Biotechnology Journal
www.biotechnology-journal.com

Biotechnol. J. 2017, 12, 1700174

DOI 10.1002/biot.201700174

Research Article

# Simplex-based optimization of numerical and categorical inputs in early bioprocess development: Case studies in HT chromatography

*Spyridon Konstantinidis, Nigel Titchener-Hooker and Ajoy Velayudhan*

The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, London, UK

Bioprocess development studies often involve the investigation of numerical and categorical inputs via the adoption of Design of Experiments (DoE) techniques. An attractive alternative is the deployment of a grid compatible Simplex variant which has been shown to yield optima rapidly and consistently. In this work, the method is combined with dummy variables and it is deployed in three case studies wherein spaces are comprised of both categorical and numerical inputs, a situation intractable by traditional Simplex methods. The first study employs in silico data and lays out the dummy variable methodology. The latter two employ experimental data from chromatography based studies performed with the filter-plate and miniature column High Throughput (HT) techniques. The solute of interest in the former case study was a monoclonal antibody whereas the latter dealt with the separation of a binary system of model proteins. The implemented approach prevented the stranding of the Simplex method at local optima, due to the arbitrary handling of the categorical inputs, and allowed for the concurrent optimization of numerical and categorical, multilevel and/or dichotomous, inputs. The deployment of the Simplex method, combined with dummy variables, was therefore entirely successful in identifying and characterizing global optima in all three case studies. The Simplex-based method was further shown to be of equivalent efficiency to a DoE-based approach, represented here by D-Optimal designs. Such an approach failed, however, to both capture trends and identify optima, and led to poor operating conditions. It is suggested that the Simplex-variant is suited to development activities involving numerical and categorical inputs in early bioprocess development.

Received 01 MAR 2017
Revised 04 MAY 2017
Accepted 11 MAY 2017
Accepted article online 12 MAY 2017

**Supporting information available online**

**Keywords:** Categorical variables · Chromatography · Filter plates · High throughput bioprocess development · RoboColumns · Simplex optimization

## 1 Introduction

Early stage bioprocess development studies are typically taking place through the adoption of an approach based on Design of Experiments (DoE) [1, 2]. Such studies, frequently aim to investigate ranges of the input variables, and to identify process relevant operating conditions to be explored in detail by further development activities. A part of these studies is also the scouting of categorical variables, or inputs. These may include culture media (e.g. [3]), or represent a choice of buffer species and chromatographic media (e.g. [4–6]). An alternative approach for performing scouting studies in early stages of bioprocess development is a grid-compatible Simplex variant [7]. Adoption of this Simplex method has been found to offer an attractive alternative to a DoE based approach for multiple case studies [7–9]. Benefits of this approach include its independence from the requirement of fitting mathematical models, the ability to deal with sparse missing values and noise levels typical of those encountered

**Correspondence:** Ajoy Velayudhan, The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, Bernard Katz Building, Gordon Street, London WC1H 0AH, UK
**Email:** a.velayudhan@ucl.ac.uk

**Abbreviations: CV,** column volume; **Cyt,** cytochrome; **DoE,** design of experiments; **FF,** fast flow; **HT,** high throughput; **mAb,** monoclonal antibody; **OA,** optical absorbance

The copyright line for this article was changed on 21st December 2017 after original online publication.

in High Throughput (HT) studies, along with the ease of experimentation due to its compatibility with non-uniform grids via hand-coding. These are, of course, combined with its ability to locate consistently optima with an accompanying efficiency comparable to approaches employing DoE techniques.

The results of a scouting study, also including categorical inputs, can potentially have a significant impact on the outcomes of future development studies. Here, a choice regarding the categorical inputs may be made and any erroneously derived conclusions will be propagated in the process development train. For example, making the wrong choice of a chromatographic resin for the purification of a product can render the performance of the optimized process suboptimal since an alternative resin could outperform the chosen one. The importance of such studies and the likelihood of a DoE approach not always reaching correct conclusions, especially when complex spaces are investigated alongside wide input ranges, have led to the consideration of the Simplex method for the concurrent optimization of numerical and categorical inputs defined over coarse grids.

Here, we assess the performance of the grid-compatible Simplex method when deployed in case studies where the inputs also involve categorical variables. Three case studies are employed to describe the application of the method and to demonstrate its results. The first is based on in silico data and serves to demonstrate an approach involving the definition of dummy variables for dealing with the categorical inputs. The latter two were developed by carrying out HT studies employing, microscale, batch and packed bed column chromatography techniques, respectively. In the former case, binding conditions were screened so as to identify those maximizing the binding of a monoclonal antibody (mAb) to a selection of chromatographic media. In the latter case, a linear gradient elution method was developed for the purification of a model system of proteins, comprised of bovine serum albumin (BSA) and cytochrome *c* (CytC), and the Simplex method aims to identify the condition leading to the maximization of an objective function including both *Throughput* and *Purity*. The case studies were also investigated via the deployment of D-Optimal designs so as to represent an approach based on DoE methodology. This study sets out to expand the applicability domain of the grid-compatible Simplex method and represents a novel application of Simplex-based methods in general.

## 2 Materials and methods

### 2.1 Gridded Simplex method

The Simplex algorithm is a direct search method which is based on the geometrical shape of a simplex [10, 11]. While the conventional Simplex methods accept continuous inputs, a variant has been developed to be compatible with discrete inputs [8]. This grid-compatible variant employs additional movements, to those of the traditional method, and rules so as to search gridded spaces, typical of those generated during HT applications in bioprocessing, while overcoming challenges relating to simplex degeneracy and oscillation. These render the method capable of identifying rapidly favorable areas within search spaces, which is then followed by their characterization through the encirclement of optima. These features stem from the nature of the modified method being a hybrid of the variable and fixed size Simplex methods [10, 11].

The application of this method in development studies leads to an iterative procedure wherein the method indicates conditions to be tested and uses the results to propose a new set of conditions until termination. Here, it is deployed retrospectively to identify optimal conditions on an already analyzed grid of conditions. This allows for the detailed assessment of the method and at the same time it simulates its intended iterative deployment.

### 2.2 Case studies

The first case study is an in silico devised scenario which aims to highlight the challenges that may be met in situations wherein numerical and categorical inputs are investigated concurrently and how they can be dealt with via the deployment of the Simplex method. The remaining two employ real experimental data from chromatography based studies. Their details are given in the next two sections. The chemicals and proteins used in these case studies were from Sigma–Aldrich Ltd (Dorset, UK) unless specified otherwise.

#### *2.2.1 Case study 1: In silico data*

Here, the gridded data were generated employing a sphere function (Eq. 1). The first two variables ($x_1$, $x_2$) are numerical in nature and span thirteen uniform levels in [–3, 3]. The inclusion of the third variable ($x_3$) in Eq. (1) aimed to simulate the effect of a categorical input with four levels (i.e. *A*–*D*). Hence, Eq. (1) takes a total of 676 values within the considered grid (i.e. 169 values for each level of $x_3$) and this function is minimized through the deployment of the Simplex method.

$$f(x)=\begin{cases}\left(x_1-1\right)^2+x_2^2+10, & x_3=A\\ \left(x_1-1\right)^2+x_2^2+20, & x_3=B\\ \left(x_1-1\right)^2+x_2^2+30, & x_3=C\\ \left(x_1-1\right)^2+x_2^2+40, & x_3=D\end{cases} \tag{1}$$

#### *2.2.2 Case study 2: Screening mAb binding conditions and resins*

The target product in this case study was a recombinant mAb expressed in CHO cells. Cell culture supernatant was spun at 35 000 × *g* for 40 min in an Avanti® J-E centrifuge (Beckman Coulter Inc., CA, USA). The resultant supernatant was diafiltrated into 5 mM Tris pH 7.0 using Vivaspin™ 20 MWCO 10000 tubes (GE Healthcare, Uppsala, Sweden) to a mAb concentration of 1.5 mg $mL^{-1}$ as measured by a protein G based HPLC method described in [12].

