frontiers | Frontiers in Manufacturing Technology

TYPE Original Research
PUBLISHED 08 January 2026
DOI 10.3389/fmtec.2025.1614335

OPEN ACCESS

EDITED BY
Kanak Kalita,
Vel Tech Dr. RR & Dr. SR Technical University, India

REVIEWED BY
Otilia Manta,
Romanian Academy, Romania
Carlo Graziani,
Argonne National Laboratory (DOE), United States

*CORRESPONDENCE
Lars Leyendecker,
lars.leyendecker@ipt.fraunhofer.de

RECEIVED 18 April 2025
REVISED 28 October 2025
ACCEPTED 28 November 2025
PUBLISHED 08 January 2026

CITATION
Leyendecker L, Gonzalez Degetau AM, Bata K, Emonts J, Schmitz A and Schmitt RH (2026) Bayesian experimental design in production engineering: a comprehensive performance and robustness study.
Front. Manuf. Technol. 5:1614335.
doi: 10.3389/fmtec.2025.1614335

Bayesian experimental design in production engineering: a comprehensive performance and robustness study

Lars Leyendecker[1]*, Ana Maria Gonzalez Degetau[1], Katharina Bata[2], Jessica Emonts[3], Angela Schmitz[4] and Robert H. Schmitt[1,5]

[1]Production Quality, Fraunhofer Institute for Production Technology IPT, Aachen, Germany, [2]Karlsruhe Institute of Technology KIT, Scientific Computing Center (SCC), Karlsruhe, Germany, [3]Department of Mechanical Engineering, University of Applied Sciences Aachen, Aachen, Germany, [4]Institute of Product Development and Engineering Design, Faculty of Process Engineering, Energy and Mechanical Systems, TH Köln—University of Applied Sciences, Köln, Germany, [5]Laboratory for Machine Tools and Production Engineering (WZL) of RWTH Aachen University, Aachen, Germany

In production engineering, the identification of optimal process parameters is essential to advance product quality and overall equipment effectiveness. Optimizing and adapting process parameters through experimental design is relevant for different phases of the life cycle of a production process: (i) design and development of new processes, (ii) failure analysis and optimization, and (iii) adaptation and calibration in series production. Existing experimental design approaches tend to be inefficient because they comprise static, non-adaptive methodologies that separate experiment design from execution and analysis. Instead, Bayesian Optimization (BO) offers an adaptive and data-efficient methodology for experimental design termed Bayesian experimental design (BED). In BED, the selection of an experiment is re-evaluated in each iteration based on previous experiment results according to an acquisition function that aims to maximize the informational content of each experiment. However, the configuration of BO algorithms for specific optimization problems requires extensive knowledge of both BO and process characteristics. The mean and covariance functions of the surrogate model, the acquisition function, and initial data sampling must be individually configured and significantly influence overall optimization performance, preventing widespread adoption in production engineering practice. To guide the configuration of BO algorithms for optimizing production processes, in this paper, we perform an extensive benchmark study with a total of 15,360 experiments. We evaluate the performance of a variety of BO algorithm configurations (including kernels, acquisition functions, and initial sampling sizes) on a total of eight optimization problems with a noiseless and a noisy variant each. The performance and robustness analysis reveals significant performance differences between individual BO algorithm configurations. The results of our benchmarking serve as empirical references based on which we derive actionable guidelines for the application of BED in production engineering.

KEYWORDS

Bayesian optimization, Bayesian experimental design, process optimization, production, manufacturing

1 Introduction

In production engineering, the technical, ecological, and economical performance of production processes depend on the parameter settings that configure the behavior of the process. To investigate the relationship (response surface) between process inputs and outputs, and therefore to find parameters that are optimal with respect to an arbitrary objective function, experimental parameter studies are performed. The goal of experimental design is to identify the set of process parameters (also called factors) that are most relevant to the performance of the process and to determine performance-optimal factor levels Freiesleben et al. (2020); Rainforth et al. (2023). Production processes are typically considered black-box systems, involve highly complex, high-dimensional design and objective spaces, and physical experimentation is time-, cost-, and resource intensive. According to the so-called polylemma of production, process optimization relies on human intuition, trial-and-error, and slow optimization cycles Schmitt and Pfeifer (2015). Traditional statistical experimental design methodologies and metaheuristics comprise full and fractional factorial Design of Experiments (DoE) Montgomery (2020); Durakovic (2017), one-factor-at-a-time (OFAT), Taguchi Method Logothetis and Wynn (1989), Response Surface Modeling Sarabia and Ortiz (2009), Latin Hypercube Sampling Tang (1993), or optimal designs Smucker et al. (2018).

Alternatively, Bayesian Optimization (BO) provides a model-based framework for adaptive experimental design using information-theoretic principles Rainforth et al. (2023). More specifically, BO is a sequential decision-making strategy for the optimization of arbitrary objective functions. In particular, BO is especially suited for optimizing expensive-to-evaluate black-box functions that i) do not have a closed-form representation, ii) do not provide function derivatives, and iii) only allow for point-wise evaluation Garnett (2023). BO consists of two core components: a surrogate model used for modeling the to-be-optimized objective function and an acquisition function that is sampled for guiding the selection of to-be-evaluated parameter sets. During optimization, the surrogate model is being continuously updated from a prior to a posterior belief by applying the Bayes theorem after new observations have been collected. The acquisition function utilizes the uncertainty quantification of the surrogate model to maximize the information gain of each experiment while balancing the exploration-exploitation trade-off. The process optimization using BO is performed until a pre-defined termination criterion (e.g., maximum number of experiments, pre-defined quality-level) is fulfilled. This concept of BO stems from early 1970s–1980s Močkus (1975); Mockus (1989) and has suffered till the recent past from computational bottlenecks hindering wide-spread application Rainforth et al. (2023). However, given the advancements of the recent years and the success of BO for hyperparameter optimization and neural architecture search, BO has regained popularity and rapid progress over the past 10 years Garnett (2023); Rainforth et al. (2023). The utilization of BO algorithms for sequential experimental design in scientific and engineering experimentation is termed Bayesian experimental design (BED). To this end, BED has been applied in material science Dieb and Tsuda (2018), manufacturing Maurya (2016), additive manufacturing Deneault et al. (2021); Guidetti et al. (2022), laser processing Duris et al. (2020), fluid dynamics Diessner et al. (2022), biotechnology Leyendecker et al. (2025); Liang and Lai (2021), plasma coating Guidetti et al. (2022) and information technology Haghanifar et al. (2020). The challenges in applying BO in manufacturing technology are, in particular, the high costs of experimentation and machine downtime, mixed variable types, collaboration with and acceptance by process experts, measurability of quality characteristics and measurement noise, and safe exploration. Additionally, a key challenge in successfully utilizing BO in production engineering is to find an optimal configuration of the BED algorithm comprising the configuration of the surrogate model and its mean and kernel functions, the acquisition function, and the initial design (number of data points to initialize the optimization). The BED configuration must be chosen depending on the characteristics of the optimization problem, i.e., the production process to be optimized, and precisely tuned to achieve optimal results.

1.1 Literature review

Previous studies have explored the impact of BED configuration, typically focusing on two components, most often the surrogate model and the acquisition function, in combination. In the field of materials science, Liang et al. Liang and Lai (2021) conducted a benchmark study evaluating different Gaussian Process-based surrogates alongside three acquisition functions, highlighting the critical role of proper initialization and exploration strategies. Diessner et al. Diessner et al. (2022) applied BED in the context of computational fluid dynamics, performing a benchmark that examined the effects of acquisition functions and initial sampling sizes. Similarly, Le Riche and Picheny (2021) investigated various surrogate models and initial design sizes on a standardized test set Finck et al. (2010). While these contributions offer valuable insights into individual components of BED configuration, none of them addresses the combined interdependencies of acquisition function, surrogate model, and initial design. This study aims to close that gap by systematically examining the interactions among these three key components in the context of production processes.

1.2 Approach and contribution of this paper

To investigate the optimization performance of different BO-configurations, we design and select a total of eight engineering-focused synthetic test functions (optimization problems) with different characteristics, complexities, and known optimal solutions. We perform a benchmark study by applying different BO-configurations to all optimization problems and compare the individual performances. As proposed by Bossek et al. Bossek et al. (2020a), we utilize the Dominated Hypervolume (HV) performance metric to consider both the success rate and efficiency of optimization. It is important to note that this study focuses on production processes, which influences certain methodological choices. In particular, since the number of adjustable parameters in such processes typically does not exceed six Ilzarbe et al. (2008); Arboretti et al. (2022), high-dimensional optimization problems are

not considered. The general goal of this investigation is to derive practical guidelines for configuring BED algorithms tailored to the optimization of production processes. The key contributions of this paper are:

1. We highlight the importance and challenges of experimental design in production engineering and propose BED as a promising data-driven methodology
2. We provide a methodology for benchmarking BO algorithms
3. By applying this methodology, we perform an extensive benchmark study across eight physically motivated test functions comprising a total of 15,360 experiments
4. The benchmarking results provide empirical references and we derive actionable guidelines for the configuration of BO and the application of BED in manufacturing process optimization
5. We outline the remaining challenges and derive further research needs to promote the adoption of BED in production engineering

The paper is structured as follows: After a brief introduction to process optimization in production and experimental design in Section 1, Materials and Methods (Section 2) outlines the fundamentals of BED and presents details of the benchmark study. In Section 3, we present the results of our study, which we further interpret and discuss in Section 4. The paper closes with a final conclusion and outlook in Section 5. For supplementary material, please refer to the Supplementary Appendix 1.

2 Methodology

This section provides the theoretical concepts of BED by first describing the fundamentals of experimental design in the engineering domain (see Section 2.1). The fundamentals of BO are given in Section 2.2 and BED is outlined in Section 2.3. In Section 2.4, we explain the scope, research aspects, and methodological approach of our benchmark study.

