A novel approach to identify optimal and flexible operational spaces for product quality control

Sam Kay [a], Mengjia Zhu [a], Amanda Lane [b], Jane Shaw [b], Philip Martin [a], Dongda Zhang [a,*]

[a] *Department of Chemical Engineering, University of Manchester, Manchester, M13 9PL, UK*
[b] *Unilever R&D Port Sunlight, Bromborough Road, Bebington, Wirral, CH63 3JW, UK*

ABSTRACT

Maintaining consistent product quality under uncertainty is essential for chemical industries, where minimising waste and meeting strict quality standards impacts overall performance. However, existing control strategies, relying on set-point control, often prove inefficient for large process uncertainties. Operational space control emerges as a promising solution, identifying reliable operational regions and enhancing process flexibility, which is particularly advantageous when achieving rigid control is challenging. This study presents a framework for extensively screening design spaces to identify optimal operational regions that satisfy quality constraints while accommodating various uncertainties prevalent in manufacturing settings. By integrating machine learning based clustering algorithms with dynamic optimisation within a real-world formulation process, the discovered operational spaces were shown to be optimal, flexible and reliable, leading to reduced batch times and energy consumption. Overall, this work outlines operational space methods as a novel advancement in modern product quality control, paving the way for resilient and economically viable industrial process designs.

1. Introduction

Product quality control is crucial for industries such as pharmaceuticals, specialty chemicals, and formulations, where stringent regulations of product quality exist to ensure safety, reliability, and good economic performance (Hicks et al., 2021). Additionally, due to the current sustainability drive, local and global governmental initiatives to reduce environmental impact have led to increasing demands and regulations being introduced; alongside rising operational costs, this has intensified the demand for tighter product quality control.

To remain competitive and combat increasing sustainability demands, there is a pressing need to develop optimised process flow diagrams (PFD) that address inefficiencies, reduce waste (Amrih and Damayanti, 2022), and enhance operational flexibility (Hill, 2007). Although the use of the term Process Flow Diagram (PFD) may vary in meaning across industrial sectors, we will be defining as is conventionally accepted in the formulation industries. A PFD is a visual representation of a process that specifies key process state variables and their required set-points to ensure consistent and controlled operation. Traditionally, manufacturing processes are operated and controlled based on the rigid state profiles (i.e. a discrete specification of state variable set-points at different stages in the process) specified in the PFD determined through numerous experimental trials

(Forbes et al., 2015; Efheij et al., 2019), which, whilst straightforward to implement and effective for stable operations, lacks the flexibility needed to manage uncertainties that may arise during process operation. Furthermore, as a consequence of its rigidity, set-point control often results in high energy consumption (Asad et al., 2017). Even in other classic set-point approaches that account for uncertainties through real-time feedback control, large sensor networks with rapid data processing are required which can be expensive and impractical hence, accommodation of uncertainties may be limited. If unaccounted for, or not appropriately accounted for, it may be experienced that these uncertainties directly lead to large deviations in the key performance indicator (KPI) and consequently, batch rejections, which is reported widely within the literature (Novaraa and Henning, 2018; Wan et al., 2012; Geletu et al., 2013). Thus, there exists a clear need for more advanced and robust optimal control strategies to maintain product quality.

So far, an extensively applied approach for effectively handling uncertainty is stochastic optimisation, which leverages probability distributions to represent uncertainties, aligning closely with real-world process variability (Zheng et al., 2014). Stochastic optimisation has seen use across a wide range of applications including complex, non-linear systems (Chen et al., 2023), where classical robust optimisation approaches, such as the use of interval uncertainty and ellipsoidal bounds often lead to overly conservative solutions and high computational costs

---

* Corresponding author.
  E-mail address: dongda.zhang@manchester.ac.uk (D. Zhang).

(Ben-Tal and Nemirovski, 2002; University, 2023). In practice, solving over the entire probability distribution is intractable so, techniques that aim to approximate the distribution, like Monte Carlo and scenario-based methods, are employed (Homem-de Mello and Bayraksan, 2014). Scenario tree optimisation is a subset of stochastic optimisation that further refines decision-making by transforming complex probabilistic problems into manageable deterministic ones, using a tree of possible uncertainty scenarios (Silvente et al., 2019; Hill, 2007). In this manner, uncertainties are propagated across the entire process, with multiple different realisations of these uncertainties being considered. Each full trajectory (scenario) of the process is therefore defined by a unique combination of uncertainty realisations (nodes), of which many can be generated to describe the overall probability distribution (i.e. scenario tree).

In addition to scenario tree optimisation approaches, flexibility analysis is often used to examine the performance of a predefined PFD. Flexibility analysis aims to identify the combination of the largest uncertainties that can be accommodated by a feasible set of state profiles so that all specified constraints are satisfied (Lima et al., 2010; Wang et al., 2016; Forster et al., 2024). Flexibility-based methods find a wide array of applications throughout the literature, being especially prominent within process engineering and reaction engineering. Many examples of which are found in Raspanti et al. (2000), where different flexibility approaches are compared for a multitude of process systems engineering case studies, taking note of the advantages and disadvantages of each. Another example is seen in the work of Bremer et al. (2017), where the flexibility of different set-points was compared for dynamic hot spot control of a carbon dioxide methanation reactor under the worst-case input feed conditions. In (Bremer and Sundmacher, 2019), this study was extended to further improve operational flexibility by applying stabilising control to unstable operating points. While flexibility analysis can establish a reliable PFD that can account for different sources of uncertainty (Swaney and Grossmann, 1985), strictly following this PFD in real operations can be challenging. Moreover, it is usually the case that optimality of the PFD is overlooked in favour of its reliability hence, in many cases, it may be considered overly conservative.

In the literature, both scenario tree and flexibility-based methods have been widely integrated into algorithms for determining optimal and reliable state set-points to combat uncertainties (Lucia et al., 2013) with reported success. Nonetheless, within the fine chemical and pharmaceutical industries it is frequently the case that the product KPI is sensitive to the optimised set-points, hence even small disturbances to the set-points may result in product quality violations. In practice, it is highly challenging to prevent such disturbances from occurring due to the presence of significant feedstock variability, human error and system inconsistencies; this makes rigid adherence to the set-points infeasible, so batch rejections are likely to be observed (Gattu and Zafiriou, 1999). Furthermore, although the set-points may be optimal with respect to the KPI, little to no consideration is given to the excessive energy expenditure necessary to operate sufficiently close to these predefined state profiles hence, such approaches are unfavourable for the current drive towards sustainable and resource-efficient operations (Bruns et al., 2020). To improve upon traditional set-point based control approaches, it becomes necessary to extend the concept of flexibility from that of reliable set-points, to that of reliable operational spaces which provide greater flexibility to the manufacturing operation.

An operational space-based control strategy directly regulates control and state variables to achieve desired outcomes within a defined operational space, rather than focusing on individual system inputs or maintaining fixed set-points. It has found considerable use within the literature in the chemical and pharmaceutical sectors in recent years, where it is often referred to as design space identification, or as an extension to flexibility analysis. In the work of Forster et al. (2024), an operational space methodology was proposed to facilitate an extension to formal flexibility analysis and was proven effective for a case study

pertaining to the temperature control of a fed-batch bioreactor producing ethanol subject to parameter uncertainties where the operational region ensured minimum ethanol concentration constraints were met. Another example is described in the work of Mortier et al. (2016) which focuses on a pharmaceutical freeze drying process that is known to incur high costs and large energy consumption, primarily due to the need to achieve stability in the product (KPI). A dynamic operational space, robust to parameter uncertainties, was designed for shelf temperature and chamber pressure, resulting in better economic performance and lower drying times. Other pharmaceutical applications are compared and their limitations are discussed in Djuris and Djuric (2017), offering suggestions to integrate operational spaces with process analytical technology (PAT) measurements for improved control and monitoring of system state, uncertainty and KPI.

Despite the significant advantages operational space control has to offer, there remains certain limitations in current methods yet to be addressed. Several notable challenges include: (i) the consideration of uncertainty is usually restricted to parameter uncertainty alone; (ii) optimality of the space is overlooked, with focus instead being on the design of specific optimal set-points; (iii) the determination of operational spaces is frequently inefficient, largely relying on computationally expensive sampling techniques; (iv) and finally, there exists no work focusing on the systematic identification of multiple spaces, which may exist for the control of a given process and might perform equally well. On top of these challenges, to the best of our knowledge, operational space methods have not been applied before into the formulation industries. However it is believed that the use of flexible operating spaces may provide significant advantages in bypassing major challenges within the sector, such as difficulties in implementing real-time product quality control and monitoring and precisely maintaining predefined state set-points.

Therefore, in this paper, the innovation is twofold; firstly, we propose a novel scenario tree optimisation framework that integrates flexible operational spaces with dynamic optimisation to achieve optimal performance across the entire design space (i.e. identifying multiple optimal and flexible operational regions), paving the way for advancements in modern process product quality control. Secondly, we pioneer the first case study investigating the effectiveness of such technique in formulation settings. The structure of the paper herein is ordered as follows. Sections 2 and 3 outline the operational space identification methodology and formulation process case study. Section 4 provides a thorough discussion of the results, and finally, Section 5 provides a conclusion to the research findings.

2. Methodology

Generally, we divide operational spaces into 3 classifications, that is: the feasible space, which is the multidimensional range of operating conditions that can satisfy process constraints; the flexible space, a subset of the feasible space which is also robust to uncertainties; and the optimal space, another subset of the feasible space in which the process is operated optimally for some pre-defined objective. In this study, we aim to propose a novel methodology for the characterisation of the intersect between the flexible and optimal spaces such that the discovered solution space contains only solutions both optimal in operation and robust to uncertainties. Furthermore, since different optimal operating conditions may exist, we aim to isolate any distinct optimal operating regions within the identified design space. Each optimal operating region is characterised by a unique optimal state profile (or a set of optimal state profiles if there are multiple state variables) and its representative bounds. In this paper, we adopt interval representation, where spaces are defined by optimal operating conditions bounded by upper and lower limits. It is then possible to rank the different operating regions contingent to some preferred metrics and heuristics (e.g., whether one is easier to achieve or maintain in practice).

