

Probabilistic Design Space Identification for Upstream Bioprocesses under Limited Data Availability

Ranjith Chiplunkar^{a,b}, Syazana Mohamad Pauzi^{a,b}, Steven Sachio^{a,b}, Maria M Papathanasiou^{a,b,*}, and Cleo Kontoravdi^{a,b,*}

^a The Sargent Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, United Kingdom

^b Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom

* Corresponding Authors: maria.papathanasiou11@imperial.ac.uk, cleo.kontoravdi98@imperial.ac.uk

ABSTRACT

Design space identification (DSId) and flexibility analysis are critical in process systems engineering, enabling efficient design of operating conditions. For bioprocess, these tasks are often hindered by the absence of reliable mechanistic models and limited experimental data. This paper presents an algorithm to address these challenges in bioprocesses. The methodology begins by constructing a Gaussian process (GP) model to predict key performance indicators (KPIs) from process inputs. Leveraging the probabilistic nature of GP predictions, we perform probabilistic design space identification (PDSId), characterizing each input point by its probability of feasibility which is the likelihood that constraints imposed on KPIs are satisfied. To visualize and analyse the feasibility space, contours at varying probability levels are identified using alpha shapes, which define deterministic boundaries corresponding to different confidence levels. This enables the quantification of volumetric process flexibility and operating ranges for each confidence level. The proposed methodology is applied to an antibody-producing Chinese hamster ovary (CHO) cell culture process, optimizing culture temperature and osmolality with respect to product yield and purity. Results are presented through probability heat maps and flexibility metrics, providing both qualitative and quantitative insights into feasibility and operational flexibility, thereby supporting informed decision-making in process design.

Keywords: Biosystems, Probabilistic design space identification, Flexibility analysis, Upstream bioprocesses

INTRODUCTION

Achieving efficient and reliable operations in complex large-scale systems, such as bioprocesses, requires carefully designed process conditions. In the biopharmaceutical industry, operational reliability is commonly assessed through key performance indicators (KPIs) like yield and purity meeting the established standards consistently. In this regard, design space identification (DSId) is a critical method for evaluating the feasibility of operating conditions. DSId involves isolating a subset of the broader knowledge space that encompasses all possible process inputs and satisfies the constraints imposed on the KPIs [1]. This subset, known as the design space, facilitates the evaluation of operating points and serves as a practical tool for process operators to assess the feasibility of the operating conditions.

Design space identification (DSId) methodologies are broadly classified into model-based and data-driven approaches. In the context of bioprocesses, model-based methods are generally preferred due to the impracticality of extensive experimentation. For instance, Yang and Ierapetritou [2] developed a modeling framework for DSId of a bioreactor producing monoclonal antibodies (mAbs). Similarly, Sachio et al. [3] proposed a model-based methodology for design space identification in chromatographic separation columns, encompassing the construction of geometric boundaries to enclose the design space and the quantification of its size to evaluate operational flexibility.

While mechanistic model-based approaches are effective in capturing the complexities of bioprocesses, they are often computationally demanding for DSId applications, primarily due to the need for extensive

sampling across the knowledge space to identify the design space accurately. To mitigate this computational burden, adaptive sampling techniques have been introduced, aiming to reduce the number of model evaluations required [4, 5]. Additionally, surrogate models have emerged as an efficient alternative for feasibility assessments, leveraging less computationally intensive techniques. Surrogate models such as Kriging regression [6] and artificial neural networks [7] have demonstrated significant potential in approximating complex process behaviors while minimizing computational overhead.

Process models are often parameterized using experimental data, which introduces stochasticity and parametric uncertainties, often related to measurement noise. As a result, the compliance of the KPI constraints cannot be determined definitively but only with a degree of confidence. Probabilistic design space identification (PDSId) addresses this by assigning each point in the knowledge space a probability of feasibility instead of a binary classification of feasible or infeasible. PDSId, however, increases computational complexity, requiring additional simulations across the uncertain parameter space. To address this, Laky et al. [8] proposed an optimization-based PDSId and flexibility analysis approach. Kusumo et al. [9] introduced a nested sampling strategy that prioritized maintaining high sample density in regions with a high probability of feasibility. Kucherenko et al. [10] developed an adaptive sampling method leveraging a metamodel that directly predicted the probability of feasibility based on the parameter space.

