Journal of Chromatography A 1759 (2025) 466193

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Machine learning enhanced process design in protein a chromatography

Andrea Galeazzi [a,b], Steven Sachio [a,b], Elizabeth Edwards [c], David Hilton [c], Maria M. Papathanasiou [a,b,*]

[a] Sargent Centre for Process Systems Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom
[b] Department of Chemical Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom
[c] GSK, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, United Kingdom

ARTICLE INFO

Keywords:
Design space identification
Protein A
Transfer learning
Machine learning
Quality by digital design

ABSTRACT

Quality by Digital Design (QbDD) employs in-silico experimentation to reduce wet-lab reliance and accelerate development. Design space identification is critical for QbDD to overcome bottlenecks and streamline process design. Traditional design space identification methods require costly, high-fidelity models. This work introduces a machine learning-enhanced design space identification approach that utilises wet-lab data. A transfer learning framework is developed under limited data availability by leveraging synthetic datasets created using off-the-shelf mechanistic models. An artificial neural network is constructed and used to classify feasible design regions under high, moderate, and low data availability (HDA, MDA, LDA). Results show strong performance in HDA for the data-driven method, with transfer learning improving accuracy in MDA and being essential in LDA. This approach demonstrates machine learning's potential to enable cost- and time-efficient process design in early-stage biopharmaceutical development.

1. Introduction

Downstream purification (DSP) of mammalian cell culture material is a highly resource-intensive phase of monoclonal antibody manufacture, with contributions up to 80% of the total production cost [1,2]. Protein A chromatography is a typical primary recovery step employed within these processes, the resin for which forms a substantial portion of these costs. Consequently, this unit step must be highly productive alongside being robust to external variations, for instance, feed stream composition variations that can originate upstream. The large number of interdependent unit operations involved during downstream processing and the need for robustness to upstream changes render DSP design a challenging task. In this setting, Quality by Design (QbD) is valuable by defining quality targets, identifying associated critical quality attributes (CQAs) and critical process parameter (CPP) controls, as well as their interactions [3–7]. QbD can enhance process understanding, reduce uncertainty, while ensuring manufacturing consistency and product quality [8–11]. Advancing this effort via computational approaches is known as Quality by Digital Design (QbDD) [12–16]. QbDD extends the rationale of QbD in the digital realm, integrating in-silico high-throughput experimentation, which can complement wet-lab datasets. This can accelerate process development by enhancing process understanding and can assist in the design of robust processes [17,18]. By computationally quantifying CPP-CQA interplay, one can identify the feasible operating region within which the process can yield products that meet the desired specifications [18,19], a process also known as design space identification. To achieve this digitally, mathematical models that describe the process dynamics are necessary. These can be high-fidelity, also known as mechanistic or white box; data-driven, also known as black-box; or a combination of the two, known as hybrid.

1.1. A portfolio of digital approaches towards bioprocess modelling

In Protein A chromatography, high-fidelity process models typically involve detailed mathematical representations of mass transfer and adsorption kinetics within the chromatography column [12,15,20–22]. Such models are often parameterised based on product- and process-specific datasets. Reparameterisation of such models to describe a different separation process can be resource- and time-intensive. It often requires domain expertise, additional experimental data, and significant computational power. Such constraints can render the transfer of high-fidelity models across processes impractical. This is highlighted during the development of the early stage process, when fast screening is necessary. At the same time, the lack of data availability challenges model reparameterisation even further, particularly in cases of material

* Corresponding author at: Sargent Centre for Process Systems Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom.
E-mail address: maria.papathanasiou11@imperial.ac.uk (M.M. Papathanasiou).

https://doi.org/10.1016/j.chroma.2025.466193
Received 31 January 2025; Received in revised form 3 July 2025; Accepted 4 July 2025
Available online 21 July 2025
0021-9673/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Nomenclature

ANN	Artificial neural network
CCCF	Clarified cell-culture fluid
CPP	Critical process parameter
CQA	Critical quality attribute
DBC	Dynamic binding capacity
DD	Data-driven
DSP	Downstream purification
F1	F1 Score
$F1_{II}$	Second order F1 Score
FN	False negative
FP	False positive
HDA	High data availability
HPLC	High-performance liquid chromatography
KPI	Key Performance Indicator
LDA	Low data availability
mAb	Monoclonal antibody
mAbs	Monoclonal antibodies
MDA	Moderate data availability
ML	Machine learning
QbD	Quality by design
QbDD	Quality by digital design
TL	Transfer learning
TN	True negative
TP	True positive

or product changes. Data-driven models, based on Machine Learning (ML), as opposed to high-fidelity process models, are developed solely on data. This can be beneficial in settings where in-depth system knowledge is not required. Nonetheless, developing and using ML models entails the inherent challenge of limited generalisation.

In this setting, transfer learning emerges as a promising solution, which can enable retention and transfer of the process knowledge, while minimising the wet-lab data required. Transfer Learning has proven successful in various fields, such as image recognition and biomedical applications, where data scarcity is common. Notably, Theodoris et al. [23] developed Geneformer, a deep learning model capable of analysing transcriptome networks in rare diseases and inaccessible tissues by leveraging extensive pretraining on broad transcriptomic datasets. Similarly, transfer learning in biopharma offers a pathway to mitigate data scarcity and accelerate the deployment of predictive models across varying systems.

Fig. 1 illustrates how a portfolio of the above-discussed digital tools can be tailored to different problem specifications. In this workflow, the first major assessment is understanding the data quality and quantity at disposal, whether it is wet-lab collected or synthetically generated from a high-fidelity model. In the case where the assessment results in poor data availability, the proposed strategy is to enhance data availability through further data collection. On the other hand, if the assessment results in good quality and quantity of data, then the availability of a digital analogue should be investigated to delineate the most effective route for achieving design space identification. When a process-specific digital analogue exists, such as a parametrised high-fidelity model, the preferred route for design space identification is through a model-based approach that harnesses the synthetic data generated from the digital analogue itself. Various methods addressing model-based design space identification challenges have been documented in the literature. For instance, Kotidis et al. [24] applied a constrained global sensitivity analysis for design space identification on an upstream bioreactor system. Kucherenko et al. [25] developed a computationally efficient framework for probabilistic design space identification using adaptive sampling to address uncertainties in process parameters and material properties. Vogg et al. [26] employed a variance-based approach to explore batch and counter-current frontal chromatography processes to evaluate process flexibility and robustness through design space exploration. Ding and Ierapetritou [27] built a novel surrogate-based design space identification for multi-column periodic counter-current chromatography, balancing computational complexity and predictive accuracy using adaptive sampling. Sachio et al. [19] implemented a machine learning-enhanced, model-based approach for design space identification in Protein A chromatography units, screening five industrial resins. Geremia et al. [28] demonstrated the use of surrogate-based feasibility analysis to expand design spaces through control strategies, mitigating uncertainty effectively.

Despite the innovation in this space, there remain several open research challenges. In this work, we tackle the following two: (1) what is the minimum amount of wet-lab data required for reliable design space identification, and (2) how can we leverage ML while transferring knowledge from white-box approaches for design space identification. To respond to these challenges, first, we assume that no high fidelity model is available, and therefore a fully data-driven design space identification route is followed (Fig. 1). Next, we assume that a high-fidelity model is available but not parametrised for the specific system at hand. For this, we employ transfer learning to retain process knowledge from a generic high-fidelity model, while using wet lab data to enhance predictive reliability of the design space. Both these approaches are applied and assessed on an industrially relevant Protein A chromatography column.

2. Experimental protocol & data

2.1. Experimental protocol

Protein A breakthrough experiments were carried out on an ÄKTA Avant 25 system from Cytiva™ (Amersham, UK) at residence times of 2, 3, and 5 min and mAbx Clarified Cell-Culture Fluid (CCCF) load concentrations of 0.74 g/L, 1.57 g/L and 3.44 g/L. The load materials at varying concentrations consisted of either untreated CCCF at 1.57 g/L, CCCF diluted to 0.74 g/L with cell culture media, or CCCF spiked with Protein A purified mAbx material to a concentration of 3.44 g/L. All materials were filtered using a 0.22 µm prior to loading onto a chromatography column.

For the Protein A breakthrough experiments a HiTrap™ MabSelect SuRe™ column from Cytiva™ (7 × 25 mm, 0.962 mL) was loaded to approximately 150 g of mAbx per L of column. The Protein A method consisted of an initial column equilibration phase, followed by loading of the CCCF load material at the specified residence time (2, 3 or 5 min). The column was then neutralised using an equilibration buffer, sanitised using a high-pH caustic solution, and again neutralised using an equilibration buffer. For all phases other than the load phase, a residence time of 5 min was used with Tris-Acetate buffers employed except during sanitisation.

Throughout the load phase of each run, flowthrough fractions of an appropriate volume were collected and the product concentration was measured by Protein A HPLC. The Protein A HPLC method used a POROS™ A 20 µm column (2.1×30 mm, 0.1 mL) from Thermo Scientific™ (Cheshire, UK), employing phosphate-based equilibration and elution buffers, a flow rate of 2 mL/min, and UV detection at 280 nm. This method consisted of an initial column equilibration phase, sample injection, a wash phase with equilibration buffer, a low-pH elution phase, and subsequent neutralisation with equilibration buffer. Flowthrough concentrations were determined by comparison of eluent peak areas to those from a calibration curve generated for Protein A purified material, of known concentration determined by UV280 absorbance, loaded at varying injection volumes.

