Engineering 5 (2019) 1077–1092

Contents lists available at ScienceDirect

Engineering

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

Research
Smart Process Manufacturing: Deep Integration of AI and Process Manufacturing—Article

Optimal Antibody Purification Strategies Using Data-Driven Models

Songsong Liu [a,b,*], Lazaros G. Papageorgiou [c,*]

[a] School of Management, Harbin Institute of Technology, Harbin 150001, China
[b] School of Management, Swansea University, Swansea SA1 8EN, UK
[c] Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, London WC1E 7JE, UK

ARTICLE INFO

Article history:
Received 31 October 2018
Revised 15 December 2018
Accepted 21 December 2018
Available online 18 October 2019

Keywords:
Antibody purification
Multiscale optimization
Antigen-binding fragment
Mixed-integer programming
Data-driven model
Piecewise linear regression

ABSTRACT

This work addresses the multiscale optimization of the purification processes of antibody fragments. Chromatography decisions in the manufacturing processes are optimized, including the number of chromatography columns and their sizes, the number of cycles per batch, and the operational flow velocities. Data-driven models of chromatography throughput are developed considering loaded mass, flow velocity, and column bed height as the inputs, using manufacturing-scale simulated datasets based on microscale experimental data. The piecewise linear regression modeling method is adapted due to its simplicity and better prediction accuracy in comparison with other methods. Two alternative mixed-integer nonlinear programming (MINLP) models are proposed to minimize the total cost of goods per gram of the antibody purification process, incorporating the data-driven models. These MINLP models are then reformulated as mixed-integer linear programming (MILP) models using linearization techniques and multiparametric disaggregation. Two industrially relevant cases with different chromatography column size alternatives are investigated to demonstrate the applicability of the proposed models.

1. Introduction

The global industry has been experiencing accelerating changes during the recent transformation of traditional manufacturing into smart manufacturing [1,2]. During the conversion process, industries face a number of challenges posed by smart manufacturing, which have attracted great attention in both academic and practitioner communities [3], particularly in the process industry [4]. Some of the challenges to be covered in this work include:

- The use and analysis of data, with a particular focus on the development of data-driven surrogate/metamodels to simplify complex processes and to enable manufacturing intelligence;
- The implementation of multiscale modeling and optimization to integrate strategic and planning decisions with operations in order to support enterprise-wide coordination and optimization;
- The development of computationally efficient models, algorithms, and tools in order to find global optimal solutions for smart manufacturing decision-making and to enable large-scale optimization.

In this work, we aim to develop optimization-based decision-making models for optimal purification strategies in the manufacturing process of an antibody product based on simple data-driven models, in an attempt to cope with the above challenges in the biopharmaceutical industry. In order to achieve better control of the processes and improve production efficiency, biopharmaceutical manufacturing process optimization problems have been investigated using different modeling and solution techniques, such as metaheuristic [5], dynamic optimization [6], evolutionary algorithm [7–9], Markov decision method [10], and mixed-integer programming [11–22]. Data-driven models—also known as surrogate models or metamodels—refer to models that are built on the basis of data, but are not dependent on theoretical knowledge of the concerned processes or systems. Data-driven models of complex processes and systems provide model simplicity and computational efficiency [23], and their integration with optimization requires less computational effort and has a broad application in the engineering field [24,25]. In particular, such models have demonstrated research benefits in the modeling and optimization of chromatography purification operations [26–29]. However, only a few attempts have been made to integrate data-driven models into optimization models for biopharmaceutical purification processes. Nagrath et al. [30] developed an artificial neural network (ANN)-based hybrid model for the optimization of preparative chromatographic processes.

* Corresponding authors.
E-mail addresses: s.liu@hit.edu.cn (S. Liu), l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou).

https://doi.org/10.1016/j.eng.2019.10.011

Pirrung et al. [31] developed ANNs from detailed mechanistic models and integrated them into the optimization of biopharmaceutical downstream processes for the maximum yield of a process with three different chromatographic columns.

Antigen-binding fragment (Fab) products are regarded as the next generation of protein-based biotherapeutics after monoclonal antibody (mAb) products, and offer many advantages due to their simpler and smaller structure [32]. There is a need to develop a cost-efficient process for Fab production in industrial practice [33], but few works on this topic exist in the literature. In this work, the multiscale optimization of the purification process of a Fab product is addressed, using microscale column chromatography experiment data for manufacturing-scale chromatography optimization. In order to achieve the most cost-efficient process, besides considering chromatography column-sizing strategies at the design level, operational decisions are taken into account—particularly the flow velocity at chromatography steps. Data-driven models are developed to estimate the chromatography throughput. When incorporating the developed data-driven models, a number of mixed-integer programming models are proposed to find the optimal Fab purification strategies, and are examined through case studies. To the best of our knowledge, this is the first work in the literature on the multiscale optimization of the purification process of Fab using data-driven models.

The rest of this paper is organized as follows: The optimization problem is described in Section 2. The data-driven models of the chromatography operations are developed in Section 3, while the proposed mathematical programming optimization models are described in Section 4. Section 5 presents industrially relevant case studies, followed by the computational results and discussion in Section 6. Finally, concluding remarks are provided in Section 7.

2. Problem statement

This work investigates the optimization of the manufacturing processes of a Fab product. Fig. 1 shows the Fab manufacturing process flowsheet studied in this work. Initially, the mammalian cells expressing the Fab are cultured in bioreactors in the upstream processing (USP) before entering into the downstream processing (DSP). In the DSP, the Fab protein product is purified by a number of operations, including centrifugation, homogenization, filtration, ultrafiltration/diafiltration (UF/DF), and three packed-bed chromatography steps, which include affinity, cation-exchange, and anion-exchange chromatography steps.

The chromatography column-sizing strategies are important to the efficiency of the whole purification process. These strategies include the number of parallel columns at each step, diameter and bed height of the columns, and number of cycles per batch, which significantly affect the cost, time, and output of the whole manufacturing process. In real practice, there are standard columns with different diameter sizes, and the bed height are set to a range of typical integer values. Thus, in this work, the column diameter and bed height are optimized from a given set of discrete alternative values.

Chromatography operations are complex process with challenges in modeling their behaviors. To optimize the chromatography operational strategies, metamodeling techniques are used in this work to mimic and predict the chromatography process performance, particularly the chromatography throughput, indicating the rate of the product output of one column within a given period, which is an important metric of the chromatography operation. To develop manufacturing-scale data-driven models, data from a microscale column chromatography laboratory experiment are collected and then fitted to obtain isotherm parameters. First-principle chromatography models are also created with scale-related parameters to capture challenges on scaling [34]. First-principle models with isotherm parameters are solved, followed by simulation runs to generate manufacturing-scale datasets using COMSOL Multiphysics simulation software [35]. The datasets include the throughput output under different input conditions of loaded mass, flow velocity, and column bed height at the two chromatography steps in the binding and elution mode—namely, the affinity and cation-exchange chromatography steps. The datasets are then used to derive data-driven models, which are incorporated into the proposed optimization models in order to reach optimal manufacturing-scale chromatography decisions. The whole procedure of the multiscale optimization approach is presented in Fig. 2; the steps presented in the last three tan-colored boxes in the figure will be described in detail in the following sections. Note that it is assumed that the considered input conditions do not affect other chromatography parameters, such as the resin's yield, binding capacity, and lifetime, which are known parameters in this problem.

In summary, the optimization problem considered in this work is described as follows:

Given:

- The process flowsheet of a Fab product;
- The number of bioreactors and their volumes, along with the bioreactor titer;
- The chromatography operation parameters, including the yield, buffer and eluent usage, dynamic binding capacity, lifetime, etc.;
- The non-chromatography operation parameters, including the yield, processing rate, buffer usage, etc.;
- Time-related data, including the processing rate, bioreaction time, annual operating time, etc.;
- Cost-related data, including the labor wage and the costs of the resin, buffer, media, equipment, etc.;
- Chromatography data from simulations based on first-principle models and microscale column experiment data;
- The candidate column diameter and bed height, and the maximum number of cycles and columns.