Research in DSId and PDSId typically assumes the availability of a mechanistic model. However, for complex systems like upstream biopharmaceutical processes, developing reliable mechanistic models is challenging due to a limited understanding of underlying mechanisms (e.g., host cell protein generation) or the need for simplifying assumptions. Consequently, data-driven approaches may be preferred as alternatives due to their ease of modeling. However, they often require extensive experimentation, which can be expensive and impractical, leading to sparse datasets. Such sparsity also means that data uncertainty becomes a significant factor that needs to be addressed.

We present a data-driven approach to address the challenges of modeling uncertainties in complex bioprocess systems. Specifically, we use Gaussian Process (GP) models to account for real-world data uncertainties stemming from measurement noise and inherent inter-experimental variability, enabling a probabilistic characterization of the design space. The framework consists of two main components. First, the GP model predicts key performance indicators (KPIs) based on input process variables, facilitating probabilistic modeling of these KPIs. The probability of feasibility is then calculated based on performance constraints, which indicates the

likelihood of meeting the KPI constraints for a given input. The second step of the framework conducts a quantitative analysis of operational flexibility for a given point. Following the method proposed by Sachio et al. [3], alpha shapes are used to define deterministic boundaries at varying levels of feasibility probabilities, allowing the quantification of process flexibility and operational ranges. This enables a detailed examination of trade-offs between process flexibility, performance, and confidence levels.

The framework is applied to experimental data from a fed-batch Chinese hamster ovary (CHO) cell culture producing a mAb to study the effects of osmolality and temperature on mAb yield and cell-derived impurity (host cell protein, HCP) generation. The results are visualized through probabilistic heat maps and flexibility metrics, offering valuable insights for process development.

METHODOLOGY

Experimental Setup

This study aims to investigate the effects of two critical upstream process parameters, culture temperature and osmolality, on HCP generation in mAb-producing CHO cell culture and their subsequent clearance in affinity and cation exchange chromatographic purification. Fed-batch culture was initiated at a seeding density of 2×10^5 cells/ml in 30 ml shake flasks using CD CHO medium (ThermoFisher Scientific, U.K.). The culture was incubated at 36.5°C with 8% CO₂ humidified air, shaking at 140 rpm. Every two days, 10% v/v EfficientFeed™ C+ AGT™ Supplement was added to each culture (ThermoFisher Scientific, U.K.). On day 5, as the culture entered the exponential phase, temperature and osmolality were manipulated while maintaining the desired viable cell count.

The study employs a central composite face-centered (CCF) design with two replicates, resulting in a total of 18 experimental runs manipulating three levels of temperature and osmolality. Specifically, temperature is varied between 32°C and 36.5°C, with 34°C as the middle level, on day 5. Osmolality is set between 410 mOsm/kg and 500 mOsm/kg, with a midpoint of 455 mOsm/kg, again on day 5 of the culture, starting from a value of 410 mOsm/kg. Osmolality was adjusted using 5M sterile NaCl, and OsmotechXT (Advanced Instruments Companies, U.K.) was used to measure the osmolality with a tolerance of 20 mOsm/kg. The culture was harvested when cell viability dropped below 80%.

Cell density and viability were determined daily using the trypan blue dye exclusion method and inverse microscopy. The extracellular mAb concentration was quantified using the biolayer interferometry BLItz system (Sartorius Stedim, U.K.). Residual HCP concentration was measured using the CHO cell HCP ELISA kit from Cygnus

Probabilistic Design Space Identification

The proposed methodology performs PDSId based on a GP model identified using the data obtained from the experiments described earlier.

