such algorithms like BO are often not easily human-interpretable and often feel random and unfounded. Conversely, space-filling design approaches enhance model generalizability but may allocate resources inefficiently to regions of limited practical interest. This inefficiency is further compounded by the curse of dimensionality, whereby the investigation space expands exponentially with increasing numbers of input factors, rendering uniform sampling impractical. Effective DoE must therefore balance these competing objectives, emphasizing relevant design subspaces while maintaining sufficient global representativeness.

Industrial applications impose additional constraints that must be addressed: input variables may exhibit discretization due to measurement limitations, categorical factors introduce discontinuities, and operational feasibility considerations and Critical Quality Attributes (CQAs) substantially constrain the accessible design space. Consequently, any newly proposed DoE framework must integrate three fundamental aims: (i) identifying high-performing regions within a multi-dimensional investigation space, (ii) ensuring sufficient coverage to reliably capture process behaviour, and (iii) adhering to practical constraints, like batch operating modes, typical of industrial bioprocessing.

In summary, there is a critical need for integrated methodologies that efficiently balance exploration, exploitation, and practical operational considerations to accelerate biopharmaceutical process development under constraints of cost and regulatory complexity. Importantly, such advancement does not necessarily require the development of novel or more sophisticated acquisition functions for BO. Rather, existing algorithms should be adapted to enhance human-machine interaction and facilitate the systematic incorporation of domain expertise, thereby improving user accessibility and comprehensibility. This emphasis on human-centered design is essential, as human decision-making remains a predominant factor in contemporary process development

workflows.

METHODS

Classical Design of Experiment Techniques

Full Factorial Design is a statistical experimental methodology in which each factor is evaluated at n discrete levels, and all possible combinations of factor levels are systematically evaluated within the investigation space (an example is presented in Figure 1 to inform model development. This approach enables investigation of both the effects of each variable. However, the experimental complexity of FFD increases exponentially with the number of factors, rendering this method impractical in high-dimensional investigation spaces [4]. Despite this limitation, Fractional Factorial Design, a subset of FFD utilizing a reduced set of the experimental runs that only mildly mitigates the problem of the number of experiments, remains widely utilized in pharmaceutical research applications. Its continued adoption is attributable to its simplicity and experimental efficiency. Furthermore, the results obtained from factorial designs can be directly interpreted by researchers without requiring the development of a formal predictive model [2].

One of the most widely employed Monte Carlo-based sampling techniques for systematic space exploration is Latin Hypercube Sampling. Applied across diverse scientific and engineering disciplines, LHS is extensively utilized for uncertainty quantification and reliability analysis [5]. The foundational methodology, stratified sampling, partitions the investigation space into distinct strata, which are weighted according to their respective conditional probabilities. In LHS, each dimension is subdivided into n_s mutually non-overlapping subspaces. From these subspaces, random pairings are performed across dimensions such that each value from each dimension is selected exactly once without replacement, yielding n_s experimental designs of dimensionality h_d. A representative example of a LHS in two dimensions is illustrated in **Figure 1.** LHS and its variants are generally preferred for their computational efficiency and demonstrated capacity to capture nonlinear behavior across the investigation space with a minimal number of experiments.

Bayesian Optimization

Bayesian optimization is employed for the optimization of computationally expensive, black-box functions, such as in materials design applications [6]. BO utilizes a probabilistic surrogate model, typically a Gaussian process (GP) regressor, to approximate the objective function and to systematically guide the selection of promising candidate points [7]. The method balances exploration (sampling regions of maximum uncertainty) and exploitation (optimizing the objective function) through the implementation of acquisition functions [6], which are subsequently maximized using standard optimization algorithms. A widely employed acquisition function is the Upper Confidence Bound (UCB), which manages the exploration-exploitation trade-off by adopting an optimistic perspective regarding uncertainty. UCB is computed according to:

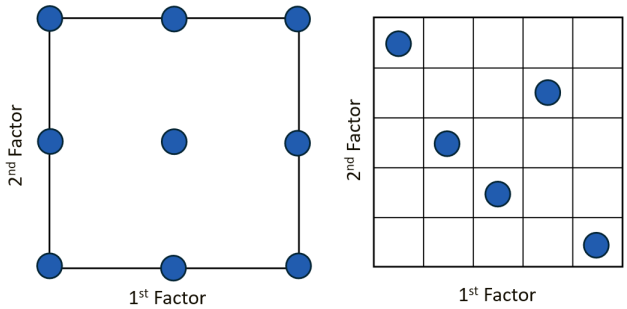

Figure 1 Example of a 3-level FFD (left) and a LHS (right) in two dimensions.