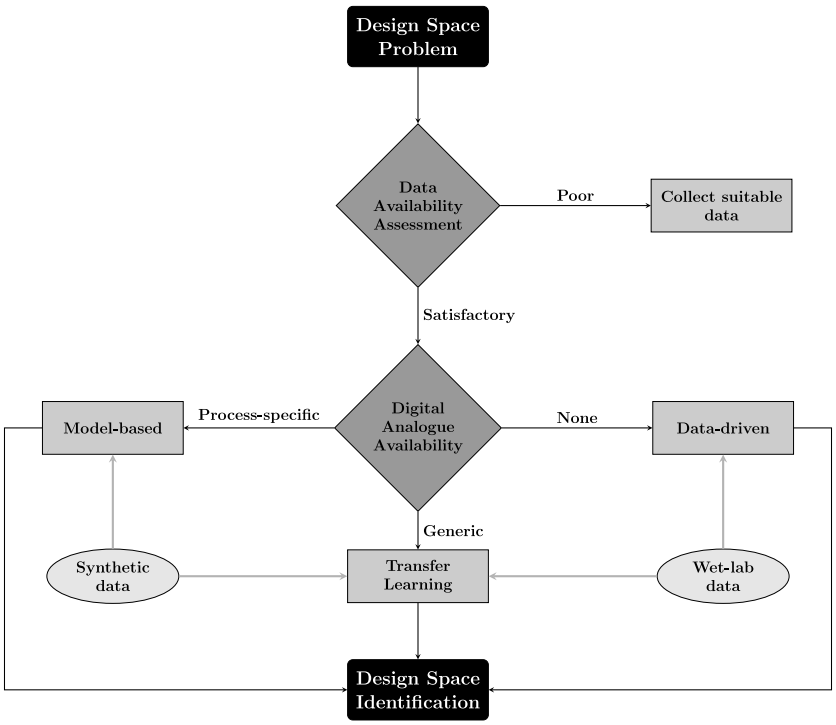


Fig. 1. Proposed workflow for tackling the design space problem, detailing decision pathways based on data availability and using model-based, data-driven, or transfer learning approaches to achieve design space identification.

Table 1
Full-factorial (3^2) experimental runs conducted with varying levels (low, medium, high) of residence time and feed mAb concentration.

Run	Residence time (min)	Feed mAb concentration (mg/mL)	Loading time (min)	Feed volumetric flow rate (mL/min)
1	2	1.53	200.00	0.481
2	3	1.54	299.69	0.321
3	5	1.54	501.04	0.192
4	2	3.36	85.72	0.481
5	3	3.36	128.44	0.321
6	5	3.41	214.74	0.192
7	2	0.69	374.22	0.481
8	3	0.73	642.18	0.321
9	5	0.72	1073.65	0.192

The fraction concentrations were normalised to the load concentration and plotted against the load ratio throughout the loading phase, having accounted for the system delay volume, to generate breakthrough curves. A full-factorial design (3^2) was employed with residence time and feed mAb concentration each set at three levels (low, medium, high) for a total of nine experimental runs, as shown in Table 1. The full-factorial design of experiments was chosen to ensure comprehensive coverage of the domain space and, just as importantly, to map its boundaries, thus enabling the machine learning methodologies proposed in this study.

The Dynamic Binding Capacity at 1% (DBC 1%) is the sole productivity metric employed in this study, defined as the load ratio when the ratio of cumulative unbound protein to total protein loaded equals 1%. This was calculated through numerical integration of the ratio of C/C_0, divided by the load ratio, where C and C_0 are the outlet and feed concentration, respectively. Moreover, DBC 1% serves as the feasibility constraint for identifying the design space. Data points meeting or exceeding this threshold (DBC $\leq$ 1%) are classified as feasible, whereas those falling below it are considered infeasible (DBC > 1%).

2.2. Experimental setup for breakthrough curve generation

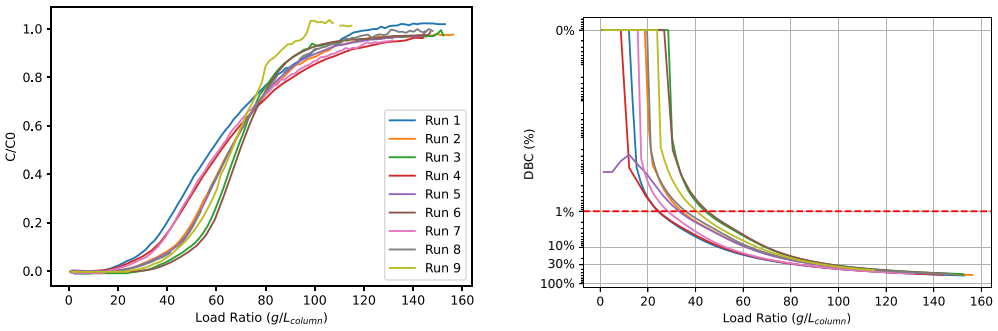
The breakthrough curves for the nine experimental runs defined in Section 2.1 of Protein A chromatography are shown in Fig. 2(a). These curves illustrate the relationship between the normalised outlet concentration (C/C_0) and the load ratio (g protein loaded per litre of column).

Breakthrough behaviour varied across the runs, with those employing progressively lower residence times exhibiting earlier breakthrough as expected. The impact of feed concentration upon breakthrough was less pronounced. All of the runs approached full saturation ($C/C_0 = 1$) at load ratios exceeding 120 g/L with DBC 1% values ranging from 40 to 60 g/L_{column}. These results exemplify the potential productivity trade-off between shorter operating times and decreased resin utilisation.

The DBC curves are shown in Fig. 2(b), reporting *DBC (%)* on a logarithmic scale versus the *Load Ratio* of the column. Notably, Run 5 deviates from the trends observed in other runs, exhibiting an initial upward shift. This can be attributed to negative values of C/C_0 recorded in the experimental data for Run 5 at low load ratios. Such values can arise due to factors like detector noise, transient system instabilities, or baseline shifts, which are common challenges in data collection processes.

To mitigate the above and ensure consistent interpretation across all runs, min–max scaling was applied to normalise the C/C_0 data between 0 and 1 for each run. Min–max scaling ensures that C/C_0 values are rescaled to a physically meaningful range, where a value of 0 corresponds to no breakthrough and a value of 1 represents full saturation. This normalisation eliminates the influence of experimental artefacts (e.g., negative values) and enhances the comparability of breakthrough profiles and DBC behaviour across all experimental conditions.

To better visualise the design space of the process, the experimental data is presented in the space defined by the selected critical process parameters at the predefined key process metric acting as a constraint. The relevant CPPs are Feed mAb Concentration (c_{in}, mg/mL), Feed Volumetric Flow Rate (Q_{in}, mL/min), and Loading Time (T_{load}, min),



(a) Breakthrough data collected from the nine experimental runs detailed in Table 1.

(b) Calculated Dynamic Binding Capacity as a function of load ratio based on the breakthrough curves.

Fig. 2. Experimental data collected from the nine experimental runs, showing breakthrough behaviour and Dynamic Binding Capacity trends. A MabSelect SuRe HiTrap was overloaded to achieve breakthrough using CCCF and fractions of the flowthrough were collected. These were analysed offline using Protein A HPLC to establish fraction concentration and therefore calculate C/C0 and construct the breakthrough curves shown.

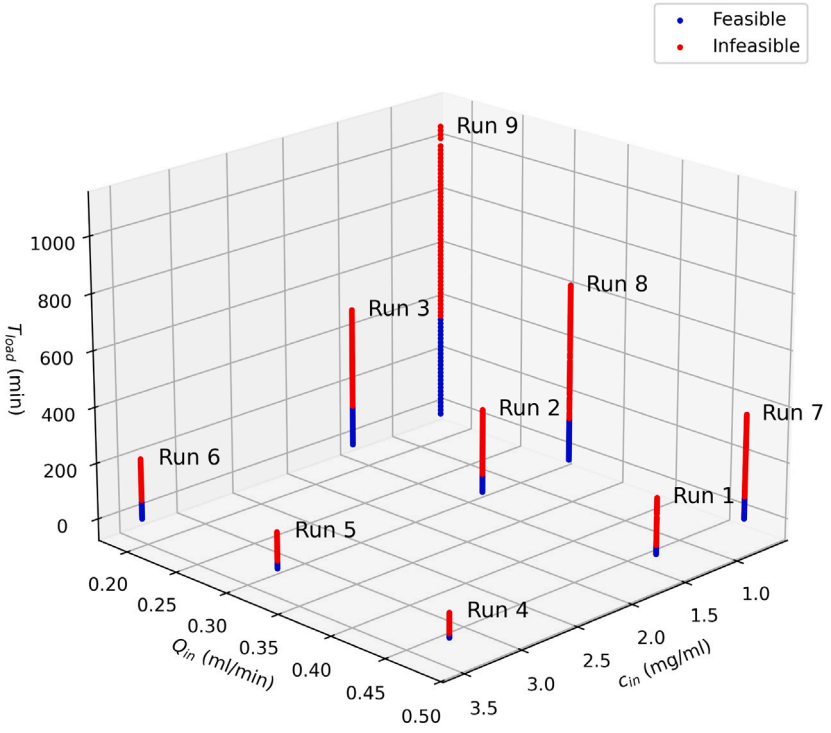


Fig. 3. Visualisation of the experimental data in the T_{load}, Q_{in}, and c_{in} space, with the DBC 1% constraint applied. Only the data points where DBC < 1% are considered feasible, in blue.

while the sole constraint is DBC 1%, evaluated using the min–max scaled C/C_0 dataset. This visualisation is presented in Fig. 3. From this figure, it is possible to assess the feasibility of each experimental run and analyse the influence of the CPPs on the key performance parameter, i.e. DBC 1%. Notably, lower feed flowrate and feed concentration lead to higher loading times, providing a greater margin to satisfy the imposed DBC constraint.

