Computers and Chemical Engineering 194 (2025) 108983

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Multi-objective reaction optimization under uncertainties using expected quantile improvement

Jiyizhe Zhang [a,d,*,1], Daria Semochkina [b,1], Naoto Sugisawa [c], David C. Woods [b], Alexei A. Lapkin [a,d,e,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, CB3 0AS, UK
[b] School of Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK
[c] Department of Basic Medicinal Sciences, Graduate School of Pharmaceutical Sciences, Nagoya University, Nagoya, 464-8601, Japan
[d] Innovation Centre in Digital Molecular Technologies, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, UK
[e] Cambridge Centre for Advanced Research and Education in Singapore (CARES Ltd), #05-05 CREATE Tower, 1 Create Way, 138602, Singapore

ARTICLE INFO

Keywords:
Multi-objective Bayesian optimization
Reaction development
Heteroscedastic noise
Machine learning

ABSTRACT

Multi-objective Bayesian optimization (MOBO) has shown to be a promising tool for reaction development. However, noise is usually inevitable in experimental and chemical processes, and finding reliable solutions is challenging when the noise is unknown or significant. In this study, we focus on finding a set of optimal reaction conditions using multi-objective Euclidian expected quantile improvement (MO-E-EQI) under noisy settings. First, the performance of MO-E-EQI is evaluated by comparing with some recent MOBO algorithms *in silico* with linear and log-linear heteroscedastic noise structures and different magnitudes. It is noticed that high noise can degrade the performance of MOBO algorithms. MO-E-EQI shows robust performance in terms of hypervolume-based metric, coverage metric and number of solutions on the Pareto front. Finally, MO-E-EQI is implemented in a real case to optimize an esterification reaction to achieve the maximum space-time-yield and the minimal E-factor. The algorithm identifies a clear trade-off between the two objectives.

1. Introduction

Multi-objective Bayesian optimization (MOBO) is a powerful tool applied in multiple stages of chemical reaction development. This includes, for example, the discovery of multi-functional molecules (J.C. and Coley, 2023), identifying the best reaction conditions to achieve high yield with good selectivity (Wang et al., 2021), and processes design where improving the yield while considering the throughput or environmental impacts at the same time (Braconi, 2023; Slattery et al., 2024). In the presence of multiple objectives that need to be optimized simultaneously, MOBO can find a set of solutions that represent a trade-off among the objectives, known as the Pareto front.

The process of MOBO starts by sampling a small number of initial data points to construct a surrogate model for each objective, which is often a Gaussian process (GP) model. The posterior of the GP provides information for an acquisition function to decide the next point to evaluate by balancing where the uncertainty of the surrogate model is large (exploration) and where the current model prediction is good (exploitation). New data is then collected, and the GP model is updated iteratively until the optimal solutions are found or the computational budget is depleted. However, one thing that cannot be neglected in reality is that the data is not always perfect and is likely to be corrupted by sometimes very high noise.

This is an issue that cannot be ignored and might degrade the performance of algorithms (Daulton et al., 2021, 2022; Letham et al., 2019). Noise can come from many sources in chemical processes, such as uncontrollable environmental variables when generating experimental data and measurement errors. Noise is also a noticeable issue for automated experimental platforms (Aldeghi et al., 2021). For example, there might be imprecision of operations by robots, or handling volatile solvents, and transferring unstable reagents which are easy to decompose. Noise is common for large-scale manufacturing and may come from fluctuations in raw chemicals or variations in process conditions (Wang and Ierapetritou, 2018).

Bayesian optimization under uncertainty is an active research area, which is being widely studied in computer simulations subject to numerical noise (Baker et al., 2022; Wang et al., 2023), compared to

* Corresponding authors.
E-mail addresses: jz596@cam.ac.uk (J. Zhang), aal35@cam.ac.uk (A.A. Lapkin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compchemeng.2024.108983
Received 25 July 2024; Received in revised form 12 December 2024; Accepted 13 December 2024
Available online 9 January 2025

noise-free settings, the noisy objective functions $y_i(\boldsymbol{x})$ can be represented by

$$y_i(\boldsymbol{x}) = f_i(\boldsymbol{x}) + \varepsilon_i, \tag{1}$$

where $f_i(\boldsymbol{x})$ are the true objective functions for $i = 1, \ldots, n$, $\boldsymbol{x}$ are a set of control variables and ε_i is additive noise which is often assumed to be Gaussian $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ (Wentzell and Brown, 2000). When σ_i is a constant, the noise is called *homoscedastic* and when $\sigma_i(\boldsymbol{x})$ varies at different $\boldsymbol{x}$ the noise is called *heteroscedastic*.

For single-objective optimization, new acquisition functions have been proposed to address noisy settings. The most widely used expected improvement (EI) acquisition function has been extended to noisy conditions, for example, expected quantile improvement (EQI) (Picheny et al., 2013) and augmented expected improvement (AEI) (William, 1982). Other acquisition functions that can intrinsically handle noise are information-based acquisition functions like predictive entropy search (PES) (Hernández-Lobato et al., 2014) or Thompson sampling (TS) (Kandasamy et al., 2018). Some other acquisition functions are risk-averse (Makarova et al., 2021) or based on knowledge gradient (Daulton et al., 2023).

In terms of MOBO with noisy observations, there are several challenges. Firstly, evaluating the performances of MOBO algorithms is not always straightforward under noisy conditions, compared to single objective optimization. Some popular performance metrics such as hypervolume can be misguided by the noisy points on the current Pareto front and hence suggest a wrong direction for optimization (Branke, 2023). While there are MOBO algorithms that can handle noise (Daulton et al., 2022; Bradford et al., 2018; Daulton et al., 2020;Semochkina et al., 2024), there is much less discussion on performance metrics when comparing different algorithms. Secondly, most approaches have been tested assuming homoscedastic noise, meaning that the variance of the noise does not depend on $\boldsymbol{x}$. In practice for chemical experiments, however, the noise is usually heterogeneous. To the best of our knowledge, the performance of these MOBO algorithms under heteroscedastic noise has not yet been compared. Lastly, so far, most studies on noisy MOBO focus on synthetic benchmark problems and practical case studies are rare, which limits the validation of these algorithms under realistic conditions and hinders their adoption in industry.

In this work, we focus on the multi-objective reaction optimization problem under noisy conditions. A recent algorithm – Euclidian expected quantile improvement (MO-E-EQI; Semochkina et al., 2024) – has been adopted to deal with heteroscedastic noise and subsequently assessed under different noise structures and magnitudes. We compare MO-E-EQI to several recent MOBO algorithms that can also handle noise via different performance metrics. Additionally, MO-E-EQI is applied to a real-life reaction optimization problem. Our study contributes to the advancement of noisy MOBO methods that are robust, efficient, and applicable to a wide range of practical problems in reaction development.

2. Materials and methods

Two reaction systems were considered in this study: an *in silico* reaction simulator for algorithm comparison under noisy conditions, and a real-world reaction system where MO-E-EQI was implemented to guide the search for the optimal conditions.

2.1. In silico reaction simulator setup

A reaction simulator was implemented for the *in silico* study to evaluate algorithm performances. It was created based on an experimental study of a nucleophilic aromatic substitution (S_NAr) reaction from the literature (Hone et al., 2017): 2,4-difluoronitrobenzene **1** reacting with pyrrolidine **2** to generate the desired product *ortho*-substituted **3**, *para*-substituted **4** and *bis*-adduct **5** as side products (Fig. 1). The reaction was conducted in a continuous flow reactor where concentrations of **1, 2, 4** and **5** at the end of the reactor were measured using an online HPLC.

Table 1
Control variables and variable ranges of the *in silico* study. The reaction scheme is shown in Fig. 1.

Variable	Unit	Range
Molar equivalent of **2**:**1**	[–]	1.0–5.0
Residence time	[min]	0.5–2.0
InitialL concentration of 2,4-difluoronitrobenzene (**1**)	[mol/min]	0.1–0.5
Temperature	[°C]	30–120

The reaction system can be described by a reactor model with plug flow assumption and reaction kinetics,

$$\begin{aligned}
\frac{dc_1}{d\tau} &= -r_1 - r_2 \\
\frac{dc_2}{d\tau} &= -r_1 - r_2 - r_3 - r_4 \\
\frac{dc_3}{d\tau} &= r_1 - r_3 \\
\frac{dc_4}{d\tau} &= r_2 - r_4 \\
\frac{dc_5}{d\tau} &= r_3 + r_4,
\end{aligned} \tag{2}$$

where r_1-r_4 are reaction rates, τ is the time that species spent along the flow reactor and c_1-c_5 are concentrations for each chemical species at different locations in the reaction tube. This system is also a widely used benchmark for multi-objective reaction optimization (Tu et al., 2022; Vel et al., 2024; Felton et al., 2021).

The optimization problem was set up based on the information in the original paper. There were four control variables for this reaction: residence time, equivalent, temperature and initial concentration, and their ranges are listed in Table 1. In the original paper (Hone et al., 2017), the way to adjust the residence time and molar equivalent is by change the flow rates of **1** and **2**, correspondingly. Two objectives were set for this reaction, to consider the reaction outcome as well as the environmental impact simultaneously: maximizing the space–time-yield (STY) and minimizing the E-factor (Sheldon et al., 2022), calculated as

$$\begin{aligned}
\text{STY} &= \frac{c_{\text{product}}}{\tau} \\
\text{E-factor} &= \frac{m_{\text{waste}}}{m_{\text{product}}},
\end{aligned} \tag{3}$$

where c_{product} is desired concentration of the product **3** at reactor outlet (unit: $g \cdot L^{-1}$); m_{product} is total mass of product (unit: g); m_{waste} is total mass of waste (total reagents mass − product mass, unit: g).

2.2. Experimental setup

To implement the algorithm for a real-life reaction optimization problem, an esterification reaction was selected as a model reaction, which is one of the most essential reactions in chemical and pharmaceutical industries. This reaction was chosen due to its wide applicability, including the synthesis of various esters for use in drug formulation and industrial applications (Gaefke et al., 2006). Furthermore, implementing this reaction in a continuous flow setup allows for enhanced reaction control and efficient heat and mass transfer. The selected reaction scheme is shown in Fig. 2, where 1,4-benzenedimethanol (**6**) reacts with acetic anhydride (**7**) to form the desired 4-hydroxymethylbenzyl acetate (**8**) and undesired 1,4-bis(acetoxymethyl)benzene (**9**).