#### 2.2.2.1 High throughput mAb CEX binding system

A Tecan Freedom Evo® 100 station, equipped with a 8-channel liquid handling arm, 1 mL dilutors, and controlled by Freedom EVOware® version 2.1 software (Tecan Group Ltd, Männedorf, Switzerland) was employed to conduct all liquid handling steps using disposable 1 mL tips (Molecular BioProducts Inc., California, USA). Two 96-well PreDictor™ CIEX screening filter plates (GE Healthcare) were employed to assess the impact of four variables on the binding of mAb in duplicate. Since these plates contain three resins (i.e. Capto™ S, SP Sepharose™ Fast Flow, and Capto MMC, all from GE Healthcare), each occupying 32 wells in a plate at volumes of 2, 6 and 6 µL per well, respectively, one of these variables was set to be *resin type*; a categorical variable with three levels (i.e. *A*–*C*, respectively). The next two variables were pH and salt concentration in a 20 mM acetate binding buffer. These were both numerical with four levels each (i.e. 4.00, 4.25, 4.50, 4.75, and 20, 40, 60, 80 mM, respectively). The fourth variable was the *salt type* used to achieve the aforementioned salt concentrations. Two salt species were chosen (i.e. NaCl and $Na_2SO_4$) and as such this variable was also categorical in nature with two levels (i.e. *A*, *B,* respectively). Hence, 32 binding buffers were tested for each of the three resins, yielding a total of 96 conditions in the resulting grid. These buffers were prepared in a 96-well deep square well plate (Fisher Scientific, Loughborough, UK) by mixing stock solutions of acetic acid, sodium acetate (VWR International Ltd, Leicestershire, UK), sodium chloride and sodium sulfate at the desired amounts, and were used to equilibrate two PreDictor plates according to the manufacturer's recommendations (i.e. experiments were run in duplicate). Additional amounts of the buffers were employed to prepare 32 feed solutions for loading the plates. For this purpose, the clarified and diafiltrated cell culture supernatant was diluted two-fold in each of the 32 buffers in a separate 96-well deep square well plate (Fisher Scientific). These feed solutions were then aliquoted to load the two 96-well PreDictor filter plates at different volumes so as to present each resin with a load challenge of 50 mg $mL^{-1}$. This challenge was chosen based on previous studies exploring the saturation of the resins at different binding conditions (data not shown). Upon completion of the loading step, the filter plates were then incubated for a period of one hour while shaking on two orbital shakers at 1100 rpm (Eppendorf UK Ltd., Stevenage, UK). At the end of the incubation period the plates were evacuated via centrifugation on an Avanti J-E centrifuge (Beckman Coulter Inc.), according to the instructions of the manufacturer, and the filtrates were collected in two new 96-well deep square well collection plates (Fisher Scientific). These flowthrough fractions were stored at 4°C before analysis using the HPLC assay described in [12]. All experiments were performed at room temperature. The determined concentrations of the analyzed fractions were employed in defining an objective function to be investigated by the deployment of the Simplex method.

### 2.2.3 *Case study 3: Optimization of binary mixture separation on cation exchange resins*

The binary mixture used in this study comprised of BSA (>96%, agarose gel electrophoresis) and CytC from bovine heart (>95%). The employed buffers were prepared using sodium acetate, acetic acid, and sodium chloride (VWR International Ltd). The miniature columns were PreDictor RoboColumns® (GE Healthcare) pre-packed with 600 µL of the cation exchange resins Capto S and Capto SP ImpRes (both from GE Healthcare). The RoboColumns were operated on a Tecan Freedom Evo 200 (Tecan Group Ltd) liquid handling station fitted with 1 mL dilutors and stainless steel fixed tips. The robotic station was equipped with an 8-channel liquid handling arm and a robot manipulator arm and was connected to an Inifinte® Pro 200 reader (Tecan Group Ltd). The hardware necessary to deploy the RoboColumns (i.e. Te-Chrom, Te-Shuttle, 96-well array plate from Tecan Group Ltd) was accompanied by hotels and carriers used to store solutions and plates. Such plates included 96-well deep square well plates (Fisher Scientific), 48-well deep square well plates (Elkay Ltd., Hampshire, UK) and 96-well UV transparent plates (Corning Lifesciences, MA, USA). The Tecan station was programmed through Freedom EVOware v2.6 (Tecan Group Ltd).

#### 2.2.3.1 High Throughput RoboColumn chromatography

For the complete automated deployment of RoboColumn chromatography on the Tecan station, a script was coded which provided a framework in which custom compiled applications convert user-defined inputs into worklists. Here, eight RoboColumns were run in parallel in each experiment and their operation followed closely the workflow reported in [13]. A liquid class with a dispense flowrate of 5 µL $s^{-1}$ was employed when dispensing into the columns giving a residence time of 2 min. The fixed stainless steel tips were sanitized, when necessary, with a protocol involving their washing with system liquid and aspiration/dispense cycles with a 0.5 M sodium hydroxide solution.

#### 2.2.3.2 Analytical methods

The analysis of the collected fractions and blanks involved the determination of the volumes of the fractions and of the concentration of BSA and CytC in each collected fraction. The former was achieved by employing the method described in [14]. Upon the determination of its volume, the height of a solution in a well (i.e. the pathlength of the absorbance measurement in cm) could also be determined and it was employed to normalize all absorbance measurements (i.e. OA cm$^{-1}$).

For the determination of the BSA and CytC amounts in the collected fractions, a method employing dual wavelength measurements was devised. Here, single component calibration curves were prepared for each of the analytes and by solving numerically a system of two equations with two unknowns (Eq. 2 and 3) their concentrations in a mixture were determined.

$$OA_{\lambda_1}/b = \alpha_{\lambda_1,0} + \alpha_{\lambda_1,1}[BSA] + \alpha_{\lambda_1,2}[BSA]^2 + \beta_{\lambda_1,0} + \beta_{\lambda_1,1}[CytC] + \beta_{\lambda_1,2}[CytC]^2 \tag{2}$$

$$OA_{\lambda_2}/b = \alpha_{\lambda_1,0} + \alpha_{\lambda_1,1}[BSA] + \alpha_{\lambda_2,2}[BSA]^2 + \beta_{\lambda_2,0} + \beta_{\lambda_2,1}[CytC] + \beta_{\lambda_2,2}[CytC]^2 \tag{3}$$

where $\lambda_1$ and $\lambda_2$ (nm) are two wavelengths at which absorbance measurements are taken, $OA_{\lambda_1}$ and $OA_{\lambda_2}$ are the blank corrected absorbances of a mixture of the two analytes at $\lambda_1$ and $\lambda_2$, $b$ is the pathlength (cm), and finally, $[BSA]$ and $[CytC]$ are the concentrations (mg mL$^{-1}$) of BSA and CytC, respectively, in a mixture. The coefficients $\alpha$ and $\beta$ were those obtained from the single component quadratic calibration curves for BSA and CytC, respectively, at 280 nm ($\lambda_1$) and 530 nm ($\lambda_2$). These were estimated by preparing standards in 20 mM acetate, pH 4.5, 20 mM NaCl over a concentration range of 0.05 mg mL$^{-1}$ to 10 mg mL$^{-1}$ for each protein (data not shown). The consideration of quadratic curves and the measurements at 530 nm were implemented to extend the working range of the assay. The method was validated by quantifying the two proteins in prepared unknown samples prior to its deployment (data not shown).

### 2.2.3.3 Linear salt gradient elution studies

The RoboColumn set up was employed in bind and elute mode to study the impact of four variables on the salt gradient elution separation of BSA (80% w/w) and CytC (20% w/w) mixtures. All runs were performed employing 20 mM acetate buffers, pH 4.5 at various sodium chloride concentrations. This particular pH was chosen since it was found to result in similar retention of both solutes on the employed cation exchange resins making their separation, via salt gradients, challenging and representative of real mixtures (data not shown). The first three considered variables were the starting sodium chloride concentration of the gradient ($C_s$), the slope of the gradient ranging from $C_s$ to 1 M of NaCl ($S$), and the column load ($L$). The fourth variable was *resin type* and was categorical in nature with two levels (i.e. $A$, $B$ for Capto SP ImpRes and Capto S, respectively). Four levels were chosen for each of the three numerical variables (i.e. 20, 60, 100 and 140 mM for $C_s$; 70, 120, 170 and 220 mM CV$^{-1}$ for $S$; and 5, 10, 15 and 20 mg mL$^{-1}$ for $L$) and their combination yielded 64 test conditions per resin. For each of these conditions and resin, two chromatograms were generated (i.e. experiments were performed in duplicate) based on a method which included the handling of the columns according to: (i) equilibration for five column volumes (CVs) with a pH 4.5, 20 mM acetate buffer at the desired $C_s$ level; (ii) wash with a pH 4.5, 20 mM acetate buffer at the desired $C_s$ level for 5 and 4 CVs for the Capto S and Capto SP ImpRes resins, respectively; (iii) strip with a pH 4.5, 20 mM acetate 1 M NaCl buffer for 4 and 2 CVs for the Capto S and Capto SP ImpRes resins, respectively; (iv) regeneration with a 1 M NaOH, 0.5 M NaCl solution for 5 and 3 CVs for the Capto S and Capto SP ImpRes resins, respectively; and finally (v) the flushing of the columns with a 20% ethanol solution for 3 CVs. For the runs employing the Capto S resin, the columns were loaded for 2 CVs and the desired loads were achieved by manipulating the concentration of the feed solutions. Conversely, for the runs employing the Capto SP ImpRes resin, the feed solution had a constant concentration (i.e. 10 mg mL$^{-1}$) and the desired loads were achieved by manipulating the loaded volume. The duration of the elution gradient in a run was determined by the specified $C_s$ and $S$. The feed solutions were prepared by dissolving the required amounts of the proteins in the corresponding equilibration (and wash) buffer. The RoboColumns were flushed with system liquid prior to the initialization of a run and prior to the flushing of the columns with the storage solution. The nominal volumes of the fractions collected by these runs were determined automatically by the deployed script and were at least of 150 µL in volume.