2.1 Fundamentals of experimental design

In production engineering, experimental design is the process of identifying key influential parameters (called factors) and determining the interaction of these factors on the output of the process and modeling the corresponding response surface. Accordingly, engineers and process operators utilize experimental design methodologies to identify the most influential process parameters and subsequently determine the optimal factor values using statistical analysis. Inherent in every optimization is the exploration-exploitation dilemma Berger-Tal et al. (2014). Accordingly, a decision must be made for each candidate selection as to whether to explore the design space or search in the vicinity of the already known best solutions. Therefore, in engineering practice, a distinction can be made between four levels of precision during execution. First, screening aims to rapidly localize the important factors in the initial design space. Second, in characterization, a narrowed search is conducted to identify the most influential factors. Third, optimization aims to determine the optimal factor levels. Finally, validation serves to ensure that the process is capable of consistently producing products that meet the predetermined quality specifications. Whereas good experimental designs ensure the validity of the optimal factor values found, excellent designs retain a high ratio between the extracted information and the invested resources Jankovic et al. (2021). Figure 1 schematically visualizes the framework for process optimization under the assumption of unknown system behavior. The behavior of the system can be described as a function $\mathbf{y}, \mathbf{o} = f(\mathbf{x}; \mathbf{z}) + \varepsilon$ that transforms the input vector $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathcal{X} \subseteq \mathbb{R}^n$ into an m-dimensional target vector $\mathbf{y} = [y_1, \ldots, y_m]^T \in \mathcal{Y} \subseteq \mathbb{R}^m$ under the potential influence of uncontrollable parameters and disturbance values $\mathbf{z} = [z_1, \ldots, z_r]^T \in \mathbb{R}^r$ and a noise term ε. In addition to the actual target variable vector $\mathbf{y}$, the process model can provide additional data in the form of process monitoring observations $\mathbf{o} = [o_1, \ldots, o_s]^T \in \mathbb{R}^s$. We assume that we do not have analytical knowledge about the process and that the system does not possess a closed-form representation, does not provide functional derivatives, and only allows for point-wise evaluation Garnett (2023). Furthermore, the optimization problem of finding $\mathbf{x}^\star = \text{argmax}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ can include multiple controllable and uncontrollable influencing factors and – depending on the number of target variables – can be either single or multi-objective. Besides the complexity of both design and objective spaces, the complexity and difficulty of the optimization problem is inherently determined by the complexity of the underlying system behavior. Depending on whether experiment design and experimentation along with experiment validation are performed in an iterative manner, a distinction can be made between sequential and non-sequential experimental design approaches.

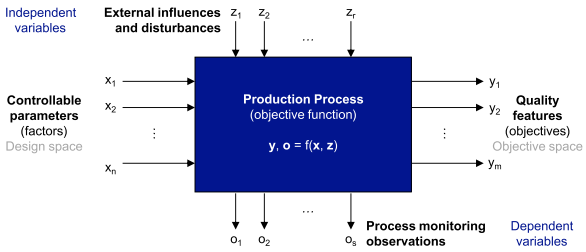

FIGURE 1
Schematic black-box production process model for experimental design.

2.2 Fundamentals of Bayesian Optimization (BO)

Optimization is an innate human behavior Garnett (2023) and optimization problems are pervasive in scientific and industrial fields that require optimization algorithms to be as efficient as possible Wang et al. (2022). In contrast to well-known metaheuristics that require large numbers of experiments and function evaluations, BO – with its model-based, adaptive, and active optimization policies – promises to be much more data-efficient in finding a global optimum (minimum or maximum) of an

unknown objective function Liang and Lai (2021). In general, the model structure of a BO algorithm comprises two core components: 1) a surrogate model (see Section 2.2.1) and 2) an acquisition function (see Section 2.2.2). The surrogate model aims to faithfully approximate the input-output behavior of the system to be optimized. The acquisition function indirectly defines the optimization policy by assessing the value of future observations and therefore guiding the parameter selection process Garnett (2023). For starting the optimization, BO requires an initial dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ that is the collection of observations of the input-output behavior of the system. In BO, the Bayes theorem is applied to incorporate a prior belief to maximize the informational content and therefore the value of each new experiment Duris et al. (2020). BO utilizes the Bayes theorem to iteratively update its prior distribution (prior) after the dataset $\mathcal{D}$ has been extended with new observations. The prior is updated to form the posterior distribution (posterior). The prior represents a belief about the behavior of the objective function f. The posterior distribution is used to compute and optimize the acquisition function in order to sample parameter combinations with high informational content for conducting new experiments.

2.2.1 Surrogate model

BO requires a probabilistic surrogate model that provides estimates and uncertainties of the objective function f Duris et al. (2020). In this work, for surrogate models, we solely consider non-parametric Gaussian processes (*GP*), the most widely adopted surrogate model. We define the GP surrogate model as $f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. To perform both prediction and uncertainty quantification, a GP utilizes a mean function $m(\mathbf{x})$ to specify the expected value of f and a covariance function $k(\mathbf{x}, \mathbf{x}')$ Duris et al. (2020); Greenhill et al. (2020). The covariance function determines the covariance between the function values $\mathbf{y}$ and $\mathbf{y}'$ corresponding to a pair of input parameters $\mathbf{x}$ and $\mathbf{x}'$ Garnett (2023). In comparison with the mean function, careful design of the covariance function is of higher criticality for the fidelity of the model and the experimentation's sample path behavior Garnett (2023).

2.2.2 Acquisition function

For the optimization of an expensive-to-evaluate function with black-box behavior, BO defines an optimization policy by introducing a substitute optimization problem utilizing a so-called acquisition function. In contrast to regular objective functions, the acquisition function is differentiable, inexpensive to evaluate and is derived from $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ Greenhill et al. (2020). Therefore, well-established numerical optimization algorithms can be utilized to iteratively optimize the acquisition function in order to propose parameter sets to be evaluated next. The acquisition function $\alpha(\mathbf{x}, \mathcal{D})$: $\mathcal{X} \rightarrow \mathbb{R}$ assigns a score to each parameter combination within the design space reflecting the value of each experiment for solving the optimization problem Garnett (2023). The acquisition function performs the trade-off between exploration and exploitation and therefore strongly influences the sample path behavior and optimization efficiency. Besides knowledge gradient, entropy search, and predictive entropy search, the most widely adopted single-objective acquisition function is expected improvement Frazier (2018).

2.3 Bayesian experimental design (BED)

Figure 2 describes the iterative operating principle of BED: Starting with the design space $\mathcal{X}$ consisting of factors and associated factor limits, which are typically predefined by process experts, as well as initial data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, BED iteratively proposes new experiments (Step 1). In each iteration, experiments are conducted using the proposed factor value vector $\mathbf{x}$ (Step 2). Formally, this step can be viewed as a query of the true objective function $f(\mathbf{x}, \mathbf{z})$ to measure the response vector. As a result of the evaluation of the experiment, the objective vector $\mathbf{y}$ is obtained. An iteration can comprise a single experiment or an arbitrary number of experiments. In these cases, one refers to single-point or batch experimentation. The results of the experiment $(\mathbf{x}, \mathbf{y})$ are added to the dataset $\mathcal{D}$ (Step 3). In this way, BED receives feedback on the result of the experiment to update the Gaussian surrogate model (Step 4). Based on the quality of the solution, the fulfillment of the termination condition is checked (Step 5). The acceptance criterion can be arbitrarily defined and, for example, take into account quality feature requirements, a maximum number of experiments, or time, cost, or resource limitations. If the acceptance criterion is satisfied, the optimization is terminated. Otherwise, optimization continues as long as the termination condition is not met. By sampling a new set of factor values through optimization of the acquisition function, the next iteration is entered.

2.4 Composition of the benchmark study

In this section we describe the scope and characteristics of our benchmark study to investigate the suitability of different BED algorithms on different types of optimization problems. The scope of this study is limited to single-objective functions, single-point evaluations, numerical, continuous and unconstrained problems and controllable parameters. The study design comprises three components: First, the BED algorithm configuration options (configuration space) (see Section 2.4.1), second, the optimization problems (performance space) (see Section 2.4.2), and third, the performance metrics (performance space) (see Section 2.4.3).

2.4.1 Definition of the configuration space

The components of BED algorithms investigated in this study comprise the kernel (also known as covariance function), the acquisition function, and the initial sampling design. Building upon the findings from the studies of Palar and Shimoyama (2019); Picheny et al. (2013); Le Riche and Picheny (2021); Liang and Lai (2021), this work considers four kernels: RBF, Matern05, Matern15, Matern25, as they belong to the standard portfolio of BO and can be applied to different processes. Each of these isotropic kernels k is also investigated with their anisotropic counterpart utilizing automatic relevance detection (ARD) Duvenaud (2014). ARD implicitly determines the relevance of the input parameters and aims to enhance the modeling accuracy and optimization efficiency.

Regarding the acquisition function α, expected improvement (EI), probability of improvement (PI), and upper confidence bound

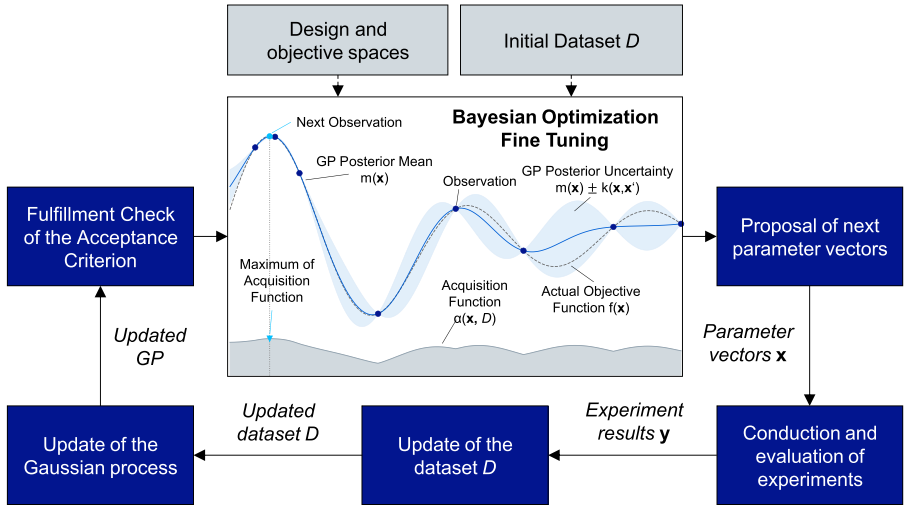

FIGURE 2
Bayesian experimental design procedure and Bayesian optimization components (based on Firas Al-Hafez (2021)).

(UCB) (with exploration scale of $\beta = 0.2$) are examined, along with a variation of EI known as Noisy Expected Improvement or NEI, which is specifically designed to handle noisy problems more effectively. For a detailed study of the presented acquisition functions, please refer to Garnett (2023).

To define the initial design, three choices must be made: the initial sampling size, the initial sampling strategy, and the number of independent runs. The initial sampling size u is a key component for determining the exploration phase of BED before fitting the GP model. Following the methodology outlined in Le Riche and Picheny (2021), this study considers three initial sampling sizes: Small (S, five trials), Medium (M, 10 trials), and Large (L, 30 trials), while maintaining a fixed total budget of 35 trials. The decision to set the initial sampling size independently of the dimensionality of the problem aligns with the study's focus on experimental design for production processes, which typically involve no more than six parameters and a maximum of 30 trials Ilzarbe et al. (2008). This restriction confines the study to low-dimensional spaces, whereas for high-dimensional problems, adjustments to the initial sampling size would be necessary.