Experimental setup and reaction conditions are shown in Fig. 3. A solution of 1,4-benzenedimethanol (**6**) (0.20 M) and Et_3N (0.20 M) in MeCN and a solution of DMAP (0.10 M) in MeCN were injected into a PTFE T-shape mixer (inner diameter: 0.5 mm) using syringe pumps. The resultant mixture was passed through the reaction tube (volume:

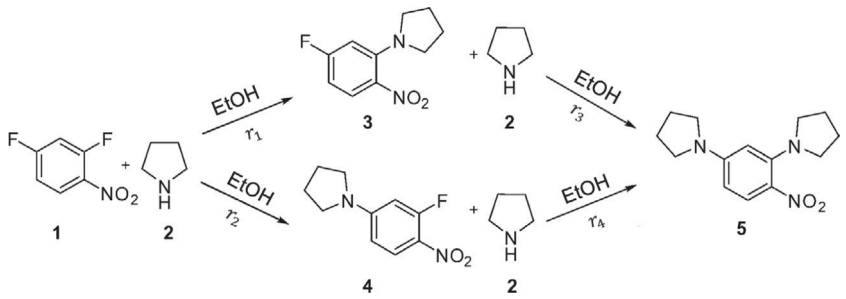

Fig. 1. S_NAr reaction of 2,4-difluoronitrobenzene (**1**) with pyrrolidine (**2**) for *in silico* study.

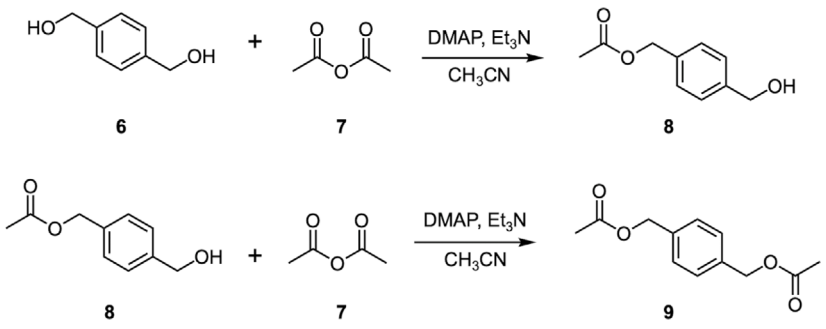

Fig. 2. Esterification reaction of 1,4-benzenedimethanol (**6**) with acetic anhydride (**7**) for experimental study.

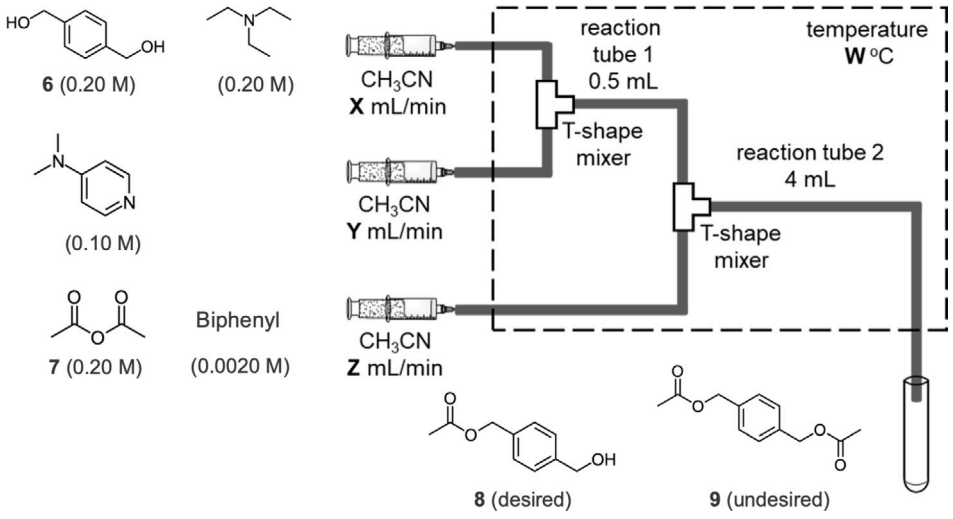

Fig. 3. Experimental setup with the esterification reaction.

0.50 mL, Vapourtec R-Series Tubing Kit). The resultant mixture and a solution of acetic anhydride (**7**) (0.20 M) and biphenyl (0.0020 M) in MeCN were injected into the second PTFE T-shape mixer (inner diameter: 0.5 mm). The resultant mixture was passed through reaction tube 2 (volume: 4.0 mL, Vapourtec R-Series Tubing Kit). Reaction tubes 1 and 2 and two T-shape mixers were immersed in a water bath with a temperature control. Syringe pumps in the system were from TriContinent and controlled by lab automation software Flab. The resultant mixture was added into saturated NH_4Cl solution and CH_3OH at room temperature to quench the reaction. Yields of the desired product 4-hydroxymethylbenzyl acetate (**8**) were determined by HPLC-UV analysis. More details can be found in Appendix A.

Four reaction variables were identified that are important for this reaction and were selected, also shown in Fig. 3. Namely, molar equivalent of 1,4-benzenedimethanol (**6**):acetic anhydride (**7**), DMAP (catalyst) loading, flow rate of acetic anhydride (**7**) and temperature. Table 2 summarized the range of each variable. The flow rate of **7** was set to be **Z** mL/min, ranging from 0.50 mL/min to 2.00 mL/min. The molar equivalent of **6**:**7** was changed by adjusting the flow rates **X** in a range of 10.00–0.50 mL/min; catalyst loading was changed by adjusting the flow rates **Y** in a range of 0.05–2.00 mL/min. The reaction tube 1 and 2 was kept in a water bath with a temperature of **W** ° C. The same objective functions as *in silico* study - maximizing space–time-yield and minimizing E-factor, were chosen as the objectives for optimizing this reaction.

3. Theory

A multi-objective optimization problem often has competing goals. Assuming the goal of simultaneous minimization (without loss of generality), such a problem can be described as:

$$\begin{array}{lll} \min & f_i(\boldsymbol{x}) & \text{for } i = 1, \ldots, n \\ \text{subject to:} & \text{lb}_k \le x_k \le \text{ub}_k & \text{for } k = 1, \ldots, v, \end{array} \tag{4}$$

Table 2
Control variables and variable ranges of the experimental study. The reaction scheme is shown in Fig. 2.

Variable	Unit	Range
Molar equivalent of **6** : **7**	[-]	1.0–5.0
Catalyst DMAP loading	[mol%]	5–50
Flow rate Z	[ml/min]	0.50–2.00
Temperature W	[°C]	25–65

where f_i are the n black-box objective functions, $\boldsymbol{x} = (x_1, \ldots, x_v)^\mathsf{T}$ represent the v input variables, and lb_k and ub_k are bounds on the input x_k. Of course, more complex equality or inequality constraints on the input space could be of interest. However, as described and implemented in this paper, the algorithm does not incorporate native input constraint handling. Constraints on the outputs are less straightforward. Although not implemented in this paper, the MO-E-EQI can deal with constraints on the objectives, Semochkina et al., 2024 incorporated upper bound constraints and that methodology could be adapted to address other types of constraints. We focus on multi-objective optimization on a Pareto front (Fonseca and Fleming, 1995), a collection of optimal solutions for each f_i such that no single objective can be improved without making at least one of the remaining objectives worse (Giagkiozis and Fleming, 2014).

For simplicity, we suppress the indexing across the multiple objective functions. We will focus on maximizing objective 1 and minimizing objective 2 in the remainder of the paper referred to as f_1 and f_2 respectively.

3.1. Euclidian distance based expected quantile improvement

To handle noisy observations, we adopted a recent approach introduced by Semochkina et al. (2024) using the **M**ulti-**o**bjective **E**uclidean **E**xpected **Q**uantile **I**mprovement (MO-E-EQI) criterion to guide our optimization. As with many MOBO methods, this method includes: (i) establishing an initial belief (prior) about the system's behaviour, building a sampling plan, calculating the responses at those points and fitting a GP model to this data in line with those prior beliefs; (ii) strategically selecting new input points $\boldsymbol{x}$ for system's evaluation based on a criterion that maximizes acquisition function; (iii) incorporating the obtained results into the existing belief, leading to a more accurate understanding of the system's optimum and its corresponding input value.

Fig. 4 illustrates an example of this process. Based on the eight noisy data points (circles) that have been already observed, a GP model was fitted with the noisy observations $y(\boldsymbol{x})$. The next point for evaluation is then selected based on the distribution of an unobserved point $Y(\boldsymbol{x})$ at a new location $\boldsymbol{x}$, characterized by the GP's mean and variance. The algorithm can decide whether to go somewhere near the current best model mean (exploitation), or somewhere with large uncertainty (exploration), and gradually move towards the optimum.

The initial belief about the function is represented by a Gaussian process (Rasmussen and Williams, 2006).

$$f(\boldsymbol{x}) \sim \mathrm{GP}\left\{\mu(\boldsymbol{x}), \kappa(\boldsymbol{x}, \boldsymbol{x}')\right\},$$

a stochastic process defined via a mean function $\mu(\cdot)$ and covariance function $\kappa(\boldsymbol{x}, \boldsymbol{x}')$. For any finite set of input vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^S$, arranged in a design matrix $X_S = \left(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^S\right)^\mathsf{T}$ a collection of random variables from this process follow a multivariate normal distribution

$$\begin{pmatrix} f(\boldsymbol{x}^1) \\ \vdots \\ f(\boldsymbol{x}^S) \end{pmatrix} \sim N\left(\boldsymbol{\mu}(X_S), K(X_S)\right),$$

with mean vector $\boldsymbol{\mu}(X_S)$ having jth entry $\mu(\boldsymbol{x}^j)$ and covariance matrix $K(X_S)$ having jk-th entry $\kappa(\boldsymbol{x}^j, \boldsymbol{x}^k)$ $(j, k = 1, \ldots, S)$.

The posterior for $f(\boldsymbol{x})$, conditional on noisy data $\boldsymbol{y}^S = \left[y(\boldsymbol{x}^1), \ldots, y(\boldsymbol{x}^S)\right]^\mathsf{T}$ is also a GP:

$$f(\boldsymbol{x}) \mid \boldsymbol{y}^S \sim \mathrm{GP}\left\{m(\boldsymbol{x}), s(\boldsymbol{x}, \boldsymbol{x}')\right\}, \tag{5}$$

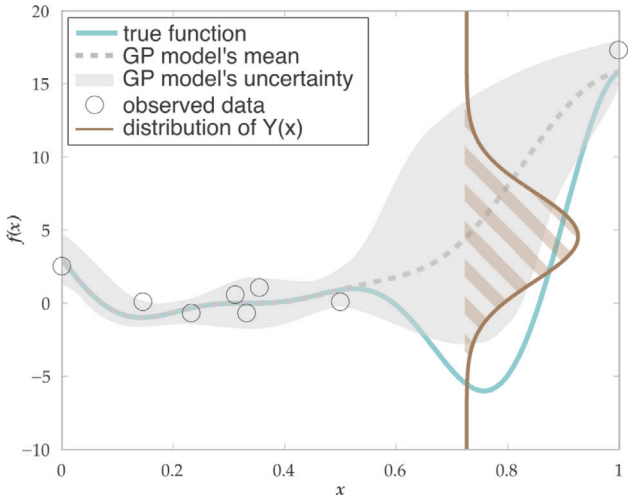

Fig. 4. Illustration of fitting a GP model with noisy observations. Based on the eight noisy observations (circles), a GP model $y(\boldsymbol{x})$ is fitted. The next point for evaluation is selected based on the distribution of an unobserved point $Y(\boldsymbol{x})$ at a new location $\boldsymbol{x}$, characterized by the GP's mean and variance. The algorithm can decide whether to go somewhere near the current best model mean (exploitation), or somewhere with large uncertainty (exploration), and gradually moves towards the optimum.

with updated mean $m(\boldsymbol{x})$ and covariance $s(\boldsymbol{x}, \boldsymbol{x}')$ functions (Semochkina et al., 2024).

