Received: 8 March 2021 | Accepted: 12 March 2021

DOI: 10.1002/asmb.2620

SPECIAL ISSUE PAPER

WILEY

Design of experiments and manufacturing design space for multi-step processes

Rosamarie Frieri[1,2] | **Marco Mariti**[2] | **Marilena Paludi**[2]

[1]Department of Statistical Sciences "P. Fortunati", University of Bologna, Bologna, Italy

[2]GSK, Technical R&D, Siena, Italy

Correspondence
Rosamarie Frieri, Department of Statistical Sciences "P.Fortunati", University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy.
Email: rosamarie.frieri2@unibo.it

Funding information
GlaxoSmithKline

Abstract

Most industrial processes are composed of multiple subsequent steps. In this article, we provide a statistical approach to design experiments and to define the manufacturing design space of multi-step processes by taking into account the complex system of interactions among steps. We consider each intermediate outcome as an additional input factor in the next step and we plan experiments following a particular sequential structure. To encompass the potential deviations from the target levels of such input factors, designs are selected according to the D-optimality in average criterion and, in order to assess their prediction capabilities, a suitable extension of the fraction of design space technique has been proposed. The manufacturing design space of the process is then defined by combining the interconnected manufacturing design spaces of the process steps and by deriving the linear combination of the process inputs that ensures the required quality standard for the final outcome. Appealing properties of this approach are also shown by the application to a three-steps biochemical process of expression and purification of a recombinant protein in which 10 input factors are included in the design.

KEYWORDS

average D-optimality, manufacturing design space, multi-step processes, quality by design

1 | INTRODUCTION

A multi-step process is a system composed of multiple subsequent stages: the quality of the final product is the result of the interactions among the steps. More specifically, the quality of the product at one stage is not only affected by the operations performed at that stage, but also by those of previous stages.[1] Dealing with multi-step processes is common in many industrial fields, but their complex structure still presents significant challenges for researchers.

In particular in pharmaceutical industry, the introduction of the quality by design (QbD)[2] paradigm has increased the demand of systematic and science-based approaches to support pharmaceutical development and manufacturing activities. QbD principles have been developed to guarantee high level of product and process understanding and, hence, high level of quality for patients.[3,4] The QbD implementation includes the identification of the critical quality attributes (CQAs) of pharmaceutical products, defined as those attributes which impact the clinical performance. Then, to ensure product quality, the manufacturing process should be designed to obtain these attributes consistently at the desired values.

The multidimensional combination and interaction of inputs and process parameters such that the final CQAs meet the desired ranges is defined as the *manufacturing design space*, *which is one of the fundamental concepts in the QbD paradigm.[2]

Clearly, the establishment of the manufacturing design space in a multi-step context should take into account that (i) the quality of the final product is the result of the combination of operations and resources employed in the subsequent steps, which are, in turn, controlled by multiple input factors/process parameters and (ii) input factors of different steps interact to determine the intermediate and the final outcome/CQA. Such structure and the lack of first principle models pose technical issues in the design, analysis, and optimization of multi-step processes.

Usually, given the complexity of the problem, the process is investigated one step at-a-time.[5] This strategy fails to detect potentially critical interactions between steps and requires a remarkable amount of experimental resources. On the other hand, by approaching the multi-step process as a big single-stage with traditional experimental designs would involve a large study, with potentially around 10 or more interacting factors, usually not affordable in practice (for biochemical processes the number of factors typically investigated in a single experiment is below 10). Indeed, a full description and understanding of the system should explore all the potential interactions between all the input factors of the process, requiring a high experimental effort. Therefore a suitable strategy to design experimental studies for multi-step processes is fundamental for scientists, especially in pharmaceutical industry in which the compliance with regulatory guidelines is often matched with the need to speed-up the development process.

Literature on multi-step processes has mainly addressed process robustness, control and optimization and discussion of case studies.[6-8] To the best of our knowledge, research on design of experiment in a multi-stage framework has been focused on split-plot designs and its variants (see the review of Yuangyai and Lin[9]). Multi-stage experimentation can be performed in a split-plot fashion, but the complexity ramp up as the number of steps increases.[5,10] In general, the proposals are not flexible in terms of number of steps involved, number of factors, number of levels for each factor, and type of experimental design used.[9]

In this article, we introduce an efficient approach to design an experimental study for a multi-step process. Motivated by QbD principles, the final objective of the design strategy is the determination of the manufacturing design space of the whole process. The procedure consists in designing a set of experiments to fit an empirical model for the final outcome. We present a new framework in which the output of each step is included as an input factor in the subsequent step and experiments are planned following a particular sequential structure. The setting of the additional input factor is subject to error since it can be set only by changing the input factors of the previous step, according to a model. Then, to fit a tentative model for each step, the D-optimal design in average[11,12] has been adopted as a design criterion. Since the error in setting input factors affects the prediction properties of the design, we propose an extension of the fraction of design space technique, which includes this effect. Once the model for the final outcome has been estimated, starting from the quality requirements on the final product, the multi-step manufacturing design space is derived by combining the interconnected manufacturing design spaces of the steps. As a result, a set of operating ranges for all the directly controlled input factors of the process is obtained. The application of the procedure is particularly relevant for—but not limited to—biochemical processes within pharmaceutical industry, for which we report an illustrative case-study example.

This article is organized as follows: in Section 2, we formalize the structure of multi-step processes and we introduce the notation, and in Section 3, we present the sequential strategy to design experiments. In Section 4, we report the method to define the multi-step manufacturing design space, while in Section 5, the complete procedure is implemented for a three-step process of expression and purification of a recombinant protein. In Section 6, we conclude with a discussion and future developments.

2 | MULTI-STEP PROCESSES AND NOTATION

Let us consider a process made of V steps, $S_1, ..., S_V$, and let us denote with $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots)'$ the set of controllable input factors of S_i, for $i = 1, \dots, V$, and with $y^{(V)}$ the output of the process (Figure 1). The main concern of multi-step systems is that the behavior of the final outcome depends both on the effect of process inputs and on the interactions among steps, namely $y^{(V)}$ depends on $\mathbf{x}^{(i)}, \forall i = 1, \dots, V$ and their interactions.

*Note that in the ICH guideline Q8 on pharmaceutical development[2] this CQAs-consistent production space is named *design space*. However, this terminology may be ambiguous in the context of this manuscript since the term design space is commonly used in the design of experiment literature to refer to the set of all the possible design points of the experiment.

FIGURE 1 Multi-step process with V steps

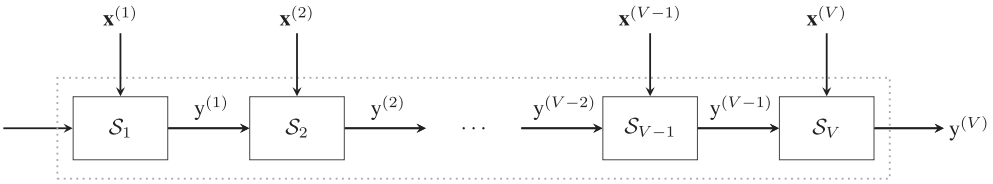


FIGURE 2 Multi-step process with V steps: new modeling

In such complex systems, mechanistic models relating inputs and outputs are not generally available so that researchers have to rely on Design of Experiment (DoE) techniques.[13] Hence, empirical models are devoted to the identification of the manufacturing design space, which then consists in learning how to set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(V)}$ in such a way that $y^{(V)}$ has the desired characteristics (e.g., quality, safety).

On the one hand, if the process is intended as a big single stage (Figure 1), the behavior of $y^{(V)}$ can be typically described by a regression model $y^{(V)} = f(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(V)}; \boldsymbol{\beta}) + \varepsilon_V$ where $f(\cdot; \boldsymbol{\beta})$ is the model response, $\boldsymbol{\beta}$ is the unknown model parameters vector and ε_V is a random error term. Designing an experiment to fit such a complex model would require, in general, a high number of runs, due to the high number of parameters to estimate. Even if few steps are involved in the process, the planning of the experiment with traditional DoE techniques[13] is hardly affordable in practice. Indeed, let us consider an experimental scenario with $V = 3$, in which the experimenter would like to study three input factors per step. Assuming $f(\cdot, \boldsymbol{\beta})$ linear in $\boldsymbol{\beta}$ involving main, quadratic and two factor interactions effects, the number of parameters to be estimated in the model for $y^{(3)}$ would be 54 (plus the intercept). On the other hand, designing experiments by not taking into account the interactions among steps (e.g., by changing input factors one-step-at-a-time and observing the final outcome) would require prohibitive experimental resources giving only partial knowledge of the process. These drawbacks call for an *ad hoc* design strategy to effectively plan experiments in a multi-step context.

Let us assume that the output of each intermediate step can be observed, so that a multi-step process can be represented as in Figure 2, where $y^{(i)}$ is the outcome of S_i for $i = 1, \dots, V$. In contrast to the setting of Figure 1, we consider the structure of the process as made of multiple components. From now on, we assume that the experimenter is interested in a single characteristic (i.e., a single CQA in the context of pharmaceutical development) of each intermediate material. This assumption is relevant for the development to our proposal (for a thorough discussion, see Section 6): only one output is carried out as an input in the next step.

