Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Design Space Exploration via Gaussian Process Regression and Alpha Shape Visualization

Elizaveta Marich[a,b], Andrea Galeazzi[a,b], Steven Sachio[a,b], Foteini Michalopoulou[a,b], and Maria M. Papathanasiou[a,b]*

[a] Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] The Sargent Centre for Process Systems Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
* Corresponding Author: maria.papathanasiou11@imperial.ac.uk.

ABSTRACT

This study introduces a novel methodology that combines Gaussian process regression (GPR) with alpha shape design space reconstruction to visualize multi-dimensional design spaces. The proposed GPR surrogate approach incorporates a kernel optimization step, employing a greedy tree search strategy to identify the optimal combinatorial kernel from a selection of base kernels. This approach efficiently evaluates design spaces around specific points of interest, enabling alpha shape reconstruction. The methodology's adaptability is demonstrated through its application to both lower-dimensional (2D and 3D) cases and more complex, higher-dimensional systems (up to 7D), showcasing its scalability and versatility. Its effectiveness is further validated by its ability to generate accurate surrogate models from limited data. Overall, this study presents a robust framework that leverages GPR surrogate modeling and alpha shape reconstruction to facilitate design space evaluation in complex, multidimensional engineering problems.

Keywords: Design Space Identification, Surrogate Modelling, Gaussian Process Regression, Alpha Shapes, Kernel Optimisation.

INTRODUCTION

The development of advanced computational tools and techniques has significantly enhanced the ability to solve challenging high-dimensional chemical engineering problems affected by uncertainty and complexities of chemical processes. Traditional computational tools are often restricted to simpler, lower-dimensional problems mainly due to numerical complexity. In the task of design space (DSp) identification, a major limitation is the substantial computational resources needed to extract the required knowledge from complex mechanistic models. Given that chemical processes are often characterized by intricate systems of equations, surrogate-based methodologies hold significant potential for enabling design space identification of complex and practical industrial scenarios by reducing the need for extensive modeling. Models like High Dimensional Model Representation (HDMR) [1] and Partial Least Squares (PLS) [2] have been directly utilized in DSp evaluation. More recently, Geremia *et al.* [3] analyzed a range of surrogate modeling techniques for DSp and found that the Gaussian Process (GP) is among the best-performing ones.

Despite surrogate models' potential to aid process design and control, producing highly accurate surrogates remains computationally demanding. This is particularly relevant when analyzing problems of increased dimensionality, where computational requirements can scale exponentially with process complexity. Moreover, meaningful analysis and visualization of the design space are crucial in high-dimensional problems to effectively evaluate quality criteria. The presented study aims to investigate the effectiveness of Gaussian process regression as a surrogate modeling technique within the context of multi-dimensional design space identification for enabling visual representation through alpha shape reconstruction.

BACKGROUND

Feasibility and Design Space Identification

Design space identification is a form of feasibility

analysis that aims to determine the feasible region of a design space, defined by the set of input variables that ensure the system output meets all specified constraints. The concept of design space identification is particularly relevant in the framework of Quality by Design (QbD), extensively applied in pharma and biopharma [4], which aims to design robust and efficient processes where quality criteria, or Key Performance Indicators (KPIs), are met in product output at varying input conditions. Design space identification is applied in this context to provide a fundamental understanding of the process operative range and is a key tool for enabling more advanced computational techniques for process enhancement, such as flexibility analysis.

The design space problem is formulated as shown in Eq. (1).

$$\begin{aligned} y &= f(\boldsymbol{\theta}) \\ \boldsymbol{\theta}_L \leq \boldsymbol{\theta} &\leq \boldsymbol{\theta}_U \\ g(\boldsymbol{y}) &\leq 0 \end{aligned} \quad (1)$$

where $\boldsymbol{y}$ is the vector of selected KPIs, f is the system model, $\boldsymbol{\theta}$ is the vector of input variables, and $\boldsymbol{\theta_L}$ and $\boldsymbol{\theta_U}$ are the vectors of the lower and upper bounds of such variables, respectively. Finally, g represents the target KPI constraints that must be satisfied.