Step 1

> **Sample a suitable number of scenarios based on uncertainty**

Step 2

> **Generate a diverse range of optimal state profiles**

Step 3

> **Cluster the state profiles to search for distinct optimal regions**

Step 4

> **Identify nominal state profiles and corresponding bounds for each optimal region**

> **Return violating batches**

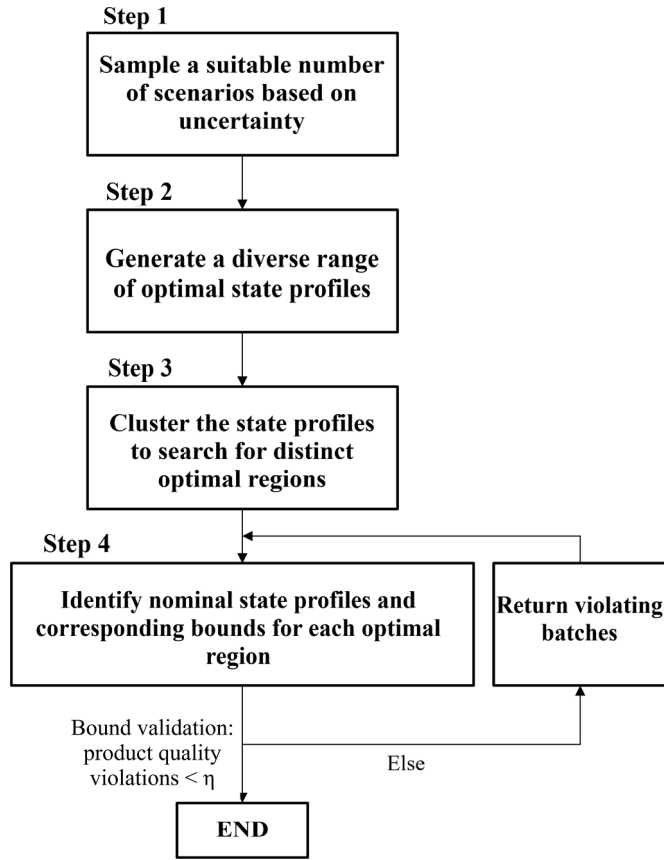Bound validation: product quality violations < η

Else

> **END**

Fig. 1. Proposed methodology for optimal space identification, terminated under the criteria that the number of product quality violations observed is less than η.

The overall methodology can be summarised into 4 steps, which are shown in Fig. 1 and elaborated in the following subsections.

2.1. General problem formulation

We assume the availability of a mathematical model to represent the dynamics of the process, where we divide the state variables into two categories: (i) $\boldsymbol{x}_{\text{flexible}}(t) \in \mathbb{R}^{n_{\text{flexible}}}$, with $t \in [0, t_f]$ represents a specific time point within the operational time horizon, denoting the vector of state variables whose values are flexible and can be directly regulated by the control inputs (e.g., maintain the temperature (a flexible state variable) of a reactor by adjusting the coolant flowrate (a control input)); (ii) $\boldsymbol{x}_{\text{dept}}(t) \in \mathbb{R}^{n_{\text{dept}}}$, denoting the vector of dependent variables whose values can be influenced by $\boldsymbol{x}_{\text{flexible}}$ and control inputs but cannot be directly altered by the control inputs. For example, the concentration of a product (a dependent state variable) may increase when both the temperature (a flexible state variable) and the concentration of an intermediate state (another dependent state variable) rise. However, unlike flexible state variables, it cannot be directly regulated through a control input. The dynamic model is expressed as follows:

$$\dot{\boldsymbol{x}}_{\text{flexible}}(t) = \boldsymbol{F}_{\text{flexible}}(t, \boldsymbol{x}_{\text{flexible}}(t), \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{u}(t))$$
$$\dot{\boldsymbol{x}}_{\text{dept}}(t) = \boldsymbol{F}_{\text{dept}}(t, \boldsymbol{x}_{\text{dept}}(t), \boldsymbol{x}_{\text{flexible}}(t), \boldsymbol{\theta}, \boldsymbol{\xi}). \tag{1}$$

Here, $\boldsymbol{u}(t) \in \mathbb{R}^m$ is the vector of control variables; $\boldsymbol{\theta} \in \mathbb{R}^p$ is the vector of fixed parameters; and $\boldsymbol{\xi} \in \mathbb{R}^r$ denotes the vector of uncertain parameters. In this paper, we assume that each source of uncertainty, ξ_i, for $i = 1, \ldots, r$, is modelled as a random variable with a distinct Gaussian distribution, i.e., $\xi_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, with μ_i represents the mean setted to the nominal value, while σ_i is the standard deviation that reflects the expected operational variability. Both parameters, μ_i and σ_i, are assumed

to be known in advance. The objective is to determine a flexible, optimal, and practical operational space for $\boldsymbol{x}_{\text{flexible}}(t)$, ensuring consistent satisfactory achievements of KPIs (e.g., product quality) and maximum process efficiency notwithstanding the uncertainties introduced by $\boldsymbol{\xi}$. We note that hereinafter, operational space is used to denote the identified flexible state space of $\boldsymbol{x}_{\text{flexible}}(t)$.

2.2. Sampling uncertainty

We utilise scenario tree analysis, as discussed in Section 1 to capture the range of uncertainties that may arise during processing. The scenario tree is constructed by sampling each source of uncertainty, ξ_i, according to its respective distribution. In practice, there exist multiple ways in which a scenario tree may be structured; the two most generic representations being depicted in Fig. 2.

The scenario tree structure described in Fig. 2(a), is often referred to as a scenario fan (Xu et al., 2012), where the number of considered uncertainties is equal for all uncertainties, ξ_1, ξ_2, and ξ_3. In contrast, an alternative branching scenario tree representation is shown in Fig. 2(b), where a different number of realisations is made for each of the uncertainties. To sample across a large number of uncertain scenarios using the scenario fan can be inefficient, thus much work within the literature focuses on the conversion of scenario fans into branching scenario trees (Pranevicius and Sutiene, 2007; Chen and Yan, 2018) (see Fig. 2b), reducing the number of scenarios that must be considered. Although useful in sparing computational expense, the use of branching scenario trees may introduce bias in how the uncertainties are represented, which often leads to a particular focus on sampling at the later stages, hence in this work we make use of the scenario fan structure to avoid any such assumptions about the uncertainties. This is especially useful in the cases where the dimension of uncertain parameters is not large, as is the case for this work.

Accurately representing uncertainty is essential for achieving robust design, control and optimisation of a process. Therefore, in theory a sufficiently large number of samples should be drawn from each uncertainty source. However, this consideration must be balanced against the increased computational load that accompanies a larger scenario tree particularly if the underlying process model is highly nonlinear and complex. In practice, a compromise is often found between these two factors. Once the scenario tree is realised, all optimisation approaches are conducted across the entire tree, ensuring that any defined constraints are upheld for each considered uncertainty realisation. In the event of high dimensional uncertainty spaces, one should consider the number of uncertainty samples necessary to well represent the uncertainty space, the computational power of the available equipment any use of parallelisation, as well as the use of scenario reduction techniques to overcome the otherwise high computational expense.

2.3. Identification of the optimal region

Since there can be multiple state profiles of $\boldsymbol{x}_{\text{flexible}}(t)$ which can provide similar performance that satisfy the process constraints and maintain KPIs within specifications, in Step 2, as illustrated in Fig. 1, we focus on refining the broad process design space to identify promising operational regions that are likely to encompass the majority of high-performing state profiles of $\boldsymbol{x}_{\text{flexible}}(t)$. Together, these will form the optimal operational region, if designed successfully (i.e., state profiles within the space will meet a predefined standard of process optimality while respecting process constraints for any uncertainties represented in the scenario tree). Specifically, for each set of high-performing state profiles, our aim is to minimise a process cost, C_{avg}, which is averaged over the process cost of all the scenarios considered within the scenario tree, while maintaining the calculated KPI value within the specifications and close to the target value specified in the PFD. In practice, the chosen cost function is highly dependent on the requirements of industry, particularly in how conservative they wish to be with respect to
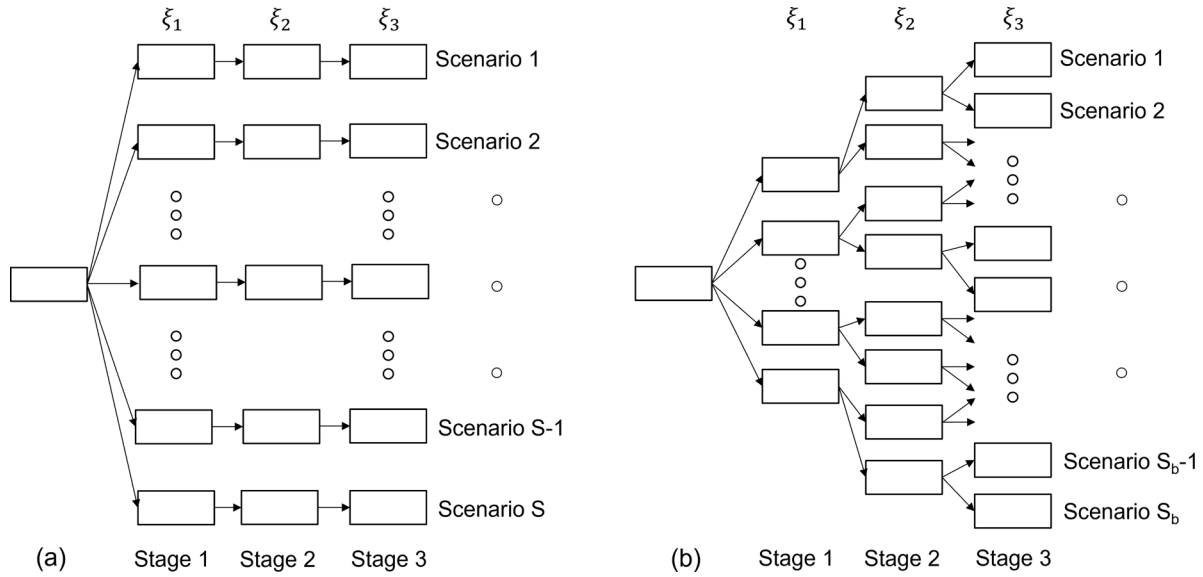
Fig. 2. General representations of a scenario tree, with each realisation of uncertainty, ξ_i, where in (a), a scenario fan is shown, and in (b), a branching scenario tree is shown.

uncertainties. In this work, even though the average scenario cost may be overly conservative due to its equal weighting associated to each potentially uncertain scenario, it is chosen to reflect the formulation industries requirements to accommodate large operational uncertainties. In other scenarios, where it is not desired to be overly conservative for the system at hand, then using alternative objectives such as the expected value may prove a better option.