Within this setting, the first step and so $y^{(1)}$, depends on $\mathbf{x}^{(1)}$ while each $y^{(i)}$ for $i = 2, \dots, V$ is affected by the input factors proper of that step, $\mathbf{x}^{(i)}$, but also by the outcome of the previous one, $y^{(i-1)}$. We then introduce a mathematical formalization of a multi-step process in which each step is characterized by a model,

$$\begin{aligned} y^{(1)} &= f_1(\mathbf{x}^{(1)}; \boldsymbol{\beta}^{(1)}) + \varepsilon^{(1)}, \\ &\vdots \\ y^{(i)} &= f_i(\mathbf{x}^{(i)}, y^{(i-1)}; \boldsymbol{\beta}^{(i)}) + \varepsilon^{(i)}, \\ &\vdots \\ y^{(V)} &= f_V(\mathbf{x}^{(V)}, y^{(V-1)}; \boldsymbol{\beta}^{(V)}) + \varepsilon^{(V)}, \end{aligned} \tag{1}$$

where for the i-th step, $f_i(\cdot)$ is the response model, $\mathbf{x}^{(i)}$ is the vector of input factors, $\boldsymbol{\beta}^{(i)}$ is the vector of the unknown model parameters and $\varepsilon^{(i)}$ is a random error. We assume $f_i(\cdot)$ linear in the parameters so that the i-th equation of (1) can be restated as $y^{(i)} = \mathbf{z}_i' \boldsymbol{\beta}^{(i)} + \varepsilon^{(i)}$, where $\mathbf{z}_i'$ is the vector (of size p_i) whose components are the term included in the model $f_i(\cdot)$. For instance, if the i-th step has two input factors and $f_i(\cdot)$ is a full quadratic model (without intercept) then $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)})$ and $\mathbf{z}_i' = (x_1^{(i)}, x_2^{(i)}, y^{(i-1)}, x_1^{(i)} \cdot x_2^{(i)}, x_1^{(i)} \cdot y^{(i-1)}, x_2^{(i)} \cdot y^{(i-1)}, [x_1^{(i)}]^2, [x_2^{(i)}]^2, [y^{(i-1)}]^2)$ with $p_i = 9$. In addition, by letting $\mathbf{y}^{(i)}$ be the vector of the n_i observations from step i, then $\mathbf{y}^{(i)} = \boldsymbol{Z}_i \boldsymbol{\beta}^{(i)} + \boldsymbol{\varepsilon}^{(i)}$ where $\boldsymbol{Z}_i$ is the design matrix expanded

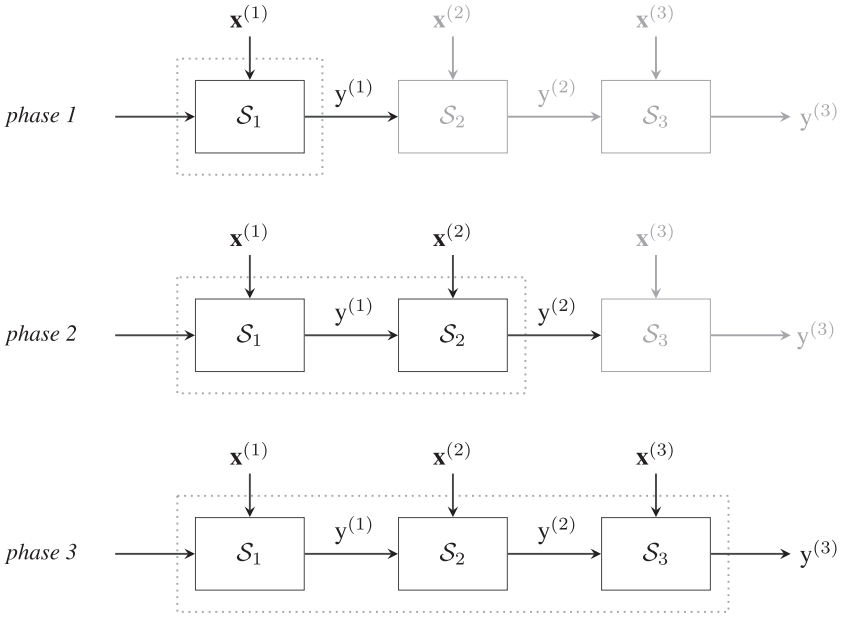


FIGURE 3 Multi-step experimental strategy for a three-step process

to the $f_i(\cdot)$ model form (i.e., it has size $n_i \times p_i$). For $\boldsymbol{\epsilon}^{(i)} = (\varepsilon_1^{(i)}, \dots, \varepsilon_u^{(i)}, \dots, \varepsilon_{n_i}^{(i)})'$ we assume that $\varepsilon_u^{(i)}$ has zero mean and variance $\sigma_i^2, \forall u = 1, \dots, n_i$ (homoschedasticity) and that the $\{\varepsilon_u^{(i)}\}$ are uncorrelated random variables. Finally, since we deal with situations in which the real experimental conditions may fluctuates around those specified in the design, that is, the actual design matrix could be different to the the planned design matrix, we denote with $\tilde{\boldsymbol{Z}}_i$ the observed design matrix expanded to the form of $f_i(\cdot)$.

3 | DESIGN OF EXPERIMENTS FOR MULTI-STEP PROCESSES

In this section, we present the design strategy to plan the experiments in multi-step processes. In general, the model for the final outcome is used to draw conclusions on the end-process material and so to derive the manufacturing design space. Following the set-up in (1), to fit an empirical model for $y^{(V)}$ we have to estimate V models. Thus, we need to generate V experimental plans: the procedure consists of a sequence of V phases where, in the generic *phase i*, the objective is to fit a model for the outcome of $\mathcal{S}_i$. On this purpose, *phase i* consists of (a) the design of the experiment for $\mathcal{S}_i$ (b) the implementation of the experiments to observe the $\mathbf{y}^{(i)}$ and finally (c) the analysis of the results for fit a model for $y^{(i)}$.

So in *phase i* (i.e., *phase* $1, \dots,$ *phase* $i-1$ have already been run), the experimental design for $\mathcal{S}_i$ is derived according to a suitable optimality criterion. Such experimental design provides target levels for $\mathbf{x}^{(i)}$ and $y^{(i-1)}$. However, since $y^{(i-1)}$ is the output of $\mathcal{S}_{i-1}$, its target levels cannot be set directly as for $\mathbf{x}^{(i)}$. The desired target values can be reached only after running the previous steps, $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{i-1}$, under proper settings of $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}$. In particular, $y^{(i-1)}$ can be controlled only through the inputs of $\mathcal{S}_{i-1}$ and the model estimates $\hat{f}_{i-1}(\cdot)$ (estimator of $f_{i-1}(\cdot)$) computed in *phase* $i-1$, but due to the prediction error intrinsic in each model, $y^{(i-1)}$ cannot be set precisely to the desired levels. More specifically, while all the input factors $\mathbf{x}^{(i)}$ in step i are subjected to a negligible experimental error, $y^{(i-1)}$ is also subjected to the error of the model for $\mathcal{S}_{i-1}$. Therefore, due to the multi-step structure of the process, each experimental plan involves one factor whose levels are set with error: the experiments should be designed by taking into account the potential deviations from the target levels that may occur due to this error.

To the best of our knowledge, the first author interested in the effect of error in setting factor levels was Box[14] and this topic has been more recently addressed by Pronzato[11] and Donev[12] in the context of optimal designs. However, robustness to error in setting factor levels of optimal designs has not been much addressed in the literature.[15]

In order to clarify the experimental strategy proposed, Example 1 and Figure 3 illustrate the multi-step procedure in the case of three steps.

Example 1. The multi-step experimental strategy is illustrated in Figure 3 in the case of a three-step process: the procedure consists of three phases. In *phase* 1 the experiments are performed on $\mathcal{S}_1$ to fit a model for $y^{(1)}$. In *phase* 2, the

experiments are performed on $\mathcal{S}_1$ and $\mathcal{S}_2$ to fit a model for $y^{(2)}$. However, note the difference between the experiments on $\mathcal{S}_1$ in *phase* 1 and in *phase* 2. In *phase* 1, the experiments on $\mathcal{S}_1$ are performed under some settings of $\mathbf{x}^{(1)}$ that are optimal to fit a model for $y^{(1)}$. In *phase* 2, the experiments on $\mathcal{S}_1$ are run once again but under different set-ups of $\mathbf{x}^{(1)}$, which are the ones ensuring $y^{(1)}$ close to the target levels optimal to fit a model for $y^{(2)}$. The same reasoning holds for *phase* 3.

Section 3.1 is focused on how to design the experiment for $\mathcal{S}_i$ at phase i and the complete and detailed procedure to implement the proposed experimental strategy from *phase* 1 up to *phase* V is reported in Section 3.2.

3.1 | Design of experiment for step *i* in phase *i*

Let us consider that *phase* 1, *phase* 2, … , *phase* $i-1$ have already been run, that is, $\hat{f}_1(\cdot)$, … ,$\hat{f}_{i-1}(\cdot)$ are available. For instance, *phase* $i-1$ is devoted to make experiments to fit a model for $y^{(i-1)}$, thus the experiments have been carried out from $\mathcal{S}_1$ up to $\mathcal{S}_{i-1}$ and the experimental conditions (values of $\mathbf{x}^{(i-1)}$ and $y^{(i-2)}$) that generated the observed $y^{(i-1)}$s are known so that $\tilde{\boldsymbol{Z}}_{i-1}$ is available and used to estimate the parameters $\hat{\boldsymbol{\beta}}^{(i-1)}$ via least squares.

We are now in *phase i*. First, we wish to design an experiment to fit an empirical model for $y^{(i)}$ (*phase i* (a)). Typically, input factors are set to target levels according to a given experimental plan. The plan is generated following optimality criteria, which are generally based on convex functions of the information matrix of the design[16] (e.g., the well-known D-optimality requires the maximization of its determinant or, equivalently, the minimization of the determinant of the inverse information matrix[17]).

However, for the i-th step,

(i) the levels of $\mathbf{x}^{(i)}$ are directly controllable and so adjustable to the target values required by the experimental plan,
(ii) the levels of $y^{(i-1)}$ can be controlled by running $\mathcal{S}_{i-1}$ and by changing $\mathbf{x}^{(i-1)}$ and $y^{(i-2)}$ according to the model $\hat{f}_{i-1}(\cdot)$.