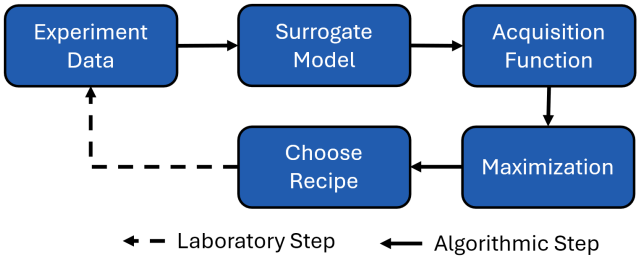

Figure 2 Overview of BO framework. BO uses the experimental data to train a surrogate model, from which an acquisition function is derived. This function is then maximized to find the most promising recipe.

$$\mathrm{UCB}(\boldsymbol{x}) = \mu(\mathbf{x}) + \beta * \sigma(\mathbf{x});\ \beta > 0 \qquad (1)$$

where μ and σ represent the mean and standard deviation of the surrogate model at point $\boldsymbol{x}$, respectively, while β is a hyperparameter that controls the exploration versus exploitation balance [8]. Alternatively, more sophisticated acquisition functions include Expected Improvement and Probability of Improvement. A representative workflow for BO is illustrated in **Figure 2**.

Batch Bayesian Optimization

The simultaneous sampling of multiple experimental designs using a BO approach is referred to as BBO [9]. BBO extends the principles of standard BO by enabling the concurrent evaluation of multiple candidate points in parallel, rather than sequentially. This parallelization offers substantial advantages in scenarios where multiple experiments or simulations can be conducted simultaneously, such as in distributed computing environments or laboratory settings equipped with multiple testing units. The primary challenge in BBO is that the acquisition function must be modified between successive point

selections to prevent redundant sampling of the same design point [9].

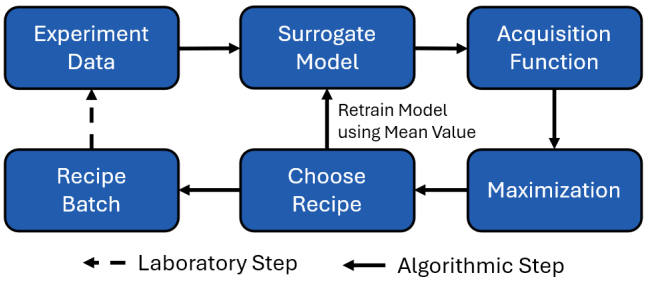

Figure 3 Overview of BBO. In q-point BO experimental data is used to generate a surrogate model from which an acquisition function is then maximised. The mean of the optimal point $\boldsymbol{x}$ is then added to the experimental batch and the surrogate model is retrained. This iterative procedure is performed until q points are sampled.

In classical BO the maximum of the acquisition function is identified and an experiment at the corresponding point $\mathbf{x}$ is conducted to obtain the true response y. This result is subsequently used to retrain the surrogate model. In q-point BO, a BBO method, no experiment is performed after the first BO iteration. The retraining procedure is performed by employing the mean prediction of the surrogate model $\mu(\boldsymbol{x})$ as a surrogate true value. After iterating this procedure n_q times, where n_q represents the batch size, all n_q experiments are executed in parallel, and the surrogate model is retrained with the empirically obtained values [9]. This approach is called BBO with Kriging Believer. The procedural flow of this algorithm is visualized in **Figure 3**.

Pareto Front Guided Sampling

In this work, a novel DoE methodology termed Pareto Front Guided Sampling was developed. The fundamental principle underlying PFGS is the generation of a Pareto front that explicitly addresses the exploration-exploitation trade-off. The methodology proceeds as follows: the surrogate model is initially trained on available experimental data. Subsequently, a Pareto front is generated to identify candidate points representing optimal compromises between competing objectives, like exploration (posterior mean) and exploitation (posterior uncertainty). An initial experimental design is selected from this Pareto front, and additional candidate designs are systematically identified based on the maximum Euclidean distance between candidate parameter vectors within the normalized, predefined investigation space. Once a sufficient number of candidate designs has been identified, the corresponding experiments are executed in a controlled laboratory setting, and the surrogate model is updated with the empirically obtained results. A comprehensive schematic representation of the PFGS algorithm is presented in **Figure 4**.