The total of 256 runs (i.e. 64 test conditions per resin assessed in duplicate) was completed in 32 experiments wherein eight RoboColumns were run in parallel. Upon their completion, the application of the analytics described in Section 2.2.3.2 allowed for the calculation of volume balances (%VBs) and mass balances (%MBs), via Eq. (4) and (5), respectively. Furthermore, Eq. (6)–(8) were also applied to calculate the *Purity*, *Yield* and *Throughput* associated with a product pool per tested condition.

$$\%VB = 100 \times \frac{\sum_i^I V_{measured,i}}{\sum_i^I V_{nominal,i}} \tag{4}$$

$$\%MB = 100 \times \frac{\sum_i^I Mass_i}{Mass\ in} \tag{5}$$

$$Yield_n = 100 \times \frac{\sum_i^I BSA\ mass_i}{Mass\ in\ BSA} \tag{6}$$

$$Purity_n = 100 \times \frac{\sum_i^I BSA\ mass_i}{\sum_i^I BSA\ mass_i + \sum_i^I Cyt\ C\ mass_i} \tag{7}$$

$$Throughput_n = \frac{\sum_i^I BSA\ mass_i}{120\left[CV_I - \left(CV_{Load} + CV_{Wash}\right)\right]} \tag{8}$$

where $n$ is the $n^{th}$ pool out of the possible $N$ product pools starting and ending at fraction numbers $i$ and $I$, respectively. In Eq. (4), $V_{\text{measured,i}}$ and $V_{\text{nominal,i}}$ correspond to the measured and nominal fraction volumes, respectively, in the $i^{th}$ fraction. In Eq. (5), $Mass_i$ corresponds to the calculated mass of a protein in the $i^{th}$ fraction (i.e. protein concentration in fraction multiplied by $V_{\text{measured,i}}$) and *Mass in* is the amount of a protein loaded to a column. In Eq. (4) and (5) the summation is over the total number of fractions in a run. In Eq. (8), the denominator of the *Throughput* (mg $s^{-1}$) was calculated as the number of column volumes at the end point of the $n^{th}$ pool ($CV_I$) offset by the duration of the load and wash stages in a given run (i.e. $CV_{\text{Load}}$ and $CV_{\text{Wash}}$, respectively). This was multiplied by 120 since the residence time was set to 2 min for all runs. Eq. (6)–(8) were further employed to obtain an objective function to be optimized by the deployment of the Simplex method.

## 2.3 Deployment of the grid compatible Simplex method

### *2.3.1 Pre-processing*

Prior to the deployment of the Simplex method, replicated measurements were averaged for each test condition for Case studies 2 and 3 so as to reduce experimental noise. Missing data were assigned highly unfavorable objective function values as a surrogate. The levels of the different variables in an optimization problem were hand-coded into ordered integers which increased monotonically with the levels of the variables. This was also applied in the case of categorical variables for the numerical assignment of their levels so as to include them in the method. For example, in Case study 1, the *A–D* levels of the third variable were hand-coded to levels 1–4 respectively, whereas in Case study 2 the *A–C* levels of the *resin type* input were hand-coded to levels 1–3. Alternative permutations of the possible level assignments for the categorical inputs were also assessed.

### *2.3.2 Objective function definition*

In Case study 1, the Simplex method was deployed to minimize the function described by Eq. (1). Likewise, in Case study 2, the Simplex method was employed to minimize the concentration of the mAb in the flowthrough fractions (i.e. $C_{\text{mAb in flowthrough}}$). In both case studies, the minimization of the objective functions was achieved by maximizing the negative objective functions since the coded Simplex method acted as a maximizer. The objective function employed in Case study 3 was a composite function consisting of the product of *Throughput* (Eq. 8) and *Purity* (Eq. 7) (i.e. *Throughput* × *Purity*). Prior to assigning an objective function value to each of the tested conditions, a screening process of the $N$ pools per condition took place so as to identify the best product pool. In this process, Eq. (6)–(8) were applied to all $N$ pools per condition and the product *Throughput* × *Purity* was calculated. For those pools where the calculated *Purity* was less than 85%, the product was set to zero. When none of the pools satisfied this constraint on *Purity*, then for such a condition, the objective function was assigned a highly unfavorable value as a surrogate, similar to the treatment of missing values. Otherwise, the pool leading to the highest product over all $N$ pools was selected and the objective function value for this condition was set to be equal to that product. The Simplex method was then deployed to maximize the composite objective function *Throughput* × *Purity*.

### *2.3.3 Evaluation of performance of Simplex method*

To assess the performance of the Simplex method in the considered case studies, a population of results was generated by deploying the method from 150 randomly defined initial simplices (or starting points) in each case study. These did not include an optimum as one of the vertices of an initial simplex. Each of these 150 searches was continued until the termination of the method and upon their completion the reached optimum was recorded along with its associated objective function value and the number of conditions evaluated. These results were employed in determining the success of the method in identifying optimal conditions and its efficiency. Visual tools, including histograms and mesh plots, were also employed for displaying the trends in the data from the case studies.

Finally, a brief comparison is made between the Simplex-based approach and an approach employing response surface methodology. For this purpose, the data from the case studies were fitted by employing regression models with their model matrix originating from D-Optimal designs [15, 16]. The fitted models included up to second order terms. Such designs were employed as they allow for a flexible selection of the calibration set and the inclusion of categorical variables.

The Simplex method was encoded in MATLAB 2015a (The MathWorks® Incorporated, MA, USA) on a dual Intel Xeon E5-2650 CPU workstation with 32 GB of RAM running Windows Server 2003 (Microsoft Corporation, WA, USA). The application of the method with the Parallel Computing Toolbox (MathWorks) allowed for the deployment of 30 simplex searches in a parallel fashion, with each lasting, typically, well below 1 min in real time. The system of linear equations (Eq. 2 and 3) was solved numerically employing MATLAB's 'fsolve' function. D-Optimal designs were built through the PLS Toolbox (v7.9.5) (Eigenvector Research Inc., WA, USA).

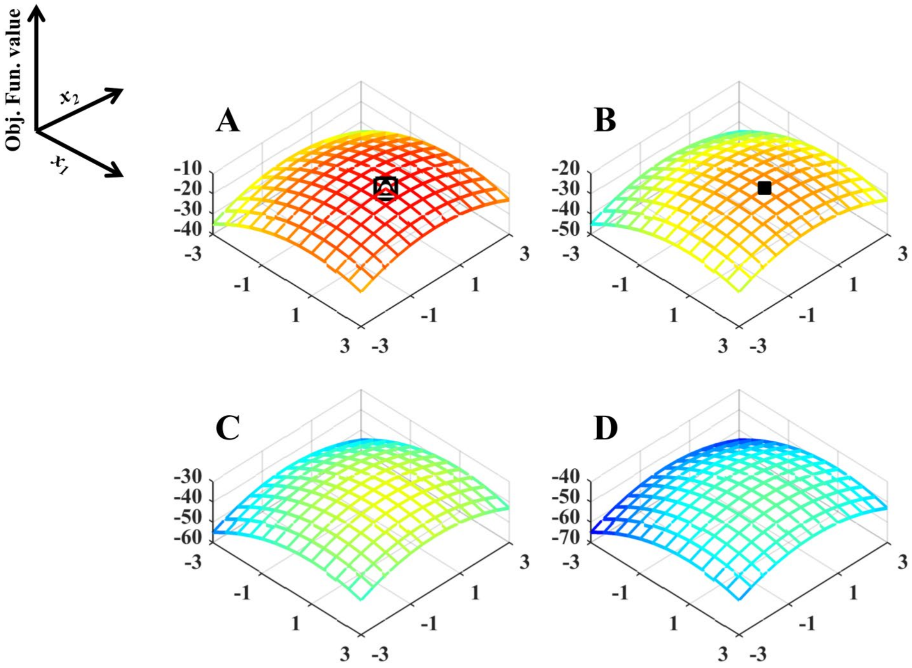


**Figure 1.** Mesh plot of the objective function value from Eq. (1) in Case study 1. Each of (**A**)–(**D**) corresponds to each of the four levels (i.e. *A–D*) of the categorical input $x_3$. (□) annotates the global optimum. (○) annotates the optimum identified by 100, 100, 72, 61.33 and 96% of the simplex searches for permutations #1–#5, respectively; (■) annotates the optimum identified by the remaining simplex searches for permutations, #3–#5 respectively; All searches converged to (○) when the dummy variable approach was employed; (▲) corresponds to the predicted optimum based on the estimations of a quadratic model calibrated on a D-Optimal design with a sample size of 77.

## 3 Results and discussion

### 3.1 Simplex method deployment

#### *3.1.1 Case study 1*

Having preprocessed the in silico data, which included the numerical assignment of the levels of $x_3$ (i.e. *A–D* assigned values of 1–4 respectively), the Simplex method was deployed from 150 randomly generated starting points to identify the combination of the three inputs leading to the minimum of the objective function (Eq. 1). For this, first, permutation, the resulting response surface (Figs. 1A–D) contained a single and well defined optimum (i.e. (□) in Fig. 1A). As a result, each search converged to the condition (1, 0, *A* or 1) (i.e. (○) in Fig. 1A) leading to a 100% success rate (permutation #1 in Table 1). This is due to the fact that with the chosen assignment of the levels of $x_3$, a monotonic relationship is imposed between the levels of the categorical input and the objective function value (Supporting information, Fig. S1A). This monotonicity was also maintained when the levels of $x_3$ are assigned values 1–4 in reverse order (i.e. *D–A*, respectively). Here, the underlying response surface also has a single and well defined optimum, but at a different coded level of the categorical input (i.e. (1, 0, *A* or 4), and the deployment of the method from an additional 150 random starting points led to the identification of this optimum with a success rate of 100% (permutation #2 in Table 1).