3. Proposed methodology

Fig. 4 illustrates the machine learning enhanced design space identification methodology proposed in this work. Based on the availability of wet-lab data and the existence of high-fidelity process models, the framework develops in two distinct pathways to achieve design space identification. The two pathways (Fig. 4) are the one developing from the existence of rich experimental datasets, and the one stemming from a lack of wet-lab data, but where a high-fidelity model is available. They are, namely, the *data-driven* (DD) and *transfer learning* (TL) routes, respectively, and already presented in Fig. 1.

These approaches are complementary, and they target different scenarios. The full data-driven route (Fig. 4, left branch) is designed to operate on the assumption of sufficient wet-lab data availability, harnessing the data directly to train the ANN in a single stage and predict the design space. Transfer learning, on the other hand, (Fig. 4, right branch) develops the ANN training in two stages. First, the network is pre-trained on synthetic data generated from a high-fidelity process model; then, wet-lab data is introduced to complete the training and fine-tune the design space prediction. A key difference of transfer learning is that the wet-lab data can be more limited in quantity and quality, compared to the data-driven approach, and the high-fidelity process model is not required to be parameterised to the specific system under study. Effectively, the transfer learning approach is designed to

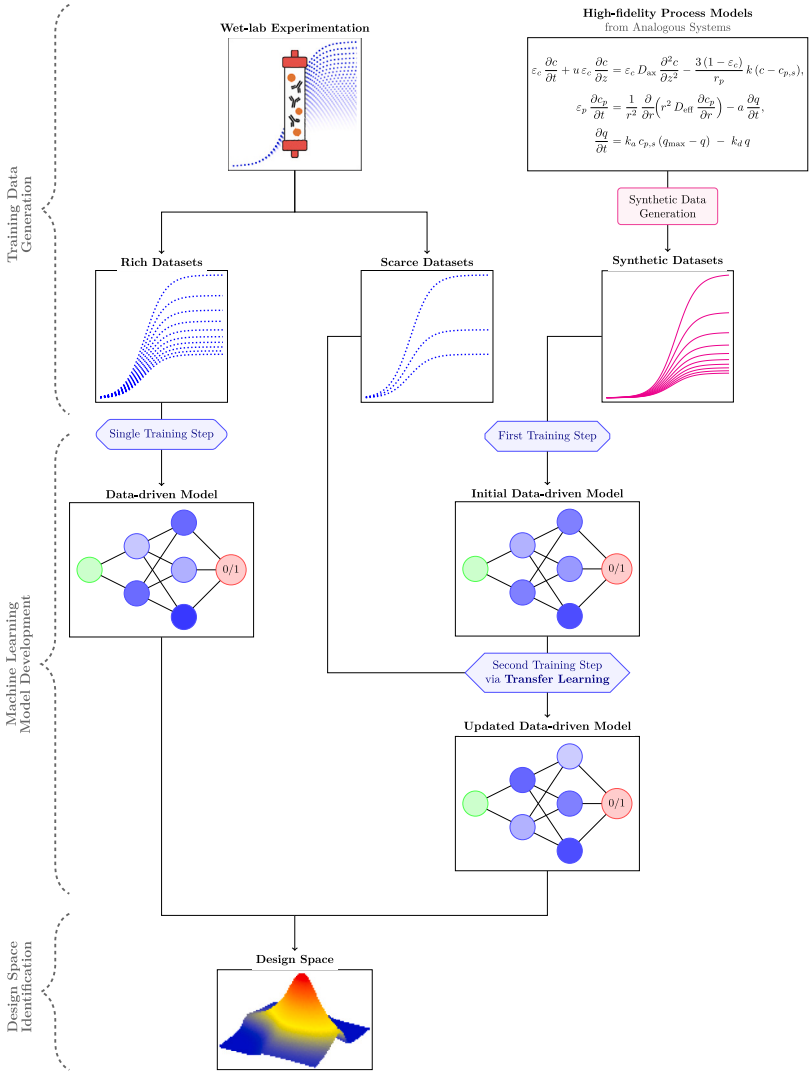


Fig. 4. Workflow breakdown of the proposed machine learning design space identification framework. The schematic highlights the fully experimental data-driven approach when rich datasets are available, and the transfer learning approach when scarce experimental datasets are coupled with high-fidelity process models of analogous systems.

leverage knowledge from different systems to make reliable predictions on the target system with limited wet-lab experimentation.

Both the data-driven and transfer learning routes presented in this work are designed for early-stage process development scenarios, where high-fidelity modelling is either unavailable or too costly to implement. The choice between the two approaches primarily depends on the availability of experimental data and/or the presence of analogous, unparameterised high-fidelity process models, as discussed in Section 3.2. Determining which approach is more suitable under specific conditions is the central aim of this work. To explore this, both methods are retrospectively applied and compared using the available wet-lab experimental data, simulating scenarios with intentionally limited data, and deliberately utilising an unparameterised high-fidelity process model that is not directly applicable to the system under investigation.

3.1. Binary classification for design space identification

In this work, the design space identification problem is formulated as a binary classification task, solved using artificial neural networks. With binary classification, the *feasibility* and *infeasibility* of the design space are represented by the *True* and *False* or 0 and 1 labels, respectively [29]. Feasibility is determined based on whether process outputs meet specific quality or performance criteria, DBC less than 1% in this case, as detailed in Section 2.2.

The input features to the ANN are the key CPPs identified from wet-lab experiments and mirrored in the synthetic data generation process: loading time (T_{load}), feed inlet flow rate (Q_{in}), and inlet product concentration (c_{in}). Loading time is selected as a CPP because it offers an intuitive and operationally relevant representation of the process conditions since it reflects the combined impact of both the residence time and load concentration variables. These selected CPPs collectively capture the factors influencing the design space.

ANNs were chosen for their ability to model complex, nonlinear relationships between input features and feasibility outcomes. However, to reflect the assumption of limited data availability and to mitigate the risk of overfitting, the network architecture is intentionally kept simple, with only a few hidden layers and a limited number of neurons. Further details on the binary classification problem formulation are included in the supplementary information.

3.2. Generic high-fidelity process model

This study employs the high-fidelity mechanistic process model proposed by Grom et al. [30] that is used as a basis to generate a synthetic dataset for pre-training the ANN as part of the TL approach (see Fig. 4).

The synthetic dataset was generated using the parameter set reported by Grom et al. [30] for the MabSelect SuRe™ resin in their original work. The differences in feedstock, column dimensions, and operational specifics make it inaccurate for the system investigated in this study. Transfer learning is applied to adapt the knowledge encoded in the intentionally inaccurate model of the analogous system to the current specific system with limited experimental data.

The Grom et al. [30] model was originally developed for Protein A chromatography of monoclonal antibodies from Chinese hamster ovary cell cultures, which describes axial and radial mass transport using a non-linear partial differential–algebraic system of equations. Key assumptions include negligible film transfer resistance, rate-limiting internal pore diffusion, Langmuir adsorption kinetics, and Danckwerts' boundary conditions [30]. The model was validated on five different resins in the original study, demonstrating robustness across a range of applications. A more detailed description of the model is provided in the supplementary information, and readers are encouraged to refer directly to the work of Grom et al. [30] for a more comprehensive discussion.

For this work, the original model [30] adopted does not entirely represent the system under investigation and would rather require extensive reparameterisation and potentially structural modifications to accomplish that. This choice is intentional, and it is necessary in order to demonstrate the methodology's potential to apply transfer learning on a relevant industrial application without prior model reparameterisation.

3.3. Design space identification accuracy quantification

The definition of suitable accuracy metrics is key for assessing the quality of a design space, especially in a machine learning context. In this study, the design space identification problem is framed as a classification task, where a model predicts the feasibility of a point within the space domain. The confusion matrix for this task comprises the following outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Here, "True" and "False" indicate whether the prediction matches the actual result, while "Positive" and "Negative" correspond to Feasible and Infeasible cases, respectively. For instance, FP indicates a point incorrectly classified as Feasible. By contrast, an experimental data point can be either classified as a pure "Positive"/"Feasible" or "Negative"/"Infeasible".

The classification performance is evaluated using standard metrics, including Accuracy, Precision, Recall, and F1 Score, a detailed description of which is provided in the supplementary information, all calculated from the elements of the confusion matrix. In addition, this study introduces a novel metric, termed the *second-order F1 Score* and denoted as $F1_{II}$. Unlike the traditional F1 Score, which relies on ground truth data, $F1_{II}$ assesses the alignment between a target model's predictions and those of a reference model, which is assumed to be a reliable approximation of the true system. This metric quantifies the discrepancy between the target model and the reference model, offering a complementary perspective on predictive performance in scenarios where experimental ground truth may be limited.

3.4. Transfer learning, training, & validation

Monitoring and preventing overfitting in machine learning training algorithms is essential, particularly in the case of ANNs. Early stopping is introduced during training to mitigate overfitting by limiting the number of epochs based on a predefined criterion. In this study, training is stopped when the F1 Score reaches or exceeds 98%, balancing accuracy and generalisability. This threshold reflects the trade-off between accuracy, which measures the model's precision on the available data, and generalisability, which assesses the model's ability to predict design space behaviour away from training points, for example, in unexplored regions. While interpolation ensures predictions remain within the bounds of observed data, extrapolation involves making predictions beyond these bounds, introducing higher uncertainty and thus should be avoided. Thus, the 98% F1 Score has been selected as it represents the highest achievable accuracy in this specific case study that does not rely excessively on interpolation. This threshold ensures that the model captures the design space effectively without compromising its ability to predict feasibility in unexplored regions. In general, each process or case study can require a different threshold to match the trade-off rationale presented here, but generalising this value is extremely challenging. This complexity is further compounded by the potential presence of random noise in the data.

