Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02597-2

Structured sampling strategies in Bayesian optimization: evaluation in mathematical and real-world scenarios

Lucas Greif[1] · **Niklas Hübschle**[1] · **Andreas Kimmig**[1] · **Simon Kreuzwieser**[1] · **Anatole Martenne**[1] · **Jivka Ovtcharova**[1]

Received: 22 August 2024 / Accepted: 19 March 2025

Abstract
This study presents a comprehensive evaluation of initial sampling techniques within the context of Bayesian Optimization (BO), a machine learning technique intended for the optimization of intricate and expensive functions. We assessed its efficacy in optimizing both theoretical benchmark functions and real-world applications. The findings reveal that, while BO is inherently robust and effective in a wide range of optimization problems, the integration of structured initial sampling methods, such as Latin Hypercube Sampling (LHS) and fractional factorial design (FFD) in the context of Design of Experiments (DoE), can significantly alter its performance. In addition, by systematically exploring different optimization strategies, the study highlights how LHS and FFD, followed by BO, can, for example, lead to substantial reductions in energy consumption—up to approximately 67.45% compared to average consumption. In conclusion, this study contributes to the growing body of knowledge on BO by demonstrating the value of early sampling techniques in enhancing BO effectiveness. This study offers a roadmap for future studies to build on in the pursuit of more efficient and effective optimization strategies in complex real-world scenarios.

Keywords Bayesian optimization · Machine learning · Latin hypercube sampling · Design of experiments · 3D printing

Introduction

In today's era of rapid technological advancement, optimizing manufacturing processes is more critical. It is crucial to improve productivity, reduce costs, and minimize energy consumption in industrial processes to foster sustainable development (Rubio et al., 2021). Moreover, companies must adapt quickly, accurately, and effectively to evolving production tasks (Iordache et al., 2019). A key aspect of this adaptation lies in selecting the right production parameters, as these directly affect defect formation, tolerances, and energy consumption in the manufactured part (Chia et al., 2022). Determining the ideal process parameters is especially challenging in high-mix-low-volume (HMLV) contexts such as additive manufacturing (Chia et al., 2022). However, ensuring production efficiency and maintaining quality standards (Becker et al., 2015) remain essential. HMLV involves producing small to moderately sized batches, typically ranging from a few units to several thousand. In this scenario, process optimization plays a pivotal role in maintaining economic feasibility and gaining a competitive edge.

Given the limited data points often available to identify the optimal process parameters in HMLV, the choice of an appropriate optimization method is critical. Bayesian optimization (BO) has garnered considerable attention because of its effectiveness in optimizing complex, expensive, or time-consuming objective functions. BO has been successfully applied in various applications, showcasing both its versatility and effectiveness. In hyperparameter tuning, reports indicate that BO reduces computational cost by approximately 30% compared to traditional grid search methods (Letham et al., 2017). Research in materials science suggests that the integration of BO can accelerate the discovery of new materials by 40%, as it allows more informed decision making and reduces the number of required experiments (Shields et al., 2021). In the context of engineering simulations, BO has reduced the number of simulations needed by

✉ Lucas Greif
Lucas.Greif@kit.edu

[1] Institute for Information Management in Engineering, Karlsruhe Institute of Technology, 76133 Karlsruhe, Germany

Published online: 28 March 2025

50%, leading to significant cost savings (Frazier & Wang, 2016). BO also accounts for environmental noise within its model, allowing it to make robust decisions even under uncertain conditions (Shields et al., 2021; Huan & Marzouk, 2011). Furthermore, BO can leverage prior knowledge about the objective function via prior distributions, thereby accelerating the optimization process by steering the search towards more promising regions of the parameter space (Frazier & Wang, 2016). Hence, its ability to capture uncertainty and integrate prior information makes BO highly versatile and widely applicable to numerous problems.

A recent literature review on BO identified seven key challenges requiring further investigation: prior-knowledge integration, optimizing high-dimensional spaces, managing constraints, conducting batch evaluations, addressing multi-objective optimization, leveraging multi-fidelity data, and handling mixed variable types (Greenhill et al., 2020).

One key challenge emphasized is the integration of prior knowledge, often achieved through initial sampling techniques. Initial sampling determines both the quality and coverage of the parameter space, directly influencing the surrogate model's predictions. Poor sampling can lead to uneven coverage, overlooking crucial regions, and weakening the initial surrogate model; which can significantly hinder overall BO performance.

Despite this importance, prior studies have largely focused on BO's overall efficiency and success without systematically investigating the impact of diverse structured sampling strategies. Existing research often overlooks the potential performance variations that stem from different sampling strategies. Although Design of Experiments (DoE) approaches are used extensively, they are typically treated as standalone methods rather than being systematically integrated into the BO framework. For example, some works used Latin Hypercube Sampling (LHS) when optimizing multi-objective problems such as vat photopolymerization (Sattari et al., 2024) or solar cell manufacturing (Liu et al., 2022), but did not statistically assess its impact compared to other sampling strategies or a baseline BO approach. Moreover, existing literature often targets highly specialized use cases, building tailored frameworks that justify the inclusion of structured sampling methods, e.g., LHS or Sobol sequences, based on domain knowledge or logical reasoning. However, these studies rarely offer a rigorous statistical evaluation of how such sampling strategies affect BO's performance. For example, in high-dimensional optimization scenarios, advanced techniques such as Multi-Objective Regionalized Bayesian Optimization (MORBO) (Daulton et al., 2022) or Cylindrical Thompson Sampling (CTS) (Rashidi et al., 2024) have demonstrated sample efficiency in large-scale problems with hundreds of dimensions. Nevertheless, these evaluations often involve a large number of iterations, which do not represent scenarios with limited computational or experimental budgets.

To address this gap, we integrate and evaluate structured sampling strategies within the BO framework in a systematic manner. In contrast to previous work, our focus is on lower-dimensional problems and significantly fewer iterations, a scenario that reflects real-world constraints in HMLV manufacturing and materials science. By analyzing the statistical impact of these sampling approaches, we aim to provide actionable insights into how initial sampling can enhance BO's performance, particularly when resources are limited.

In summary, our key contributions are:

- We establish a comprehensive evaluation framework to examine sampling strategies in mathematical functions with diverse characteristics, allowing systematic comparison.
- We perform a detailed statistical analysis to reveal performance differences when distinct sampling strategies are integrated into BO.
- Using a 3D printing case study and materials science datasets, we demonstrate the practical value of structured sampling in real-world optimization problems.

By bridging these gaps, our study offers a deeper understanding of how sampling strategies interact with BO, ultimately providing actionable insights to optimize complex and costly processes.

Theoretical foundations and related works

Bayesian optimization

BO is a machine learning-based method designed to optimize black-box functions that are costly to evaluate (Greenhill et al., 2020; Jones et al., 1998; Shahriari et al., 2016). It is particularly effective when the objective function lacks a closed-form expression and evaluations are expensive, such as in hyperparameter tuning of machine learning models or experimental design.

BO uses a surrogate probabilistic model to represent the unknown objective function $f(\mathbf{x})$. This surrogate model offers a posterior distribution over potential functions that align with the observed data, encapsulating both the mean prediction $\mu(\mathbf{x})$ and the uncertainty $\sigma^2(\mathbf{x})$ at each location within the domain (Frazier, 2018). This probabilistic approach enables BO to effectively balance exploration and exploitation by employing an acquisition function, which guides the choice of subsequent sampling points to locate the global optimum efficiently.

Given a dataset of observations $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \ldots, n\}$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the input

parameters and $y_i = f(\mathbf{x}_i) + \epsilon_i$ are the corresponding noisy evaluations with $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$, the GP surrogate model predicts the posterior mean and variance at a new point $\mathbf{x}$ using the kernel function $k(\mathbf{x}, \mathbf{x}')$ (Frazier, 2018).

The acquisition function $\alpha(\mathbf{x}; \mathcal{D})$ leverages the predictions of the surrogate model to guide the selection of the next evaluation point:

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x}; \mathcal{D}).$$

Common acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB), each balancing exploration and exploitation differently. For example, the UCB acquisition function is defined as (Srinivas et al., 2010):

$$\alpha_{\text{UCB}}(\mathbf{x}; \mathcal{D}) = \mu(\mathbf{x}) + \kappa\sigma(\mathbf{x}),$$

where $\kappa > 0$ controls the trade-off between exploration (sampling points with high uncertainty $\sigma(\mathbf{x})$) and exploitation (sampling points with low predicted mean $\mu(\mathbf{x})$).

BO operates iteratively, repeating the following steps until convergence criteria are met (Plöckinger et al., 2022):

1. *Surrogate model update*: update the GP model with the current dataset $\mathcal{D}$ to obtain the posterior distribution over $f(\mathbf{x})$.
2. *Acquisition function maximization*: solve

   $$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x}; \mathcal{D})$$

   to determine the next point to evaluate.
3. *Objective function evaluation*: evaluate the true objective function with possible noise:

   $$y_{\text{next}} = f(\mathbf{x}_{\text{next}}) + \epsilon.$$

4. *Dataset augmentation*: augment the dataset:

   $$\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{x}_{\text{next}}, y_{\text{next}})\}.$$

This process efficiently navigates the trade-off between exploration and exploitation, converging towards the global optimum of $f(\mathbf{x})$ (Mockus, 1989).

BO has been extended to handle various problem settings, including constrained optimization, multi-objective optimization, and high-dimensional spaces (Neumann-Brosig et al., 2018). In constrained BO, constraints are modeled probabilistically and the acquisition function incorporates the probability of satisfying these constraints (Gelbart et al., 2014). Multi-objective BO seeks to optimize several objective functions simultaneously by finding a set of Pareto-optimal solutions (Knowles, 2006).

The convergence criterion for BO can include a maximum number of function evaluations, a threshold for improvement between iterations, or limits to the computational budget (Plöckinger et al., 2022; Cashore et al., 2016). These criteria ensure that the optimization process terminates appropriately, especially when dealing with expensive evaluations.

Design of experiments and sampling

Efficient sampling strategies are essential in BO, especially when the number of evaluations is limited by cost or time constraints. By carefully selecting initial sample points, we can build a more accurate surrogate model with fewer evaluations, thereby improving the optimization process. This approach is particularly valuable in real-world parameter optimization problems, where only a limited number of initial samples are feasible.

Table 1 presents several techniques to initiate a surrogate model with sampled data.

Each method offers distinct advantages and drawbacks, and the optimal approach depends heavily on the specific features of the problem at hand. Two critical factors to consider are the available sampling budget and the dimensionality of the parameter space. A sufficiently large sample size guarantees comprehensive coverage of the design space, but as the number of parameters increases, the resources needed for detailed investigation may become prohibitive. When the sampling budget is limited, carefully selecting an appropriate sampling technique is essential to avoid overlooking critical regions.

In situations where minimal prior knowledge is available and the dimensionality remains modest, methods such as random sampling or LHS can provide an accessible starting point, delivering broad coverage without imposing strict assumptions. When computational budgets are severely constrained, LHS typically offers a more structured and reliable alternative to purely random approaches, ensuring more uniform coverage and potentially improving initial surrogate performance. If existing insights suggest that optima cluster near design-space boundaries, factorial and especially fractional factorial designs can be highly valuable. By systematically varying parameters at a carefully chosen subset of levels, these approaches help detect promising regions quickly and highlight how parameters interact to form local minima or maxima. In moderate-dimensional problems, fractional factorial designs are often more practical than full-factorial approaches because they mitigate the exponential growth in required samples.

Conversely, when no clear information exists regarding the location of the optimum, space-filling methods such as maximin or quasi-random sequences (e.g., Sobol sampling) can distribute points more evenly across the design space, helping to prevent the surrogate model from miss-

Table 1 Design of experiments methodologies

Method	Description	Advantages	Disadvantages
Factorial design	Evaluates all combinations of selected parameter levels (e.g., high and low values for each parameter)	Systematic exploration; identifies main effects and interactions between parameters Navid et al. (2018)	Number of required samples grows exponentially with the number of parameters; impractical for limited budgets Zhou et al. (2020)
Latin hypercube sampling	Divides each parameter range into intervals and samples each interval once, ensuring uniform coverage across dimensions	Good space-filling properties with limited samples; reduces sampling variance Stein (1987)	Effectiveness decreases in very high-dimensional spaces with few samples Deutsch and Deutsch (2012)
Maximin sampling	Selects points to maximize the minimum distance between any two samples, spreading them evenly across the design space	Ensures diverse sampling; good for constructing accurate surrogate models Xiong et al. (2009)	Computationally intensive to determine optimal points; may not be feasible with very limited samples or high-dimensional spaces Vořechovský and Eliáš (2020)
Random sampling	Randomly selects points within the design space	Simple to implement; no prior assumptions needed Renardy et al. (2021)	May result in uneven coverage; can miss important regions due to randomness Renardy et al. (2021)
Sobol sequence sampling	Uses low-discrepancy sequences to generate quasi-random, uniformly distributed samples	Provides uniform coverage; deterministic and reproducible; efficient for integrating over the space Renardy et al. (2021)	May require a specific number of samples to achieve desired uniformity; less flexible with very few samples Kucherenko et al. (2015)

ing critical but unexpected regions. The choice becomes more nuanced as dimensionality increases. Although LHS and Sobol sequences often remain useful, their effectiveness diminishes if the sampling budget cannot increase proportionally with dimensionality. In higher dimensions, additional strategies can be needed, such as adapting existing space-filling techniques or integrating domain-specific knowledge, to maintain efficiency.

Ultimately, the decision relies on balancing these trade-offs. If prior experience or preliminary analyses guide attention to particular regions of the parameter space, factorial or fractional factorial designs can quickly delve into those areas. If the distribution of optima is unknown and the parameter space is large, methods that prioritize coverage over structure offer a more robust initial exploration. In all cases, the choice of a sampling technique is critically dependent on available resources, the dimensionality of the problem, and the degree of confidence in hypothesizing where the optimum may lie within the design space.

Related works

In recent years, BO has become increasingly prominent in optimizing complex systems, especially where evaluations are expensive or time-consuming. This section reviews the relevant literature on BO's applications in various domains, with an emphasis on different sampling strategies and their role in optimizing manufacturing processes.

Application of Bayesian optimization in real-world scenarios

An increasing number of studies employ BO for real-world problems. For example, a physics-constrained multi-objective Bayesian Optimization (MOBO) algorithm was developed to optimize vat photopolymerization (VPP) of thermoplastics by identifying parameter sets that balance printing speed and material strength (Sattari et al., 2024). This work adopted Latin Hypercube Sampling (LHS) based on a recommendation for process optimization in solar cell manufacturing (Liu et al., 2022). However, the authors limited their scope to adjusting ink compositions and concluded that integrating both printing parameters and ink compositions would be too complex, as parameter scales differ significantly and would necessitate a larger experimental workload. In addition, they did not compare the impact of LHS sampling against a baseline BO approach without structured sampling.

Another study that recommended LHS sampling proposed a BO framework with probabilistic constraints, incorporating domain knowledge (e.g., film quality evaluations) and prior experimental data to optimize perovskite solar cell fabrication via Rapid Spray Plasma Processing (RSPP) (Liu et

al., 2022). The researchers performed initial sampling with LHS and then used a Gaussian Process (GP) surrogate model with an anisotropic kernel and an Upper Confidence Bound (UCB) acquisition function in subsequent BO iterations. The framework achieved a maximum power conversion efficiency (PCE) of 18.5% within 100 optimization iterations, surpassing both standard BO and conventional methods such as LHS alone, one-variable-at-a-time sampling (OVATS), and factorial sampling. However, the study did not present robust statistical tests to confirm the significance of these performance differences, and it relied on a single kernel and acquisition function. However, the research highlights the potential to integrate structured sampling and domain knowledge into BO, while demonstrating the need for more extensive statistical analysis and broader benchmarking.

In another example, BO was compared with evolutionary algorithms (EAs), such as the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), to optimize chlorine dosage in water distribution systems (WDS) (Moeini & Abokifa, 2024). The goal was to balance disinfection effectiveness with minimizing byproduct formation using multi-species water quality (MS-WQ) models. BO demonstrated superior computational efficiency and faster convergence, requiring fewer evaluations than GA and PSO to reach better optimization results. Although the study highlighted BO's sensitivity to parameters such as kernel length scale and acquisition functions, GA and PSO exhibited lower sensitivity but required more evaluations. The findings underscore the potential of BO for real-time WDS optimization, although exploring alternative surrogate models could further improve its robustness.