The operational time horizon is discretised into N intervals. To reduce the control efforts, we assume $\boldsymbol{x}_{\text{flexible}}(t)$ is held constant within each interval (that is, piecewise constants), i.e., $\boldsymbol{x}_{\text{flexible}}(t) = \boldsymbol{x}_k^{\text{flexible}}$, for $t \in [t_k, t_{k+1}]$ and $k = 0 \dots, N-1$, where $t_k = k\Delta t$ and $\Delta t = \frac{t_f}{N}$. Also, due to physical limitations imposed on the control actions (i.e., $\boldsymbol{u}(t)$), $\boldsymbol{x}_k^{\text{flexible}}$ are also bounded within the range of $[\underline{\boldsymbol{x}}_k^{\text{flexible}}, \overline{\boldsymbol{x}}_k^{\text{flexible}}]$, for $k = 0, \dots, N-1$. In other words, if the flexible states, $\boldsymbol{x}_k^{\text{flexible}}$, are assigned values within this range, it is possible to adjust the control inputs to reach these assigned values. The process can be subject to constraints, represented by $\boldsymbol{g}(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}})$, for $s = 1, \dots, S$, where S denotes the number of uncertain scenarios in the realised scenario tree. Additionally, it is presumed that the initial states ($\boldsymbol{x}_0^{\text{dept}}$) are known.

We can define the optimisation problem as follows to identify the optimal state trajectories of $\boldsymbol{x}^{\text{flexible}}$:

$$
\begin{aligned}
&\min_{\substack{\boldsymbol{x}_k^{\text{flexible}} \\ k=0,\dots,N-1}} && \lambda_1 \cdot \text{obj}_1 + \lambda_2 \cdot C_{\text{avg}} \\
&\text{s.t.} && \text{for } k = 0, \dots, N-1, \quad s = 1, \dots, S, \\
& && \text{obj}_1 = \frac{1}{S} \sum_{s=1}^{S} (y_{\text{KPI},s} - y_{\text{KPI,T}})^2 \\
& && C_{\text{avg}} = \frac{1}{S} \sum_{s=1}^{S} C_s \\
& && y_{\text{KPI},s} = h\big(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}}\big) \\
& && \boldsymbol{x}_{k+1,s}^{\text{dept}} = \boldsymbol{f}_{\text{dept}}\big(\Delta t, \boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}}, \boldsymbol{\theta}, \boldsymbol{\xi}_s\big) \\
& && \boldsymbol{g}\big(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}}\big) \leq 0 \\
& && \boldsymbol{x}_0^{\text{dept}} = \boldsymbol{x}^{\text{dept}}(0) \\
& && \underline{\boldsymbol{x}}_k^{\text{flexible}} \leq \boldsymbol{x}_k^{\text{flexible}} \leq \overline{\boldsymbol{x}}_k^{\text{flexible}}, \quad \underline{y}_{\text{KPI}} \leq y_{\text{KPI},s} \leq \overline{y}_{\text{KPI}},
\end{aligned}
\tag{2}
$$

where λ_1 and λ_2 are weighting parameters for each objective; obj_1 is the average sum-of-squared deviations from KPIs across all the scenarios;

and C_s is the process cost of each scenario. $h(\cdot)$ is the functional transformation between the process states and the KPIs, with $y_{\text{KPI},s}$ and $y_{\text{KPI,T}}$ being the resulting KPI value for Scenario s and the target KPI value, respectively. $\boldsymbol{f}_{\text{dept}}(\cdot)$ represents the numerical integration of the dependent states. And $\boldsymbol{g}(\cdot)$ constitutes the process inequality constraints (note, equality constraints can be represented as two inequality constraints).

To identify a diverse set of high-performing state trajectories and explore a wide potential operational space, once the initial optimisation of Problem (2) is complete, new high-performing state profiles can be determined by re-optimising Problem (2) with an added penalty function, p, in the objective function (i.e., $\lambda_1 \cdot \text{obj}_1 + \lambda_2 \cdot C_{\text{avg}} + p$). The added penalty aims to maximise differences between existing and new state trajectories. This penalty may take the form:

$$
p = \lambda_3 \cdot \sum_{d=1}^{D_{\text{curr}}} \sum_{j=1}^{n_{\text{flexible}}} \sum_{k=0}^{N-1} \frac{\left(x_{\text{flexible}\,j,k} - x_{\text{flexible}\,j,k,d}^*\right)}{\Delta \overline{x}_{\text{flexible}\,j,k}},
\tag{3}
$$

where λ_3 is the weighting parameter to control the penalty function's strength, with larger values promoting greater diversity among solutions and, thus, more distinct optimal regions. D_{curr} is the number of previous iterations. Here, $x_{\text{flexible}\,j,k,d}^*$ represents the discretised optimal state value from an identified state trajectory, whilst $x_{\text{flexible}\,j,k}$ is the state value in the current optimisation. $\Delta \overline{x}_{\text{flexible}\,j,k}$ denotes the maximum range of $x_{\text{flexible}\,j,k}$ across identified trajectories, normalising the values to account for state variables across different scales.

In the application of this approach, and with a sufficient number of iterations, D_{max}, one can screen the entire design space to identify a large number of high-performing state profiles that satisfy all feasibility constraints under uncertainty. However, due to the inherent complexity of the dynamic optimisation problem, in practice the user must decide on the number of iterations to perform. A smaller number of iterations may yield only a limited set of high-performing state profiles, while a larger number could become prohibitively computationally expensive.

2.4. Clustering the identified optimal regions

After completing Step 2 (see Fig. 1), clustering algorithms are employed to group the high-performing state profiles into different regions, each one representing a potential optimal operational space. Clustering algorithms, which use unsupervised machine learning, are designed to handle clusters of varying densities and shapes. To ensure reliable results, it is essential to validate and analyse the identified clusters across

multiple algorithms, focusing on their characteristics and differences in relation to the features of high-performing state profiles that fall within each cluster (Anand and Kumar, 2022). In this study, several machine learning clustering algorithms, including: k-means (Na et al., 2010), DB-SCAN (density-based spatial clustering of applications with noise) (Khan et al., 2014), and spectral clustering (Jia et al., 2014), were utilised and compared to ensure the consistency of the recommended number of clusters.

The k-means algorithm is an extensively used simple and effective technique, the objective of which is to minimise the summed distance between the cluster centres and any available data points (Na et al., 2010) for a given number of clusters, thus determining their optimal positions. As one of the most popular algorithms, k-means has found broad application, from categorisation of crude oils based on their physiochemical properties (Sancho et al., 2022), to separation of waste water treatment data into different regions for improved modelling of the chemical oxygen demand (Ay and Kisi, 2014). For a more detailed explanation of the algorithm, interested readers are referred to Ikotun et al. (2023).

The second algorithm used was DBSCAN, which is a density-based method designed to cluster data points through density-based connectivity analysis (Khan et al., 2014). A cluster is formed if there are enough points within a neighbourhood of a given radius, and this algorithm is particularly effective at handling irregular shapes and outliers. Additionally, DBSCAN is able to suggest the optimal number of clusters, which is not possible using k-means or spectral clustering. Due to these advantages, DBSCAN is commonly applied to complex systems such as in Rovira et al. (2022), where it is used to distinguish different regions of a reactor using a lower dimensional representation of key input variables. Another examples is provided in Lörzing et al. (2024), where DBSCAN is used to identify anisotropic clusters for discovery of underlying structural organisation in cells. Interested readers are directed to Bushra and Yi (2021) for further description of the algorithm.

The final algorithm compared was spectral clustering, which is another promising technique for describing clusters not well-separated and with complex shapes (Rodriguez et al., 2019). This technique depends on the construction of an affinity matrix to determine the similarity between two data points, for which the eigenvalues and eigenvectors can be used to divide the data into different clusters. In the literature spectral clustering is typically used for systems with complex relations such as in Lin et al. (2021), for which different molecular junction conformations are analysed in vast quantities of junction break data to fully explore the information hidden in largely unexplored datasets. A further application is shown in Fujiwara et al. (2011) for process monitoring of manufacturing processes by clustering data of similar production devices, removing the need for statistical modelling of each process. For a more comprehensive explanation of the algorithm, readers are referred to Von Luxburg (2007).

Overall, in this work, the use of k-means, DBSCAN and spectral clustering together allows us to cover a wide range of possible cluster properties, with each method accounting for potential drawbacks of the rest.

2.5. Establishing nominal state profiles and bounds

In Steps 2 and 3 of Fig. 1, approximations for some number of unique optimal operational regions are established. At this stage, we need to refine these regions by determining the nominal state profiles for each of these regions along with their corresponding lower and upper bounds for each flexible state variable. In other words, an optimal operational region for a specific flexible state variable is defined by its nominal profile and associated upper and lower bounds, such that remaining within these bounds ensures feasibility under the considered uncertainties.