GP model to predict KPIs

In this study, we assume that the uncertainties in the system model and measurement noise can be modeled using Gaussian distributions. Hence, we adopt the GP modeling framework to model the KPIs. GP regression is a nonparametric, nonlinear modeling technique that is particularly effective in scenarios with limited data and inherent model uncertainty. Its strong interpolation capabilities make it well-suited for cases where data is obtained through factorial designs that emphasize the extreme ends of input variability.

Let y and u represent the KPIs and input actions, respectively. The Gaussian process model used to predict the KPIs is expressed as follows.

$$y = f(u) + e \quad (1)$$

Here, $f(\cdot)$ represents the system model and e represents the measurement noise which is Gaussian with a mean of zero and variance of σ^2 , represented as $\mathcal{N}(0, \sigma^2)$. A GP model learns the function $f(\cdot)$ as a Gaussian probability distribution function rather than a deterministic entity. Hence, the predictions of KPIs would also be gaussian distributions. Let \mathcal{U}_T and \mathcal{Y}_T represent training data used to train the GP model. For any new point u_* in the knowledge space, the KPI prediction is a Gaussian distribution with the mean μ_* and variance Σ_* expressed as follows:

$$\mu_* = \mu(u_*) + \Sigma(u_*, \mathcal{U}_T)(\Sigma(\mathcal{U}_T, \mathcal{U}_T) + \sigma_e^2 \cdot I_{N \times N})^{-1}(\mathcal{Y}_T - \mu(\mathcal{U}_T)) \quad (2)$$

$$\Sigma_* = \Sigma(u_*, u_*) - \Sigma(u_*, \mathcal{U}_T)(\Sigma(\mathcal{U}_T, \mathcal{U}_T) + \sigma_e^2 \cdot I_{N \times N})^{-1}\Sigma(\mathcal{U}_T, u_*) \quad (3)$$

Here, $\mu(\cdot)$ and $\Sigma(\cdot)$ represent the mean and covariance functions of the prior distribution of the function. For the sake of brevity, the detailed explanation of GP modeling is omitted here, and readers are directed to Schulz et al. [11] for further information.

Estimating the probability of feasibility

Let us consider the DSId problem for the system with the following model and the constraints.

$$y = f(u); \quad u \in [u_l, u_u] \quad (4)$$

$$g(y) < 0 \quad (5)$$

In Eq. (4), $u \in [u_l, u_u]$ represents the knowledge space and Eq. (5) represents the constraints on the system. In this work we consider the case where the constraints are

directly imposed on the KPIs y which are represented as given below.

$$y_l \leq y \leq y_u \quad (6)$$

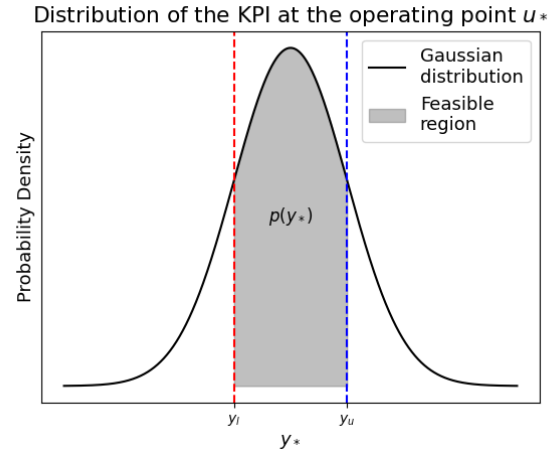


Figure 1. Depiction of the calculation of the probability of feasibility.