In particular, the levels of $y^{(i-1)}$ are set through $\hat{f}_{i-1}(\mathbf{x}^{(i-1)}, y^{(i-2)}; \hat{\boldsymbol{\beta}}^{(i-1)}) = \tilde{\boldsymbol{z}}'_{i-1}\hat{\boldsymbol{\beta}}^{(i-1)}$. If following a design criterion τ_u is the target level for $y^{(i-1)}$ in the u-th experimental run and t_u is the design point such that $\hat{f}_{i-1}(t_u) = \tau_u$, then the point estimator $\hat{f}_{i-1}(t_u) = t'_u\hat{\boldsymbol{\beta}}^{(i-1)}$ is unbiased with variance

$$\eta_u^2 = \sigma_{i-1}^2(1 + t'_u(\tilde{\boldsymbol{Z}}'_{i-1}\tilde{\boldsymbol{Z}}_{i-1})^{-1}t_u), \tag{2}$$

which can be estimated by replacing σ_{i-1}^2 with the sample variance $\hat{\sigma}_{i-1}^2$ (note that the choice of t_u may not be unique, see Remark 1). Then, experimental data in the u-th run will be generated by $Y_u^{(i-1)} = \tau_u + e_u$ rather than by τ_u, where e_u is the error in setting $y^{(i-1)}$ in the u-th observation. By assuming e_u normally distributed with zero mean and variance η_u^2 (and also $\mathbb{E}[e_u \cdot e_j] = 0$ for $u \neq j$), the level of $y^{(i-1)}$ becomes a random variable, that is, $Y_u^{(i-1)} \sim N(\tau_u, \eta_u^2)$. Thus, the planned $\boldsymbol{Z}_i$ is composed of deterministic elements, associated to the input factors $\mathbf{x}^{(i)}$, and random elements associated to $Y^{(i-1)}$, that is, the generic row of $\boldsymbol{Z}_i$ is $\mathbf{z}'_i = (x_1^{(i)}, x_2^{(i)}, \dots, Y^{(i-1)}, x_1^{(i)} \cdot x_2^{(i)}, \dots, x_1^{(i)} \cdot Y^{(i-1)}, \dots, [x_1^{(i)}]^2, \dots, [Y^{(i-1)}]^2, \dots)$.

Remark 1. Consider the target level τ_u for $y^{(i-1)}$ in the u-th experimental run. There may be many possible choices of the design points t_u that satisfies $\hat{f}_{i-1}(t_u) = \tau_u$. From a statistical perspective, the choice of t_u can lead to different η_u^2, that is, to different magnitudes of the prediction variance. A natural choice is then to take the t_u for which the corresponding η_u^2 is the lowest. The selection of the most appropriate t_u could be also driven by practical considerations like cost reasoning or operative purposes.

Remark 2. Note that in *phase i* (a), the planned $\boldsymbol{Z}_i$ has random components due to the still unobserved outcomes of $\mathcal{S}_{i-1}$. Since we work under the assumption that intermediate outcomes can be measured, when the experiment on $\mathcal{S}_{i-1}$ in *phase i* (b) takes place, the realizations of $Y^{(i-1)}$ can be observed. Thus, the experimental conditions that generate $\boldsymbol{y}^{(i)}$ are not random: $\tilde{\boldsymbol{Z}}_i$ is the matrix $\boldsymbol{Z}_i$ conditional on the realizations of the $Y^{(i-1)}$s, and $\tilde{\boldsymbol{Z}}_i$ is not random (see also Donev[12]). Therefore least squares estimates of $\boldsymbol{\beta}^{(i)}$ can can be obtained in *phase i* (c) by $\hat{\boldsymbol{\beta}}^{(i)} = (\tilde{\boldsymbol{Z}}'_i\tilde{\boldsymbol{Z}}_i)^{-1}\tilde{\boldsymbol{Z}}'_i\boldsymbol{y}^{(i)}$.

In *phase i* (a) the design and so the choice of the target levels for both input factors set directly and with error of $\mathcal{S}_i$ should be made according to some optimality considerations. In the presence of error in setting the levels of some input factors, the average D-optimal criterion has been proposed.[11,12] Average D-optimal design is based on minimizing the expected value of the determinant of the random inverse information matrix, that is,

$$\min \quad \mathbb{E}[\det(\boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1}]. \tag{3}$$

The distribution of $\det(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}$ is very complex and, in general, not available in closed-form, so that we compute $\mathbb{E}[\det(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}]$ with Monte Carlo approximation. To implement the criterion in (3), we have extended the Fedorov's exchange algorithm.[17] Starting from an initial design of a given size, at each iteration the algorithm will select a point of the design to be removed and exchanged with a new point from a candidate set. The point is chosen as the one that gives the best improvement in terms of the criterion in (3). This procedure is iterated until no further exchanges are found to improve the criterion more than a given small threshold (e.g., 10^{-10}). The pseudo code is reported in the Appendix and the R code is available upon request to the first author.

The criterion in (3) provides a design which is optimal in average. The real experimental conditions recorded in $\tilde{\boldsymbol{Z}}_i$ are a single realization of $\boldsymbol{Z}_i$. However, in the next section we report a suitable extension of the fraction of design space (FDS) technique[18] to evaluate the impact of input factors set with error on the prediction capabilities of the design. In the exposition of Section 3.1.1, we refer to the *i*-th step of the process in *phase i*.

3.1.1 | Prediction properties of designs with error in factor levels

The most used measures of the prediction properties of a response surface design are the scaled prediction variance (*SPV*) and the unscaled prediction variance (*UPV*), which, for a given extended design matrix $\tilde{\boldsymbol{Z}}_i$ and a point $\tilde{\mathbf{z}}_0$ (expanded to the model space), are given by $SPV(\tilde{\mathbf{z}}_0) = n_i\tilde{\mathbf{z}}_0'(\tilde{\boldsymbol{Z}}_i'\tilde{\boldsymbol{Z}}_i)^{-1}\tilde{\mathbf{z}}_0$ and $UPV(\tilde{\mathbf{z}}_0) = SPV(\tilde{\mathbf{z}}_0)/n_i$ respectively. Zahran et al.[18] introduced the fraction of design space technique to both assess the prediction capability of a single design and make comparisons between competing designs. They present a graphical method, the FDS plot, to quantify the fraction of the design space with *SPV* less than or equal than any *SPV* values. We have extended such technique to evaluate designs in presence of error in factor levels. Let us assume to be in *phase i* (a). In this case, due to the random $\boldsymbol{Z}_i$, in any $\tilde{\mathbf{z}}_0$, instead of a single value we get a distribution of $UPV(\tilde{\mathbf{z}}_0)$. As analogous to the *UPV*, which is sorted and displayed in the FDS plot, we consider $QPV_{1-\alpha}$, defined, for each design point, as the $1-\alpha$ quantile of $UPV(\tilde{\mathbf{z}}_0)$. Hence, for a given fraction of the design space γ, the corresponding $QPV_{1-\alpha}$ value denotes that the $1-\alpha\%$ of the possible designs occurrence have the $\gamma\%$ of the design space with *UPV* at or below that value (a typical value used in experiments is $\gamma\% = 80\%$).

For numerical comparisons, we consider the half-width of the confidence interval for the predicted mean (divided by the expected variability on the tentative model σ_i^2), computed with respect to the $1-\alpha$ quantile of the *UPV* distribution. In particular, by letting $d_\gamma = q_\gamma(UPV) \cdot t_{0.975,\ n_i-p_i}$, where q_γ is the γ-quantile of the *UPV* and $t_{0.975,n_i-p_i}$ is the 0.975 quantile of the student-t distribution with $n_i - p_i$ degrees of freedom, each d_γ value indicates that the $\gamma\%$ of the design space is precise enough to predict the mean within $\pm\hat{\sigma}_i d_\gamma$. To accommodate the error in factor levels, following similar reasoning, we define

$$d_{\gamma,\alpha} = q_\gamma(\sqrt{QPV_{1-\alpha}}) \cdot t_{1-0.05/2,\ n_i-p_i},$$

whose value denotes that the $1-\alpha\%$ of the realizations of the designs have the $\gamma\%$ of the design space precise enough to predict the mean within $\pm\hat{\sigma}_i d_{\gamma,\alpha}$.

Notice that the usefulness of $QPV_{1-\alpha}$ and $d_{\gamma,\alpha}$ is not limited to multi-step processes. Also in the case of a single-step process, in which for some reason one or more input factors are set with error, these tools may be adopted to evaluate the prediction capabilities of the experimental designs.

3.2 | Experimental strategy for multi-step processes

In this section, we report the complete procedure that starting from $\mathcal{S}_1$ provides a model for $\mathcal{S}_V$. According to our proposal, the experiments are performed following a particular sequential structure. The experimental strategy is articulated in a sequence of phases in which the same operations are substantially repeated by adding each time one more step as follows (see also Table 1).

phase 1 *(a)* **Design the experiment for step 1.** The experimental plan can be derived in this case according to a classical designs (i.e., a design that does not consider input factors set with error, like D-optimal design, central composite design).

(b) **Perform the experiment on step 1.** Perform the n_1 runs on $\mathcal{S}_1$.

(c) **Analyze the results from step 1.** Estimate the parameters $\boldsymbol{\beta}^{(1)}$ of the response model $f_1(\cdot)$ and the variance of the observations σ_1^2.

phase 2 *(a)* **Design the experiment for step 2.** Use information from *phase* 1 *(c)* to find the D-optimal design in average for step 2 as follows. Fix a range $[y_L^{(1)}, y_U^{(1)}]$ of values to explore in input for $\mathcal{S}_2$, choose target levels within this interval (e.g., *low*, *mid*, and *high* level) and compute the corresponding variance of the error in setting $y^{(1)}$ to the desired levels (e.g., $\eta^2_{low}, \eta^2_{mid}, \eta^2_{high}$). Implement D-optimality in average to find the optimal design for $\mathcal{S}_2$ (with n_2 runs); compute $\boldsymbol{Z}_2$.