Guidetti et al. (2022) introduced a batch Bayesian Optimization (BBO) framework specifically suited for advanced manufacturing methods such as atmospheric plasma spraying and fused deposition modeling (Guidetti et al., 2022). BBO is sample-efficient, which is a critical advantage when physical experiments are expensive. This framework employs a novel acquisition function that effectively balances exploration and exploitation in batch settings, enabling the simultaneous evaluation of multiple sets of process parameters.

Besides tackling high-dimensional optimization directly, recent studies highlight the importance of balancing computational efficiency with performance. Sampling methods have been employed to handle the exponential expansion of the search space in high-dimensional challenges. Two notable methods are Multi-Objective Regionalized Bayesian Optimization (MORBO) (Daulton et al., 2022) and Cylindrical Thompson Sampling (CTS) (Rashidi et al., 2024), both employing distinct approaches to sampling. MORBO performs BO concurrently in multiple local regions of the design space, using a coordinated strategy that efficiently identifies diverse globally optimal solutions. It has shown notable improvements in sample efficiency on several synthetic benchmarks and real-world cases, such as optical display and vehicle design problems with up to 146 and 222 parameters, respectively. CTS, on the other hand, begins by sampling a modest number of points to determine which dimensions most influence the objective, thereby learning a reduced-dimensional trust region. Subsequent sampling iterations focus on these subspaces, updating them adaptively as optimization progresses. CTS typically needs 10 to 30 points to learn a reduced subspace, even in problems with hundreds of parameters, and often completes the entire optimization with fewer than 200 or 300 total points.

Applications of design of experiments methods

Joseph (2012) introduced a DoE-based Interpolation Technique (DoIt) that employs experimental designs to sample the parameter space more effectively (Joseph, 2012). DoIt was shown to estimate posterior distributions using fewer evaluations than methods like Monte Carlo or Markov Chain Monte Carlo, making it particularly resource-efficient for high-dimensional problems. Morishita and Kaneko (2022) proposed a clustering-based initial sampling approach for combinatorial optimization tasks, including chemical compound design, to ensure a uniform sample distribution between clusters (Morishita & Kaneko, 2022). Their results demonstrated up to a 5% reduction in the number of experiments needed to achieve optimal solutions compared to random sampling or D-optimality.

Huan and Marzouk (2011) developed a Bayesian experimental design framework for nonlinear systems, emphasizing optimal experimental setups for inference (Huan & Marzouk, 2011). They leveraged polynomial chaos expansions and Monte Carlo sampling to compute the expected information gain, guiding the selection of the samples. This approach proved to be effective for non-linear parameter inference problems, highlighting the potential of systematic sampling strategies such as LHS.

Beyond these specific optimization tasks, other research has explored parameter selection strategies for additive manufacturing (AM) processes such as fused deposition modeling (FDM). A study applied the Taguchi method to assess 27 parameter combinations, each with three replicates, using Analysis of Variance (ANOVA) to pinpoint the most influential sustainability factors (Camposeco-Negrete, 2020). The results indicated that the build plate and nozzle temperatures were most critical and that the layer height varied in importance depending on the product geometry. In a related project on sustainable manufacturing, another group integrated machine learning with a design of experiment approach to enhance pharmaceutical 3D printing sustainability (Li et al., 2023). Using a Definitive Screening Design (DSD), they identified and ranked the parameters that affect the environmental impact, then implemented ML models to

predict CO_2 emissions. This methodical combination of ML and DOE helped identify key variables that influence sustainability outcomes.

Existing benchmarking studies

In addition to domain-specific research, various benchmark investigations have compared multiple optimization methodologies. For example, a study evaluated seven metaheuristic algorithms in 11 mathematical functions that model different real-world challenges (Ismail & Halim, 2017). The functions exhibited diverse complexities, including modality, separability, discontinuity, and pronounced surface effects.An additional benchmark analysis in materials science evaluated various BO variants, as well as surrogate models including Gaussian processes with either isotropic or anisotropic kernels, and random forests (RF), for the purpose of optimizing material properties (Liang et al., 2021). The researchers measured factors such as the convergence rate, the enhancement factor, and the acceleration factor, underscoring the importance of aligning acquisition functions and surrogate models with the problem at hand. They also suggested evaluating RF kernels with LHS sampling, anticipating improved performance given RF's minimal assumptions about data. However, the study provided limited statistical evidence to confirm these improvements.

Research gap and contribution

Despite considerable advancements in BO and the growing interest in structured sampling techniques, multiple gaps persist in current research. First, although many studies incorporate strategies like LHS or Sobol sequences into BO, these integrations often remain ad hoc, lacking systematic comparisons with baseline BO methods. As a result, the influence of structured sampling on BO performance, particularly in low-iteration contexts with limited experimental or computational resources, remains ambiguous. Second, rigorous statistical validation is often missing. Although performance gains are frequently reported, few studies include formal hypothesis testing, confidence intervals, or non-parametric evaluations to separate genuine improvements from random variability or problem-specific peculiarities. Third, most research concentrates on high-dimensional or high-iteration scenarios typified by advanced frameworks like MORBO and CTS, which do not necessarily reflect the practical constraints many industrial or materials science applications face. Finally, there is a shortage of studies investigating the transferability of structured sampling methods outside their initial domain-specific contexts. Because many of these frameworks are custom-built and grounded in domain-centric arguments, their broader applicability is uncertain. Furthermore, there is no widely recognized benchmarking strategy for comparing sampling approaches across a diverse set of tasks, from synthetic mathematical functions to real manufacturing processes, further limiting the generalizability of current findings.

Our research addresses these deficiencies through a systematic and statistical evaluation of structured sampling methods integrated into the BO framework. We compare various sampling strategies with baseline BO approaches under controlled experimental conditions, testing mathematical functions selected to represent different levels of problem complexity. This methodology ensures that the results are not overly dependent on a single domain-specific case. We also embed statistical methods such as hypothesis testing and confidence intervals to validate whether observed performance gains exceed random noise or incidental effects. Unlike some prior research, we focus on settings with limited iteration budgets, reflecting the real constraints practitioners encounter due to time, cost, or resource limitations. Through this approach, our goal is to generate insights that are directly applicable even in resource-constrained environments. Ultimately, by systematically evaluating and benchmarking these methods, we seek to offer a robust foundation for choosing structured sampling strategies that can improve BO performance in a variety of real-world contexts.

Methodology

Scope and methods

Initially, the scope of the study must be determined by selecting the surrogate models, acquisition functions, and sampling methods that will be analyzed subsequently.

Surrogate models

We used four surrogate models: Gaussian Processes (GP) with isotropic and anisotropic kernels, Tree-structured Parzen Estimator (TPE), and Random Forests (RF).

Gaussian Processes (GP) are widely used in BO due to their probabilistic nature and flexibility. For GPs with isotropic kernels, the covariance function depends only on the Euclidean distance between points:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x}_p - \mathbf{x}_q\|^2}{2l^2}\right),$$

where σ_f^2 is the signal variance and l is the length scale. The isotropic kernel assumes that the function varies uniformly across all dimensions of the input space. The main assumption for isotropic kernels is that the objective function is smooth and stationary, meaning its statistical properties are invariant to translations in the input space. Using this covari-

ance function, the GP provides a predictive mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ for a new point $\mathbf{x}$:

$$\mu(\mathbf{x}) = \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y},$$
$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*,$$

where $\mathbf{K}$ is the covariance matrix of the training data, $\mathbf{k}_*$ is the covariance vector between the new point and the training data, σ_n^2 is the noise variance, and $\mathbf{y}$ is the vector of objective values from the training data.

In contrast, GPs with anisotropic kernels allow the covariance function to vary along different dimensions, providing greater flexibility for modeling functions where certain input dimensions are more influential than others. The covariance function for an anisotropic kernel is defined as:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(- \sum_{d=1}^{D} \frac{(x_{p,d} - x_{q,d})^2}{2 l_d^2} \right),$$

where l_d is the length scale for the d-th dimension. The anisotropic kernel assumes that the objective function is smooth and sufficiently differentiable, typically at least twice differentiable, to ensure proper gradient-based estimation of directional dependencies. This kernel is particularly suited for problems where the function exhibits varying rates of change along different dimensions.

Furthermore, we have used Tree-structured Parzen Estimator (TPE) as surrogate model. TPE is a probabilistic model particularly effective in high-dimensional optimization problems (Watanabe, 2023). Rather than directly formulating the objective function, TPE divides the search space into two regions: one corresponding to "good" configurations (objective values below a threshold) and another for "bad" configurations (values above the threshold). The objective function is modeled using two conditional probability densities, $g(y \mid \mathbf{x})$ for good configurations and $l(y \mid \mathbf{x})$ for bad configurations. The next evaluation point is chosen to maximize the ratio of these densities:

$$\gamma(\mathbf{x}) = \frac{g(y \mid \mathbf{x})}{l(y \mid \mathbf{x})}.$$

TPE assumes that the objective function exhibits clustering behavior, where certain regions of the parameter space correspond to significantly better values than others. This assumption allows TPE to focus on sampling in regions with a high potential for improvement.

Random Forests (RF) provide a non-parametric alternative to GPs and TPE. RFs aggregate predictions from an ensemble of decision trees, making them particularly robust for non-stationary objective functions. Traditionally, the prediction for a new point $\mathbf{x}$ is calculated as the mean of the predictions of all trees:

$$\mu(\mathbf{x}) = \frac{1}{n_{\text{tree}}} \sum_{k=1}^{n_{\text{tree}}} \hat{h}_k(\mathbf{x}),$$

where $\hat{h}_k(\mathbf{x})$ is the prediction from the k-th tree. In this study, we employ a median-based approach to enhance robustness against outliers and skewed distributions (Roy & Larocque, 2012):

$$\tilde{\mu}(\mathbf{x}) = \text{median}\left\{ \hat{h}_k(\mathbf{x}) : k = 1, 2, \ldots, n_{\text{tree}} \right\}.$$

The variability in predictions is captured using the median absolute deviation (MAD):

$$\tilde{\sigma}(\mathbf{x}) = \text{median}\left\{ \left| \hat{h}_k(\mathbf{x}) - \tilde{\mu}(\mathbf{x}) \right| : k = 1, 2, \ldots, n_{\text{tree}} \right\}.$$

RFs assume minimal structure on the objective function, allowing them to handle irregularities such as non-stationarity and heteroscedasticity effectively. We used $n_{\text{tree}} = 100$ and bootstrap = True which was shown to be suitable parameters for RF kernels (Liang et al., 2021).

Acquisition functions

For the acquisition functions expected improvement (EI) (Mockus et al., 1978), probability of improvement (POI) (Kushner, 1964), upper confidence bound (UCB) (Srinivas et al., 2010), and Stochastic Monte Carlo (SMC) (Vangelatos et al., 2021; Sheikh & Marcus, 2022) were considered. These acquisition functions can be expressed as follows:

For Expected Improvement (EI):

$$\alpha_{\text{EI}}(\mathbf{x}) = (f(\mathbf{x}_{\text{best}}) - \mu_t(\mathbf{x}))\, \Phi(Z) + \sigma_t(\mathbf{x}) \phi(Z)$$
$$\text{where } Z = \frac{f(\mathbf{x}_{\text{best}}) - \mu_t(\mathbf{x})}{\sigma_t(\mathbf{x})},$$

where $f(\mathbf{x}_{\text{best}})$ is the best observed value, $\mu_t(\mathbf{x})$ and $\sigma_t(\mathbf{x})$ are the predicted mean and standard deviation of the function $f(\mathbf{x})$ at point $\mathbf{x}$, $\Phi(Z)$ is the cumulative distribution function, and $\phi(Z)$ is the probability density function of the standard normal distribution.

For probability of improvement (POI):

$$\alpha_{\text{POI}}(\mathbf{x}) = \Phi\left(\frac{f(\mathbf{x}_{\text{best}}) - \mu(\mathbf{x}) - \kappa}{\sigma(\mathbf{x})} \right),$$

where Φ is the cumulative distribution function of the standard normal distribution, and κ is a tunable parameter.

For upper confidence bound (UCB):

$$\alpha_{\text{UCB}}(\mathbf{x}) = \mu(\mathbf{x}) - \kappa \sigma(\mathbf{x}),$$

where $\mu(\mathbf{x})$ is the predicted mean, $\sigma(\mathbf{x})$ is the predicted standard deviation, and κ controls the exploration-exploitation tradeoff.

For stochastic Monte Carlo (SMC):

$$\alpha_{\text{SMC}}(\mathbf{x}) = \mu(\mathbf{x}) - r(\mathbf{x}),$$

where $r(\mathbf{x})$ is sampled from $U(0, 2\sigma(\mathbf{x}))$, with $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ being the predicted mean and standard deviation returned by the surrogate model at $\mathbf{x}$. This approach samples from a truncated uniform distribution to introduce stochasticity in the acquisition process, encouraging diverse exploration.

Sampling strategies

Our attention is on two sampling methods: fractional factorial design (FFD) and Latin Hypercube Sampling (LHS).

Factorial Design provides systematic insights into the main effects and potential interactions between parameters, making it suitable for initial exploration when the design space is relatively small or when specific factor levels are of interest. In a full factorial design, denoted as 2^k, where k represents the number of factors each at two levels (typically designated as "low" and "high"), every possible combination of factor levels is evaluated. This results in a total of 2^k experimental runs. For example, with $k = 3$ factors, the full factorial design would involve $2^3 = 8$ experimental conditions. However, when the number of factors is large, evaluating all possible combinations becomes impractical or too expensive (Tsao & Patel, 2013). To address this, a fractional factorial design selects a fraction of the full factorial runs, typically a half, quarter, or another fraction, thus reducing the number of experiments needed. This reduction is achieved by confounding certain higher-order interactions, which are presumed to be negligible, allowing the study to focus on the most significant effects with fewer trials.

LHS ensures an even distribution across the design space with a limited number of trials, making it particularly effective for achieving good space coverage within the constraints of a small experimental budget. In LHS, each of the k dimensions of the parameter space is divided into N equally probable intervals, and one sample point is randomly selected from each interval. Mathematically, for a sample size N and k parameters, LHS ensures that for the i-th sample point $\mathbf{x}_i$:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik}),$$

where each x_{ij} is randomly selected from the j-th interval. This method guarantees that the full range of each parameter is explored, reducing the likelihood of clustering samples and improving the surrogate model's accuracy.

Using FFD, we can systematically evaluate interactions among a reduced subset of parameters, thereby optimizing the use of limited computational budgets while still identifying critical factors that influence performance. This approach is particularly valuable when preliminary insights, or even domain intuition, suggest that high-value regions may lie near particular boundaries or when assessing multifactor interactions is essential to refining the search domain. In contrast, LHS provides a more evenly distributed coverage of the parameter space, ensuring that each region is sampled more systematically than with purely random methods, even if prior knowledge is scarce. Furthermore, these methods are complementary: FFD targets organized blends of parameter levels to detect important effects and interactions, whereas LHS focuses on even sampling throughout the entire design domain. Consequently, the integration of both approaches will also be assessed.

Evaluation framework

The overall framework of the study is shown in Fig. 1.

In the *Initialization* phase, BO is utilized to identify the optimal combination of the surrogate model, the acquisition function, and, where applicable, the κ parameter. This phase systematically explores four acquisition functions: POI, UCB, EI, and SMC, with κ values ranging from 0 to 10. The *Data Generation* phase involves simulating optimization scenarios for various black-box functions. The optimization methods tested include BO without prior sampling, Random Search (RS), and BO initialized with structured sampling strategies, namely LHS, FFD, or a combination of LHS and FFD. Each optimization method is evaluated on a set of benchmark functions with varying modalities and dimensions. This phase involves repeating the optimization process 1000 times for each combination of function and method. In the *Evaluation* phase, the performance of the different methods is statistically assessed using bootstrapping and non-parametric tests. The Kruskal–Wallis test is applied to detect significant differences between methods. If significant differences are found, Dunn's test is used to conduct pairwise comparisons. This ensures a robust statistical analysis of the performance variations between methods. Finally, the framework proceeds to the *Validation* phase, where the approach is tested in both real-world and open-data scenarios. A practical case study involving 3D printing is conducted. Here, BO is used to minimize energy consumption while maintaining part quality. The study evaluates the effects of key process parameters, such as the printing temperature and the height of the layer, on energy efficiency and demonstrates the practical utility of the optimization framework. Additionally, the methodology is validated on publicly available materials science datasets. These datasets include diverse optimization challenges, such as improving yield in nanoparticle synthesis and optimizing structural properties in additive manufacturing.