Determine:

- Chromatography column-sizing strategies, such as the column diameter, bed height, and number of columns at each chromatography step;
- Operational strategies, such as the liner velocities, loaded product mass, and number of cycles at the affinity and cation-exchange chromatography steps;

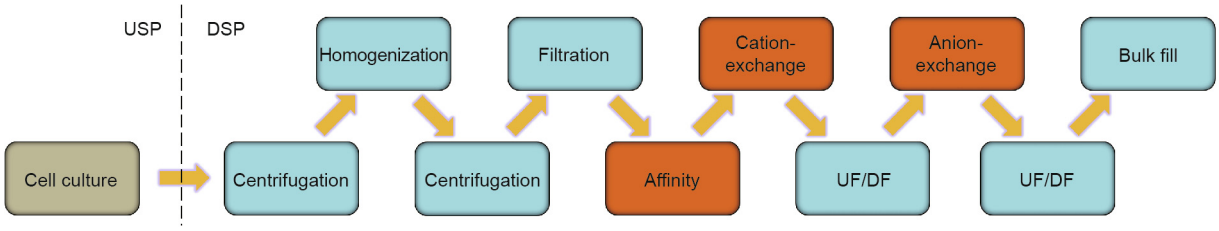

Fig. 1. A Fab manufacturing process. Orange boxes represent chromatography operations. USP: upstream processing; DSP: downstream processing; UF/DF: ultrafiltration/diafiltration.

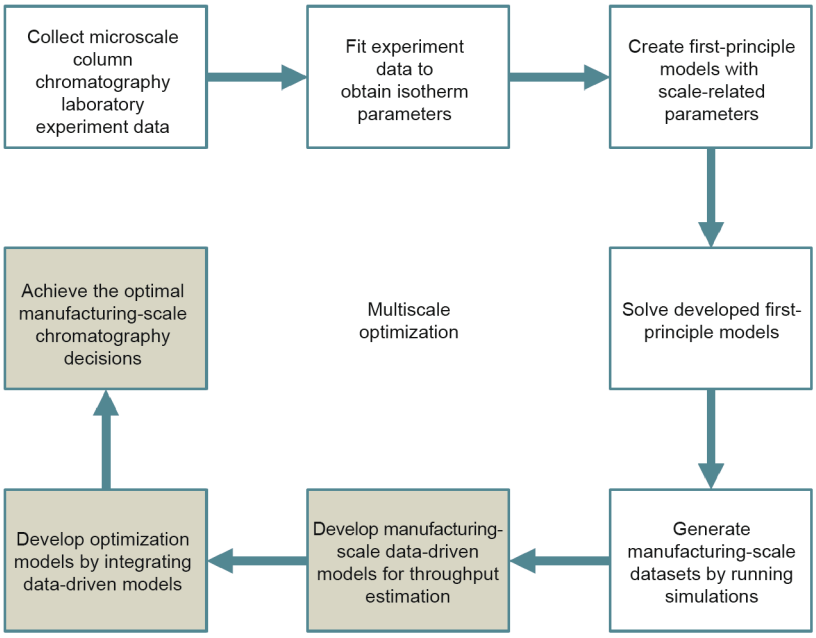

Fig. 2. Procedure for the multiscale optimization of chromatographic strategies.

- The number of total completed batches;
- The annual total processing time;
- The annual total production output;
- The annual total cost.

So as to minimize the cost of goods (COG) per gram of the Fab product—that is, the ratio of the annual total cost to the annual total production output.

3. Data-driven models

In this section, simple data-driven models are developed for chromatography performance based on simulation datasets. In this work, as a key performance criterion, the chromatography throughput is considered as the output of the models. Only the affinity and cation-exchange chromatography steps are modeled, due to the limited available data. Both the datasets for the affinity and cation-exchange chromatography steps are based on a single chromatography column with a diameter of 1 m. Three key variables influencing the chromatography throughput are considered in the datasets as the inputs of the data-driven models, namely: the loaded mass, flow velocity, and column bed height.

A number of widely used methods are implemented in order to achieve accurate and simple models, including linear regression, support vector regression (SVR) [36], kriging [37], pace regression [38], response surface methodology (RSM) [39], and piecewise linear regression [40]. To estimate the prediction accuracy of these methods, cross-validation is performed. Given a dataset, n-fold cross-validation randomly splits the samples into n subsets of equal size. Then $(n-1)$ subsets of samples are used in the training, and the remaining set is used to validate the prediction accuracy of the obtained data-driven models. In this work, 10 rounds of five-fold cross-validation are performed by generating random sample splits, and the mean absolute error (MAE) over all 50 testing sets is used as the final error metric for comparison of the prediction accuracy. Linear regression, SVR, kriging, and pace regression are implemented in WEKA machine learning software [41] with default settings, while RSM and piecewise linear regression are run in GAMS [42] using the CPLEX mixed-integer linear programming (MILP) solver. Table 1 presents the prediction error results obtained after running the cross-validation on the datasets for affinity chromatography with 3081 samples, and for cation-exchange chromatography with 2847 samples.

Table 1 shows that the piecewise linear regression method gives the best prediction accuracy among all the tested methods. The piecewise linear regression method creates a model to separate samples into multiple complementary intervals on one input variable, with the flexibility of each interval being fitted by its own linear regression function. Considering its ease of modeling and understanding, the piecewise linear regression method is chosen to create the final data-driven model for chromatography throughput estimation, where all samples are used in the training process. In the procedure of the piecewise linear regression [40], each input variable in turn serves once as the partition variable. For each partition variable, an MILP model is solved to determine the breakpoint of the partition variable between only two intervals, and the variable corresponding to the minimum training error is kept as the partition variable. Until the termination criterion is met, the number of intervals is increased and MILP models are solved iteratively with the same partition variable. Following this procedure, the following two models are obtained to estimate throughput, TP_s^1, at chromatography step s (af for affinity and ce for cation-exchange):

$$TP_{af}^1 = \begin{cases} 0.1914 \cdot LM_{af}^1 + 0.3570 \cdot V_{af} - 12.0477 \cdot H_{af} + 230.1318, \text{ if } 654 < LM_{af}^1 \le 1643 \\ -0.003294 \cdot LM_{af}^1 + 0.008982 \cdot V_{af} + 43.5598 \cdot H_{af} - 649.3012, \text{ if } 1643 < LM_{af}^1 \le 2629 \\ 0, \text{ if } 2629 < LM_{af}^1 \le 7069 \end{cases} \quad (1)$$

Table 1
MAE comparison of different methods.

Method	Affinity chromatography	Cation-exchange chromatography
Linear regression	121.40	777.26
SVR	111.31	700.98
Kriging	115.39	720.67
Pace regression	115.50	712.93
RSM	90.45	527.31
Piecewise linear regression	34.68	233.39

$$TP^1_{ce} = \begin{cases} 0.1287 \cdot LM^1_{ce} + 2.3940 \cdot V_{ce} - 51.4883 \cdot H_{ce} + 895.2814, \text{ if } 3142 < LM^1_{ce} \le 12242 \\ -0.002489 \cdot LM^1_{ce} + 0.05041 \cdot V_{ce} + 260.5745 \cdot H_{ce} - 3890.1058, \text{ if } 12242 < LM^1_{ce} \le 19698 \\ 0, \text{ if } 19698 < LM^1_{ce} \le 28274 \end{cases} \tag{2}$$

where the superscript 1 of the variables refers to the column with a diameter of 1 m. In the obtained models, the loaded mass, LM^1_s, at chromatography step s is determined to separate three intervals by the procedure given in Ref. [40], which provides a smaller prediction error than the other two variables, V_s and H_s, which are the linear velocity and the column bed height, respectively.