Here, y_l and y_u represent the lower and upper limits on the KPIs. With every point in the knowledge space being associated with its corresponding predicted KPI distribution, as described in the previous section, we proceed to estimate the probability of feasibility for the corresponding input point. In PDSId, the probability of feasibility for a given input u_* is defined as the probability with which the constraints specified in Eq. (6) are satisfied. Mathematically, it can be expressed as:

$$p(u_*) = \int_{y_l}^{y_u} p(y_* | u_*) dy_* \quad (7)$$

The above equation calculates the area under the distribution of the KPI at the operating point of u_* in the feasibility region. The quantity $p(u_*)$ hence denotes the probability at which the given operating point u_* is guaranteed to satisfy the constraints imposed on the KPIs. For the case of a Gaussian distribution and a single KPI, the probability of feasibility is represented as follows:

$$p(u_*) = \Phi\left(\frac{y_u - \mu_*}{\Sigma_*^{0.5}}\right) - \Phi\left(\frac{y_l - \mu_*}{\Sigma_*^{0.5}}\right) \quad (8)$$

Here, $\Phi(\cdot)$ represents the cumulative distribution function of a Gaussian distribution. The calculation of $p(u_*)$ is depicted in Fig. 1 where the shaded area between y_l and y_u represents the probability of feasibility. To characterize whole knowledge space with a probability of feasibility, we sample uniformly from the knowledge space using Sobol sequence. For each sample of $u_* \in [u_l, u_u]$ generated, the predicted KPI distribution is obtained using Eq. (2) and (3), following which the probability of feasibility is evaluated as per Eq. (8). This results in every possible

point in the knowledge space being characterized by a probability of feasibility. The results are visualized through the usage of probability heat maps which are presented in the Results and Discussion section.

While the probability of feasibility provides valuable information regarding feasibility, to facilitate further analysis, we draw contours at different probability levels and quantify the operational flexibility at each level, as described in the next section.

Quantitative Flexibility Assessment

To further enhance process understanding, we propose a deterministic quantitative flexibility assessment approach that builds upon the identified probabilistic design space. This method enables the quantification of process flexibility by defining deterministic boundaries at varying levels of probability of feasibility and computing a volumetric flexibility index for each level. By integrating both probabilistic information and geometric representations, this approach provides a structured way to assess the robustness and adaptability of the design space.

To achieve this, we employ alpha shapes as a geometric tool to delineate deterministic boundaries of the probabilistic design space. An alpha shape defines the spatial extent of a given set of points and can represent both convex and non-convex regions by adjusting a parameter known as the alpha radius. At large alpha radius values, the alpha shape becomes identical to a convex hull. As the alpha radius decreases, the alpha shape becomes increasingly non-convex. Alpha shapes can be exploited to identify design spaces as described in Sachio et al [3]. Briefly, the method has three major steps:

Step 1. Model sampling. The GP model is sampled via quasi-random Sobol sampling to obtain a dense dataset containing 16384 (2^{14}) input combinations and their corresponding probability of feasibility. In this work, we are working with 2D problems and hence 16384 samples is more than enough for an accurate analysis [3].

Step 2. Point cloud classification. A set of n_p different levels of feasibility probability is defined which is used as the constraints to characterize n_p alpha design spaces [3]. For a single feasible probability level, for example $\geq 95\%$, the dataset is screened by separating the points which satisfy $\geq 95\%$ from those which violate it. The group of points which satisfy this constraint forms the satisfied point cloud. The rest of the points form the violated point cloud.

Step 3. Alpha shape construction. An alpha shape is formed by solving for the alpha radius which defines the space occupied by the satisfied point cloud and does not capture any violated points. The resulting alpha shape is the alpha design space, with an explicit definition of the boundary and quantifiable size (volumetric flexibility metric). The schematic of the proposed methodology is presented in Fig. 2.