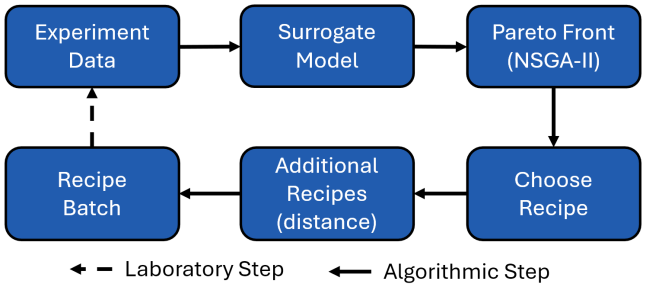

Figure 4 Overview of PFGS framework. The PFGS algorithm uses experiments to train the surrogate model from which a Pareto front is generated. The recipes are selected and experiments performed.

Pareto Front Generation

The Pareto front is generated by simultaneously optimizing the mean prediction of the surrogate model (representing exploitation) and the corresponding standard deviation (representing exploration) using a Multi Objective Optimization (MOO) Algorithm. The resulting Pareto front is subsequently refined through the elimination of candidate points from two distinct regions, as illustrated in **Figure 5**. First, candidate points characterized by low mean prediction are systematically removed, as these regions are of limited interest for identifying optimal process conditions. Second, candidate points located in regions of low standard deviation are removed. This elimination strategy serves two purposes. On the one hand it mitigates the influence of process measurement noise by focusing on dominant trends. On the other hand it ensures that the resulting sampling points are appropriately distributed throughout the investigation space, leveraging the property that standard deviation increases with distance from previously sampled experimental designs.

Recipe Selection

To determine the next experimental design in the batch, a candidate point must be selected from the remaining Pareto front (white region in **Figure 5**). In principle, all remaining points on the Pareto front are equally suitable for sampling. However, the specific selection depends on user preferences and objectives. The selection of the first recipe employs the utopia point as a reference. The utopia point is defined as the hypothetical point simultaneously maximizing both the mean prediction and standard deviation. Ideally, this utopia point (green) would be directly sampled; however, since it does not exist in the investigation space, the geometrically nearest point is selected using the Euclidean distance calculated in the normalized parameter space. For subsequent selections within the batch setting, the remaining candidate designs are sequentially identified from the Pareto front by maximizing the Euclidean distance (in the input space)

to previously selected experimental designs and executed experiments.

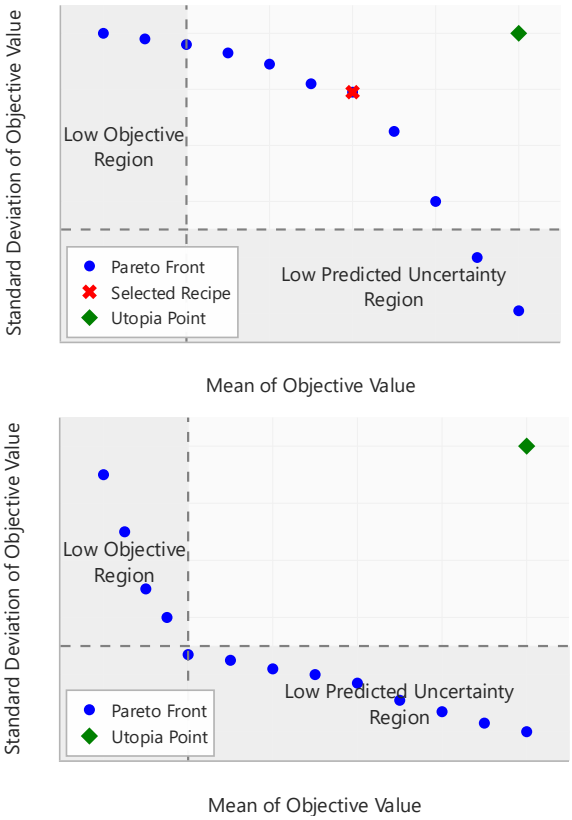

Figure 5 (top) The generated Pareto front using the model's mean and standard deviation. The gray areas are not sampled due to the low objective values or low uncertainty. (bottom) The Pareto front shown in blue is moving towards lower standard deviations as the model's knowledge is improving, resulting in the case that no point of the Pareto front is selected.

Stopping Criteria

The algorithmic structure enables termination of experimental design generation when the surrogate model achieves a specified level of accuracy within the target region of interest. **Figure 5** illustrates the Pareto front at an early sampling stage, where numerous candidate points occupy the white region from which the algorithm draws samples. However, as iterations progress and the model improves through successive refinement, the Pareto front gradually shifts toward regions of lower standard deviation. Eventually, no candidate points remain within the white sampling region.