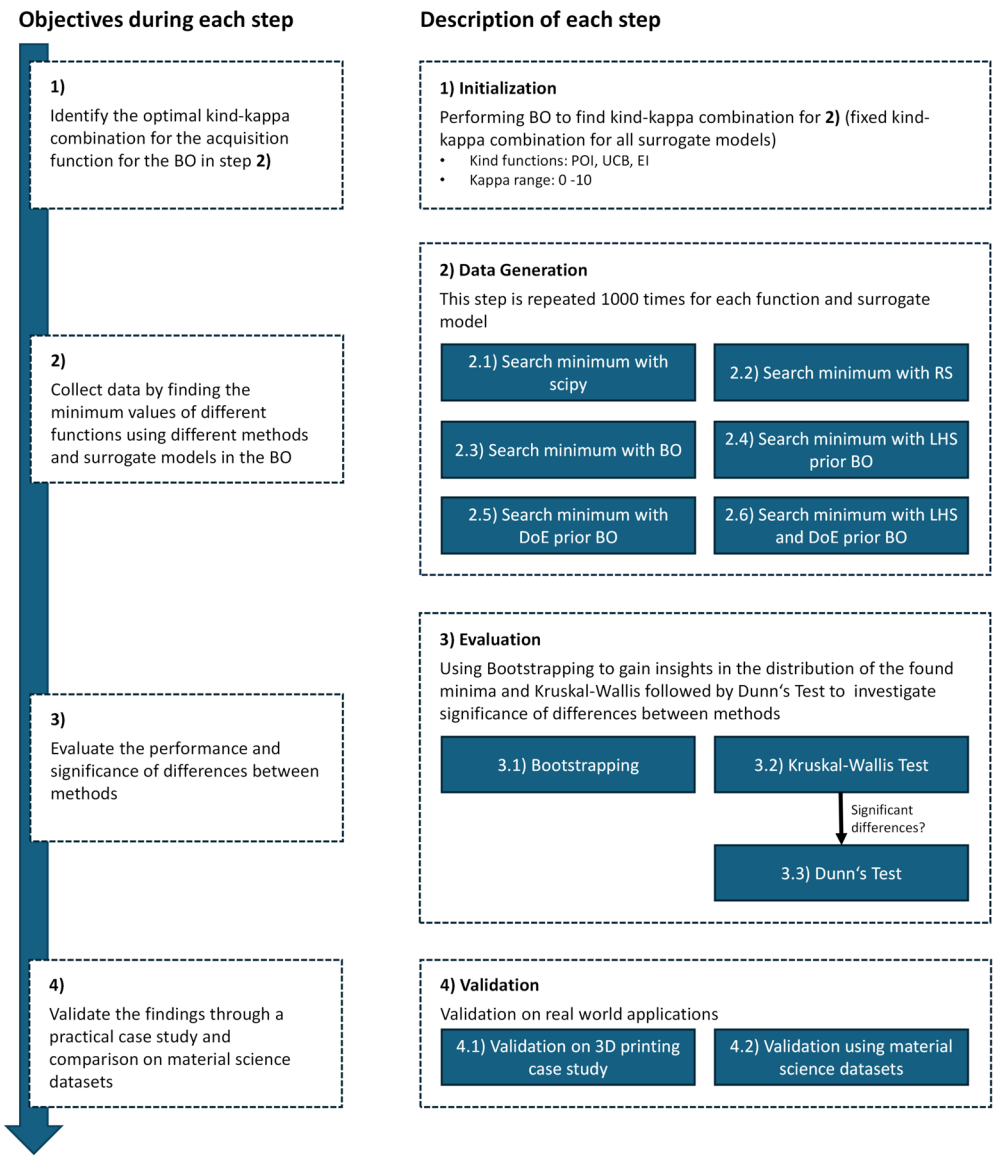


Fig. 1 Overall framework of the study

For our BO implementations, we use the public library BayesOpt[1] as documented in Martinez-Cantin (2014), GPyOpt[2] along with pyDOE3[3] for FFD and scipy.stats[4] to generate LHS data.

Initialization

Before starting the main BO procedure, we needed to determine the optimal combination of the surrogate model, the acquisition function, and, when applicable, the parameter κ.

We considered four surrogate models: GP with isotropic and anisotropic kernels, TPE, and RF, along with four acquisition functions: POI, UCB, EI, and SMC. Although we explored values of κ ranging from 0 to 10, it is important to note that κ is only meaningful for the acquisition functions that incorporate it into their formulation. In particular, κ affects the balance between exploration and exploitation in both UCB and POI, whereas EI and SMC do not incorporate the parameter κ.

To identify the best-performing configuration, we conducted an additional BO routine. Each candidate setup—defined by a particular combination of surrogate model, acquisition function, and (if relevant) κ—was evaluated

[1] https://github.com/rmcantin/bayesopt.

[2] https://github.com/SheffieldML/GPyOpt.

[3] https://github.com/relf/pyDOE3.

[4] https://github.com/scipy/scipy.

using the following loss function:

$$L = \sum_{i=1}^{i} \left| \frac{\text{mean}(\text{found\_min}_i)}{\text{true\_max}_i - \text{true\_min}_i} \right|,$$

where found_min_i denotes the list with the found minimum for each function i from each optimization iteration in this BO cycle with the proposed kind and κ value, true_max_i denotes the maximum of function i over the parameter space, and true_min_i denotes the minimum of function i over the parameter space.

where found_min_i denotes the distribution of minimum values obtained for the i-th test function across multiple independent optimization runs, and true_max_i, true_min_i represent the known maximum and minimum values of the i-th test function over its parameter space.

In this preliminary phase, we performed 100 BO function evaluations. For each candidate configuration, we performed 10 runs per test function, allowing 20 function evaluations per run to emphasize computational efficiency and identify minima with minimal overhead.

After computing the loss L for each evaluated set-up, we selected the configuration that yielded the lowest overall loss. This configuration—comprising the chosen surrogate model, the acquisition function, and, if applicable, the value of κ—was then used to guide our main suite of BO experiments.

Data generation

In order to evaluate the performance of the five methods, we initially simulated the scenario of determining the best parameter set for real experiments by utilizing mathematical functions as black-box functions, where the goal was to find the minimum value. Therefore, our optimization methods were not provided with information about the mathematical functions themselves but only with the function values in the proposed variable configurations $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$. Consequently, numerical minimization techniques that rely on gradient information are not applicable. Each optimization method was executed separately for each function, performing 1000 iterations for each combination. Our objective is to employ functions with differing properties, making them appropriate for evaluating algorithms that generalize optimization.

To reflect the varying bounds typically encountered in real-world scenarios, we selected different ranges for the variables listed in Table 2.

Different parameters naturally exhibit different ranges due to their specific physical, economic, or operational significance. By choosing varied bounds, we aim to improve flexibility and robustness.

Table 2 Bounds for each variable

Variable	Variable range
x_1	$[-0.5,\ 1]$
x_2	$[-20,\ 10]$
x_3	$[-0.5,\ 0.5]$
x_4	$[-20,\ 30]$
x_5	$[-5,\ 10]$
x_6	$[-10,\ 5]$

Table 3 outlines the attributes of each function used, including dimensionality (d), key properties, and modality. Generally, the Rosenbrock function exhibits unimodal behavior in lower dimensions ($d \leq 2$). However, for dimensions greater than 4, it is shown to be multimodal (Shang & Qiu, 2006).

Table 4 presents the mathematical formulas for each function.

Figure 2 illustrates the graphical representations of the benchmark functions, in order to facilitate understanding of the different functions and their characteristics.

In conclusion, by treating these mathematical functions as black-box functions, we aim to mimic different real-world application scenarios where the internal workings of the function are unknown, but optimization is still required.

Evaluation

To assess how well the methods optimize mathematical functions, we employ statistical techniques to examine the significance of the differences between the methods. Furthermore, to compare the minima found by each method with the actual minima of the parameter space, we use the "L-BFGS-B" algorithm (Zhu et al., 1997) implemented in the SciPy optimization library. The L-BFGS-B algorithm is particularly well suited for this task due to its ability to efficiently handle bound constraints while maintaining a balance between memory usage and computational cost.

For visualization and initial evaluation, a bootstrapping analysis was performed. Bootstrapping is a resampling method employed to approximate the sampling distribution of a statistic by drawing repeated samples with replacement from the initial dataset (Mooney & Duval, 1993). This method is particularly useful when the underlying distribution is unknown or when parametric assumptions are in doubt. Let $X = \{x_1, x_2, \ldots, x_n\}$ be our sample data. We generate B bootstrap samples X_b^* by sampling with replacement from X, where $b = 1, 2, \ldots, B$. For every bootstrap sample X_b^*, the corresponding statistic θ_b^* is calculated. The empirical distribution of $\{\theta_1^*, \theta_2^*, \ldots, \theta_B^*\}$ then provides the bootstrap estimate of the statistic's distribution. By applying bootstrapping, we can gain insight into the variability and

Table 3 Functions and their properties

Function	Dimensionality	Significant properties	Modality
Ackley Ackley (1987)	d	A function considered relatively straightforward because its form features a single funnel	Multimodal
Griewank Griewank (1981)	d	Many regularly distributed local minima and maxima, bowl-shape for large input spaces	Multimodal
Rastrigin Mühlenbein et al. (1991)	d	Overall shape is flatter than Ackley's function, fewer minima	Multimodal
Rosenbrock De Jong (1975)	d ∈ [4, 30]	Steep valley shape	Multimodal
Schaffer F7 Dieterich and Hartke (2012)	d	Concentric barriers need to be overcome to reach the global minimum	Multimodal
Schwefel Schwefel (1981)	d	Less symmetric, many local minima and maxima, no overarching slope directing towards the global minimum	Multimodal

Table 4 Functions and their equations

Function	Equation
Ackley	$f(x) = -20 \cdot \exp\left(-0,2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)\right) + 20 + e$
Griewank	$f(x) = \sum_{i=1}^{d} \frac{x_i^2}{4000} - \prod_{i=1}^{d} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$
Rastrigin	$f(x) = 10 \cdot d + \sum_{i=1}^{d} \begin{cases} 10 \cdot x_i^2, & \text{if } x_i > 5.12 \text{ or } x_i < -5.12 \\ x_i^2 - 10 \cdot \cos(2\pi x_i), & \text{if } -5.12 \le x_i \le 5.12 \end{cases}$
Rosenbrock	$f(x) = \sum_{i=1}^{d-1} \left(100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2\right)$
Schaffer F7	$f(\mathbf{x}) = \frac{1}{d}\sum_{i=1}^{d} \left(\sqrt{x_i^2 + x_{i+1}^2} \cdot \left(\sin\left(50 \cdot \left(\sqrt{x_i^2 + x_{i+1}^2}\right)^{0.2}\right) + 1\right)\right)^2$
Schwefel	$f(x) = 418.9829 \cdot d - \sum_{i=1}^{d} x_i \sin(\sqrt{\lvert x_i \rvert})$

distribution of the minima found by our optimization methods. This facilitates a more robust comparison by allowing us to compute confidence intervals and visualize the distribution of the results.

In addition, further statistical procedures, whether parametric or non-parametric, are applied to assess the data produced by different techniques. Parametric methods necessitate certain assumptions about data distribution, such as a normal distribution, whereas non-parametric methods do not rely on these assumptions and are instead based on the rankings or distributions obtained directly from the data. Consequently, if the assumptions required for parametric methods are met, they tend to be more efficient and accurate, rendering them a better choice than non-parametric methods.

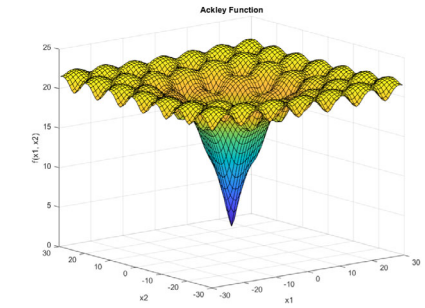
We presuppose that the assumptions of data independence and normal distribution might not be fully met. To investigate this, we examine the QQ plots of the residuals in our data.Furthermore, to assess the normality of the dataset, we utilize the Shapiro–Wilk test. This test examines the null hypothesis that the data follow a normal distribution (Razali & Wah, 2011). The test statistic W is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

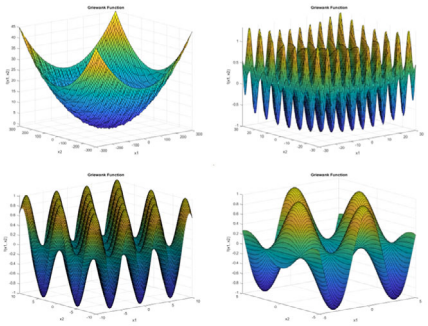
where $x_{(i)}$ represent the ordered values within the sample (order statistics), $\bar{x}$ indicates the sample mean, a_i are constants obtained from the expected values of the order statistics of a standard normal distribution along with their covariance matrix, and n signifies the size of the sample.

A lower value of W indicates a departure from normality. The p-value corresponding to W is used to decide whether to reject the null hypothesis.

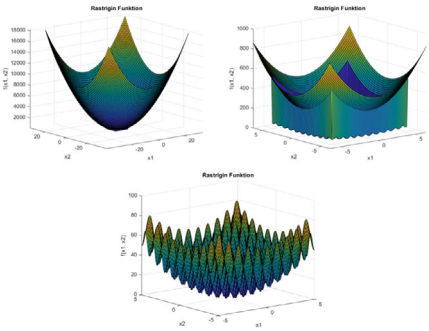
Thus, we resort to non-parametric approaches. We employ a mix of the Kruskal–Wallis test and Dunn's test to assess the significance of mean differences. Initially, we examine whether there are significant mean differences between the methods at a significance level of $\alpha = 0.05$ for function evaluations of 10, 20, 30, and 100 per function. If significant differences are found, we use Dunn's test to determine which pairs of methods have a significant difference at the $\alpha = 0.05$ level.



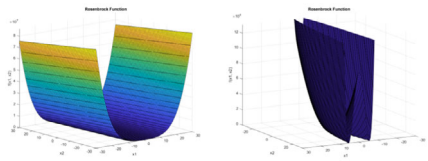
(a) Ackley Function



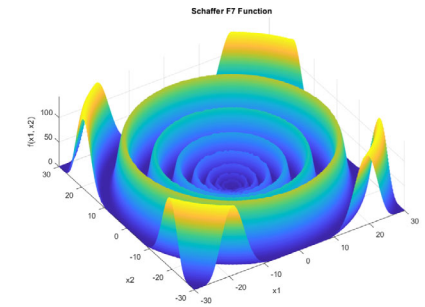
(b) Griewank Function



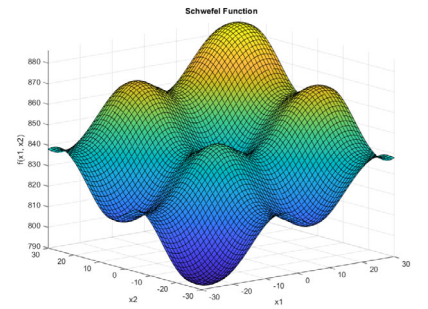
(c) Rastrigin Function



(d) Rosenbrock Function



(e) Schaffer F7 Function



(f) Schwefel Function

Fig. 2 Function plots

Considering the framework of our experiments, it is reasonable to state that the assumptions required for the Kruskal–Wallis test are satisfied. In our experiment, each function is optimized independently. Furthermore, we use the same kind and κ combination to perform the BO, so we can assume that our data are independent and identically distributed. Moreover, when considering each function and method, the optimization process remains consistent across function evaluations, with differences only in the stochastic elements of the algorithms, such as the initial point of the BO. Therefore, we can conclude that the minima identified for each method and function are derived from the same stochastic distribution function, and therefore all the assumptions for the Kruskal–Wallis and Dunn test are satisfied.

The test statistic of the Kruskal–Wallis test (Ostertagova et al., 2014) is then given by

$$H := \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{1}{n_j} \left(W_j - \frac{n_j(N+1)}{2} \right)^2 ,$$

where W_j is the rank sum of the j-th sample, n_j the number elements in thee j-th sample, $N := \sum_{j=1}^{k} n_j$ and $k = 5$ denotes the number of groups, in our case our five methods. Under our null hypothesis that the different groups come from the same probability distribution, the test statistic in the distribution converges to the chi-square distribution with parameter $k - 1$ degrees of freedom. We therefore reject the

null hypothesis if

$$H > \chi^2_{k-1;1-\alpha}$$

applies. In this context, $\chi^2_{k-1;1-\alpha}$ represents the $(1-\alpha)$-quantile of a chi-squared distribution having $k-1$ degrees of freedom.