(b) **Perform the experiments on step 1 and step 2.** Perform the n_2 runs on $\mathcal{S}_1$ and $\mathcal{S}_2$. The input factors levels of $\mathcal{S}_1$ are set to obtain the outputs of $\mathcal{S}_1$ close to the target levels computed as optimal for $\mathcal{S}_2$. At the end of $\mathcal{S}_1$ record the realizations of $Y_u^{(1)}$, $\forall u = 1, \ldots, n_2$ (which are different from those obtained at the end of the *phase* 1), so that the real experimental conditions (i.e., the observed design matrix) that generate the $\boldsymbol{y}^{(2)}$ are known; compute $\tilde{\boldsymbol{Z}}_2$.

(c) **Analyze the results from step 2.** Estimate the parameters $\boldsymbol{\beta}^{(2)}$ of the response model $f_2(\cdot)$ and the variance of the observations σ_2^2.

⋮ ⋮

phase i (a) **Design the experiment for step 2.** Use information from *phase* $i-1$ *(c)* and implement D-optimality in average to find the optimal design for $\mathcal{S}_i$ (with n_i runs); compute $\boldsymbol{Z}_i$.

(b) **Perform the experiments on step 1, step 2, … step *i*.** Perform n_i runs on $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i$. For step $\mathcal{S}_{i-1}$ factors levels are set to obtain the output of $\mathcal{S}_{i-1}$ close to the target levels computed as optimal for $\mathcal{S}_i$ and factors' levels for $\mathcal{S}_{i-2}, ..., \mathcal{S}_1$ are then set accordingly. At the end of $\mathcal{S}_{i-1}$ record the realizations of $Y_u^{(i-1)}$, $\forall u = 1, \ldots, n_i$ (which are different from those obtained at the end of *phase* $i-1$) so that the real experimental conditions (i.e., the observed design matrix) that generate the $\boldsymbol{y}^{(i)}$ are known; compute $\tilde{\boldsymbol{Z}}_i$.

(c) **Analyze the results from step *i*.** Estimate the parameters $\boldsymbol{\beta}^{(i)}$ of the response model $f_i(\cdot)$ and the variance of the observations σ_i^2.

⋮ ⋮

phase V (a) **Design the experiment for step *V*.** Use information from *phase* $V-1$ *(c)* and implement D-optimality in average to find the optimal design for $\mathcal{S}_V$ (with n_V runs); compute $\boldsymbol{Z}_V$.

(b) **Perform the experiments on step 1, step 2, …, step *V*.** Perform n_V runs on all the process steps and proceed similarly to *phase i (b)*; compute $\tilde{\boldsymbol{Z}}_V$.

(c) **Analyze the results from step *V*.** Estimate the parameters $\boldsymbol{\beta}^{(V)}$ of the response model $f_V(\cdot)$ and the variance of the observations σ_V^2.

TABLE 1 Strategy of experimentation for V-step processes

Phase	Experiment	Input	How is set	Design criterion	Output	For next step
phase 1	n_1 runs on $\mathcal{S}_1$	$\mathbf{x}^{(1)}$	directly	Classic DoE	observe $\boldsymbol{y}^{(1)}, \tilde{\boldsymbol{Z}}_1$, compute $\hat{f}_1(\cdot)$	$[y_L^{(1)}, y_U^{(1)}]$ η_u^2
phase 2	n_2 runs on $\mathcal{S}_1$ and $\mathcal{S}_2$	$\mathbf{x}^{(2)}$ $y^{(1)}$	directly by changing $\mathbf{x}^{(1)}$ through $\hat{f}_1(\cdot)$	Average D-optimal ($\boldsymbol{Z}_2$)	observe $\boldsymbol{y}^{(2)}, \tilde{\boldsymbol{Z}}_2$; compute $\hat{f}_2(\cdot)$	$[y_L^{(2)}, y_U^{(2)}]$ η_u^2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
phase i	n_i runs on $\mathcal{S}_1, ..., \mathcal{S}_i$	$\mathbf{x}^{(i)}$ $y^{(i-1)}$	directly by changing $\mathbf{x}^{(i-1)}$ and $y^{(i-2)}$ through $\hat{f}_{i-1}(\cdot)$	Average D-optimal ($\boldsymbol{Z}_i$)	observe $\boldsymbol{y}^{(i)}, \tilde{\boldsymbol{Z}}_i$; compute $\hat{f}_i(\cdot)$	$[y_L^{(i)}, y_U^{(i)}]$ η_u^2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
phase V	n_V runs on $\mathcal{S}_1, ..., \mathcal{S}_V$	$\mathbf{x}^{(V)}$ $y^{(V-1)}$	directly by changing $\mathbf{x}^{(V-1)}$ and $y^{(V-2)}$ through $\hat{f}_{V-1}(\cdot)$	Average D-optimal ($\boldsymbol{Z}_V$)	observe $\boldsymbol{y}^{(V)}, \tilde{\boldsymbol{Z}}_V$; compute $\hat{f}_V(\cdot)$	$[y_L^{(V)}, y_U^{(V)}]$ η_u^2

TABLE 2 Number of parameters to be estimated: single-stage DoE versus multi-step DoE (S_2 and S_3 involve an additional input factor)

Controlled inputs			Number of parameters	
S_1	S_2	S_3	Single-stage DoE	Multi-step DoE
2	2	2	28	26
3	3	3	55	40
4	4	4	91	57
F	F	F	$(9F^2+9F+2)/2$	$(3F^2+13F+14)/2$

The complete procedure requires $n_1+n_2+\ldots+n_V$ runs where the n_1 runs are performed only on step 1 (*phase* 1), the n_2 runs are performed on step 1 and step 2 (*phase* 2) and so on. At the end of *phase* V, we obtain V fitted models describing each step of the process. An example for a case study is reported in Section 5 for $V=3$.

In the next example, we show the potential advantages of our proposal in terms of required experimental effort.

Example 2. In the case of a three-step process with three input factors per step and a full tentative quadratic model with intercept for $y^{(3)}$, we report in Table 2 the number of parameters to be estimated in the case the experimenter considers the process as a single-stage (single-stage DoE) and in the proposed multi-step approach (multi-step DoE). We compared the two procedures as the number of controllable input factors per stage (F) increases.

Clearly a larger number of parameters to estimate requires a higher number of runs. However, note that the single-stage and the multi-step procedures cannot be directly compared in terms of the number of runs. This number depends on the design criterion adopted and on the required prediction properties of the experimental design. The main advantage of our set-up is related to the number of steps involved in each experiment: according to the single-stage DoE, each of the planned runs involve all the process' steps, whereas adopting our proposal the number of runs carried out from the first up to the last step is limited, inducing, in general, a gain in terms of experimental resources. This example also highlights how the saving of experimental resources brought by the multi-step DoE increases as the number of controllable input factors per stage increases. Finally, note also that under the single-stage approach, the single model for the final outcome it is likely very complex. In this case, the assumption of a full-quadratic tentative model may often be too stringent.

4 | MANUFACTURING DESIGN SPACE DEFINITION FOR MULTI-STEP PROCESSES

Once the procedure of Section 3.2 has been performed, V fitted models, one for each outcome, are available. These models are used to define a set of interconnected acceptable ranges expressed in terms of the directly controlled input factors of the process. We will refer to *multi-step* manufacturing design space (denoted by DS) to intend the manufacturing design space of the entire process and to *individual* manufacturing design space as this space for a single step (denoted by DS_i for S_i).

Starting from the last step, let us assume that the quality target for the final outcome of the process requires $y^{(V)} \in R$. Then, from $\hat{f}_{V-1}(\cdot)$, the corresponding multidimensional combination of $\mathbf{x}^{(V)}$ and $y^{(V-1)}$ such that $y^{(V)} \in R$ is derived. This region is usually restricted by a confidence/prediction interval on the fitted response, giving an acceptable (more robust) smaller region for $\mathbf{x}^{(V)}$ and $y^{(V-1)}$. Then, the determined range for $y^{(V-1)}$ becomes the quality target used to derive the individual manufacturing design space for S_{V-1}, so that DS_{V-1} is defined from $\hat{f}_{V-1}(\cdot)$ and is expressed in terms of $\mathbf{x}^{(V-1)}$ and $y^{(V-2)}$.

The same reasoning is iterated until the manufacturing design space for $y^{(1)}$, which will be expressed in terms of $\mathbf{x}^{(1)}$, is derived. The combination of the acceptable ranges of the directly controllable input factors gives the multi-step manufacturing design space

$$DS = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(V)} \text{ such that } y^{(V)} \in R\}. \tag{4}$$

A similar procedure has been adopted in the work of Eon-Duval et al.[19] but the authors do not address two fundamental issues. First, they do not formally provide a procedure to model a multi-step process and second they do not take into account the multi-step structure of the process in the designing of experiments.

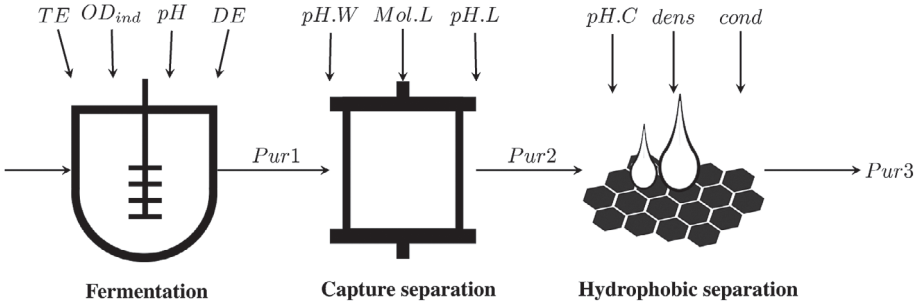


FIGURE 4 Three-step process of expression and purification of a recombinant protein

Remark 3. Two potential presentations of the manufacturing design space are reported in the guidelines.[2] (i) It can be defined by a non-linear combination of inputs' ranges that makes $y^{(V)} \in R$. In this case, the manufacturing design space is explained by mathematical equations describing relationships between inputs that lead to successful outputs. While this approach allows the maximum operative range to achieve the required quality standard, it makes the manufacturing design space be a complex set. Otherwise (ii) the manufacturing design space can be defined as a smaller region, based on a linear combination of input factors. Even if this approach is more limiting, it is often preferred in the applications due to operational simplicity. For this reason, in this article we adopt definition in (ii). In principle, the experimenter could select any sub-region based on a linear combination of inputs; often this choice is driven by scientific and practical considerations. For example in Figure 8, the manufacturing design space as in (i) is the yellow region, whereas the manufacturing design space as in (ii) is the red-delimited rectangle.