4. Results & discussion

4.1. Design space identification based on synthetic data

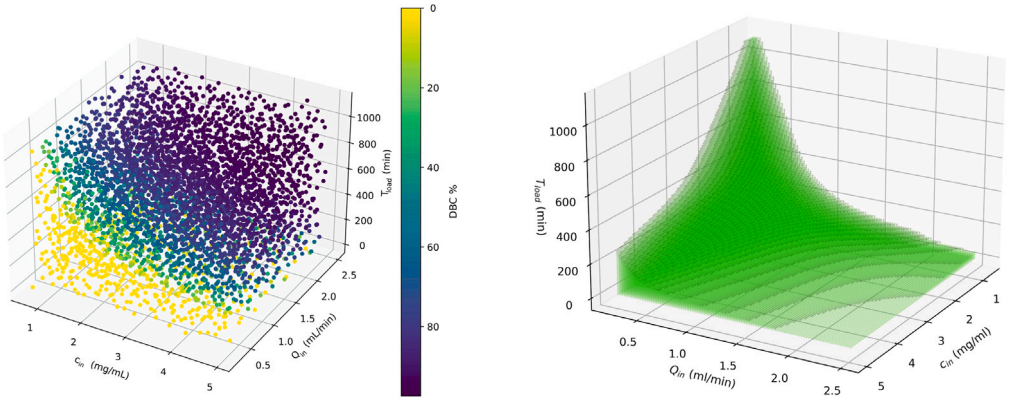
The application of the machine learning framework and the training of the artificial neural network directly on the synthetic dataset serve two key purposes. First, this approach demonstrates the ANN's ability to accurately identify the design space under ideal conditions, where data availability is virtually unlimited. Second, it develops a pre-trained ANN that serves as the foundation for the transfer learning methodology, upon which further training is performed using the specific wet-lab data.

The chromatography process model [30] for the MabSelect SuRe™ was sampled using a Sobol' sequence [31] with 2^{12} (4096) points across three input variables: Feed mAb Concentration (c_{in}, mg/mL), Feed Volumetric Flow Rate (Q_{in}, mL/min), and Loading Time (T_{load}, min). Details on the implementation of the Sobol method are provided in the Supplementary Information. The sampling covered a broad domain, ranging from $[0.2, 0.7, 10]$ to $[2.5, 5, 1100]$ (mg/mL, mL/min, min), respectively.

As previously defined (Section 2.1), the constraint used to distinguish between feasible and infeasible regions throughout the study is the DBC 1% process metric, applied consistently to both the synthetic and experimental data. While DBC is often evaluated at 10% breakthrough ($DBC_{10\%}$) in practice, the stricter $DBC_{1\%}$ level was selected here to rigorously evaluate the ability to identify feasible operating regions under more demanding performance requirements. Consequently, operating points are considered "feasible" throughout this work if they meet or exceed the $DBC_{1\%}$ target, as determined from the experimental breakthrough data.

The resulting dataset is depicted in Fig. 5(a). Given the high availability of data points, 10% of the dataset was reserved for testing without significantly affecting accuracy. The early stopping criterion of $F1 \geq 98\%$ was evaluated on this test set, as described in Section 3.4.

Table 2 summarises the accuracy metrics introduced in Section 3.3 for the identified design space, while Fig. 5(b) illustrates the predicted feasible region. Notably, the training F1 Score is slightly lower than the testing F1 Score, potentially due to the early stopping criterion of F1 Score >98%. However, this difference is not substantial enough to undermine the validity of the early-stopping strategy. Moreover, sacrificing a small degree of accuracy for improved generalisability is often advisable, as it enhances prediction reliability across the entire design domain, as discussed in Section 3.4. These results confirm that the adopted machine learning approach is effective at design space identification under conditions of sufficient data availability. This performance serves as a critical benchmark for extending the methodology when applied to experimental wet-lab data in the subsequent stages of this study.



(a) Synthetic data generated through Sobol' sampling from the generic process model of Grom et al. [30].

(b) Resulting design space identified at $DBC1\%$. The green area represents the predicted feasible region.

Fig. 5. Synthetic data generated (a) and predicted design space (b) using the machine learning methodology applied to the synthetic dataset.

Table 2
Design space identification accuracy metrics on synthetic data.

	Whole set	Train set	Test set
Accuracy	99.51%	99.48%	99.76%
Precision	95.24%	94.95%	97.73%
Recall	100.00%	100.00%	100.00%
F1 Score	97.56%	97.41%	98.85%

4.2. Design space identification based on wet-lab data

To compare the performance of the data-driven and transfer learning approaches illustrated in Fig. 4, a series of scenarios with differing levels of wet-lab data availability, exhibiting varying degrees of data quality and quantity, were simulated. In these scenarios, variations in data quality were represented by selectively removing specific experimental runs from the training dataset, whereas variations in data quantity corresponded to changes in the total number of runs included. Thus, both machine learning approaches were applied to each simulated scenario. The resulting performance was quantitatively evaluated using the F1 Score and $F1_{II}$, as defined in Section 3.3. The $F1_{II}$ metric required a reference model, which was generated using the data-driven approach trained on the full set of experimental runs.

Training of the neural networks followed the procedure outlined in Section 3.4, including early stopping once the training F1 Score reached or exceeded 98%. The reference model thus constructed using the full experimental dataset achieved an F1 Score of 98.14%, and the resulting shape is shown in Fig. 6. The predicted design space indicates good agreement with the experimental data and robustness, provided by the smooth contours of the feasible design space. This can underscore the reliability of the data-driven approach to produce reliable design spaces directly on experimental data when the data availability is satisfactory.

For the transfer learning approach, the base model is the neural network trained on the synthetic data generated from the original Grom et al. [30] model, and whose results are discussed in Section 4.1. This model, pre-trained on a different system, was subsequently fine-tuned using only the available wet-lab data, enabling the adaptation to the specific characteristics of the target system. Fig. 7 shows a comparison between the design spaces created with the data-driven design space identification approach: the original synthetic design space from Grom et al. [30] (Fig. 7(a)), used as the base model for transfer learning, and the experimental design space derived from the current study (Fig. 7(b)). The synthetic design space (Fig. 7(a)) fills a larger portion of the domain compared to the experimental one (Fig. 7(b)). A similar distinction can also be seen by comparing Figs. 5(b) and 6.

From a morphological point of view, the two design spaces are similar in shape. Nonetheless, the design space identified prior to the application of transfer learning (Fig. 7(a)) does not accurately describe the updated wet-lab dataset (Fig. 7(b)). For this, transfer learning is used as a means to reparameterise the "initial data-driven model" (Fig. 4, right-hand side) using experimental data of the system at hand. The model aims to iteratively converge on a good fit by finding a set of parameters that can accurately describe the relationship between the new experimental data and the existing model. An alternative to the presented approach would be to reparameterise the initial mechanistic model. In this case, and due to the behaviour of the outlier runs 8 and 9 in particular, this would require extensive model restructuring, including reformulation of the equation set. In an environment of accelerated process development, such an exercise would hinder the overall process lifecycle and would potentially increase the amount of experimental data required.

4.2.1. Data availability influence on machine learning design space identification

To assess the applicability and performance of both the DD and TL approaches, scenarios were simulated by progressively reducing the size of the training dataset through the iterative removal of experimental runs. This procedure generated multiple data splits by deliberately leaving out one or more available data points. Each data split defines which experimental runs are used for training the machine learning models and which are reserved for testing. For example, a valid split might exclude a single experimental run from training and assign it to testing, or vice versa. Each split is assigned a unique identifier (Split ID) to enable consistent cross-referencing across tables and figures, as shown in Table 3. The nine available experimental runs result in $\sum_{k=1}^{9}\binom{9}{k}$ possible splits, where $\binom{9}{k}=\frac{9!}{k!(9-k)!}$, equating to 2^9-1, or 511 unique train/test combinations. Due to the impracticality of analysing all combinations, a subset of 14 representative splits was selected for this study. These splits are listed in Table 3, their corresponding accuracy metrics are reported in Table 4, and the resulting second-order F1 Scores are visualised in Fig. 8.

Three distinct regions of data availability are defined to contextualise the splits: "High Data Availability" (HDA), "Moderate Data Availability" (MDA), and "Low Data Availability" (LDA). High data availability corresponds to more than six runs used for training, moderate data availability includes between six and three runs, and low data availability consists of three or fewer runs, provided at least one of them is Run 9. These regions are discussed in greater detail in Section 4.2.3.