In order to incorporate the above obtained data-driven models into the optimization models for decision-making, they need to be reformulated by introducing a binary variable. The binary variable, $O_{s,r}$, is defined to be equal to 1 if the loaded mass at chromatography step s lies within interval r; the throughput output is then obtained from the corresponding liner function. As there is only one interval to be selected, the binary variable should satisfy Eq. (3):

$$\sum_r O_{s,r} = 1, \ \forall s \in \{\text{af}, \text{ce}\} \tag{3}$$

The value of the separate input, LM^1_s, in this model should be between the two breakpoints ($bp_{s,r}$) of the selected interval, which can be formulated as the following linear equation:

$$\sum_r bp_{s,r-1} \cdot O_{s,r} + \varepsilon \le LM^1_s \le \sum_r bp_{s,r} \cdot O_{s,r}, \ \forall s \in \{\text{af}, \text{ce}\} \tag{4}$$

where ε is a small number to separate two successive intervals at the breakpoints. The general expression of the throughput output is as follows:

$$TP^1_s = \sum_r \left(\beta^{LM}_{s,r} \cdot LM^1_s + \beta^V_{s,r} \cdot V_s + \beta^H_{s,r} \cdot H_s + \beta^0_{s,r} \right) \cdot O_{s,r}, \quad \forall s \in \{\text{af}, \text{ce}\} \tag{5}$$

where $\beta^{LM}_{s,r}$, $\beta^V_{s,r}$, $\beta^H_{s,r}$, and $\beta^0_{s,r}$ are parameters in the function. When interval r' is selected at step s, that is, $O_{s,r'} = 1$, Eq. (4) becomes $bp_{s,r'-1} < LM^1_s \le bp_{s,r'}$. In this case, the loaded mass lies in the interval r'; Eq. (5) then ensures that the throughput is equal to the output of the linear function on the selected interval r':

$$TP^1_s = \beta^{LM}_{s,r'} \cdot LM^1_s + \beta^V_{s,r'} \cdot V_s + \beta^H_{s,r'} \cdot H_s + \beta^0_{s,r'}, \text{ if } bp_{s,r'-1} \le LM^1_s \le bp_{s,r'} \tag{6}$$

The above throughput regression model will be incorporated into the optimization models of Fab manufacturing processes in the next section.

4. Optimization models

In this section, we will use the above data-driven models to develop two mixed-integer nonlinear programming (MINLP) models of the purification processes of a Fab product, using different modeling methods of alternative column sizes. These MINLP models are then reformulated as MILP models using exact linearization techniques and multiparametric disaggregation.

4.1. MINLP model A

MINLP model A is formulated based on the model provided in Ref. [16] for a mAb manufacturing process. In this model, a number of alternative column volume sizes are first generated from combinations of the given discrete column diameter and bed height. The column volume size i at chromatography step s ($cv_{s,i}$) corresponds to a specific diameter size ($dm_{s,i}$) and bed height ($h_{s,i}$) among the given alternatives.

4.1.1. Column volume

The total column volume (TCV_s) at chromatography step $s \in CS$ (CS is the set of chromatography steps, CS ={af, ce, ae}, ae stands for anion-exchange) is determined by the number of columns ($CN_{s,i}$) in the selected size multiplied by the corresponding column volume:

$$TCV_s = \sum_i cv_{s,i} \cdot CN_{s,i}, \ \forall s \in CS \tag{7}$$

By introducing a binary variable, $X_{s,i}$, for the selection of column size i at chromatography step s, the following constraints can ensure that only one column size can be selected:

$$\sum_i X_{s,i} = 1, \ \forall s \in CS \tag{8}$$

$$CN_{s,i} \le \max CN_s \cdot X_{s,i}, \ \forall s \in CS, i \tag{9}$$

where $\max CN_s$ refers to the maximum allowed number of columns.

In each batch, the available resin volume at a chromatography step s—that is, the total column volume (TCV_s) multiplied by the number of cycles per batch (CYN_s)—must be sufficient to process all protein mass entering into that step (M_{s-1}), and the required resin volume (RV_s) is determined by the dynamic binding capacity (dbc_s) and resin utilization factor (μ).

$$CYN_s \cdot TCV_s \ge RV_s, \ \forall s \in CS \tag{10}$$

$$RV_s = \frac{M_{s-1}}{dbc_s \cdot \mu}, \ \forall s \in CS \tag{11}$$

4.1.2. Product mass

The initial product mass (M_0) entering into the DSP process is equal to the bioreaction titer ($titer$) multiplied by the bioreactor working volume—that is, the bioreactor volume (brv) times the working volume ratio (α).

$$M_0 = titer \cdot \alpha \cdot brv \tag{12}$$

The product protein mass going out from step s is equal to the mass from the previous one, step (s – 1), multiplied by the corresponding yield of step s, yd_s.

$$M_s = yd_s \cdot M_{s-1}, \ \forall s \tag{13}$$

The annual product output (AP) is the product mass after the bulk fill step ($s = \text{bf}$):

$$AP = \sigma \cdot BN \cdot M_{\text{bf}} \tag{14}$$

where BN is the number of completed batches, upper bounded by the maximum allowed batch number, and σ is the batch success rate.

4.1.3. Product volume

The initial product volume entering into the DSP (PV_0) is equal to the working volume of the bioreactor, formulated as follows:

$$PV_0 = \alpha \cdot brv \tag{15}$$

For the first four steps of the process, including the first centrifugation ($s = \text{ct}_1$), homogenization ($s = \text{ho}$), second centrifugation ($s = \text{ct}_2$), and filtration ($s = \text{fi}$) steps, the product volume remaining after step s (PV_s) is equivalent to the product volume entering into this step:

$$PV_s = PV_{s-1}, \ \forall s \in \{\text{ct}_1, \text{ho}, \text{ct}_2, \text{fi}\} \tag{16}$$

At the affinity ($s = \text{af}$) and cation-exchange chromatography ($s = \text{ce}$) steps, the product volume is equal to the eluent volume, while at the anion-exchange chromatography ($s = \text{ae}$) step, the product volume does not change.

$$PV_s = ecv_s \cdot CYN_s \cdot TCV_s|_{s\neq ae} + PV_{s-1}|_{s=ae},\ \forall s \in CS \tag{17}$$

where ecv_s is the eluent volume to column volume ratio.

At the first UF/DF step ($s = uf_1$), the flush volume is added to the product volume entering the step:

$$PV_{uf_1} = (fvr + 1) \cdot PV_{ce} \tag{18}$$

where fvr is the flush volume ratio at this step. At the second UF/DF step ($s = uf_2$), the remaining product volume is the mass divided by the filling concentration, $fconc$:

$$PV_{uf_2} = \frac{M_{uf_2}}{fconc} \tag{19}$$

4.1.4. Buffer volume

The buffer volume added in each step (BV_s) is defined as follows:

$$BV_s = bvr_s \cdot PV_{s-1},\ \forall s \in \{ct_1, ct_2\} \tag{20}$$

$$BV_s = 0,\ \forall s \in \{ho, fi\} \tag{21}$$

$$BV_s = bcv_s \cdot CYN_s \cdot TCV_s,\ \forall s \in CS \tag{22}$$

$$BV_{uf_1} = fvr \cdot PV_{ce} \tag{23}$$

$$BV_{uf_2} = dvr \cdot PV_{uf_2} \tag{24}$$

where bvr_s is the buffer volume ratio at centrifugation step s and bcv_s is the buffer volume to column volume ratio at chromatography step s, and fvr and dvr are the flush volume ratio and diafiltration volume ratio at the first and second UF/DF steps, respectively.