Within various sampling strategies, pseudo-random Sobol sampling is chosen due to its ability to effectively cover the parameter space under the specific conditions encountered. Due to the stochastic nature of the Sobol algorithm, it is decisive to perform multiple independent runs for each BO configuration on each test function. Following the recommendations of Mersmann et al. (2010), a total of ten independent runs are conducted, using ten different random seeds. Each random seed is applied to each BO configuration, ensuring that all configurations start with the same initial data and that no configuration benefits from random fluctuations. A detailed description of the Sobol algorithm can be found in Section 1.1 of the Supplementary Material.

2.4.2 Definition of the problem space

To examine the performance of different BO algorithm configurations on different optimization problems, we create artificial datasets utilizing a total of eight analytic test functions (see Figure 4). Four of these eight test functions, namely, *F1, F2, F3, F4*, are mathematical problems. The remaining four (*AdaptedBranin, Borehole, OTLCircuit, WingWeight*) originate from Forrester et al. (2008); Surjanovic and Bingham (2021) and comprise physically motivated optimization problems. Please refer to the Supplementary Material Section 1.2 for a detailed description of the optimization problems.

Consistent with previous benchmark studies Qin et al. (2021); Picheny et al. (2013); Palar and Shimoyama (2019); Gan et al. (2021), this research investigates both noiseless and noisy versions of objective functions. To introduce random noise to the output of the noisy functions, a noise level of 0.1 is employed, following the approach outlined in Qin et al. (2021) and Gan et al. (2021). In each trial, a pseudorandom number ranging between 0 and 0.1 with a uniform distribution is generated. This number is then multiplied with the standard deviation (SD) of the objective of the respective function and added to the output value. The SD of the function's objective is calculated on the basis of a random uniform sampling of size 10,000. This methodology allows for a controlled examination and comparison of the effects of noise. It is essential to emphasize that the noise in this work is homoscedastic, meaning it does not depend on the sequential course of the experiment. Heteroscedastic noise is not considered in this study.

2.4.3 Definition of the performance space

In this study, we focus on three key metrics to evaluate the performance of individual BED configurations on the optimization problems: solution quality, robustness, and efficiency. We employ the multi-objective dominated hypervolume (HV) metric, as proposed by Bossek et al. (2020b). According to Equation 1, the HV metric integrates robustness (measured by the probability of failure p_f) and efficiency (measured by the running time of successful experiments r_s). Lower values of p_f and r_s yield higher HV values, indicating superior overall performance. It

should be noted that the HV metric was chosen because of its efficiency-robustness trade-off, since the number of required experiments constitutes the key cost driver in the optimization of production engineering systems. The HV metric is illustrated in Figure 3 (left) and calculated according to Algorithm 1.

$$\mathrm{HV} = (1 - r_s)\left(1 - p_f\right) \tag{1}$$

Let I be a test function
Let $\theta(k, \alpha, u)$ be a BO configuration
Let $i \in \{1, \ldots, w\}$ be a single trial
Let u be the number of Sobol trials, $U = \{1, \ldots, u\}$
Let v be the number of Bayesian trials, $V = \{u + 1, \ldots, v\}$
Let $w = u + v$ be the total budget, $W = U \cup V$
Let $j \in \{1, \ldots, m\}$ be a single run of total runs m
for each combination θ, I **do**
  for each trial i **do**
    Calculate relative deviation $(\Delta y_i)_j$ according to (2)
  end for
  for each run j **do**
    if the algorithm finds an optimal solution within $\pm\, \Delta y^*$ according to (3) **then**
      Set $\mathrm{success}_j = 1$
    else
      Set $\mathrm{success}_j = 0$
    end if
  end for
  Define successful runs set S according to (4)
  Calculate probability of failure $p_f^{\theta,I}$ according to (5)
  Calculate running time $r_s^{\theta,I}$ according to (6)
  Calculate dominated Hypervolume $\mathrm{HV}^{\theta,I}$ according to (7)
end for

Algorithm 1. Calculate Dominated Hypervolume (HV).

- The relative deviation is calculated as the normalized difference between the known optimal solution y^* and the observed solution y_i at each trial i for each run j (Equation 2). This normalization is performed to ensure comparability between different test functions.

$$\left(\Delta y_i\right)_j = \left|\frac{y^* - y_i}{y^*}\right| \tag{2}$$

- A run j is considered successful if the algorithm has found an optimal solution within the tolerance range of $\pm\, \Delta y^*$ relative to the known optimal solution for each test function (Equation 3). In this work a value of $\Delta y^* = 0.05$ is utilized for all test functions, corresponding to an optimization of 95%.

$$\mathrm{success}_j = \begin{cases} 1 & \text{if} \quad \exists \Delta y_i < \Delta y^* \quad | \quad i \in V \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

- Set of successful runs (Equation 4):

$$S = \left\{ j \in \{1, \ldots, m\} \quad \text{with} \quad \mathrm{success}_j = 1 \right\} \tag{4}$$

- The probability of failure $p_f^{\theta,I}$ for one BO configuration θ and a test function I over all m runs can be defined as Equation 5:

$$p_f^{\theta,I} = 1 - \frac{1}{m} \sum_{j=1}^{m} \mathrm{success}_j \tag{5}$$

- The running time $r_s^{\theta,I}$ for one BO configuration θ and a test function I is the first successful trial i within the Bayesian trials where the set tolerance was achieved (Equation 6). It is aggregated through all successful runs S and normalized to the total budget w. This last step sets the reference time T defined in the HV equation of Bossek et al. (2020b) to 1.

$$r_s^{\theta,I} = \frac{1}{w} \cdot \frac{1}{|S|} \sum_{j \in S} \arg\min_{i \in V} \left(\Delta y_i\right)_j \tag{6}$$

- Finally, the dominated Hypervolume $\mathrm{HV}^{\theta,I}$ of a BO configuration θ on a test function I can be calculated as follows (Equation 7).

$$\mathrm{HV}^{\theta,I} = \left(1 - r_s^{\theta,I}\right)\left(1 - p_f^{\theta,I}\right) \tag{7}$$

Higher values of HV indicate better performance of the BO configuration, while lower values suggest inefficiency, lack of robustness, or a combination of both. The minimal HV-value is 0, with $p_f = 1.0$ and $r_s = 1.0$, while the maximum HV-value is given at $p_f = 0$ and a minimal running time of $r_s = 0.14$, resulting in a maximum of HV= 0.86. The minimal running time of $r_s = 0.14$ is the result of dividing the minimum number of initial samples by the fixed budget $(5/35 \approx 0.142)$.

In order to compare different BO configurations and assess their suitability on the various test functions, we utilize a classification approach based on the resulting HV and taking into account the interrelationships between kernels, acquisition functions and initial dataset size. Each configuration (i.e., each kernel, acquisition function, and initial dataset size) is classified as qualified if it results in a good performance for optimizing the test function, non-qualified if it leads to poor performance, or undetermined if there is no clear outcome regarding its performance. This classification approach is depicted in Algorithm 2.

Since each individual configuration appears in multiple combinations with other configuration parameters, simple aggregation techniques, such as computing the median HV across all runs, may distort the actual performance. For example, if the RBF kernel performs well when paired with EI, NEI, and UCB, but poorly with PI, its overall median HV may be skewed downward, thus underrepresenting its true capability. While a variance analysis could reveal the degree to which a configuration influences outcomes, it does not provide insight into the quality of the performance itself, which is essential for this study. Therefore, a more nuanced classification method is applied that considers both performance level and variability across combinations.

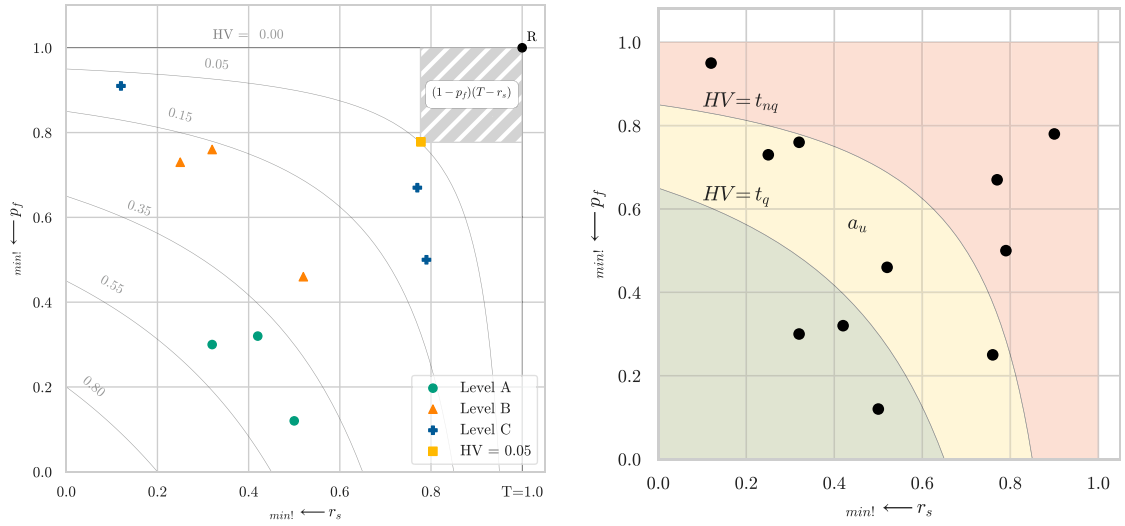

FIGURE 3
Dominated Hypervolume (HV) adapted from Bossek et al. (2020b) (left) and HV span plot with classification areas (green: qualified, yellow: undetermined, red: non-qualified) (right) (p_f: probability of failure, r_s: running time of successful runs, T: optimization budget, t_q: non-qualified threshold, t_{nq}: qualified threshold, a_u: width of undetermined classification area).

```
for each test function i in the problem space I do
  Initialize lists: Q ← []//Qualified configurations
  Initialize lists: NQ ← []//Non-qualified configurations
  Initialize lists: UD ← []//Undetermined configurations
  Let c be a component of a BO configuration (c being a kernel, acquisition function or initial sampling size)
  for all trials where the configuration contains c in θ do
    Compute median of HV^{θ,I} for the trials
    if median of HV^{θ,I} is in non-qualified area then
      Add c to NQ
    end if
  end for
  for all configurations where c in θ and not in NQ do
    Compute median of HV^{θ,I}
    if median of HV^{θ,I} is in qualified area then
      Add c to Q
    else
      Add c to UD
    end if
  end for
end for
```

Algorithm 2. Classify configurations.

