Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Probabilistic design spaces from small DoEs – *A boundary-focused workflow using quantile surrogates*

Tobias Overgaard[ab]*, Emmanouil Papadakis[b], Maria-Ona Bertran[b], and Maria M. Papathanasiou[cd]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[b] CMC & Product Supply, PS API, Novo Nordisk A/S, 4400 Kalundborg, Denmark
[c] The Sargent Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, United Kingdom
[d] Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom
* Corresponding Author: tobov@dtu.dk.

ABSTRACT

Probabilistic design spaces enable pharmaceutical manufacturers to balance regulatory compliance and operational efficiency. In this context, the "edge-of-failure" separating compliant from non-compliant operation is not a fixed line, but an uncertain region driven by model parameter uncertainty. Traditional methods typically map this probability of failure across the whole design space, which is a computationally expensive task. We propose a more direct approach, reformulating the problem using quantile functions to search for a deterministic boundary at a desired confidence level. These functions are approximated via Gaussian process surrogates using a novel adaptive sampling strategy. An industrial peptide acylation case study identifies the probabilistic design space using 13 experiments versus 18 from a traditional DoE; a 28% reduction. The framework quantifies trade-offs between operational robustness and yield optima.

Keywords: Probabilistic design space, Adaptive sampling, Pharmaceutical manufacturing, Surrogate modeling

INTRODUCTION

In pharmaceutical manufacturing, a design space (DS) is *"the multidimensional combination and interaction of material attributes and process parameters that have been demonstrated to provide assurance of quality"* [1]. Once approved, a DS allows manufacturers to adjust process settings within its bounds without additional regulatory submissions, enabling operational flexibility [2]. A DS can be expressed mathematically as

$$\{\mathbf{x} \in \mathcal{K} \mid \forall \ell \in \{1, \dots, L\}, g_\ell(\mathbf{x}, f(\mathbf{x}, \theta)) \leq 0\}, \quad (1)$$

where $\mathcal{K} \subseteq \mathbb{R}^P$ is the knowledge space, and the process model $\mathbf{y} = f(\mathbf{x}, \theta)$ predicts CQAs $\mathbf{y} \in \mathbb{R}^Q$ with fixed model parameters $\theta \in \mathbb{R}^R$. Each constraint g_ℓ represents a CQA specification limit or a process constraint.

Ignoring model parameter uncertainty yields a deterministic design space (DDS), as in Eq. (1). In practice, parameter uncertainty can reduce or reshape the DS, motivating methods to define a probabilistic design space (PDS). In this context, the model parameters follow a conditional density $p(\theta \mid \mathbf{x})$, which can be inferred via first-order Taylor series approximation [3] or Bayesian parameter estimation [4]. Under model parameter uncertainty, the constraints in Eq. (1) become probabilistic

$$\{\mathbf{x} \in \mathcal{K} \mid \forall \ell \in \{1, \dots, L\}, \mathbb{P}(g_\ell(\mathbf{x}, f(\mathbf{x}, \boldsymbol{\Theta})) \leq 0) \geq C\}, \quad (2)$$

where $\boldsymbol{\Theta}$ is a random variable with density $p(\theta \mid \mathbf{x}), C$ is a reliability level, and

$$\mathbb{P}(g_\ell(\mathbf{x}, f(\mathbf{x}, \boldsymbol{\Theta})) \leq 0) = \int_{\{\theta \mid g_\ell(\mathbf{x}, f(\mathbf{x}, \theta)) \leq 0\}} p(\theta \mid \mathbf{x}) \, \mathrm{d}\theta. \quad (3)$$

The central challenge in PDS identification is the computational cost of evaluating Eq. (3). The objective of this paper is to propose a method that bypasses this cost by reformulating Eq. (2) as a deterministic problem.

Sampling-based approaches to PDS identification

Wet-lab experimentation is a costly and time-consuming task in pharmaceutical development. Therefore, the literature favors model-based methods for mapping the DS [5]. Knowledge-driven models are often preferred over purely empirical ones as they embed physical principles, require less calibration data, and extrapolate more reliably beyond the training domain [6].

Standard PDS identification typically employs a Monte Carlo framework [7]. The knowledge space $\mathcal{K}$ is discretized, and the model $f(\mathbf{x}, \boldsymbol{\theta})$ is evaluated by drawing samples of $\boldsymbol{\theta}$ at each discrete point $\mathbf{x}$. The feasibility probability in Eq. (3) is estimated as the fraction of simulations satisfying all CQA constraints. The PDS is then defined as the set of inputs $\mathbf{x}$ where this probability meets the reliability level C.