Alpha shapes DSp Reconstruction

Various methodologies exist for identifying and visualizing design spaces. Among them, the alpha shape method has demonstrated high accuracy in capturing both convex and non-convex feasibility regions in two- and three-dimensions [5]. Unlike convex hull methods, alpha shapes provide a flexible geometric representation that adapts to the underlying dataset structure. An alpha shape defines a boundary that encapsulates a set of points based on a shape-controlling parameter, the alpha radius $\alpha_r \in [0, +\infty]$. This parameter determines the level of detail in the reconstructed shape, allowing for fine or coarse representations of the feasibility region depending on the dataset S and application requirements [6].

Gaussian Process Regression

In this study, Gaussian Process Regression (GPR) is employed as a surrogate model to approximate the system function (f in Eq. (1)). GPR is a non-parametric probabilistic approach that assumes the underlying function follows a multivariate Gaussian distribution, characterized by a mean function, $m(x)$, and a covariance function, called kernel, $k(x, x')$. Common kernel functions include the Radial Basis Function (RBF, also known as the Squared Exponential, SE), Rational Quadratic (RQ), Periodic (Per), and Linear (Lin) kernels. This approach enables flexible modeling of complex, non-linear relationships without requiring an explicit parametric form, instead relying on kernel functions to capture dependencies [7]. The kernel function plays a critical role in capturing relationships between data points, allowing GPR to make interpolative predictions. To model complex dependencies, composite kernels, formed by combining base kernels through addition or multiplication, can be employed, generating effective covariance functions while enhancing model expressiveness, as depicted in **Figure 1**.

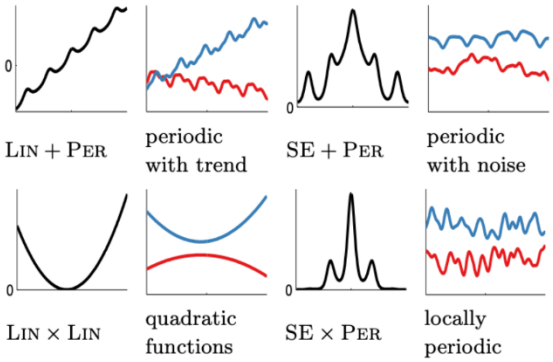

Figure 1: Examples of addition and multiplication of base kernels, in black, and the functions they are able to capture, colored, adapted from [8].

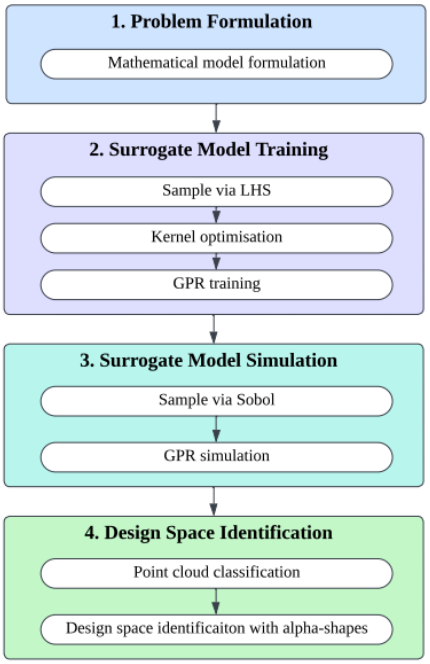

Figure 2: Design Space Identification Framework Schematic.

METHODOLOGY

The methodology proposed in this work, shown in **Figure 2**, is aimed at the identification of the design space of high-dimensional constrained mathematical problems through Gaussian process regression surrogate modeling. In the context of chemical engineering,