In both of these cases, the described treatment of the categorical input leads to a simplification of the optimiza-

**Table 1.** Success rate of Simplex method in locating the global optimum in Case studies 1–3. Nine permutations are investigated for assigning values to the levels of the categorical inputs $x_3$ (*A–D*) in Case study 1 and *resin type* in Case studies 2 and 3 (*A–C* and *A–B*, respectively). Results are based on 150 searches initializing from random points.

| Case study | Permutation | Levels | | | | % Success rate |
|---|---|---|---|---|---|---|
| | | *A* | *B* | *C* | *D* | |
| 1 | 1 | 1 | 2 | 3 | 4 | 100.00 |
| | 2 | 4 | 3 | 2 | 1 | 100.00 |
| | 3 | 1 | 4 | 3 | 2 | 72.00 |
| | 4 | 4 | 1 | 2 | 3 | 61.33 |
| | 5 | 3 | 1 | 4 | 2 | 96.00 |
| 2 | 6 | 1 | 2 | 3 | NA | 100.00 |
| | 7 | 1 | 3 | 2 | NA | 92.67 |
| | 8 | 2 | 1 | 3 | NA | 63.33 |
| 3 | 9 | 1 | 2 | NA | NA | 100.00 |

tion problem due to the monotonic trend. However, for a categorical variable with four levels, a total of 4! permutations can be assessed for the numerical assignment of the levels. For an additional three of these permutations (i.e. permutations #3–#5 in Table 1), the monotonicity was disrupted and additional spurious optima emerged along the categorical input. For permutations #3 and #4, the two optima were on either end of the categorical input (Supporting information, Figs. S1B and S1C, respectively) whereas for permutation #5 the two optima were one level apart (Fig. S1D). For these permutations, the Simplex method was less successful in identifying the optimum (Table 1). Here, searches that did not converge at the global optimum were instead stranded at a local optimum (i.e. (■) in Fig. 1B). The greater success rate of the method for permutation #5 is due to the fact that in this case the two optima lay close to each-other (Fig. S1D) and were less well separated compared to permutations #3 and #4. As a result, the encirclement movements that the method carries out once an optimum is identified led to a higher chance of locating the global optimum once the local optimum had been reached for permutation #5.

These results highlight the impact of the treatment of the categorical inputs in an arbitrary fashion on the method; additional, spurious, optima may emerge and as a result the identification of optimal conditions can become more challenging. To counter this challenge, an alternative approach was adopted for dealing with the categorical inputs and it was based on the definition of dummy variables.

#### 3.1.1.1 Categorical input consideration via dummy variables

Dummy variables are adopted in regression analysis when dealing with categorical inputs [17]. Here, a categorical input with $\zeta$ levels is converted to $\zeta$-1 dummy binary variables ($d_{\zeta-1}$). For example, a categorical input with three levels (i.e. $A$–$C$) is converted according to: $A \rightarrow (0,0)$, $B \rightarrow (1,0)$, and $C \rightarrow (0,1)$. Hence, the input space is expanded by one dimension since the categorical input is replaced by three coordinates of ($d_1$, $d_2$). Such coding enforces the comparison of one reference level against the remaining ones without assuming the existence of an order in the levels of the categorical variable.

The impact of this approach, for dealing with categorical inputs, on the performance of the Simplex method was considered by deploying an additional 150 randomly initialized searches in Case study 1. For this purpose, three dummy variables were defined ($d_1$, $d_2$, $d_3$) and the four levels of $x_3$ were assigned as follows: $A \rightarrow (0,0,0)$, $B$ (1,0,0), $C \rightarrow (0,1,0)$, and $D \rightarrow (0,0,1)$. As a result the 3D grid was converted to a 5D grid. Since the Simplex method is applicable to grids, and the definition of the three dummy variables led to $2^3$ combinations of their levels, the existing grid was complemented by the missing coordinates in this new 5D grid. For these conditions, the objective function was set to a highly unfavorable value as a surrogate. Hence, a 5D grid was generated with 13 × 13 × 2 × 2 × 2 = 1352 conditions from which half were treated as missing values and which included a single optimum at the coordinate (1, 0, 0, 0, 0) (i.e. (1, 0, $A$) in the original 3D grid). With the inclusion of the dummy variables, the Simplex method reached the optimum with a success rate of 100% which was as expected since in this scenario no false optima emerged from the handling of the categorical input $x_3$.

Such a result serves to demonstrate the applicability of an approach employing dummy variables in dealing with categorical inputs during the deployment of the Simplex method and the ability of the latter to locate optima consistently in spaces involving such inputs.

### 3.1.2 *Case study 2*

The second case study aimed to investigate the conditions leading to the strongest binding of a mAb on three resins as a function of the pH, salt concentration ([Salt]) and *salt type* (Fig. 2). For this purpose, the duplicated concentrations of the mAb in the 96 flowthrough fractions were averaged for assessing trends and for deploying the Simplex method, as described in Section 2.3.1. Coefficients of variation were predominately below ≈5% with a few exceptions reaching to ≈10%.

Within the spanned ranges of pH and [Salt], stronger binding was observed for resin Capto S (Figs. 2A and 2D) than for resins SP Sepharose FF (Figs. 2B and 2E) and Capto MMC (Figs. 2C and 2F) for both of the considered salt species, i.e. NaCl (Figs. 2A–C) and $Na_2SO_4$ (Figs. 2D–F). For both salts, conditions of low [Salt] and pH led to a reduced binding of the mAb to the Capto S resin (Figs. 2A and 2D). This indicates a binding mechanism similar to the one observed in [18, 19]. This trend is more pronounced for $Na_2SO_4$ (Fig. 2D) since the response surface has a ridge along the pH/[Salt] plane with the strongest binding occurring at high pH (4.75) and low [Salt] (20 mM) (i.e. (□) in Fig. 2D). This is consistent with the higher ionic strength of this salt compared to NaCl. For the latter salt species, the binding trends on Capto S were also complex since two optima occur; one at an intermediate pH and high [Salt], and a second one at a high pH and low [Salt] (i.e. (♦) and (▼) in Fig. 2A, respectively). These binding trends were weakly emulated by the second cation exchange resin, SP Sepharose FF (Figs. 2B and 2E) whereas the third considered resin, Capto MMC, a multimodal chromatography resin, exhibited different binding trends even if only the ionic interactions were active in the employed conditions (Figs. 2C and 2F). Hence, the results of this case study represent a complex system wherein there is a clear trend regarding the impact of the categorical input *resin type*, whereas the trends regarding the remaining three inputs are less apparent; there exist multiple optima along the pH and salt concentration plane and one salt does not clearly lead to better or worse binding compared to the other.