To classify the BO configurations effectively, the HV values ranging from 0 to 0.86 are divided into three distinct areas: non-qualified NQ (red), qualified Q (green), and undetermined UD (yellow) (Figure 3 (right)). Each black point represents an experiment with a specific BO configuration $\theta = (K, \alpha, u)$. The limits of the areas, represented by the thresholds t_{nq} and t_q, are determined based on the distribution of HV values obtained from the experiments for each test function. The width of the undetermined area, denoted by a_u, also varies according to the characteristics of the test function. To determine the non-qualified threshold t_{nq} and the qualified threshold t_q, a comparative and adaptive approach is followed based on the obtained benchmark results in noiseless and noisy cases, separately. For t_{nq}, the configuration variable with the lowest median HV value is identified, and the threshold is placed just above the 50th percentile of the median of the single configurations. This approach ensures that configurations with poor performance are eliminated from consideration. For t_q, the variables with the best distribution of HV are taken into account, and the threshold is set below the 25th percentile of the best configurations. By adopting this approach, configurations that show superior performance are identified.

It is important to note that the thresholds can be adjusted according to specific requirements of the user, such as demanding higher efficiency or robustness. However, in this study, the comparative approach is chosen to provide practical and general recommendations for the selected BO configurations. At the conclusion of the evaluation, each individual configuration is classified as either qualified, non-qualified, or undetermined for each test function. This analysis allows for a more informed and nuanced assessment of the performance of each configuration in optimizing the test functions. By classifying the configurations in this manner, practical recommendations can be made regarding the suitability of different BO configurations for specific test functions in terms of robustness and efficiency.

2.4.4 Summary of the benchmark study

As a summary, Figure 4 provides a final overview of the key characteristics of our study. It encompasses the problem space characteristics, BO algorithm configurations and performance

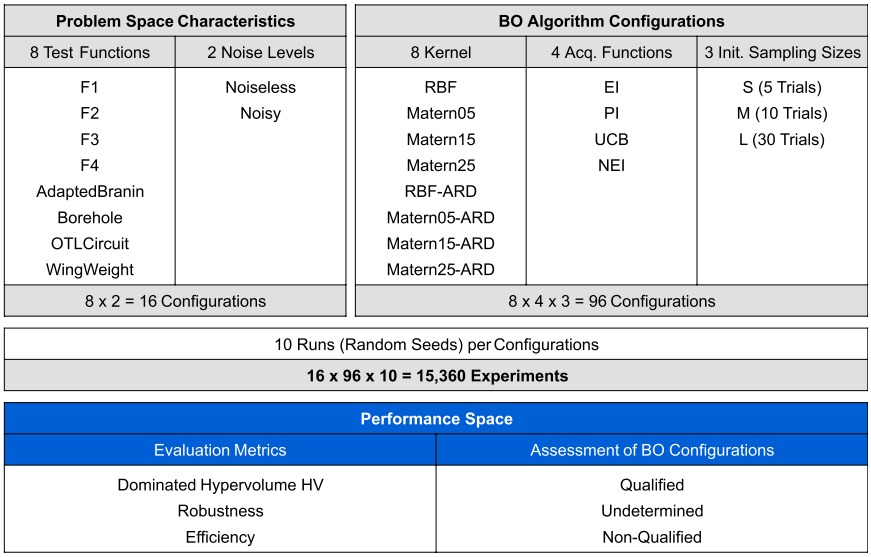

FIGURE 4
Overview of the benchmark study.

metrics. The evaluation process involves 16 optimization problems, consisting of eight functions with two noise levels. In the BO configuration space, a total of 96 configurations are tested, which is obtained by combining eight kernels, four acquisition functions, and three initial sampling sizes in a full-factorial manner. This results in a thorough evaluation of 1,536 configurations. In order to achieve statistical significance, ten experiments with different random seeds are carried out and evaluated for each configuration. In total, the study comprises 15,360 experiments. The performance metrics include both the metrics for each individual experiment and the classification metrics used to compare the BO configurations with each other. This comprehensive approach allows for a thorough investigation of BO algorithms and provides valuable insights for making informed decisions when selecting suitable configurations for different optimization tasks.

The study is conducted in the following way: First, the BO configuration $\theta = (K, \alpha, u)$ is defined, and a random seed is set to ensure stochastic robustness. Using the Sobol sampling strategy, the test function is evaluated with an initial sampling size of $u = S, M, L$ to collect the initial dataset. Subsequently, the GP model with the kernel k is fitted. The acquisition function α is called to select candidates for the next evaluation of the test function, yielding the corresponding objective value. We define a fixed budget of trials for each experiment run across all problems, enabling a comparison of the problems and their complexity. The choice of the fixed budget is based on the study of Ilzarbe et al. (2008) who investigated the use of DOE in different engineering applications. Of the 77 reviewed articles, 77% ran a number of trials less or equal to 30 per experiment. Based on this review, a fix budget of 35 trials per experiment is set for all configurations and test functions. If the number of trials has not exceeded the total budget of 35 trials, an additional trial is performed. This process is repeated until a total of 10 runs with different random seeds have been executed.

3 Results

In this section, the results of the benchmark study (Section 2.4) are outlined according to the following structure: In Section 3.1, preliminary results provide an overview of the analysis and narrow subsequent examination. Subsequently, in Section 3.2, emphasis is placed on evaluating the responsiveness of the test functions. This analysis offers insight into the overall optimization level achievable for each test function, irrespective of specific BO configurations. The overarching goal is to uncover the importance of selecting appropriate BO configurations for specific test functions, while underscoring variations in their optimization capabilities. In Section 3.3, a detailed examination of individual BO configurations is conducted with a focus on specific test functions. This analysis leads to the qualification of individual kernels, acquisition functions, and initial sampling sizes.

3.1 Preliminaries

In this Section, an initial overview of the performance of the BO configurations is provided. The aim of these preliminary observations is to gain a first impression of the results and identify any emerging trends in the data, regardless of the specific test function being examined. This analysis helps to narrow down the focus of the study and identify areas of interest for further investigation.

Figure 5 displays the resulting HV of all experiments conducted on each test function, considering both noisy and noiseless scenarios. The HV values are plotted based on two key components: the probability of failure p_f and the running time of successful experiments r_s, as depicted in Figure 3. Each test function has the same number of points (8 kernels x 4 acquisition functions x 3 sizes x 2 noise levels = 192), and

in some functions, they overlap. It is important to note that the points in all three subfigures represent the same data, but are differentiated by color labels based on the specific configuration. The upper subfigure provides insights into the performance of different kernels, the middle subfigure examines the impact of various acquisition functions, and the lower subfigure explores the influence of the initial sampling size. This evaluation yields three main findings:

- The test functions F1, F2, and OTLCircuit consistently exhibit higher HV values (lower values of p_f and r_s) in all combinations of BO configurations. This suggests that these functions are comparatively simpler compared to the other test functions.
- Upon closer examination of the initial sampling sizes, it is evident that all experiments conducted with a large initial dataset exhibit higher running times and consequently lower values of HV. Across all test functions, no experiment with this configuration surpasses an HV value of 0.15, regardless of the choice of acquisition function or kernel. This observation can be mathematically explained by the fact that none of the experiment points have values of r_s below 0.85. This is due to the minimum running time achievable by the (L)-experiments occurring at trial 30, which when divided by the total budget of 35 trials (as explained in Equation 6), results in a value of 0.85.
- No other obvious trend can be observed with respect to single kernels or acquisition functions across the different test functions. While there is a tendency of underperformance of Matern05 for F1 or F2, this trend is not consistently observed in all other test functions. The absence of obvious trends, apart from the ones mentioned earlier, highlights the need for further investigation and analysis to gain a deeper understanding of the performance of different configurations and their interrelationships in various test functions.

Based on these findings, the subsequent evaluation will only include small (S) and medium (M) initial sampling sizes, excluding large ones (L). It has been demonstrated that larger initial sampling sizes result in lower efficiency without offering significant advantages in robustness.

3.2 Analysis of the responsiveness of the test functions

In this Section, the results of the responsiveness of the test functions under optimization with the selected BO configurations are shown. The evaluation considers the overall performance metric HV, the robustness measured by the probability of failure p_f, and the efficiency measured by the running time r_s. Additionally, the analysis distinguishes between noiseless and noisy data, providing insights into the noise sensitivity of the optimization for each test function. Section 3.2.1 addresses the average optimization level that can be achieved for each test function, elucidating the similarities or differences in the optimization capability of the BO configurations. Section 3.2.2 zooms in on individual aspects of robustness and efficiency.

3.2.1 Optimization level

The overall optimization level for each test function is depicted in Figure 6. It presents boxplot distributions of the HV values for all experiments, categorized by the eight test functions and distinguishing between noiseless and noisy data. The position of the box, indicated by the median HV value, represents the level of optimization that can be achieved for each function. The length of the box, represented by the interquartile range (IQR), provides insights into the range of HV values covered by the BO configurations. Functions with lower IQRs indicate that a wide range of BO configurations can achieve results close to the median HV value, indicating simplicity of the function. On the other hand, functions with higher IQRs suggest that not all configurations are equally effective in optimizing them, indicating a greater variability in performance and the need for more specific configurations. A rough grouping can be made according to the comparison between HV median and IQR:

A. F1, F2, OTLCircuit show higher HV and smaller IQR values in both noiseless and noisy cases.
B. F4, WingWeight show medium HV and higher IQR values, similar in noiseless and noisy cases.
C. F3, AdaptedBranin, Borehole show lower HV and higher IQR values and differences between noiseless and noisy cases.

We observe the following results for each of these groups:

A. The test functions F1, F2, and OTLCircuit exhibit median HV values of approximately 0.7, with OTLCircuit having more outliers towards lower HV values, suggesting there are some configurations that are clearly less qualified than the others. The IQR of all three functions is less than 0.12, indicating a similar optimization potential with most of the BO configurations. Notably, for these three functions, the optimization results on noisy data appear to be similar to those on noiseless data, suggesting a low sensitivity to noise in these particular functions.
B. The test functions F4 and WingWeight show median HV values around 0.5. For F4, the IQR is 0.19 in noiseless cases and 0.11 in noisy cases. In the case of WingWeight, the IQR is 0.28 for both noiseless and noisy cases, with the noiseless case slightly skewed towards higher HV values. The long whiskers of both functions towards lower HV values indicate the lower performance of some BO configurations.
C. The functions F3, AdaptedBranin, and Borehole exhibit a more heterogeneous group in terms of optimization results. For F3, the median HV values are around 0.2, and the IQR is approximately 0.2 in both noiseless and noisy cases, making it the test function with the poorest overall performance. In contrast, AdaptedBranin and Borehole exhibit median HV values of 0.56 and 0.45, respectively, with larger IQRs of nearly 0.4 in noiseless cases. When considering noisy data, the median HV values decrease to 0.2 for both functions, and the IQRs reduce to approximately 0.14. Among all the functions, the difference in performance between noiseless and noisy cases is most pronounced for these two functions. These observations highlight the complexity and sensitivity of the optimization process for these particular test functions.

FIGURE 5
HV over all test functions for different kernel functions. **(a)** HV over all test functions for different kernel functions, acquisition functions. **(b)** HV over all test functions for different acquisition functions, and initial dataset sizes. **(c)** HV over all test functions for different initial dataset sizes.