The total required buffer volume in each batch (BBV) is the sum of the buffer volume at all steps, and the annual buffer volume (ABV) is the total buffer volume of all the completed batches.

$$BBV = \sum_s BV_s \tag{25}$$

$$ABV = BN \cdot BBV \tag{26}$$

4.1.5. Processing time

The processing times (T_s) at the affinity and cation-exchange chromatography steps are determined by the mass output of each column divided by its throughput (TP_s):

$$T_s = \frac{M_s}{TP_s \cdot \sum_i CN_{s,i}},\ \forall s \in \{af, ce\} \tag{27}$$

At the anion-exchange chromatography step, the processing times for the loading product (PLT) and adding buffer (BAT) are calculated separately using the volumetric flow rate (VFR) to obtain the processing time at the step.

$$T_{ae} = PLT + BAT \tag{28}$$

$$PLT = \frac{PV_{uf_1}}{VFR \cdot \sum_i CN_{ae,i}} \tag{29}$$

$$BAT = \frac{CYN_{ae} \cdot bcv_{ae} \cdot \sum_i cv_{ae,i} \cdot X_{ae,i}}{VFR} \tag{30}$$

$$VFR = \frac{1}{1000} \cdot vel \cdot \pi \cdot \sum_i \left(\frac{dm_{ae,i}}{2}\right)^2 \cdot X_{ae,i} \tag{31}$$

where vel is the linear velocity of flow at the anion-exchange chromatography step. The processing time at the bulk fill step is assumed to be constant, while at the other non-chromatography steps, the process time is equal to the corresponding product volume divided by the processing rate (pr_s):

$$T_s = \frac{PV_s}{pr_s},\ \forall s \in \{ct_1, ho, ct_2, fi, uf_1, uf_2\} \tag{32}$$

The processing time of one batch, BT, is the total processing time for all steps, divided by the shift duration (sfd) and the number of shifts per day (sfn):

$$BT = \frac{\sum_s T_s}{sfd \cdot sfn} \tag{33}$$

The annual processing time (AT) is the total processing time of all batches:

$$AT = BN \cdot BT \tag{34}$$

which is limited by the annual operating time (aot) minus the seed train bioreaction time (st) and the bioreaction time (brt) of a single batch, $aot - st - brt$.

4.1.6. Data-driven model

The throughput of the 1 m-diameter column is calculated from the piecewise linear regression model obtained in the previous section, including Eqs. (3) and (4). As the selected bed height at chromatography step s is expressed as $\sum_i h_{s,i} \cdot X_{s,i}$ in this model, Eq. (5) is modified as follows:

$$TP_s^1 = \sum_r \left(\beta_{s,r}^{LM} \cdot LM_s^1 + \beta_{s,r}^{V} \cdot V_s + \beta_{s,r}^{H} \cdot \sum_i h_{s,i} \cdot X_{s,i} + \beta_{s,r}^{0}\right) \cdot O_{s,r},\ \forall s \in \{af, ce\} \tag{35}$$

Considering that the throughput of a chromatography column could be regarded as the product of the protein density, linear velocity of flow, and column area, it is assumed that the throughput is proportional to the column area. Thus, it is also proportional to the column diameter squared. In this case, the relationship between the throughput of the selected column and that of the 1 m-diameter column (TP_s^1) is formulated as follows:

$$TP_s = \frac{\sum_i (dm_{s,i})^2 \cdot X_{s,i}}{(refDM)^2} \cdot TP_s^1,\ \forall s \in \{af, ce\} \tag{36}$$

where $refDM$ refers to the reference diameter, which is 100 cm in this work.

In the data-driven regression model, the mass loaded to the 1 m-diameter column is used to proportionally calculate the actual loaded mass to the selected column (LM_s).

$$LM_s = \frac{\sum_i (dm_{s,i})^2 \cdot X_{s,i}}{(refDM)^2} \cdot LM_s^1,\ \forall s \in \{af, ce\} \tag{37}$$

where the loaded mass, LM_s, is defined as the product mass entering each column in each cycle, defined as follows:

$$LM_s = \frac{M_{s-1}}{CYN_s \cdot \sum_i CN_{s,i}},\ \forall s \in \{af, ce\} \tag{38}$$

4.1.7. Objective function

In this work, the objective is to minimize COG per gram, which equals the annual total cost (AC) divided by the annual total production output (AP). All cost terms included in the annual total cost calculation and the related constraints are presented in Supplementary data. The objective function is as follows:

$$COG = \frac{AC}{AP} \tag{39}$$

Thus, the proposed MINLP model A minimizes Eq. (39), subject to the constraints of Eqs. (3), (4), and (7–38), and Eqs. (S1–S19) in Supplementary data.

4.2. MILP model A*

Next, the obtained MINLP model is reformulated as an MILP model for ease of computation. The new linear constraints are presented below.

4.2.1. Integer variable discretization

Similar to the work in Refs. [17–19], in order to facilitate the linearization of nonlinear terms in other constraints, the integer variables $CN_{s,i}$, CYN_s, and BN are discretized and expressed by binary variables, as shown in Eqs. (40–44).

$$CN_{s,i} = \sum_{j=1}^{\max CN_s} j \cdot W_{s,i,j}, \ \forall s \in CS, i \tag{40}$$

$$\sum_{j=1}^{\max CN_s} W_{s,i,j} = X_{s,i}, \ \forall s \in CS, i \tag{41}$$

$$CYN_s = \sum_{k=1}^{\max CYN_s} k \cdot Y_{s,k}, \ \forall s \in CS \tag{42}$$

$$\sum_{k=1}^{\max CYN_s} Y_{s,k} = 1, \ \forall s \in CS \tag{43}$$

$$BN = \sum_{n=1}^{\log_2 \max BN} 2^{n-1} \cdot Z_n \tag{44}$$

where $W_{s,i,j}$, $Y_{s,k}$, and Z_n are binary variables introduced for discretisation of the above integer variables, and j and k are the indices of column number and cycle number, respectively.

4.2.2. Column volume linearization

Based on Eqs. (42) and (43), the nonlinear constraint Eq. (10) can be reformulated as follows:

$$\sum_{k=1}^{\max CYN_s} k \cdot \overline{YV}_{s,k} \geq RV_s, \ \forall s \in CS \tag{45}$$

$$\overline{YV}_{s,k} \leq \max TCV_s \cdot Y_{s,k}, \ \forall s \in CS, k = 1, \ldots, \max CYN_s \tag{46}$$

$$\sum_{k=1}^{\max CYN_s} \overline{YV}_{s,k} = TCV_s, \ \forall s \in CS \tag{47}$$

where an auxiliary variable $\overline{YV}_{s,k} \equiv Y_{s,k} \cdot TCV_s$ is introduced and defined, and $\max TCV_s$ is the maximum total column volume at chromatography step s.