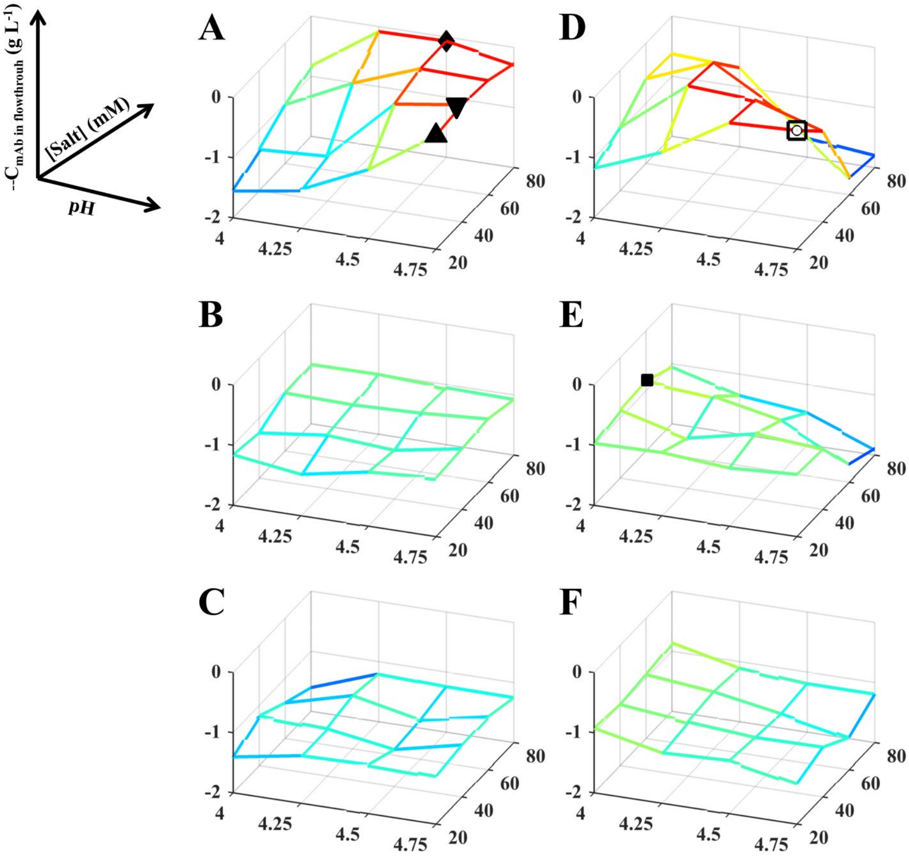


**Figure 2.** Mesh plot of the negative average mAb concentrations in the flowthrough fractions ($C_{\text{mAb flowthrough}}$) in Case study 2. Each of (**A**)–(**F**) corresponds to a combination of the categorical inputs *salt type* and *resin type*: (**A**) NaCl salt and Capto S resin; (**B**) NaCl salt and SP Sepharose FF resin; (**C**) NaCl salt and Capto MMC resin; (**D**) $Na_2SO_4$ salt and Capto S resin; (**E**) $Na_2SO_4$ salt and SP Sepharose FF resin; (**F**) $Na_2SO_4$ salt and Capto MMC resin. (□) annotates the global optimum; (♦) annotates the optimum identified by 36.67% of the simplex searches for permutation #8 whereas (▼) annotates a local optimum for the *salt type* and *resin type* combination in (**A**); (○) annotates the optimum identified by 100, 92.67, and 63.33% of the simplex searches for permutations #6–#8, respectively; (■) annotates the optimum identified by the remaining simplex searches for permutation #7; All searches converged to (○) when the dummy variable approach was employed; (▲) corresponds to the predicted optimum based on the estimations of a quadratic model calibrated on a D-Optimal design with a sample size of 66.

The Simplex method was deployed to identify the combination of the two numerical (pH and [Salt]) and two categorical (*salt type* and *resin type*) inputs minimizing the average mAb concentration in the flowthrough fractions (i.e. $C_{\text{mAb in flowthrough}}$). The impact of the treatment of the *resin type* input was also assessed by adopting three different permutations for the assignment of values to its three levels, *A*–*C*, corresponding to Capto S, SP Sepharose FF and Capto MMC, respectively (permutations #6–#8 in Table 1). This was not implemented for the *salt type* input since only two levels were present (i.e. NaCl or *A* and $Na_2SO_4$ or *B*) and it was treated as a dichotomous categorical input (i.e. for this input, levels *A* and *B* were assigned values of 1 and 2 respectively). For this purpose, 150 searches were deployed from random starting points for the three permutations considered in this case study.

The results in Table 1 for Case study 2 demonstrated further the importance of the treatment of the categorical inputs for the performance of the Simplex method since its success rate varied from ≈63% to 100%. High success rates in locating the global optimum (i.e. (□) in Fig. 2D) were returned for those cases wherein the first level of the categorical input *resin type* (i.e. *A* or Capto S) was placed on the boundary of the searched space (i.e. permutations #6 and #7). In the first case, all searches reached to a condition coinciding with the global optimum (i.e. (○) in Fig. 2D) whereas in the second case, ≈7% of the simplex searches became stranded at a local optimum (i.e. (■) in Fig. 2E) for which the *resin type* input was at level *B* (or SP Sepharose FF). The remaining searches also converged to the global optimum (i.e. (○) in Fig. 2D). By contrast, the lowest success rate corresponded to the permutation placing the first level of the categorical input *resin type* within the boundaries of the searched space (i.e. permutation #8 in Table 1). Here, ≈63% of the searches converged to the same condition as the global optimum whereas the remaining ones converged to one of the best two binding conditions observed for Capto S and NaCl (i.e. (○) and (♦) in Figs. 2D and 2A, respectively).

The difference in the ability of the Simplex method to converge to the global optimum for permutations #6 and #7 (Table 1) was attributed to the loose similarity of the binding trends between Capto S and SP Sepharose FF and the dissimilarity of their trends to those observed for Capto MMC. For permutation #7, starting a search from an initial simplex spanning only the Capto MMC and SP Sepharose FF resins, and in particular with most of its vertices lying on the pH/[Salt] plane of the SP Sepharose FF resin for either salt, prevented the method from following a gradient leading to the pH/[Salt] plane of the Capto S resin; the uniformity of the binding trends was disrupted by the intermediate level of the *resin type* input being Capto MMC in permutation #7. This is similar to the behavior in Case study 1 where the monotonicity of the objective function values was disrupted depending on the coding of the levels of the categorical input (e.g. Supporting information, Figs. S1B and S1C).

The higher success rates in locating the global optimum for permutations #6 and #7 compared to permutation #8 (Table 1) was attributed to the fact that in the former two, the global optimum lies at the boundary of the search space and there exists a clear improvement in the objective function as the level of the *resin type* input changes from its intermediate value (i.e. 2) to its lower limit (i.e. 1). Hence, the method follows a gradient that guides the deployed searches towards the boundary of the space which then leads to a detailed sampling of the corresponding subspace. This behavior is described by Fig. S2 (Supporting information); the first and second bars indicate that the majority of the vertices of the formed simplices, across all 150 searches, lie on levels *A*, *B* and *A*, *C* for permutations #6 and #7, respectively. Likewise, Figs. S3A and S4A show the conditions selected by two searches converging to the global optimum for permutations #6 and #7, respectively. Here, the majority of the evaluated conditions in the two grids lie on the first two assigned levels (i.e. 1 and 2) of the input *resin type*. Conversely, in the case of permutation #8, the simplex searches distributed the evaluated conditions more uniformly across the three levels of the *resin type* input (i.e. third bar in Fig. S2 and S4B) since here the optimum lies within the boundaries of the searched space along the input *resin type*.

The adoption of the dummy variable approach in this case study also resulted in a uniform distribution of the evaluated conditions across the levels of the *resin type* input (Supporting information, fourth bar in Fig. S2 and S3B). Here, the three levels of the *resin type* input were converted into two dummy variables according to *A* (or Capto S) → (0,0), *B*