The nominal profiles (i.e. optimised state variable set-points at different stages in the process) for the flexible states are identified with a dynamic optimisation formulation outlined in Problem (2), with the search space now confined to the clustered region identified in Step 3

(see Fig. 1). This ensures that the solution lies within the appropriate cluster; thus, $\underline{x}_k^{\text{flexible}}$ and $\overline{x}_k^{\text{flexible}}$ in Problem (2) are altered to represent the characteristics of the cluster (i.e. specific lower and upper bound of the cluster). Upon obtaining the nominal profiles, we propose a two-step algorithm to independently find the corresponding upper and lower bounds by maximising the distance between them while continuing to meet process constraints. The optimisation strategy for finding the lower bound is shown in Eq. (4):

$$\max_{\substack{\boldsymbol{x}_k^{\text{flexible,lb}} \\ k=0,\dots,N-1}} \min_{j=1}^{n_{\text{flexible}}} \min_{k=0}^{N-1} \underline{w}_{j,k} \cdot \left(x_{j,k}^{\text{flexible,nominal}} - x_{j,k}^{\text{flexible,lb}}\right)$$

$$\begin{aligned}
\text{s.t.} \quad & \text{for } k = 0,\dots,N-1, \quad s = 1,\dots,S, \\
& \underline{w}_{j,k} = \frac{1}{x_{j,k}^{\text{flexible,nominal}} - \underline{x}_{j,k}^{\text{flexible}}} \\
& y_{\text{KPI},s} = h\left(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}}\right) \\
& \boldsymbol{x}_{k+1,s}^{\text{dept}} = \boldsymbol{f}_{\text{dept}}\left(\Delta t, \boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible,lb}}, \boldsymbol{\theta}, \boldsymbol{\xi}_s\right) \\
& \boldsymbol{g}\left(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible,lb}}\right) \leq 0 \\
& \boldsymbol{x}_0^{\text{dept}} = \boldsymbol{x}^{\text{dept}}(0) \\
& \underline{x}_k^{\text{flexible}} \leq \boldsymbol{x}_k^{\text{flexible,lb}} \leq \boldsymbol{x}_k^{\text{flexible,nominal}}, \quad \underline{y}_{\text{KPI}} \leq y_{\text{KPI},s} \leq \overline{y}_{\text{KPI}},
\end{aligned}$$

(4)

where Problem (4) aims to maximise the minimum distance between the nominal flexible state, $x_{j,k}^{\text{flexible,nominal}}$ and the lower bound of flexible state, $x_{j,k}^{\text{flexible,lb}}$, across all flexible states variables and timesteps. The normalisation parameter, $\underline{w}_{j,k}$, normalises the contribution of different flexible states to the objective function. By designing the objective function to maximise the minimum distance between $\boldsymbol{x}_k^{\text{flexible,nominal}}$ and $\boldsymbol{x}_k^{\text{flexible,lb}}$, we aim to distribute the bounds as evenly as is attainable across the full trajectory for each flexible state variable, reducing the likelihood of requiring precise set-point control at any single time-step.

The upper bounds ($\boldsymbol{x}_k^{\text{flexible,ub}}$) are derived using a similar approach, as follows:

$$\max_{\substack{\boldsymbol{x}_k^{\text{flexible,ub}} \\ k=0,\dots,N-1}} \min_{j=1}^{n_{\text{flexible}}} \min_{k=0}^{N-1} \overline{w}_{j,k} \cdot \left(x_{j,k}^{\text{flexible,ub}} - x_{j,k}^{\text{flexible,nominal}}\right)$$

$$\begin{aligned}
\text{s.t.} \quad & \text{for } k = 0,\dots,N-1, \quad s = 1,\dots,S, \\
& \overline{w}_{j,k} = \frac{1}{\overline{x}_{j,k}^{\text{flexible}} - x_{j,k}^{\text{flexible,nominal}}} \\
& y_{\text{KPI},s} = h\left(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible}}\right) \\
& \boldsymbol{x}_{k+1,s}^{\text{dept}} = \boldsymbol{f}_{\text{dept}}\left(\Delta t, \boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible,ub}}, \boldsymbol{\theta}, \boldsymbol{\xi}_s\right) \\
& \boldsymbol{g}\left(\boldsymbol{x}_{k,s}^{\text{dept}}, \boldsymbol{x}_k^{\text{flexible,ub}}\right) \leq 0 \\
& \boldsymbol{x}_0^{\text{dept}} = \boldsymbol{x}^{\text{dept}}(0) \\
& \boldsymbol{x}_k^{\text{flexible,nominal}} \leq \boldsymbol{x}_k^{\text{flexible,ub}} \leq \overline{\boldsymbol{x}}_k^{\text{flexible}}, \quad \underline{y}_{\text{KPI}} \leq y_{\text{KPI},s} \leq \overline{y}_{\text{KPI}}.
\end{aligned}$$

(5)

This two-step algorithm produces upper and lower bounds that account for uncertainties whilst simultaneously respecting process constraints, providing an optimal operational region for which one may relax control of the design variables whilst achieving desired process performance and meeting desired product quality specifications. Although the process optimal performance target (process cost) is not explicitly considered in Eqs. (4) and (5), the use of $\underline{x}_k^{\text{flexible}}$ and $\overline{x}_k^{\text{flexible}}$ identified for each cluster allows for a high probability that the process still operates close to its optimal performance.

As this two-step algorithm independently determines the upper and lower bounds, each is guaranteed to comply with constraints individually. Nevertheless, it is not ensured that all state profiles sampled within

these bounds will also act accordingly. Consequently, it becomes essential that these bounds are validated by randomly sampling state profiles between them and recording instances that violate constraints. The recorded samples are then reintroduced into the framework through implementing a penalty function in the objective function of Eq. (4) and Eq. (5), effectively narrowing the bounds so that violating state profiles are unable to be sampled. The penalty function, q, is defined as follows:

$$q = \sum_{z=1}^{c_v} Q_z, \text{ with } Q_z = \begin{cases} 1 & \text{if } \forall k : \boldsymbol{x}_k^{\text{flexible,lb}} \leq \boldsymbol{x}_{k,z}^{\text{flexible}} \leq \boldsymbol{x}_k^{\text{flexible,ub}} \\ 0 & \text{Otherwise,} \end{cases} \quad (6)$$

where c_v is the total number of recorded constraint-violating state profiles ($\boldsymbol{x}_{k,z}^{\text{flexible}}$) from the recordings made in the validation step; and Q_z is the penalisation parameter, which equals 1 if $\boldsymbol{x}_{k,z}^{\text{flexible}}$ can be sampled from the identified region, and 0 if not. As shown in Fig. 1, the process is terminated once the violation rate of the validation set is less than or equal to a specific threshold, η. In reality, considering interdependence between operational variables may expand the discovered operational spaces, however taking such relationships into account on an industrial process could cause practical decision making to become convoluted for operators to achieve and maintain. Therefore, this is not taken into account in this work, instead favouring the two step dynamic optimisation followed by subsequent refinements of the bounds.

The formulation of the penalty described in Eq. (6) is advantageous as it is an effective brute force approach to ensure improvements in the violation rates observed when sampling within the optimal operational region, and is generalisable across other systems. However, the specific formulation shown may risk overly conservative reductions in bound areas hence, for simpler systems it may be preferred to make use of a more elegant penalty function or directly reoptimise the identified bounds over the problematic uncertain scenarios. In practice, therefore, one should consider a trade-off between practicality of the solution, and its conservativeness.

3. Case study

To demonstrate the proposed framework described in Sections 2.2–2.5, an in-silico case study was developed based on a dynamic model constructed to simulate a real industrial formulation process, for which a simplified schematic of the main mixing unit is shown in Fig. 3. The case study involves a multiphase non-Newtonian liquid product formulation, produced through a batch mixing process in which the PFD is defined via a series of ingredient additions made by process operators, alongside a set of processing conditions such as temperature and shear rate. The product KPI is difficult to control due to the complexity of the multiphase mixing mechanisms and cannot be measured online. Therefore, identifying a suitable operational space to guarantee process flexibility and product quality is of critical importance.

Typically, the state variables one has access to in order to achieve these objectives are restricted by existing plant operations and procedures. In this study, the sequence of ingredient additions is predetermined by the established procedure, as is the final ingredient ratio and mass. Therefore, the types of flexible state variables are selected from standard processing variables, such as shear rate, pressure, temperature, and flowrate. In this case, we select two of such variable types (process parameters), and an additional one referring to the timing of ingredient additions, resulting in a total of 3 types of flexible state variables.

Recently, a mechanistic model has been constructed to simulate the formulation process (Rogers et al., 2024), and the model was validated to have high accuracy over different process scales. The model proposed by Rogers et al. (2024) simplifies the liquid product's multiphase structure, describing it through a set of key material transformations and interactions within the formulation, with ingredient concentrations serving as dependent state variables in the model. A full description of the model's dynamics and the system under study is found in Rogers et al.

(2024), and the model equations are depicted in Eq. (7),

$$\frac{\mathrm{d}X_1(t)}{\mathrm{d}t} = -5r_1(t) - 10r_3(t)$$
$$\frac{\mathrm{d}X_2(t)}{\mathrm{d}t} = -2r_1(t)$$
$$\frac{\mathrm{d}X_3(t)}{\mathrm{d}t} = r_1(t) - r_3(t) - 3r_2(t)$$
$$\frac{\mathrm{d}X_4(t)}{\mathrm{d}t} = r_2(t) \quad (7)$$
$$\frac{\mathrm{d}X_5(t)}{\mathrm{d}t} = r_3(t)$$

with

$$r_1(t) = a_1 \cdot X_6(t) \cdot (c_1 - X_7(t)) \cdot X_1(t) \cdot X_2(t)$$
$$r_2(t) = a_2 \cdot X_6(t) \cdot \left(X_3(t) - \frac{X_4(t)}{b_2 \cdot X_6(t) \cdot (X_7(t) - c_2) + d} \right)$$
$$r_3(t) = a_3 \cdot X_6(t) \cdot \left(X_3(t) \cdot X_1(t) - \frac{X_5(t) \cdot X_7(t)}{b_3} \right).$$