In order to compare the different methods pairwise, we now consider Dunn's test along with the Bonferroni correction (Dinno, 2015). Dunn's test operates by making pairwise comparisons between the ranks of the groups, adjusting for multiple comparisons through the Bonferroni correction. The test statistic of the Dunn's test is calculated as follows.

$$z_i = \frac{\frac{1}{n_i}W_i - \frac{1}{n_j}W_j}{\sigma_i},$$

where W_i, W_j denotes the rank sum of the $i-th$ and $j-th$ sample as before and σ_i is the standard deviation with ties correction, which is calculated as

$$\sigma_i = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r}(\tau_s^3 - \tau_s)}{12(N-1)}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Here, the number of tied ranks is denoted by r and the number of observations that are tied to the s-th value is denoted by τ_j. Then our test statistic z_i follows approximately the standard distribution. After calculating the Bonferroni adjusted p-value, we look at each method and count whether it is better, i.e. a smaller mean and a significant difference.

Validation

After finalizing the theoretical framework of our optimization approaches, we proceeded to validate their performance under realistic conditions. The overarching goal of this validation phase was twofold: first, to assess the robustness of the optimization strategies in the presence of noise and process variability, and second, to confirm their effectiveness in practical applications. To achieve this, we implemented and tested the proposed methods in a controlled 3D printing scenario using polylactic acid (PLA) as the printing material. By systematically varying process parameters and monitoring energy consumption, we could evaluate how well the optimization techniques minimize resource use without compromising print quality. Beyond this initial application, we further verified the generalizability of our methods by benchmarking them on five distinct materials science datasets. These datasets reflect a variety of manufacturing processes and optimization objectives, thereby providing a comprehensive assessment of the adaptability of the methods. In the following paragraphs, we detail the 3D printing experiments and their parameter setups, as well as the procedures and datasets used in our broader materials science benchmarks.

Fig. 3 Designed part

Before detailing the specific experiments and data sources used, it is important to clarify the nature of the functions we seek to model. In the context of additive manufacturing (AM), these functions often represent complex, multi-faceted relationships between controllable input parameters (e.g., temperature, print speed, material composition) and output performance measures (e.g., energy consumption, mechanical strength, dimensional accuracy). Such relationships typically lack simple closed-form expressions and are instead treated as "black-box" functions, constructed from empirical data and surrogate modeling techniques. By characterizing the underlying response surfaces in this manner, we can guide our optimization strategies more effectively, ensuring that they adapt not only to specific 3D printing conditions but also scale to other manufacturing processes with minimal assumptions. This same principle applies to the materials science datasets: while each dataset may reflect distinct physical processes and objectives, the complex, interdependent nature of their parameters and outputs similarly calls for a data-driven, model-agnostic approach that can generalize beyond the particularities of any one domain.

3D Printing

3D printing of PLA involves variables such as printing temperature, print speed, infill density, and printing bed temperature, all of which influence the energy consumption and thus the carbon footprint of the printing process. Our objective was to find optimal settings that reduce energy consumption while maintaining print quality.

The designed part is shown in Fig. 3. This component is derived from a case study related to the German research initiative AIBetOn3D,[5] which seeks to improve 3D printing technology for construction materials and consequently mitigate their environmental impact. The energy consumption was monitored in real time using a power meter (TAPO P110) connected to the 3D printer (Prusa MK4).

To approximate the energy consumption of the 3D printer, the trapezoidal rule was applied to the power consumption data over time. This method involves summing up the areas of trapezoids formed under the power-time curve, which pro-

[5] AI-assisted 3D printing for building materials.

Table 5 Bounds for the 3D printing experiments

Parameter	Bounds
Infill mass proportion	0–30%
Printing temperature	190–230 °C
Build plate temperature	30–60 °C
Layer height	0.1-−0.3 mm
Print speed	40–150 mm/s

vides an approximation of the total energy used. The formula for the Trapezoidal Rule is given by:

$$\text{Energy Consumption} \approx \sum_{i=1}^{n} \frac{(t_i - t_{i-1})}{2}(P_i + P_{i-1})$$

where t_i represents the time stamps and P_i represents the power readings on those time stamps.

Setting a unique geometry allows us to directly identify the impact of a parameter modification on energy consumption. However, this approach lacks generalizability, as certain features that a parameter could theoretically impact more significantly than another may be underrepresented. This could lead to optimization being reliable only for the specific part on which the model was based. To mitigate this bias, we selected parameters that are more generalizable by nature. For example, instead of modifying the infill percentage parameter available in the slicer software, we opted to modify the infill mass over the total mass ratio, which is independent of the geometry.

The chosen parameters and their respective bounds are listed in Table 5.

To comprehensively evaluate the energy consumption optimization for 3D printing, we conducted a total of 80 experiments. These experiments were divided into different strategies to assess the effectiveness of various optimization methods.

We performed:

1. 20 experiments solely using BO.
2. 8 experiments using LHS followed by 12 experiments with BO.
3. 8 experiments using FFD followed by 12 experiments with BO.
4. 8 experiments using FFD and LHS, followed by 4 experiments with BO.

Material science

To enhance the robustness of our findings, we evaluated the sampling strategies on five publicly accessible materials science datasets, following the benchmarking approach outlined in (Liang et al., 2021). Each dataset was standardized into an optimization framework where the objective was formulated as a global minimization problem. Table 6 presents the characteristics of the datasets.

The datasets were derived from high-throughput experimental systems and represent a variety of challenges in materials science. BO was used during the data collection process for some datasets, including P3HT/CNT, AgNP, Perovskite, and AutoAM, to guide the selection of subsequent experimental conditions in their respective materials optimization campaigns.

Figure 4 demonstrates the probability density function of objective values for each dataset.

The datasets reveal varied distributions, underscoring the discrepancies in complexity and variability across the optimization landscapes.

In order to execute the BO framework, we calculate the Euclidean distance between each point in the dataset and all others. The design space be represented by a set of points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$. The Euclidean distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2},$$

where $x_{i,k}$ and $x_{j,k}$ are the k-th components of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

For each point $\mathbf{x}_i \in \mathcal{X}$, the closest neighbor $\mathbf{x}_j$ is determined by solving:

$$\mathbf{x}_j = \arg \min_{\mathbf{x} \in \mathcal{X}, \mathbf{x} \neq \mathbf{x}_i} d(\mathbf{x}_i, \mathbf{x}).$$

The point $\mathbf{x}_j$ is then used as feedback in the BO framework for the subsequent optimization iteration.

Results

Hyperparameter optimization of the acquisition function

After performing the preliminary BO routine to determine the best configuration for each surrogate model, we identified the acquisition function and the value κ (when applicable) that minimized the loss L defined earlier. Table 7 summarizes these results for the four surrogate models considered: Gaussian Processes (GP) with isotropic and anisotropic kernels, Tree-structured Parzen Estimators (TPE), and Random Forests (RF).

In the case of the isotropic GP, the best performance was obtained with the SMC acquisition function, which does not require κ, allowing for a flexible exploration-exploitation trade-off driven by stochastic sampling. The anisotropic

Table 6 Material science datasets and their properties

Dataset	d	Domain	Process	Objective	Size
AgNP Mekki-Berrada et al. (2021)	5	Silver nanoparticles	Flow synthesis	Yield	3295
AutoAM Deneault et al. (2021)	4	Materials manufacturing	3D Printing	Shape score	100
Crossed Barrel Gongora et al. (2020)	4	3D printed structure	Mechanical toughness	Strength	1800
P3HT Dieterich and Hartke (2012)	5	Organic materials	Spin coating	Multimodal	233
Perovskite Sun et al. (2021)	3	Thin film perovskite	Spin coating	Stability score	139

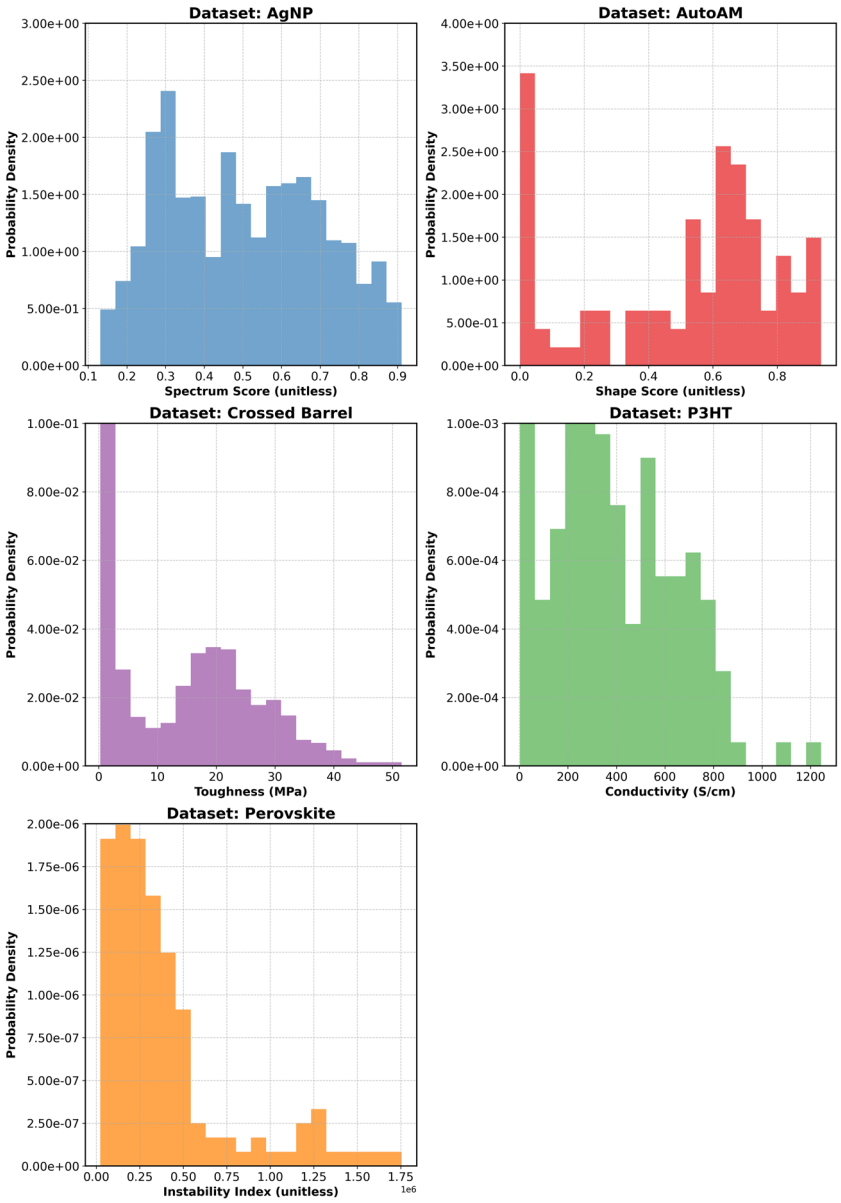


Fig. 4 Histogram of objective values of the Material Science Datasets

Table 7 Best-performing configurations for each surrogate model, including the chosen acquisition function and, if applicable, the κ value, along with the resulting loss L

Surrogate model	Acquisition function	κ	**L**
GP (isotropic)	SMC	–	1.185
GP (anisotropic)	POI	7.39	1.178
TPE	UCB	6.94	1.194
RF	UCB	6.83	1.201

GP model, on the other hand, performed best with POI at $\kappa = 7.39$. This relatively high κ increased the improvement threshold, encouraging exploration to find more substantial improvements beyond the current best solutions. For TPE and RF, both achieved their lowest loss values using UCB with κ values of approximately 6.94 and 6.83, respectively. In these cases, the elevated κ discouraged sampling high-uncertainty regions, leaning the search more toward exploitation of areas already identified as promising. In conclusion, the most effective arrangement for every surrogate model exhibited a distinct equilibrium between exploration and exploitation. This equilibrium was influenced by both the selection of the acquisition function and the adjustment of κ.

Mathematical function optimization

In this section, we present a comprehensive analysis of the optimization methods applied to a series of benchmark mathematical functions. We begin by examining their convergence behavior using a representative case. Figure 5 shows the convergence behavior of the Ackley function and isotropic kernel. The x-axis represents the iteration count, while the y-axis shows the average best value obtained by each method. The analysis includes confidence intervals to provide insight into variability between trials.

We introduce a series of predefined checkpoints at iterations 10, 20, 30, and 100, using them as reference points to highlight how performance evolves over time. Checkpoint 1 (CP1) at iteration 10 represents the early stage of optimization, where initial differences in performance between the methods can be observed. Checkpoint 2 (CP2) at iteration 20 highlights the mid-way point of rapid convergence, as most methods show significant improvement at this stage. Checkpoint 3 (CP3) at iteration 30 marks a point where convergence begins to stabilize for some methods, providing a clearer picture of their relative performance. Finally, Checkpoint 4 (CP4) at iteration 100 serves as the final checkpoint, where the differences in sampling diminish. At each checkpoint, we examine the statistical properties of the results to determine whether the differences observed between methods are statistically significant. To ensure the validity of this comparison, we first evaluated the underlying distributional assumptions. Examination of QQ-Plots indicated that the residuals do not follow a normal distribution. To confirm this, we performed the Shapiro–Wilk test at a 5% significance level, finding that 342 of the 384 tests rejected the normality assumption. As a result, analysis of variance and other parametric methods are inappropriate. Therefore, we employ nonparametric tests, specifically the Kruskal–Wallis test followed by Dunn's test, to assess differences in the medians of the methods. With the statistical framework established and the appropriate nonparametric tests identified, we now proceed to present the results for each of the selected checkpoints. For each iteration milestone, we will:

1. Deploy bootstrapping to illustrate the performance trends for each function and kernel.
2. Identify general observation trends across all functions and kernels, noting any observed convergence patterns.
3. Conduct kernel-specific analyzes to determine whether certain strategies are particularly advantageous for specific kernels.
4. Apply the Kruskal–Wallis and Dunn tests to evaluate the statistical significance of performance differences.

General observations at iteration 10

By the 10th iteration, as shown in Fig. 6, clear performance patterns begin to emerge. The performance of FFD and LHS prior BO is not examined because this method necessitates 18 iterations for initialization, making it unfeasible to assess after just 10 iterations. BO using FFD as a prior consistently achieves lower bootstrap mean values and tighter confidence intervals than other methods across all kernels for the Schaffer F7, Rosenbrock, and Schwefel function. When comparing FFD prior BO against all alternatives, the Kruskal–Wallis test at the 5% significance level confirms that these differences are statistically significant for the Schwefel, Schaffer F7, and Rosenbrock function. However, for the Schaffer F7 function, significant differences with the anisotropic kernel occur only when comparing FFD prior BO to LHS prior BO. Subsequent Dunn's tests further confirm these findings, indicating that at iteration 10, FFD prior BO outperforms the other methods.

In comparison, RS demonstrates performance on par with most BO methods. However, generally wide confidence intervals highlight the variability that emerges from the limited evaluation steps at this iteration count.

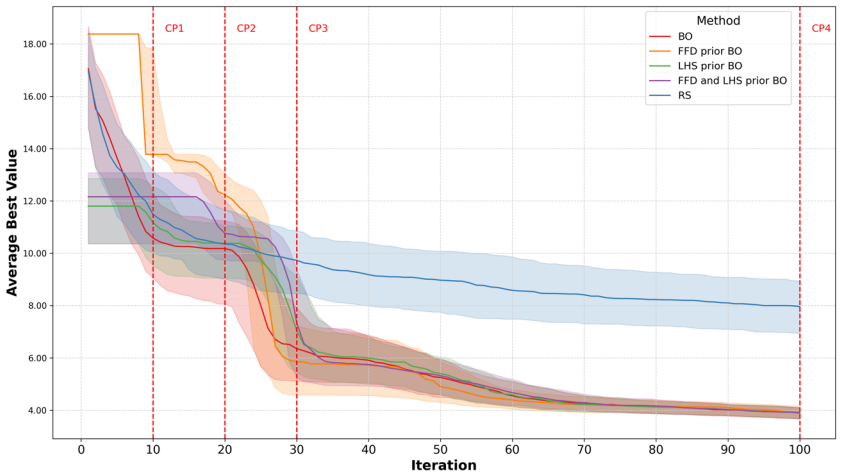


Fig. 5 Convergence analysis of optimization methods

For the Ackley and Griewank function, isotropic kernels using LHS or without priors, as well as anisotropic kernels without priors, yield lower mean values. Additionally, for the Rastrigin function, anisotropic kernels without priors attain the lowest mean values.