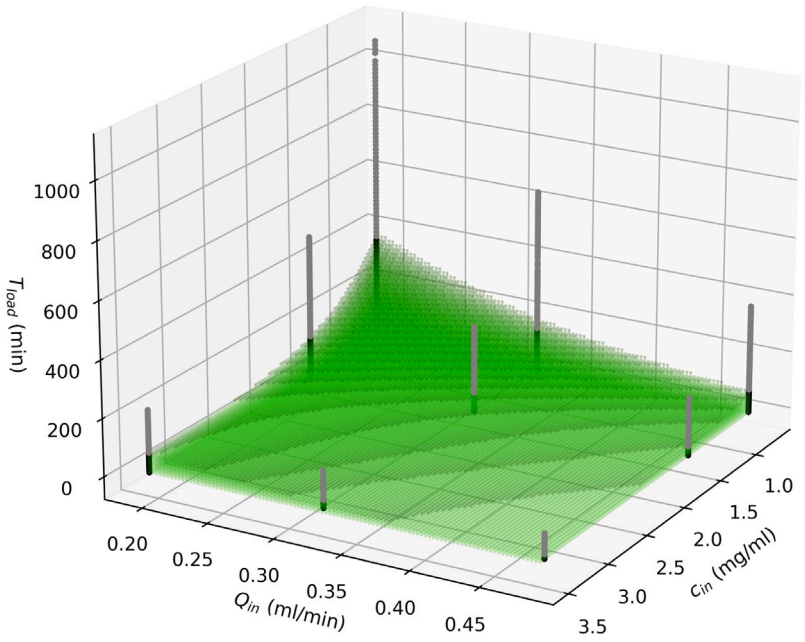
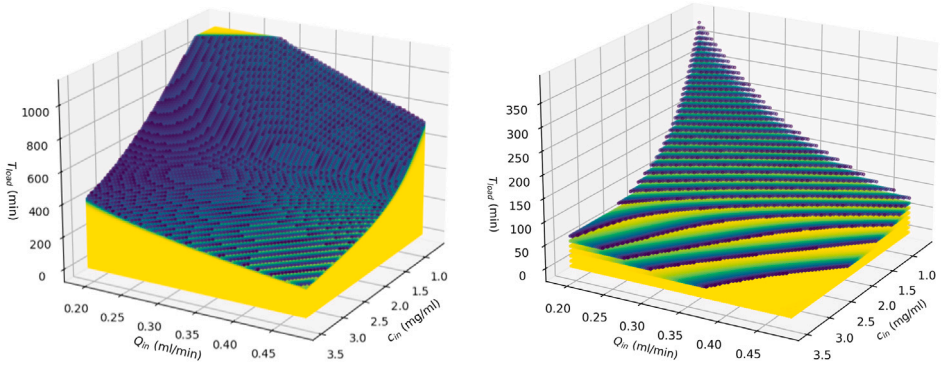


Fig. 6. Design space identified with the purely data-driven method trained on the full experimental dataset, used as the reference for calculating the $F1_{II}$ Score. The green area represents the predicted feasible region, while the grey bars depict the experimental data. Light grey indicates infeasible data, and dark grey represents feasible data. For detailed experimental data visualisation, see Fig. 3.



(a) Synthetic data design space. (b) Wet-lab data design space.

Fig. 7. Comparison of design spaces identified through the data-driven approach: synthetic data generated from the Grom et al. process model, used as the initial model for transfer learning, (a) and wet-lab experimental data from this study (b). The synthetic design space has been cropped to match the axes of the wet-lab data for direct comparison.

The results of this comparative analysis reveal several noteworthy findings. As shown in Table 4, the data-driven method alone achieves high accuracy ($\gtrsim 95\%$) in the HDA region, effectively addressing the design space identification problem. In this region, the addition of transfer learning does not significantly enhance accuracy, as the second-order F1 Scores for Split IDs 1 through 5 exhibit minimal differences between the two methods. However, transfer learning improves predictive accuracy in the MDA region, where 4 to 7 out of the 9 experimental runs are removed for testing, leaving just two runs for training in the worst-case scenario. This becomes particularly pronounced in the LDA region, as seen in Splits 13 and 14, where the F1 Scores for transfer learning are substantially higher compared to the data-driven method.

Split 9 represents a notable outlier where the application of transfer learning results in reduced predictive accuracy. As shown in Fig. 13, the transfer learning method over-predicts the boundaries of the feasible region at high concentrations and flow inlet. The reason for this outlier behaviour, where transfer learning is diminishing accuracy rather than enhancing it, as also noted in Fig. 8, finds an explanation in the core of the TL methodology and the training applied. In this particular data split of Split 9, the corner corresponding to high concentrations and high flow inlet is not supported by experimental data. Consequently, neither the data-driven nor the transfer learning approaches can harness information from that region during training. However, transfer learning carries prior knowledge from pre-training on the synthetic data generated using the purposely unparameterised model of Grom et al. [30], as explained in Section 3.2. This transferred knowledge suggests the existence of a feasible region in that corner. The difference between the synthetic and experimental data can be seen by comparing the respective design spaces shown in Fig. 7, where the region of feasibility in that area is larger in the synthetic data (Fig. 7(a)) compared to the wet-lab data (Fig. 7(b)). As a result, the transfer learning method (Fig. 13) over-predicts the existence of a feasible region, while the data-driven method under-predicts it, having no prior knowledge to influence its interpolation. Although both methods naturally show lower accuracy in this no-data region, the extent of over-prediction by the transfer learning model exceeds the under-prediction of the data-driven model,

Table 3
Comparative training/test splits definition. *Region* indicates the amount of data available per split: "High Data Availability" (HDA), "Moderate Data Availability" (MDA), and "Low Data Availability" (LDA).

Region	Split ID	Training run									N Test runs
		1	2	3	4	5	6	7	8	9	
HDA	1	✓	✗	✓	✓	✓	✓	✓	✓	✓	1
	2	✓	✓	✗	✓	✓	✓	✓	✓	✓	1
	3	✓	✓	✓	✓	✓	✓	✓	✓	✗	1
	4	✓	✗	✓	✗	✓	✓	✓	✓	✗	3
	5	✓	✗	✓	✓	✗	✓	✓	✗	✓	3
MDA	6	✗	✓	✗	✓	✗	✓	✓	✗	✓	4
	7	✗	✗	✗	✓	✗	✓	✓	✗	✓	5
	8	✗	✓	✓	✗	✗	✗	✗	✓	✓	5
	9	✗	✗	✓	✗	✗	✗	✗	✓	✓	6
LDA	10	✓	✗	✗	✗	✗	✗	✗	✗	✓	7
	11	✗	✓	✗	✗	✗	✗	✗	✗	✓	7
	12	✗	✗	✗	✓	✗	✗	✗	✗	✓	7
	13	✗	✗	✓	✗	✗	✗	✗	✗	✓	7
	14	✗	✗	✗	✗	✗	✗	✗	✓	✓	7

thereby explaining the greater drop in F1 Score accuracy for transfer learning compared to data-driven modelling (Fig. 8 and Table 4). In contrast, Split 8 (Fig. 12) demonstrates a case where the same prior information transferred from the base model to the experimental data enhances accuracy instead of reducing it.

In the HDA region, Split 3 (Fig. 9) offers an insightful example. Both methods achieve an F1 Score of approximately 96% (see Table 4), which can be deemed satisfactory. A closer inspection of Fig. 9 reveals that both predicted design spaces miss a significant portion of the feasible region. This issue arises from the exclusion of Run 9 during training. This could indicate two key points: (i) F1 Score alone is not a reliable indicator of true accuracy, emphasising the need for visual inspection (as further discussed in Section 4.2.2); and (ii) not all experimental runs contribute equally to the model's performance. In this case, Run 9 is particularly critical due to its larger feasible region compared to other runs.

In the MDA region, Splits 5 and 6 reveal interesting outcomes. Split 5 (Fig. 10) achieves an $F1_{II}$ Score of approximately 95% for both methods, demonstrating strong predictive accuracy. Split 6 (Fig. 11), however, highlights the advantage of transfer learning, as it significantly enhances the predictive accuracy of the design space compared to the data-driven method alone under identical data availability conditions. This result underscores the efficiency of transfer learning in optimising data usage.

In the LDA region, Split 14 (Fig. 15) shows a noteworthy behaviour in the prediction using the data-driven approach. Given the scarcity of data points, the design space shape is visibly inverted and greatly overpredicted in regions distant from the training runs (Runs 8 and 9). In contrast, the transfer learning approach still provides a meaningful and satisfactory prediction of the design space (Fig. 15b) under the same training regime. This trend is also evident in Split 13 (Fig. 14), where the DD prediction is highly unreliable, while the TL prediction is able to capture the underlying trend without the support of additional data.

These results demonstrate the increased reliability of the transfer learning methodology, which relies on prior knowledge to build design space predictions. Although the numerical values of the F1 Scores for the LDA case do not exceed 90%, the predictions generated with TL offer a representation that may still be sufficient for decision-making in the early stages of process development. This is evident in the TL predictions for Splits 13 and 14 (Figs. 14 and 15), where the design space successfully captures the expected shape even when extrapolating far beyond the wet-lab data used for training. This reliability can help guide experimentation and process design in the early stages and serves as an effective tool for enhancing the screening of materials (e.g. resins) or process conditions with limited data.

Table 4
Accuracy results per each Split selected. *Region* indicates the amount of data available per split: "High Data Availability" (HDA), "Moderate Data Availability" (MDA), and "Low Data Availability" (LDA).