The fundamental idea behind the expected quantile improvement (EQI) acquisition function is 'tunable precision' where a particular observation's precision can be adjusted using more computational budget. In a physical system, noise can come from various sources, including environmental variables that are outside of our control and measurement errors. In this paper for *in silico* experiments, noise was added to the model predictions to imitate noise in a physical system. Mathematically, tunable precision can be implemented by taking a sample mean of N independent Monte Carlo drawings:

$$\bar{y}_N(\boldsymbol{x}) = \frac{1}{N}\sum_{r=1}^{N} y_r(\boldsymbol{x}) = \frac{1}{N}\sum_{r=1}^{N}\left[f(\boldsymbol{x}) + \varepsilon_r\right] = f(\boldsymbol{x}) + \frac{1}{N}\sum_{r=1}^{N}\varepsilon_r, \tag{6}$$

where $\varepsilon_r \sim \mathcal{N}(0, \sigma^2)$. The estimate of the variance σ^2 can be calculated as

$$\hat{\sigma}^2_N(\boldsymbol{x}) = \frac{1}{N-1}\sum_{r=1}^{N}\left[y_r(\boldsymbol{x}) - \hat{y}_N(\boldsymbol{x})\right]^2 \tag{7}$$

and an estimated variance of a Monte Carlo sample mean $\bar{y}_N(\boldsymbol{x})$ is $\hat{\sigma}^2_N(\boldsymbol{x})/N$.

Single-objective EQI to address noisy data was introduced by Picheny et al. (2013) defined as

$$\mathrm{EQI}\left[\boldsymbol{x}^{S+1}, \sigma^2(\boldsymbol{x}^{S+1})\right] = \mathrm{E}_{Q^{S+1}}\left[\left(q^S(\boldsymbol{x}^\star) - Q^{S+1}(\boldsymbol{x}^{S+1})\right)^+\right], \tag{8}$$

where $(z)^+ = \max(0, z)$, $q^S(\boldsymbol{x}) = m(\boldsymbol{x}) + \Phi^{-1}(\beta)s(\boldsymbol{x})$ is the β-quantile from the current GP posterior (see, for example, Rasmussen and Williams (2006) equation 7 for the details on the posterior distribution of a GP) with $\beta \in [0.5, 1)$, $\boldsymbol{x}^\star = \mathrm{argmin}_{\boldsymbol{x}\in X_S} q^S(\boldsymbol{x})$ and $Q^{S+1}(\boldsymbol{x}^{S+1})$ is the corresponding β-quantile when one additional observation $y(\boldsymbol{x}^{S+1})$ is added to the data set. This was later extended to the multi-objective case by Semochkina et al. (2024) and the equivalent equation for MO-E-EQI could be found in Equation (2.10) of that paper.

The variance of the next observation is represented here by the variance $\sigma^2(\boldsymbol{x}^{S+1})$. This should constitute the expected variance at the new input $\boldsymbol{x}^{S+1}$ if the simulator is run or the experiment is conducted. It was shown by Picheny et al. (2013) that $Q^{S+1}(\boldsymbol{x}^{S+1})$ follows a different Gaussian distribution. The posterior mean and variance are

$$m_{Q_{S+1}}\left(\boldsymbol{x}^{S+1}, \sigma^2(\boldsymbol{x}^{S+1})\right) = m(\boldsymbol{x}^{S+1}) + \Phi^{-1}(\beta)\sqrt{\frac{\sigma^2(\boldsymbol{x}^{S+1}) \times s^2(\boldsymbol{x}^{S+1})}{s^2(\boldsymbol{x}^{S+1}) + \sigma^2(\boldsymbol{x}^{S+1})}}, \tag{9}$$

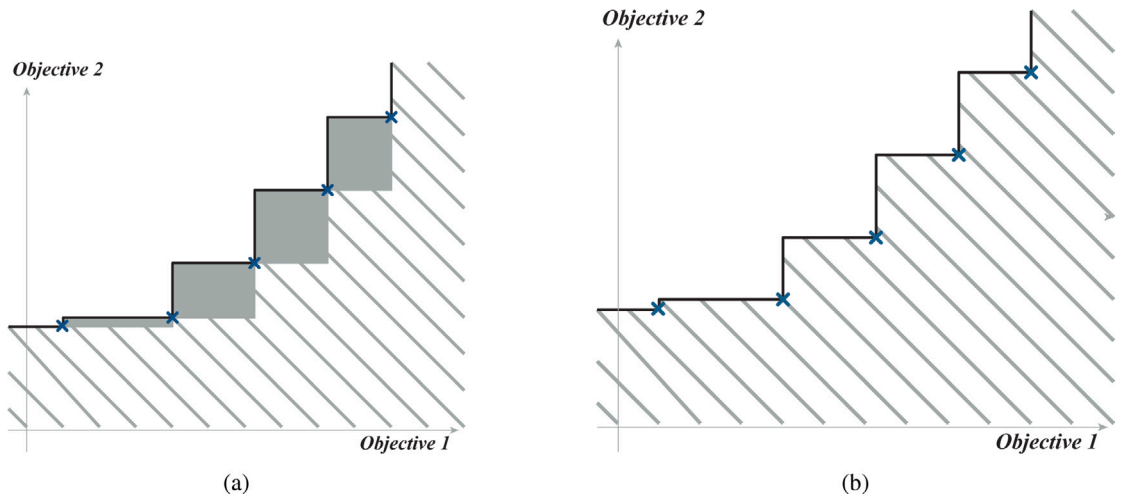

Fig. 5. Improvements possible from a single point in the Pareto set (Keane, 2006). (a). Aggressive setting: a Pareto set of five non-dominated points (blue crosses) for a problem with two objectives. The solid line is the Pareto front. The shaded area shows where new points would augment the Pareto front, while the hatched area is where new points would dominate and replace the existing set of non-dominated points. (b). Non-aggressive setting: the integration area is extended to include the parts where new solutions could be added to the current Pareto set without replacing any of the current points.

and

$$s^2_{Q_{S+1}}\left(\boldsymbol{x}^{S+1}, \sigma^2(\boldsymbol{x}^{S+1})\right) = \frac{\left[s^2(\boldsymbol{x}^{S+1})\right]^2}{s^2(\boldsymbol{x}^{S+1}) + \sigma^2(\boldsymbol{x}^{S+1})}, \tag{10}$$

where $m(\boldsymbol{x}^{S+1})$, $s^2(\boldsymbol{x}^{S+1})$ are the GP's mean and variance respectively and $\sigma(x)$ is the observational noise from (1). In reality an estimate from (7) is used as a proxy for existing observations and an estimate based on the current GP fit is calculated for future observations. This is discussed in detail in Section 4.1.1.

The choice of β tunes the level of reliability wanted on the final result, setting $\beta = 0.5$ means that the algorithm will compare design points based on the GP model's mean only in (9) and uncertainty from (10). With a higher β, EQI penalizes designs with high uncertainty and becomes more conservative. It should be noticed that, to calculate EQI using (8), future noise $\sigma^2(\boldsymbol{x}^{S+1})$ is required at any $\boldsymbol{x}^{S+1}$ as $\sigma^2(\boldsymbol{x}^{S+1})$ also influence the performance of EQI.

To quantify the improvement of EQI for two objectives, two criteria of MO-E-EQI are illustrated in Fig. 5 by plotting a Pareto front set against two objectives. The horizontal and vertical solid lines represent the Pareto front. The expected improvement is calculated by using the probability of improvement (area of hatched area) and Euclidian distance between the centroid of the area and each member in the Pareto set (Keane, 2006). This means that for each new potential input, $\boldsymbol{x}^{S+1}$, the expected improvement is calculated as the expectation of the distance between the quantile of the closest Pareto front point and the quantile of the potential improved point with respect to the current joint probability distribution of the response associated with that input point (from the corresponding GPs' distributions). The exact expression and derivation can be found in Equation (2.10) of the Semochkina et al. (2024) paper. This expectation is a double integral, the limits of which are dictated by the area of interest as demonstrated by Fig. 5. If a new point falls into the hatched area, it will dominate and replace at least one current member in the Pareto set, whereas if it is placed in the shaded hatched area, it will augment current Pareto solutions. Therefore, by changing the area of integration, MO-E-EQI can change between an aggressive setting by adding fewer points on the Pareto front, or a non-aggressive setting by adding more in-between points. MO-E-EQI is calculated for each point on a comprehensive grid and the point that maximizes the MO-E-EQI objective function is selected as the one to be sequentially added to the design. Appendix B describes detailed steps of the algorithm.

3.2. Related noisy MOBO algorithms

The performance of MO-E-EQI was benchmarked with some state-of-the-art noisy MOBO methods which will be introduced briefly here.

*q*NParEGO (Noisy Pareto Efficient Global Optimization (Daulton et al., 2020; Letham et al., 2019)) is a batch variant of ParEGO (Knowles, 2006) suitable for noisy settings. *q*NParEGO converts a multi-objective problem into a single-objective problem by applying a random scalarization of the objectives. By choosing a different weight vector for scalarization at each iteration of the search, an approximate Pareto front can be built up gradually.

*q*NEHVI (Noisy Expected Hypervolume Improvement (Daulton et al., 2023)) is based on the expected hypervolume improvement criterion for noisy settings.

TSEMO (Thompson Sampling Efficient Multiobjective Optimization (Bradford et al., 2018)) is based on the Thompson sampling heuristic that samples from each GP using spectral sampling. This leads to individual functions from which an approximate Pareto set can be found using the NSGA-II algorithm. The next evaluation point is then selected from the Pareto set based on the largest hypervolume improvement.

3.3. Implementation details

3.3.1. Initial design

We start MOBO with some initial sample points set by a space-filling strategy. These initial points help to map the entire design space and provide initial information on the design space. In the following investigations, 20 initial sampling points were generated using Maximum Projection design, MaxPro (Ba and Joseph, 2018), which is one type of Latin hypercube design that maximizes space-filling properties on projections with respect to all possible subsets of factors. Different numbers of initial sample points and sampling strategies were compared in Appendix C.