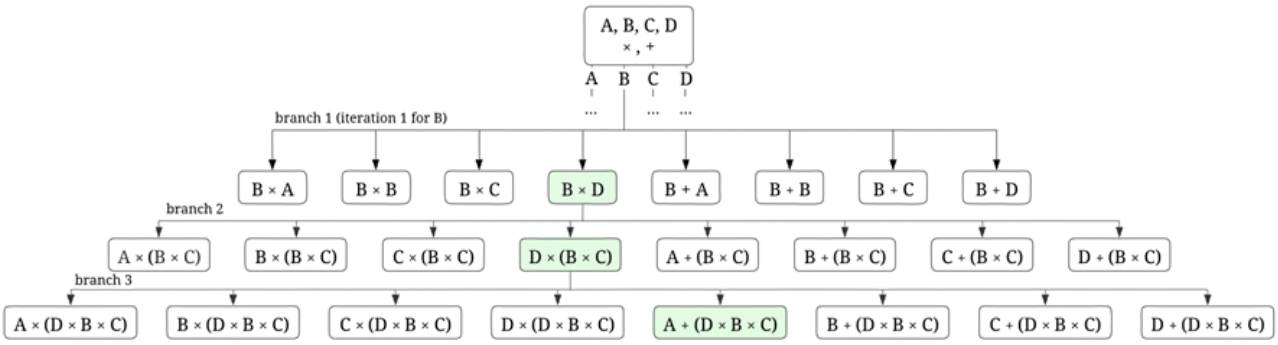

Figure 3: Schematic of the kernel optimisation algorithm.

the investigated variables represent design decisions, with the objective function being the process design and/or product specification. This methodology can be integrated as part of the QbD framework by facilitating the exploration of design space and understanding of how process disturbances, process variables, and design variables influence the chemical process and its KPIs.

The computations for this study are conducted in Python and performed using an Intel® Core™ i7–13,700 CPU @ 2.10 GHz. Nevertheless, the proposed framework can be easily implemented and adapted to different programming platforms with minimal changes to the procedure.

Stage 1: Problem Formulation

The first step of the methodology is defining the objective function, $f(\boldsymbol{x})$, and its constraints, $g(\boldsymbol{x})$, for the analysis. In chemical engineering applications, the objective function may represent the chemical process under investigation or an optimization function that evaluates relevant KPIs. An example of a mathematical problem formulation is presented in **Table 1**.

Table 1: Example of a mathematical problem analyzed.

The Modified Gomez Function (D=3)	
Objective function and constraints	**Search space**
$\min f(\boldsymbol{x}) = -\sin(4\pi x_1) + 2sin^2(2\pi x_2) + 4sin^3(\pi x_3)$ $s.t. \quad g_1(\boldsymbol{x}) = f \leq 0$	$[-1, 1]^3$

Stage 2: Surrogate Model Training and Kernel Optimisation

The problem specifications define the search space within which the knowledge space is sampled using Latin Hypercube Sampling (LHS) for the vector of investigated variables $\boldsymbol{x} = [x_1, x_2, \ldots, x_{n_x}]$ of size n_x. The sampled data and computed corresponding outputs constitute the training data on which the surrogate model learns about the system. This training data is thus fed to a kernel optimization algorithm for selecting the best-fit GPR kernel.

The kernel optimization algorithm utilizes a combinatorial method for new kernel construction with a greedy tree search algorithm, as presented by Galeazzi *et al.* [8]. The greedy tree search selects optimal solutions at each step, ensuring the strategy remains computable within reasonable time and computational constraints. New kernels are constructed using four basic kernels (RBF, RQ, Periodic, Linear) and operators (multiplication, addition) to capture intricate model behaviors, as described above. This process is illustrated in the first branch for kernel B in **Figure 3**. Once a branch is composed of all potential kernels, they are evaluated based on the mean absolute error (MAE) parameter, and the kernel with the minimum MAE from the iteration is recorded and used as the basis for generating options in the next branch. This search continues until the user-specified tree depth is achieved, limited to five iterations for this study. All best kernels from each branch are stored, and the final best kernel is selected from this list, thus balancing the model's complexity and performance while mitigating the risk of overfitting that could arise with additional iterations without yielding significant improvements.

Stage 3: Surrogate Model Testing

To gauge and test the surrogate model performance, after its training, a dense sample set is generated from the benchmark case studies function and MAE and Mean Absolute Percentage Error (MAPE) are re-evaluated for accuracy. Several reference graphs of the true function are generated to provide a visual benchmark, further aiding in the validation of the surrogate's performance. In a practical scenario where a computationally expensive high-fidelity model is implied this step would be skipped. However, in this study, this step is key to understanding the methodology potential.