This termination condition can occur at two distinct stages of the algorithm. First, termination may occur during batch selection, where insufficient candidate points remain in the sampling region to complete the full batch. For example, if five experimental designs were requested but only three viable candidates remain in the white region, the algorithm terminates early, resulting in an adapted (reduced) batch size. This indicates that continued sampling of the investigation space is not justified for the current model state. To obtain the remaining designs, either the elimination barrier that constrains sampling to high-performing regions must be relaxed, or a higher accuracy threshold must be set. Alternatively, the three available designs can be executed, and sampling can resume in the subsequent iteration. Second, termination may occur when no candidate points remain in the white sampling region at the initiation of an iteration. In this case, the algorithm ceases entirely, indicating that the surrogate model has achieved a sufficient level of training for the specified objectives.

Case Study

To demonstrate the efficacy of the proposed PFGS algorithm, two case studies of varying complexity were selected. To facilitate interpretability and explainability, the dimensionality was intentionally constrained to two input parameters (pH and temperature) and one output parameter (product concentration) across both case studies. The studies should represent simplified practical objectives for bioprocesses. In both cases no noise was added to the data.

Bioprocess

The first case study is based on a simplified Chinese Hamster Ovary (CHO) cell bioprocess model described in [10] that models a product concentration by varying pH and temperature. A visualization of the process response surface is presented in **Figure 6**. This case study presents the challenge of a relatively shallow gradient across most of the investigation space, which consequently renders optimization computationally challenging.

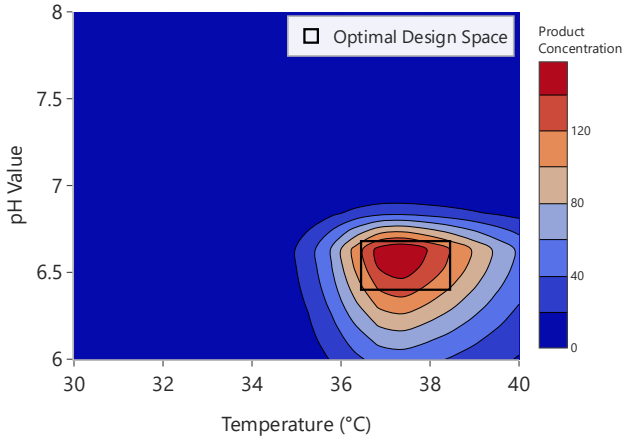

Figure 6 The optimization landscape of the bioprocess optimization problem.

Adapted Himmelblau Function

A second case study was developed using the well-established Himmelblau function [11]. The function was adapted to a maximization problem with an investigation space equivalent to that employed in the first case study. This case study presents the distinct challenge of systematically identifying and characterizing multiple local maxima. From a practical perspective, the global maximum (located in the upper left region) may be characterized by excessively constrained input variable ranges, rendering implementation infeasible in an industrial context. Conversely, the broader local maxima positioned in the right region of **Figure 7** offer more favorable practical operating conditions.

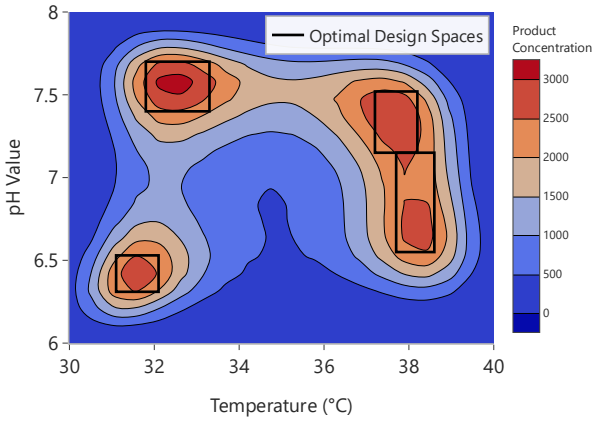

Figure 7 The adapted Himmelblau function to the optimization capabilities of the algoritms when multiple optima are present.

Experimental Setup

To evaluate the algorithm's performance, the optimal design regions are identified and delineated by black boxes. These regions are then used to generate test datasets by systematic sampling from the ground truth function. The resulting test datasets are presented in **Figure 8**.