While robust and widely applicable, this approach carries a computational cost proportional to the product of grid points and parameter samples, which often proves prohibitive for complex models or large knowledge spaces. To combat this, dimension reduction techniques like partial least squares (PLS) are used to project high-dimensional inputs into a tractable latent space prior to PDS mapping [19]. Computational overhead can also be managed through targeted exploration; nested sampling [8], for instance, accelerates convergence by iteratively discarding low-probability points in favor of candidates near the target threshold C. Furthermore, adaptive surrogate-driven sampling [9] can reduce simulation requirements by learning a surrogate of the feasibility probability and direct new evaluations to the most informative regions (e.g., near the feasibility boundary [10]). Common surrogates include Gaussian process (GP) regression [11] to drive acquisition via uncertainty estimates, or artificial neural networks (ANNs) [12], which have recently been deployed to approximate expensive mechanistic models under material uncertainty [20]. Alternative approaches replace the inner sampling over $\boldsymbol{\theta}$ by a fast optimizer inspired by flexibility analysis [13], though it tends to yield more conservative PDS estimates.

After identifying the PDS, a common next step is to inscribe the largest axis-aligned (hyper)rectangle within it - the *acceptable operating region* (AOR) - as this facilitates simple communication about process flexibility to regulatory authorities. However, a fundamental limitation persists across the aforementioned methods: they rely on grid discretization to generate probability maps. If explicit analytical expressions for the boundary constraints existed, AOR identification could be formulated directly as a volume-maximizing problem, akin to flexibility index calculations [13]. Because current sampling-based approaches lack these explicit expressions, AOR identification typically demands complex post-processing via computational geometry [5]. This highlights a critical gap: *the need for a methodology that generates explicit boundary constraints to streamline AOR identification and reveal flexibility-limiting boundary points.*

Quantile-based PDS formulation

To provide practitioners with explicit boundaries, we propose a new formulation of the PDS, inspired by methods from reliability-based design optimization [14, 17]. We utilize the quantile function, which is defined as

$$Q_\ell^{(C)}(\mathbf{x}) = \inf\{q \in \mathbb{R} \mid \mathbb{P}(g_\ell(\mathbf{x}, f(\mathbf{x}, \boldsymbol{\Theta})) \leq q) \geq C\}, \quad (4)$$

for each $\ell = 1, \ldots, L$. Effectively, $Q_\ell^{(C)}$ acts as the inverse of Eq. (3). Leveraging the properties of $Q_\ell^{(C)}$, we can recast the probabilistic constraints in Eq. (2) into

$$\mathbb{P}(g_\ell(\mathbf{x}, f(\mathbf{x}, \boldsymbol{\Theta})) \leq 0) \geq C \;\Leftrightarrow\; Q_\ell^{(C)}(\mathbf{x}) \leq 0, \quad (5)$$

for all $0 < C \leq 1$. Using Eq. (5), PDS identification turns into a deterministic problem, where the zero-level set $\{\mathbf{x} \mid Q_\ell^{(C)}(\mathbf{x}) = 0\}$ explicitly defines the "edge-of-failure" boundary at a fixed reliability level C. Unlike raw probabilities, the quantile constraint provides a clearer, more useful signal for targeted experimentation: its sign indicates feasibility, and its magnitude measures the distance to the boundary.

The challenge lies in learning a surrogate for these quantiles based on data from a small DoE. This work addresses three practical questions:

(i) **Surrogate construction:** *How do we build quantile surrogates that account for input and model uncertainty?* We employ GP surrogates trained with a novel three-phase adaptive sampling scheme: (1) sample along the PDS boundary, (2) target intersection points between the boundary and the AOR to identify binding constraints, and (3) sample near KPI optima to balance performance with robustness.

(ii) **Experiment selection**: *Which experiments most effectively reduce uncertainty?* We target the intersection of the AOR and the feasibility boundary to pinpoint edge-of-failure zones.

(iii) **Stopping criteria:** *How many experiments are required?* We monitor the convergence of the size of the AOR to determine a minimum viable dataset.

The remainder of this paper details the methodology, applies it to an industrial API case study, and validates the results against traditional approaches.

METHODOLOGY

This paper introduces a novel methodology to identify PDSs using quantile surrogates, as summarized in Fig. 1. The core innovation is to replace expensive Monte Carlo simulations with a surrogate model that predicts the quantiles of the CQA constraints to account for both process and model uncertainty. To use the methodology, we assume the following two requirements are met:

(a) **Process model:** A mechanistic model f exists, mapping process parameters $\mathbf{x}$ and model parameters $\boldsymbol{\theta}$ to the relevant CQAs.

(b) **Uncertainty characterization**: We assume a small initial DoE has already been performed. This data

is used to fit the model f and will result in a conditional probability density $p(\boldsymbol{\theta} \mid \mathbf{x})$ that quantifies the uncertainty in the model parameters.