Stage 4: Design Space Identification

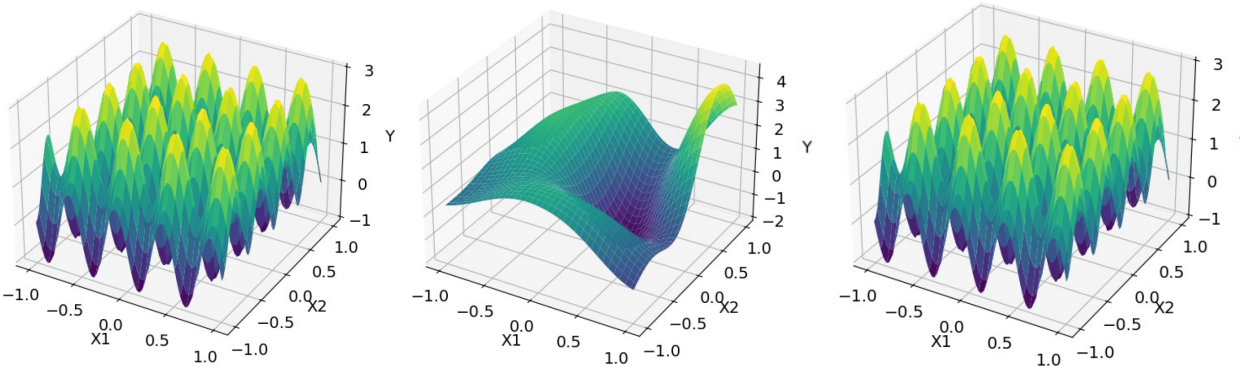

Figure 4: (a) Original function, (b) GPR surrogate prediction without kernel optimization, and (c) GPR surrogate prediction with kernel optimization.

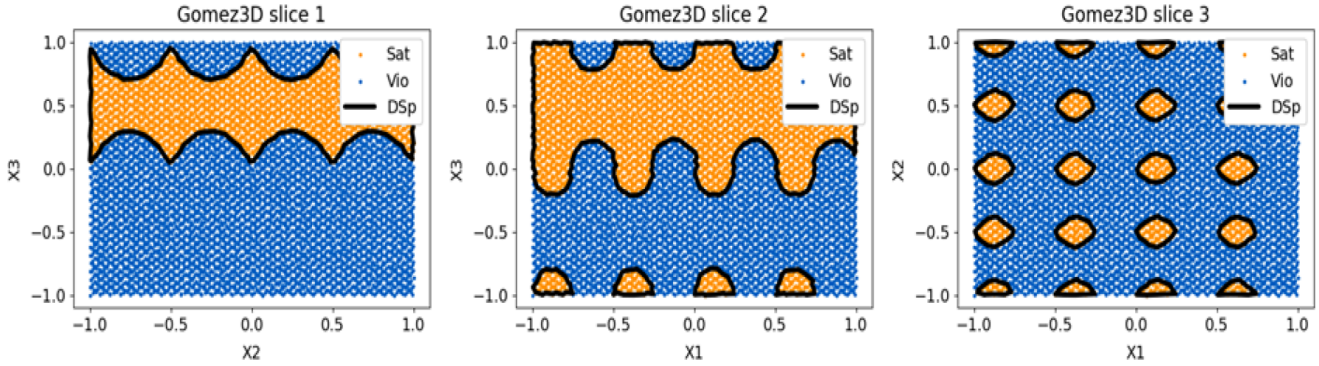

Figure 5: Design space (DSp) of the 3D modified Gomez function reconstructed with alpha shapes on 2D slices.

In this stage, the feasibility problem under investigation is developed. In the context of chemical engineering, it involves the selection of design decision variables and KPIs for analysis requirements for which the acceptable region of operation will be defined. In this study, where mathematical functions are employed, the feasibility function is equivalent to the objective function defined in Stage 1, $f(\boldsymbol{x})$, constrained to a user-defined value for DSp evaluation.