Region	Split ID	N Test runs	F1 Score			$F1_{II}$ Score		
			DD	TL	Diff	DD	TL	Diff
HDA	1	1	89.4%	91.3%	+1.9%	96.9%	96.6%	−0.3%
	2	1	89.8%	98.1%	+8.3%	97.3%	97.8%	+0.5%
	3	1	76.2%	73.2%	−3.0%	96.6%	95.7%	−0.8%
	4	3	77.8%	77.8%	0.0%	92.0%	91.5%	−0.5%
	5	3	91.4%	92.9%	+1.5%	95.7%	95.8%	+0.2%
MDA	6	4	84.2%	91.3%	+7.0%	88.7%	94.5%	+5.7%
	7	5	86.8%	86.4%	−0.4%	87.1%	90.2%	+3.0%
	8	5	68.8%	84.3%	+15.5%	89.4%	94.9%	+5.5%
	9	6	68.3%	55.5%	−12.8%	87.9%	78.6%	−9.3%
LDA	10	7	62.7%	85.2%	+22.5%	78.3%	85.1%	+6.7%
	11	7	65.4%	86.8%	+21.4%	82.3%	90.4%	+8.1%
	12	7	56.8%	63.4%	+6.6%	64.5%	71.7%	+7.1%
	13	7	44.6%	71.9%	+27.3%	57.3%	83.7%	+26.4%
	14	7	41.6%	55.1%	+13.5%	21.6%	63.2%	+41.6%

4.2.2. *Selecting effective accuracy metrics*

In the context of this comparative analysis, where different data splits are evaluated to understand the methodology's effectiveness, it is important to establish reliable metrics that can be used to clearly gauge the performance of the system. While the F1 Score can serve as a useful guide for assessing prediction accuracy, it might fail to provide a comprehensive assessment of the design space accuracy, as demonstrated earlier. The second-order F1 Score (derived in Section 3.3) offers valuable insights into how predictions align with the true design space and, in most cases, provides a more accurate representation compared to the first-order F1 Score. However, its primary limitation lies in its retrospective nature, as it can only be applied once sufficient data has been collected to evaluate scenarios with fewer data points, as shown in the comparative analysis in Section 4.2. This constraint makes the second-order metric impractical for online applications, where no reference data-driven model is available. Consequently, this highlights a broader challenge in the application of data-intensive machine learning methodologies, which are inherently interpolative and lack a clear distinction between training and testing data, complicating the detection of overfitting. Addressing this issue remains an open research challenge.

When calculated on small test sets, the F1 Score may not accurately reflect the true quality of the predicted design space. For instance, Splits 3 and 4 exhibit relatively low F1 Scores despite achieving high second-order F1 Scores (see Table 4), indicating that the predicted design space closely resembles the ideal reference. This underscores the importance of evaluating both metrics in tandem to gain a comprehensive understanding of predictive accuracy.

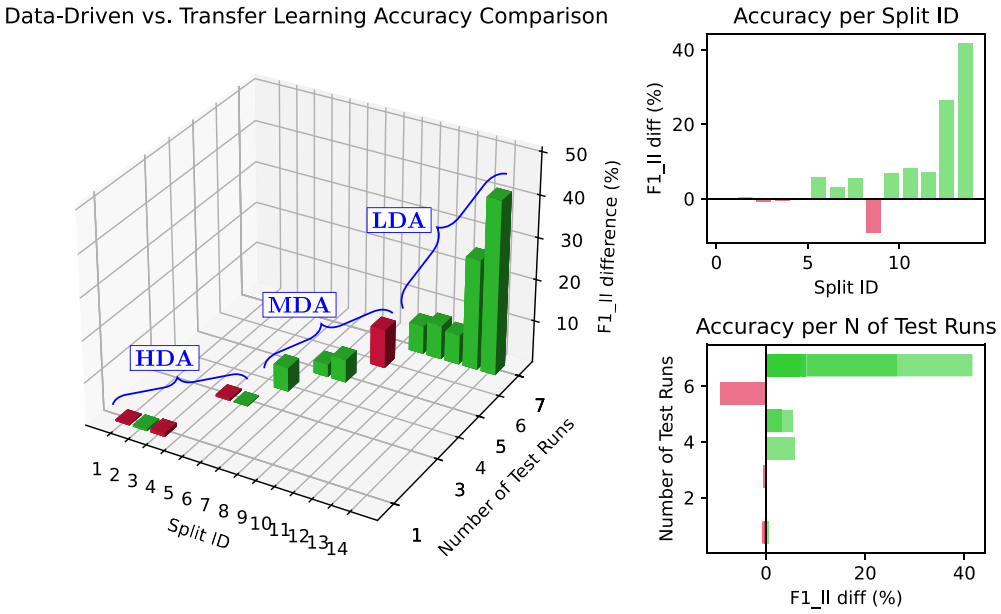
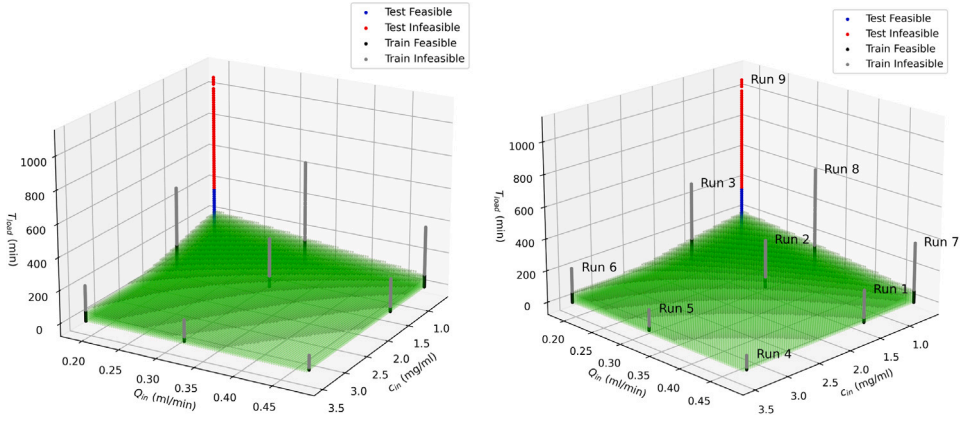
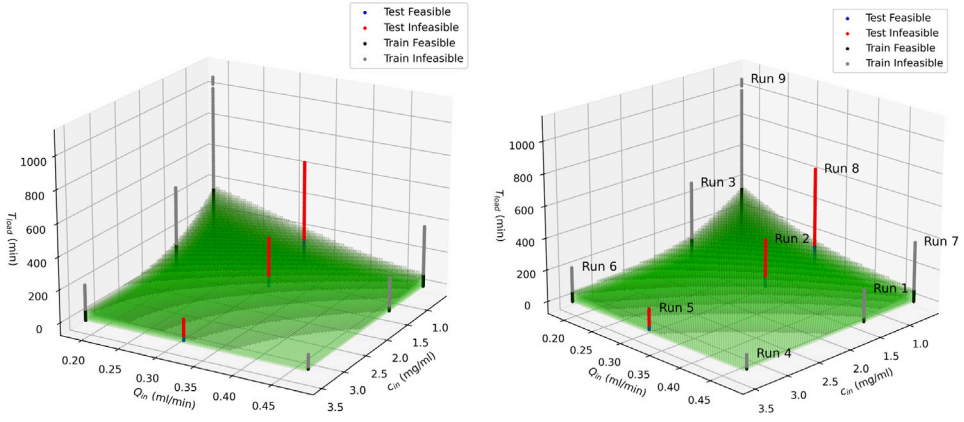


Fig. 8. Graphical analysis of the accuracy differences reported in Table 4 per each split between data-driven and transfer-learning. Boxes in green indicate a more accuracy % for the transfer learning method while in red the opposite. *Region* indicates the amount of data available per split: "High Data Availability" (HDA), "Moderate Data Availability" (MDA), and "Low Data Availability" (LDA).



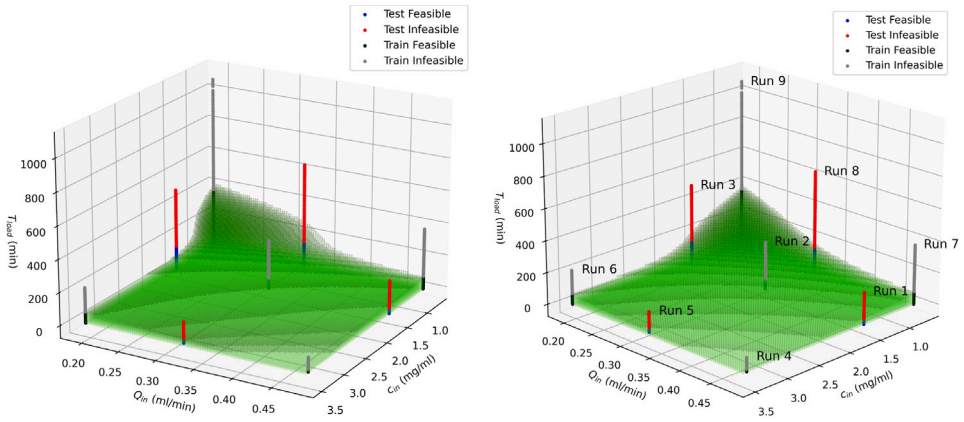
(a) Data-driven (b) Transfer learning

Fig. 9. Split 3 (high data availability).



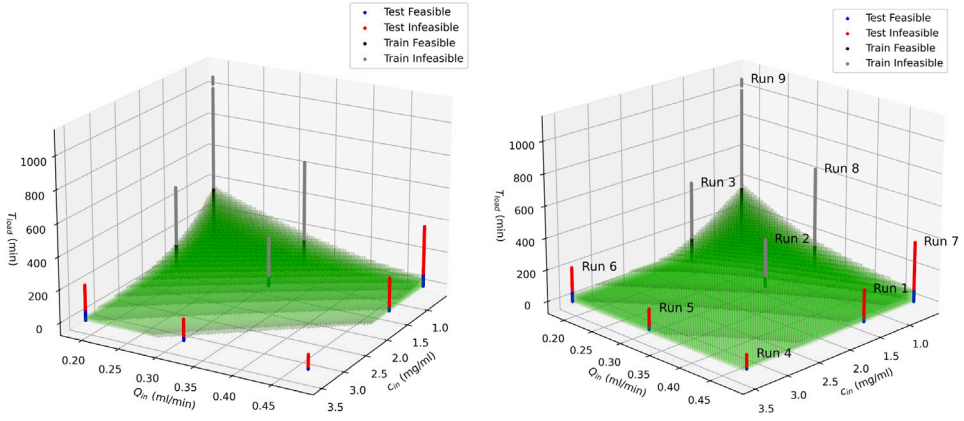
(a) Data-driven (b) Transfer learning

Fig. 10. Split 5 (high data availability).