3.3.2. Noise settings

Measurements of chemical experiment outcomes may differ even when the experiment is repeated under the same conditions. This means that the recorded measurements will spread around some mean value or the true response. It is common to represent noisy measurements as having a normal distribution with a certain mean and a standard deviation, where the mean represents the true response. The more replicated experiments and their response measurements there are, the better one can capture the true distribution of the response and the

Table 3
Heteroscedastic noise settings for the *in silico* case study with linear and loglinear noise structures.

Noise structure	α	β
Linear-1	0.01	–
Linear-2	0.05	–
Linear-3	0.10	–
Linear-4	0.20	–
Loglinear-1	0.85	−1.70
Loglinear-2	1.20	−1.30

spread of the noisy measurements around the true response. This is often modelled as additive Gaussian noise $\varepsilon \sim N(0, \sigma^2)$, where σ is the standard deviation. If the noise appears to be non-Gaussian in practice, further adjustments need to be considered (Picheny et al., 2022).

More often, chemical measurements are heteroscedastic and the standard deviation of the Gaussian noise changes at different conditions $\boldsymbol{x}$,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2(\boldsymbol{x})).$$

To evaluate the performances of algorithms under heteroscedastic noise, we implemented two types of noise structures with different magnitudes for the *in silico* study.

Linear noise structure is the most commonly assumed in literature (Jalali et al., 2017; Wang and Ierapetritou, 2017) where a linear relationship is set between the standard deviation of noise and the function value. Four noise magnitudes were chosen $\alpha = 0.01, 0.05, 0.10$ and 0.20 for comparison.

$$\sigma(\boldsymbol{x}) = \alpha \cdot f(\boldsymbol{x}). \tag{11}$$

In addition, to mimic the noise structure close to chemical measurements, we adopted Horwitz's rule from analytical chemistry, which suggests a linear relationship between the log standard deviation of noise and the log value of the function value (Huang et al., 2006)

$$\log_{10}[\sigma(\boldsymbol{x})] = \alpha \cdot \log_{10}[f(\boldsymbol{x})] + \beta. \tag{12}$$

Two noise magnitudes of loglinear noise were chosen here. Loglinear-1 with $\alpha = 0.85$ and $\beta = -1.70$ follows an empirical setting suggested by Albert and Horwitz (1997) and Loglinear-2 with $\alpha = 1.20$ and $\beta = -1.30$ was chosen to be a larger noise magnitude. Parameters of noise settings can be found in Table 3, an illustration of different noise structures and magnitudes can be found in Appendix D.

3.3.3. Performance metrics

To compare the performances of MOBO algorithms, several metrics have been proposed in literature to evaluate how close the solutions are to the Pareto front and how evenly the solutions are distributed along the frontier, as shown in Figs. 6(a) and 6(b). In addition, the number of solutions on the Pareto front is also important because it provides more choices to decision-makers. In this study, MOBO algorithms were evaluated using hypervolume-based metric, coverage metric and the number of Pareto optimal solutions for comparisons.

The hypervolume-based (HV) metric is widely recognized as a unary value which is able to measure the closeness of the solutions to the optimal set (Zitzler and Thiele, 1999). The hypervolume metric calculates the volume of the objective space covered by members of an obtained Pareto set P bounded by a reference point R (see Fig. 6(c)):

$$\mathrm{HV} = \bigcup_{p=1}^{N} v_p, \ p \in P, \tag{13}$$

where v_p is the area of each of the rectangles the space is broken into by the reference point and the Pareto front points.

The coverage metric was proposed by Lewis et al. (2009) to quantify how well a set of Pareto front solutions covers the objective space. It divides the objective space into several radial sectors originating from the reference point $R(r_1, r_2)$. The value of the coverage metric is calculated as the ratio of these sectors with at least one Pareto front solution to the total number of sectors, defined as:

$$\Psi = \frac{1}{N} \sum_{n=1}^{N} \Psi_n, \tag{14}$$

where N is the number of sectors and

$$\Psi_n = \begin{cases} 1, & \text{if } p_i \in P \text{ and } \alpha_{i-1} \le \tan\left[\frac{|r_1 - f_1(x)|}{|r_2 - f_2(x)|}\right] \le \alpha_i \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Here P is the Pareto optimal solutions and α_i is the radial angles of the sections. Both objectives need to be scaled to [0,1] to calculate this metric. As an example in Fig. 6(d), the solution space is divided into six sections and four contain at least one Pareto front solution. Therefore the coverage metric is calculated to be 66.7%. In the following algorithm comparison, the same number of nine sections were used for all algorithms.

3.3.4. Implementation details

In this study, all benchmaking was performed on a laptop equipped with an Apple M1 chip, 16 GB of RAM. The code for implementing the algorithms can be found online at: https://github.com/sustainable-processes/MO-E-EQI.

Implementation of MO-E-EQI was in R (R Core Team, 2022). Sequential design maximizing MO-E-EQI was implemented in the R package `MOEEQI` (Semochkina, 2024a) and `DiceOptim` (Picheny and Ginsbourger, 2014). To incorporate noisy observations into our algorithm, GPs were fitted using the `km` function in the `DiceKriging` (Roustant et al., 2021) package providing the `noise.var` argument, specifying the corresponding variance for each observational point. A Python version of MO-E-EQI is currently being developed and will be provided on the Github repository.

For algorithm comparison, *q*NEHVI, *q*NParEGO were implemented in Python using Botorch (Balandat et al., 2020) and TSEMO was implement through Summit (Felton et al., 2021). Gaussian process models were trained using `HeteroskedasticSingleTaskGP` in Botorch. An interactive app was designed using R Shiny (Semochkina, 2024b) for algorithm performance comparison, see Appendix E. This interactive app visualizes the performance of various algorithms used in the *in silico* example across different noise levels and algorithm modifications. By plotting combinations of algorithm results, the tool allows users to interactively explore the relationship between inputs and outputs in the objective space. This is particularly valuable for practitioners as the input space is often neglected when plotting optimization results.

4. Results and discussion

4.1. In silico algorithms comparison under noise

MO-E-EQI is first compared with other MOBO algorithms under noisy conditions; namely, *q*NParEGO, *q*NEHVI, TSEMO. The sequential Latin hypercube sampling strategy (named LHS Space Filling) was also included for comparison as a baseline. For the following comparison, the algorithms started with 20 initial sampling points with two objectives to maximize STY and to minimize E-factor.

The GPs work best when covariates are normalized to the unit cube and outcomes are standardized (i.e. zero mean, unit variance). To achieve that, prior to fitting a GP, all inputs were scaled to $[0, 1]$ and all the outputs from the training data were standardized, using the sample mean and sample variance. Those samples' means and variances were subsequently used to standardize all model outputs in the sequential part of the algorithm.

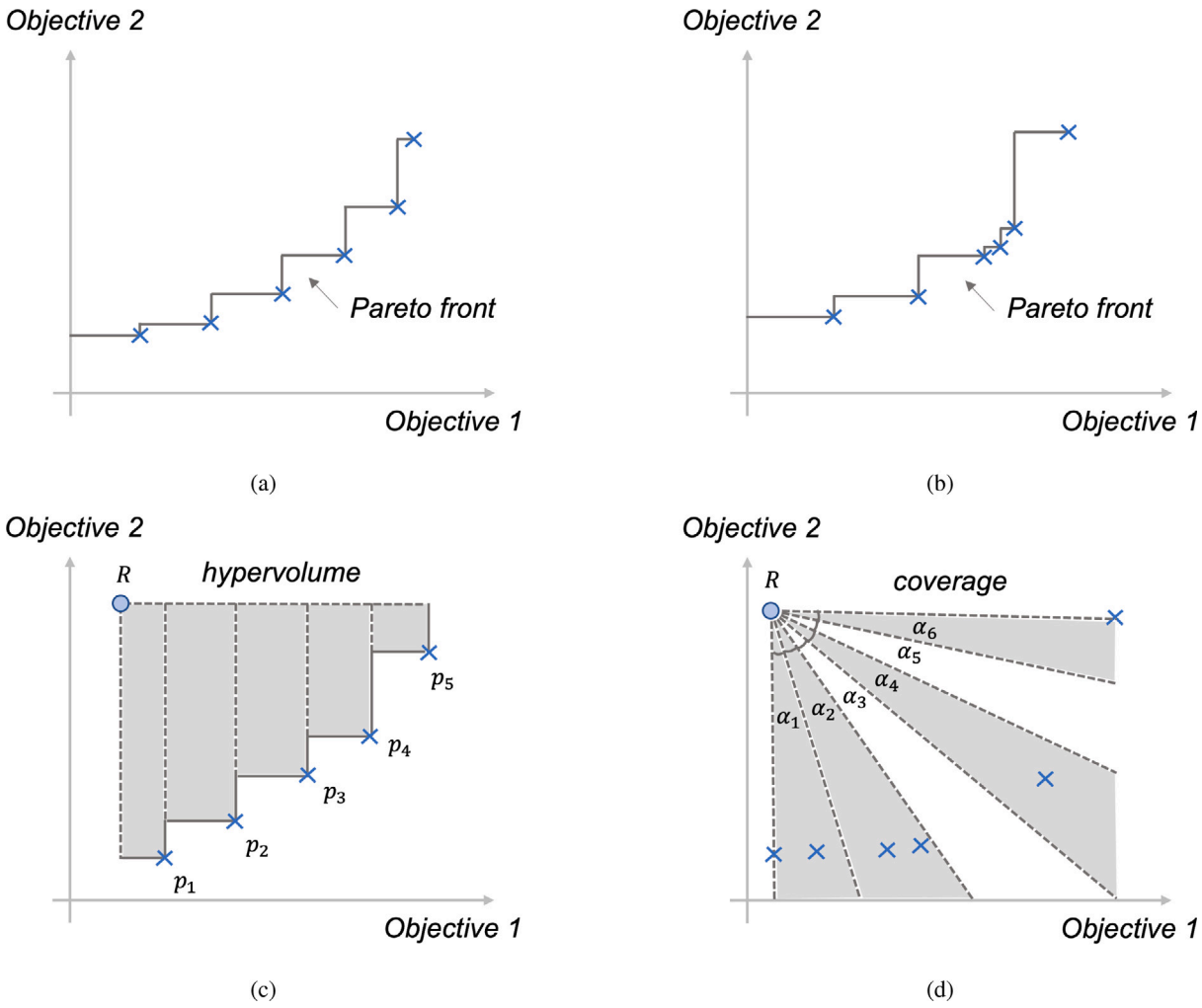

Fig. 6. Pareto front solutions of multi-objective optimization and evaluation metrics. (a). Pareto front solutions with a uniform spread; (b). Pareto front solutions with a poor spread; (c). Illustration of the hypervolume metric; (d). Illustration of the coverage metric.