In summary, the test functions exhibit varying levels of optimization difficulty across three groups. Group A functions are relatively easier to optimize, Group B functions pose moderate challenges with no significant noise-related differences, while Group C functions, especially noisy cases, present the highest complexity and varied optimization success. These findings emphasize the influence of function complexity and noise on BO performance. This underlines the relevance of a precise

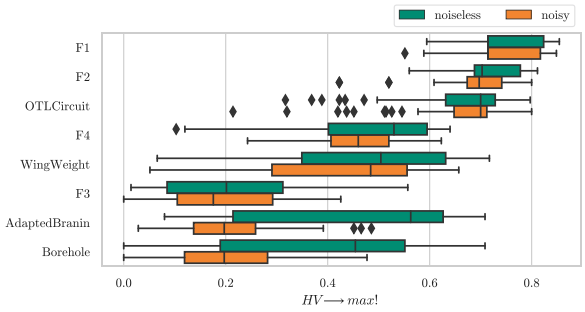

FIGURE 6
Distribution of HV values for test functions, given noiseless and noisy data. Higher values are better.

algorithm configuration when dealing with complex and noisy optimization problems.

3.2.2 Robustness and efficiency

To gain insights into the robustness and efficiency in all BO configurations, a closer look at the two components of HV is taken: the probability of failure p_f and the running time r_s. Figure 7 provides visual representations of the distribution of p_f and r_s for each test function over all BO configurations. Keeping the grouping defined above, following observations can be made:

A. F1 and F2 consistently achieve a 100% success rate ($p_f = 0.0$) across almost all configurations, indicating their high robustness in optimization, independent of the BO configuration used. Similarly, the OTLCircuit function demonstrates a 90% success rate with 75% of the configurations, while the outliers should be considered as non-robust configurations. In terms of efficiency, these three functions can be effectively optimized with 75% of the configurations, as they are able to reach the predefined tolerance level within 0.3 of the total budget. These observations hold true for both noiseless and noisy cases, highlighting the resilience of these functions to noise interference.
B. F4 exhibits an overall efficiency of around $r_s = 0.46$ across all configurations. While there are some configurations that are not qualified in terms of robustness, most of them perform well in this regard. On the other hand, the WingWeight function shows a narrow IQR for the running time, suggesting that the majority of configurations achieve a high efficiency around $r_s = 0.4$. However, there are some outliers indicating that certain configurations may struggle to reach optimal efficiency. In terms of the probability of failure, WingWeight displays a wider IQR, indicating a greater variation in robustness across different BO configurations. This highlights the importance of carefully selecting the appropriate BO configuration for this particular test function, as there can be significant differences in performance and robustness among the various configurations.
C. In terms of efficiency, both AdaptedBranin and Borehole exhibit similar behavior. The median values of the running time (r_s) approximate to 0.4 for noiseless cases and shift towards 0.6 for noisy cases. Furthermore, all configurations achieve these efficiency levels with a relatively narrow interquartile range (IQR) of less than 0.17. On the other hand, the F3 function appears to be less affected by noise in terms of efficiency, with median r_s values of around 0.6 for both noiseless and noisy cases.

When it comes to robustness, noiseless AdaptedBranin and Borehole show a probability of failure (p_f) of 0.1 with 50% of the BO configurations, indicating relatively high robustness. However, the remaining configurations exhibit lower levels of robustness, as reflected in the wider IQR values. In the noisy cases, AdaptedBranin and Borehole have p_f median values of 0.5, with IQR values of 0.2 and 0.4, respectively. For the F3 function, the probability of failure is consistently 0.5 in both noiseless and noisy cases, with wider IQR values of 0.6 and 0.4, respectively.

From these observations, it can be concluded that the selection of qualified BO configurations is decisive to achieve a high level of robustness for these test functions. Specifically, for AdaptedBranin and Borehole, certain configurations demonstrate good performance in terms of efficiency and robustness, while others may lead to suboptimal results. Therefore, careful consideration of the specific BO configuration is essential to ensure effective optimization for these test functions.

Table 1 provides the numerical values of the results discussed earlier, including the median values and IQRs (in parentheses) of HV, p_f, and r_s for each test function, keeping the established groups. These values offer a quantitative representation of the overall performance of the BO configurations on the selected test functions. It can be observed that F1 shows the best optimization results with highest HV values, while F3 exhibits the worst performance under all test functions. Additionally, a notable difference can be noticed between the noiseless and noisy cases for AdaptedBranin and Borehole, particularly in terms of robustness. The values for probability of failure in the noisy cases are substantially higher compared to the noiseless cases, indicating that the presence of noise has a significant impact on the performance of BO configurations on these functions. This highlights the need to carefully consider the influence of noise when optimizing these test functions using BO.

Following conclusions can be drawn out of this analysis:

A. F1, F2, and OTLCircuit can be considered relatively simple functions in both noiseless and noisy cases. Regardless of the specific BO configurations used (excluding outliers), these functions exhibit robustness and efficiency.
B. F4 and WingWeight represent the next level of complexity. There are no significant differences between noiseless and noisy cases. While most configurations achieve the desired efficiency, there are some configurations that lack the desired robustness.
C. F3, AdaptedBranin, and Borehole are the most complex functions in the problem space. They exhibit lower efficiency levels compared to other functions. Moreover, the behavior differs significantly between noiseless and noisy cases, and certain BO configurations demonstrate higher levels of robustness and efficiency. This underscores the importance of examining the configurations in greater detail.

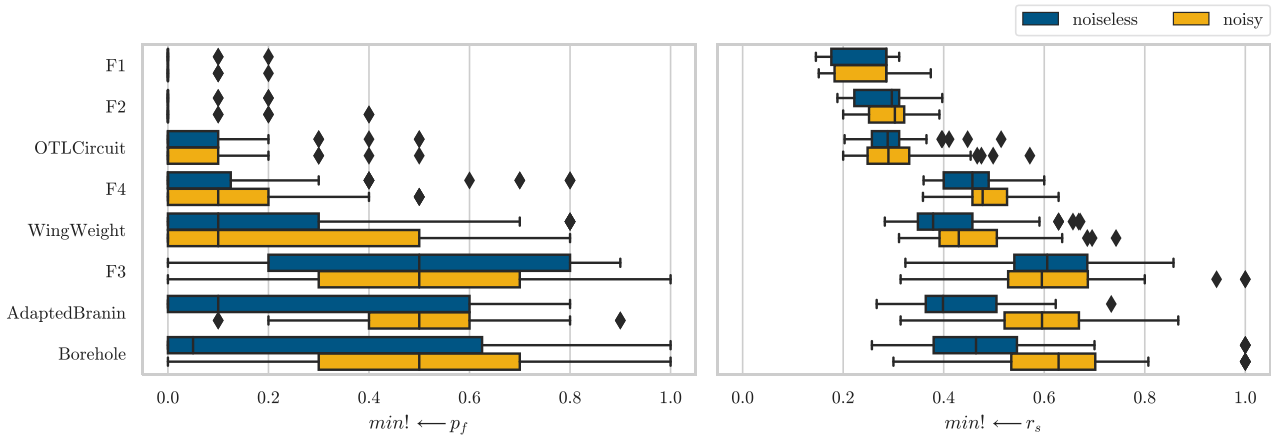

FIGURE 7
Distribution of probability of failure and running time for test functions, given noiseless and noisy data. Lower values are superior.

3.3 Analysis of optimization performance of BO configurations

After examining the general responsiveness of the test functions in Section 3.2, a more detailed analysis is conducted on the individual BO configurations. The objective is to determine the appropriateness of each configuration for optimizing specific test functions, following the systematic approach described in Section 2.4.3. Streamlining this process, the analysis seeks to identify whether certain single kernels, acquisition functions, or initial sampling sizes can be clearly classified as qualified or non-qualified, irrespective of their combination. The thresholds for each test function are shown in Figure 8. Configurations with HV values inside the red are classified as non-qualified. Configurations with HV values in the yellow are classified as undetermined. Configurations with HV values in the green area are classified as qualified for optimization. For each test function, the end of the green area marks the maximal optimization achieved by the best BO configuration on that function. The test functions are grouped according to Section 3.2, which provides insights into the complexity of optimizing each test function.

Group A (F1, F2, OTLCircuit) has a wide non-qualified area, indicating that optimizations under $HV = 0.7$ are not acceptable for this kind of functions and that most of the BO configurations achieve an optimization level of higher than $HV = 0.75$. The thresholds of second group B (F4, WingWeight) shift in the middle, with wider undetermined areas. This suggests that the best optimization to achieve with appropriate BO configurations lies on the qualified area with values of HV between 0.6 and 0.7, while choosing an inappropriate BO configuration results in HV values between 0 and 0.46. Group C (F3, AdaptedBranin, Borehole) presents the most complex scenario. With significant differences between non-qualified and qualified areas, the optimization of these functions is expected to be lower, with varying ranges between 0.55 and 0.7 depending on the function. This leads to the fact that choosing the right combination of BO configuration is essential for achieving the high optimization possible. To address these inquiries, a detailed analysis of F3 and Borehole in Sections 3.3.1 and 3.3.2 is conducted. Further results for the other test functions can be found in Supplementary Appendix 1.3. By analyzing the BO configurations in their different combinations and classifying them into the qualification areas, an assessment on the single kernels, acquisition functions and initial sampling sizes can be made. This is based on the statistical procedure described on Section 2.4.3, which reduces the non-qualified configurations when they present non-optimal results irrespective of their combination.

3.3.1 F3 - Analysis and classification

In this section, the results for F3 are presented. The classification ranges are as follows: non-qualified, $[0.0, t_{nq}]$ undetermined, $]t_{nq}, t_q[$ and qualified, $[t_q, 0.86]$, with the thresholds $t_{nq} = 0.15$ and $t_q = 0.31$. The configurations falling within each area are identified and analyzed to provide insights into their performance on the F3 test function. In Figure 9a, the results of the classification approach for the noiseless F3 function are depicted. Among the kernels, RBF, Matern05, and Matern15 are classified as non-qualified, as are UCB and PI among the acquisition functions, and medium (M) initial sampling size. After excluding these configurations, the following configurations can be classified as qualified: small (S) initial sampling size, EI and NEI for the acquisition functions, and all kernels with ARD. The classification for the Matern25 kernel remains undetermined.

In Figure 9b, the results of the classification approach for the noisy F3 function are presented. In this case, all isotropic kernels are excluded, as well as UCB, PI, and the M initial sampling size. The overall optimization performance is lower than in noiseless cases. Among the remaining configurations, RBF-ARD, Matern15-ARD and Matern25-ARD among the kernels, NEI among the acquisition functions and small initial sampling size can be classified as qualified. The classification for the rest of the configurations remains undetermined. Indeed, it is interesting to observe that the same trend can be observed between noiseless and noisy cases for the F3 function. The noisy case shifts all values into lower optimization levels, indicating the impact of noise on the

TABLE 1 Overview of HV, p_f, and r_s median values for each test function and BO configurations (IQR values in parenthesis).