4.2.3. Annual production linearization

By introducing another auxiliary variable, $\overline{ZM}_{s,n} \equiv Z_n \cdot M_s$, to express the bilinear term in Eq. (14), this constraint can be reformulated by the following equations [43]:

$$AP = \sum_{n=1}^{\lceil \log_2 \max BN \rceil} \sigma \cdot 2^{n-1} \cdot \overline{ZM}_{\mathrm{bf},n} \tag{48}$$

$$\overline{ZM}_{\mathrm{bf},n} \leq titer \cdot \alpha \cdot brv \cdot Z_n, \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{49}$$

$$\overline{ZM}_{\mathrm{bf},n} \leq M_{\mathrm{bf}}, \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{50}$$

$$\overline{ZM}_{\mathrm{bf},n} \geq M_{\mathrm{bf}} - titer \cdot \alpha \cdot brv \cdot (1 - Z_n), \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{51}$$

4.2.4. Product and buffer volume linearization

According to Eqs. (46) and (47), Eqs. (17) and (22) can be rewritten as the following two linear equations, respectively:

$$PV_s = ecv_s \cdot \sum_{k=1}^{\max CYN_s} k \cdot \overline{YV}_{s,k} \Big|_{s \neq \mathrm{ae}} + PV_{uf_1}\big|_{s=\mathrm{ae}}, \ \forall s \in CS \tag{52}$$

$$BV_s = bcv_s \cdot \sum_{k=1}^{\max CYN_s} k \cdot \overline{YV}_{s,k}, \ \forall s \in CS \tag{53}$$

$\overline{ZV}_n \equiv Z_n \cdot BBV$ is defined using the following equations:

$$\overline{ZV}_n \leq \max BBV \cdot Z_n, \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{54}$$

$$\overline{ZV}_n \leq BBV, \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{55}$$

$$\overline{ZV}_n \geq BBV - \max BBV \cdot (1 - Z_n), \ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{56}$$

Thus, Eq. (26) can be rewritten as follows:

$$ABV = \sum_{n=1}^{\lceil \log_2 \max BN \rceil} 2^{n-1} \cdot \overline{ZV}_n \tag{57}$$

4.2.5. Processing time linearization

Eq. (27) can be reformulated to include a nonlinear term of the product of one integer variable, $CN_{s,i}$, and two continuous variables, T_s and TP_s. Two auxiliary variables are introduced: $\overline{TTP}_s \equiv T_s \cdot TP_s$ and $\overline{WTTP}_{s,i,j} \equiv W_{s,i,j} \cdot T_s \cdot TP_s$. $\overline{WTTP}_{s,i,j}$ can be determined by $\overline{TTP}_s$ using the following equations:

$$\sum_i \sum_{j=1}^{\max CN_s} j \cdot \overline{WTTP}_{s,i,j} = M_s, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{58}$$

$$\overline{WTTP}_{s,i,j} \leq \max TP_s \cdot \max T_s \cdot W_{s,i,j}, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, i, \ j = 1, \ldots, \max CN_s \tag{59}$$

$$\sum_i \sum_{j=1}^{\max CYN_s} j \cdot \overline{WTTP}_{s,i,j} = \overline{TTP}_s, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{60}$$

Next, $\overline{TTP}_s$ is reformulated using multiparametric disaggregation [44–46], where the throughput, TP_s, is expressed as a multiparametric sum of active decimal powers determined by binary variable $BTP_{s,d,q}$ and continuous variable $CTP_{s,q} \in [0, 1]$, where d is the digit position ranging from p to $\max p$, and q is the number for power d.

$$TP_s = \sum_{d=p}^{\max p} \sum_{q=0}^{9} 10^d \cdot q \cdot BTP_{s,d,q} + \sum_{q=0}^{1} 10^p \cdot q \cdot CTP_{s,q}, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{61}$$

$$\sum_{q=0}^{9} BTP_{s,d,q} = 1, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, \ d = p, \ldots, \max p \tag{62}$$

$$\sum_{q=0}^{1} CTP_{s,q} = 1, \ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{63}$$

Based on the above equations, variable $\overline{TTP}_s$ is disaggregated into a number of non-negative continuous variables $\overline{BTTP}_{s,d,q}$ and $\overline{CTTP}_{s,q}$:

$$\overline{TTP}_s = \sum_{d=p}^{\max p} \sum_{q=0}^{9} 10^d \cdot q \cdot \overline{BTTP}_{s,d,q} + \sum_{q=0}^{1} 10^p \cdot q \cdot \overline{CTTP}_{s,q}, \quad \forall s \in \{\text{af}, \text{ce}\} \tag{64}$$

$$\overline{BTTP}_{s,d,q} \le \max T_s \cdot BTP_{s,d,q}, \quad \forall s \in \{\text{af}, \text{ce}\}, d = p, \ldots, \max p,\ q = 0, \ldots, 9 \tag{65}$$

$$\overline{CTTP}_{s,q} \le \max T_s \cdot CTP_{s,q},\ \forall s \in \{\text{af}, \text{ce}\},\ q = 0, 1 \tag{66}$$

$$\sum_{q=0}^{9} \overline{BTTP}_{s,d,q} = T_s,\ \forall s \in \{\text{af}, \text{ce}\}, d = p, \ldots, \max p \tag{67}$$

$$\sum_{q=0}^{1} \overline{CTTP}_{s,q} = T_s,\ \forall s \in \{\text{af}, \text{ce}\} \tag{68}$$

Note that multiparametric disaggregation is a relaxation method, and the above reformulation provides a close approximation of the original equations, and generates a lower bound of the optimum.

By defining $\overline{WT}_{s,i,j} \equiv W_{s,i,j} \cdot PLT$ using the following constraints:

$$\overline{WT}_{\text{ae},i,j} \leqslant \max T_{\text{ae}} \cdot W_{\text{ae},i,j},\ \forall i,\ j = 1, \ldots, \max CN_{\text{ae}} \tag{69}$$

$$\sum_i \sum_{j=1}^{\max CN_{\text{ae}}} \overline{WT}_{\text{ae},i,j} = PLT \tag{70}$$

The following linear constraint replaces Eq. (29):

$$\frac{1}{1000} \cdot \sum_i \sum_{j=1}^{\max CN_{\text{ae}}} j \cdot vel \cdot \pi \cdot \left(\frac{dm_{\text{ae},i}}{2}\right)^2 \cdot \overline{WT}_{\text{ae},i,j} = PV_{\text{uf}_1} \tag{71}$$

Using another auxiliary variable, $\overline{XY}_{s,i,k} \equiv X_{s,i} \cdot Y_{s,k}$, and the following constraints are equivalent to Eq. (30):

$$BAT = \sum_i \sum_{k=1}^{\max CYN_{\text{ae}}} \frac{bcv_{\text{ae}} \cdot CV_{\text{ae},i} \cdot k \cdot \overline{XY}_{\text{ae},i,k}}{(1/1000) \cdot vel \cdot \pi \cdot (\text{dm}_{\text{ae},i}/2)^2} \tag{72}$$

$$\sum_i \overline{XY}_{\text{ae},i,k} = Y_{\text{ae},k},\ \forall k = 1, \ldots, \max CYN_{\text{ae}} \tag{73}$$

$$\sum_{k=1}^{\max CYN_{\text{ae}}} \overline{XY}_{\text{ae},i,k} = X_{\text{ae},i}, \forall i \tag{74}$$

Based on Eq. (44), Eq. (34) can be rewritten as follows:

$$AT = \sum_{n=1}^{\lceil \log_2 \max BN \rceil} 2^{n-1} \cdot \overline{ZT}_n \tag{75}$$

where $\overline{ZT}_n \equiv Z_n \cdot BT$, satisfying:

$$\overline{ZT}_n \le (aot - st - brt) \cdot Z_n,\ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{76}$$

$$\overline{ZT}_n \le BT,\ \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{77}$$

$$\overline{ZT}_n \ge BT - (aot - st - brt) \cdot (1 - Z_n), \quad \forall n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{78}$$

4.2.6. Data-driven model reformulation

The bilinear terms of the regression models in Eqs. (4) and (35) can be rewritten as the following linear constraints [47]:

$$bp_{s,r-1} + \varepsilon - \max LM_s \cdot (1 - O_{s,r}) \le LM_s^1 \le bp_{s,r} + \max LM_s \cdot (1 - O_{s,r}),\ \forall s \in \{\text{af}, \text{ce}\}, r \tag{79}$$

$$\beta_{s,r}^{LM} \cdot LM_s^1 + \beta_{s,r}^{V} \cdot V_s + \beta_{s,r}^{H} \cdot \sum_i h_{s,i} \cdot X_{s,i} + \beta_{s,r}^{0} - \max TP_s \cdot (1 - O_{s,r}) \le TP_s^1 \le \beta_{s,r}^{LM} \cdot LM_s^1 + \beta_{s,r}^{V} \cdot V_s + \beta_{s,r}^{H} \cdot \sum_i h_{s,i} \cdot X_{s,i} + \beta_{s,r}^{0} + \max TP_s \cdot (1 - O_{s,r}), \quad \forall s \in \{\text{af}, \text{ce}\}, r \tag{80}$$