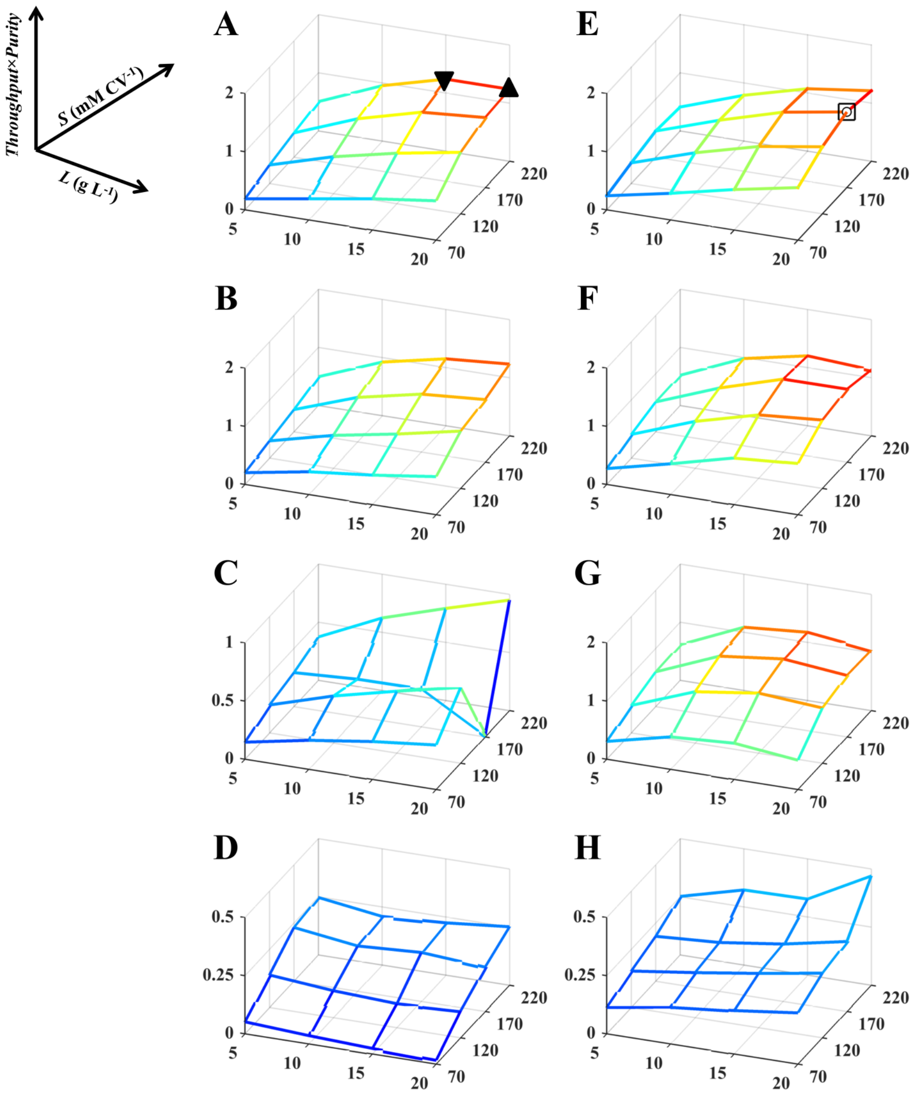


**Figure 3.** Mesh plot of the average composite objective function *Throughput* × *Purity* in Case study 3. Each of (**A**)–(**H**) corresponds to a combination of the inputs $C_s$ and *resin type*: (**A**) $C_s$ of 20 mM and Capto SP ImpRes resin; (**B**) $C_s$ of 60 mM and Capto SP ImpRes resin; (**C**) $C_s$ of 100 mM and Capto SP ImpRes resin; (**D**) $C_s$ of 140 mM and Capto SP ImpRes resin; (**E**) $C_s$ of 20 mM and Capto S resin; (**F**) $C_s$ of 60 mM and Capto S resin; (**G**) $C_s$ of 100 mM and Capto S resin; (**H**) $C_s$ of 140 mM and Capto S resin. (□) annotates the global optimum; (○) annotates the optimum identified by 100% of the simplex searches for permutation #9; (▼) annotates the optimum for the $C_s$ and *resin type* combination in (**A**); (▲) corresponds to the predicted optimum based on the estimations of a quadratic model calibrated on a D-Optimal design with a sample size of 42. The low objective function values in (**C**) for a slope, *S*, of 170 mM CV$^{-1}$ was due to a gross experimental error in the preparation of the employed buffers.

(or SP Sepharose FF) → (1,0) and *C* (or Capto MMC) → (0,1). Similar to Case study 1, the existing 4D grid in Case study 2 was consequently expanded to a 5D grid and it was complemented by the addition of 32 surrogate points to ensure its completeness. While the distribution of the evaluated conditions is similar to the one observed for permutation #8, the incorporation of the two dummy variables in the Simplex method led to a 100% success rate since all searches converged to the same condition as the global optimum (i.e. (○) in Fig. 2D). Hence, as was the case for the in silico data in Case study 1, this approach was also found to be highly successful in this case study, which employed real data, while eliminating the impact of the numerical assignment of the levels of the *resin type* input.

### 3.1.3 Case study 3

The last study investigated the separation of a binary mixture of model proteins (BSA and CytC) with the RoboColumn microscale HT chromatography technique. A total of 128 chromatograms were developed for each of the two resins (i.e. 64 conditions in duplicate for resins Capto S and Capto SP ImpRes). The calculated volume balances (Eq. 4) across all runs, per resin, were found to be on average 95.3 ± 2.0% and 96 ± 2.9% for resins Capto S and SP ImpRes, respectively. Good closure was also observed, on average, for the mass balances (Eq. 5) for both BSA (i.e. 88.7 ± 5.1 and 90.8 ± 5.9% for Capto S and SP ImpRes, respectively) and CytC (i.e. 98.6 ± 7.5 and 96.3 ± 4.9% for Capto S and SP ImpRes, respectively). The better mass balance closures for CytC compared to BSA were attributed to the better sensitivity of the dual wavelength method (Section 2.2.3.2) for the former solute due to its specific absorbance at 530 nm.

Upon the application of Eq. (7) and (8), and the screening of the candidate product pools as described in Section 2.3.2, the composite objective function *Throughput* × *Purity* was calculated (subject to a purity constraint of 85%) for each condition (Fig. 3). The objective function values for each of the two resins appear to be comparable (i.e. Figs. 3A–D and 3E–H for Capto SP ImpRes and Capto S, respectively) with the Capto S resin delivering the highest objective function value (i.e. (□) in Fig. 3E) at a low initial salt concentration (i.e. 20 mM in $C_s$), intermediate gradient slope (i.e. 170 mM $CV^{-1}$ in $S$) and high load (i.e. 20 g $L^{-1}$ in $L$) (Fig. 4B). The condition returning the highest *Throughput* × *Purity* value for the Capto SP ImpRes resin (i.e. (▼) in Fig. 3A and 4A) differed from the aforementioned one since here intermediate and high values were needed for $L$ (i.e. 15 g $L^{-1}$) and $S$ (i.e. 220 mM $CV^{-1}$) respectively, whereas the starting salt concentration was the same (i.e. $C_s$ at 20 mM). Since the product pools associated with the two conditions, as depicted in the denominator of Eq. (8), were approximately of equal size (i.e. 3.50 and ≈3.52 CVs in Figs. 4A and 4B, respectively; a zoomed in version of Fig. 4 and the product pools can be found in Supporting information, Fig. S5) the differences in the achieved *Throughput* were attributed to the BSA amount present in the product pool (i.e. numerator in Eq. 8). This amount was part of a trade-off affecting the value of the composite objective function; for Capto SP ImpRes, a high purity was preferred over a high BSA amount whereas for Capto S the opposite was true. This trade-off is expressed through the relationship between the load ($L$) and the gradient slope ($S$) since the response surface in Fig. 3 gave evidence of both a significant interaction between the two inputs and a quadratic trend for either input.

Despite the existence of these two optima and the non-linear trends in the response surface, the deployment of the Simplex method in this case study led to a 100% success rate since all simplex searches converged to a grid point occupied by the global optimum (i.e. (○) in Fig. 3E). Here, the categorical variable, *resin type*, had two levels ($A$ and $B$), similar to the categorical variable *salt type* in Case study 2, and was therefore also treated as a dichotomous variable; no dummy variable approach was implemented, instead levels $A$ and $B$ were assigned values of 1 and 2 respectively (permutation #9 in Table 1).

## 3.2 D-Optimal based DoE analysis

The performance of the Simplex method in locating optima in the investigated spaces was compared against a DoE approach. For this purpose, D-Optimal response surface designs were employed. Here, the model matrix was set to include up to quadratic terms and the categorical inputs were treated with dummy variables unless they were dichotomous. The sample size of these designs, for each case study, was set to 77, 66 and 42; the average number of conditions evaluated by the deployment of the Simplex method with the dummy variable approach, where applicable, in the three case studies (Supporting information, Figs. S6A to S6C for Case studies 1 to 3, respectively). In the first case study, the regression model yielded excellent results, as expected, since all variance was accounted for by its predictions. Hence, it captured entirely accurately the trends of the response and the location of the optimum (i.e. (▲) in Fig. 1A). The calibrated regression models were less successful in predicting the objective function values across the whole space for Case studies 2 and 3. In Case study 2 the capacity of the model to capture trends was low since a coefficient of determination (%$R^2$) of ≈63% was obtained due to the inability of the model to characterize the curvature in the data (Fig. S7). The predicted optimum (i.e. (▲) in Fig. 2A) lay on the correct salt concentration, pH and *resin type*, but indicated the use of NaCl instead of $Na_2SO_4$ salt. Hence, follow-up experiments would focus on the wrong salt and would result in suboptimal binding of the mAb to the Capto S resin. Conversely, in the third case study, the model captured the trends in the data more accurately (Supporting information, Fig. S8), as indicated by a returned %$R^2$ of ≈82%. However, the higher accuracy of

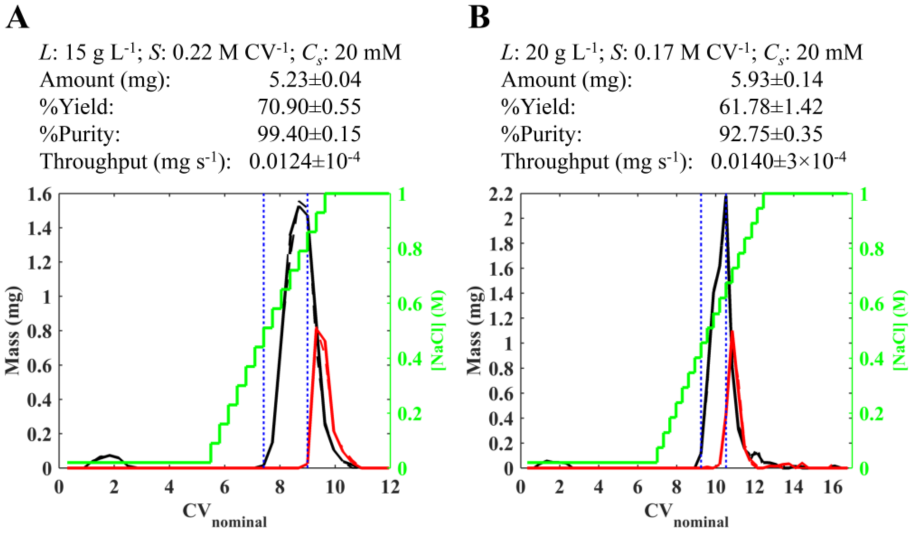


**Figure 4.** Chromatograms obtained from running the RoboColumns at conditions maximizing the objective function *Throughput* × *Purity*, subject to a purity constraint of 85%, for the separation of the BSA (black – and -- lines for each replicate) and cytochrome *c* (red – and -- lines for each replicate) mixture on resins: (**A**) Capto SP ImpRes; and (**B**) Capto S. In both (**A**) and (**B**): Left y-axis shows the mass of each protein in each fraction; Right y-axis shows the employed NaCl linear gradient approximated by a series of steps (green line); x-axis shows the run duration in column volumes based on the nominal fraction volumes (CV$_{\text{nominal}}$); Vertical blue lines indicate the start and end points of the selected product pool. The text at the top of **A** and **B** summarizes the corresponding condition in terms of load ($L$), gradient slope ($S$) and starting salt concentration ($C_s$), and the returned BSA amount, Yield, Purity and Throughput (± 1 standard deviation) from the product pool.

the model was not accompanied by an accurate characterization of the optimum. Here, the model predicted that the optimal separation of the two proteins employed resin Capto SP ImpRes instead of Capto S (i.e. (▲) in Fig. 3A). Furthermore, the model indicated that the separation on this resin could be improved by increasing the load and the gradient slope and consequently it did not approximate the optimal separation on Capto SP ImpRes (i.e. (▼) in Fig. 3A) either. Hence, similar to the observations from Case study 2, the adoption of a D-Optimal design and regression modeling would guide additional development work towards sub-optimal directions by failing to yield reliable results.