Kernel-specific observations at iteration 10

1. *Anisotropic kernel:* BO without priors achieves the lowest bootstrap mean values for the Ackley, Griewank, and Rastrigin function. The Kruskal–Wallis test at the 5% level indicates significant differences.
2. *Isotropic kernel:* when FFD does not produce the best result, subsequent BO phases are outperformed by LHS sampling or no prior approaches. Only for the Rosenbrock function does the Kruskal–Wallis test at the 5% level show significant differences when comparing LHS-prior BO to BO without priors.
3. *RF kernel:* apart from scenarios favoring FFD, LHS sampling leads for every tested function except Schaffer F7. The Kruskal–Wallis test at the 5% level indicates significant differences for the Ackley and Rastrigin function when comparing LHS-prior BO to both BO without priors and FFD prior BO, and for the Rosenbrock and Schwefel function when comparing LHS-prior BO to BO without priors.
4. *TPE kernel:* apart from conditions favoring FFD, LHS sampling outperforms no-prior scenarios at this iteration. However, the Kruskal–Wallis test at the 5% significance level indicates significant differences only for the Rosenbrock and Schaffer F7 function.

General observations at iteration 20

By the 20th iteration, as depicted in Fig. 7, a more nuanced performance landscape emerged. For the Ackley, Rastrigin, and Rosenbrock function, anisotropic kernels without priors consistently achieve superior performance. A Kruskal–Wallis test at the 5% level indicates significant differences when comparing BO without priors (anisotropic kernel) to all other surrogate models. In contrast, for the Griewank function, isotropic kernels with LHS priors produce lower mean values. Although these results differ significantly from all RF and TPE kernel methods, they are not significantly different from the anisotropic no-prior approach. For the Schaffer F7 function, the RF, TPE, and anisotropic kernels outperform the isotropic kernel; comparing the best methods of each kernel shows that these three are significantly better than the isotropic kernel, although there are no significant differences between them. With the Schwefel function, anisotropic kernels without priors and isotropic kernels across all methods yield the best results. However, there is no significant difference between these approaches. Notably, RS has now been surpassed by a wide range of kernel and sampling configurations.

Kernel-specific observations at iteration 20

1. *Anisotropic kernel:* without priors, the anisotropic kernel continues to deliver leading performance across multiple functions. However, confidence intervals are generally wider than those of other kernels. The Kruskal–Wallis test at the significance level 5% confirms significant differences for the Ackley and Griewank function.
2. *Isotropic kernel:* performance varies more widely. In some cases, no-prior configurations yield the best outcomes, while in others, LHS sampling is more effective. For the Rosenbrock and Rastrigin function, FFD priors remain competitive.
3. *RF kernel:* LHS sampling frequently provides favorable results. However, the differences between approaches remain small, indicating that several methods produce closely aligned results. The Kruskal–Wallis test at the

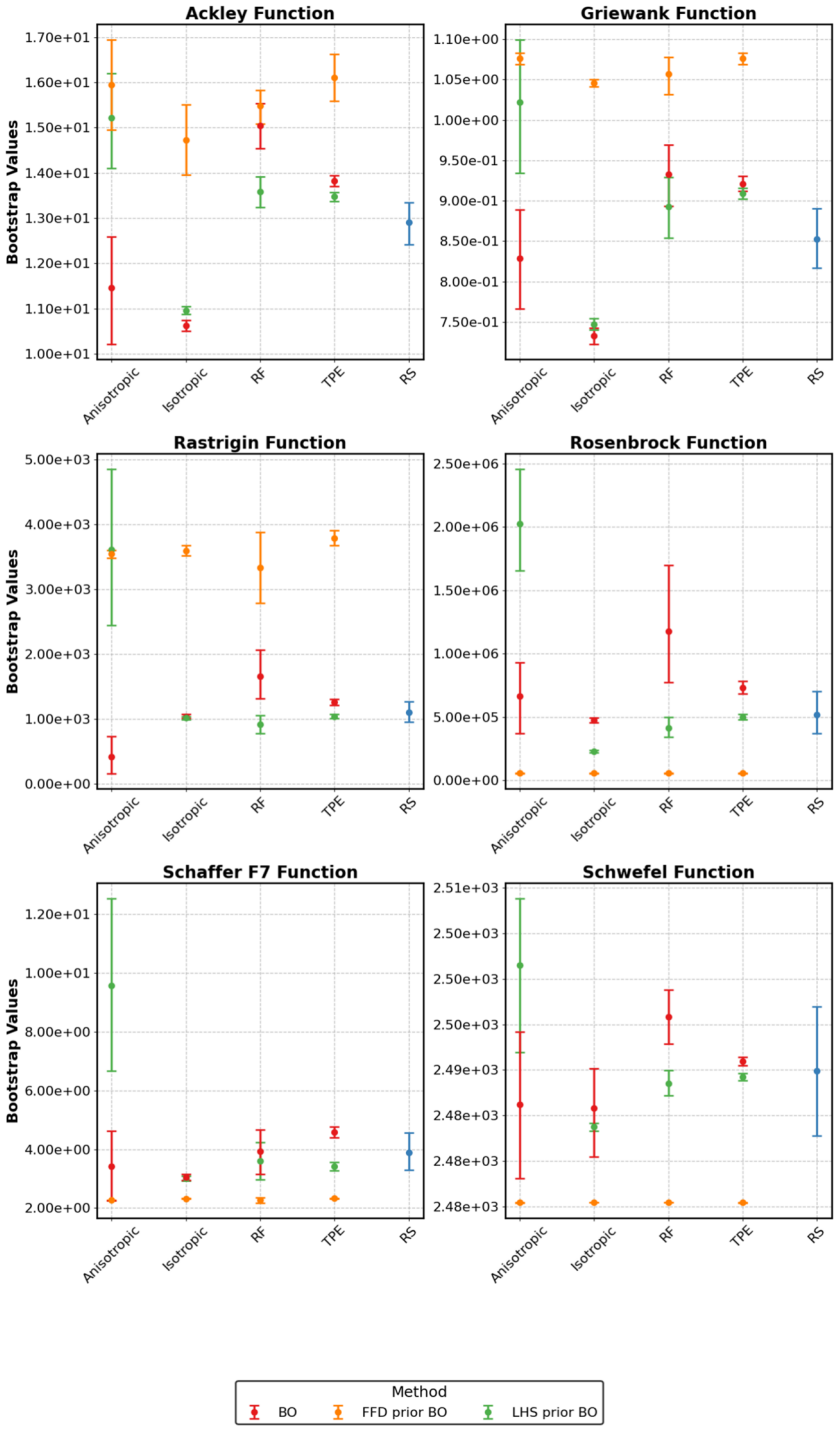


Fig. 6 Performance comparison at iteration 10

5% level shows a significant difference only for the Rastrigin function.

4. *TPE kernel:* for the Schaffer F7, Schwefel, and Rosenbrock function, configurations involving FFD or FFD combined with LHS priors tend to yield the best performance. However, for other functions, no clear or consistent pattern emerges at this point.

General observations at iteration 30

By the 30th iteration, as depicted in Fig. 8, the observed performance patterns exhibit increased clarity. For the Ackley, Rastrigin, Rosenbrock, Griewank and Schwefel function, anisotropic kernels without priors achieve lower bootstrap means. For the Ackley, Rastrigin, Rosenbrock, and Schwefel function, the Kruskal-Wallis test at the 5% level indicates

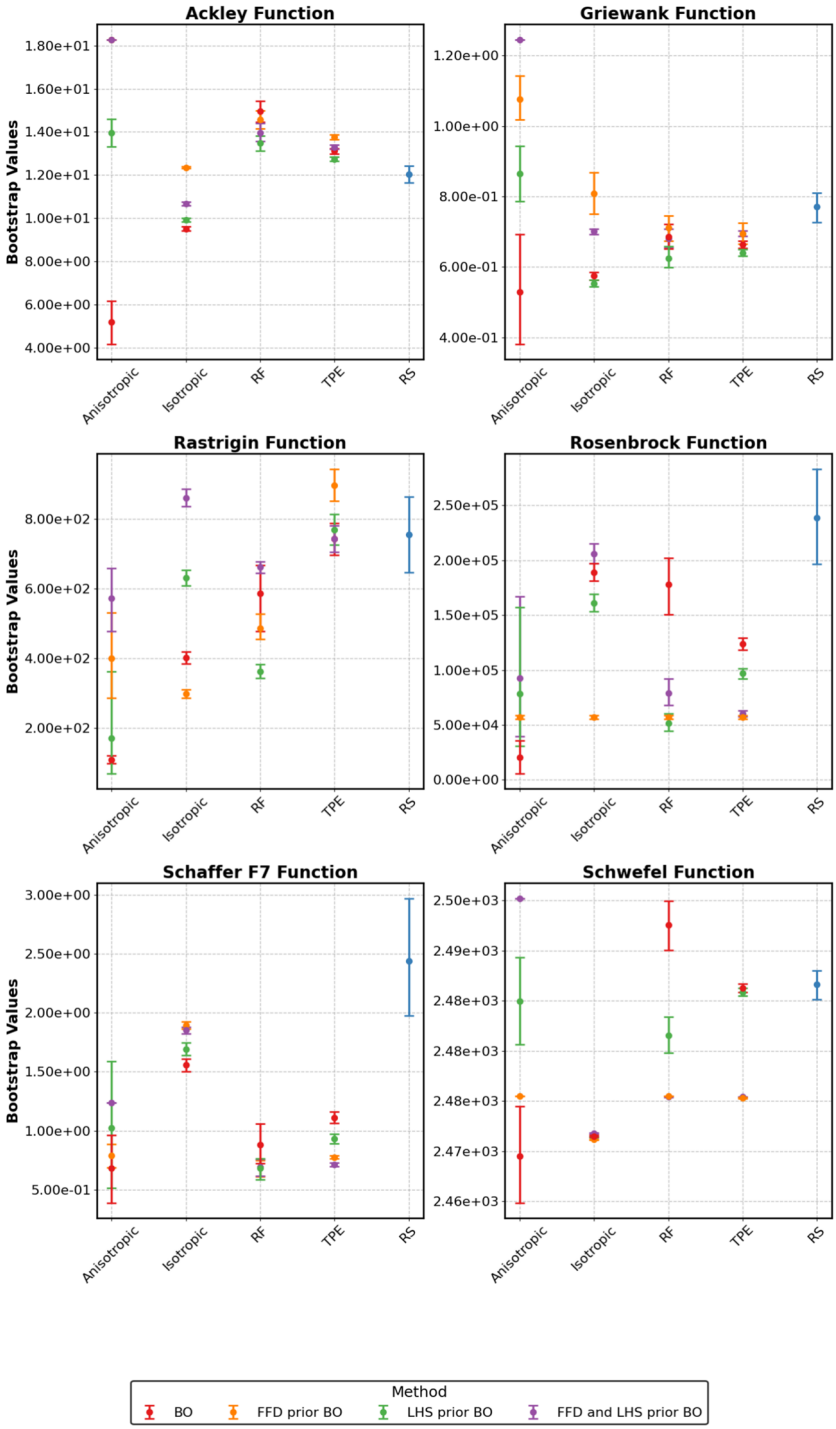


Fig. 7 Performance comparison at iteration 20

significant differences when comparing BO without priors (anisotropic kernel) to all other surrogate models. For the Schaffer F7 function, anisotropic kernels frequently appear among the top performers, although TPE and RF kernels match or closely approach these outcomes in specific configurations. Meanwhile, RS remains generally outperformed by multiple kernel-prior combinations, indicating a more pronounced gap at this iteration stage, suggesting limited improvement and confirming that more structured approaches (Kruskal–Wallis test at the level 5% indicates significant differences when comparing the best method of each surrogate model with RS).

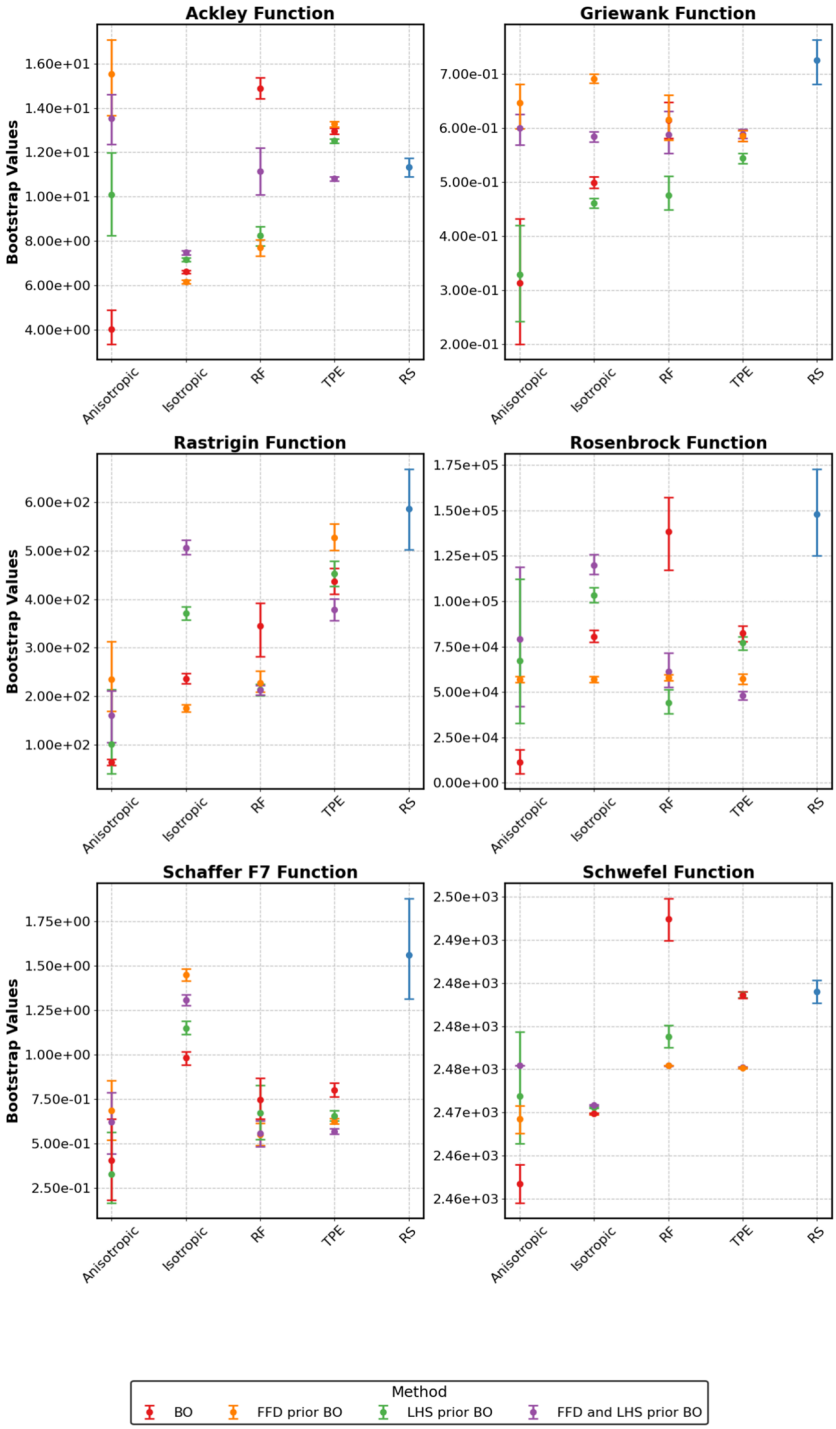


Fig. 8 Performance comparison at iteration 30

Kernel-specific observations at iteration 30

1. Anisotropic Kernel: Without priors or when combined with LHS priors, consistently yields lower bootstrap mean values across multiple test functions. However, it often exhibits relatively wide confidence intervals, underscoring that there is still residual variability at this stage. The Kruskal–Wallis test at the 5% significance level confirms significant differences for the Ackley, Rosenbrock, and Schwefel function when comparing BO without priors to the other methods.
2. Isotropic Kernel: Performance is function-dependent. Although at times no priors or LHS priors prove advantageous, certain functions still favor FFD priors. In general, isotropic kernels show moderate variability, resulting in mixed results across the set of test functions.

3. RF Kernel: LHS sampling continues to provide favorable results, although the differences among various prior configurations are less pronounced than at earlier iterations. However, this kernel generally maintains a position slightly behind the anisotropic configurations on most functions (for the Griewank and Rosenbrock function, the Kruskal–Wallis test at the 5% level indicates significant differences when comparing LHS prior to BO to the other methods).
4. TPE kernel: although anisotropic kernels often lead, TPE kernels achieve competitive results, especially for certain functions like Schaffer F7. For the TPE kernel, combined FFD and LHS priors can yield lower mean values for the Rastrigin, Schwefel, Schaffer F7, and Ackley function (for the Ackley, Rosenbrock, and Schaffer F7 the Kruskal–Wallis test at the level 5% indicates significant differences).