4.1.1. Hypervolume distance comparison

Under a significant level of noise, it is not uncommon that a single model run could significantly overestimate the true Pareto front. The most popular metric in the literature (Riquelme et al., 2015) – hypervolume (HV) – is inadequate to estimate how well the identified Pareto front is compared to the true Pareto front. With that in mind, we calculate the volume of the difference between the Pareto front identified by an algorithm and the true Pareto front. This takes into account both under- and over-estimation and combines these into one metric based on HV. A drawback of taking the difference between the true HV and the one identified by the algorithm under high noise is that where both under- and over-estimation are present, they could potentially cancel each other out and not provide a true picture of how far the identified Pareto front is from the true one (see Fig. 7(b) for details).

We use the R package `sf` (Pebesma et al., 2023) to create complex polygons, using a reference point and the set of points on the Pareto fronts, take the difference between those polygons, take the union of under- and over-estimating polygons and calculate the volume of the resulting polygon as our measure. To construct the HV distance for this example, a reference point of (STY = 13, E-factor = 4) was used.

One of the major drawbacks of noisy observations is that a single observation is unreliable as an estimate of the truth. The *q*NEHVI algorithm deals with that by integrating out the uncertainty of the GP model when calculating the expected HV improvement. However, if the ultimate goal is to identify the Pareto front correctly, a single observation may not be sufficient. We compared the performance of *q*NEHVI, *q*NParEGO, TSEMO, LHS Space Filling and a single-run-based MO-E-EQI algorithm. Even though all five algorithms were implemented for a single run only, unlike *q*NEHVI, *q*NParEGO and TSEMO, MO-E-EQI allows repeated observations. This means that an algorithm can choose to return to an existing input location and request the model to be run again. In that case, the model observations will be combined into one observation (Semochkina et al., 2024). Additionally, we compared the performance of algorithms based on multiple model runs. This means that for every input, the simulator is run multiple times and the sample mean is adopted as an 'observation' when a GP is fitted. That also means that, if the algorithm desires to repeat an observation (i.e. run the simulator at the same input again), the simulator will again be run multiple times and the sample mean will be calculated as the secondary observation. The two observations will then be combined according to Semochkina et al. (2024) Equation (2.8).

A crucial factor in the performance of these algorithms is accurate noise estimation. Since they rely on noisy observations, understanding the noise level and structure is essential. Noise can be estimated during experiments through repeated measurements or obtained from external sources, such as expert knowledge. If repeated experiments are infeasible, prior noise estimates are necessary. However, these algorithms cannot be applied effectively if no noise estimation or related information is available.

The important feature of the MO-E-EQI algorithm is the estimation of the future noise incorporated in the expected improvement calculation. This requires us to predict the level of noise at an unobserved

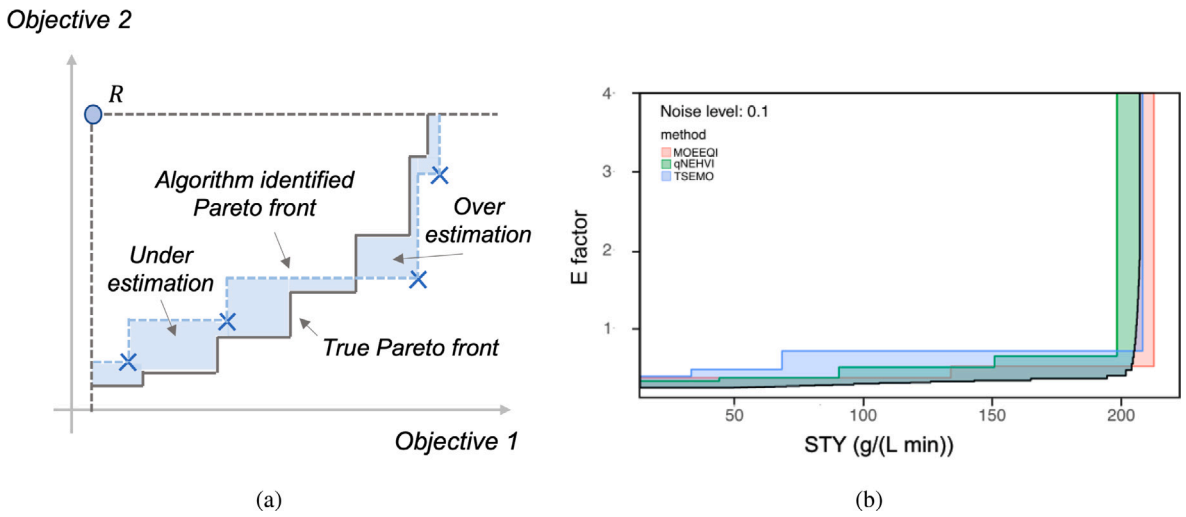

Fig. 7. Hypervolume distance between the identified Pareto front and true Pareto front. This consists of multiple complex polygons. (a). Illustration of reference point and under-, over-estimation of Pareto front. (b). An example of comparing algorithms using hypervolume distance.

input location. We remind our reader that a direct linear/loglinear relationship between the noise standard deviation and the expected value of the response was assumed. Under that assumption, the future noise on each step of the algorithm was calculated using the current fitted GPs' means as a guide for our mean response and the noise was calculated based on the relationship with a specific noise magnitude. This means replacing $f(\boldsymbol{x})$ with $m(\boldsymbol{x})$ in Eqs. (11) and (12).

Another feature of the MO-E-EQI algorithm is the quantile level $\beta \in [0.5, 1)$. We ran the algorithm with $\beta = (0.6, 0.7, 0.8, 0.9)$. The results for all values of β can be found in Appendix F, and for this case, the choice of β was not a major factor in the algorithm's performance compared to the differences with other algorithms' performances. Here we present the comparison of the algorithm for $\beta = 0.7$. The resulting HV distance for all five algorithms and MO-E-EQI with $\beta = 0.7$ are presented in Fig. 8.

It shows that MO-E-EQI, *q*NEHVI, *q*NParEGO, TSEMO show significant efficiency than simple sequential strategy LHS Space Filling. MO-E-EQI finds the optimal solutions quickly followed by *q*NEHVI, while *q*NParEGO and TSEMO show relatively slower HV distance decrease. Comparing between single- and multiple-runs, MO-E-EQI works better with multiple model runs. When increasing the noise level from 0.01 to 0.20, all algorithms show worse performance. It is noticeable for a high noise setting, that the HV distance for the *q*NEHVI algorithm drops and then starts growing again after about 10 iterations. This is potentially due to the overestimation of the true Pareto front. In that case, when a new input is suggested and the noisy observation overestimates the true response, the Pareto front identified by the algorithm can start moving further away from the true Pareto front.

Results of loglinear noise structure can be found in Fig. 9. In general, the trend of algorithms' performance was found to be similar to the linear case. It is interesting to notice for high noise level Loglinear-2 with MO-E-EQI single-run, surprisingly *q*NParEGO shows better HV distance performance than any other algorithms.

We further investigated this by plotting all 20 repeated runs of *q*NParEGO and MO-E-EQI, see Fig. 10. It is notable that the solutions of *q*NParEGO tend to aggregate at the bottom right corner (Fig. 10(a)), while MO-E-EQI suggests more diverse results for both objectives (Fig. 10(b)). Algorithms based on Euclidean expected improvement can effectively target under-sampled areas. This space-filling characteristic can be beneficial in optimization (Wagner et al., 2010), as it ensures a more comprehensive exploration of the solution space.

The MO-E-EQI may not always be the best-performing method in terms of HV distance metric, particularly in cases with high noise and single runs. As mentioned before, all the methods presented in this paper rely on some understanding of noise: either estimated or provided from other sources. Single-run settings are less realistic in terms of accurate noise gauging for an unknown system. The value of repeated observations is widely perceived (Gilmour and Trinca, 2012). In fact, the framework used to fit GPs for *q*NEHVI and *q*NParEGO methods acknowledges the important role the replications play in estimating noise (Binois et al., 2018). Without accurate noise estimates, prior assumptions become critical. This can lead to MO-E-EQI underperforming compared to other methods with strong assumptions on the noise. If those assumptions are accurate, such approaches can outperform single-run estimation. However, incorrect variance models can undermine robustness. Multiple-run approaches with estimated variance are generally more robust.

Due to the properties of the current reaction simulator, the true Pareto front is quite perpendicular. This indicates that improving one objective does not strongly influence the other, which also have been noticed from the literature (Vel et al., 2024). This means that if there are points occupied at the right bottom corner (at maximum STY) and closed to the true Pareto front, HV distance will decrease quickly. Even though MO-E-EQI adds more points for the other objective, this does not show a significant decrease of HV distance as *q*NParEGO does. Therefore, although HV can describe how close the solutions are to the real Pareto front, it cannot describe how spread the solutions are. As a summary, for real applications, it should be noticed that when the noise level is high, multiple observations are necessary for MO-E-EQI to have a good estimation of the noise level. As the structure of the Pareto front cannot be known *a priori*, the algorithms need to be evaluated from different perspectives and this will be further explained in the following sections.

4.1.2. Coverage comparison

To evaluate how solutions spread on the objective space, we adopted the radial coverage metric (Lewis et al., 2009). This metric specifically describes how well the solutions cover the whole range of the objective space.