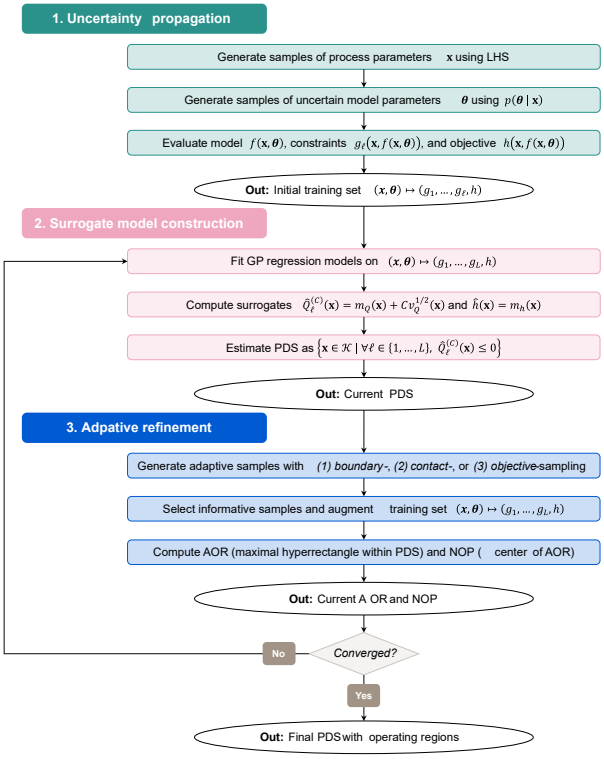

Figure 1. PDS identification methodology.

Step 1: Uncertainty propagation

The first step generates an initial training set to capture how parameter uncertainty propagates to the constraints g_ℓ, as well as to the objective function h, which describes process KPIs like yield or capacity. To obtain analytical expressions of uncertainty, we require Gaussian model parameters. Since this may not be true in general, we assume a transformation T (e.g., Box-Cox) exists such that the transformed model parameters $\mathbf{Z} \coloneqq T(\boldsymbol{\Theta})$ are distributed as $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma})$. This leads to analytical results for the model uncertainty and avoids expensive Monte Carlo sampling in subsequent steps [16].

We generate an initial training set $\mathcal{D}_0$ by sampling I_0 process parameter points using Latin hypercube sampling (LHS) for uniform coverage. At each process point, we draw J_0 model parameter realizations from the conditional density. In practice, we typically find that $I_0 \in [5, 15]$ and $J_0 \in [1, 10]$ are sufficient. The model f is evaluated at these points to compute the constraints and objective. The surrogate fits to the output w_{ij}, which is defined according to the function being modeled:

$$w_{ij} = \begin{cases} g_\ell(\mathbf{x}_i, T^{-1}(\mathbf{z}_{ij})), & \text{constraint model l,} \\ h(\mathbf{x}_i, T^{-1}(\mathbf{z}_{ij})), & \text{objective model.} \end{cases} \tag{6}$$

This yields an initial training set of input-output pairs $\mathcal{D}_0 = \{(\mathbf{v}_{ij}, w_{ij})\}_{1 \le i \le I_0, 1 \le j \le J_0}$, where the input vector $\mathbf{v}_{ij} \coloneqq [\mathbf{x}_i;\ \mathbf{z}_{ij}]$ concatenates both the deterministic process settings and the transformed model parameters.

Step 2: Surrogate model construction

Given the initial training set $\mathcal{D}_0$, we employ GP regression to model the constraint functions g_ℓ and objective h. GP regression is preferred for its strength in uncertainty quantification [11]. The GP assumes a zero-mean prior and a squared exponential (SE) kernel function with automatic relevance determination (ARD) [15]. This allows the model to learn unique length-scales for each input dimension, effectively identifying the importance of each process and model parameter.

Analytical moment matching: Standard GP prediction assumes deterministic inputs. However, our inputs include uncertain model parameters $\mathbf{Z}$. While mapping a Gaussian input through a nonlinear GP generally yields a non-Gaussian predictive distribution, we approximate the resulting distribution with a Gaussian using moment matching [16]. As a result, we can derive closed-form expressions for the mean $\tilde{m}_w(\mathbf{x}^*)$ and variance $\tilde{s}_w(\mathbf{x}^*)$

$$\begin{aligned} \tilde{m}_w(\mathbf{x}^*) &\coloneqq \boldsymbol{\beta}^\top \mathbf{q}(\mathbf{x}^*), \\ \tilde{s}_w(\mathbf{x}^*) &\coloneqq \alpha^2 - \operatorname{tr}\Big((\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{Q}(\mathbf{x}^*)\Big) \\ &\quad + \boldsymbol{\beta}^\top \mathbf{Q}(\mathbf{x}^*) \boldsymbol{\beta} - \tilde{m}_w^2(\mathbf{x}^*). \end{aligned} \tag{7}$$

We refer the reader to [16] for the definitions of $\boldsymbol{\beta}, \mathbf{q}$ and $\mathbf{Q}$ and the full derivation of Eq. (7). The other parameters include the signal variance α^2 of the SE kernel and the observation noise σ_ε^2. The covariance matrix $\mathbf{K}$ is computed by evaluating the SE kernel across all input pairs.