5 | CASE STUDY: MANUFACTURING DESIGN SPACE FOR A THREE-STEP BIOCHEMICAL PROCESS

In this section, we implement our proposal to an illustrative case-study. We considered a biochemical process commonly used in pharmaceutical industry to produce and purify recombinant proteins expressed by *Escherichia coli*. The process consists of three separate steps (Figure 4):

- *Fermentation*, where the *Escherichia coli* culture is grown and the recombinant protein is expressed in the bacterial cells;
- *Capture separation*, which is the first purification step where the target recombinant protein is captured in the column and a first portion of impurities is removed (DNA fragments, HCP, endotoxin etc);
- *Hydrophobic separation*, which is the last purification step, where the hydrophobic interactions are used to separate our target recombinant proteins from other residual proteins.

For the fermentation step, four input factors were identified as potentially critical process parameters and included in the study: trace element concentration in the fermentation media (*TE*), optical density of induction (OD_{ind}), pH of the fermentation media (*pH*), which is maintained fixed during the whole fermentation and the duration of the expression phase (*DE*), that is, the time between the induction point and the end of fermentation. Since the protein purity from the fermentation is expected to impact the performance of the purification steps, the output selected to be included as input factor for the capture separation step is the protein purity in the capture load material (denoted by *Pur*1). In the capture step, instead, we considered three process parameters and *Pur*1 as input factors. The three capture process parameters are the following: pH level of the wash (*pH.W*), the molarity, and the pH of the load material (*Mol.L* and *pH.L* respectively). Also in this case, the purity of the target protein (*Pur*2) has been identified as the output of the capture to be included as input in the hydrophobic separation. In this latter step, three process parameters have been identified as potentially critical for the final purity: the pH of the material loaded in the column (*pH.C*), the protein concentration, and the conductivity of the loaded material (*dens* and *cond*, respectively). Finally, we include the purity of the incoming material (*Pur*2) as additional input factor.

Other steps, like the final filtration or the centrifugation at the end of the fermentation steps were not considered in our study, however the approach can be easily extended to include them. The aim of the study is to determine the multi-step manufacturing design space corresponding, in this case, to the region of input factors/process parameters

TABLE 3 Step 1: Fermentation

(A) Input factors of Step 1 and ranges

$\mathbf{x^{(1)}}$	*Input*	*Range*
$x_1^{(1)}$	*TE*	[1.3, 1.7]
$x_2^{(1)}$	OD_{ind}	[3.0, 7.0]
$x_3^{(1)}$	*pH*	[6.2, 6.8]
$x_4^{(1)}$	*DE*	[5.0, 7.0]

(B) Output of Step 1, coded level in brackets.

Output		*Input*						
Level	*Pur*1	OD_{ind}		*pH*		*DE*		η_u^2
low	13.32	7.00	(+1.00)	6.80	(+1.00)	5.00	(−1.00)	0.71^2
medium	15.54	6.70	(+0.85)	6.41	(−0.30)	6.35	(+0.35)	0.68^2
high	17.75	4.52	(−0.24)	6.20	(−1.00)	7.00	(+1.00)	0.70^2

which can guarantee a final purity of the target protein (*Pur*3) above 88%. More specifically, in the following, the protein purity is defined as the relative amount of the recombinant protein of interest with respect to the total amount of all the components present in the material, that is, host cell proteins, fragmented recombinant protein, aggregated forms. The methods used to determine the protein purity are: SDS-PAGE after fermentation and reverse phase HPLC after capture separation and after hydrophobic separation.

In Section 5.1, we report the design and analysis of the experiments for the three steps, and in Section 5.2, we derive the multi-step manufacturing design space.

5.1 | Design and analysis of three-step process

STEP 1: Fermentation

For each input factor of the fermentation, five levels were identified within the ranges reported in Table 3(A).

To determine the relationship between *Pur*1 and the four fermentation parameters, a modified face-centered central composite design with 34 runs was defined (three extra runs were added as confirmation runs). From the results reported in Table B1, the fitted model for *Pur*1 (in terms of the coded unit) is

$$\hat{Pur}1 = 16.32 - 0.52 \cdot OD_{ind} - 0.44 \cdot pH + 0.94 \cdot DE - 1.10 \cdot OD_{ind}^2, \tag{5}$$

with $\hat{\sigma}_1^2 = 0.64^2$ and $R^2 = 0.74$ (predictive $R^2 = 0.63$). We used the model in (5) to determine the combination of OD_{ind}, *pH*, and *DE* such that *Pur*1 reaches its minimum value, 13.32 (for $OD_{ind} = 1, pH = 1, DE = -1$), and maximum, 17.75 (for $OD_{ind} = -0.24, pH = -1, DE = 1$). Within this interval, we selected three target levels for *Pur*1, low, medium, and high, in which we compute η_u^2, as in (2). These information are summarized in Table 3(B).

STEP 2: Capture separation

In order to fit a quadratic tentative model with two factor interactions effects, for each of the four factors of this step (see Table 4(A)), we consider three levels, $[-1, 0, 1]$. As far as *Pur*1 is concerned, it is set by changing fermentations input factors through the model in (5).

According to the criterion in (3), we generated D-optimal designs with different number of runs (from 29 to 33), with $\eta_{+1}^2 = 0.316^2$, $\eta_0^2 = 0.305^2$, and $\eta_{-1}^2 = 0.321^2$. In this case, we select the design with $d_{0.8,\,0.95} = 1.49$ which corresponds to the design with 33 runs. In Figure 5, we report the corresponding FDS plot. The black curve is related to the *UPV* values

TABLE 4 Step 2: Capture separation

(A) Input factors of Step 2 and ranges			
	Input	***Range***	
$x_1^{(2)}$	*pH.W*	[6.80, 7.20]	set directly
$x_2^{(2)}$	*Mol.L*	[70.00, 80.00]	set directly
$x_3^{(2)}$	*pH.L*	[6.80, 7.20]	set directly
$y^{(1)}$	*Pur*1	[13.32, 17.75]	set indirectly

(B) Output of Step 2, coded levels in brackets						
Output		***Input***				
Level	***Pur2***	***Mol.L***		***Pur1***		η_u^2
low	72.90	80.00	(+1.00)	12.27	(−1.00)	3.30^2
medium	84.48	78.00	(+0.06)	15.14	(+0.05)	3.19^2
high	96.06	74.00	(−0.20)	18.31	(+1.00)	3.39^2

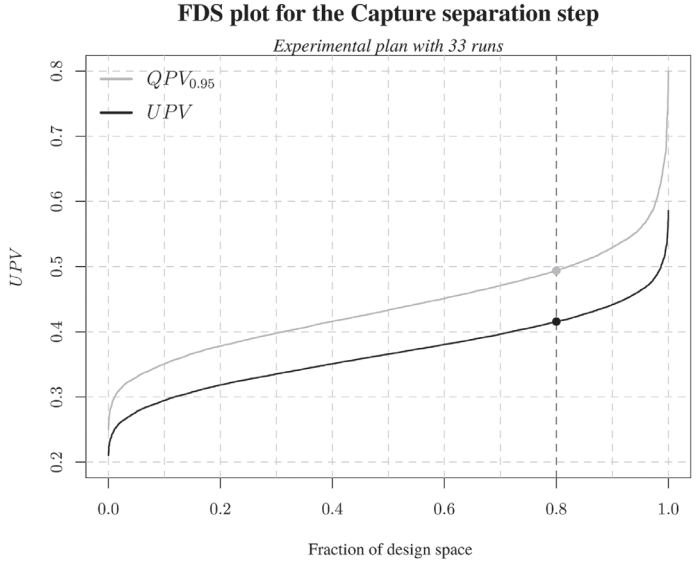


FIGURE 5 Fraction of design space plot for the 33 runs design of the capture separation step

obtained in the ideal case of no error in setting *Pur*1, while the gray curve refers $QPV_{0.95}$ (Section 3.1.1). For the chosen design, the *UPV* value for a fraction of the design space $\gamma = 0.8$ (gray dot) indicates that the 95% of the possible realizations of this design have the 80% of the design space with $UPV \leq 0.49$. Whereas, if ideally we could run the experiment by setting precisely all input factors, the 80% of the design space would have had $UPV \leq 0.42$ (black dot).

The generated design provides target levels for *pH.W*, *Mol.L*, *pH.L*, and *Pur*1. In order to proceed with the experiment, we rearrange the experimental plan in terms of the directly controllable input factors: target levels for *Pur*1 become target levels for OD_{ind}, *pH*, and *DE*. The 33 runs are independently performed on the fermentation and the capture steps. Clearly, the *Pur*1 values achieved in the fermentations will be different from the target values due to the model error (e.g., 13.36 is just prediction given by the model, experimental values are expected within $13.36 \pm e_u$). From the results (reported in Table B2), we estimate the following model (in terms of coded unit) for the *Pur*2 outcome:

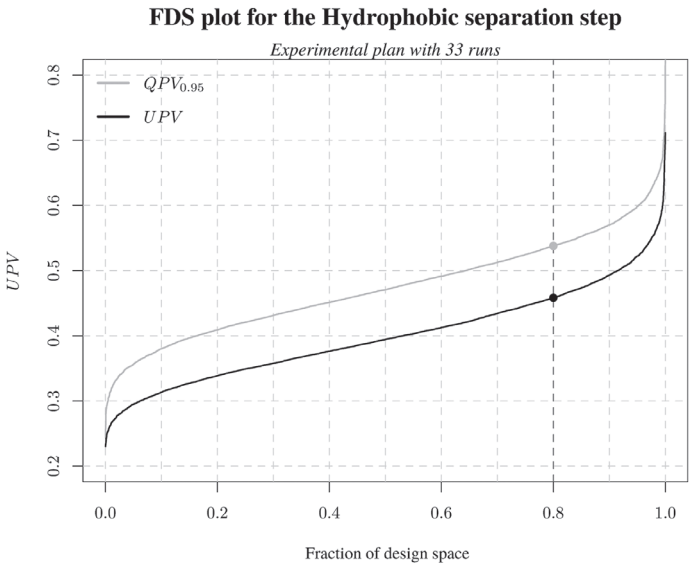
$$\hat{Pur}2 = 87.80 - 1.91 \cdot Mol.L + 8.08 \cdot Pur1 - 4.91 \cdot Mol.L^2, \tag{6}$$

with $\hat{\sigma}_2^2 = 3.1^2$ and $R^2 = 0.80$ (predictive $R^2 = 0.75$). Following the same procedure of the previous step, by using the model in (6), we derived the combination of *Mol.L* and *Pur*1 such that *Pur*2 reaches its minimum and maximum (see Table 4(B)). Within this range, we selected three target levels, low, medium, and high for *Pur*2 and we compute the corresponding variance of the error for a future prediction by (2).