Results are shown in Fig. 11. MO-E-EQI shows a better coverage (around 50% of the Pareto front) than other algorithms for both linear and loglinear noise structures with various noise magnitudes. *q*NParEGO shows the worst performance in terms of coverage as the solutions tend to aggregate into the right-bottom corner. When comparing *q*NParEGO and MO-E-EQI with Loglinear-1 noise, the coverage metric was 0.3 for *q*NParEGO and 0.6 for MO-E-EQI. It is also notable that there is no clear trend of how noise structure and magnitude affect the

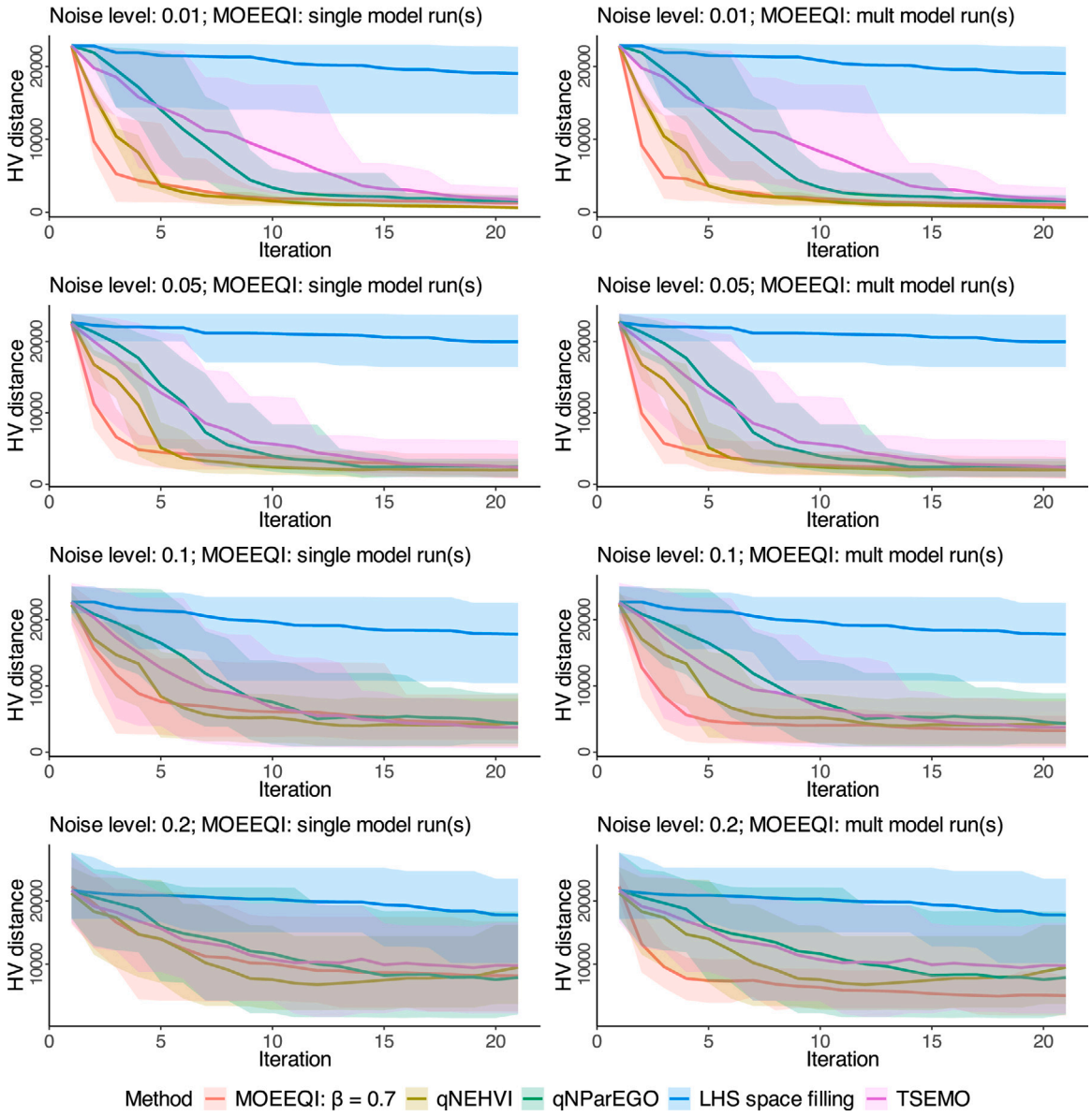

Fig. 8. Comparison of hypervolume distances between MO-E-EQI, *q*NEHVI, *q*NParEGO, TSEMO and LHS Space Filling under linear noise. Solid lines represent the average hypervolume distances over 20 repeated simulations with shading giving 95% confidence intervals.

coverage, for MO-E-EQI, the coverage dropped by 10% from Linear-1 to Linear 4 and 22% from Loglinear-1 to Loglinear-2.

One thing to note is that this coverage metric only describes how distributed the solutions are, no matter how close they are to the Pareto front. LHS Space Filling strategy shows quite high scores of coverage metric, although its performance of hypervolume distance is very poor from Fig. 9. This demonstrates that the evaluation of noisy MOBO algorithms cannot just depend on a single aspect.

4.1.3. Number of Pareto optimal solutions comparison

As part of this study, the numbers of Pareto front solutions are compared. This is a piece of useful information for real applications as it will provide chemists/engineers with more choices of solutions. Three settings of the MO-E-EQI were considered here, both aggressive (Fig. 5(a)) and non-aggressive (Fig. 5(b)) criteria with 20 and 40 sequential optimization steps:

(1) MO-E-EQI-1: total 20 optimization steps, 20 aggressive steps;

(2) MO-E-EQI-2: total 20 optimization steps, 10 aggressive steps followed by 10 non-aggressive steps;

(3) MO-E-EQI-3: total 40 optimization steps, 20 aggressive steps followed by 20 non-aggressive steps.

Fig. 12 compares the numbers of solutions on the Pareto front for different algorithms and noise settings. On average, *q*NEHVI and MO-E-EQI-2 give more solutions on the Pareto front for 20 optimization steps. Fewer solutions were observed when the noise level was raised (for cases of Linear-4 and Loglinear-2). By using a mix of aggressive and non-aggressive criteria of MO-E-EQI, augmenting the current Pareto front is encouraged by the algorithm, therefore, an increased number of solutions for MO-E-EQI-2 were found compared to MO-E-EQI-1. In addition, running the MO-E-EQI longer with 40 steps lead to an improved number of Pareto front solutions in general. For relatively low noise MO-E-EQI-3 found more solutions on the Pareto front (average 16, 12, 10, 16 for noise Linear 1–3 and Loglinear-1, respectively), comparing to MO-E-EQI-2 (average 9, 8, 7, 8 for noise Linear 1–3 and Loglinear-1, respectively). When the noise level is high for Linear-4 and Loglinear-2, the number of solutions did not show obvious improvement when increasing the optimization steps.

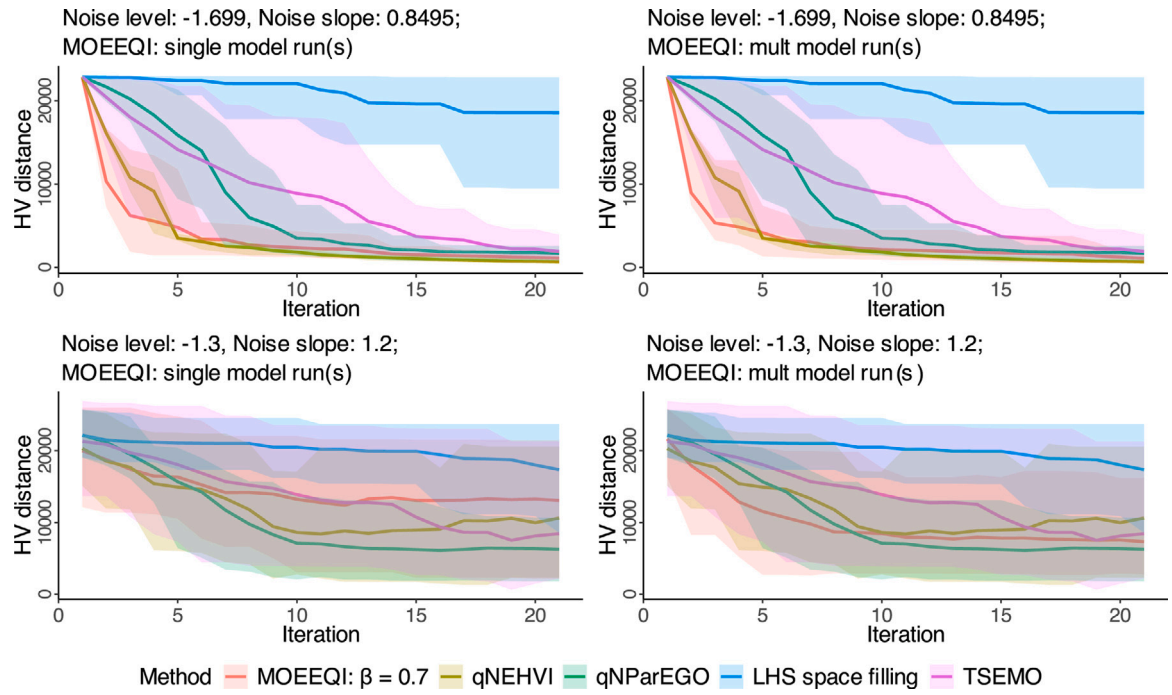

Fig. 9. Comparison of hypervolume distances between MO-E-EQI, *q*NEHVI, *q*NParEGO, TSEMO and LHS Space Filling under loglinear noise. Solid lines represent the average hypervolume distances over 20 repeated simulations with 95% confidence interval.

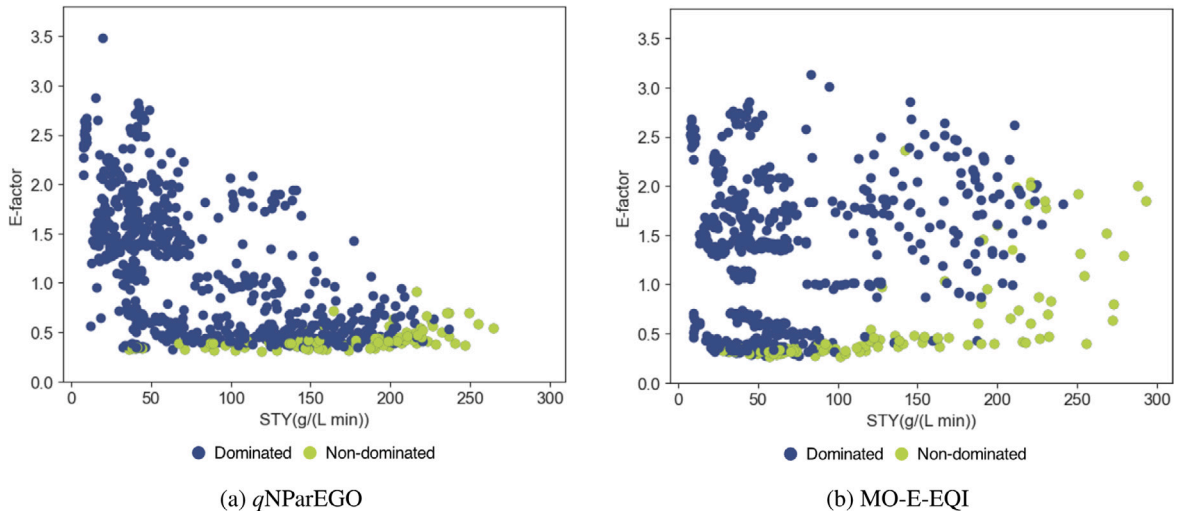

Fig. 10. Results of 20 repeated runs for *q*NParEGO and MO-E-EQI with Loglinear-2 noise structure. (a). Results of *q*NParEGO; (b). Results of MO-E-EQI. For both figures, the optimization started with 20 initial points and optimized for another 20 steps; the blue points are the dominated solutions from all 20 repeated runs and the green points are 20 Pareto front solutions all plotted at once.

In practice, experts can decide on the proportion of aggressive and non-aggressive criteria to get different numbers of optimal solutions given the overall budget, depending on the problem-specific goals.