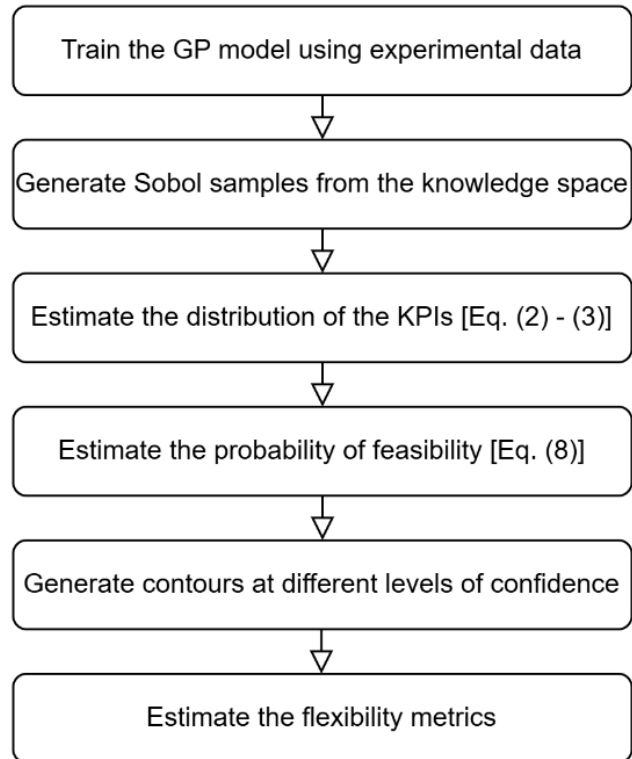


Figure 2. Schematic representation of the proposed methodology.

RESULTS AND DISCUSSION

The experimental data comprises 18 data points derived from a 2-factor, 3-level factorial design with two replicates for each operating condition. The inputs are the Osmolality (O) and Temperature (T) while the outputs are the Yield and the ratio of HCP to mAb concentration. A general heuristic regarding GP regression is that it requires 10 points per regressor, which would be 20 for two regressors. This is close to 18 points; however, 9 of these are at the same input conditions.

Of the 18 points, 12 are used to train the GP model, and 6 are reserved for validation. Gaussian process regression with a co-regionalized RBF kernel is trained in Python. The average values for mAb yield and HCP-to-mAb concentration ratio (HCP/mAb) in the experiments are 1.25 and 0.25, respectively. Therefore, the KPI constraints are defined such that the minimum mAb yield is 1.25, and the maximum HCP/mAb ratio is 0.25. Consequently, the constraints for the KPIs are as follows:

$$1.25 \leq \text{Yield}; \text{HCP/mAb} \leq 0.25 \quad (9)$$

KPI Predictions in the Knowledge Space

First, the GP model is trained, and the Sobol samples generated from the knowledge space are used to predict

the distribution of the KPIs. Figure 3 illustrates the mean of the KPIs for each input in the knowledge space, providing a convenient way to visualize KPI variability within the knowledge space. It can be observed that higher temperatures and mid-level osmolality values generally favor higher yields, with mid-level osmolality also resulting in lower HCP concentrations.