The design space is first defined by declaring a nominal point, which is user-selected and can represent any point of interest, such as an optimum. To enable the application of alpha shape reconstruction on higher-dimensional problems, a slicing technique is applied to create lower-dimensional planar slices around the point of interest. As the first stage of DSp identification, the surrogate model is densely resampled around the selected nominal point of interest within the boundaries applied. The slicing function is applied at this stage to create planes for DSp reconstruction. The number of orthogonal slices generated is determined only by the dimensionality of the problem and whether 2D or 3D plots are desired for visualization. Finally, the *dside* package is utilized on each slice for design space identification with alpha shape reconstruction. For a more detailed explanation of the package working principles, the reader is referred to Sachio *et al.* [9].

RESULTS

Kernel Optimization in GPR

The quality of design space evaluation is directly linked to the accuracy of predictions made by the surrogate model. Therefore, the initial step in the methodology is to ensure that the GPR can effectively capture even the most intricate relationships within the data. Typically, a basic kernel such as the RBF or Matern 5/2 is employed in GPR analysis. However, this is often insufficient for capturing the complexities of intricate data trends. In chemical engineering, experimentally or computationally collected data often exhibit behaviors that are not characterized a priori, necessitating further analysis to confirm whether the selected kernel is appropriate for the given data type. To avoid this laborious analysis and to create a versatile framework capable of handling any data complexity, kernel optimization was incorporated into the methodology. This optimization allows the construction of a composite kernel, enabling the capture of intricate data trends.

The effectiveness of the developed kernel optimization function was demonstrated using the modified Gomez function, proposed by Geremia *et al.* [3] and shown in **Table 1**. Both the optimized and non-optimized cases

were trained using the same LHS training set of size 38, ensuring that the training set and its size did not influence the accuracy of the surrogate predictions. The original function (**Figure 4.a**), generated using true values, serves as the benchmark for evaluating the GPR surrogate models, illustrating a significant improvement in prediction quality with and without kernel optimization, **Figures 4.c** and **4.b** respectively. This improvement is reflected in the performance metrics reported in **Table 2**, where near-perfect predictions were generated with kernel optimization. GPR training grows cubically as the number of samples increases and this optimization succeeds in reducing the number of training points needed. However, the time required for model training grows linearly with the number of kernels investigated, presenting a trade-off question between computational efficiency and predictive performance.

Table 2: Kernel optimization function performance.

	MAE	R^2	Accuracy	F1 score
Without kernel optimisation	1.43	0.47	73.89%	73.97 %
With kernel optimisation	0.01	1.00	99.87%	99.87 %

Application to 2D and 3D problems

The application to 2D and 3D case studies is useful in understanding the performance of the proposed methodology and developing heuristics for sample size selection. The latter is of particular importance as with increasing dimensionality the need for balancing computational demand and prediction accuracy arises. Results on the 2D study adapted from Sasena *et al.* [10] showed that a small number of training samples, n = 10D, was sufficient to achieve a 100% F1 score resulting in the DSp being captured fully. However, as expected, with increased dimensionality, and, therefore, problem complexity, the training set requirement increased.

The methodology's efficiency was further validated in the more complex 3D case study using the Modified Gomez function adapted from Geremia *et al.* [3]. As the dimensionality and complexity of the problem increased, a larger number of training samples was required, with an acceptable MAPE achieved only after the fourth incremental increase, where $n = 2(10D) = 60$. As shown in **Table 3**, the model struggled to achieve a good fit with smaller sample sizes, with MAPE ranging from nearly 550% to approximately 140%. This difficulty was primarily due to the surrogate's challenge in capturing the rapid fluctuations in the function's output. A significant improvement in the accuracy of the GPR predictions occurred between the third (n = 53) and fourth (n = 60) iterations. This change is likely due to the search space becoming sufficiently well-sampled, enabling the model to capture the nonlinearities of the original function more effectively, thereby reducing the model's uncertainty. The low standard deviations observed across 10 runs for the n = 60 for MAE (0.00163) and MAPE (0.00521) confirm that the methodology consistently produces reliable and accurate results.

While the metrics generally improved with an increase in the training samples, an exception was noted in the second incremental increase (n = 45), where MAPE increased despite improvements in MAE and R^2. This anomaly might be attributed to the overestimation of very small actual values, which disproportionately impacts MAPE due to its sensitivity to such values.

Table 3: Methodology results for the 3D case study across different training set sizes, n. Increments refer to the percentile increase of the initial 10D training set.