Group	Test function	HV ↑		p_f ↓		r_s ↓	
		Noiseless	Noisy	Noiseless	Noisy	Noiseless	Noisy
A	F1	0.71	0.71	0.00	0.00	0.29	0.29
	F1 (IQR)	(0.11)	(0.10)	(0.00)	(0.00)	(0.11)	(0.10)
	F2	0.70	0.70	0.00	0.00	0.30	0.30
	F2 (IQR)	(0.09)	(0.07)	(0.00)	(0.00)	(0.09)	(0.07)
	OTLCircuit	0.70	0.70	0.00	0.00	0.29	0.29
	OTLCircuit (IQR)	(0.10)	(0.06)	(0.10)	(0.10)	(0.05)	(0.08)
B	F4	0.53	0.46	0.00	0.10	0.46	0.48
	F4 (IQR)	(0.19)	(0.11)	(0.13)	(0.20)	(0.09)	(0.07)
	WingWeight	0.50	0.48	0.1	0.1	0.38	0.43
	WingWeight (IQR)	(0.28)	(0.27)	(0.30)	(0.50)	(0.11)	(0.11)
C	F3	0.20	0.18	0.50	0.50	0.61	0.60
	F3 (IQR)	(0.23)	(0.19)	(0.60)	(0.40)	(0.15)	(0.16)
	AdaptedBranin	0.56	0.20	0.10	0.50	0.40	0.60
	AdaptedBranin (IQR)	(0.41)	(0.12)	(0.60)	(0.20)	(0.14)	(0.15)
	Borehole	0.45	0.20	0.05	0.50	0.46	0.63
	Borehole (IQR)	(0.36)	(0.16)	(0.63)	(0.40)	(0.17)	(0.17)

performance of the BO configurations. However, despite this shift, the clear recommendations for qualified configurations remain consistent. For both noiseless and noisy cases, kernels with ARD, EI, or NEI and a small (S) initial sampling size are recommended for optimizing the F3 function. This highlights the robustness and effectiveness of these configurations, making them reliable choices for practical applications.

3.3.2 Borehole - Analysis and classification

In this section, the results for Borehole are presented. The qualification thresholds are set to $t_{nq} = 0.19$ and $t_q = 0.54$ with minimal and maximal HV values of zero and 0.86. The configurations falling within each area are identified and analyzed to provide insights into their performance on the Borehole test function. Figure 10a illustrates the results of the classification approach for the noiseless Borehole function. Based on the analysis, the Matern05 kernel and the PI acquisition function are classified as non-qualified. Among the initial sampling sizes, both small (S) and medium (M) sizes present similar results. After excluding the non-qualified configurations RBF-ARD, Matern15-ARD, and Matern25-ARD are classified as qualified. All other configurations fall into the undetermined area, indicating that their performance on the Borehole function is lower and requires a dedicated combination of the other configuration variables.

In Figure 10b, the results of the classification approach for the noisy Borehole function are presented. According to the analysis, the isotropic RBF and Matern05 kernels, as well as the PI acquisition function, are classified as non-qualified for the noisy Borehole function. All other configurations fall into the undetermined area, with no single configuration being qualified for optimization with higher values than HV = 0.56. However, it is worth noting that there is one outlier at HV = 0.48, indicating that the combination of Matern15-ARD, NEI, and small initial sampling size achieves a comparatively higher performance, although none of these configurations can be single classified as qualified. There is a clear difference in the reduction of performance between the noiseless and noisy cases. The noisy Borehole function shows a significant decrease in optimization performance compared to its noiseless counterpart. However, the differences between the single configurations in the noisy case are lower than in the noiseless one.

4 Discussion

In this section, the results for all test functions are discussed. We evaluate the performance of the BO configurations on all test functions and provide actionable guidelines. Table 2 presents the classification results of all single BO configurations on the noiseless and noisy test functions. While the classification of BO configurations is performed for each individual test function, some general observations can be made that apply across different test functions. These general observations provide valuable insights into the overall performance of certain configurations, enabling users to make informed decisions and tailor their BO algorithms more effectively to specific optimization tasks. In the following, the results are discussed individually for kernels, acquisition functions, and initial sampling sizes.

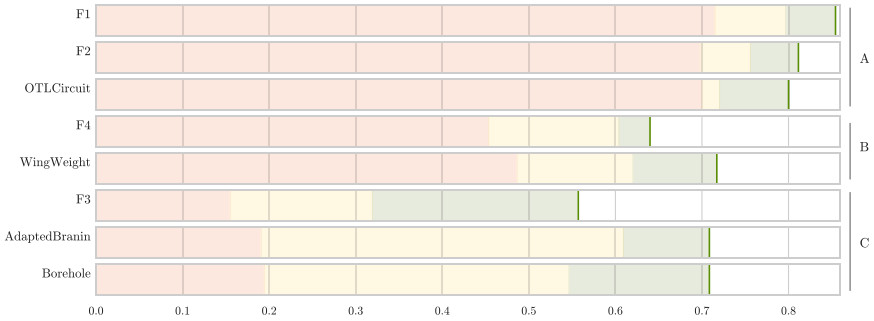

FIGURE 8
HV thresholds for all test function classifications.

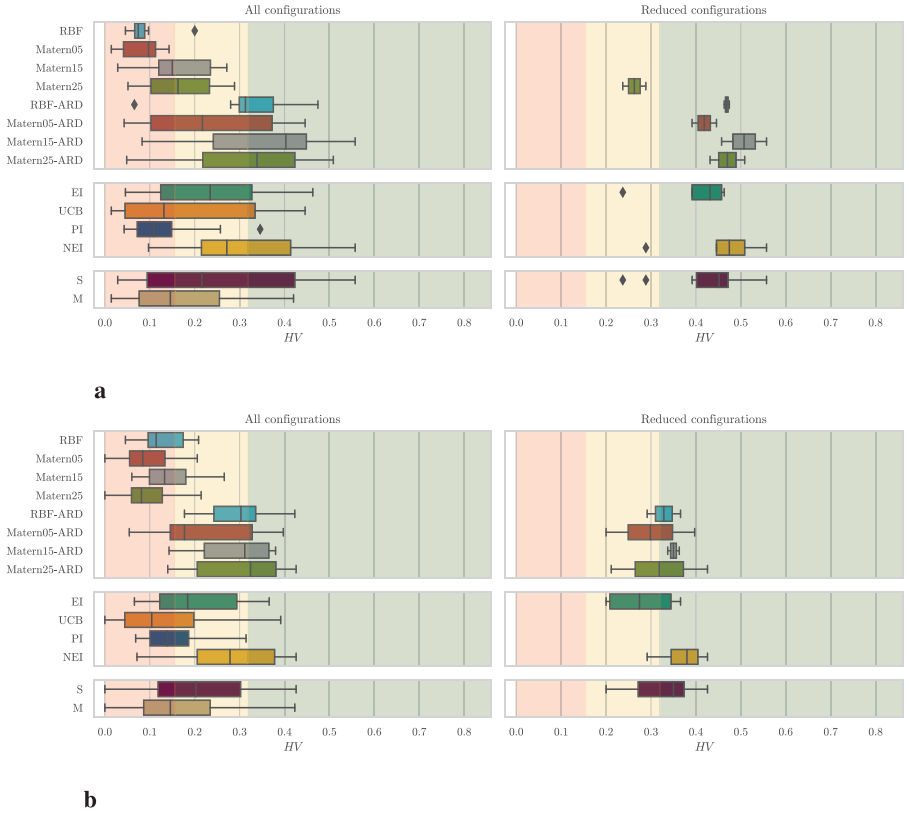

FIGURE 9
Classification of BO configurations for the noiseless F3. **(a)** Noiseless F3 and the noisy F3. **(b)** Noisy F3 test functions.

4.1 Kernels

Across the test functions, the Matern05 kernel is consistently classified as non-qualified, indicating its poor performance in optimizing the selected problems. However, in the case of AdaptedBranin and F4, Matern05 is categorized as undetermined, suggesting that its performance is not clearly better or worse than other configurations. For AdaptedBranin, it is undetermined for both noiseless and noisy cases, and for F4, it is undetermined only for noisy cases. This indicates that the performance of Matern05 on these two functions is not as straightforward as in other cases, and it may require further investigation to understand its behavior. Overall, Matern05 shows inferior performance across most test functions, making it a less preferable choice for optimizing these problems. Matern05-ARD performs poorly in all test functions and is never classified as the best option in both noiseless and noisy cases. For group C, it seems to be a plausible option in noiseless cases, but not a good option in noisy ones. Its overall performance is consistently inferior compared to other configurations, reinforcing the observation that Matern05-ARD is not a general recommended choice for optimizing

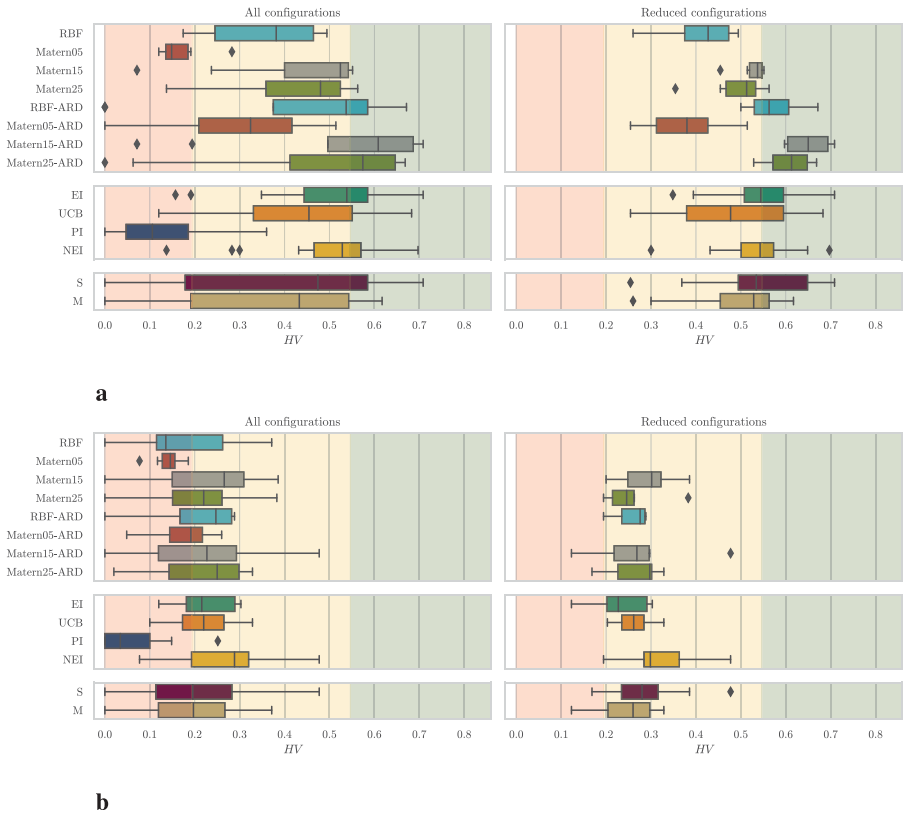

FIGURE 10
Classification of BO configurations for the noiseless Borehole. **(a)** Noiseless Borehole and the noisy Borehole. **(b)** Noisy Borehole test functions.