By the 100th iteration, as depicted in Fig. 9, the influence of the initial sampling diminishes across all kernels and functions. According to the Kruskal–Wallis test at the significance level of 5%, significant differences remain only for the Schwefel function (when using RF and TPE kernels) and for the Rosenbrock and Griewank function (when using the isotropic kernel). Across all test functions, anisotropic kernels continue to produce notably low bootstrap mean values. In contrast, RS remains markedly outperformed, underscoring the advantages of BO methods at this advanced stage of the optimization process.

Validation

3D printing

Figure 10 presents a comparative analysis of energy consumption in different optimization strategies for 3D printing processes. The strategies include BO, LHS followed by BO, FFD followed by BO, and a combination of LHS followed by FFD and then BO. Each subplot provides a detailed view of energy consumption trends over 20 function evaluations for the respective method.

In the upper-left subplot, the trend of energy consumption is displayed for the sole application of BO. The y axis, which ranges from 5 to 40 Wh, indicates the energy used, while the x-axis represents the function evaluation number. The minimum energy consumption, 8.63 Wh, is observed in the 20th function evaluation, marked by a red point and an annotation.

The top right subplot illustrates the energy consumption when using LHS followed by BO. The initial LHS phase concludes at the 8th function evaluation, as indicated by the green dashed line. Subsequent function evaluations reflect the optimization phase through BO. The minimum energy consumption, 8.88 Wh, occurs at the 10th function evaluation.

The bottom left subplot shows the energy consumption trends for FFD followed by BO. The transition from the FFD phase to the BO phase is marked by a blue dashed line at the 8th function evaluation. This approach demonstrates the most significant reduction in energy consumption, with the minimum value recorded at 5.82 Wh during the first function evaluation, highlighting the initial effectiveness of the FFD strategy.

The bottom right subplot presents the energy consumption for a sequential strategy combining LHS, FFD, and BO. The initial FFD phase ends at the 8th function evaluation (blue dashed line), followed by the LHS phase concluding at the 16th function evaluation (green dashed line). This approach identifies the same minimum energy consumption of 5.82 Wh in the first function evaluation.

We conducted five additional tests under the same conditions to verify the results against noise. The findings were 6.02 ± 0.27 Wh, indicating that noise is not a major issue in this research.

The mean energy usage for all data points is 17.89 Wh, accompanied by a standard deviation of 7.12 Wh. The lowest observed energy usage is 5.82 Wh. This signifies a notable decrease, since the minimum value is about 67.45% less than the mean.

When optimizing manufacturing process parameters for energy efficiency, it is critical to evaluate whether the long-term savings justify the initial efforts and energy expenditures during the optimization phase. The trial runs required for parameter refinement often lead to additional energy consumption before the optimal state is achieved. To quantify when these initial costs are compensated, we have calculated the break-even point (BEP), as shown in Fig. 11.

In an extreme scenario, it is assumed that all parts produced during the optimization phase are discarded. This assumption contributes to a higher initial energy overhead, effectively increasing the y-intercept of the energy consumption curve for the optimization scenario. Although the start-up cost (in terms of energy) is elevated, the energy consumption rate after reaching the optimized state decreases, resulting in a lower slope for the optimized curve once the process reaches steady-state conditions. Thus, after the BEP, the incremental energy consumption per part is significantly lower than it would have been if standard conditions had been maintained. In our case, the analysis indicates that the BEP occurs after approximately 30 parts are produced using the optimized parameters. Beyond this production volume, the cumulative energy saved by operating under optimized conditions exceeds the excess energy consumed during the optimization process. Consequently, once the BEP is reached, continued production in the optimized settings

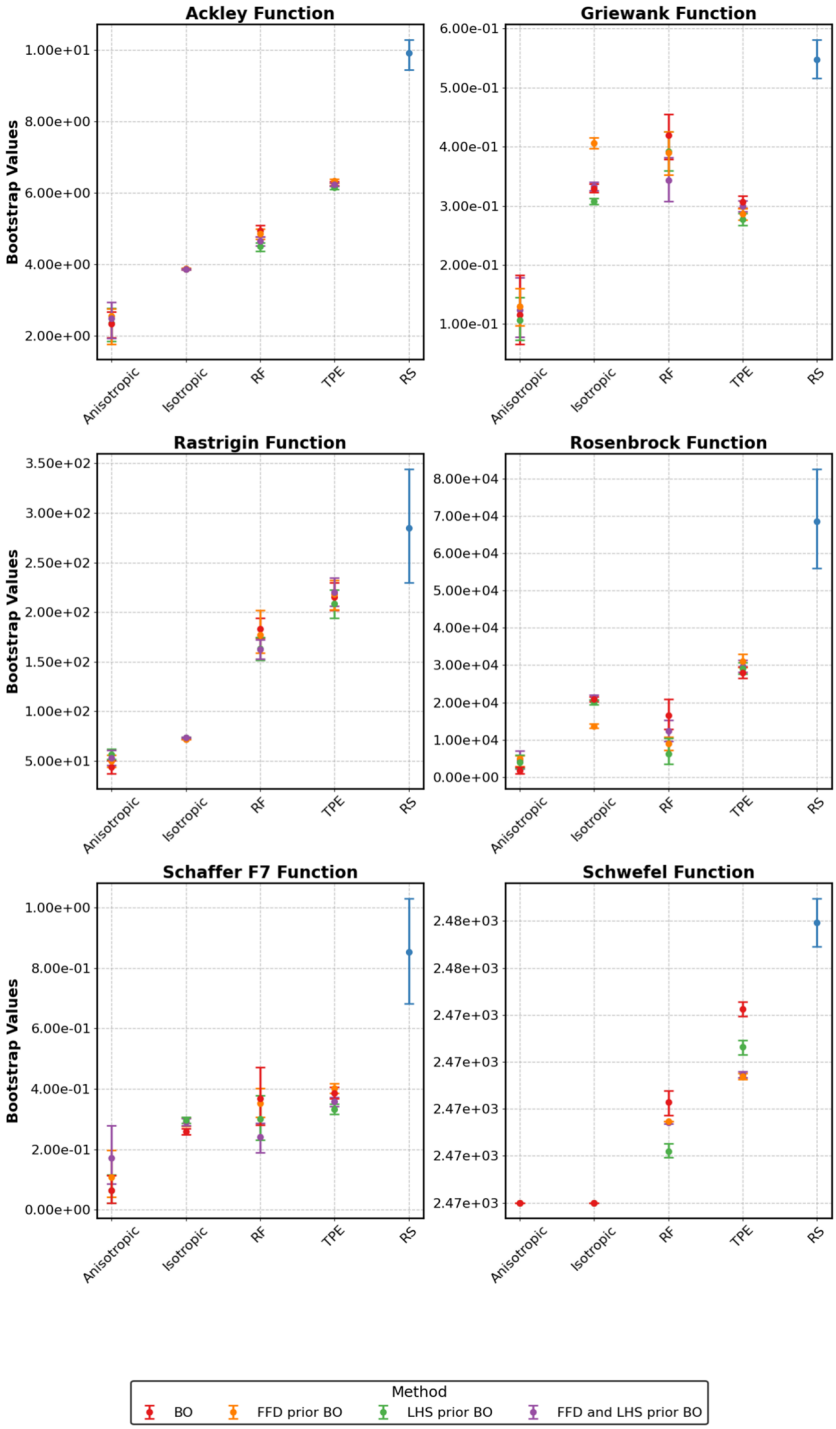


Fig. 9 Performance comparison at iteration 100

results in a net reduction in energy consumption compared to maintaining the standard parameter conditions. Furthermore, the time required to achieve these optimized parameters must also be considered. Iterative testing and data analysis needed to refine process parameters can introduce machine downtime and extend lead times. Although these additional time investments do not directly factor into the calculation of energy consumption, they do influence the practical feasibility and economic rationale for pursuing parameter optimization.

Moreover, visual inspection was carried out on all printed components, and no alterations in quality were observed. This indicates that the energy-optimized process does not compromise the quality of the product in our case study and for our specific product.

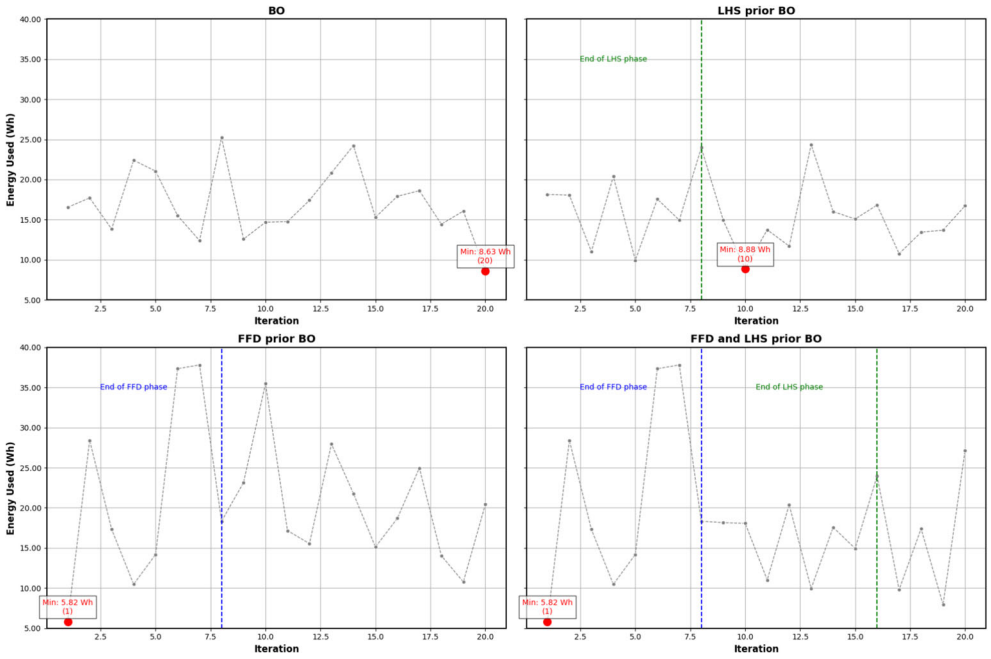
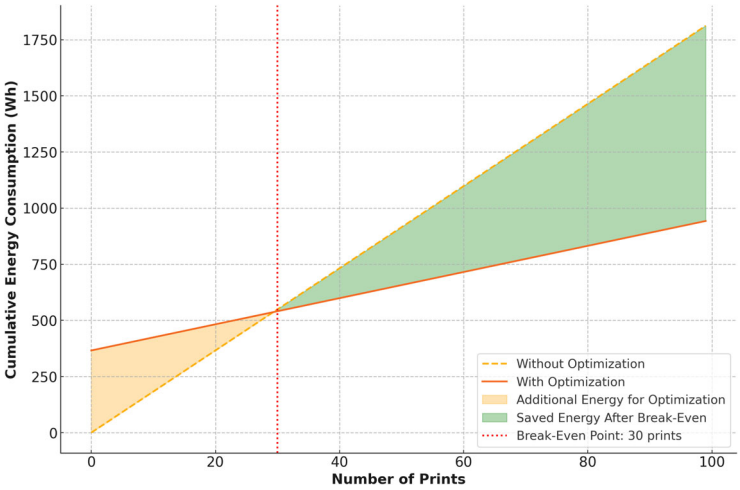


Fig. 10 Results case study

Fig. 11 BEP when all parts for optimization are discarded



To further analyze the impact of different input parameters on energy consumption, we performed a clustering analysis based on Principal Component Analysis (PCA). This approach allows us to identify patterns and groupings within the data that are not immediately apparent from the function evaluation plots alone. By reducing the dimensionality of the data and highlighting key contributing factors, PCA provides a deeper understanding of how various parameters influence energy consumption.

Figure 12 demonstrates the results of the k-means clustering of the energy consumption data based on PCA.

In order to determine the optimal number of clusters, we refer to the number of clusters at the first non-linear inflection point in the plot of within-cluster sums of squares ranked by size (Liu et al., 2022). Thus, we conclude that there are three clusters.

The clusters are visually represented on a two-dimensional plane defined by the first two principal components (PC1 and PC2). Principal components are linear combinations of the original features, and their loadings indicate the contribution of each feature to the components. Mathematically, the principal components are defined as follows.

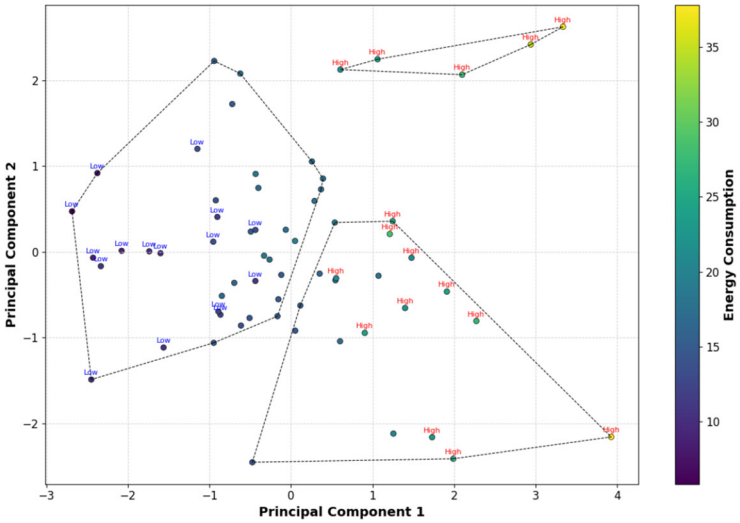


Fig. 12 PCA Clustering of energy consumption data with highlighted clusters

$$\begin{aligned}\text{PC1} = {} & 0.369 \times \text{Bed Temperature} \\ & + 0.693 \times \text{Energy Used} - 0.611 \times \text{Layer Height} \\ & - 0.061 \times \text{Print Speed} - 0.076 \\ & \times \text{Infill Mass Proportion} \\ & + 0.030 \times \text{Nozzle Temperature}\end{aligned}$$

$$\begin{aligned}\text{PC2} = {} & 0.473 \times \text{Infill Mass Proportion} - 0.493 \\ & \times \text{Nozzle Temperature} + 0.689 \times \text{Print Speed} \\ & + 0.199 \times \text{Bed Temperature} + 0.104 \times \text{Energy Used} \\ & + 0.087 \times \text{Layer Height}\end{aligned}$$

PC1 is primarily influenced by the energy used and the height of the layer, with positive contributions from the temperature of the bed and negative contributions from the height of the layer. PC2 is mainly driven by the proportion of infill mass and the printing speed, with negative contributions from the nozzle temperature.

Each point in the plot corresponds to a specific set of parameter values and its respective energy consumption, which is indicated by the color scale on the right. The clusters are enclosed by dashed convex hulls, and significant points within the clusters are annotated to highlight regions of high and low energy consumption.

Cluster 1, represented by moderate to high PC1 and low PC2 values, is likely to exhibit higher energy consumption due to its high PC1 values. The key features associated with this cluster include higher bed temperature, a lower layer height, moderate to high energy usage, a lower proportion of infill mass, and a faster print speed, along with a higher nozzle temperature. This is visually confirmed, as points within this cluster are predominantly marked as "High" in red, indicating higher energy consumption.

Cluster 2, characterized by low PC1 and low to high PC2 values, is likely to exhibit lower energy consumption due to its low PC1 values. The key features associated with this cluster include lower bed temperature, higher layer height, lower energy usage, higher infill mass proportion, and print speed, along with a lower nozzle temperature. This is visually confirmed, as points within this cluster are predominantly marked as "Low" in blue, indicating lower energy consumption.

Cluster 3, with high PC1 and PC2 values, is also likely to exhibit high energy consumption. The key features associated with this group include balanced values of bed temperature, layer height, infill mass proportion, print speed, and nozzle temperature. This cluster displays a mix of high and low energy consumption points, indicating a balance of parameter values.

Material science

The bootstrap analysis results for the five datasets are visualized in Fig. 13, comparing the performance of four sampling strategies integrated with the BO framework: LHS prior BO, FFD prior BO, FFD and LHS prior BO, and BO (baseline).

The methods are evaluated across isotropic, TPE, RF, and anisotropic kernels, with the bootstrap values plotted alongside their respective confidence intervals.

By evaluating these optimization approaches on a range of materials science datasets, we observe a more heterogeneous performance landscape than that seen with synthetic benchmark functions.

For the AgNP dataset, across all kernels, FFD with or without LHS priors consistently achieve the lowest bootstrap mean values. At the 5% significance level, the Kruskal–Wallis test reveals significant differences for each method when compared to FFD prior BO or FFD and LHS prior BO, with the exception of the RF Kernel and LHS prior BO.