Quantile surrogate construction: Using these analytical moments, we construct the C-level quantile surrogate. For a required reliability level C (e.g., 95%), the surrogate is defined as the upper confidence bound of the predictive distribution

$$\tilde{Q}_\ell^{(C)}(\mathbf{x}) = \tilde{m}_w(\mathbf{x}) + \Phi^{-1}(C) \sqrt{\tilde{s}_w(\mathbf{x})}, \tag{8}$$

where Φ^{-1} is the quantile function of the Gaussian distribution. The PDS can then be directly approximated as

$$\Big\{\mathbf{x} \in \mathcal{K} \mid \forall \ell \in \{1, \dots, L\}, \hat{Q}_\ell^{(C)}(\mathbf{x}) \le 0\Big\}, \tag{9}$$

thereby accounting for both input and model uncertainty. Unlike Eq. (8), the objective surrogate is expressed as $\hat{h}(\mathbf{x}) = \tilde{m}_w(\mathbf{x})$, as it approximates only the expected value rather than a quantile. At this stage, all surrogates are built on $\mathcal{D}_0$, and the next step is to refine them near AOR-critical regions and the process optimum.

Step 3: Adaptive refinement

Surrogates based on space-filling designs often lack accuracy at the specific boundaries that define

feasibility. We define a three-phase adaptive sampling strategy to combat this. At each iteration, the algorithm selects a new design point $\mathbf{x}^*$ and parameter realization $\mathbf{z}^*$ to evaluate, refits the surrogates, and updates the PDS.

- **Phase 1: Boundary-guided sampling:** Early iterations focus on sharpening the PDS boundaries. We search for process settings where the constraints are closest to their critical values (the zero-level set). This amounts to minimizing the worst-case constraint violation, i.e.., $\max_{\ell} |\hat{Q}_{\ell}^{(C)}(\mathbf{x})|$. To enable gradient-guided optimization, we build a smooth proxy of the worst-case constraint violation, given by

$$r(\mathbf{x}) = \frac{1}{\eta} \log\Big(\sum_{\ell=1}^{L} \exp\big(\eta\, \hat{Q}_{\ell}^{(C)}(\mathbf{x})\big)\Big), \tag{10}$$

  where $\eta > 0$ is a sharpness parameter. As $\eta \to \infty, r(\mathbf{x})$ approaches $\max_{\ell} \hat{Q}_{\ell}^{(C)}(\mathbf{x})$. We optimize

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} r^2(\mathbf{x}), \tag{11}$$

  to target regions where $r(\mathbf{x}) \approx 0$, which correspond to the zero-level set of the worst-case constraint.
- **Phase 2: Contact-guided sampling**: Once the quantile boundary is defined, we focus on the AOR - the largest hyperrectangle inscribed within the PDS. The accuracy of the AOR depends on the so-called *contact points*, which we define as the points where the rectangle intersects the PDS boundary. We first solve an optimization problem to maximize the AOR volume. We then sample the edge points of this hyperrectangle that are predicted to lie on the quantile boundary. This refines the "edge-of-failure" regions that ultimately limit the size of the operating window.
- **Phase 3: Objective-guided sampling**: Finally, the strategy aims at optimizing process performance. We maximize the objective surrogate subject to the quantile constraints. This ensures the PDS is accurate in the region of optimal operation, verifying that the identified optimum is truly robust to uncertainty.

The Next Iteration: Once the best process point $\mathbf{x}^*$ is identified (from any of the three phases), we must select which parameter realization $\mathbf{z}^*$ to simulate. Simply sampling the mean is insufficient. Instead, we draw candidates from the parameter distribution $\mathcal{N}(\boldsymbol{\mu}(\mathbf{x}^*), \boldsymbol{\Sigma})$ and select the one that minimizes the *boundary uncertainty*. Mathematically, we identify active constraint $\ell^* \in \{1, \dots, L\}$ (defined as the one nearest to $\mathbf{x}^*$) and compute

$$\mathbf{z}^* = \operatorname*{argmin}_{\mathbf{z}} \frac{|m_{\ell^*}([\mathbf{x}^*;\mathbf{z}])|}{\sqrt{s_{\ell^*}([\mathbf{x}^*;\mathbf{z}])}}, \tag{12}$$

where m_{ℓ^*} and s_{ℓ^*} are the GP posterior mean and variance for constraint ℓ^* [15]. This criterion selects parameters where the constraint prediction is near the boundary (small $|m_{\ell^*}|$) yet highly uncertain and poorly informed (large s_{ℓ^*}) [17]. If multiple parameter samples are desired (as in Step 1), we select those close to the minimum.