For this case study, $\boldsymbol{x}^{\text{dept}}(t) = [X_1(t), \ldots, X_5(t)]^T$ depicts a vector of 5 dependent process states (ingredient concentrations) described by the 3 material transformations ($r_1(t)$, $r_2(t)$, and $r_3(t)$) through time, t; $\boldsymbol{x}^{\text{flexible}}(t) = [X_6(t), X_7(t), T_{\text{add}}(t)]^T$ refers to 3 flexible state variables that can be directly regulated by their respective control actions, where $T_{\text{add}}(t)$ is a variable used to determine the time at which each ingredient is added into the process. Here, $\boldsymbol{a}$, $\boldsymbol{b}$ and d are constant parameters determined through parameter estimation; and $\boldsymbol{c}$ represents physical constants with known values. As shown in Eq. 7, $\boldsymbol{x}^{\text{flexible}}$ can directly influence $\boldsymbol{x}^{\text{dept}}$ (i.e. ingredient concentrations), and can be directly regulated by the control inputs; thus their state profiles are specified in a PFD. Additionally, y_{KPI} (an important product quality indicator) can be dynamically predicted as follows:

$$y_{\text{KPI}} = h(\boldsymbol{x}^{\text{dept}}(t_f)) = (\boldsymbol{x}^{\text{dept}}(t_f))^\top \boldsymbol{A} \, \boldsymbol{x}^{\text{dept}}(t_f) + \boldsymbol{\beta}^\top \boldsymbol{x}^{\text{dept}}(t_f) + \gamma, \quad (8)$$

where $\boldsymbol{x}^{\text{dept}}(t_f)$ is the value of $\boldsymbol{x}^{\text{dept}}$ at the end of the operational time, t_f; $\boldsymbol{A}$ and $\boldsymbol{\beta}$ denote the matrix coefficients for the quadratic terms, and vector of coefficients for the linear terms, respectively; and γ is a constant parameter. In total, there exists 14 model parameters which were parameterised to describe the dynamics of batch operations. In this work, each of the flexible state variables are then discretised into 5 decision variables, each accounting for a specific mixing stage defined by the 5 ingredient additions, and thus are piecewise constant over the entire batch time, as noted in Section 2.3 (i.e., 15 total decision variables). Hereinafter, the design space in this work consists of 15 dimensions, and the discretised model will be used to simulate the process response to different state profiles, providing a case study for evaluating the effectiveness of the proposed operational space identification method for complex, non-linear, dynamic systems. Hereafter, we will refer to the three flexible state variables as Process Parameter 1, 2 and 3.

In this study, the primary objective is to robustify the process to uncertainties such that the end batch quality is maintained within an optimal specification range ($\pm 10\%$) from the desired value, $y_{\text{KPI,T}}$. On top of this, a secondary objective is introduced to minimise the batch cycle time t_f, so that process economics are enhanced, and process sustainability is improved though reducing energy expenditure; this defines the cost, C_{avg}, in Eq. (2).

The accurate representation of process uncertainties is critical to deriving the optimal operational space. If one overestimates expected process variations, then the operational regions identified in Section 2.5 will be overly conservative; conversely, if underestimated, then the regions found will not be robust to uncertainties in practice, leading to excessive end product quality violations. Therefore, to discover the sources of process uncertainty, and their typical distributions, historical data must be thoroughly examined. In this work, three main sources of uncertainty were observed and realised using 14 uncertain model parameters, and 15 uncertain decision variables. The first source is variability in compo-
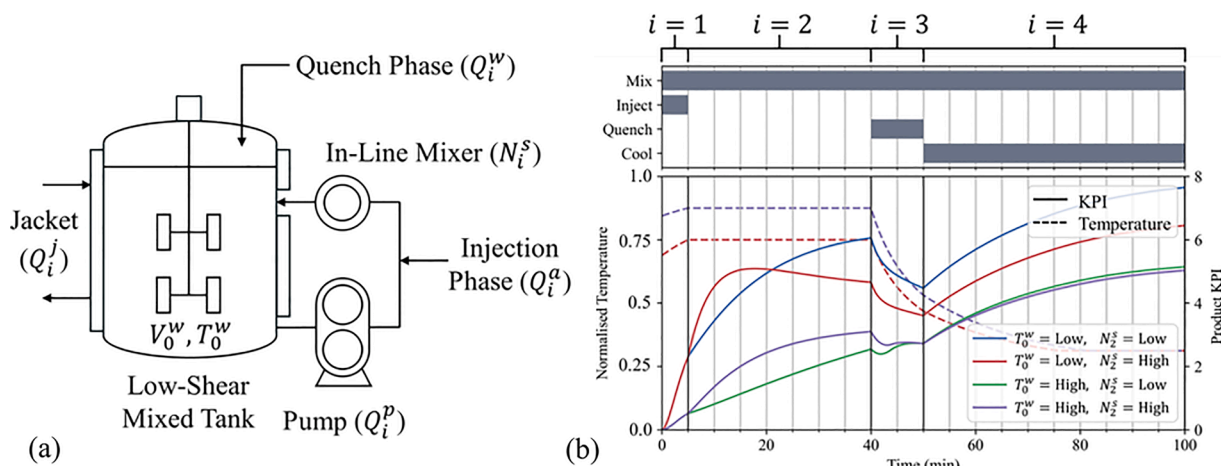
Fig. 3. Simplified representation of the batch mixing equipment used (a), and four simulated KPI profiles obtained using the system model (b) (Rogers et al., 2024).

sition of the input feed, which is reflected in the estimated parameters of the model proposed by Rogers et al. (2024). The second source of uncertainty is identified as human error, primarily introduced into the system through challenges in precisely following the timing for ingredient additions as indicated by the plant process flow diagram. The final source of uncertainty is system inconsistencies, referring to deviations of the state values from their designated set-points. The uncertainty levels assumed for the three sources are 10 %, 20 %, and 5 % per standard deviation, respectively, based on the historical data. Exceptions were made, however, for three model parameters, being b_3 and two of the matrix coefficients in matrix, $\mathbf{A}$, which were all found to be much less effected by feedstock variability so were assigned an uncertainty level of 3 % per standard deviation.

We note that all optimisation strategies in this paper were completed in Python 3.10.9 using Scipy v1.10.0's differential evolutionary algorithm, with Scipy's solve_ivp integration function using the LSODA numerical integrator. Mathematical operations and data formatting were completed using NumPy v1.23.5, Pandas v1.5.3, Matplotlib v3.7.0 and seaborn v0.12.2. The computer specifications used in this study were: Intel core i7-13700H 13th Gen @ 2.40 GHz, 32 GB of RAM.

4. Results and discussion

4.1. Results of initial design space screening

In this study, the first two steps of the proposed framework aim to efficiently screen the entire design space to identify potential optimal operational regions. These identified regions are then thoroughly examined, refined, and validated in Steps 3 and 4 (in Fig. 1). To balance computational cost and facilitate rapid screening, the scenario tree is kept moderately sized, requiring only a small set of initial high-performing state profiles strategically distributed across the entire design space to ensure diversity. Consequently, 100 uncertain scenarios were generated to construct the scenario tree by considering all types of uncertainties, and 40 initial high-performing state profiles were determined using the dynamic optimisation framework described in Section 2.3 with 15 decision variables, each of which ensured tight regulation over product quality and sought to minimise batch cycle time. Although 40 high-performing state trajectories may seem sparse, the specific optimisation problem implemented aims to maximise the distance between each high-performing trajectory, thus allowing for a small number of high-performing solutions to sufficiently search and accurately determine the operational space. On top of this, as more iterations are completed, one would expect the computational cost to increase due to an increased complexity of the problem introduced by the penalty term in Eq. (3).

Table 1
Normalised absolute values for the optimised decision variable's mean, μ, and standard deviation, σ, (across all high-performing state profiles) for Process Parameters 1, 2 and 3.

	Decision variable	Normalised mean	Normalised standard deviation
Process Parameter 1	1	0.522	0.433
	2	0.553	0.353
	3	0.520	0.385
	4	0.512	0.365
	5	0.403	0.371
Process Parameter 2	6	0.662	0.330
	7	0.608	0.342
	8	0.613	0.248
	9	0.727	0.192
	10	0.741	0.202
Process Parameter 3	11	0.100	0.008
	12	0.084	0.006
	13	0.084	0.007
	14	0.099	0.007
	15	0.084	0.006

Hence, there exists a compromise between further exploring the optimal space and the computational cost this entails.

To evaluate if the dynamic optimisation framework designed in Section 2.3 was able to generate a wide array of high-performing solutions spanning the entire range of the total design space, the normalised mean value and standard deviation of each decision variable is summarised in Table 1. From Table 1, it is clear that the objective of the optimisation strategy in Section 2.3 to determine a range of diverse high-performance solutions has been achieved for all the 10 decision variables which represent the process operating conditions at different mixing phases. In contrast, the 5 decision variables associated with ingredient addition times exhibited consistency across different high-performing state profiles, typically converging toward their lower bounds. This trend can be attributed to the direct correlation between these decision variables and the batch cycle time. In industrial practice, mixing periods are often extended beyond necessity as a conservative measure. However, in the optimisation problem focused in this work, these durations were minimised, in line with the objective defined by Eq. (2).

4.2. Results of optimal operational space clustering

Since it was found that the ingredient addition times were similar across all the high-performing state profiles, they were fixed at their lower bounds and thus the operational space is reduced from 15 to 10
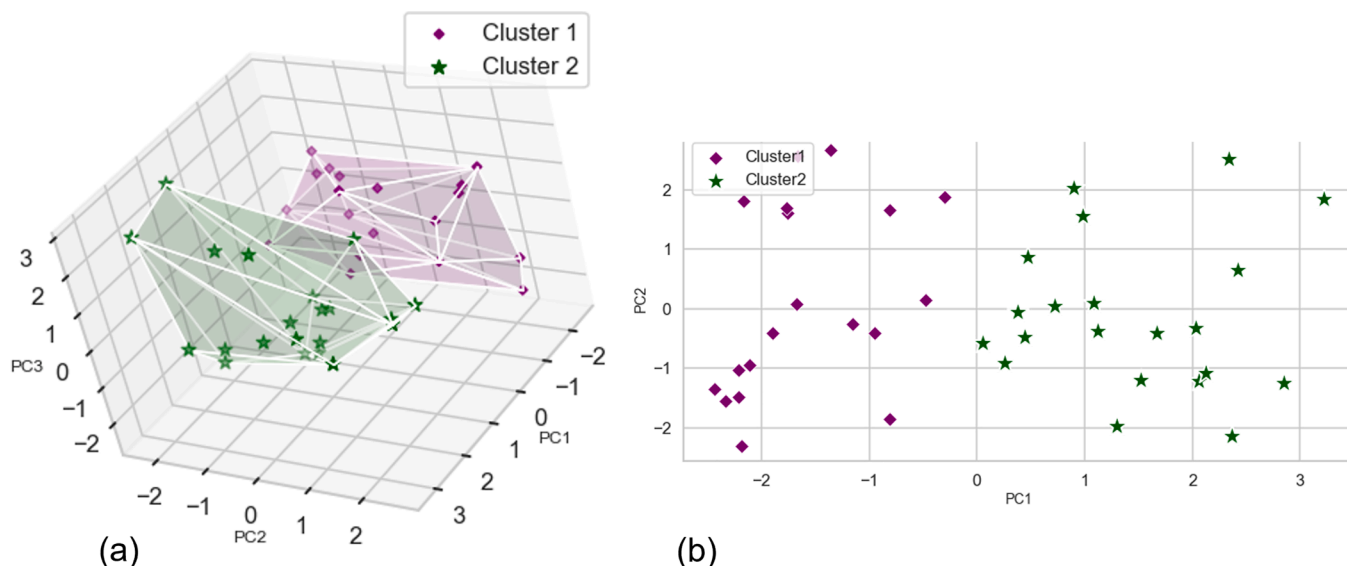
Fig. 4. Results of the clustering analysis showing the two identified clusters in a reduced 3 dimensional (a), and 2 dimensional (b) PCA space.