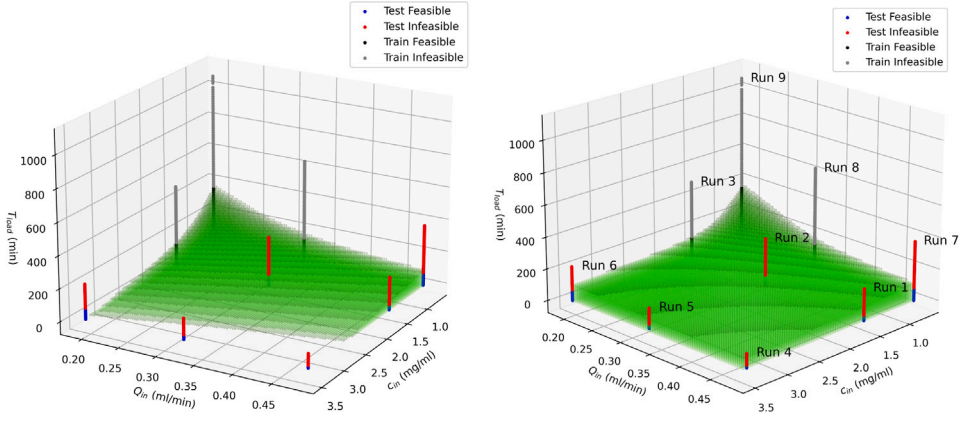
(a) Data-driven (b) Transfer learning

Fig. 11. Split 6 (moderate data availability).



(a) Data-driven (b) Transfer learning

Fig. 12. Split 8 (moderate data availability).



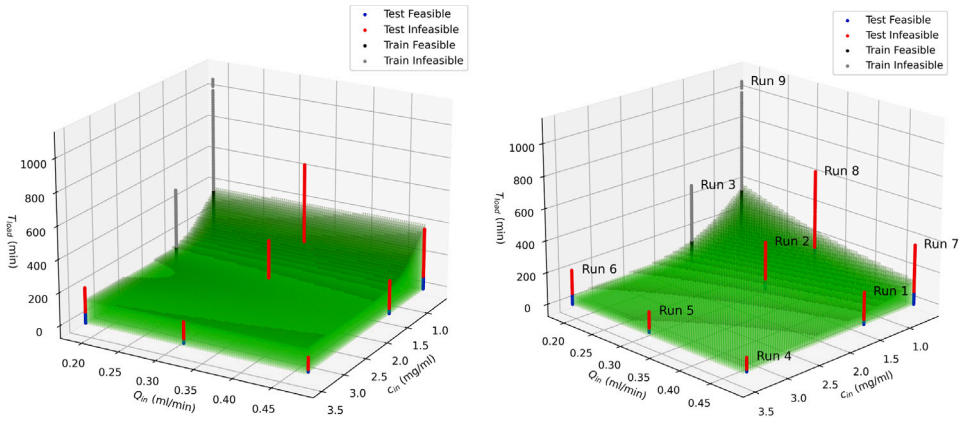
(a) Data-driven (b) Transfer learning

Fig. 13. Split 9 (moderate data availability).

4.2.3. Data quality and quantity assessment

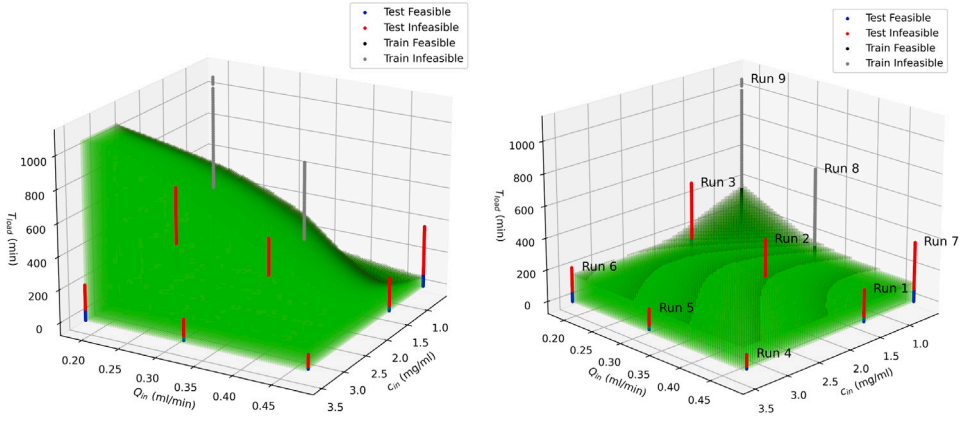
The analysis of the results on wet-lab data provides insights into the interplay between data quantity and quality in achieving accurate design space identification. The insights discovered for each of the three regions delineated above (HDA, MDA, and LDA) can give guidance on how to assess data availability and modelling strategies, answering the questions posed in the framework proposed above in Fig. 1.

Data quality can be considered even more critical than data quantity in certain scenarios. For example, excluding Run 9 from the splits in the comparative analysis (Section 4.2) has a significant impact, even if the remaining eight runs are included in the training. This is because Run 9 intrinsically carries much more information than most other runs. Specifically, its feasibility range, under the applied constraints, is considerably larger. Consequently, most splits in the comparative



(a) Data-driven (b) Transfer learning

Fig. 14. Split 13 (low data availability).



(a) Data-driven (b) Transfer learning

Fig. 15. Split 14 (low data availability).

analysis include Run 9, as its absence typically renders the resulting design space predictions meaningless.

In the HDA region, transfer learning is generally not recommended, as the trade-off between modelling investment and accuracy improvement is unfavourable. In the MDA region, transfer learning is beneficial for improving prediction accuracy and enhancing the design space, although the data-driven method alone may suffice depending on the use case and accuracy requirements. In the LDA region, the data-driven method alone is unreliable due to its interpolative nature, making transfer learning essential for creating meaningful and accurate design spaces.

5. Conclusions

Design space identification is a key tool within the framework of quality by design, and particularly, quality by digital design. Often, the reliance on high-fidelity process models to define these spaces can present significant challenges. In such scenarios, data-driven strategies based solely on experimental data may be favoured. However, due to their dependence on the quality and quantity of available data, the prediction reliability of such methods must be enhanced. Transfer learning achieves this by harnessing pre-existing knowledge derived from process modelling of analogous yet significantly different systems to improve data efficiency, prediction reliability, and the accuracy of design spaces.

In this work, we develop and compare a design space identification methodology based on machine learning and transfer learning. To build the design space, the method can either be applied directly to experimental data or integrated with transfer learning by leveraging knowledge coming from analogous systems. The case study investigated is that of an antibody Protein A affinity chromatography using a single column packed with the MabSelect SuRe™ resin. The machine learning problem is formulated as a binary classification task, where an artificial neural network is trained on wet-lab data collected from the experimental apparatus. Transfer learning is introduced as a tool to enable the transfer of knowledge from synthetic data generated by a high-fidelity process model parametrised for a significantly different but analogous system. The results demonstrated that the artificial neural network classifier is reliable and accurate in predicting design spaces in data-dense scenarios. This was confirmed through testing on a synthetic dataset generated from a process model reported in the literature for a similar system. Building on this cornerstone finding, the methodology was extended in a purely data-driven fashion to wet-lab data collected in this study, where accurate results were achieved when a large amount of data was used for training. To benchmark the methodology's capabilities, a comparative analysis was performed to evaluate its performance across different levels of data availability, mimicking scenarios where data is limited.

We demonstrate that machine learning, and by extension, transfer learning, can offer an efficient way of performing model adaptation and recalibration. Compared to conventional mechanistic approaches,

hybrid methodologies, as the one proposed here, can help tackle challenges related to mathematical restructuring of the model formulation. This becomes highly relevant in cases of outlier behaviours, as the example runs 8 and 9 in this work. The application of transfer learning significantly enhances the data efficiency of data-driven methods, reducing the need for a priori process knowledge and making these methods particularly valuable in early-stage process development.

CRediT authorship contribution statement

Andrea Galeazzi: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Steven Sachio:** Formal analysis, Data curation. **Elizabeth Edwards:** Writing – original draft, Validation, Investigation, Data curation. **David Hilton:** Writing – review & editing, Validation, Investigation, Data curation. **Maria M. Papathanasiou:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

Funding from the UK EPSRC for the i-PREDICT: Integrated adaPtive pRocEss DesIgn and ConTrol (Grant EP/W035006/1) is gratefully acknowledged. The authors greatly acknowledge Andrej Pohar and Matic Grom for sharing the process model utilised in this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.chroma.2025.466193.

Data availability

The data that has been used is confidential.