## 3.3 Evaluation of the Simplex method and the dummy variable approach

### *3.3.1 Identification of optima*

The results in Table 1 indicate the impact of arbitrary level assignments of the categorical inputs on the performance of the method. This was clearly demonstrated by Case study 1 where the monotonic behavior between the objective function values and the levels of the input $x_3$ could be severely disrupted. The same, but less evidently, applied for Case study 2 between the objective function values and the categorical input *resin type* since here assigning adjacent numerical levels to Capto S and SP Sepharose FF, both resins being ion exchangers, led to a higher convergence to the global optimum compared to the alternative (Table 1).

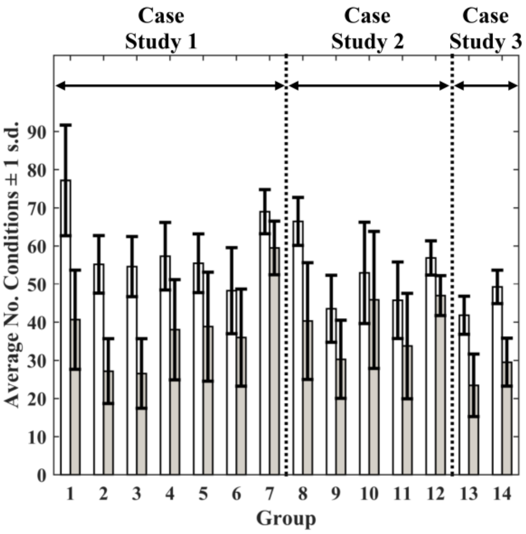
To overcome this challenge, an alternative approach was adopted wherein the categorical inputs were replaced by dummy variables. In Case study 1, three such dummy variables were defined, due to the four levels of the categorical input, and the deployment of the Simplex method was entirely successful in locating the real optimum. In the second case study the deployment of the method led to a 100% success rate while dealing seamlessly with dummy variables, due to the *resin type* input, a dichotomous input (i.e. *salt type*), and numerical inputs. The last case study included a single, dichotomous, categorical input (i.e. *resin type*). As a result, a dummy variable approach was not necessary and the deployment of the method still identified the global optimum with a 100% success rate.

By comparison, the implementation of a regression analysis approach, employing a response surface D-Optimal design, was not entirely successful in Case studies 2 and 3, where real data where employed; in both case studies the models' estimations failed to at least point towards the correct levels of the categorical inputs. Hence, planning future development activities based on the predictions of these models would be wasteful; they would employ wrongly chosen salts and resins and would need to be repeated upon the elucidation of the correct selection. This observation is similar to conclusions drawn from other studies wherein higher order regression models were also considered [9].

### *3.3.2 Simplex method efficiency*

The adoption of a dummy variable approach, when multilevel categorical inputs are involved, leads to an increase in dimensionality and an accrued increase in the average number of conditions evaluated by the method compared to the situation where no dummy variables are defined (i.e. white bars of Group 1 versus Groups 2–6, and white bars of Group 8 versus groups 9–11 in Fig. 5 for Case studies 1 and 2, respectively). This increase is, however, attributed to the encirclement movements that the method carries out prior to its termination. If the method was terminated once the optimum had been reached, and hence did not engage in its encirclement, then the adoption of the dummy variable approach is directly comparable to those cases where no dummy variables are employed (i.e. grey bars of Group 1 versus Groups 2–6, and grey bars of Group 8 versus groups 9–11 in Fig. 5 for Case studies 1 and 2, respectively). This is in agreement with the features of the Simplex methods in general since the addition of inputs (or dimensions) in an optimization problem does not affect significantly the number of points evaluated by the method; only the initialization of the method and the shrink movements are affected by the increased dimensionality. At the same time, it needs to be highlighted that the inclusion of dummy variables led to a more balanced search of a space across the categorical inputs, as indicated by Figs. S2 and S3B (Supporting information), but at the same time, the Simplex method still maintains its ability to screen out unfavorable conditions rapidly since, for example, in Fig. S3B only a narrow range of the numerical inputs is sampled. Therefore, any increase in the number of selected conditions cannot be considered as wasteful.

Figure 5 also assesses a different regime of the deployment of the Simplex method in cases wherein categorical inputs are present. This is the separation of the considered space into subspaces, according to the levels of the categorical inputs, and the separate deployment of the method for each such subspace. For example, in Case study 1, four sub-spaces could be defined (one for each level of the $x_3$ input) whereas in Case study 2 six spaces can be distinguished due to the two and three levels of the *salt* and *resin type* categorical inputs (i.e. 2 × 3 = 6 combinations). Then, an overall average number of conditions evaluated by the method can be obtained by summing over the results of the deployments for each space. This is indicated by Groups 7, 11 and 14 in Fig. 5 for Case studies 1 to 3, respectively. Comparing Groups 2–6, 9–10 and 13 to the aforementioned three groups in Fig. 5, and in particular the grey bars (i.e. early termination of method without optimum encirclement), shows that the concurrent consideration of all levels of the categorical inputs can lead to a more efficient deployment of the method. This behavior is more apparent when the number of the levels of the categorical inputs is high (i.e. Groups 2–6 versus Group 7 and Groups 9, 10 versus Group 11 in Fig. 5 for Case studies 1 and 2, respectively), whereas when the overall number of levels is low, the efficiency of the two approaches is similar (i.e. Group 13 versus Group 14 in Fig. 5 for Case study 3). For Case studies 1 and 2 such a regime can also be compared against the dummy variable approach (i.e. Group 1 versus Group 7 and Group 8 versus Group 12 in Fig. 5 for Case studies 1 and 2, respectively) and similar conclusions can be drawn.



| Group | Case Study | Approach/Level assignment |
|---|---|---|
| **1** | **1** | **Dummy variable approach** |
| 2 | 1 | $x_3$: *A* → 1; *B* → 2; *C* → 3; *D* → 4 |
| 3 | 1 | $x_3$: *A* → 4; *B* → 3; *C* → 2; *D* → 1 |
| 4 | 1 | $x_3$: *A* → 1; *B* → 4; *C* → 3; *D* → 2 |
| 5 | 1 | $x_3$: *A* → 4; *B* → 1; *C* → 2; *D* → 3 |
| 6 | 1 | $x_3$: *A* → 3; *B* → 1; *C* → 4; *D* → 2 |
| 7* | 1 | 150 Searches for each level of $x_3$ |
| **8** | **2** | **Dummy variable approach** |
| 9 | 2 | *Resin type: A* → 1; *B* → 2; *C* → 3 |
| 10 | 2 | *Resin type: A* → 2; *B* → 1; *C* → 3 |
| 11 | 2 | *Resin type: A* → 1; *B* → 3; *C* → 2 |
| 12* | 2 | 150 Searches for each of 6 combinations of *salt type* and *resin type* levels |
| **13** | **3** | **Dummy variable approach** |
| 14* | 3 | 150 Searches for each level of *resin type* |

**Figure 5.** Bar plot of the average number of unique conditions (±1 standard deviation) selected by the 150 simplex searches per case study and assessed assignment of values to the levels of the categorical inputs. The groups in the x-axis are summarized in the right hand side table. White bars calculated based on continuing a search until termination. Grey bars calculated based on terminating a search once the global optimum has been identified. Groups 7, 12 and 14 simulate a scenario in which an experimental space is divided into sub-spaces according to the levels of the categorical inputs and their individual exploration by the Simplex method. For these groups the bars and error bars are calculated as a sum of means and a square root of sum of variances, respectively. These are obtained from 150 random searches per sub-space (i.e. four in Case study 1, six in Case study 2 and two in Case study 3).

Finally, a brief comparison can be made between the efficiency of the Simplex method and an approach employing response surface methodology. While a D-Optimal design was employed previously, alternative designs, such as central composite designs, could also be employed in the presence of categorical inputs. This would require however the generation of such a design for each combination of the levels of the categorical variables. Hence in Case study 2, six composite designs could be employed which would require a total of 54 conditions for evaluation whereas 30 conditions would be required in Case study 3 for two composite designs. These are both lower than the conditions required by the Simplex method with the dummy variables throughout the majority of the simplex searches (Supporting information, Figs. S6B and S6C). However, by preventing the method from encircling the located optima, ≈80% of the deployed searches evaluate less than 54 and 30 conditions for Case studies 2 and 3, respectively (Figs. S6D and S6E, respectively). Hence, while the Simplex method would be less efficient than the described DoE based approach, if left to terminate naturally, it is within the ability of the method to be at least as efficient as such an approach by preventing it from encircling the located optima.