The new simulation is added to the training set

$$\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \begin{cases} \{\mathbf{v}^*, g_{\ell}(\mathbf{x}^*, T^{-1}(\mathbf{z}^*))\}, & \textit{constraint } l, \\ \{\mathbf{v}^*, h(\mathbf{x}^*, T^{-1}(\mathbf{z}^*))\}, & \textit{objective.} \end{cases} \tag{13}$$

and all surrogates are refitted. Convergence is assessed differently for each phase: An error-based stopping criterion (ESC) is used for boundary sampling (see [18] for details), AOR size stability is used for contact sampling, and optimizer convergence is used for the objective phase. When one phase converges, we either advance to the next or analyze the final PDS. If not, we can add more digital samples or update with new experimental data.

CASE STUDY

In this paper, we consider an acylation reaction step, which is a critical synthetic step in the production of a long-acting peptide-based API. The process attaches a hydrophobic side chain to the peptide intermediate to form the final API. Several competing reactions must be balanced to achieve high product purity and yield while minimizing the formation of impurities.

Reaction scheme and knowledge-driven model

The mechanistic model comprises seven coupled ordinary differential equations tracking the concentrations of the species: SM (starting material), P (desired product), A (impurity from undesired acylation), HI (hydrophobic impurities from di-acylation), HSC (hydrolyzed side chain reagent), SC (side chain reagent), and P^* (degraded product). The reaction scheme is:

$\mathrm{SM} + \mathrm{SC} \xrightarrow{k_1} \mathrm{P}$ (desired acylation at primary site)

$\mathrm{SM} + \mathrm{SC} \xrightarrow{k_2} \mathrm{A}$ (undesired acylation at secondary site)

$\mathrm{P} + \mathrm{SC} \xrightarrow{k_2} \mathrm{HI}$ (di-acylation at secondary site)

$\mathrm{A} + \mathrm{SC} \xrightarrow{k_1} \mathrm{HI}$ (di-acylation at primary site)

$\mathrm{SC} \xrightarrow{k_3} \mathrm{HSC}$ (hydrolysis of side chain reagent)

$\mathrm{P} \xrightarrow{k_4} \mathrm{P}^*$ (product degradation)

We assume that reactions at the same site happen at identical rates regardless of the other site's state, so only two acylation rates are required. We also assume that both rates depend on pH and temperature, $k_i = k\_i(\mathrm{pH}, T)$ for $i = 1, 2$. The hydrolysis and degradation rates follow $k_i = k_i(T)10^{\mathrm{pH}-14}$ for $i = 3, 4$, reflecting their hydroxide-catalyzed mechanisms. All rate parameters are log-transformed and estimated using least-squares regression on the 18 available LHS experiments

$$\log\big(k_1(\mathrm{pH},T)\big) = \beta_{1,1} + \beta_{1,2}/T + \beta_{1,3}\mathrm{pH} + \beta_{1,4}\mathrm{pH}^2 + \varepsilon_1,$$
$$\log\big(k_2(\mathrm{pH},T)\big) = \beta_{2,1} + \beta_{2,2}/T + \beta_{2,3}\mathrm{pH} + \beta_{2,4}\mathrm{pH}^2 + \varepsilon_2,$$
$$\log\big(k_3(T)\big) = \beta_{3,1} + \beta_{3,2}/T + \varepsilon_3,$$
$$\log\big(k_4(T)\big) = \beta_{4,1} + \beta_{4,2}/T + \varepsilon_4, \qquad (14)$$

where the βs denote regression coefficients, and the εs are zero-mean Gaussian errors. The $1/T$ term captures Arrhenius temperature dependence, and log-transformation ensures approximate Gaussianity (as required by Step 1 of the methodology), with means parameterized by pH and T. As an example, take $\mathrm{pH} = 11.5, T = 2.50°\mathrm{C}$. Then $\log([k_1,k_2,k_3,k_4]^\top) \sim \mathcal{N}(\boldsymbol{\mu}(11.5,2.50),\boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu}(11.5,2.50) = [-4.26,-8.11,1.11,-0.950]^\top,$$
$$\boldsymbol{\Sigma} = 10^{-3}\,\mathrm{diag}(5.96,0.856,915,269). \qquad (15)$$

For brevity (and model confidentiality), we do not report the full distributions across all (pH,T). By construction, the rates are modeled as independent random variables (no covariance entries in $\boldsymbol{\Sigma}$).

RESULTS AND DISCUSSION

We demonstrate the methodology on the peptide acylation case study. Based on the reaction scheme, a standard kinetic model was formulated as a system of ordinary differential equations. With this, Requirement (a) of the methodology (**process model**) is fulfilled. Furthermore, the fitted rate parameters in Eq. (14) follow a Gaussian distribution, thereby meeting Requirement (b) (**uncertainty characterization**).