Eq. (36) can be linearized using a new auxiliary variable, $\overline{XTP}_{s,i} \equiv X_{s,i} \cdot TP_s^1$, and the following constraints:

$$TP_s = \frac{\sum_i (dm_{s,i})^2 \cdot \overline{XTP}_{s,i}}{(\text{ref}DM)^2},\ \forall s \in \{\text{af}, \text{ce}\} \tag{81}$$

$$\overline{XTP}_{s,i} \le \max TP_s \cdot X_{s,i},\ \forall s \in \{\text{af}, \text{ce}\}, i \tag{82}$$

$$\sum_i \overline{XTP}_{s,i} = TP_s^1,\ \forall s \in \{\text{af}, \text{ce}\} \tag{83}$$

Similarly, using the auxiliary variable $\overline{XLM}_{s,i} \equiv X_{s,i} \cdot LM_s^1$, Eq. (37) is equivalent to:

$$LM_s = \frac{\sum_i (dm_{s,i})^2 \cdot \overline{XLM}_{s,i}}{(\text{ref}DM)^2},\ \forall s \in \{\text{af}, \text{ce}\} \tag{84}$$

$$\overline{XLM}_{s,i} \le \max LM_s \cdot X_{s,i},\ \forall s \in \{\text{af}, \text{ce}\}, i \tag{85}$$

$$\sum_i \overline{XLM}_{s,i} = LM_s^1,\ \forall s \in \{\text{af}, \text{ce}\} \tag{86}$$

Eq. (38) includes a nonlinear term involving two integer variables and one continuous variable, which can be linearized as follows:

$$\sum_{j=1}^{\max CN_s} \sum_{k=1}^{\max CYN_s} j \cdot k \cdot \overline{WYLM}_{s,i,j,k} = M_{s-1},\ \forall s \in \{\text{af}, \text{ce}\},\ i \tag{87}$$

$$\sum_{k=1}^{\max CYN_s} \overline{WY}_{s,i,j,k} = W_{s,i,j},\ \forall s \in \{\text{af}, \text{ce}\}, i,\ j = 1, \ldots, \max CN_s \tag{88}$$

$$\sum_i \sum_{j=1}^{\max CN_s} \overline{WY}_{s,i,j,k} = Y_{s,k},\ \forall s \in \{\text{af}, \text{ce}\},\ k = 1, \ldots, \max CYN_s \tag{89}$$

$$\overline{WYLM}_{s,i,j,k} \le \max LM_s \cdot \overline{WY}_{s,i,j,k}, \quad \forall s \in \{\text{af}, \text{ce}\}, i,\ j = 1, \ldots, \max CN_s, k = 1, \ldots, \max CYN_s \tag{90}$$

$$\sum_i \sum_{j=1}^{\max CN_s} \sum_{k=1}^{\max CYN_s} \overline{WYLM}_{s,i,j,k} = LM_s,\ \forall s \in \{\text{af}, \text{ce}\} \tag{91}$$

where there are two auxiliary variables, $\overline{WY}_{s,i,j,k} \equiv W_{s,i,j} \cdot Y_{s,k}$, and $\overline{WYLM}_{s,i,j,k} \equiv W_{s,i,j} \cdot Y_{s,k} \cdot LM_s$.

4.2.7. Objective function linearization

The calculation of consumables cost in Eq. (S9) in Supplementary data involves nonlinearities. By introducing auxiliary variable $\overline{ZYV}_{s,k,n} \equiv Z_n \cdot \overline{YV}_{s,k}$, Eq. (S9) can be reformulated as follows:

$$CC = \sum_{s \in CS} \sum_{n=1}^{\log_2 \max BN} \sum_{k=1}^{\max CYN_s} \frac{2^{n-1} \cdot k \cdot of \cdot rpc_s \cdot \overline{ZYV}_{s,k,n}}{l_s} \tag{92}$$

where *of* refers to the resin overpacking factor.

$$\overline{ZYV}_{s,k,n} \le \max TCV_s \cdot Z_n,\ \forall s \in CS, k = 1, \ldots, \max CYN_s, \quad n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{93}$$

$$\overline{ZYV}_{s,k,n} \le \overline{YV}_{s,k},\ \forall s \in CS, k = 1, \ldots, \max CYN_s, \quad n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{94}$$

$$\overline{ZYV}_{s,k,n} \geq \overline{YV}_{s,k} - \max TCV_s \cdot (1 - Z_n),\ \forall s \in CS,\ k = 1, \ldots, \max CYN_s, n = 1, \ldots, \lceil \log_2 \max BN \rceil \tag{95}$$

where rpc_s and l_s are the resin price and resin lifetime at chromatography step s, respectively.

In the objective function Eq. (39), the annual cost can be expressed as the product of COG per gram and the annual production, which can be written as follows:

$$\overline{COGAP} = AC \tag{96}$$

where auxiliary variable $\overline{COGAP} \equiv COG \cdot AP$. Using multiparametric disaggregation and introducing new auxiliary variables $\overline{BCOGAP}_{d,q} \equiv COG \cdot BAP_{d,q}$ and $\overline{CCOGAP}_q \equiv COG \cdot CAP_q$, the variables AP and $\overline{COGAP}$ can be disaggregated as follows:

$$AP = \sum_{d=p}^{\max p} \sum_{q=0}^{9} 10^d \cdot q \cdot BAP_{d,q} + \sum_{q=0}^{1} 10^p \cdot q \cdot CAP_q \tag{97}$$

$$\overline{COGAP} = \sum_{d=p}^{\max p} \sum_{q=0}^{9} 10^d \cdot q \cdot \overline{BCOGAP}_{d,q} + \sum_{q=0}^{1} 10^p \cdot q \cdot \overline{CCOGAP}_q \tag{98}$$

$$\sum_{q=0}^{9} BAP_{d,q} = 1,\ \forall d = p, \ldots, \max p \tag{99}$$

$$\sum_{q=0}^{1} CAP_q = 1 \tag{100}$$

$$\overline{BCOGAP}_{d,q} \leq \max COG \cdot BAP_{d,q},\ \forall d = p, \ldots, \max p, q = 0, \ldots, 9 \tag{101}$$

$$\overline{CCOGAP}_q \leq \max COG \cdot CAP_q,\ \forall q = 0, 1 \tag{102}$$

$$\sum_{q=0}^{9} \overline{BCOGAP}_{d,q} = COG,\ \forall d = p, \ldots, \max p \tag{103}$$

$$\sum_{q=0}^{1} \overline{CCOGAP}_q = COG \tag{104}$$

Thus, the reformulated MILP model A* includes the constraints shown in Eqs. (3), (7–9), (11–13), (15), (16), (18–21), (23–25), (28), (32), (33), and (40–104); and Eqs. (S1–S8) and (S10–S19) in Supplementary data, with the variable COG as the objective.

4.3. MINLP model B

In this section, an alternative MINLP model B is introduced. By generating a set of column volume sizes, the models A and A* result in a large number of variables and equations, which hinder the computation of the proposed models. To overcome this shortcoming, in the new MINLP model B, we get rid of the discrete column volume sizes, and introduce an integer variable, H_s, for the bed height, and a binary variable, $E_{s,m}$, to indicate whether diameter size, $\widetilde{dm}_m$, is selected. In addition, the subscript i is removed from the chromatography column number variable, and variable $\widetilde{CN}_s$ expresses the number of columns at chromatography step s, which is upper bounded by $\max CN_s$. Thus, the number of discrete variables in model B is reduced with improved computational efficiency. Based on the proposed MINLP model A presented above, a number of new variables and constraints are developed for the MINLP model B, as introduced below.