## 4 Conclusions

Experimental studies during the early stages of bioprocess development are tasked with identifying promising operating conditions from a large range of alternatives. Such studies often focus on both numerical and categorical inputs. Examples of the latter include buffer species, solvents, media etc. and their inclusion aims to select an appropriate system which will be part of future experiments. Making such a selection with confidence at an early stage of the development train is therefore of critical importance. The applicability of the grid-compatible Simplex method for concurrent investigation of categorical and numerical inputs was investigated through three case studies; one employing in silico data and two based on experimental data. It was observed that the combination of the method with an approach dealing with the categorical inputs via the definition of dummy variables led to the identification of optimal conditions with high success rates and in an efficient fashion. The inclusion of dummy variables was shown to avoid the generation of spurious optima, which may emerge from arbitrary level assignments of the categorical inputs, and the method allowed for seamless consideration of numerical inputs along with various types of categorical inputs (i.e. multi-level and/or dichotomous). By contrast, an approach employing response surface designs, such as D-Optimal designs, failed to capture the data trends and to identify the optima accurately. The results presented here further support the view that the grid compatible Simplex variant is an attractive approach for early-phase bioprocess development and demonstrate its suitability for deployment in studies employing both numerical and categorical variables. This enhances the wide applicability of the method and represents a novel development for Simplex methods since they are not traditionally applicable to such mixed optimization problems.

*A. Velayudhan acknowledges a Manufacturing Fellowship Award from the Engineering and Physical Sciences Research Council (EPSRC) of the United Kingdom. The assistance of Maria Parau and José Mauricio Martín Bufájer in the generation of the data in the third case study is duly acknowledged. The provision of the mAb material from Muazzam Khan is also gratefully acknowledged.*

*The authors declare no financial or commercial conflict of interest.*

## Nomenclature

| | | |
|---|---|---|
| $\%MB$ | (–) | Mass balance |
| $\%R^2$ | (–) | Coefficient of determination |
| $\%VB$ | (–) | Volumetric balance |
| [Salt] | (mM) | Salt concentration in binding buffer |
| $A$ | (–) | First level of categorical input |
| $\alpha$ | (–) | Standard curve coefficients for BSA |
| $B$ | (–) | Second level of categorical input |
| $\beta$ | (–) | Standard curve coefficients for Cytochrome C |
| $BSA\ Mass_i$ | (mg) | BSA mass in $i^{th}$ fraction |
| $C$ | (–) | Third level of categorical input |
| $C_{\mathrm{mAb\ in\ flowthrough}}$ | (g $L^{-1}$) | mAb concentration in flowthrough |
| $C_s$ | (–) | Starting salt concentration in gradient |
| $CV_I$ | (–) | End point of product pool in column volumes |
| $CV_{\mathrm{Load}}$ | (–) | Duration of column load in column volumes |
| $CV_{\mathrm{nominal}}$ | (–) | Duration of run in column volumes |

| | | |
|---|---|---|
| $CV_{\mathrm{Wash}}$ | (–) | Duration of column wash in column volumes |
| $CytC\ Mass_i$ | (mg) | Cytochrome C mass in $i^{th}$ fraction |
| $D$ | (–) | Fourth level of categorical input |
| $d$ | (–) | Dummy variable |
| $i$ | (–) | Start fraction number of product pool |
| $I$ | (–) | End fraction number of product pool |
| $L$ | (g $L^{-1}$) | Load |
| $b$ | (cm) | Pathlength |
| $Mass\ in$ | (mg) | Protein mass in column load |
| $Mass_i$ | (mg) | Protein mass in $i^{th}$ fraction |
| $n$ | (–) | $n^{th}$ product pool |
| $N$ | (–) | Number of possible product pools |
| $resin\ type$ | (–) | Categorical input in Case studies 2 and 3 |
| $S$ | (mM $CV^{-1}$) | Gradient slope |
| $salt\ type$ | (–) | Categorical input in Case study 2 |
| $V_{\mathrm{measured,i}}$ | (mL) | Determined volume in $i^{th}$ fraction |
| $V_{\mathrm{nominal,i}}$ | (mL) | Nominal volume in $i^{th}$ fraction |
| $x_1$ | (–) | Numerical input in Case study 1 |
| $x_2$ | (–) | Numerical input in Case study 1 |
| $x_3$ | (–) | Categorical input in Case study 1 |
| $\zeta$ | (–) | Number of levels in categorical input |
| $\lambda$ | (nm) | Wavelength |

## 5 References

[1] Kumar, V., Bhalla, A., Rathore, A. S., Design of experiments applications in bioprocessing: Concepts and approach. *Biotechnol. Progr.* 2014, *30*, 86–99.

[2] Mandenius, C.-F., Brundin, A., Bioprocess optimization using design-of-experiments methodology. *Biotechnol. Progr.* 2008, *24*, 1191–1203.

[3] Zartman, J., Restrepo, S., Basler, K., A high throughput template for optimizing *Drosophila* organ culture with response surface methods. *Development* 2013, *140*, 667–674.

[4] Chollangi, S., Parker, R., Singh, N., Li, Y., Borys, M., Li, Z., Development of robust antibody purification by optimizing protein-A chromatography in combination with precipitation methodologies. *Biotechnol. Progr.* 2015, *112*, 2292–2304.

[5] Heldin, E., Grönlund, S., Shanagar, J., Hallgren, E. et al., Development of an intermediate chromatography step in an insulin purification process. The use of a High Throughput Process Development approach based on selectivity parameters. *J. Chromatogr. B* 2014, *973*, 126–132.

[6] McDonald, P., Tran, B., Williams, C. R., Wong, M., Zhao, T. et al., The rapid identification of elution conditions for therapeutic antibodies from cation-exchange chromatography resins using high-throughput screening. *J. Chromatogr. A* 2016, *1433*, 66–74.

[7] Chhatre, S., Konstantinidis, S., Ji, Y., Edwards-Parton, S., Zhou, Y., Titchener-Hooker, N. J., The simplex algorithm for the rapid identification of operating conditions during early bioprocess development: Case studies in Fab' precipitation and multimodal chromatography. *Biotechnol. Bioeng.* 2011, *108*, 2162–2170.

[8] Konstantinidis, S., Chhatre, S., Velayudhan, A., Heldin, E., Titchener-Hooker, N. J., The hybrid experimental simplex algorithm – An alternative method for 'sweet spot' identification in early bioprocess development: Case studies in ion exchange chromatography. *Anal. Chim. Acta* 2012, *743*, 19–32.

[9] Konstantinidis, S., Welsh, J. P., Roush, D. J., Velayudhan, A., Application of Simplex-based experimental optimisation to challenging bioprocess development problems: case studies in downstream processing. *Biotechnol. Progr.* 2016, *32*, 404–419.

[10] Nelder, J. A., Mead, R., A simplex method for function minimization. *Comput. J.* 1965, *7*, 308–313.

[11] Spendley, W., Hext, G. R., Himsworth, F. R., Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics* 1962, *4*, 441–461.

[12] Tarrant, R. D. R., Velez-Suberbie, M. L., Tait, A. S., Smales, C. M., Bracewell, D. G., Host cell protein adsorption characteristics during protein a chromatography. *Biotechnol. Progr.* 2012, *28*, 1037–1044.

[13] Welsh, J. P., Petroff, M. G., Rowicki, P., Bao, H. et al., A practical strategy for using miniature chromatography columns in a standardized high-throughput workflow for purification development of monoclonal antibodies. *Biotechnol. Progr.* 2014, *30*, 626–635.

[14] McGown, E., Hafeman, D., Multichannel pipettor performance verified by measuring pathlength of reagent dispensed into a microplate. *Anal. Biochem.* 1998, *258*, 155–157.

[15] De Aguiar, P. F., Bourguignon, B., Khots, M. S., Massart, D. L., Phan-Than-Luu, R., D-Optimal designs. *Chemom. Intell. Lab. Syst.* 1995, *30*, 199–210.

[16] Fedorov, V. V., Studden, W. J., Klimko, E. M. (Eds), *Theory of Optimal Experiments*, Academic Press, New York 1972.

[17] Cohen, J., Cogen, P., West, G. W., Aiken, L. S., Riegert, D. (Eds.), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates Inc., London 2003.

[18] Harinarayan, C., Mueller, J., Ljunglöf, A., Fahrner, R. et al., An exclusion mechanism in ion exchange chromatography. *Biotechnol. Bioeng.* 2006, *95*, 775–787.

[19] Urmann, M., Graalfs, H., Joehnck, M., Jacob, L. R., Frech, C., Cation-exchange chromatography of monoclonal antibodies: Characterization of a novel stationary phase designed for production-scale purification. *mAbs* 2010, *2*, 395–404.