Design space characterization

We characterize the PDS by two CQAs - hydrophobic impurities $\leq 2\%$ (constraint 1) and purity $\geq 70\%$ (constraint 2) over $\mathrm{pH} \in [10.5, 12.5]$ and $T \in [-3, 10]°\mathrm{C}$. We set $\mathbf{x} \coloneqq [\mathrm{pH}, T]^\top$ and $\mathbf{z} \coloneqq \log([k_1,k_2,k_3,k_4]^\top)$ and calibrate with 13 experiments (this choice is justified later). The reliability level is set at $C = 0.975$. For each sample of $\mathbf{x}$, we draw five $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}),\boldsymbol{\Sigma})$. The surrogate is initialized with 10 LHS points in $\mathbf{x}$, then refined adaptively.

Fig. 2 shows the PDS evolution at 15/30/40 iterations using the three-phase sampling strategy. At iteration 15 (Fig. 2**(a)**), boundary-guided samples (black ×) cluster on $\hat{Q}_1^{(C)} = 0$ and $\hat{Q}_2^{(C)} = 0$ (solid lines), quickly capturing the quantile boundary. Separation from the mean contours (dashed lines) reflect model uncertainty.

By iteration 30 (Fig. 2**(b)**), the AOR (black □) is identified, and contact-focused samples (orange ×) intersect the contours. The PDS geometry changes only slightly, with the mean purity contour shifting toward the quantile, indicating reduced uncertainty; the AOR's upper-left corner remains unsampled because it is unconstrained. The nominal operating point (NOP) lies around $\mathrm{pH} \approx 11.8$ and $T \approx 1.6°\mathrm{C}$ (black ★).

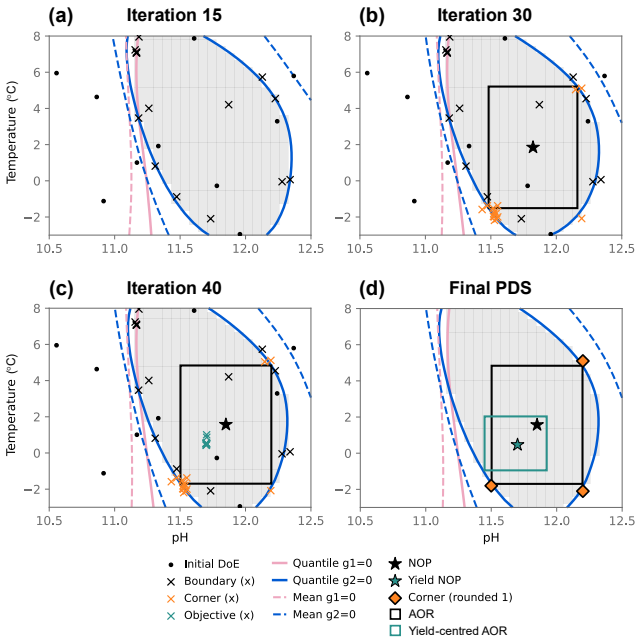

Figure 2. Snapshots of the PDS identification scheme.

At iteration 40 (Fig. 2**(c)**), we introduce a yield objective. Objective-guided samples (green ×) cluster near $\mathrm{pH} \approx 11.7$ and $T \approx 0.5°\mathrm{C}$, producing a yield-centered NOP (green ★). The AOR has shifted slightly to the right, possibly reflecting selection among equally sized rectangles.

The resulting PDS (Fig. 2**(d)**) shows two sets of NOPs and AORs: the yield-centered AOR (green □) has area ≈ 1.8 versus ≈ 4.6 for the maximal-flexibility AOR (black □), quantifying the robustness–performance trade-off. Maximizing yield narrows the window and raises violation risk, yet the flexibility-oriented AOR still contains the yield optimum, enabling optimal operation within robust ranges. Three edge-of-failure points (orange ♦) are found as local averages of the corner-point clusters. Using these as suggestions for follow-up experiments will most efficiently sharpen the boundary between feasible and infeasible operation.

Convergence analysis and experimental strategy

Fig. 3 shows convergence of the three-phase adaptive sampling. The error-based stopping criterion (ESC) in Fig. 3**(a)** converges within ~10 boundary-guided iterations: the purity error (e_2) drops from ~18% to ~7.5% and remains below 10%, while the impurity constraint (e_1) is reliably identified within five iterations. We extend to 15 iterations to confirm stability of the estimates.

The NOP stabilizes after 15 iterations of contact-guided sampling (30 iterations in total). Another 10 iterations of objective-guided sampling shows that the NOP remains locked at $\mathrm{pH} \approx 11.8$ and $T \approx 1.6°\mathrm{C}$ (Fig. 3**(b)**). Overall, we ran 40 iterations (five parameter draws each)

plus 10 LHS burn in runs, totaling 210 evaluations of f.