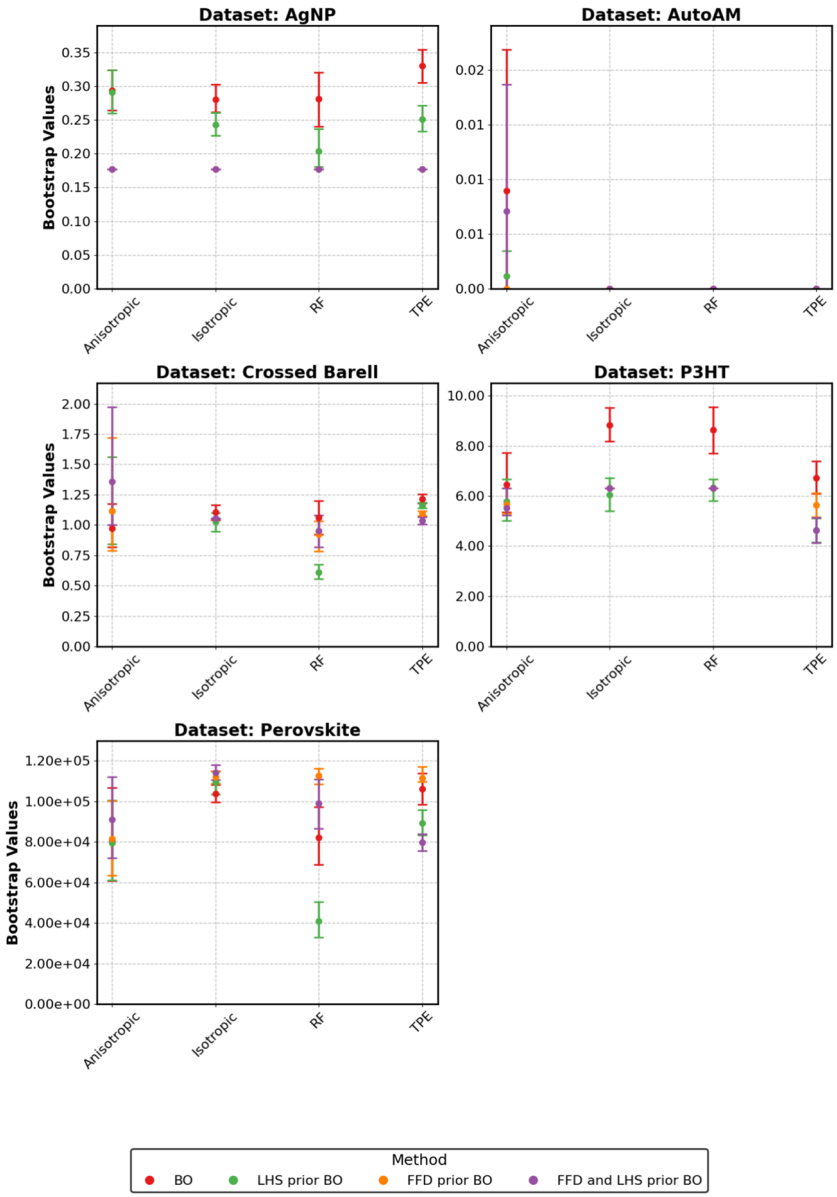


Fig. 13 Performance comparison at iteration 20 of Material Science Datasets

The AutoAM dataset exhibited a unique scenario in which the RF, TPE, and isotropic kernels achieved perfect minimization results across all sampling methods. In contrast, the anisotropic kernel did not fully converge. However, at the 5% significance level, the Kruskal–Wallis test did not show any significant differences.

For the Crossed Barell dataset, the most effective approach was LHS prior BO when paired with the RF kernel. The TPE kernel favored the combination of FFD and LHS prior. For both isotropic and anisotropic kernels, differences across sampling methods were negligible. At the 5% significance level, the Kruskal-Wallis test confirms a significant difference when comparing the LHS prior BO with the RF kernel against all other methods.

Within the P3HT dataset, the TPE kernel, when combined with FFD and LHS priors, yields the most favorable results. In general, the variance among the different kernels is minimal. Nevertheless, BO without priors often underperforms in comparison to methodologies that incorporate sampling strategies. At the 5% significance level, the Kruskal-Wallis test does not reveal significant differences for the anisotropic kernel. For the isotropic and RF kernel, LHS prior BO, FFD prior BO and FFD and LHS prior BO all show significant differences when compared to BO without priors. For the TPE kernel only FFD and LHS prior BO show significant differences when compared to BO without priors.

The Perovskite dataset exhibited a pattern similar to that of the Crossed Barrel dataset. In this case, the LHS prior BO approach coupled with the RF kernel proved to be the most

effective. Additionally, the TPE kernel showed a preference for the FFD and LHS prior combination. Differences across sampling methods for both isotropic and anisotropic kernels were minimal. The Kruskal–Wallis test confirmed a significant difference at the 5% significance level when comparing the LHS prior BO with the RF kernel to all other methods.

Dunn's test was executed for each notable difference across all datasets. Nonetheless, it revealed no changes in the trends of the bootstrap values.

Discussion

The results highlight the significant impact of initial sampling strategies on the performance of BO. Techniques such as FFD and LHS provide structured initial points that alter the surrogate model's ability to approximate the optimization landscape, particularly in the early stages. Our statistical analysis, based on the Kruskal-Wallis and subsequent Dunn's tests, confirms that these sampling strategies not only yield statistically significant differences among methods but also translate into meaningful performance improvements in practical terms.

Performance differences across kernels for mathematical functions

Across the benchmark functions, our nonparametric analysis consistently revealed statistically significant differences among the optimization methods at various iteration checkpoints. The Kruskal-Wallis test—used due to the non-normality of the residuals–consistently indicated overall differences ($p < 0.05$) in median performance across the five methods. Subsequent Dunn's pairwise comparisons provided further granularity, demonstrating that the structured sampling methods (FFD and LHS) often outperform both random search (RS) and baseline BO without priors in early iterations.

For example, at iteration 10, Dunn's test showed that FFD prior BO achieved significantly lower median values for functions with pronounced boundary effects such as Schaffer F7, Schwefel, and Rosenbrock ($p < 0.01$). This suggests that the strategic edge exploration offered by FFD can quickly guide the optimizer toward promising regions. In contrast, for functions like Ackley and Griewank, where internal exploration is critical, the differences between structured sampling and non-prior methods were less pronounced, underscoring that the effectiveness of a sampling method is strongly tied to the specific nature of the optimization landscape.

The anisotropic kernel, with its ability to assign unique length scales to individual dimensions, consistently excelled in multimodal settings. For functions such as Ackley and Rastrigin, even in the absence of structured priors, the anisotropic kernel demonstrated strong convergence. However, wider confidence intervals indicate sensitivity to function-specific complexities, implying that while these kernels are flexible, their performance can be influenced by overfitting in high-dimensional scenarios. In contrast, isotropic kernels benefited markedly from LHS initialization, especially in regular multimodal functions, since the even distribution of starting points mitigated their inherent limitation of relying on a single uniform length scale.

TPE kernels, when combined with FFD prior sampling, leveraged the diversity of the initial sampling to efficiently identify promising regions, particularly for complex multimodal functions such as Schaffer F7. The partitioning of the search space into "good" and "bad" configurations in TPE synergized well with structured sampling, driving superior convergence by iteration 20. However, TPE performance showed a higher dependency on effective priors, as its robustness diminished when prior information was noisy or sparse.

RF kernels, due to their nonparametric nature, were less assumption bound and therefore versatile. LHS initialization significantly improved their exploratory capacity in functions with many local minima, such as Schwefel. However, the slower convergence rates observed for RF—compared to anisotropic kernels in functions such as Rosenbrock and Schaffer F7—highlight the need to compensate for RF's limitations through robust sampling techniques.

Importantly, our analyses indicate that the performance discrepancies between sampling strategies are more pronounced during early iterations. With increasing BO steps, the differences diminish, suggesting that, while structured sampling is crucial when evaluation budgets are limited, the choice of surrogate model ultimately plays a more dominant role in later iterations.

Performance differences for real-world applications

Our real-world application of BO to optimize energy consumption during 3D printing clearly demonstrates the importance of structured initial sampling. In this study, we optimized parameters such as print speed, bed temperature, layer height, and infill mass proportion to reduce energy usage while maintaining print quality. Among the approaches evaluated, those that used FFD prior—alone or in combination with LHS—resulted in the largest reductions in energy consumption.

The advantage of FFD lies in its structured approach to initial sampling. By strategically probing combinations at critical factor levels, FFD captures essential interactions between parameters early on. This comprehensive initial coverage can, through meticulous design or favorable chance, identify a near-optimal parameter setting. In such cases, the subsequent BO phase has little room for further improvement

because the promising region has already been effectively mapped out.

Furthermore, PCA clustering confirms that, among the 3D printing parameters, bed temperature and layer height are the most critical influencers of energy consumption. Specifically, high bed temperatures and low layer heights consistently correlate with higher energy usage, while more balanced settings lead to moderate consumption.

The application of BO on five diverse materials science datasets provided additional insight into the effectiveness of different sampling strategies and kernel choices. In particular, no single sampling strategy consistently outperformed the others across all datasets. For the AgNP dataset, FFD prior BO significantly outperformed other methods regardless of the surrogate model used. The pronounced boundary effects in this dataset favored FFD, which efficiently explored edge cases where optimal solutions were likely to be found. In contrast, the AutoAM dataset, characterized by a simpler optimization landscape, did not show significant performance differences between sampling strategies and surrogate models, suggesting that in less complex search spaces, sophisticated sampling strategies may offer limited additional benefits.

For the P3HT dataset, the best performance was achieved using TPE combined with LHS and FFD sampling. The multimodal nature of this dataset probably benefited from TPE's ability to model multiple peaks effectively. Meanwhile, for the Perovskite and Crossed Barrel datasets, prior LHS BO paired with an RF kernel delivered the most favorable outcomes. The capability of the RF kernel to handle non-linear relationships and complex interactions probably contributed to its superior performance in these cases.

Overall, these findings indicate that while structured sampling strategies such as FFD and LHS can yield substantial benefits - especially in challenging, boundary-sensitive landscapes - the optimal combination of sampling method and surrogate model is highly dependent on the specific characteristics of the optimization problem at hand.

General implications and recommendations

The absence of a universally superior sampling strategy across all datasets and mathematical functions highlights the intricate relationship between the sampling method, the surrogate model, and the specific characteristics of the optimization landscape. Different kernels and sampling methods excel under varying conditions due to factors such as boundary effects, complexity of the search space, dimensionality, and data size.

The real-world applications highlight that BO tends to avoid placing multiple values simultaneously at the corners of the parameter space when limiting the function evaluation to 20. This characteristic of BO can be advantageous, as it avoids redundant exploration in areas less likely to contain the optimum, but it also highlights a limitation when the optimum is near the edges. In our study, mathematical functions often had a minimum near (0, ..., 0), typically towards the center of the space, but real-world examples showed that the optimum could be at an edge, where pure BO might struggle to explore effectively.

Optimization landscapes with significant boundary phenomena, which can be found in the AgNP dataset or in the Schaffer F7, Schwefel and Rosenbrock function, benefit from sampling strategies such as FFD that emphasize edge exploration.

Dimensionality also plays a key role in shaping optimization outcomes. As the dimensionality of the search space increases, the effectiveness of structured sampling methods such as FFD may diminish due to the exponential growth in the number of sample points required to maintain coverage. In such cases, strategies such as LHS, which offer a balance between uniformity and diversity, may become more favorable.

Lastly, the findings highlight the iterative nature of BO. Early iterations benefit significantly from structured sampling, which establishes a foundation for the surrogate model. As iterations progress, the choice of the surrogate model becomes a prominent choice, particularly when capturing fine-grained features of the optimization landscape.

In summary, our findings recommend a tailored approach:

- For early stage optimization in constrained environments, structured sampling is statistically proven to alter convergence and possibly achieve lower objective values.
- As the optimization process continues, the influence of the initial sampling strategy decreases and the focus should shift toward selecting a surrogate model that best captures the fine-grained features of the landscape.
- The interplay between sampling methods and kernel choices is critical.

Limitations and future work

One limitation of this study is the fixed choice of the combination of the surrogate model and the acquisition function in the BO process. Although the initialization phase was designed to identify the most effective pairing of the surrogate model and the acquisition function for all functions, it is possible that a more granular combination could have provided additional benefits. For instance, different acquisition functions may perform best for specific functions, which our fixed pairing approach may not fully capture. However, the general differences in loss between acquisition functions were not substantial, suggesting that the impact of this limitation may not be significant in practice. This fixed pairing approach was

necessary because expanding the scope to include all possible combinations would have made the derivation of key insights much more challenging. With four surrogate models and four sampling strategies, a broader scope would have required an exponential increase in pairwise comparisons to analyze the interplay between methods. Such a scenario would have significantly complicated the identification of meaningful trends and conclusions. However, future studies that confirm this assumption would be beneficial in confirming our findings and ensuring the robustness of our approach.

Furthermore, this study focused on single-objective optimization in different datasets, prioritizing simplicity and clarity in evaluating the effectiveness of various methods. Although this approach allowed for a streamlined comparison of sampling strategies and surrogate models, it does not fully address the complexities of practical scenarios where multiple objectives often need to be optimized simultaneously. For instance, in real-world cases, objectives such as minimizing energy consumption and maximizing product quality frequently conflict, requiring trade-offs that cannot be captured in a single-objective framework. However, starting with single-objective optimization provides a strong foundation for understanding and benchmarking methods. It allows for a controlled exploration of the parameter space and helps establish baseline insights into the effectiveness of the techniques. Extending this work to multi-objective optimization (MOO) is a logical next step but introduces additional challenges. Modeling multiple objectives and their interactions is inherently more complex and less streamlined than single-objective cases. For example, the Pareto front, a set of optimal trade-off solutions, can be difficult to compute and interpret, particularly when additional variables and constraints are introduced.

In our case study, we assumed that the ranges of selected parameters, spanning the print temperature, the print speed, the build plate temperature, the layer height, and the infill mass proportion, are adequately representative of the typical conditions encountered in commercial and research-oriented PLA printing setups. Although these ranges may not capture the full breadth of industrial-scale systems, diverse material formulations, or specialized geometries, they align with common usage scenarios and current practical guidelines. While this limited parameter space may restrict the straightforward application of our findings to all possible 3D printing scenarios, it guarantees that our analysis stays relevant to a broad category of applications using comparable equipment and materials. We also acknowledge that the experimental measurements are influenced by inherent noise and potential systematic errors. The energy consumption data, collected via a commercially available power meter, may be subject to minor calibration deviations, ambient environmental fluctuations, and filament variability. Moreover, the trapezoidal numerical approximation used to compute total energy consumption from discrete power measurements assumes relatively smooth changes between recorded data points. This simplification may overlook small, transient variations in power draw. However, given the scale of our experiments and the consistency in measurement methods, these noise sources are likely to exert only a limited influence on the relative comparisons between parameter settings. The key trends in energy usage should remain robust, as each experimental configuration was tested under similar conditions and instrumentation, ensuring that any measurement biases or noise are at least partially offset by relative consistency in the trials.

An important consideration in optimization is the challenge posed by high-dimensional parameter spaces. High-dimensional spaces are notoriously difficult for optimization algorithms to navigate because of the curse of dimensionality, which describes the exponential increase in the volume of the search space with the addition of each dimension. Although the benchmark functions used in this study were limited to six dimensions and the case studies focused on 3–5 dimensions, practical real-world problems can involve a much greater number of dimensions and intricate constraints. This issue is particularly pronounced in industrial processes and machine learning applications, where optimization of dozens or even hundreds of parameters is common. The sheer size and complexity of high-dimensional spaces can dramatically reduce optimization efficiency, making it challenging to adequately explore the parameter space and increasing the risk of converging to suboptimal solutions. Furthermore, the interactions among parameters in high-dimensional settings are often intricate and nonlinear, which can exacerbate these difficulties. Subsequent research ought to concentrate on formulating and incorporating approaches that address these issues, employing structured sampling techniques specifically designed for this problem. Another avenue for future work is to scale up the parts printed to include variability in the parameter spaces since it is likely that the optimal parameter space changes dynamically when increasing or decreasing the dimension of the part.