the test functions. Given its consistently poor performance, it is advisable to avoid using Matern05-ARD as a kernel configuration when applying BO to these test functions. RBF performs satisfactorily in optimizing the simple functions of group A. In group B, it is classified as qualified for noiseless functions and undetermined for noisy ones. However, for the complex group C, RBF encounters challenges in optimizing F3, remains undetermined for AdaptedBranin, and is undetermined for noiseless Borehole and non-qualified for its noisy case. Overall, RBF shows decent performance for simple functions but struggles in more complex and noisy scenarios. The RBF-ARD and Matern25-ARD kernels do indeed exhibit similar behaviors in many cases. For group A, they both perform well on both noiseless and noisy functions. However, for groups B and C, they show the trend of performing worse on noisy functions compared to noiseless ones. Matern25-ARD struggles particularly on the noisy functions F4, WingWeight, and AdaptedBranin, being consistently classified as non-qualified in these cases. On F3 and Borehole, Matern25-ARD is classified as qualified in noiseless cases and undetermined in noisy cases. On the other hand, RBF-ARD achieves the highest performance on F3 in both noisy and noiseless cases. It generally presents a qualified-undetermined classification in noiseless-noisy cases and is only classified as non-qualified in the noisy case of F4. Overall both RBF-ARD and Matern25-ARD kernels show potential for optimizing noiseless functions, while presenting an acceptable performance on noisy functions.

Matern25 performs well in group A (simple functions) but encounters difficulties in handling noisy cases for the more complex groups B and C. It is classified as undetermined for the Borehole function in both noiseless and noisy cases. Similar to RBF-ARD and Matern25-ARD, the performance of Matern25 is mixed, with drops in optimization observed in noisy conditions. While Matern25 shows satisfactory performance for simple functions, its ability to handle noise and complexity diminishes for more challenging optimization problems. This findings go hand in hand with the ones of Palar and Shimoyama (2019) and Le Riche and Picheny (2021). Matern15 and Matern15-ARD configurations show promising performance. They are classified as qualified for most test functions, offering good results. The isotropic Matern15 kernel outperforms the anisotropic one in group B (F4 and WingWeight), while the anisotropic version performs better in group C (F3, AdaptedBranin, and Borehole). Indeed, and as a general conclusion, group B seems to be optimized better by isotropic kernels and group C by anisotropic ones. Further investigation about the problems' landscape is needed, to adequately recommend a certain kernel for a given problem. However as a general recommendation, Matern15 and Matern15-ARD seem suitable for optimizing a wide range of problems.

4.2 Acquisition functions

Among all acquisition functions, PI consistently performs worse than all the other options, with the exception of

TABLE 2 Classification for noiseless (0) and noisy (1) test functions. Green: qualified, yellow: undetermined, red: non-qualified. Kernels with ARD are grayed out because, in one-dimensional functions like F2, there is no difference between isotropic and anisotropic kernels.

Configuration		F1		F2*		OTLCircuit		F4		WingWeight		F3		AdaptedBranin		Borehole	
	Noise	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Kernels	RBF																
	Matern05																
	Matern15																
	Matern25																
	RBF-ARD																
	Matern05-ARD																
	Matern15-ARD																
	Matern25-ARD																
Acq. Functions	EI																
	UCB																
	PI																
	NEI																
Init.Size	S																
	M																
	Group	A						B				C					

F4 where it is classified in noiseless and noisy cases as undetermined. In the noisy case, UCB and EI are worse, considered non-qualified. These observations highlight the general recommendation of not using PI as a default choice for an unknown process without further investigation. This finding is consistent with previous works Benjamins et al. (2022); Ath et al. (2021), which have also explained the poor performance of PI due to its greedy nature.

UCB has a better performance than PI, but still encounter difficulties in several functions. In both noiseless and noisy F1 and F3 is UCB classified as non-qualified, as well as for noisy F4. It remains undetermined for AdaptedBranin and Borehole and further simpler noisy functions. The performance of UCB could be due to its fixed β-parameter, which may prioritize exploitation over exploration, limiting its ability to effectively explore the search space and find the global optimum. As a result, UCB is generally not recommended as a default choice for an unknown process without investigating the influence of the β parameter and carefully tuning it for specific optimization tasks. This finding contradicts the outcomes of studies by Qin et al. (2021) and Diessner et al. (2022), who reported better results with UCB than with EI. Further investigation must be made regarding this acquisition function in production fields. Among all acquisition functions, EI and NEI stand out as the best performing options. EI performs consistently well in all test functions, but it encounters difficulties in optimizing noisy F4 and is undetermined in noisy WingWeight, F3 and both noiseless AdaptedBranin and Borehole. On the other hand, NEI shows the best performance across all test functions, both in noiseless and noisy cases, and although is undetermined in some functions, it is not outperformed by any other acquisition function. As a result, NEI appears to be the better default choice for new, unknown processes, providing robust and efficient optimization capabilities.

4.3 Initial sampling sizes

In group A, the small initial sampling size (S) clearly outperforms the medium one (M) in both noiseless and noisy cases. Similarly, in group C, the smaller sampling size tends to perform better than the medium one. Only in the noisy cases of group B, medium initial sampling seems to represent a better option than small initial sampling sizes. This could be due to the lack of exploration at the beginning of the experiment. In general, based on

the performance across six of eight test functions (groups A and C), the best option would be to begin with a small (S) initial dataset and prioritize the efficiency of the algorithm, underlining the state of the art presented in the introduction. If, during the course of the experiment, it is observed that the optimization is not sufficient, a couple of exploratory trials could be implemented in the mid-time to compensate for the possible lack of exploration at the initial steps. Such adaptive approaches to enhance a better balance between exploration and exploitation are under investigation Benjamins et al. (2022); Hoffman et al. (2010).

4.4 Summary of actionable guidelines for applying BED in manufacturing

In summary, we deduce the following findings and guidelines regarding the configuration of the kernel, acquisition function, and initial sampling size. In general, it appears that there is no one-fits-all solution for the different optimization problems, but rather that the different characteristics of the optimization problems place different demands on BO configuration. For kernels, RBF presents a reasonable choice for simple test functions, while we recommend Matern15-ARD as a reasonable default option for complex optimization problems. In principle, it is advisable to use anisotropic (ARD) kernels for more complex problems, such as those typically encountered in manufacturing. Since noise negatively impacts optimization performance, process and measurement noise should be minimized by precisely calibrating both actuators and sensors of the manufacturing process. With regard to acquisition functions, it appears that the exploration behavior has a significant influence on optimization performance, especially in the case of complex problems. Based on their exploratory behavior, we recommend EI and NEI as qualified default options, while PI and the investigated UCB configuration are not suggested. In terms of initial sampling size, we recommend keeping the additionally randomly generated experiment data small and instead leaving the search for the optimum to the BO algorithm with a sufficiently exploratory acquisition function. Already existing datasets for which no further experiments need to be conducted should nevertheless be utilized to initialize the BO algorithms. In addition, it is decisive to perform screening and characterization trials prior to optimization to determine the most critical parameters and associated parameter ranges, thus keeping the dimensionality of the optimization problem as small but also as influential as possible.

5 Conclusion

Optimization of production processes is an ongoing challenge for manufacturing companies in order to continuously improve product quality and process productivity, increase the overall equipment effectiveness, and thus remain economically competitive. Process optimization is becoming more complex given that a rising number of process parameters and objectives must be precisely adjusted to each other (e.g., due to the growing efficiency concerns, tighter quality specifications, and shorter product life cycles). Traditional experimental design methods are no longer able to cope with the increasing complexity of process optimization. With Bayesian Optimization (BO), Bayesian Experimental Design (BED) has evolved as an adaptive, data-driven approach to efficiently find optimal parameters in black-box optimization problems in the engineering domain. However, to successfully utilize BO in engineering use-cases, BO algorithms have to be precisely configured to the given problem. To investigate the performance of individual configurations of the BO algorithm for different optimization problems and to unravel insights that allow the derivation of practical guidelines, we designed and conducted a BED benchmark study comprising a total of 15,360 experiments.

As a result of our study, we present an extensive performance and robustness analysis that unveils significant performance differences between individual BO algorithms on different optimization problems. The results of the benchmark study provide empirical references and actionable guidelines for the configuration of BED. The study advocates BED as an adaptive, data-efficient tool for optimizing process parameters, achieving 95% precision within a budget of 35 iterations using the best-qualified configurations at various levels of complexity. We show that there is no universally optimal BO configuration. For complex optimization problems, particularly in manufacturing, anisotropic kernels such as Matern15-ARD are recommended, while exploration-oriented acquisition functions like EI or NEI offer robust default choices. Randomly generated initial experiments should be kept small and instead leave the search for the optimum to a sufficiently exploratory BO algorithm.

Furthermore, the results underscore the significant role of benchmark studies in not only identifying optimal BO configurations but also highlighting an existing research gap in terms of understanding the interplay between the characteristics of production processes and BED performance. The performance of the BO configuration unveils distinct intrinsic patterns in various test functions, indicating shared responses among certain test functions to the optimization process. Importantly, the results of our study fails to unravel a clear relationship between the characteristics of optimization problems and the performance of BO configurations. This suggests that current characteristics do not adequately capture the inherent patterns of the response of test functions to optimization. A focus of applied research must therefore lie on the investigation and identification of production process and data characteristics that correlate strongly with the performance of different BO algorithms, allowing one to make a profound configuration decision for successfully applying BED to new optimization problems. The limitations of our study include its focus on single-objective optimization and that it does not include a further examination of different hyperparameter sets for both kernels and acquisition functions. To further establish BED in production engineering practice and to accelerate process optimization and reduce development costs, our further research focuses on the collaboration between BED and domain experts comprising the integration of expert knowledge into BED. We examine the extension to multi-objective optimization cases and investigate the communication between domain experts to increase comprehensibility, and thus facilitate user acceptance and widespread adoption in industry applications.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

LL: Visualization, Writing – original draft, Formal Analysis, Project administration, Methodology, Validation, Conceptualization, Investigation, Data curation, Supervision, Writing – review and editing, Software. AG: Writing – review and editing, Formal Analysis, Writing – original draft, Methodology, Visualization, Conceptualization, Validation, Investigation, Software, Data curation. KB: Writing – review and editing, Methodology, Conceptualization, Supervision. JE: Writing – review and editing, Conceptualization, Data curation, Methodology. AS: Writing – review and editing, Supervision. RS: Writing – review and editing, Funding acquisition, Resources.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Founded by "ICNAP - International Center for Networked, Adaptive Production". A Fraunhofer Initiative.