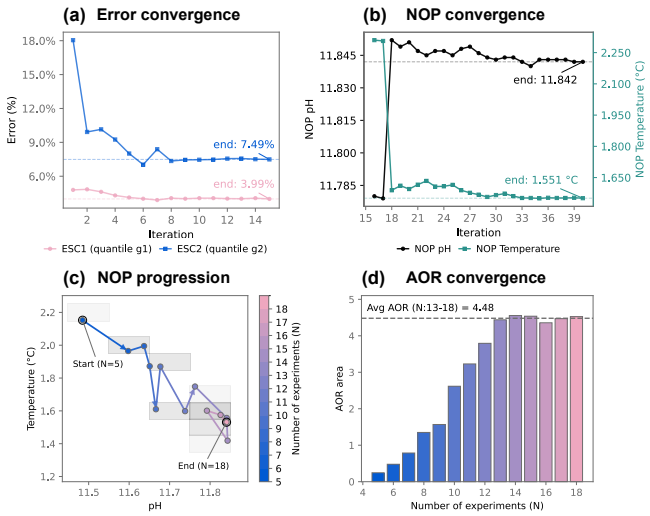

Figure 3. Convergence analysis.

Fig. 3**(c)** uses the edge-of-failure points (orange ♦) to guide laboratory experimentation. We start with $N = 5$ experiments (the minimum to estimate Eq. (14)) and iteratively add new experiments by choosing (pH, T) settings near these points. As N increases, the NOP converges; a color gradient shows experiment order, and the gray background partitions points into bins by rounding them to one decimal place; darker shading indicates a higher concentration of points. By $N = 10$, the NOP is near its final location, with most runs at $pH \approx 11.8$ and $T \approx 1.6$°C.

Fig. 3**(d)** confirms that 13 experiments suffice to estimate the PDS: tracking the AOR area for $N \in [5, 18]$ shows stabilization at $N = 13$, with an average area of ~ 4.48; additional runs will not expand the feasible region.

Validation and practical considerations

We validated the PDS against a DDS derived from standard Monte Carlo [7]. Rather than sampling across an uncertain parameter space, we fix the log-transformed rates at their expected values $\boldsymbol{\mu}(\mathbf{x})$ and compute $\max(g_1, g_2)$ on a 250×250 grid (totaling 62500 evaluations of f). We report that the PDS mean boundary closely aligns with the DDS boundary, while the PDS quantile boundary is strictly enclosed by the DDS, ensuring a conservative buffer against parameter uncertainty.

Implementation alongside traditional workflows: Traditional (large factorial) DoEs lack adaptivity and expend runs in infeasible or low-information regions; our approach concentrates experiments at feasibility boundaries to most effectively reduce uncertainty and define the operating window.

In process development, a small confirmation study is often included between process characterization and validation to address residual gaps; the proposed methodology in this paper can guide this stage. Furthermore, it naturally accommodates incremental updates with new data from scale-up and commercial manufacturing, making it well-suited for lifecycle management.

Batch-wise experimentation: As shown in Fig. 3, targeting edge-of-failure points reduced experiments from 18 to 13; an important saving given costly intermediates/reagents and reaction times of up to 17 hours. In practice, single-run experimentation is often infeasible, as samples are sent for analysis at central facilities. This can delay results for hours or days. Batch-sequential design can address this by bundling the experiments, with adaptive refinement between batches.

Model extensions: CQA data may exhibit significant analytical variation. To account for this uncertainty, implementing stochastic differential equations (rather than ordinary differential equations) could provide a more robust analysis by explicitly modeling measurement noise alongside process dynamics.

CONCLUSION

We present a quantile-based, adaptive methodology for probabilistic design space identification that efficiently characterizes feasible operating regions under parameter uncertainty; it yields explicit boundary constraints, identifies edge-of-failure points for targeted experimentation, and provides convergence criteria for reliable characterization.

Applied to industrial API manufacturing, the methodology identified the probabilistic design space in 13 rather than 18 experiments (28% reduction), saving costly intermediates and time. It also computed the largest inscribed operating window at the most flexible point (AOR area ≈ 4.5) and at the yield optimum (AOR area ≈ 1.8), capturing the robustness–performance trade-off.

Future extensions include: (i) strategies for cases where the process optimum lies on the feasibility boundary (vanishing flexibility); and (ii) multi objective formulations using convex combinations of yield, capacity, and unit cost to map trade-offs and define operating windows spanning multiple modes. A mode-switching policy could alternate between these optima as conditions change, enabling real-time optimization while maintaining compliance. In summary, quantile-based design space identification provides a fast route to process characterization that balances product quality and process efficiency.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from Novo Nordisk A/S, PS API manufacturing PhD program.