TABLE 5 Step 3: Hydrophobic separation; input factors and ranges

Input		Range	
$x_1^{(3)}$	*pH.C*	[6.40, 7.00]	set directly
$x_2^{(3)}$	*dens*	[2.00, 8.00]	set directly
$x_3^{(3)}$	*cond*	[80.00, 110.00]	set directly
$y^{(2)}$	*Pur*2	[92.90, 96.06]	set indirectly

FIGURE 6 Fraction of design space plot for the 33 runs design of the hydrophobic separation step



STEP 3: Hydrophobic separation

The input factors of the hydrophobic separation step are reported in Table 5 together with the selected ranges. As regards *Pur*2, it can be controlled by tuning the inputs of the capture and fermentation steps according to the models in (5) and (6).

Assuming a tentative model with quadratic terms and two factor interactions, we proceeded as for step 2: we rescaled the input factors to the range $[-1; 1]$ and consequently the variance of the error in setting *Pur*2 (obtaining $\eta_{+1}^2 = 0.29^2$, $\eta_0^2 = 0.27^2$, and $\eta_{-1}^2 = 0.28^2$). Among D-optimal designs in average, with 30-33 runs, by comparing FDS plots and the expected half-width values, we selected the experimental plan with 33 runs. For this plan, the expected half-width of the mean predicted from the model is below 1.54 in 80% of the design space. For the same proportion of the design space, the 95% of the experimental plan will have $UPV \leq 0.54$, whereas $UPV \leq 0.46$ if the experiment would have been run with no error (FDS plot in Figure 6).

The experimental plan is then rearranged in order to express the target values for *Pur*2 as target values for *Mol.L* and *Pur*1. In turn, target levels for *Pur*1 are translated into target levels for OD_{ind}, *pH*, and *DE*. The values obtained experimentally for *Pur*1 and *Pur*2 have been measured and recorded (data reported in Table B3) and have been used to fit the following model (in terms of coded units) for *Pur*3,

$$\hat{Pur}3 = 88.81 + 0.77 \cdot pH.C + 1.58 \cdot dens + 4.38 \cdot Pur2 + 3.75 \cdot dens^2 + 1.76 \cdot pH.C \cdot Pur2, \quad (7)$$

with $\hat{\sigma}_3^2 = 2.01^2$ and $R^2 = 0.80$ (predictive $R^2 = 0.69$).

Since *Pur*3 is a CQA of the final outcome we now proceed to the manufacturing design space definition. For this three-step process, three interconnected models are available, one for *Pur*1, one for *Pur*2, and one for *Pur*3.

5.2 | Multi-step manufacturing design space determination

In this study, the manufacturing design space is the region of the input factors' space which can consistently guarantee a final protein purity at or above 88% (often called specification limits in this kind of studies). Therefore, by the model

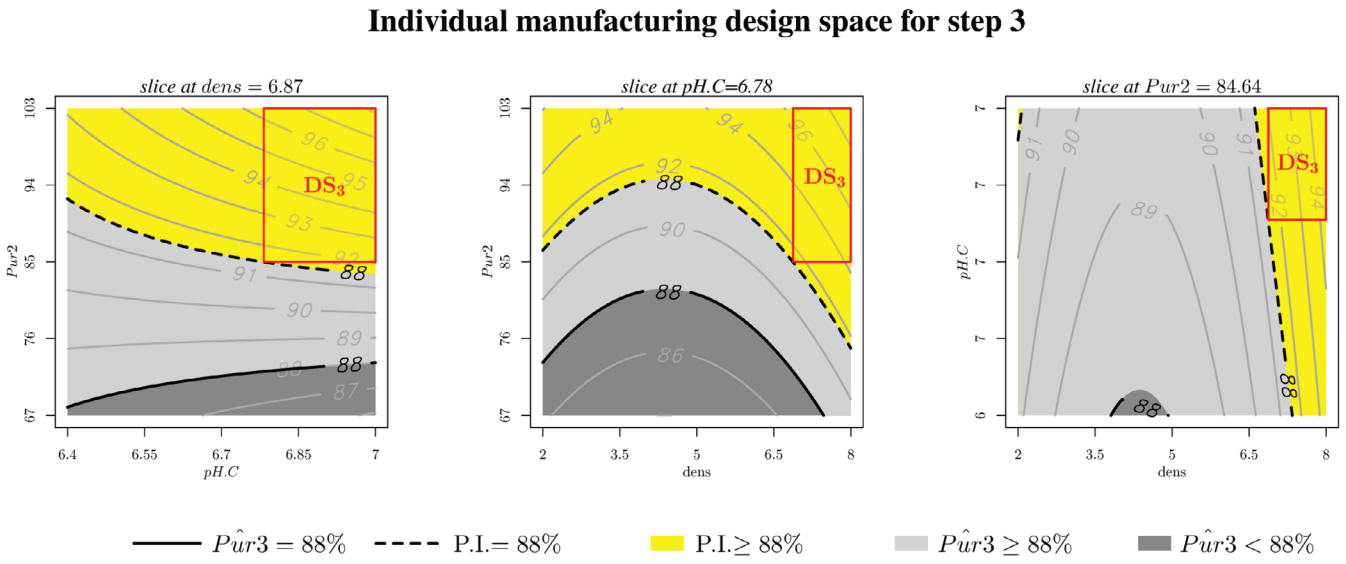


FIGURE 7 Manufacturing design space for the hydrophobic separation step

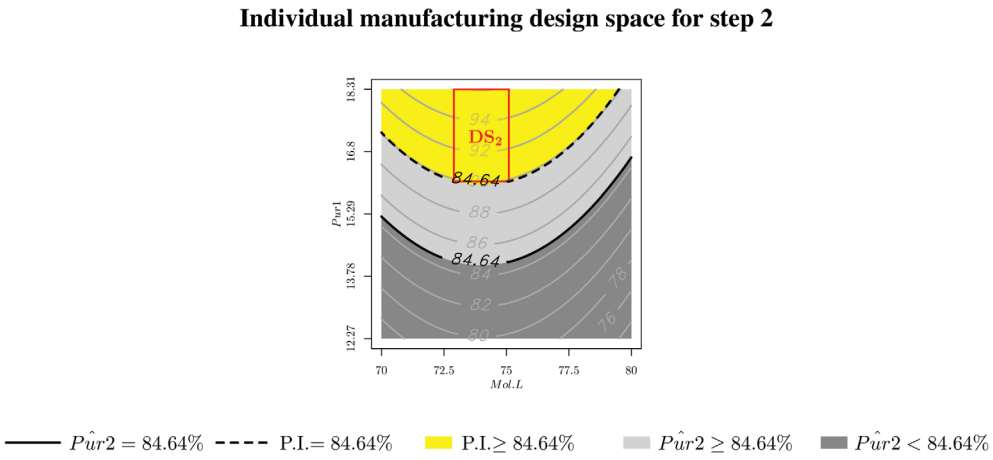


FIGURE 8 Manufacturing design space for the capture separation step

in (7) and by graphical optimization we selected the region of *pH.C*, *dens*, and *Pur*2 such that $\hat{Pur}3 \geq 88\%$. To take into account the uncertainty on the model predictions, this region is typically reduced following operative considerations. More specifically we considered the region restricted by the 95% one-sided prediction interval (P.I.) on a single future observation of *Pur*3, as shown in Figure 7, but other intervals can be considered as well, like, for example, tolerance intervals.[20] In Figure 7, the solid black curve is the model prediction such that $\hat{Pur}3 = 88\%$, while the dashed black curve is the corresponding bound given by the prediction interval equal to 88% We highlight in yellow the region of the input factors such that specification limits are satisfied. In this case, within this region, we identified a suitable sub-region of operating conditions (see Remark 3)—the red-bordered rectangles—so that the manufacturing design space is defined by the linear combination of the significant input factors for the hydrophobic step and the outcome of the capture step as follows,

$$DS_3 = \{dens \in [6.87, 8.00], pH.C \in [6.82, 7.00]\ Pur2 \geq 84.64\%\}.$$

Now, $Pur2 \geq 84.64\%$ becomes the specification limit for the capture step: the resulting manufacturing design space is identified by the suitable region delimited by the one-sided 95% prediction interval on *Pur*2, as shown in Figure 8 and it can be defined as