dimensions. The clustering stage was therefore completed based on the 10 decision variables (relating to process operating conditions at different mixing phases) alone. When the three different clustering algorithms were used, a range of optimal clusters were suggested via a sensitivity analyses making use of the silhouette method; with k-means and spectral clustering both suggesting two, and there being negligible difference between using two or three with DBSCAN.

In order to ascertain whether the third cluster provided any additional benefit, the silhouette method (Celebi and Aydin, 2016) was employed. In general, the silhouette method provides a measure of how close the data-points in a specific cluster are to each other, relative to the data points in other clusters. In this manner, it is possible to understand whether each of the clusters is uniquely well representative of the data points belonging to it. Specifically, a score of -1 is assigned to data points for which are poorly clustered, and a score of 1 is assigned to those that are well clustered; these scores can then be averaged over all data points to provide an overall assessment of the clustering. The average silhouette scores obtained using two and three clusters were 0.25 and 0.24, respectively, which advises that the third cluster provides no additional benefit, so the optimal number is two. Hence, two clusters were deemed appropriate, each of which corresponding to two distinct subregions within the approximated optimal space. Fig. 4 shows the two subregions via a reduced principal component (PCA) space; however, the clustering itself was not determined using the principal components, PCA was used solely for visualisation purposes.

Upon investigation of the two clusters seen in Fig. 4, it was found that a major behavioural difference was observed regarding the patterns of a specific process parameter (flexible state variable) that is changed at each mixing stage (represented by 5 decision variables). More specifically, in Cluster 2, a high value of this process parameter is always maintained across the initial timesteps of the process, whereas Cluster 1 is characterised by a large decrease of this process parameter at the beginning of the process. Given that the investigation of the subregions yielded distinguishable patterns for each, this evidences that there are indeed two separate optimal operational regions. As a result, proceeding with the two subregions individually is appropriate.

4.3. Results of optimal operational space refinement

Finally, to refine the two optimal operational regions, it is essential to identify the nominal state profiles and establish the corresponding lower and upper bounds for each decision variable within the two subregions.

Thus, the optimisation problem in Eq. (2) was adopted to identify the nominal state profiles as described in Section 2.5. Fig. 5 shows a comparison of estimated nominal trajectories of the two optimal regions for the 10 decision variables.

From Fig. 5, we can observe distinctly different state pathways offered by each subregions' nominal trajectories. This is particularly noticeable at the beginning of the batch where Subregion 1 is characterised by a low magnitude of Process Parameter 2 for which is compensated by a large value in Process Parameter 1; the opposite to that which is seen in Subregion 2. Each of these state pathways, when validated on 100 previously unseen randomly sampled uncertainties, boasted a violation rate of 0 % with respect to the batch quality specifications. This directly indicates the superior performance of these nominal state profiles and suggests the effectiveness of the currently proposed methodology.

Subsequently, bound estimation was carried out as described in Section 2.5 to refine the boundary of each optimal operational region and the result is shown in Fig. 6. By comparing with the two subregions, a major difference is observed in each region via the trade-off between the size of the operable spaces for the two process parameters. Specifically, Optimal Region 1 incorporates a 44.9 % decrease in its size of Process Parameter 1's operable space when compared to Optimal Region 2 but compensates for this loss by yielding an increase of 42.2 % in the operable space for Process Parameter 2. Crucially, it is possible to provide a ranking of which optimal operational region is preferred based on whether the manufacturing site has access to better control of Process Parameter 1 or 2 (i.e., the flexible state variables) via the available control inputs on site; where Optimal Region 2 would be preferred for the latter, and Optimal Region 1 would be preferred for the former.

As mentioned in Section 2.5, it is necessary to validate the identified operational regions by sampling a sufficient number of random state profiles between the upper and lower bounds and observing instances that violate end product quality constraints (i.e., $\pm 10\%$ from the KPI target, $y_{\mathrm{KPI,T}}$). In this study, we generated 1000 random sets of state profiles within each optimal operational region which were then tested using 100 unseen uncertain scenarios (i.e. 10^5 tests in total). In order to use these violating instances to compare the performance of the different optimal operational regions, we define the batch failure rate. If one considers that every one of the 10^5 tests represents a simulated batch in full (subject to different uncertainties) then, the probability of a batch failing, $P_{\mathrm{failure}} = P(y_{\mathrm{KPI,s}} > 1.1 \cdot y_{\mathrm{KPI,T}}) \cup P(y_{\mathrm{KPI,s}} < 0.9 \cdot y_{\mathrm{KPI,T}})$, is simply the fraction of these simulated batches that encountered a product quality violation. From this, Optimal Regions 1 and 2 were determined to
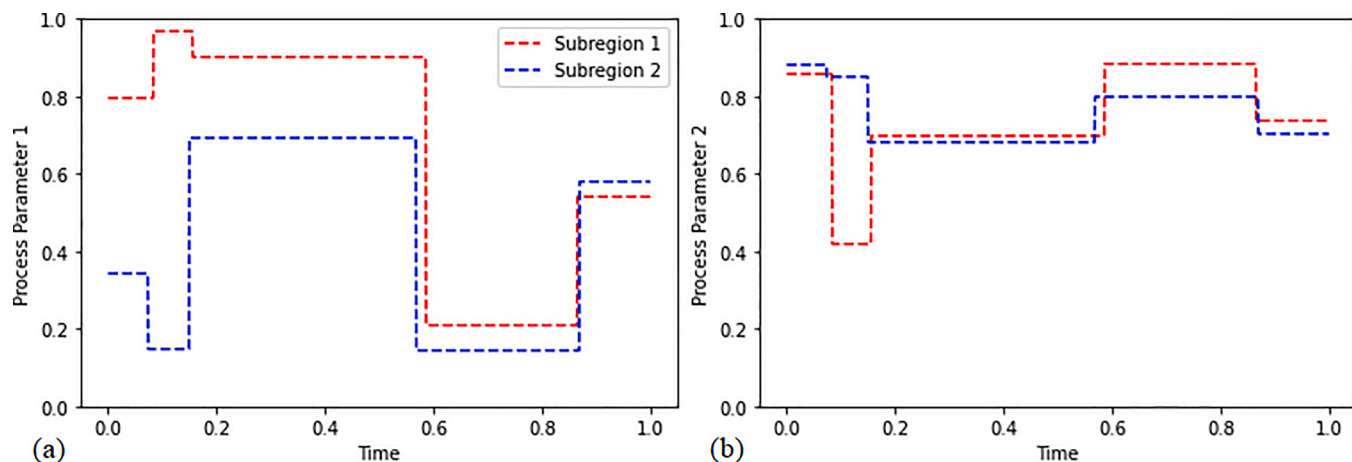
Fig. 5. Normalised nominal state trajectories for Process Parameter 1 (a) and 2 (b), for the two subregions identified.
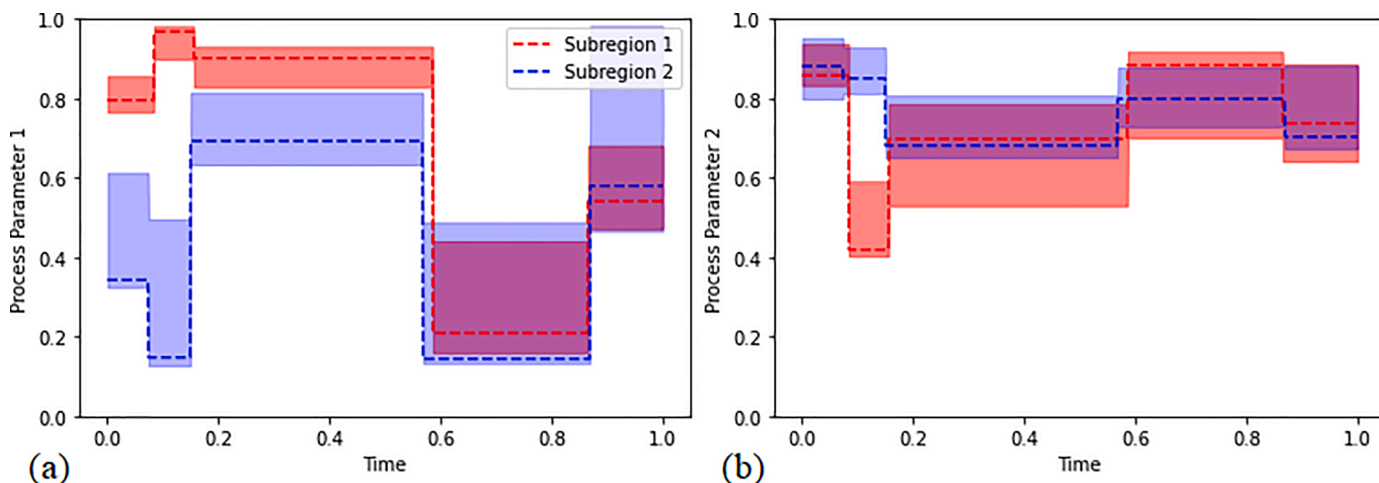

Fig. 6. Normalised state trajectories and bounds of Subregions 1 and 2, for Process Parameter 1 (a) and Process Parameter 2 (b).

possess batch failure rates of 0.9 % and 0.1 %, respectively. From further analysis of the violating instances, it is found that, out of the 100 unseen scenarios, the maximum number that a specific set of state profiles violated constraints for was 10 and 3, with the averages being 4.3 and 1.5 scenarios for Optimal Regions 1 and 2, respectively. Furthermore, out of these violating batches, the maximum and average percentage change from the desired product quality value was 41 % and 15.8 % for Optimal Region 1, and 47 % and 16.1 % for Optimal Region 2.