4.2. Experimental case study

For the real-world application, MO-E-EQI was implemented to guide the optimization of an esterification reaction. Four variables, flow rate, catalyst (DMAP) loading, equivalent and temperature were adjusted to maximum STY while minimize E-factor simultaneously. Initial 20 sampling points were generated by MaxPro and experimental data with three samples was collected for each condition. Based on these data, GPs were trained and optimization was sequentially continued for a further 40 steps with $\beta = 0.6$. The non-aggressive method was used for the MO-E-EQI criterion.

Fig. 13(a)–(c) shows reaction optimization results, also in Appendix G. As shown in Fig. 13(a), the MO-E-EQI identified Pareto front efficiently and showed a very clear trade-off between STY and E-factor. The Pareto front consisted of 20 solutions, with STY ranging from 2.66 g/(L·min) to 195.04 g/(L·min) while the E-factor ranged from 2.89 to 8.47. All the data on this Pareto front highlighted compromised solutions between STY and E-factor, which means that STY cannot be improved without worsening E-factor and vice versa.

Fig. 13(b) presents optimization trajectories for two objective functions. When the algorithm started, the first sequential 10 steps were more exploratory and favoured large STY solutions. Then it became more exploitative by adding more points around the region near the minimum E-factor. It should be noted that although we set the experiment budget to be 60, optimization results were not improved since step 53, which indicated that the optimization had already stalled.

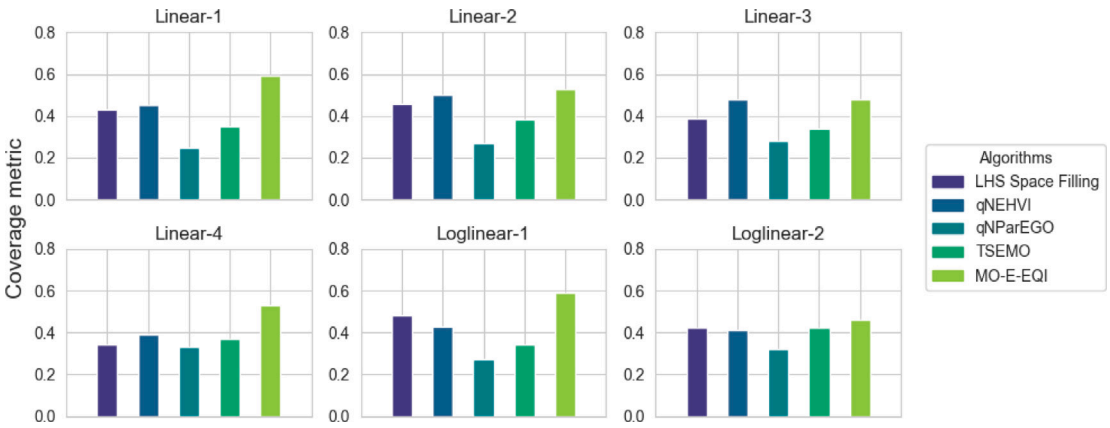

Fig. 11. Comparison of coverage metric between MO-E-EQI, *q*NEHVI, *q*NParEGO, TSEMO and LHS Space Filling over different noise structures and magnitudes. Results are the average values of coverage metric over 20 repeated runs (MO-E-EQI with multiple model runs and $\beta = 0.6$).

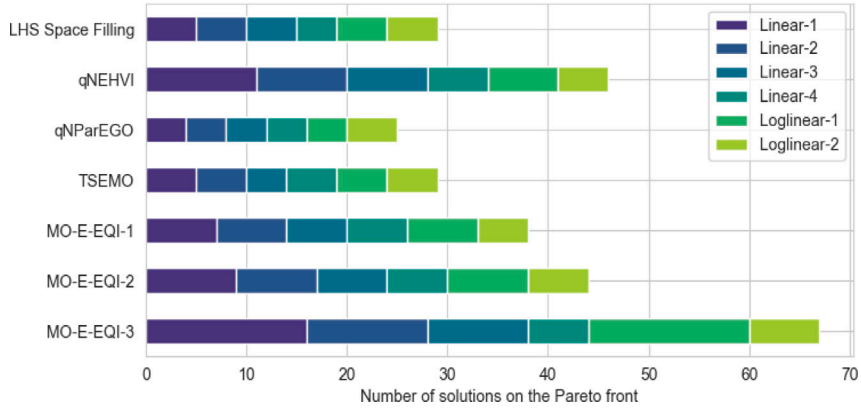

Fig. 12. Number of solutions on Pareto front with different algorithms and noise settings. Results are the average number over 20 repeated runs (MO-E-EQI with multiple model runs and $\beta = 0.6$).

The impact of four reaction variables was visualized in a three-dimensional plot for three main influencing factors (equivalent, flow rate and DMAP loading), as shown in Fig. 13(c). The sizes and colours of the data points represent STY and E-factor respectively, and temperature is represented using the linewidth of the points. At low equivalents, the E-factor was always low but STY was also low. When increasing the equivalent from moderate to high, more desired product was produced but at the same time more waste was generated, which resulted in a decrease of the E-factor but favoured STY. Large DMAP loading accelerated the reaction and increased STY but generated more waste. The influence of temperature mainly changed the reaction rate but was not that obvious. In addition, it can be noticed that data points were quite sparse in the high equivalent from 4 to 5 compared to the low equivalent from 1 to 3, where the algorithm was more explorative in the high equivalent region.

5. Conclusions

Multi-objective Bayesian optimization (MOBO) has been shown to be a useful tool for reaction development. In this study, we focus on finding the optimal reaction conditions under heteroscedastic noise using multi-objective Euclidian expected quantile improvement (MO-E-EQI). The algorithm was first compared with some recent noisy MOBO algorithms in an *in silico* study with multiple noise structures and magnitudes. Then MO-E-EQI was further implemented to guide experiments in a real case and a clear Pareto front was identified successfully.

The *in silico* study shows that noise does make a difference and affects the algorithm performances. High noise degrades the performance of all algorithms involved in the study. However, no significant difference was observed for different noise structures with linear and loglinear cases. For MOBO under noise, it is shown that metrics to evaluate algorithm performances can be problematic. Overall, MO-E-EQI shows robust performances in terms of HV distance and solution coverage compared to other algorithms. The selection of aggressive and non-aggressive criteria of MO-E-EQI can tune the number of solutions on the Pareto front based on the need.

In the experimental case study, an esterification reaction was selected with four continuous variables: equivalent, temperature, flow rate and catalyst loading. Two objective functions were set as space–time-yield and E-factor. MO-E-EQI was able to find the Pareto front efficiently and identified a clear trade-off between the two objectives. A set of Pareto front solutions generated with STY ranging from 2.66 g/(L·min) to 195.04 g/(L·min) while the E-factor ranged from 2.89 to 8.47, which left researchers to decide which condition to choose for further development. Notably, MO-E-EQI shows efficient performances under noise, however, this may be at a cost of increasing the number of experimental measurements by repetitions. It is therefore more suitable for automated experimental platforms or large-scale manufacturing where cheap sensor data is available.

For real-life applications of MOBO, sometimes the complexity of the problem and the noise levels are hard to estimate *a priori*. Most of the time, noise structures are more complicated than simplified Gaussian noise that is assumed in most studies and different assumptions on noise

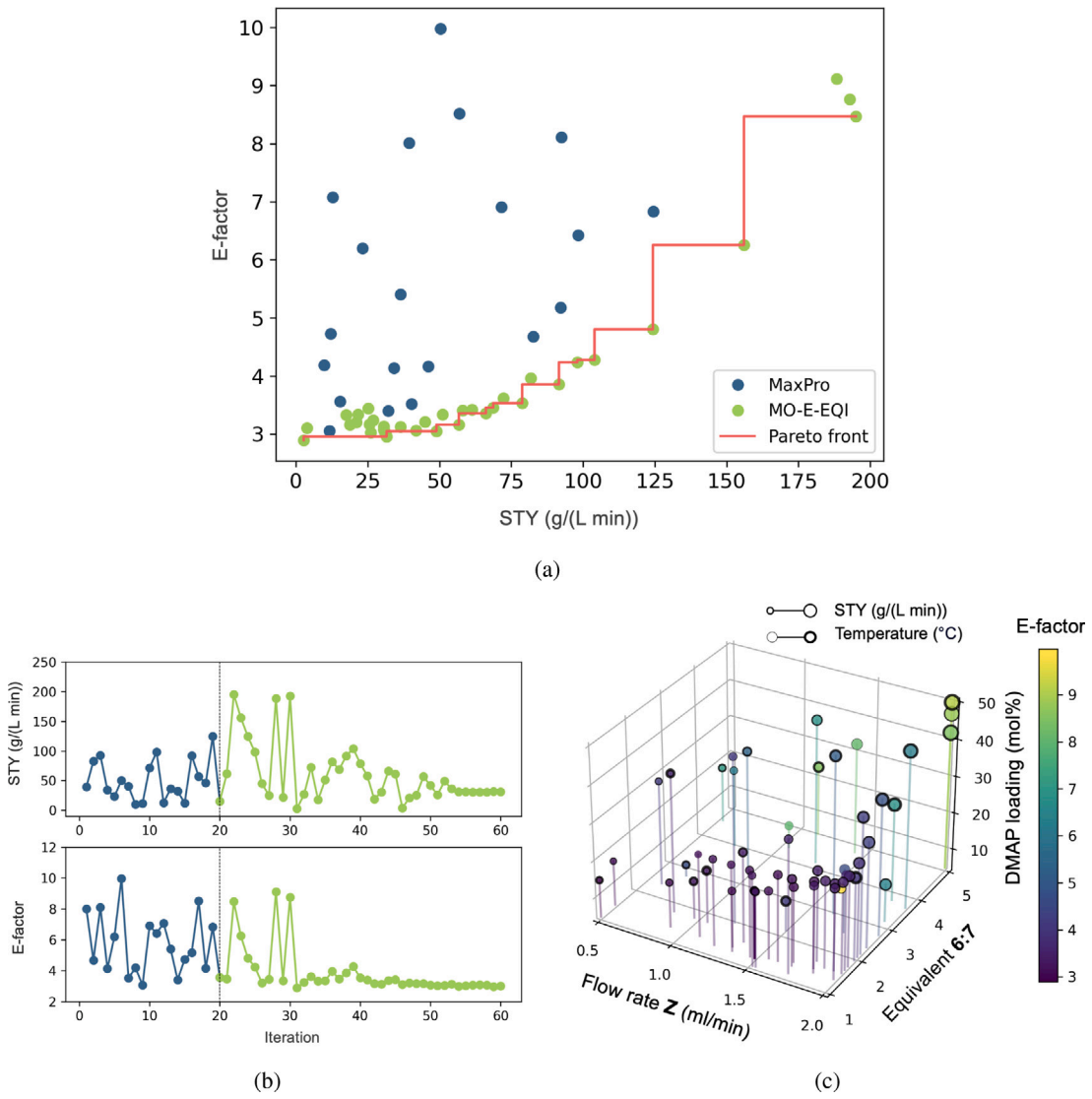

Fig. 13. Optimization results of a real-world esterification reaction. (a). Pareto front solutions; (b). Iterations of two objectives; (c). Three-dimensional optimization results versus variables.

structures need to be further considered. In addition, designing performance metrics for MOBO under noise is challenging but meaningful to ensure algorithm performances for robust applications.