$$DS_2 = \{Mol.L \in [72.90, 75.10],\ Pur1 \geq 16.02\%\}.$$

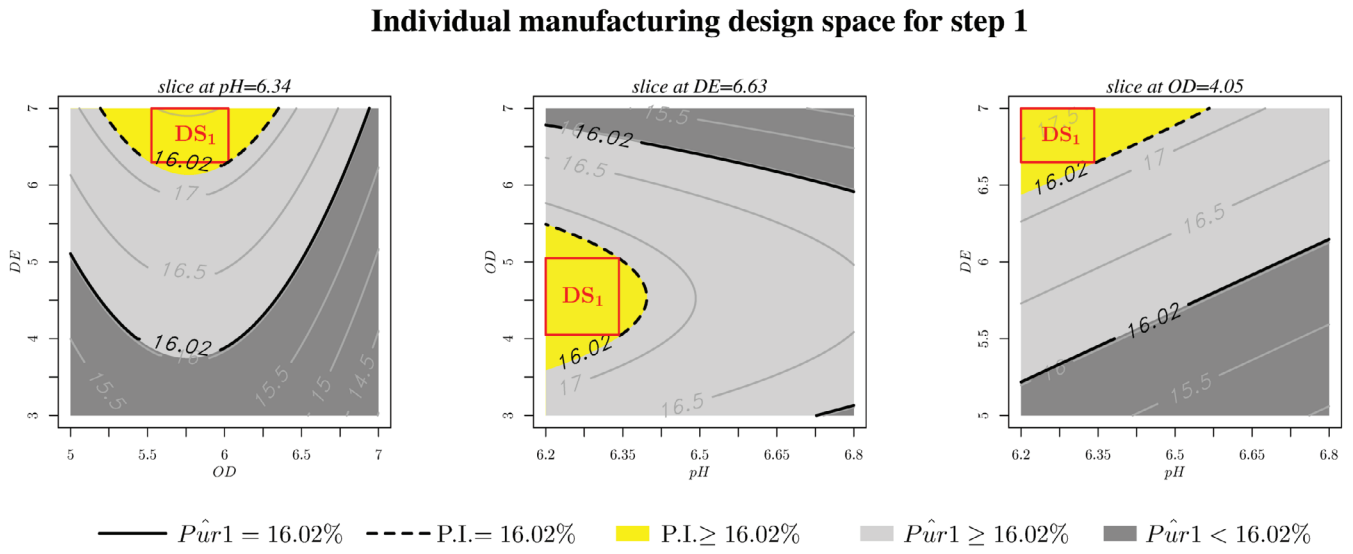


FIGURE 9 Manufacturing design space for the fermentation step

The same procedure is repeated for the fermentation step (see Figure 9), whose manufacturing design space is given by

$$DS_1 = \{OD \in [4.05, 5.05],\ pH \in [6.63, 7.00],\ DE \in [3.00, 6.05]\}.$$

The multi-step manufacturing design space is then derived by combining the individual ones of each process step, obtaining

$$DS = \{dens \in [6.87, 8.00], pH.C \in [6.82, 7.00], Mol \in [72.90, 75.10],$$
$$OD_{ind} \in [4.05, 5.05],\ pH \in [6.63, 7.00],\ DE \in [3.00, 6.05]\}.$$

6 | DISCUSSION AND CONCLUSIONS

In this article, we provide a statistical approach to define the manufacturing design space of a process composed of multiple steps. Since the multi-step manufacturing design space does consider the interactions among subsequent steps, it leads to a very good process understanding, it guarantees quality and safety of products and faster and more consistent product development. In the context of pharmaceutical industry, these aspects consistently guarantee a drug product with the desired properties leading to benefits for patients. Moreover, the multi-step manufacturing design space increases both manufacturing flexibility and process robustness, which are crucial for reducing costs and batch discarding. Indeed, in the first place, working within the manufacturing design space is not considered as a change, whereas movements outside would normally initiate a regulatory post approval. In the second place, it is well-known that a manufacturing design space that spans the entire process can increase the operational flexibility.[2] The experimental effort required by the procedure, although moderate, is still affordable by scientists, making our proposal a valid compromise between process knowledge and experimental resources. Despite our motivating set-up is related to pharmaceutical processes, our methodology is general and it can be considered for multi-step experimentation in various industrial fields.

In the three-steps fermentation and purification process considered, three experiments for three steps are required: overall, 10 input factors were involved in the study. If we would have approached this process as a big single-stage, the experiment would have had 10 input factors, resulting in 66 parameters to estimate (in the case of a full-quadratic model for the final outcome). In principle, the D-optimal design with 100 runs could have also been a possible solution to design the experiment. However, by considering the process as a big single-stage, each of the 100 runs would have involved all the three process' steps. According to our proposal only the last set of experiments, namely, 33 runs, has to be performed

on the whole process. Therefore, the multi-step design strategy requires a limited number of runs for the experiments carried out from the first up to the last step.

The multi-step approach relies on the assumption that the outcomes of the intermediate steps can be measured to derive intermediate models. The quality of these models depends on many elements and, of course, on the features of the step itself. The impact of an input factor set with error on the intermediate models, and so on the model for the final outcome, strongly depends on how many input factors has the step in view of one factor set with error and on the effect (linear, quadratic, etc.) that the input factor set with error is expected to have on the outcome, that is, on the tentative model.[12] Finally, the quality of an intermediate model is also affected by the observed realization of the planned design matrix. We would recommend a case-by-case simulation study to evaluate this impact. We would also advise to evaluate the prediction properties of the design by taking into account the possible deviations from the target levels that could occur in setting the input factor with error, as suggested in Section 3.1.1. This prevents an overestimation of the prediction properties of the design.

Notice that the design strategy proposed here can be also applied to contexts in which, at the end one step (or more) it is not possible to stop and observe the outcome. In these cases, such step can be simply merged with the subsequent one and considered as a single-stage in the design of experiment and in the manufacturing design space.

The design strategy presented in this article is based on planning one experimental design for each step including, as additional input factor, the output of the previous one. This setting requires the selection of a single outcome/CQA of the previous step that the experimenter is interested to study, in interactions with the input factors of the current one. This assumption is essential in practice: an input—which is actually an output of the previous step, say y_A—is set by running the previous step and by adjusting its inputs to achieve the desired values for y_A given in the experimental plan. In the case the experimental design would include two set indirectly inputs—say y_A and y_B—as they are controlled by changing the same inputs of the previous step, they cannot be, in general, set independently to the desired level. Essentially, if the levels of y_A are appropriately changed in the experiment, the levels of y_B can only be observed. Thus, at the end of each step, multiple CQAs can be still measured and monitored, but the experimental plan will only be optimal for the selected one. In many industrial fields, scientific knowledge and discussion with process' experts should help to identify the appropriate CQA to be included as an input for the next step. Otherwise, the two outcomes could be treated separately into two different experiments but the procedure would employ a quite large amount of experimental resources. Further research is surely requested on this point.

In addition, the multi-step framework provides hints for future research in many directions. We focused on processes in which the behavior of the outcome of each step can be well approximated by linear models in the parameters and we consider only the interactions among subsequent steps. This framework encompasses several practical situations, however, the procedure could be appropriately extended, with increasing complexity, to relax these assumptions. Moreover, since first principle models may eventually exist just for one of the steps that makes up the process, one of the main direction for future studies is the extension of our proposal to accommodate both mechanistic and empirical models. As regards the design criterion, in our framework, it should take into account that one input factors is not set directly. As a starting point we adopted the average D-optimality proposed in the literature but this problem offers insights for further research to derive alternatives design criteria suitable for our set-up. Future developments will be also dedicated to the optimization of the total number of experimental runs to be performed on each step of the process.

ACKNOWLEDGEMENTS

We wish to thank Alessandro Pieri and Eva Grassi of GSK TRD Drug Substance team, who provided a very useful support to this work. This work was sponsored by GlaxoSmithKline Biologicals SA. We are grateful to the referees for their helpful comments, which led to a substantially improved version of the article.

CONFLICT OF INTERESTS

The authors have declared the following conflict of interests: Rosamarie Frieri is a PhD student at the University of Bologna and participates in a post graduate studentship program at GSK, with Marilena Paludi as her supervisor. Marilena Paludi and Marco Mariti are employees of the GSK group of companies.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the Appendix of this article.

ORCID

Rosamarie Frieri https://orcid.org/0000-0003-2707-183X

REFERENCES

1. Shi J, Zhou S. Quality control and improvement for multistage systems: a survey. *IIE Trans*. 2009;41(9):744-753.
2. ICH guideline Q8 (R2) on pharmaceutical development. U.S. Department of Health and Human Services, Food and Drug Administration; November 2009.
3. Yu L. Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm Res*. 2008;25:781-791.
4. Jiang C, Flansburg L, Ghose S, Jorjorian P, Shukla AA. Defining process design space for a hydrophobic interaction chromatography (HIC) purification step: application of quality by design (QbD) principles. *Biotechnol Bioeng*. 2010;107(6):985-997.
5. Tyssedal J, Kulahci M. Experiments for multi-stage processes. *Qual Technol Quant Manag*. 2015;12(1):13-28.
6. Bjørkestøl K, Sivertsen E, Næs T. Optimization of two-step batch processes and the method of compensation for random error. *J Chemom*. 2012;26(6):311-321.
7. Beshah S, Desta G. Multi-Stage and multi-response process optimization in Taguchi method. *JEEA*. 2015;33:1-12.
8. Moslemi A, Seyyed-Esfahani M. Robustness indices in multistage problems. *Qual Reliab Eng*. 2017;33:1211-1224.
9. Yuangyai C, Lin DK. Understanding multistage experiments. *Int IJEDPO*. 2013;3(4):384-409.
10. Kulahci M, Tyssedal J. Split-plot designs for multistage experimentation. *J Appl Stat*. 2017;44(3):493-510.
11. Pronzato L. Information matrices with random regressors. application to experimental design. *J Stat Plan Inference*. 2002;108:189-200.
12. Donev A. Design of experiments in the presence of errors in factor levels. *J Stat Plan Inference*. 2004;126(2):569-585.
13. Montgomery DC. *Design and Analysis of Experiments*. New York, NY: John Wiley & Sons; 2008.
14. Box GEP. The effects of errors in the factor levels and experimental design. *Technometrics*. 1963;5(2):247-262.
15. Jensen WA. Open problems and issues in optimal design. *Qual Eng*. 2018;30(4):583-593.
16. Silvey S. *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Amsterdam, Netherlands: Springer; 1980.
17. Fedorov V. *Theory of Optimal Experiments*. New York, NY: Academic Press; 1972.
18. Zahran A, Anderson-Cook CM, Myers RH. Fraction of design space to assess prediction capability of response surface designs. *J Qual Technol*. 2003;35(4):377-386.
19. Eon-Duval A, Valax P, Solacroup T, et al. Application of the quality by design approach to the drug substance manufacturing process of an Fc fusion protein: towards a global multi-step design space. *J Pharm Sci*. 2012;101(10):3604-3618.
20. Krishnamoorthy K, Mathew T. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons; 2008.
21. Miller AJ, Nguyen NK, Algorithm AS. 295: a fedorov exchange algorithm for D-optimal design. *J R Stat Soc Ser C Appl Stat*. 1994;43(4):669-677.