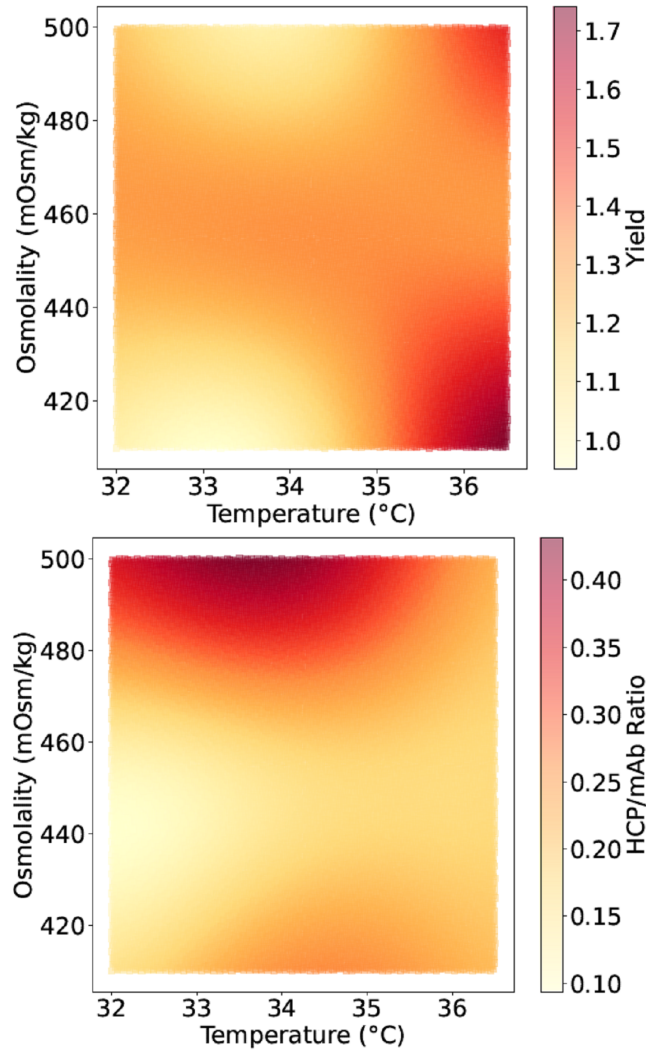


Figure 3. Heat map depicting the means of the predicted KPIs.

However, there are operating points, such as $[T, O] \approx [32 - 33, 440-440]$ which can achieve high purity but low yield. Hence, one must combine all the KPI constraints to better assess these trade-offs, which is the DS representation of the feasibility analysis.

Probabilistic Design Space Identification

To visualize the design space with respect to both KPIs, the probability of feasibility is estimated for each point in the knowledge space, and the results are presented in Fig. 4. This heatmap, shows regions with higher

probabilities of satisfying the constraints in warmer red tones, while lower probability regions are represented in yellow tones.

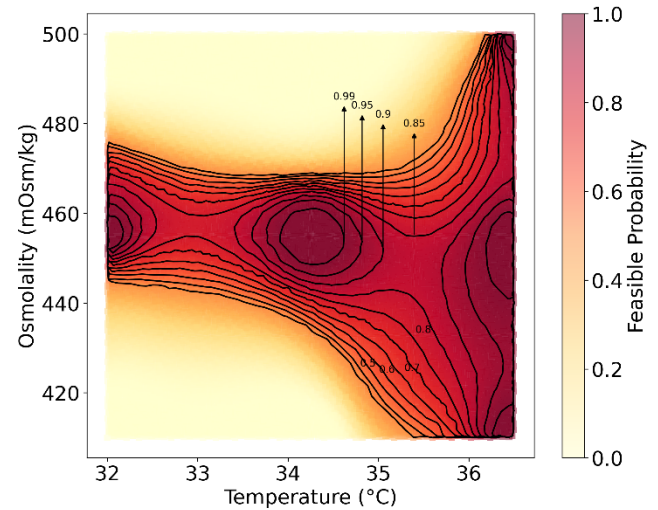


Figure 4. Probability heat map with the contours at different levels of confidence.

The results indicate that the constraints in Eq. (9) are likely to be satisfied in the regions where $[T, O]$ are $[34.25, 455]$ and $[36.5, 410-500]$. Thus, higher temperatures allow for more variability in osmolality while satisfying the KPI constraints. This is because higher temperatures promote cell growth and productivity, while lower temperatures may prolong the culture duration but can also lead to increased host cell protein accumulation. To gain further insights, the alpha shapes-based boundary detection method is applied to obtain contours at different probability levels. A contour at a confidence level α encloses all points with a probability of feasibility greater than or equal to α . In Fig. 4, contours are provided for probability values of $[0.99, 0.95, 0.90, 0.85, 0.80, 0.77, 0.70, 0.65, 0.60, 0.55, 0.5]$, with the lowest level set at 0.5, as this is typically the threshold for accepting or rejecting an operating point. The heatmap representation provides a convenient way to visualize the design space and derive insights into how feasibility varies across the knowledge space.

Flexibility Assessment

To quantify the information depicted in the contours, the volumes of the space enclosed by each contour can be calculated. This provides a means of assessing the flexibility of operation at different confidence levels. The volumes for the heatmap in Fig. 4 are presented in Fig. 5. These volumes are normalized to represent the fraction of the total knowledge space enclosed within the contour. For instance, operating at a probability of feasibility of at least 90% probability of feasibility offers flexibility within up to 10% of the total knowledge

space. In addition to assessing flexibility, this approach allows for the evaluation of the trade-off between flexibility and confidence level, where a higher confidence level reduces flexibility and vice versa.