Increment	n	MAE	MAPE	R^2
Starting	30	1.09952	549.45%	0.67
+25%	38	1.03456	441.92%	0.72
+50%	45	1.01836	514.45%	0.73
+75%	53	0.26165	143.90%	0.98
+100%	60	0.00451	1.74%	1.00
Benchmark	300	0.00002	0.01%	1.00

It is important to note that while the methodology is capable of analyzing the system as a 3D problem, as demonstrated by the results for n = 60, 2D planar slicing offers a more understandable visualization of the problem and provides further insights into the method's performance. Such DSp slices are presented in **Figure 5**.

Application to high-dimensional problems

The results from both 2D and 3D case studies convincingly demonstrate that the proposed methodology utilizing GPR for surrogate modeling and alpha shapes for DSp reconstruction is effective. Based on the analysis of the training size required to encapsulate complex functions, training sample sizes for high dimensional problems were selected following the $2(10D)$ rule.

For higher-dimensional analysis common engineering problems for optimization, listed in **Table 4,** were adapted from Bouhlel *et al.* [11], namely the Gas Transmission Compressor Design (GTCD), Pressure Vessel Design (PVD), and Speed Reducer 7 (RS7) case studies. These cases were chosen because the parameters span different scales, potentially complicating analysis during kernel optimization. Larger-magnitude parameters may disproportionately influence the kernel, causing numerical deviations and reducing model accuracy. However,

the proposed methodology mitigated these issues, achieving high accuracy (**Table 4**) and demonstrating robustness despite parameter variability.

Table 4: Methodology results for the high-dimensional case studies.

Name	D	n	MAE	MAPE	R^2	F1 Score
GTCD	4	80	2.79E+01	3.86%	1	99.91%
PVD	4	80	1.20E-02	9.09%	1	100.00%
RS7	7	140	9.26E-01	0.02%	1	92.59%

The application of the $2(10D)$ rule was effective for 4D cases but proved less efficient in the 7D study, as expected. Higher-dimensional problems typically require exponentially larger training sets, a trend consistent with findings from the 2D and 3D studies. The 7D results exhibited significant deviations across slices, with a tendency for false negatives in half of them, though no false positives were observed. This suggests that the surrogate model struggled to converge optimally in certain slices, indicating that a denser sample set might be needed for accurate design space evaluation. However, this would make alpha shape reconstruction computationally prohibitive. By reducing the 7D problem to 3D through slicing, the methodology enabled DSp evaluation with alpha shapes, demonstrating its efficiency in handling complex design spaces, as shown in **Figure 6**.

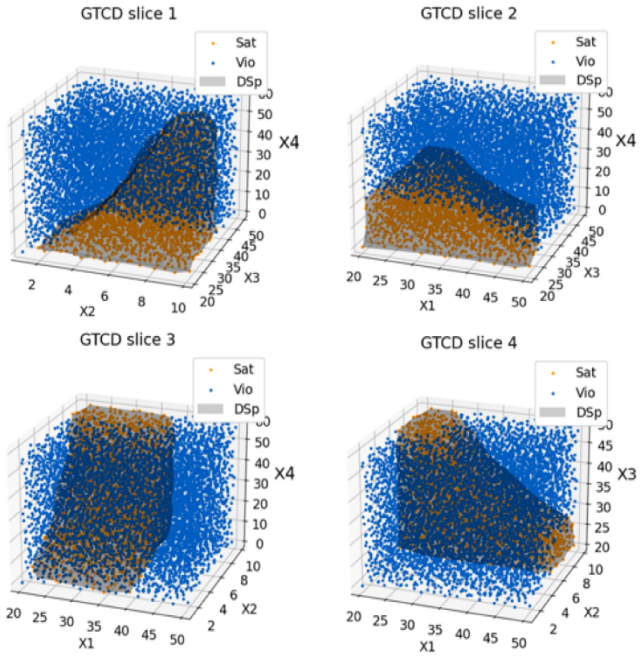

Figure 6: DSp of the 4D GTCD Function Reconstructed with Alpha Shapes on 3D Slices.