References

[1] Daniel G. Bracewell, Mili Pathak, Guijun Ma, Anurag Rathore, Re-use of protein a resin: Fouling and economics, BioPharm Int. 28 (3) (2015).
[2] Kristi Budzinski, Megan Blewis, Philip Dahlin, Daniel D'Aquila, Julia Esparza, Jack Gavin, Sa V. Ho, Clarice Hutchens, David Kahn, Stefan G. Koenig, Robert Kottmeier, Jeff Millard, Matt Snyder, Brad Stanard, Lixin Sun, Introduction of a process mass intensity metric for biologics, New Biotechnol. (ISSN: 1871-6784) 49 (2019) 37–42, http://dx.doi.org/10.1016/j.nbt.2018.07.005.
[3] Anurag S. Rathore, Helen Winkle, Quality by design for biopharmaceuticals, Nature Biotechnol. (ISSN: 1087-0156, 1546-1696) 27 (1) (2009) 26–34, http://dx.doi.org/10.1038/nbt0109-26.
[4] Anurag S. Rathore, Roadmap for implementation of quality by design (QbD) for biotechnology products, Trends Biotechnol. (ISSN: 0167-7799) 27 (9) (2009) 546–553, http://dx.doi.org/10.1016/j.tibtech.2009.06.006, 1879-3096.
[5] Anurag S. Rathore, QbD/PAT for bioprocessing: Moving from theory to implementation, Curr. Opin. Chem. Eng. (ISSN: 2211-3398) 6 (2014) 1–8, http://dx.doi.org/10.1016/j.coche.2014.05.006.
[6] Z. Kis, K. Tak, D. Ibrahim, S. Daniel, D. van de Berg, M.M. Papathanasiou, B. Chachuat, C. Kontoravdi, N. Shah, Quality by design and techno-economic modelling of RNA vaccine production for pandemic-response, Comput. Aided Chem. Eng. (ISSN: 1570-7946) 49 (2022) 2167–2172, http://dx.doi.org/10.1016/B978-0-323-85159-6.50361-4.
[7] R.S. Nunavath, M.T. Singh, A. Jain, M. Chakma, R. Arivuselvam, M.S.T.A. Azeeze, Quality by design in pharmaceuticals: A review of its impact on regulatory compliance and product quality, Drug Res. (ISSN: 2194-9379) 74 (1) (2023) 18–23, http://dx.doi.org/10.1055/a-2185-4916.
[8] Lawrence X. Yu, Pharmaceutical quality by design: Product and process development, understanding, and control, Pharm. Res. (ISSN: 0724-8741) 25 (4) (2008) 781–791, http://dx.doi.org/10.1007/s11095-007-9511-1.
[9] G.V. Reklaitis, J. Khinast, F. Muzzio, Pharmaceutical engineering science—New approaches to pharmaceutical development and manufacturing, Chem. Eng. Sci. (ISSN: 0009-2509) 65 (21) (2010) iv–vii, http://dx.doi.org/10.1016/j.ces.2010.08.041.
[10] Lawrence X. Yu, Gregory Amidon, Mansoor A. Khan, Stephen W. Hoag, James Polli, G.K. Raju, Janet Woodcock, Understanding pharmaceutical quality by design, AAPS J. (ISSN: 1550-7416) 16 (4) (2014) 771–783, http://dx.doi.org/10.1208/s12248-014-9598-3.
[11] Anurag S. Rathore, Sumit K. Singh, Production of protein therapeutics in the quality by design (QbD) paradigm, in: Zuben E. Sauna, Chava Kimchi-Sarfaty (Eds.), Protein Therapeutics, Springer International Publishing, Cham, ISBN: 978-3-319-41818-6, 2017, pp. 41–67, http://dx.doi.org/10.1007/7355_2015_5004.
[12] H. Narayanan, M.F. Luna, M. von Stosch, M.N. Cruz Bournazou, G. Polotti, M. Morbidelli, A. Butté, M. Sokolov, Bioprocessing in the digital age: The role of process models, Biotechnol. J. (ISSN: 1860-6768) 15 (1) (2020) http://dx.doi.org/10.1002/biot.201900172.
[13] H. Narayanan, F. Dingfelder, I. Condado Morales, B. Patel, K.E. Heding, J.R. Bjelke, T. Egebjerg, A. Butté, M. Sokolov, N. Lorenzen, P. Arosio, Design of biopharmaceutical formulations accelerated by machine learning, Mol. Pharm. (ISSN: 1543-8384) 18 (10) (2021) 3843–3853, http://dx.doi.org/10.1021/acs.molpharmaceut.1c00469.
[14] H. Narayanan, F. Dingfelder, A. Butté, N. Lorenzen, M. Sokolov, P. Arosio, Machine learning for biologics: Opportunities for protein engineering, developability, and formulation, Trends Pharmacol. Sci. (ISSN: 0165-6147) 42 (3) (2021) 151–165, http://dx.doi.org/10.1016/j.tips.2020.12.004.
[15] Francesco Destro, Massimiliano Barolo, A review on the modernization of pharmaceutical development and manufacturing – Trends, perspectives, and the role of mathematical modeling, Int. J. Pharm. (ISSN: 0378-5173) 620 (2022) 121715, http://dx.doi.org/10.1016/j.ijpharm.2022.121715.
[16] R.V. Barenji, R.E. Hariry, D. Demirkol, T.U. Daim, Research landscape analysis for quality in Pharma 4.0 era, Technol. Soc. (ISSN: 0160-791X) 76 (2024) http://dx.doi.org/10.1016/j.techsoc.2024.102472.
[17] Harini Narayanan, Mattia Sponchioni, Massimo Morbidelli, Integration and digitalization in the manufacturing of therapeutic proteins, Chem. Eng. Sci. (ISSN: 0009-2509) 248 (2022) 117159, http://dx.doi.org/10.1016/j.ces.2021.117159.
[18] Steven Sachio, Cleo Kontoravdi, Maria M. Papathanasiou, A model-based approach towards accelerated process development: A case study on chromatography, Chem. Eng. Res. Des. (ISSN: 02638762) 197 (2023) 800–820, http://dx.doi.org/10.1016/j.cherd.2023.08.016.
[19] Steven Sachio, Blaž Likozar, Cleo Kontoravdi, Maria M. Papathanasiou, Computer-aided design space identification for screening of protein A affinity chromatography resins, J. Chromatogr. A (ISSN: 0021-9673) 1722 (2024) 464890, http://dx.doi.org/10.1016/j.chroma.2024.464890.
[20] Fabian Steinebach, Monica Angarita, Daniel J. Karst, Thomas Müller-Späth, Massimo Morbidelli, Model based adaptive control of a continuous capture process for monoclonal antibodies production, J. Chromatogr. A (ISSN: 0021-9673) 1444 (2016) 50–56, http://dx.doi.org/10.1016/j.chroma.2016.03.014.
[21] C. Varsakelis, S. Dessoy, M. von Stosch, A. Pysik, Show me the money! Process modeling in pharma from the investor's point of view, Processes 7 (9) (2019) http://dx.doi.org/10.3390/pr7090596.
[22] S. Zobel-Roos, A. Schmidt, L. Uhlenbrock, R. Ditz, D. Köster, J. Strube, Digital twins in biomanufacturing, Adv. Biochem. Eng. Biotechnol. 176 (2021) 181–262, http://dx.doi.org/10.1007/10_2020_146.
[23] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, Patrick T. Ellinor, Transfer learning enables predictions in network biology, Nature (ISSN: 1476-4687) 618 (7965) (2023) 616–624, http://dx.doi.org/10.1038/s41586-023-06139-9.
[24] Pavlos Kotidis, Panagiotis Demis, Cher H. Goey, Elisa Correa, Calum McIntosh, Stefania Trepekli, Nilay Shah, Oleksiy V. Klymenko, Cleo Kontoravdi, Constrained global sensitivity analysis for bioprocess design space identification, Comput. Chem. Eng. (ISSN: 00981354) 125 (2019) 558–568, http://dx.doi.org/10.1016/j.compchemeng.2019.01.022.
[25] Sergei Kucherenko, Dimitrios Giamalakis, Nilay Shah, Salvador García-Muñoz, Computationally efficient identification of probabilistic design spaces through application of metamodeling and adaptive sampling, Comput. Chem. Eng. (ISSN: 00981354) 132 (2020) 106608, http://dx.doi.org/10.1016/j.compchemeng.2019.106608.
[26] Sebastian Vogg, Thomas Müller-Späth, Massimo Morbidelli, Design space and robustness analysis of batch and counter-current frontal chromatography processes for the removal of antibody aggregates, J. Chromatogr. A (ISSN: 0021-9673) 1619 (2020) 460943, http://dx.doi.org/10.1016/j.chroma.2020.460943.
[27] Chaoying Ding, Marianthi Ierapetritou, A novel framework of surrogate-based feasibility analysis for establishing design space of twin-column continuous chromatography, Int. J. Pharm. (ISSN: 03785173) 609 (2021) 121161, http://dx.doi.org/10.1016/j.ijpharm.2021.121161.

[28] Margherita Geremia, Fabrizio Bezzo, Marianthi G. Ierapetritou, Design space determination of pharmaceutical processes: Effects of control strategies and uncertainty, Eur. J. Pharmaceut. Biopharmaceut. (ISSN: 0939-6411) 194 (2024) 159–169, http://dx.doi.org/10.1016/j.ejpb.2023.12.008.

[29] Christopher M. Bishop, Pattern Recognition and Machine Learning, in: Information Science and Statistics, Springer, ISBN: 978-0-387-31073-2, 2008.

[30] Matic Grom, Mirijam Kozorog, Simon Caserman, Andrej Pohar, Blaž Likozar, Protein A affinity chromatography of Chinese hamster ovary (CHO) cell culture broths containing biopharmaceutical monoclonal antibody (mAb): Experiments and mechanistic transport, binding and equilibrium modeling, J. Chromatogr. B (ISSN: 15700232) 1083 (2018) 44–56, http://dx.doi.org/10.1016/j.jchromb.2018.02.032.

[31] I.M. Sobol', On the distribution of points in a cube and the approximate evaluation of integrals, USSR Comput. Math. Math. Phys. (ISSN: 0041-5553) 7 (4) (1967) 86–112, http://dx.doi.org/10.1016/0041-5553(67)90144-9.