4.3.1. Column volume

In model B, the total column volume is calculated using the variables of bed height (H_s), diameter selection ($E_{s,m}$), and number of columns ($\widetilde{CN}_s$).

$$TCV_s = \frac{1}{1000} \cdot \pi \cdot \sum_m \left(\frac{\widetilde{dm}_m}{2} \right)^2 \cdot E_{s,m} \cdot H_s \cdot \widetilde{CN}_s,\ \forall s \in CS \tag{105}$$

In addition, only one diameter size can be selected at each chromatography step.

$$\sum_m E_{s,m} = 1,\ \forall s \in CS \tag{106}$$

4.3.2. Processing time

By replacing $\sum_i CN_{s,i}$ by $\widetilde{CN}_s$ in Eqs. (27) and (29), the following constraint can be obtained:

$$T_s \cdot TP_s \cdot \widetilde{CN}_s = M_s,\ \forall s \in \{\text{af}, \text{ce}\} \tag{107}$$

$$PLT = \frac{PV_{\text{uf}_1}}{VFR \cdot \widetilde{CN}_{\text{ae}}} \tag{108}$$

Similarly, BAT and VFR are formulated using the new integer variable, H_s, and the binary variable, $E_{s,m}$:

$$BAT = \frac{CYN_{\text{ae}} \cdot bcv_{\text{ae}} \cdot (1/1000) \cdot \pi \cdot \sum_m \left(\widetilde{dm}_m / 2 \right)^2 \cdot E_{\text{ae},m} \cdot H_{\text{ae}}}{VFR} \tag{109}$$

$$VFR = \frac{1}{1000} \cdot vel \cdot \pi \cdot \sum_m \left(\frac{\widetilde{dm}_m}{2} \right)^2 \cdot E_{\text{ae},m} \tag{110}$$

4.3.3. Data-driven model

Eq. (38) can be written using the new column number variable as follows:

$$LM_s = \frac{M_{s-1}}{CYN_s \cdot \widetilde{CN}_s},\ \forall s \in \{\text{af}, \text{ce}\} \tag{111}$$

Similar to Eqs. (36) and (37), the throughput and loaded mass of the selected columns can be calculated from those of the 1 m-diameter columns.

$$TP_s = \frac{\sum_m \left(\widetilde{dm}_m \right)^2 \cdot E_{s,m}}{(\text{ref}DM)^2} \cdot TP_s^1,\ \forall s \in \{\text{af}, \text{ce}\} \tag{112}$$

$$LM_s = \frac{\sum_m \left(\widetilde{dm}_m \right)^2 \cdot E_{s,m}}{(\text{ref}DM)^2} \cdot LM_s^1,\ \forall s \in \{\text{af}, \text{ce}\} \tag{113}$$

4.3.4. Cost

In the cost calculation, Eq. (S13) for fixed capital investment (FCI) in Supplementary data should be replaced by the following nonlinear constraint:

$$FCI = lang \cdot (1 + gef) \cdot \left(brc \cdot brn + \sum_{s \in CS} \sum_m \widetilde{cc}_{s,m} \cdot \widetilde{CN}_s \cdot E_{s,m} + oe\lambda \cdot brc \cdot brn \right) \tag{114}$$

where $lang$ is the Lang factor; gef is the general equipment factor; brn is the number of bioreactors; brc is the cost of one bioreactor; $oe\lambda$ is the ratio of other equipment cost to the bioreactor cost; $\widetilde{cc}_{s,m}$ is the chromatography column cost of diameter size m at chromatography step s.

Thus, the proposed MINLP model B minimizes Eq. (39), subject to the constraints shown in Eqs. (3–5), (10–26), (28), (32–34), and (105–114); and Eqs. (S1–S12) and (S14–S19) in Supplementary data.

4.4. MILP model B*

In MILP model B*, all nonlinear constraints of MINLP model B are linearized. Besides the linear constraints presented in MILP model A*, the newly developed ones are given below.

4.4.1. Integer variable discretization

At first, integer variable $\widetilde{CN}_s$ is discretized using binary variable $F_{s,j}$, which represents whether or not j columns are used at chromatography step s, as follows:

$$\widetilde{CN}_s = \sum_{j=1}^{\max CN_s} j \cdot F_{s,j},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{115}$$

$$\sum_{j=1}^{\max CN_s} F_{s,j} = 1,\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{116}$$

4.4.2. Column volume linearization

In order to linearize Eq. (105), an auxiliary variable, $\overline{EFH}_{s,m,j} \equiv E_{s,m} \cdot F_{s,j} \cdot H_s$, is introduced, with the following constraints:

$$\overline{EFH}_{s,m,j} \le \max H_s \cdot \overline{EF}_{s,m,j},\ \forall s \in CS, m,\ j = 1, \ldots, \max CN_s \tag{117}$$

$$\sum_m \sum_{j=1}^{\max CN_s} \overline{EFH}_{s,m,j} = H_s,\ \forall s \in CS \tag{118}$$

where $\overline{EF}_{s,m,j} \equiv E_{s,m} \cdot F_{s,j}$ is defined as follows:

$$\sum_m \overline{EF}_{s,m,j} = F_{s,j},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\},\ j = 1, \ldots, \max CN_s \tag{119}$$

$$\sum_{j=1}^{\max CN_s} \overline{EF}_{s,m,j} = E_{s,m},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, m \tag{120}$$

Therefore, Eq. (105) is reformulated to the following linear constraint:

$$TCV_s = \sum_m \sum_{j=1}^{\max CN_s} j \cdot \pi \cdot \left(\frac{\widetilde{dm}_m}{2}\right)^2 \cdot \overline{EFH}_{s,m,j},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{121}$$

4.4.3. Processing time linearization

According to Eqs. (61–68), the term $T_s \cdot TP_s$ can be expressed by $\overline{TTP}_s$. Therefore, $\overline{FTTP}_{s,j}$ is introduced to express $F_{s,j} \cdot \overline{TTP}_s$, with the following constraints:

$$\overline{FTTP}_{s,j} \le \max TP_s \cdot \max T_s \cdot F_{s,j},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\},\ j = 1, \ldots, \max CN_s \tag{122}$$

$$\sum_{j=1}^{\max CN_s} j \cdot \overline{FTTP}_{s,j} = \overline{TTP}_s,\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{123}$$

Thus, Eq. (107) is rewritten as follows:

$$\sum_{j=1}^{\max CN_s} j \cdot \overline{FTTP}_{s,j} = M_s,\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{124}$$

Based on the introduction of auxiliary variable $\overline{EFT}_{s,m,j} \equiv \overline{EF}_{s,m,j} \cdot PLT$ and Eqs. (119) and (120), Eq. (108) can be linearized as follows:

$$\frac{1}{1000} \cdot \sum_m \sum_{j=1}^{\max CN_{\mathrm{ae}}} j \cdot vel \cdot \pi \cdot \left(\frac{\tilde{dm}_m}{2}\right)^2 \cdot \overline{EFT}_{\mathrm{ae},m,j} = PV_{\mathrm{uf}_1} \tag{125}$$

$$\overline{EFT}_{\mathrm{ae},m,j} \le \max T_{\mathrm{ae}} \cdot \overline{EF}_{\mathrm{ae},m,j},\ \forall m,\ j = 1, \ldots, \max CN_{\mathrm{ae}} \tag{126}$$

$$\sum_m \sum_{j=1}^{\max CN_{\mathrm{ae}}} \overline{EFT}_{\mathrm{ae},m,j} = PLT \tag{127}$$