REFERENCES

1. FDA. Guidance for Industry: Q8(R2) Pharmaceutical Development. 2009.
2. Facco P, Dal Pastro F, Meneghetti N, Bezzo F, Barolo M. Bracketing the design space within the knowledge space in pharmaceutical product development. Ind. Eng. Chem. Res. 54:5128-5138 (2015). https://doi.org/10.1021/acs.iecr.5b00863
3. Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: state of the art. Chemical Engineering Science 63:4846-4872 (2008). https://doi.org/10.1016/j.ces.2007.11.034
4. Coleman MC, Block DE. Bayesian parameter estimation with informative priors for nonlinear systems. AIChE Journal 52:651-667 (2005). https://doi.org/10.1002/aic.10667
5. Sachio S, Kontoravdi C, Papathanasiou MM. A model-based approach towards accelerated process development: a case study on chromatography. Chemical Engineering Research and Design 197:800-820 (2023). https://doi.org/10.1016/j.cherd.2023.08.016
6. Moshiritabrizi I, McMullen JP, Wyvratt BM, McAuley KB. A comparative study of strategies for incorporating uncertainty in design space determination for pharmaceutical manufacturing. Ind. Eng. Chem. Res. 64:21658-21668 (2025). https://doi.org/10.1021/acs.iecr.5c03037
7. García-Muñoz S, Luciani CV, Vaidyaraman S, Seibert KD. Definition of design spaces using mechanistic models and geometric projections of probability maps. Org. Process Res. Dev. 19:1012-1023 (2015). https://doi.org/10.1021/acs.oprd.5b00158
8. Kusumo KP, Gomoescu L, Paulen R, García Muñoz S, Pantelides CC, Shah N, Chachuat B. Bayesian approach to probabilistic design space characterization: a nested sampling strategy. Ind. Eng. Chem. Res. 59:2396-2408 (2019). https://doi.org/10.1021/acs.iecr.9b05006
9. Kucherenko S, Giamalakis D, Shah N, García-Muñoz S. Computationally efficient identification of probabilistic design spaces through application of metamodeling and adaptive sampling. Comput. Chem. Eng. 2020;132:106608. doi:10.1016/j.compchemeng.2019.106608Geremia
10. Geremia M, Bezzo F, Ierapetritou MG. A novel framework for the identification of complex feasible space. Computers & Chemical Engineering 179:108427 (2023). https://doi.org/10.1016/j.compchemeng.2023.108427
11. Ding C, Ierapetritou M. A novel framework of surrogate-based feasibility analysis for establishing design space of twin-column continuous chromatography. International Journal of Pharmaceutics 609:121161 (2021). https://doi.org/10.1016/j.ijpharm.2021.121161
12. Metta N, Ramachandran R, Ierapetritou M. A novel adaptive sampling based methodology for feasible region identification of compute intensive models using artificial neural network. AIChE Journal 67: (2020). https://doi.org/10.1002/aic.17095
13. Laky D, Xu S, Rodriguez JS, Vaidyaraman S, García Muñoz S, Laird C. An optimization-based framework to define the probabilistic design space of pharmaceutical processes with model uncertainty. Processes 7:96 (2019). https://doi.org/10.3390/pr7020096
14. Moustapha M, Sudret B, Bourinet JM, Guillaume B. Quantile-based optimization under uncertainties using adaptive kriging surrogate models. Struct Multidisc Optim 54:1403-1421 (2016). https://doi.org/10.1007/s00158-016-1504-4
15. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press (2005). https://doi.org/10.7551/mitpress/3206.001.0001
16. Deisenroth MP. Efficient Reinforcement Learning using Gaussian Processes. PhD thesis, KIT; 2010.
17. Kim J, Song J. Quantile surrogates and sensitivity by adaptive gaussian process for efficient reliability-based design optimization. Mechanical Systems and Signal Processing 161:107962 (2021). https://doi.org/10.1016/j.ymssp.2021.107962
18. Wang Z, Shafieezadeh A. ESC: an efficient error-based stopping criterion for kriging-based reliability analysis methods. Struct Multidisc Optim 59:1621-1637 (2018). https://doi.org/10.1007/s00158-018-2150-9
19. Bano G, Facco P, Bezzo F, Barolo M. Probabilistic design space determination in pharmaceutical product development: a bayesian/latent variable approach. AIChE Journal 64:2438-2449 (2018). https://doi.org/10.1002/aic.16133
20. Meng Q, Bogle D, Charitopoulos VM. Probabilistic design space exploration and optimization via bayesian approach for a fluid bed drying process. European Journal of Pharmaceutical Sciences 210:107116 (2025). https://doi.org/10.1016/j.ejps.2025.107116