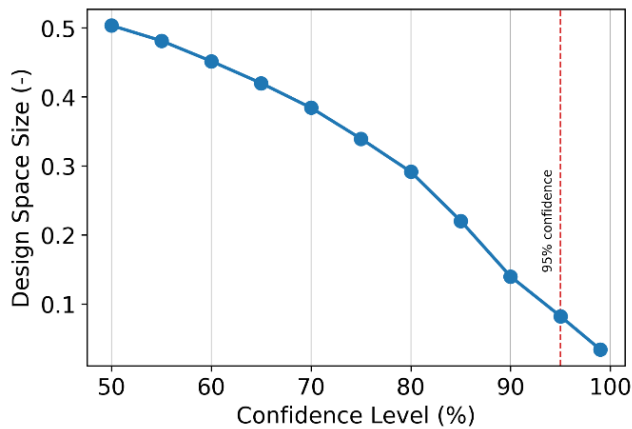


Figure 5. Design space size at various levels of confidence as a metric of flexibility in operation.

Design space for different KPI constraints

Finally, if the DS for a particular KPI constraint does not meet the desired criteria, alternative constraints can be explored to assess feasibility and flexibility for different operating goals. Figure 6 illustrates this exploration across four scenarios. To maintain clarity, we plot the contours of 50% probability of feasibility, and the flexibility metrics for all the levels mentioned earlier are shown in Fig. 7.

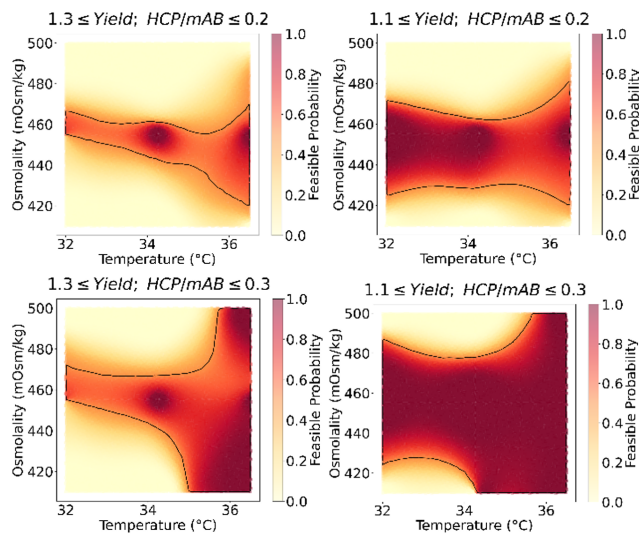


Figure 6. Probability heat map for different KPI constraints

The first scenario represents a stringent case, requiring a yield of at least 1.3 and an HCP/mAb ratio of at most 0.2. In the second scenario, the yield constraint is

relaxed to 1.1, which increases the DS. In this case, mid-level osmolality and lower temperatures are preferred. The third scenario sets the minimum yield to 1.3, while relaxing the HCP/mAb constraint to 0.3. Here, higher temperatures and lower osmolality are generally favored. Finally, both yield and purity constraints are relaxed to 1.1 and 0.3, respectively, resulting in a significantly larger DS. These trends align with empirical expectations. This can also be observed in Fig. 7, which compares the size of the design space at different levels of probability of feasibility for various KPI constraints. The most ambitious constraints, such as high yield and low impurity, lead to a smaller design space at every probability level, while relaxed constraints result in a larger design space. This representation provides a bird's-eye view of the trade-offs between the required probability of feasibility, operational flexibility, and the ambitiousness of the KPI constraints. Experimental validation of these computational findings is underway.