Therefore, the effectiveness of the methodology is shown without any additional refinements of the bounds. Even though the probability of batch failure is low and may be acceptable for specific applications, particularly if one favours a larger operational region, in this study, we aim to achieve both optimal and flexible operational regions within which, most sampled sets of state trajectories are entirely robust to uncertainties. Hence, a tolerance is defined for a sampled trajectory failure rate, that is, the fraction of the 1000 randomly sampled sets of state profiles that experience any product quality violations when the 100 uncertain batches are simulated for it. Here, we aim to reduce the sampled trajectory failure rates of each optimal operational region below 1 %. This means refinement of the bounds is necessary for both optimal operational regions.

In order to further refine the operational regions, the penalty described in Eq. (6) in Section 2.5 was employed. Fig. 7 shows the refined operational bounds for both subregions, each after two iterations of refining the bounds. From Fig. 7, it can be seen that the operational region associated with the second subregion is of much greater area compared

to the first region's operational region. To be precise, the decrease in area of the operational bounds for Subregions 1 and 2 between Figs. 6 and 7 was 51.3 % and 24.6 % for Process Parameter 1, and 78.9 % and 39.4 % for Process Parameter 2, respectively. Moreover, a direct comparison of refined bound areas between Subregions 1 and 2 reveals that the operational bounds of Subregion 1 are 67.1 % and 59.2 % smaller than those of Subregion 2, for Process Parameters 1 and 2, respectively. Using the same random state trajectory and uncertainty sampling approach as described earlier (10^5 simulated batches), we can validate the refined bounds. When the bounds are validated, we obtain that Subregions 1 and 2 posses batch failure rates of 0 % and 0.009 % (out of 10^5 batches), respectively. Hence, in practice, it may be concluded that, in the case where a small probability of violations is acceptable, enforcing a zero violation rate is a suboptimal policy since it may significantly reduce the operational region to effectively set-point control. Instead, one should permit a small violation rate in order to achieve a large increase in flexibility of the operational region.

An investigation of the violations observed in Subregion 2's bounds reveals that the maximum number of uncertain scenarios that violated constraints, for any set of state profiles, was 1. Similarly, the maximum and average percentage change from the desired KPI value are 21.9 % and 11.5 %, respectively. Both subregions met the pre-defined tolerance of the sampled trajectory failure rate, being 0 % and 0.9 % (out of 1000 randomly sampled sets of state profiles) for Subregions 1 and 2, respectively. Consequently, the operational bounds for both subregions were shown highly robust to uncertainties.
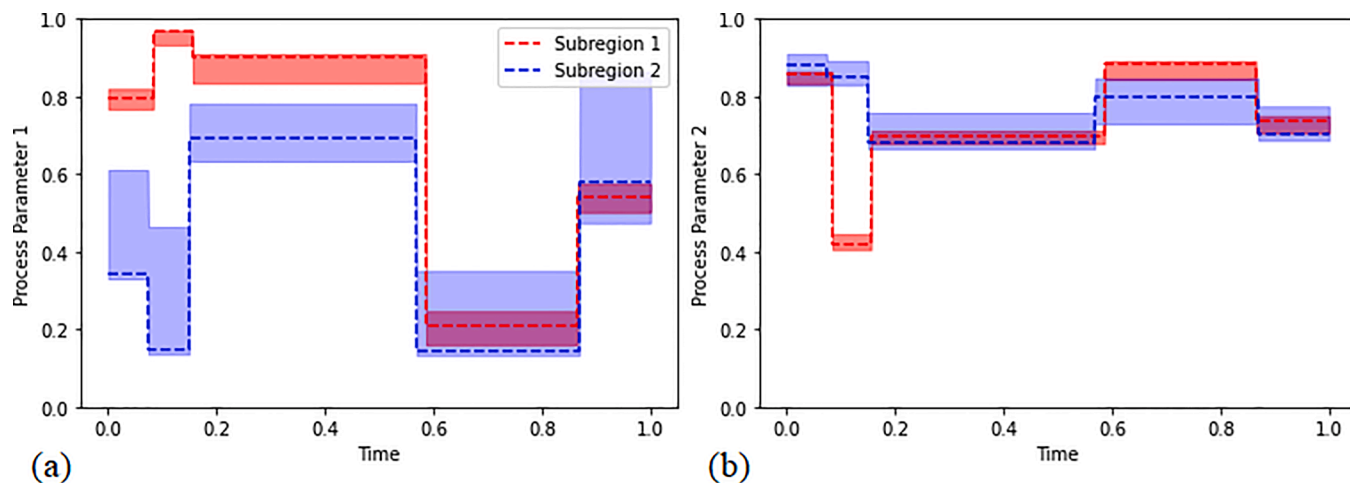
Fig. 7. Normalised state trajectories and refined bounds of Subregions 1 and 2, for Process Parameter 1 (a) and Process Parameter 2 (b).

Overall, based on this case study, the framework's ability to generate several distinct optimal regions has been proven, each of which provides operational regions practically robust to uncertainties which can be analysed and ranked based on the characteristics preferred for a specific application. Furthermore, both operational regions provided a significant decrease in batch time compared to the current set-point control approach used in the manufacturing site, that is 38.6 % in both cases. Through a rigorous investigation into the performance of each operational region, it would likely be recommended that Region 2 is more advantageous in execution due to its larger operational flexibility and similar robustness compared to Region 1. Although, in this work the methodology was defined for a dynamic batch system, it should be emphasised that this can also be applied for continuous systems, hence increasing the methodologies utility.

4.4. Flexibility analysis

In the previous section, we explored the robustness and optimal performance of each operational region. Now, we wish to investigate the flexibility of each of the nominal state profiles (i.e. the optimal state trajectories identified within each operational space) for the purpose of providing further information about the two operational regions. In particular, since it is known that raw material uncertainty contributes most to batch violations, we focus on testing the flexibility of each nominal state profile to this form of uncertainty. In this work, the method used to identify the combination of largest uncertainties that still results in no constraint violation is to apply a similar two-step optimisation framework as discussed in 2.5. Specifically, we maximise the distance between the nominal process model parameter values (material uncertainty is represented by the process model parametric uncertainty as explained in Section 3), $\xi^{nominal}$, and the optimisation variables representing the largest parameter uncertainty, ξ^{ub} or ξ^{lb}. With reference to Equation (5), one would exchange the flexible state variable terms, $x_{j,k}^{\text{f,flexible,ub}}$, for the respective uncertain parameters, ξ_i, $\forall i = 1 \dots N_\xi$, where N_ξ represents the total number of uncertain parameters. Using this method, the greatest positive and negative deviations from the nominal parameter values that continue to result in feasible operation can be found. The results of which are shown in Table 2 for all 14 uncertain parameters.

After analysis of the results in Table 2, it is ascertained that Subregion 2 possesses greater flexibility with respect to raw material uncertainties, hence it is likely to provide more reliable operation in practice. Additionally, it is noted that both sets of nominal state profiles provide high robustness to parameter uncertainty, with 41 % of the reported values in Table 2 being greater than the maximum uncertainties considered in this study (20 % for 2 standard deviations), and 71 % of the values be-

Table 2
The largest allowable positive (ub) and negative (lb) %deviations from the nominal parameter values, for the two subregion's nominal state profiles.

Parameter	Max %change from nominal			
	Subregion 1		Subregion 2	
	ub	lb	ub	lb
1	18.9	19.2	39.9	21.0
2	19.8	38.1	39.0	39.0
3	39.9	18.9	18.9	21.2
4	26.3	18.1	19.1	21.1
5	5.5	5.2	5.7	6.3
6	5.5	5.2	5.7	6.3
7	18.3	17.2	18.9	21.0
8	5.5	5.1	5.7	6.3
9	19.9	19.0	33.6	25.4
10	39.4	37.4	40.0	39.7
11	35.4	36.9	22.4	37.1
12	18.3	17.2	19.0	21.4
13	18.3	17.2	19.0	21.0
14	18.3	17.2	19.0	21.1

ing greater than 18 %. Explicitly, this means that the subregions are all robust to even the most extreme uncertainties expected to occur in operation. It can be seen that for model parameters 5, 6, and 8, both region's nominal state profiles are vulnerable to their large deviations, however, as discussed in Section 3, these parameters have previously been identified not sensitive to the change of raw materials, thus in practice it is unlikely that uncertainties arise from these model parameters.

4.5. Computational details of the case study

Table 3 shows the computation times for each of the individual steps in the framework for the specific purpose they achieve, where each step refers to the steps shown in Fig. 1.

Naturally, the computation times of these steps may differ depending on the specification of computational facility they are executed on.

In Section 4.1, the use of a moderate number of high-performing state trajectories is discussed, where the improvement in optimal space identification versus the increased computational cost limits the number of high-performing trajectories that can be generated. Although, ideally, with a large enough number of high-performing state profiles being obtained, it may be possible even discover additional optimal operational spaces, for this study, the computational expense for generating a large number of high-performance solutions would render this idea impractical. This can be clearly reasoned from Table 3, as it takes an average of

Table 3

Computational expense for each respective step of the methodology.