Three distinct algorithms were compared in this study: LHS, BO, and the proposed PFGS. For both BO and PFGS, batch implementations were additionally evaluated. An initial sampling phase was conducted using LHS for all approaches, and a GP was trained on the resulting experimental data. Following initialization, the Relative Root Mean Squared Error (RRMSE) shown in **Equation 2** was calculated using the two previously defined test datasets.

$$RRMSE = \frac{\sqrt{mean((y_{predicted} - y_{test})^2)}}{std(y_{test})} \quad (2)$$

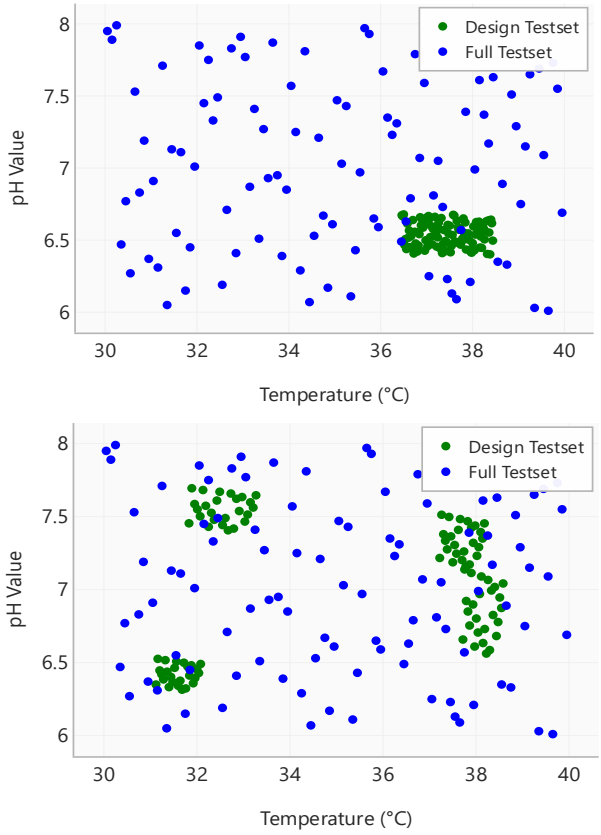

Figure 8 The test set for the classical Bioprocess (top) and the adapted Himmelblau function (bottom). The blue points correspond to the full dataset and the green ones to the design dataset.

$y_{predicted}$: test set values predicted with GP

y_{test}: true test set values

Subsequently, experimental designs were generated according to the respective algorithmic approaches, a GP was trained and the RRMSE was recalculated after each sample or batch of samples using the testset in order to track model improvement.

Implementation

In this study the data was sampled using LHS sampling implementation of the pyDOE2 package [12]. For all algorithms a GP was trained using botorch [13]. The MOO of the PFGS was performed using the NSGA-II algorithm [14] of the pymoo package [15], and BBO was used in the Kriging Believer setting with and UCB acquisition function with $\beta = 1.96$ and maximized using particle swarm optimization [16] with the pyswarm package [17]. The code is available online on the Autonomous Industrial Systems Laboratory GitHub Page.

RESULT & DISCUSSION

Results were generated using the same five random

seeds initializations across the five independent runs for each algorithm. All plots display the mean RRMSE calculated from the five replicate runs and the 95% confidence interval. However, for the PFGS algorithm in the first case study, the stopping criterion was occasionally satisfied prior to completion of all five runs, resulting in mean RRMSE values calculated from fewer than five independent replicates. For automated boundary selection in PFGS, threshold values constraining the objective function and uncertainty metrics must be specified. In the first case study, a minimum objective value of 20 and a standard deviation threshold of 10 were established. For the adapted Himmelblau function, minimum objective and standard deviation thresholds of 1500 and 1000, respectively, were employed. These threshold values were derived from the initial Pareto front. However, in a human-in-the-loop setting, these parameters could be iteratively adjusted during each sampling iteration. For reproducibility and standardization purposes, such manual adjustments were not implemented in this study.

Bioprocess

The first bioprocess case study (**Figure 9**) clearly demonstrates that algorithmic guidance is essential for efficient identification of the optimal region within the investigation space. LHS exhibited the poorest performance, which is attributable to the relatively small region of practical interest compared to the total investigation space size. Consequently, LHS allocated experimental resources inefficiently to regions of negligible interest.

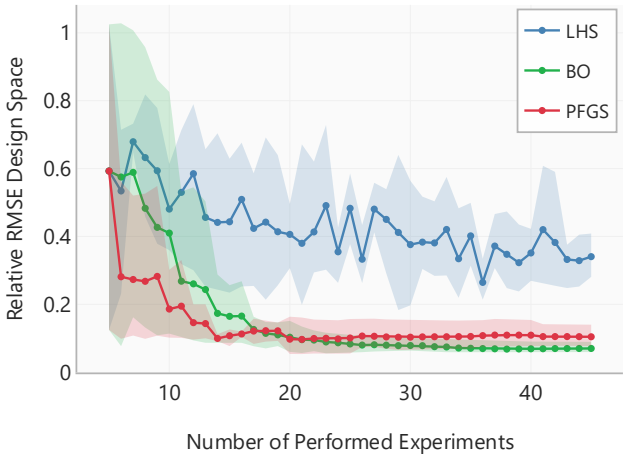

Figure 9 Bioprocess case study: Performance on the Design space test set of LHS, BO and PFGS with one experiment taken at each iteration.