CONCLUSIONS

This study presents a methodology that integrates surrogate modeling with alpha shape design space identification to reconstruct complex, high-dimensional design spaces. The framework employs Gaussian Process Regression (GPR) with an optimized kernel selection strategy to capture intricate relationships in complex mathematical systems. GPR enables accurate modeling and representation of feasibility regions, while alpha shape design space identification facilitates the slicing of multidimensional design spaces into two- or three-dimensional projections around a point of interest, allowing for multi-perspective visualization. Its adaptability to different design criteria makes it a powerful tool for evaluating process variations and optimizing design decisions.

The methodology's effectiveness is validated across mathematical problems ranging from two to seven dimensions, demonstrating its scalability and applicability to both low- and high-dimensional challenges.

Ongoing research is focusing on transferring this methodology to real-world problem cases. Future research should tackle refining sample size selection to develop a universal heuristic that balances surrogate model accuracy with GPR computational costs.

ACKNOWLEDGEMENTS

Funding from the UK EPSRC for the i-PREDICT: Integrated adaptive pRocEss DesIgn and ConTrol (Grant EP/W035006/1) is gratefully acknowledged.

REFERENCES

1. I. Banerjee and M. G. Ierapetritou, 'Model Independent Parametric Decision Making', Annals of Operations Research, vol. 132, no. 1–4, pp. 135–155, Nov. 2004, doi: 10.1023/B:ANOR.0000045280.55945.e8.
2. P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, and M. Barolo, 'Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development', Ind. Eng. Chem. Res., vol. 54, no. 18, pp. 5128–5138, May 2015, doi: 10.1021/acs.iecr.5b00863.
3. M. Geremia, F. Bezzo, and M. G. Ierapetritou, 'A novel framework for the identification of complex feasible space', Computers & Chemical Engineering, vol. 179, p. 108427, Nov. 2023, doi: 10.1016/j.compchemeng.2023.108427.
4. A. S. Rathore and H. Winkle, 'Quality by design for biopharmaceuticals', Nat Biotechnol, vol. 27, no. 1, pp. 26–34, Jan. 2009, doi: 10.1038/nbt0109-26.
5. I. Banerjee and M. G. Ierapetritou, 'A novel feasibility analysis approach based on dimensionality reduction and shape reconstruction',

in Computer Aided Chemical Engineering, vol. 20, L. Puigjaner and A. Espuña, Eds., in European Symposium on Computer-Aided Process Engineering-15, 38 European Symposium of the Working Party on Computer Aided Process Engineering, vol. 20. , Elsevier, 2005, pp. 85–90. doi: 10.1016/S1570-7946(05)80136-1.
6. H. Edelsbrunner and E. P. Mücke, 'Three-dimensional alpha shapes', ACM Trans. Graph., vol. 13, no. 1, pp. 43–72, Jan. 1994, doi: 10.1145/174462.156635.
7. C. E. Rasmussen and C. K. I. Williams, Gaussian processes for machine learning. in Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2006.
8. A. Galeazzi, F. de Fusco, K. Prifti, F. Gallo, L. Biegler, and F. Manenti, 'Predicting the performance of an industrial furnace using Gaussian process and linear regression: A comparison', Computers & Chemical Engineering, vol. 181, p. 108513, Feb. 2024, doi: 10.1016/j.compchemeng.2023.108513.
9. S. Sachio, C. Kontoravdi, and M. M. Papathanasiou, 'A model-based approach towards accelerated process development: A case study on chromatography', Chemical Engineering Research and Design, vol. 197, pp. 800–820, Sep. 2023, doi: 10.1016/j.cherd.2023.08.016.
10. M. J. Sasena, P. Papalambros, and P. Goovaerts, 'Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization', Engineering Optimization, vol. 34, no. 3, pp. 263–278, Jan. 2002, doi: 10.1080/03052150211751.
11. M. A. Bouhlel, N. Bartoli, R. G. Regis, A. Otsmane, and J. Morlier, 'Efficient global optimization for high-dimensional constrained problems by using the Kriging models combined with the partial least squares method', Engineering Optimization, vol. 50, no. 12, pp. 2038–2053, Dec. 2018, doi: 10.1080/0305215X.2017.1419344.