How to cite this article: Frieri R, Mariti M, Paludi M. Design of experiments and manufacturing design space for multi-step processes. *Appl Stochastic Models Bus Ind*. 2021;37:942–960. https://doi.org/10.1002/asmb.2620

APPENDIX A. D-OPTIMAL DESIGN IN AVERAGE

In exchange algorithms, starting from an initial design, each design point is considered for exchange with each of the point of a candidate list. The selected pair of points to exchange is the one which gives the best improvement in terms of the chosen design criterion (in this case is the pair which most decreases the expected determinant). This procedure is iterated until no further improvement in the criterion can be obtained by a pairwise exchange.[21] The algorithm finds local optimum so the procedure is usually repeated for multiple initial designs.

Notation:

design points: $\tilde{\mathbf{z}}_i$ for $i = 1, \ldots, n$
candidate points: $\tilde{\mathbf{z}}_j$ for $i = 1, \ldots, N$
$\delta(\tilde{\mathbf{z}}_i) = \tilde{\mathbf{z}}_i'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\mathbf{z}}_i$
$\delta^2(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) = \mathbf{z}_i'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\mathbf{z}}_j$
$\Delta_{ij} = \delta(\tilde{\mathbf{z}}_j) - \delta(\tilde{\mathbf{z}}_i) - \delta(\tilde{\mathbf{z}}_i)\delta(\tilde{\mathbf{z}}_j) + \delta^2(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j)$
$D_{ij} = [\det(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}) + (1 + \Delta_{ij})]^{-1}$, namely, the change in $\det(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}$ which would be obtained by switching $\tilde{\mathbf{z}}_i$ with $\tilde{\mathbf{z}}_j$
t: small threshold (e.g., 10^{-10})

nsim: number of Monte Carlo simulations. To select *nsim* we run a series of preliminary simulation studies in which we check the distributions of D_{ij} and their expectations for increasing *nsim*. In our case study example *nsim*=1000 (as also in Donev[12]) was enough to provide reasonable stable results.

```
generate a (random) start design with n points
while  −(D_new − D_old) ≤ t {
    for 1, … , nsim {
      generate a realization* of Z_old → Z̃_old
      generate a realization* of the N candidate points
      compute D_old = det(Z̃'_old Z̃_old)^−1
        for i in 1, … , n {
          for j in 1, … , N {
            compute Δ_ij
            compute D_ij
        } }
    }
compute 𝔼[D_ij]
    select i and j minimizing 𝔼[D_ij] (if more than one best exchange, select one randomly)
    exchange z_i with z_j so that Z_old ⟶ Z_new
    update the determinant D_new = det(Z'_new Z_new)^−1
}
```

*a random draw from normal distribution (with zero mean and variance in (2)) is added to each target level of the input factor set with error in the main effect column of $\boldsymbol{Z}_{old}$; quadratic and two factor interactions effects are then computed accordingly.

APPENDIX B. EXPERIMENTAL DATA

See Tables B1 to B3.

TABLE B1 Experimental results from the fermentation step

run	*Pur*1	*TE*	OD_{ind}	*pH*	*DE*
1	16.20	1.00	−1.00	1.00	1.00
2	15.80	0.00	0.00	0.00	0.00
3	16.10	1.00	1.00	−1.00	1.00
4	16.80	0.00	0.00	0.00	0.00
5	14.80	1.00	−1.00	−1.00	−1.00
6	17.50	0.00	0.00	0.00	0.00
7	15.10	−0.50	−0.50	−0.67	−0.50
8	16.60	−1.00	−1.00	−1.00	1.00
9	12.55	1.00	1.00	1.00	−1.00
10	15.45	−1.00	1.00	1.00	1.00
11	13.10	−1.00	1.00	−1.00	−1.00
12	14.50	−1.00	−1.00	1.00	−1.00
13	16.10	−1.00	−1.00	−1.00	−1.00
14	16.70	0.00	0.00	0.00	0.00
15	15.00	1.00	1.00	1.00	1.00
16	15.30	1.00	1.00	1.00	1.00
17	16.03	0.50	0.50	0.67	0.50
18	16.70	0.00	0.00	0.00	0.00
19	17.80	1.00	−1.00	−1.00	1.00

TABLE B1 (Continued)

run	*Pur*1	*TE*	OD_{ind}	*pH*	*DE*
20	15.30	1.00	1.00	−1.00	−1.00
21	13.70	−1.00	1.00	1.00	−1.00
22	15.40	−1.00	−1.00	1.00	1.00
23	14.90	1.00	−1.00	1.00	−1.00
24	17.20	−1.00	0.00	0.00	0.00
25	15.00	0.00	0.00	1.00	0.00
26	16.00	0.00	0.00	−1.00	0.00
27	16.50	0.00	0.00	0.00	0.00
28	16.10	0.00	0.00	0.00	0.00
29	17.60	0.00	0.00	0.00	1.00
30	16.20	−0.50	0.50	−1.00	0.50
31	16.50	1.00	0.00	0.00	0.00
32	14.90	0.00	1.00	0.00	0.00
33	14.80	0.00	0.00	0.00	−1.00
34	15.70	0.00	−1.00	0.00	0.00

TABLE B2 Experimental results from the capture separation step

run	*Pur*2	*pH.W*	*Mol.L*	*pH.L*	*Pur*1
1	74.86	−1.00	1.00	1.00	−0.35
2	89.50	−1.00	1.00	1.00	0.67
3	80.65	−1.00	0.00	−1.00	−0.86
4	90.51	1.00	0.00	−1.00	0.61
5	87.95	0.00	0.00	1.00	−0.13
6	96.20	0.00	−1.00	−1.00	0.87
7	80.42	1.00	−1.00	−1.00	−0.60
8	91.94	1.00	−1.00	0.00	0.62
9	89.72	−1.00	−1.00	−1.00	0.88
10	77.44	0.00	1.00	−1.00	−0.11
11	79.59	1.00	−1.00	1.00	−0.11
12	90.10	−1.00	−1.00	0.00	0.80
13	91.95	−1.00	1.00	−1.00	0.64
14	95.41	0.00	0.00	0.00	0.64
15	75.04	1.00	1.00	1.00	0.09
16	85.45	1.00	0.00	0.00	−0.79
17	85.11	−1.00	−1.00	1.00	−0.05
18	80.47	0.00	−1.00	1.00	−1.00
19	75.30	1.00	1.00	−1.00	−0.52
20	78.76	1.00	0.00	1.00	−0.81
21	88.63	1.00	1.00	1.00	0.66
22	82.42	1.00	−1.00	−1.00	−0.07
23	90.66	1.00	−1.00	1.00	0.93
24	88.27	1.00	1.00	−1.00	0.93
25	78.79	−1.00	1.00	1.00	−0.89
26	81.14	−1.00	1.00	0.00	−0.14
27	77.55	0.00	−1.00	0.00	−0.93
28	89.40	−1.00	0.00	1.00	0.54
29	80.41	−1.00	1.00	−1.00	−0.18
30	82.22	−1.00	−1.00	1.00	−0.46
31	91.46	0.00	−1.00	1.00	1.00
32	70.64	1.00	1.00	1.00	−0.77
33	79.55	−1.00	−1.00	−1.00	−0.61

TABLE B3 Experimental results from the hydrophobic separation step

run	*Pur3*	*pH.C*	*cond*	*dens*	*Pur2*
1	86.10	−1.00	1.00	−1.00	−0.80
2	91.23	1.00	1.00	1.00	−0.60
3	91.18	−1.00	−1.00	0.00	0.60
4	89.47	1.00	1.00	−1.00	−0.30
5	97.89	0.00	−1.00	1.00	0.70
6	91.98	−1.00	−1.00	1.00	−0.90
7	96.63	1.00	−1.00	−1.00	0.70
8	93.02	−1.00	−1.00	−1.00	1.00
9	98.47	1.00	1.00	1.00	0.40
10	91.18	0.00	1.00	1.00	−0.70
11	95.68	1.00	−1.00	−1.00	0.70
12	90.43	−1.00	1.00	1.00	−0.10
13	90.08	−1.00	−1.00	−1.00	−0.40
14	90.64	−1.00	0.00	−1.00	0.10
15	88.61	0.00	1.00	0.00	0.70
16	95.72	−1.00	1.00	−1.00	1.00
17	97.11	0.00	−1.00	1.00	0.50
18	86.20	0.00	0.00	0.00	−1.00
19	96.68	−1.00	1.00	1.00	0.30
20	85.62	1.00	1.00	−1.00	−0.90
21	85.80	−1.00	1.00	0.00	−0.60
22	84.39	1.00	0.00	0.00	−0.40
23	90.25	0.00	−1.00	−1.00	0.00
24	100.65	1.00	−1.00	1.00	0.40
25	90.95	−1.00	0.00	1.00	−0.90
26	88.49	−1.00	−1.00	0.00	−0.20
27	97.38	1.00	0.00	1.00	0.60
28	91.58	−1.00	0.00	1.00	0.90
29	95.61	1.00	1.00	0.00	0.60
30	90.14	1.00	−1.00	−1.00	−0.70
31	89.79	1.00	−1.00	1.00	−0.60
32	92.57	1.00	1.00	−1.00	0.60
33	91.64	−1.00	1.00	−1.00	0.70