Figure 10 presents the results for the batch optimization setting, where five parallel experiments were conducted simultaneously in each iteration. Both PFGS and BBO demonstrated comparable performance trajectories. When these batch results are compared to the sequential sampling approach (one experiment per iteration, shown in **Figure 9**), a modest decline in model performance is observed upon transition to the batch setting. This performance reduction is a well-established phenomenon in batch optimization, arising from the delayed model updating schedule; the surrogate model is updated only after completion of all five experiments within each batch, rather than after each experiment.

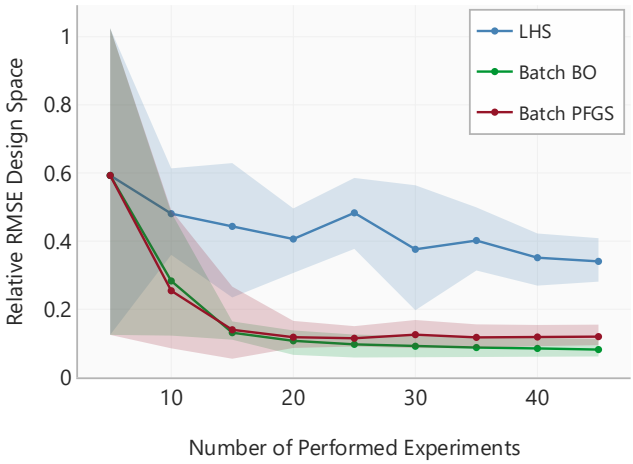

Figure 10 Performance on the Design space test set of LHS, Batch BO and Batch PFGS with five experiments taken at each iteration.

Adapted Himmelblau Function

This case study presents a more complex landscape characterized by three local optima and one global optimum. BBO encounters limitations in this setting, as it is fundamentally designed to identify a single maximum rather than to comprehensively map regions of interest. This limitation is evident in **Figure 11**, where the BO algorithm has a high RRMSE. As shown in **Figure 12** BO converges to a single optimum and focuses subsequent sampling efforts on local exploitation rather than exploring the different optima like PFGS. LHS demonstrated substantially improved performance in this case compared to the first example, attributable to the distributed spatial arrangement of the optima across the investigation space. However in needs to be mentioned that LHS is not suited in a practical setting due to its noniterative nature. Finally, PFGS allocates experimental resources more efficiently and does not wastefully concentrate sampling in subregions of negligible interest. It demonstrated comparable performance to LHS while possessing the distinct advantage of identifying and systematically sampling all four optima.

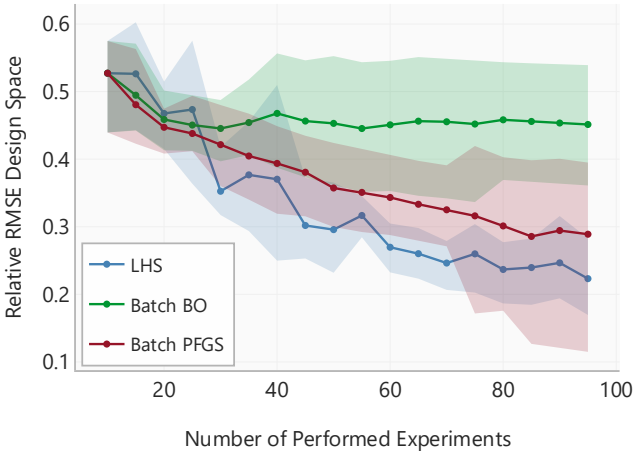

Figure 11 Himmelblau Case Study: Performance on the Design space test set of LHS, BBO and PFGS with one experiment taken at each iteration.