CRediT authorship contribution statement

Jiyizhe Zhang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Daria Semochkina:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Naoto Sugisawa:** Writing – review & editing, Methodology, Investigation, Formal analysis. **David C. Woods:** Writing – review & editing, Conceptualization. **Alexei A. Lapkin:** Writing – review & editing, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was in part supported by EPSRC, United Kingdom funded project EP/S019472/1 "Chembots: digital-chemical-robotics to convert code to molecules and complex systems" and EPSRC, United Kingdom EP/W031019/1 "Bio-derived and Bio-inspired Advanced Materials for Sustainable Industries (VALUED)". The work was enabled by the Innovation Centre in Digital Molecular Technologies (*i*DMT), an ERDF co-funded project. The work was supported by the grant-in-aid for JSPS Fellows (22KJ1553 to N. Sugisawa).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compchemeng.2024.108983.

Data availability

Data will be made available on request.

References

Albert, R., Horwitz, W., 1997. A heuristic derivation of the horwitz curve. Anal. Chem. 69, 789–790.

Aldeghi, M., Häse, F., Hickman, R.J., Tamblyn, I., Aspuru-Guzik, A., 2021. Golem: an algorithm for robust experiment and process optimization. Chem. Sci. 12 (44), 14792–14807.

Ba, S., Joseph, V.R., 2018. MaxPro: Maximum projection designs. In: R Package Version 4.1-2.

Baker, E., Barbillon, P., Fadikar, A., Gramacy, R.B., Herbei, R., Higdon, D., Huang, J., Johnson, L.R., Ma, P., Mondal, A., Pires, B., Sacks, J., Sokolov, V., 2022. Analyzing stochastic computer models: A review with opportunities. Statist. Sci. 37 (1), 64–89.

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A.G., Bakshy, E., 2020. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. Adv. Neural Inf. Process. Syst. 33, 21524–21538.

Binois, M., Gramacy, R.B., Ludkovski, M., 2018. Practical heteroscedastic Gaussian process modeling for large simulation experiments. J. Comput. Graph. Stat 27 (4), 808–821.

Braconi, E., 2023. Bayesian optimization as a valuable tool for sustainable chemical reaction development. Nat. Rev. Methods Primers 3 (1), 74.

Bradford, E., Schweidtmann, A.M., Lapkin, A., 2018. Efficient multiobjective optimization employing gaussian processes, spectral sampling and a genetic algorithm. J. Glob. Optim 71 (2), 407–438.

Branke, J., 2023. Performance metrics for multi-objective optimisation algorithms under noise. arXiv preprint. https://arxiv.org/abs/2206.03301.

Daulton, S., Balandat, M., Bakshy, E., 2020. Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. Adv. Neural Inf. Process. Syst. 33, 9851–9864.

Daulton, S., Balandat, M., Bakshy, E., 2021. Parallel Bayesian optimization of multiple noisy objectives with expected hypervolume improvement. Adv. Neural Inf. Process. Syst. 34, 2187–2200.

Daulton, S., Balandat, M., nBakshy, E., 2023. Hypervolume knowledge gradient: a lookahead approach for multi-objective bayesian optimization with partial information. In: ICML. PMLR, pp. 7167–7204.

Daulton, S., Cakmak, S., Balandat, M., Osborne, M.A., Zhou, E., Bakshy, E., 2022. Robust multi-objective Bayesian optimization under input noise. In: ICML. PMLR, pp. 4831–4866.

Felton, K.C., Rittig, J.G., Lapkin, A.A., 2021. Summit: benchmarking machine learning methods for reaction optimisation. Chem.-Methods 1 (2), 116–122.

Fonseca, C.M., Fleming, P.J., 1995. An overview of evolutionary algorithms in multiobjective optimization. Evol. Comput. 3 (1), 1–16.

Gaefke, G., Enkelmann, V., Hoger, S., 2006. A practical synthesis of 1 4-diiodo-2, 5-bis(chloromethyl)benzene and 1 4-diiodo-2, 5-bis(bromomethyl)benzene. Synth. 2971–2973.

Giagkiozis, I., Fleming, P.J., 2014. Pareto front estimation for decision making. Evol. Comput. 22 (4), 651–678.

Gilmour, S.G., Trinca, L.A., 2012. Optimum design of experiments for statistical inference. J. R. Stat. Soc. C: Appl. Stat 61 (3), 345–401.

Hernández-Lobato, J.M., Hoffman, M.W., Ghahramani, Z., 2014. Predictive entropy search for efficient global optimization of black-box functions. Adv. Neural Inf. Process. Syst. 27.

Hone, C.A., Holmes, N., Akien, G.R., Bourne, R.A., Muller, F.L., 2017. Rapid multistep kinetic model generation from transient flow data. React. Chem. Eng. 2 (2), 103–108.

Huang, D., Allen, T.T., Notz, W.I., Zeng, N., 2006. Global optimization of stochastic black-box systems via sequential kriging meta-models. J. Global Optim. 34, 441–466.

Jalali, H., Van Nieuwenhuyse, I., Picheny, V., 2017. Comparison of kriging-based algorithms for simulation optimization with heterogeneous noise. European J. Oper. Res. 261 (1), 279–301.

J.C., Fromer., Coley, C.W., 2023. Computer-aided multi-objective optimization in small molecule discovery. Patterns 4 (2), 1–17.

Kandasamy, K., Krishnamurthy, A., Schneider, J., Poczos, B., 2018. Parallelised Bayesian Optimisation Via Thompson Sampling, vol. 84, PMLR, pp. 133–142.

Keane, A.J., 2006. Statistical improvement criteria for use in multiobjective design optimization. AIAA J. 44 (4), 879–891.

Knowles, J., 2006. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems, Vol. 10, No. 1. IEEE, pp. 50–66.

Letham, B., Karrer, B., Ottoni, G., Bakshy, E., 2019. Constrained Bayesian optimization with noisy experiments. Bayesian Anal. 14, 2.

Lewis, A., Mostaghim, S., Scriven, I., 2009. Asynchronous Multi-Objective Optimisation in Unreliable Distributed Environments.. Springer, Berlin, pp. 51–78.

Makarova, A., Usmanova, I., Bogunovic, I., Krause, A., 2021. Risk-averse heteroscedastic bayesian optimization. Adv. Neural Inf. Process. Syst. 34, 17235–17245.

Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Müller, K., Pedersen, T.L., Baston, D., Dunnington, D., 2023. Simple Features for R. R package version 1.0-15.

Picheny, V., Ginsbourger, D., 2014. Noisy kriging-based optimization methods: a unified implementation within the DiceOptim package. Comput. Statist. Data Anal. 71, 1035–1053.

Picheny, V., Ginsbourger, D., Richet, Y., Caplin, G., 2013. Quantile-based optimization of noisy computer experiments with tunable precision. Technometrics 55, 2–36.

Picheny, V., Moss, H., Torossian, L., 2022. Bayesian quantile and expectile optimisation. PMLR, pp. 1623–1633.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning, vol. 1, MIT Press, Cambridge MA.

Riquelme, N., Von Lücken, C., Baran, B., 2015. Performance metrics in multi-objective optimization. In: Latin Amer. Comput. Conf.. CLEI, IEEE, pp. 1–11.

Roustant, O., Ginsbourger, D., Deville, Y., 2021. DiceKriging: Kriging methods for computer experiments. In: R Package Version 1.6.0.

Semochkina, D., 2024a. MOEEQI: Multi-objective euclidian expected quantile improvement. R package. https://github.com/StatsDasha/MO-E-EQI.

Semochkina, D., 2024b. ParetoPal: Rshiny app for Pareto front comparison. https://statsdasha.shinyapps.io/ParetoPal/.

Semochkina, D., Forrester, A.I., Woods, D.C., 2024. Multi-objective optimization using expected quantile improvement for decision making in disease outbreaks. SIAM/ASA J. Uncertain. Quantif. http://dx.doi.org/10.1137/24M1633625, (in press).

Sheldon, R.A., Bode, M.L., Akakios, S.G., 2022. Metrics of green chemistry: Waste minimization. Curr. Opin. Green Sustain. Chem. 33, 100569.

Slattery, A., Wen, Z., Tenblad, P., Sanjosé-Orduna, J., Pintossi, D., den Hartog, T., Noël, T., 2024. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. Science 383 (6681), eadj1817.

Tu, B., Gandy, A., Kantas, N., Shafei, B., 2022. Joint entropy search for multi-objective Bayesian optimization. Adv. Neural Inf. Process. Syst. 35, 9922–9938.

Vel, Aravind Senthil, Cortes-Borda, Daniel, Felpin, Francois-Xavier, 2024. A chemist's guide to multi-objective optimization solvers for reaction optimization. React. Chem. Eng..

Wagner, T., Emmerich, M., Deutz, A., Ponweiser, W., 2010. On expected-improvement criteria for model-based multi-objective optimization. In: Proceedings of the International Conference on Parallel Problem Solving from Nature. Springer, pp. 718–727.

Wang, Y., Chen, T.-Y., Vlachos, D.G., 2021. NEXTorch: a design and Bayesian optimization toolkit for chemical sciences and engineering. J. Chem. Inf. Model. 61 11, 5312–5319.

Wang, Z., Ierapetritou, M., 2017. A novel surrogate-based optimization method for black-box simulation with heteroscedastic noise. Ind. Eng. Chem. Res. 56 (38), 10720–10732.

Wang, Z., Ierapetritou, M., 2018. Constrained optimization of black-box stochastic systems using a novel feasibility enhanced kriging-based method. Comput. Chem. Eng. 118, 210–223.

Wang, X., Jin, Y., Schmitt, S., Olhofer, M., 2023. Recent advances in Bayesian optimization. ACM Comp. Surv. 55 (13), 1–36.

Wentzell, P.D., Brown, C.D., 2000. Signal processing in analytical chemistry. Encyc. Anal. Chem. 11, 9764–9800.

William, Horwitz, 1982. Evaluation of analytical methods used for regulation of foods and drugs. Anal. Chem. 54 (1), 67–76.

Zitzler, E., Thiele, L., 1999. IEEE transactions on Evolutionary Computation, IEEE transactions on Evolutionary Computation, vol. 3 (no. 4).257–271,