	Computation times	
Step	Function	Order of magnitude
Step 1	Per scenario tree generation	2 min
Step 2	Per optimum generated	2 h
Step 3	Clustering of optimum state profiles using k-means	1 min
Step 4	Single nominal control strategy + Full bounds	4 h
Flexibility analysis	Full bounds per nominal trajectory	3 min

2 hours to complete the optimisation problem when solving optimisation Problem (2) in Step 2. While other steps require negligible time to execute, Step 2 alone takes several days to complete. Despite the high computational cost, it is important to note that PFD development is typically performed offline in practice. As such, computational time is not a primary concern when determining the operational space for process design and optimisation.

5. Conclusion

In conclusion, operational space control provides a robust and flexible solution for integrating uncertainties into process product quality control to reliably meet key process constraints and product KPI. By combining flexibility and optimality, this approach supports the discovery of operational regions that are not only reliable but also high-performing across a wider range of uncertain scenarios. Unlike traditional set-point control, which restricts operations to narrow margins, operational space control enables broader process operation, allowing for adjustments in response to varying conditions without compromising on efficiency or quality, and reducing energy consumption.

A significant advantage of the currently proposed framework lies in its ability to systematically identify multiple optimal operational regions within the entire design space. This feature allows process operators to gain valuable insights into a system's operability, highlighting regions where the process is inherently more stable and easier to control. Such knowledge supports more informed decision-making and can enhance overall operational performance. The proposed framework was validated through a case study involving a dynamic batch mixing process for the production of a consumer goods formulation. Results from this study indicate that the framework effectively manages uncertainties for a real-world application, yielding economically advantageous process designs and reliable operational strategies that adhere to critical process constraints and key product quality.

Overall, this work introduces a novel combination of optimal and flexible design space identification with dynamic optimisation under uncertainty. The demonstrated methodology offers broad potential for application in various industries, particularly in contexts where conventional control approaches struggle to accommodate the inherent variability of complex processes, or where it is difficult to achieve strict compliance with set-points. Consequently, this operational space control framework represents a meaningful advancement in modern process quality control, pioneering more robust, adaptive, and economically viable process systems in industries facing increasing demands for sustainability, efficiency, and product quality.

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.

Data availability

The authors do not have permission to share data.

References

Amrih, P., Damayanti, R.W., 2022. Pharma 4.0 quality management challenge: a literature review. In: Proceedings of 12th Annual International Conference on Industrial Engineering and Operations Management (IEOM) Conference, pp. 606–615.

Anand, S.K., Kumar, S., 2022. Experimental comparisons of clustering approaches for data representation. ACM Comput. Surv. (CSUR) 55 (3), 1–33.

Asad, H.S., Yuen, R. K.K., Huang, G., 2017. Multiplexed real-time optimization of HVAC systems with enhanced control stability. Appl. Energy 187, 640–651.

Ay, M., Kisi, O., 2014. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J. Hydrol. 511, 279–289.

Ben-Tal, A., Nemirovski, A., 2002. Robust optimization–methodology and applications. Math. Program. 92, 453–480.

Bremer, J., Rätze, K. H.G., Sundmacher, K., 2017. Co2 methanation: optimal start-up control of a fixed-bed reactor for power-to-gas applications. AlChE J. 63 (1), 23–31.

Bremer, J., Sundmacher, K., 2019. Operation range extension via hot-spot control for catalytic CO 2 methanation reactors. React. Chem. Eng. 4 (6), 1019–1037.

Bruns, B., Herrmann, F., Polyakova, M., Grünewald, M., Riese, J., 2020. A systematic approach to define flexibility in chemical engineering. J. Adv. Manuf. Process. 2 (4), e10063.

Bushra, A.A., Yi, G., 2021. Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. IEEE Access 9, 87918–87935.

Celebi, M.E., Aydin, K., 2016. Unsupervised Learning Algorithms. Vol. 9. Springer.

Chen, L., Dong, T., Peng, J., Ralescu, D., 2023. Uncertainty analysis and optimization modeling with application to supply chain management: a systematic review. Mathematics 11 (11), 2530.

Chen, Z., Yan, Z., 2018. Scenario tree reduction methods through clustering nodes. Comput. Chem. Eng. 109, 96–111.

Djuris, J., Djuric, Z., 2017. Modeling in the quality by design environment: regulatory requirements and recommendations for design space and control strategy appointment. Int. J. Pharm. 533 (2), 346–356.

Efheij, H., Albagul, A., Albraiki, N.A., 2019. Comparison of model predictive control and PID controller in real time process control system. In: 2019 19th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA). IEEE, pp. 64–69.

Forbes, M.G., Patwardhan, R.S., Hamadah, H., Gopaluni, R.B., 2015. Model predictive control in industry: challenges and opportunities. IFAC-PapersOnLine 48 (8), 531–538.

Forster, T., Vázquez, D., Moreno-Palancas, I.F., Guillén-Gosálbez, G., 2024. Algebraic surrogate-based flexibility analysis of process units with complicating process constraints. Comput. Chem. Eng. 184, 108630.

Fujiwara, K., Kano, M., Hasebe, S., 2011. Correlation-based spectral clustering for flexible process monitoring. J. Process Control 21 (10), 1438–1448.

Gattu, G., Zafiriou, E., 1999. A methodology for on-line setpoint modification for batch reactor control in the presence of modeling error. Chem. Eng. J. 75 (1), 21–29.

Geletu, A., Klöppel, M., Zhang, H., Li, P., 2013. Advances and applications of chance-constrained approaches to systems optimisation under uncertainty. Int. J. Syst. Sci. 44 (7), 1209–1232.

Hicks, A., Johnston, M., Mowbray, M., Barton, M., Lane, A., Mendoza, C., Martin, P., Zhang, D., 2021. A two-step multivariate statistical learning approach for batch process soft sensing. Digit. Chem. Eng. 1, 100003.

Hill, K., 2007. Industrial development and application of biobased oleochemicals. Pure Appl. Chem. 79 (11), 1999–2011.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. 622, 178–210.

Jia, H., Ding, S., Xu, X., Nie, R., 2014. The latest research progress on spectral clustering. Neural Comput. Appl. 24, 1477–1486.

Khan, K., Rehman, S.U., Aziz, K., Fong, S., Sarasvady, S., 2014. Dbscan: past, present and future. In: The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014). IEEE, pp. 232–238.

Lima, F.V., Jia, Z., Ierapetritou, M., Georgakis, C., 2010. Similarities and differences between the concepts of operability and flexibility: the steady-state case. AlChE J. 56 (3), 702–716.

Lin, L., Tang, C., Dong, G., Chen, Z., Pan, Z., Liu, J., Yang, Y., Shi, J., Ji, R., Hong, W., 2021. Spectral clustering to analyze the hidden events in single-molecule break junctions. J. Phys. Chem. C 125 (6), 3623–3630.

Lörzing, P., Schake, P., Schlierf, M., 2024. Anisotropic DBSCAN for 3d SMLM data clustering. J. Phys. Chem. B 128 (33), 7934–7940.

Lucia, S., Finkler, T., Engell, S., 2013. Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. J. Process Control 23 (9), 1306–1319.

Homem-de Mello, T., Bayraksan, G., 2014. Monte carlo sampling-based methods for stochastic optimization. Surv. Oper. Res. Manag. Sci. 19 (1), 56–85.

Mortier, S. T. F.C., Van Bockstal, P.-J., Corver, J., Nopens, I., Gernaey, K.V., De Beer, T., 2016. Uncertainty analysis as essential step in the establishment of the dynamic design space of primary drying during freeze-drying. Eur. J. Pharm. Biopharm. 103, 71–83.

Na, S., Xumin, L., Yong, G., 2010. Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. IEEE, pp. 63–67.

Novaraa, F.M., Henning, G.P., 2018. Resilient scheduling under uncertain processing times: a hybrid CP/TOC approach. In: Computer Aided Chemical Engineering. Elsevier. Vol. 44, pp. 1261–1266.

Pranevicius, H., Sutiene, K., 2007. Scenario tree generation by clustering the simulated data paths. In: Proc 21st Eur Conf on Model and Simul (ECMS).

Raspanti, C.G., Bandoni, J.A., Biegler, L.T., 2000. New strategies for flexibility analysis and design under uncertainty. Comput. Chem. Eng. 24 (9–10), 2193–2209.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L. d.F., Rodrigues, F.A., 2019. Clustering algorithms: a comparative approach. PLoS One 14 (1), e0210236.

Rogers, A.W., Lane, A., Mendoza, C., Watson, S., Kowalski, A., Martin, P., Zhang, D., 2024. Integrating knowledge-guided symbolic regression and model-based design of experiments to automate process flow diagram development. Chem. Eng. Sci. 300, 120580.

Rovira, M., Engvall, K., Duwig, C., 2022. Identifying key features in reactive flows: a tutorial on combining dimensionality reduction, unsupervised clustering, and feature correlation. Chem. Eng. J. 438, 135250.

Sancho, A., Ribeiro, J.C., Reis, M.S., Martins, F.G., 2022. Cluster analysis of crude oils with k-means based on their physicochemical properties. Comput. Chem. Eng. 157, 107633.

Silvente, J., Papageorgiou, L.G., Dua, V., 2019. Scenario tree reduction for optimisation under uncertainty using sensitivity analysis. Comput. Chem. Eng. 125, 449–459.

Swaney, R.E., Grossmann, I.E., 1985. An index for operational flexibility in chemical process design. Part I: formulation and theory. AlChE J. 31 (4), 621–630.

University, P., 2023. Robust optimization with ellipsoidal uncertainty and its applications. https://www.princeton.edu/. Accessed: 2024-11-14.

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416.

Wan, J., Marjanovic, O., Lennox, B., 2012. Disturbance rejection for the control of batch end-product quality using latent variable models. J. Process Control 22 (3), 643–652.

Wang, H., Mastragostino, R., Swartz, C. L.E., 2016. Flexibility analysis of process supply chain networks. Comput. Chem. Eng. 84, 409–421.

Xu, D., Chen, Z., Yang, L., 2012. Scenario tree generation approaches using k-means and LP moment matching methods. J. Comput. Appl. Math. 236 (17), 4561–4579.

Zheng, Q.P., Wang, J., Liu, A.L., 2014. Stochastic optimization for unit commitment—A review. IEEE Trans. Power Syst. 30 (4), 1913–1924.