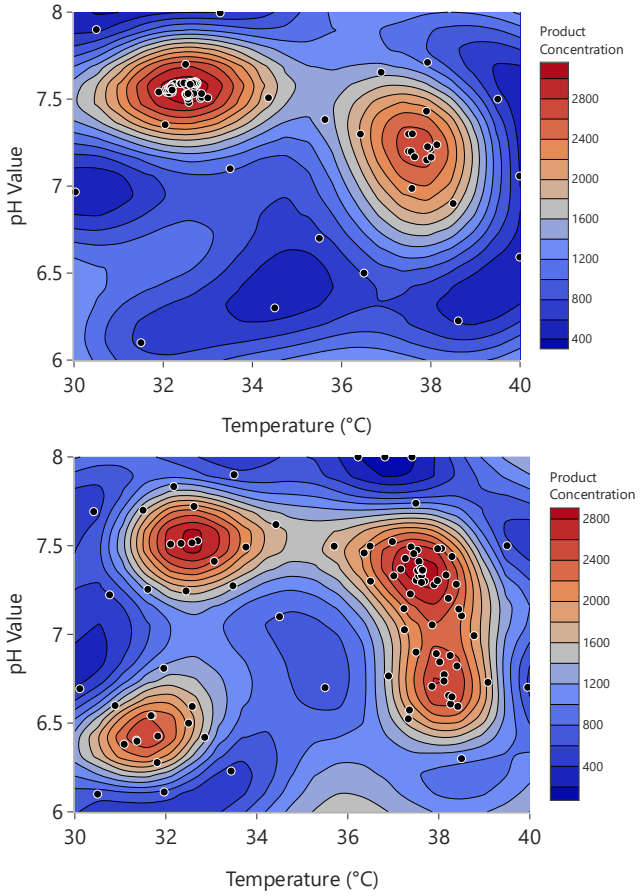

Figure 12 The location of sampled points (95 total) in the investigation space for one seed of Batch BO (top) and Batch PFGS (bottom). Additionaly a contour plot of the trained GP on those samples is visible.

The performance across the entire investigation space is illustrated in **Figure 13**. Both optimization-focused algorithms demonstrate comparatively poor performance in regions of low objective values. This outcome reflects the inherent trade-off previously described wherein improved accuracy in high-performing regions is achieved at the cost of reduced sampling coverage in regions of low objective values. However, this sampling bias is not problematic from a practical standpoint, since in industrial applications, experimental resources devoted to regions of limited practical interest represent an inefficient allocation.

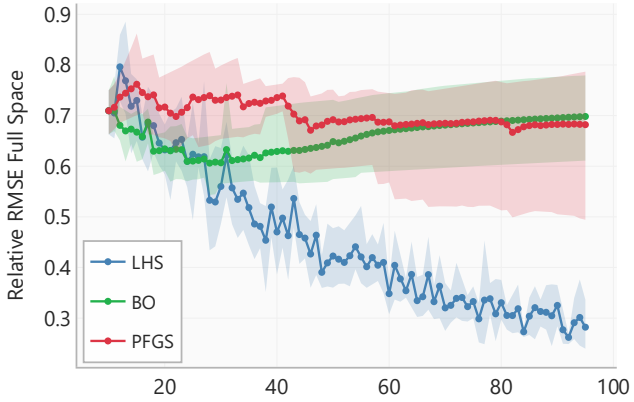

Figure 13 The performance of the algorithms on the full space (blue points in **Figure 8** (bottom) for the Himmelblau case study).

CONCLUSION

This work presents Pareto Front Guided Sampling, a novel DoE methodology that systematically addresses the fundamental trade-off between exploitation and exploration in nonlinear bioprocess optimization. The proposed algorithm generates a Pareto front representing simultaneous optimization of the surrogate model's predictive accuracy (mean prediction) and uncertainty quantification (standard deviation), enabling intelligent selection of experimental designs that balance the identification of high-performing with consideration of undersampled regions.

In the first bioprocess case study, characterized by shallow gradients across the investigation space and one optima, PFGS showed similar performance to BO, efficiently allocating experimental resources to regions of practical interest. However, the ability of PFGS to better distribute experiments might still lead to increased performance relative to BO, particularly in high-dimensional settings or under substantial process noise; these aspects were not investigated in this study and remain directions for future work. In the second case study with multiple local and global optima, PFGS outperformed BO and was able to uniquely identify and systematically sampling all stationary points, a capability unavailable to classical BO approaches optimized for single-point convergence. Furthermore, the batch implementation of PFGS

maintains performance while accommodating the practical requirement for parallel experimentation, a critical consideration in industrial bioprocessing workflows. It is crucial to emphasize that PFGS does not outperform classical BO strategies in a strict sense. Rather, within bioprocessing contexts, broader coverage of the optima landscape is often desirable, as it facilitates identification of optimal design spaces. In this framework, the uncertainty threshold effectively biases the acquisition toward increased exploration.