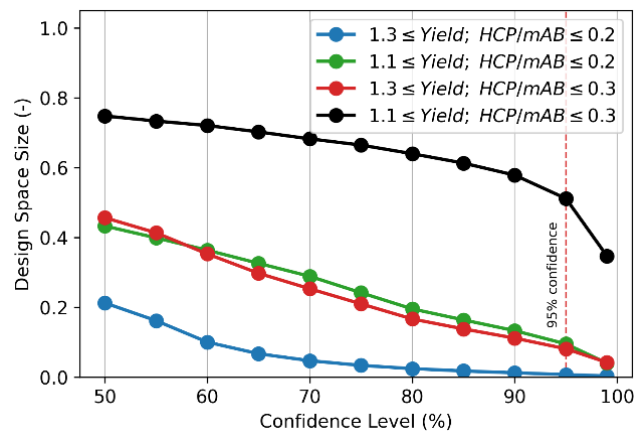


Figure 7. Comparing the design space sizes for various KPI constraints

CONCLUSION

Assessment of the feasibility of operating points and the flexibility of those operations is vital for novel and complex processes such as bioprocesses. This paper presents a methodology to achieve these goals for systems with limited understanding. The methodology includes three key features: estimating the probability of feasibility of an operating point, identifying the design space for a given probability of feasibility or confidence level, and quantifying the flexibility of operations at a given confidence level. The obtained results are presented through probabilistic heat maps and flexibility metrics, providing actionable insights for process development scientists. In the presented CHO cell study, initial experimental data was used to develop a GP model of how two process conditions, temperature, and osmolality, affect recombinant product yield and purity.

Subsequent system analysis using the proposed methodology provided insights for balancing purity-yield trade-offs and guide further experimentation and process design. The future research focus is on performing additional experiments to confirm the DS and improve the accuracy of the model.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from the Engineering and Physical Sciences Research Council U.K. (EP/X024156/1 and EP/W035006/1). Support from the UKRI Impact Acceleration Account (EP/X52556X/1) is also gratefully acknowledged. Additionally, Mohamad Pauzi, S. acknowledges the Ministry of Higher Education, Malaysia, for awarding a scholarship through the IPTA Academic Training Scheme, as well as Universiti Teknologi MARA, Malaysia, for the Academic Staff Scholarship.

REFERENCES

1. Kasemiire A, Avohou HT, De Bleye C, Sacre PY, Dumont E, Hubert P, Ziemons E. Design of experiments and design space approaches in the pharmaceutical bioprocess optimization. *European Journal of Pharmaceutics and Biopharmaceutics*. 2021 Sep 1;166:144-54.
2. Yang O, Ierapetritou M. mAb production modeling and design space evaluation including glycosylation process. *Processes*. 2021 Feb 9;9(2):324.
3. Sachio S, Kontoravdi C, Papathanasiou MM. A model-based approach towards accelerated process development: A case study on chromatography. *Chemical Engineering Research and Design*. 2023 Sep 1;197:800-20.
4. Ito K, Couckuyt I, d'Ippolito R, Dhaene T. Design space exploration using Self-Organizing Map based adaptive sampling. *Applied Soft Computing*. 2016 Jun 1;43:337-46.
5. Zhao F, Grossmann IE, García-Muñoz S, Stamatis SD. Design space description through adaptive sampling and symbolic computation. *AIChE Journal*. 2022 May;68(5):e17604.
6. 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*. 2021 Nov 20;609:121161.
7. 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*. 2021 Feb;67(2):e17095.
8. 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*. 2019 Feb 14;7(2):96.
9. 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. *Industrial & Engineering Chemistry Research*. 2019 Nov 26;59(6):2396-408.
10. Kucherenko S, Giamalakis D, Shah N, García-Muñoz S. Computationally efficient identification of probabilistic design spaces through application of metamodeling and adaptive sampling. *Computers & Chemical Engineering*. 2020 Jan 4;132:106608.
11. Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of mathematical psychology*. 2018 Aug 1;85:1-6.

© 2025 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See <https://creativecommons.org/licenses/by-sa/4.0/>