Acknowledgements

This work is based on the master's thesis of Ana Maria Gonzalez Degetau entitled "Bayesian Machine Learning for Data-driven Optimization in Production Processes: A Benchmark Study" as part of the joint ICNAP research project "evolve" between Fraunhofer Institute for Production Technology IPT, Fraunhofer Institute for Laser Technology ILT and Fraunhofer Institute for Microbiology and Applied Ecology IME, and in cooperation with the Institute of Product Development and Engineering Design, Faculty of Process Engineering, Energy and Mechanical Systems of TH Köln - University of Applied Sciences.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmtec.2025.1614335/full#supplementary-material

References

Al-Hafez, F. (2021). Finding the optimal learning rate using bayesian optimization.

Arboretti, R., Ceccato, R., Pegoraro, L., and Salmaso, L. (2022). Design of experiments and machine learning for product innovation: a systematic literature review. *Qual. Reliab. Eng. Int.* 38, 1131–1156. doi:10.1002/qre.3025

Ath, G., Everson, R., Rahat, A., and Fieldsend, J. (2021). Greed is good: exploration and exploitation trade-offs in Bayesian optimisation. *ACM Trans. Evol. Learn. Optim.* 1, 1–22. doi:10.1145/3425501

Benjamins, C., Raponi, E., Jankovic, A., van der Blom, K., Santoni, M. L., Lindauer, M., et al. (2022). Pi is back! switching acquisition functions in bayesian optimization

Berger-Tal, O., Nathan, J., Meron, E., and Saltz, D. (2014). The exploration-exploitation dilemma: a multidisciplinary framework. *PLOS ONE* 9, 1–8. doi:10.1371/journal.pone.0095693

Bossek, J., Doerr, C., and Kerschke, P. (2020a). "Initial design strategies and their effects on sequential model-based optimization: an exploratory case study based on bbob," in *Proceedings of the 2020 genetic and evolutionary computation conference*, 778–786doi. doi:10.1145/3377930.3390155

Bossek, J., Kerschke, P., and Trautmann, H. (2020b). A multi-objective perspective on performance assessment and automated selection of single-objective optimization algorithms. *Appl. Soft Comput.* 88, 105901. doi:10.1016/j.asoc.2019.105901

Deneault, J. R., Chang, J., Myung, J., Hooper, D., Armstrong, A., Pitt, M., et al. (2021). Toward autonomous additive manufacturing: bayesian optimization on a 3d printer. *MRS Bull.* 46, 566–575. doi:10.1557/s43577-021-00051-1

Dieb, T. M., and Tsuda, K. (2018). in *Machine learning-based experimental design in materials science*. Editor I. Tanaka (Singapore: Springer Singapore), 65–74. doi:10.1007/978-981-10-7617-6_4

Diessner, M., O'Connor, J., Wynn, A., Laizet, S., Guan, Y., Wilson, K., et al. (2022). Investigating bayesian optimization for expensive-to-evaluate black box functions: application in fluid dynamics. *Front. Appl. Math. Statistics* 8, 1076296. doi:10.3389/fams.2022.1076296

Durakovic, B. (2017). "*Design of experiments application, concepts, examples: state of the art*," 5. Periodicals of Engineering and Natural Sciences PEN. International University of Sarajevo. doi:10.21533/pen.v5i3.145

Duris, J., Kennedy, D., Hanuka, A., Shtalenkova, J., Edelen, A., Egger, A., et al. (2020). Bayesian optimization of a free-electron laser. *Phys. Rev. Lett.* 124, 2825. doi:10.1103/PhysRevLett.124.124801

Duvenaud, D. K. (2014). *Automatic model construction with Gaussian processes*. Doctoral dissertation, University of Cambridge. Available online at: https://api.semanticscholar.org/CorpusID:107112403

Finck, S., Hansen, N., Ros, R., and Auger, A. (2010). Real-parameter black-box optimization benchmarking 2010: presentation of the noiseless functions

Forrester, A., Sbester, A., and Keane, A. J. (2008). *Engineering design via surrogate modelling*. Chichester, UK: John Wiley and Sons. doi:10.1002/9780470770801

Frazier, P. I. (2018). A tutorial on bayesian optimization

Freiesleben, J., Keim, J., and Grutsch, M. (2020). Machine learning and design of experiments: alternative approaches or complementary methodologies for quality improvement? *Qual. Reliab. Eng. Int.* 36, 1837–1848. doi:10.1002/qre.2579

Gan, W., Ji, Z., and Liang, Y. (2021). Acquisition functions in bayesian optimization. *Int. Conf. Big Data & Artif. Intell. and sotware Eng.* 2, 129–135. doi:10.1109/icbase53849.2021.00032

Garnett, R. (2023). *Bayesian optimization*. United Kingdom: TJ Books Limited, Padstow Cornwall.

Greenhill, S., Rana, S., Gupta, S., Vellanki, P., and Venkatesh, S. (2020). Bayesian optimization for adaptive experimental design: a review. *IEEE Access* 8, 13937–13948. doi:10.1109/ACCESS.2020.2966228

Guidetti, X., Rupenyan, A., Fassl, L., Nabavi, M., and Lygeros, J. (2022). Advanced manufacturing configuration by sample-efficient batch bayesian optimization. *IEEE Robotics Automation Lett.* 7, 11886–11893. doi:10.1109/LRA.2022.3208370

Haghanifar, S., McCourt, M., Cheng, B., Wuenschell, J., Ohodnicki, P., and Leu, P. W. (2020). Discovering high-performance broadband and broad angle antireflection surfaces by machine learning. *Optica* 7, 784. doi:10.1364/OPTICA.387938

Hoffman, M., Brochu, E., and Freitas, N. (2010). *Portfolio allocation for bayesian optimization*. UAI. doi:10.48550/arXiv.1009.5419

Ilzarbe, L., Álvarez, M. J., Viles, E., and Tanco, M. (2008). Practical applications of design of experiments in the field of engineering: a bibliographical review. *Qual. Reliab. Eng. Int.* 24, 417–428. doi:10.1002/qre.909

Jankovic, A., Chaudhary, G., and Goia, F. (2021). Designing the design of experiments (doe) – an investigation on the influence of different factorial designs on the characterization of complex systems. *Energy Build.* 250, 111298. doi:10.1016/j.enbuild.2021.111298

Le Riche, R., and Picheny, V. (2021). Revisiting bayesian optimization in the light of the coco benchmark. *Struct. Multidiscip. Optim.* 64, 3063–3087. doi:10.1007/s00158-021-02977-1

Leyendecker, L., Nausch, H., Wergers, C., Scheffler, D., and Schmitt, R. H. (2025). Bayesian experimental design for optimizing medium composition and biomass formation of tobacco by-2 cell suspension cultures in stirred-tank bioreactors. *Front. Bioeng. Biotechnol.* 13, 1617319. doi:10.3389/fbioe.2025.1617319

Liang, Q., and Lai, L. (2021). "Scalable bayesian optimization accelerates process optimization of penicillin production," in *NeurIPS 2021 AI for science workshop*.

Logothetis, N., and Wynn, H. P. (1989). *Quality through design: experimental design, off-line quality control and Taguchi's contributions*, 7. Oxford: Clarendon Press.

Maurya, A. (2016). "Bayesian optimization for predicting rare internal failures in manufacturing processes," in *2016 IEEE international conference on big data (big data)* (IEEE), 2036–2045. doi:10.1109/BigData.2016.7840827

Mersmann, O., Preuss, M., Trautmann, H., Ppsn, X. I., Schaefer, R., Cotta, C., et al. (2010). "Benchmarking evolutionary algorithms: towards exploratory landscape analysis," in *Parallel problem solving from Nature* (Berlin, Heidelberg: Springer Berlin Heidelberg), 73–82.

Močkus, J. (1975). "On bayesian methods for seeking the extremum," in *Optimization techniques IFIP technical conference Novosibirsk, July 1–7, 1974*. Editors G. Goos, J. Hartmanis, P. Brinch Hansen, D. Gries, C. Moler, G. Seegmüller, et al. (Berlin, Heidelberg: Springer Berlin Heidelberg), 400–404. doi:10.1007/3-540-07165-2_55

Mockus, J. (1989). *Bayesian approach to global optimization: theory and applications, vol. 37 of mathematics and its applications Soviet series*. Dordrecht: Kluwer Acad. Publ.

Montgomery, D. C. (2020). *Design and analysis of experiments*. tenth edition edn. Hoboken, NJ: Wiley.

Palar, P. S., and Shimoyama, K. (2019). Efficient global optimization with ensemble and selection of kernel functions for engineering design. *Struct. Multidiscip. Optim.* 59, 93–116. doi:10.1007/s00158-018-2053-9

Picheny, V., Wagner, T., and Ginsbourger, D. (2013). A benchmark of kriging-based infill criteria for noisy optimization. *Struct. Multidiscip. Optim.* 48, 607–626. doi:10.1007/s00158-013-0919-4

Qin, N., Zhou, X., Wang, J., and Shen, C. (2021). "Bayesian optimization: model comparison with different benchmark functions," in *2021 international conference on signal processing and machine learning (CONF-SPML) (IEEE)*, 329–333. doi:10.1109/CONF-SPML54095.2021.00071

Rainforth, T., Foster, A., Ivanova, D. R., and Smith, F. B. (2023). Modern bayesian experimental design

Sarabia, L., and Ortiz, M. (2009). "1.12 - response surface methodology," in *Comprehensive chemometrics*. Editors S. D. Brown, R. Tauler, and B. Walczak (Oxford: Elsevier), 345–390. doi:10.1016/B978-044452701-1.00083-1

Schmitt, R., and Pfeifer, T. (2015). *Qualitätsmanagement: Strategien - Methoden - Techniken*, 5. München: Hanser eLibrary. doi:10.3139/9783446440821

Smucker, B., Krzywinski, M., and Altman, N. (2018). Optimal experimental design. *Nat. Methods* 15, 559–560. doi:10.1038/s41592-018-0083-2

Surjanovic, S., and Bingham, D. (2021). Virtual library of simulation experiments: test functions and datasets

Tang, B. (1993). Orthogonal array-based latin hypercubes. *J. Am. Stat. Assoc.* 88, 1392–1397. doi:10.1080/01621459.1993.10476423

Wang, X., Jin, Y., Schmitt, S., and Olhofer, M. (2022). Recent advances in bayesian optimization