The algorithmic framework incorporates automatic termination criteria, enabling the method to cease experimentation once the surrogate model achieves specified accuracy thresholds within regions of interest. This feature substantially reduces experimental burden and development costs while respecting practical constraints inherent to industrial bioprocesses, including parameter discretization, categorical factors, and critical quality attributes.

OUTLOOK

Future extensions of this work will incorporate constrained optimization and multi-objective formulations to address more complex bioprocess scenarios requiring simultaneous optimization of multiple competing performance criteria. Additionally, multi-fidelity optimization strategies leveraging experimental data from distinct reactor types and scales are envisioned, enabling more efficient knowledge transfer across different bioprocessing modalities. In this study, the analysis was constrained to two-dimensional investigation spaces to facilitate the interpretability and validation of the PFGS methodology. However, future investigations should systematically evaluate algorithm performance in high-dimensional investigation spaces representative of realistic bioprocess scenarios with numerous controllable parameters.

ACKNOWLEDGEMENTS

This research was conducted during a Master's thesis research at ETH Zurich in collaboration with DataHow AG and was subsequently further developed during doctoral studies at Imperial College London. This work was supported by financial contributions from DataHow AG and the Department of Chemical Engineering Scholarship from Imperial College London.

AUTHOR IDENTIFIERS

Author ORCIDs:
Stricker S: 0009-0008-2261-8430
Francisco dos Santos L: 0000-0001-5931-4272
Wirnsperger C: 0009-0000-8511-1428
Butté A: 0000-0003-3506-3792
Del Rio Chanona A: 0000-0003-0274-2852
Mercangöz M: 0000-0002-4449-0414
Gosàlbez G. G: 0000-0001-6074-8473

REFERENCES

1. Kasemiire A, Avohou HT, De Bleye C, Sacre PY, Dumont E, Hubert P, Ziemons E. Design of experiments and design space approaches in the pharmaceutical bioprocess optimization. European Journal of Pharmaceutics and Biopharmaceutics 166:144-154 (2021). https://doi.org/10.1016/j.ejpb.2021.06.004
2. Chowdary KPR, Ravi Shankar K. Optimization of pharmaceutical product formulation by factorial designs: case studies. Journal of Pharmaceutical Research 15:105 (2016). https://doi.org/10.18579/jpcrkc/2016/15/4/108815
3. Möller J, Kuchemüller KB, Pörtner R. Bioprocess intensification with model-assisted doe-strategies for the production of biopharmaceuticals. Physical Sciences Reviews 9:2925-2945 (2023). https://doi.org/10.1515/psr-2022-0105
4. Antony J. Full factorial designs. Design of Experiments for Engineers and Scientists :63-85 (2014). https://doi.org/10.1016/b978-0-08-099417-8.00006-7
5. Shields MD, Zhang J. The generalization of latin hypercube sampling. Reliability Engineering & System Safety 148:96-108 (2016). https://doi.org/10.1016/j.ress.2015.12.002
6. P. I. Frazier, “A Tutorial on Bayesian Optimization, ” Jul. 08, 2018, *arXiv*.
7. Frazier PI, Wang J. Bayesian optimization for materials design. Springer Series in Materials Science :45-75 (2015). https://doi.org/10.1007/978-3-319-23871-5_3
8. Archetti F, Candelieri A. The acquisition function. SpringerBriefs in Optimization :57-72 (2019). https://doi.org/10.1007/978-3-030-24494-1_4
9. N. Hunt, “Batch Bayesian Optimization, ” 2020.
10. Siska M, Pajak E, Rosenthal K, del Rio Chanona A, von Lieres E, Helleckes LM. A guide to bayesian optimization in bioprocess engineering. Biotech & Bioengineering 123:805-830 (2026). https://doi.org/10.1002/bit.70129
11. D. M. Himmelblau, *Applied Nonlinear Programming*. McGraw-Hill, 1972.
12. R. Sjögren and D. Svensson, *pyDOE2 Design of Experiments for Python*. (2018).
13. M. Balandat *et al.*, “BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization, ” Dec. 08, 2020, *arXiv*.
14. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-

II. IEEE Trans. Evol. Computat. 6:182-197 (2002). https://doi.org/10.1109/4235.996017
15. Blank J, Deb K. Pymoo: multi-objective optimization in python. IEEE Access 8:89497-89509 (2020). https://doi.org/10.1109/access.2020.2990567
16. Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks 4:1942-1948 (None). https://doi.org/10.1109/icnn.1995.488968
17. James V Miranda L. Pyswarms: a research toolkit for particle swarm optimization in python. JOSS 3:433 (2018). https://doi.org/10.21105/joss.00433