Conclusion

This study presented a comprehensive evaluation of initial sampling techniques within the BO framework, focusing on their effectiveness in optimizing both synthetic benchmark functions and practical applications. The investigation revealed that, while BO is inherently robust and effective in a wide range of optimization problems, the integration of structured initial sampling methods, such as LHS and FFD, can significantly alter its performance. The study demonstrated that initial sampling strategies can significantly influence the performance of the subsequent BO process. Although these strategies can improve the effectiveness of BO in some cases,

they can also hinder it in others. This variability highlights the importance of carefully selecting and adapting the sampling method to the specific situation at hand. It also underscores that neglecting the choice of sampling strategy can lead to suboptimal overall performance.

The practical application of these methods in optimizing the energy consumption of 3D printing processes demonstrated their relevance in the real world. By systematically exploring different optimization strategies, the study highlighted how LHS and FFD, followed by BO, can lead to substantial reductions in energy consumption. This not only underscores the practical utility of these methods in manufacturing, but also points to their potential in promoting sustainable practices by minimizing resource usage. Furthermore, the application of the methods to materials science datasets demonstrated that even with just 20 optimization iterations, convergence to values very close to the minimum was achieved.

However, the study also identified limitations and areas for future research. The observed performance variability across different design spaces highlights the importance of carefully tailoring the choice of initial sampling techniques and BO configurations-such as kernel type, acquisition function, and sampling strategy -to the specific problem being addressed. This underscores the need for further investigation into adaptive or problem-specific kernel selection strategies to enhance the flexibility and effectiveness of BO methods.

Additionally, while the focus on single-objective optimization offers valuable and clear insights, it does not fully address the complexities of many real-world problems where multiple conflicting objectives must be considered simultaneously. Future research should prioritize the development and application of multi-objective optimization frameworks capable of balancing factors such as quality, cost, and performance within a single cohesive model. Expanding the scope of testing to include more diverse and higher-dimensional parameter spaces, along with conducting extensive case studies across a broader range of domains, will be essential to validate and refine the generalizability and robustness of the findings presented in this study.

In conclusion, this study contributes to the growing body of knowledge on BO by demonstrating the value of initial sampling techniques in enhancing the effectiveness of BO. The findings have significant implications for both theoretical research and practical applications, offering a roadmap for future studies to build upon in the pursuit of more efficient and effective optimization strategies in complex, real-world scenarios.

Author contributions Conceptualization, L.G.; writing—original draft preparation, L.G. and N.H.; writing—review and editing, L.G., N.H., A.K., S.K. and A.M.; validation, A.M.; visualization, N.H. and A.K.; supervision, J.O.; All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The research is funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) in research Project (No. 03LB2041E).

Availability of data and materials Data will be made available on request.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article. The authors have non-financial interests to disclose.

Ethical approval Not applicable.

References

Ackley, D. H. (1987). *A connectionist machine for genetic Hillclimbing*. Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4613-1997-9

Becker, J. M. J., Borst, J., & Veen, A. V. D. (2015). Improving the overall equipment effectiveness in high-mix-low-volume manufacturing environments. *Cirp Annals-Manufacturing Technology, 64*, 419–422. https://doi.org/10.1016/J.CIRP.2015.04.126

Camposeco-Negrete, C. (2020). Optimization of printing parameters in fused deposition modeling for improving part quality and process sustainability. *The International Journal of Advanced Manufacturing Technology, 108*(7), 2131–2147. https://doi.org/10.1007/s00170-020-05555-9

Cashore, J., Vishwanath, A., & Leeds, M. (2016). Multi-step lookahead Bayesian optimization for adaptive data sampling. In *Proceedings of the workshop on data-efficient machine learning at the 33rd international conference on machine learning (ICML 2016)*, New York City. https://sites.google.com/site/dataefficientml/.

Chia, H., Wu, J., Wang, X., & Yan, W. (2022). Process parameter optimization of metal additive manufacturing: A review and outlook. *Journal of Materials Informatics, 2*, 18. https://doi.org/10.20517/jmi.2022.18

Daulton, S., Eriksson, D., Balandat, M., & Bakshy, E. (2022). Multi-objective Bayesian optimization over high-dimensional search spaces. In Cussens, J., & Zhang, K. (eds.) *Proceedings of the 38th conference on uncertainty in artificial intelligence. Proceedings of machine learning research*, (Vol. 180, pp. 507–517). PMLR. https://proceedings.mlr.press/v180/daulton22a.html.

De Jong, K. A. (1975). *An analysis of the behavior of a class of genetic adaptive systems* (Ph.D thesis, University of Michigan). https://doi.org/10.7302/10966.

Deneault, J. R., Chang, J., Myung, J., Hooper, D., Armstrong, A., Pitt, M., & Maruyama, B. (2021). Toward autonomous additive manufacturing: Bayesian optimization on a 3d printer. *MRS Bulletin, 46*, 566–575. https://doi.org/10.1557/s43577-021-00051-1

Deutsch, J., & Deutsch, C. (2012). Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning and Inference, 142*, 763–772. https://doi.org/10.1016/J.JSPI.2011.09.016

Dieterich, J. M., & Hartke, B. (2012). Empirical review of standard benchmark functions using evolutionary global optimization. *Applied Mathematics, 3*, 1552–1564. https://doi.org/10.4236/am.2012.330215

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal, 15*(1), 292–300. https://doi.org/10.1177/1536867X1501500117

Frazier, P. I. (2018). Bayesian optimization. *Recent Advances in Optimization and Modeling of Contemporary Problems*. https://doi.org/10.1287/educ.2018.0188

Frazier, P. I., & Wang, J. (2016). Bayesian optimization for materials design. In T. Lookman, F. Alexander, & K. Rajan (Eds.), *Information science for materials discovery and design* (pp. 45–75). Cham: Springer. https://doi.org/10.1007/978-3-319-23871-5_3

Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian optimization with unknown constraints. In: *Proceedings of the 30th conference on uncertainty in artificial intelligence (UAI 2014)* (pp. 250–259). AUAI Press. https://doi.org/10.5555/3020751.3020778.

Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., Morgan, E. F., & Brown, K. A. (2020). A Bayesian experimental autonomous researcher for mechanical design. *Science Advances, 6*(15), 1708. https://doi.org/10.1126/sciadv.aaz1708

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

Griewank, A. O. (1981). Generalized descent for global optimization. *Journal of Optimization Theory and Applications, 34*(1), 11–39. https://doi.org/10.1007/BF00933356

Guidetti, X., Rupenyan, A., Fassl, L., Nabavi, M., & Lygeros, J. (2022). Advanced manufacturing configuration by sample-efficient batch Bayesian optimization. *IEEE Robotics and Automation Letters, 7*(4), 11886–11893. https://doi.org/10.1109/LRA.2022.3208370

Huan, X., & Marzouk, Y. (2011). Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics, 232*, 288–317. https://doi.org/10.1016/j.jcp.2012.08.013

Huan, X., & Marzouk, Y. (2011). Simulation-based optimal Bayesian experimental design for nonlinear systems. *ournal of Computational Physics, 232*, 288–317. https://doi.org/10.1016/j.jcp.2012.08.013

Iordache, M., Costea, A., & Malea, C. (2019). Methods for modernization of processing equipment. *IOP Conference Series: Materials Science and Engineering, 564*, 012074. https://doi.org/10.1088/1757-899X/564/1/012074

Ismail, I., & Halim, A. H. (2017). Comparative study of meta-heuristics optimization algorithm using benchmark function. *International Journal of Electrical and Computer Engineering (IJECE), 7*(3), 1643–1650. https://doi.org/10.11591/ijece.v7i2.pp1103-1109

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization, 13*(4), 455–492. https://doi.org/10.1023/A:1008306431147

Joseph, V. R. (2012). Bayesian computation using design of experiments-based interpolation technique. *Technometrics, 54*(3), 209–225. https://doi.org/10.1080/00401706.2012.680399

Knowles, J. (2006). ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation, 10*(1), 50–66. https://doi.org/10.1109/TEVC.2005.851274

Kucherenko, S., Albrecht, D., & Saltelli, A. (2015). Exploring multi-dimensional spaces: a comparison of latin hypercube and quasi monte carlo sampling techniques. arXiv preprint arXiv:1505.02350.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering, 86*(1), 97–106. https://doi.org/10.1115/1.3653121

Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2017). Constrained Bayesian optimization with noisy experiments. arXiv:1706.07094. https://doi.org/10.1214/18-BA1110.

Li, H., Alkahtani, M. E., Basit, A. W., Elbadawi, M., & Gaisford, S. (2023). Optimizing environmental sustainability in pharmaceutical 3d printing through machine learning. *International Journal of Pharmaceutics, 648*, 123561. https://doi.org/10.1016/j.ijpharm.2023.123561

Liang, Q., Gongora, A., Ren, Z., Tiihonen, A., Liu, Z., Sun, S., Deneault, J., Bash, D., Mekki-Berrada, F., Khan, S., et al. (2021). Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *NPJ Computational Materials, 7*(1), 188. https://doi.org/10.1038/s41524-021-00656-9

Liu, Z., Rolston, N., Flick, A. C., Colburn, T. W., Ren, Z., Dauskardt, R. H., & Buonassisi, T. (2022). Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule, 6*(4), 834–849. https://doi.org/10.1016/j.joule.2022.03.003

Liu, T., Shryane, N., & Elliot, M. (2022). Attitudes to climate change risk: Classification of and transitions in the UK population between 2012 and 2020. *Humanities and Social Sciences Communications, 9*(1), 1–15. https://doi.org/10.1057/s41599-022-01287-1

Martinez-Cantin, R. (2014). Bayesopt: A Bayesian optimization library for nonlinear optimization, experimental design and bandits. *Journal of Machine Learning Research, 15*(1), 3735–3739. https://doi.org/10.5555/2627435.2697061

Mekki-Berrada, F., Ren, Z., Huang, T., Wong, W. K., Zheng, F., Xie, J., Tian, I. P. S., Jayavelu, S., Mahfoud, Z., Bash, D., et al. (2021). Two-step machine learning enables optimized nanoparticle synthesis. *NPJ Computational Materials, 7*(1), 55. https://doi.org/10.1038/s41524-021-00520-w

Mockus, J. (1989). *Bayesian approach to global optimization: Theory and applications. Mathematics and its applications* (Vol. 37). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-0909-0.

Mockus, J., Tiesis, V., & Zilinskas, A. (1978). *The application of Bayesian methods for seeking the extremum*. North-Holland Publishing Company. https://www.researchgate.net/publication/248818761_The_application_of_Bayesian_methods_for_seeking_the_extremum.

Moeini, M., & Abokifa, A. (2024). Chlorine dosage management in drinking water systems: Comparing Bayesian optimization to evolutionary algorithms. *Journal of Hydroinformatics, 26*(11), 2720–2738. https://doi.org/10.2166/hydro.2024.090

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference. Quantitative applications in the social sciences* (Vol. 95). SAGE. https://uk.sagepub.com/en-gb/eur/book/bootstrapping.

Morishita, T., & Kaneko, H. (2022). Initial sample selection in Bayesian optimization for combinatorial optimization of chemical compounds. *ACS Omega, 8*, 2001–2009. https://doi.org/10.1021/acsomega.2c05145

Mühlenbein, H., Schomisch, M., & Born, J. (1991). The parallel genetic algorithm as function optimizer. *Parallel computing, 17*(6–7), 619–632. https://doi.org/10.1016/S0167-8191(05)80052-3

Navid, A., Khalilarya, S., & Abbasi, M. R. (2018). Diesel engine optimization with multi-objective performance characteristics by non-evolutionary Nelder–Mead algorithm: Sobol sequence and Latin hypercube sampling methods comparison in doe process. *Fuel, 226*, 97–110. https://doi.org/10.1016/j.fuel.2018.04.142

Neumann-Brosig, M., Schneider, S., Vargas, D. V., & Bäck, T. (2018). Data-efficient optimization using Bayesian optimization. In *International conference on parallel problem solving from nature* (pp. 123–134). Springer. https://doi.org/10.1007/978-3-319-99253-2_10.

Ostertagova, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal–Wallis test. *Applied Mechanics and Materials, 611*, 115–120. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Plöckinger, M., Kechagia, P. T., & Affenzeller, M. (2022). Bayesian optimization for black-box problems constrained by simulation time or available budget. *Engineering Optimization, 54*(3), 425–441. https://doi.org/10.1080/0305215X.2021.1922100

Rashidi, B., Johnstonbaugh, K., & Gao, C. (2024). Cylindrical Thompson sampling for high-dimensional Bayesian optimization. In Dasgupta, S., Mandt, S., & Li, Y. (Eds.), *Proceedings of The 27th international conference on artificial intelligence and statistics* (Vol. 238, pp. 3502–3510). PMLR. https://proceedings.mlr.press/v238/rashidi24a.html.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics,2*(1), 21–33. https://www.nrc.gov/docs/ML1714/ML17143A100.pdf.

Renardy, M., Joslyn, L. R., Millar, J. A., & Kirschner, D. (2021). To sobol or not to sobol? The effects of sampling schemes in systems biology applications. *Mathematical Biosciences*. https://doi.org/10.1016/j.mbs.2021.108593

Roy, M.-H., & Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics, 24*(4), 993–1006. https://doi.org/10.1080/10485252.2012.715161

Rubio, F., Llopis-Albert, C., & Valero, F. (2021). Multi-objective optimization of costs and energy efficiency associated with autonomous industrial processes for sustainable growth. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2021.121115

Sattari, K., Wu, Y., Chen, Z., Mahjoubnia, A., Su, C., & Lin, J. (2024). Physics constrained multi-objective Bayesian optimization to accelerate 3d printing of thermoplastics. *Additive Manufacturing*. https://doi.org/10.1016/j.addma.2024.104204

Schwefel, H.-P. (1981). *Numerical optimization of computer models*. Wiley. https://archive.org/details/numericaloptimiz0000schw.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE, 104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Shang, Y.-W., & Qiu, Y.-H. (2006). A note on the extended Rosenbrock function. *Evolutionary Computation, 14*(1), 119–126. https://doi.org/10.1162/106365606776022733

Sheikh, H. M., & Marcus, P. S. (2022). Bayesian optimization for mixed-variable, multi-objective problems. *Structural and Multidisciplinary Optimization, 65*(11), 331. https://doi.org/10.1162/106365600568202

Shields, B. J., Stevens, J. M., Li, J., Parasram, M., Damani, F. N., Alvarado, J. I. M., Janey, J., Adams, R. P., & Doyle, A. (2021). Bayesian reaction optimization as a tool for chemical synthesis. *Nature, 590*, 89–96. https://doi.org/10.1038/s41586-021-03213-y

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th international conference on machine learning (ICML-10), Haifa, Israel* (pp. 1015–1022). https://icml.cc/Conferences/2010/papers/422.pdf.

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics, 29*(2), 143–151. https://doi.org/10.1080/00401706.1987.10488205

Sun, S., Tiihonen, A., Oviedo, F., Liu, Z., Thapa, J., Zhao, Y., Hartono, N. T. P., Goyal, A., Heumueller, T., Batali, C., et al. (2021). A data fusion approach to optimize compositional stability of halide perovskites. *Matter, 4*(4), 1305–1322. https://doi.org/10.1016/j.matt.2021.01.008

Tsao, J., & Patel, M. H. (2013). An intuitive design pattern for sequentially estimating parameters of a 2k factorial experiment with active confounding avoidance and least treatment combinations. *Computers & Industrial Engineering, 66*, 601–613. https://doi.org/10.1016/j.cie.2013.08.005

Vangelatos, Z., Sheikh, H. M., Marcus, P. S., Grigoropoulos, C. P., Lopez, V. Z., Flamourakis, G., & Farsari, M. (2021). Strength through defects: A novel Bayesian approach for the optimization of architected materials. *Science Advances, 7*(41), 2218. https://doi.org/10.1126/sciadv.abk2218

Vořechovský, M., & Eliáš, J. (2020). Modification of the maximin and phi criteria to achieve statistically uniform distribution of sampling points. *Technometrics, 62*, 371–386. https://doi.org/10.1080/00401706.2019.1639550

Watanabe, S. (2023). Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint. https://arxiv.org/abs/2304.11127.

Xiong, F., Xiong, Y., Chen, W., & Yang, S. (2009). Optimizing Latin hypercube design for sequential sampling of computer experiments. *Engineering Optimization, 41*, 793–810. https://doi.org/10.1080/03052150902852999

Zhou, W., Yang, J., & Liu, M.-Q. (2020). Optimal maximin l2-distance Latin hypercube designs. *Journal of Statistical Planning and Inference*. https://doi.org/10.1016/j.jspi.2019.11.006

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software, 23*(4), 550–560. https://doi.org/10.1145/279232.279236