In addition, with $\overline{YH}_{s,k} \equiv Y_{s,k} \cdot H_s$, the following constraints can replace Eq. (109) in the MILP model:

$$BAT = \sum_{k=1}^{\max CYN_{\mathrm{ae}}} \frac{bc\, v_{\mathrm{ae}} \cdot k \cdot \overline{YH}_{\mathrm{ae},k}}{vel} \tag{128}$$

$$\overline{YH}_{\mathrm{ae},k} \le \max H_{\mathrm{ae}} \cdot Y_{\mathrm{ae},k},\ \forall k = 1, \ldots, \max CYN_s \tag{129}$$

$$\sum_{k=1}^{\max CYN_s} \overline{YH}_{\mathrm{ae},k} = H_{\mathrm{ae}} \tag{130}$$

4.4.4. Data-driven model reformulation

Similar to Eq. (80), the data-driven model constraint given in Eq. (5) can be reformulated to linear constraints [47], as follows:

$$\begin{aligned} &\beta_{s,r}^{LM} \cdot LM_s^1 + \beta_{s,r}^{V} \cdot V_s + \beta_{s,r}^{H} \cdot H_s + \beta_{s,r}^{0} - \max TP_s \cdot (1 - O_{s,r}) \\ &\le TP_s^1 \le \beta_{s,r}^{LM} \cdot LM_s^1 + \beta_{s,r}^{V} \cdot V_s + \beta_{s,r}^{H} \cdot H_s + \beta_{s,r}^{0} + \\ &\max TP_s \cdot (1 - O_{s,r}),\ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, r \end{aligned} \tag{131}$$

Eq. (111) is linearized to Eq. (132) by introducing auxiliary variables $\overline{FYLM}_{s,j,k} \equiv F_{s,j} \cdot Y_{s,k} \cdot LM_s$ and $\overline{FY}_{s,j,k} \equiv F_{s,j} \cdot Y_{s,k}$.

$$\sum_{j=1}^{\max CN_s} \sum_{k=1}^{\max CYN_s} j \cdot k \cdot \overline{FYLM}_{s,j,k} = M_{s-1},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{132}$$

$$\sum_{k=1}^{\max CYN_s} \overline{FY}_{s,j,k} = F_{s,j},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\},\ j = 1, \ldots, \max CN_s \tag{133}$$

$$\sum_{j=1}^{\max CN_s} \overline{FY}_{s,j,k} = Y_{s,k},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, k = 1, \ldots, \max CYN_s \tag{134}$$

$$\begin{aligned} &\overline{FYLM}_{s,j,k} = \max LM_s \cdot \overline{FY}_{s,j,k},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, \\ &j = 1, \ldots, \max CN_s, k = 1, \ldots, \max CYN_s \end{aligned} \tag{135}$$

$$\sum_{j=1}^{\max CN_s} \sum_{k=1}^{\max CYN_s} \overline{FYLM}_{s,j,k} = LM_s,\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{136}$$

In Eqs. (112) and (113), the nonlinearities involve the product of $E_{s,m}$ and a continuous variable. With the introduction of the auxiliary variables, $\overline{ETP}_{s,m}$ and $\overline{ELM}_{s,m}$, the following linear equations are used instead as linear constraints:

$$TP_s = \frac{\sum_m \left(\widetilde{dm}_m\right)^2 \cdot \overline{ETP}_{s,m}}{(\mathrm{ref}DM)^2},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{137}$$

$$\overline{ETP}_{s,m} \le \max TP_s \cdot E_{s,m},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\}, m \tag{138}$$

$$\sum_m \overline{ETP}_{s,m} = TP_s^1,\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{139}$$

$$LM_s = \frac{\sum_m \left(\widetilde{dm}_m\right)^2 \cdot \overline{ELM}_{s,m}}{(\mathrm{ref}DM)^2},\ \forall s \in \{\mathrm{af}, \mathrm{ce}\} \tag{140}$$

$$\overline{ELM}_{s,m} \leq \max CLM_s \cdot E_{s,m},\ \forall s \in \{af, ce\}, m \quad (141)$$

$$\sum_m \overline{ELM}_{s,m} = LM_s^1,\ \forall s \in \{af, ce\} \quad (142)$$

4.4.5. Cost linearization

At last, in order to reformulate the fixed capital investment formula, $\overline{EF}_{s,m,j}$ can be used to linearize Eq. (114) as follows:

$$FCI = lang \cdot (1 + gef) \cdot \left(brc \cdot brn + \sum_{s \in CS} \sum_m \sum_{j=1}^{\max CN_s} \widetilde{cc}_{s,m} \cdot k \cdot \overline{EF}_{s,m,j} + oe\lambda \cdot brc \cdot brn \right) \quad (143)$$

Overall, MILP model B* includes constraints shown in Eqs. (3), (11–13), (15), (16), (18–21), (23–25), (28), (32), (33), (42–57), (61–68), (75–79), (92–104), (106), and (115–143); and Eqs. (S1–S8), (S10–S12), and (S14–S19) in Supplementary data.

In summary, the equations of all four proposed models are summarized in Table 2.

5. Case study

In this section, the four proposed optimization models are applied to industrially relevant case studies to examine their performances. The process flowsheet of this example is shown in Fig. 1, where there are one bioreactor and three chromatography steps: affinity, cation-exchange, and anion-exchange chromatography steps. For the chromatography column-sizing decisions, the number of columns utilized at each chromatography step is limited to four, while a maximum of 10 cycles are allowed in each batch. Two cases with different alternatives of chromatography column diameter and bed height are considered. Case 1 includes 10 alternative discrete diameters varying from 50 to 200 cm and 11 alternative discrete bed height between 15 and 25 cm, while Case 2 has 26 alternative discrete diameters between 50 and 300 cm and 21 alternative discrete bed height taking integer values ranging from 10 to 30 cm. In models A and A*, where discrete column volume sizes are used, there are 110 alternatives in Case 1 and 546 alternatives in Case 2. The detailed alternative column diameters and bed height are given in Table 3.

The chromatography resin utilization factor μ is 0.95. The linear velocity of the flows at the affinity and cation-exchange chromatography steps is limited between 200 and 600 cm·h^{-1}, while the linear velocity of the flows at the anion-exchange chromatography step is fixed at 300 cm·h^{-1}. Other parameters for the three chromatography steps are given in Table 4.

The *aot* of the process is 340 d, and the batch success rate σ is 90%. Process parameters of the non-chromatography steps are provided in Table 5. The cost-relevant data can be found in Supplementary data (Table S1).

The proposed optimization models were implemented in GAMS 24.7 [42] on a 64-bit Windows 7-based machine with an Intel Core i5-3330 3.00 GHz processor and 8.0 GB RAM, using BARON as the MINLP solver and CPLEX as the MILP solver. The central processing unit (CPU) time for each model was limited to 1 h.

6. Results and discussion

The proposed models were applied to the above two cases with different column size alternatives. The computational results are presented and discussed in this section.

6.1. Case 1

The model statistics and computational results of the four proposed models for Case 1 are presented in Table 6, in which all four models reach their global optimal solutions. Note that although the reported optimal objectives of all four models are the same to the first decimal place, the solutions of the MILP models are actually very close approximations of the global optima of the correspond-

Table 2
Model summary.

Constraints	MINLP A	MILP A*	MINLP B	MILP B*
Integer variable discretization	—	Eqs. (40–44)	—	Eqs. (42–44), (115), (116)
Column volume	Eqs. (7–11)	Eqs. (7–9), (11), (45–47)	Eqs. (10), (11), (105), (106)	Eqs. (11), (45–47), (106), (117–121)
Product mass	Eqs. (12–14)	Eqs. (12), (13), (48–51)	Eqs. (12–14)	Eqs. (12), (13), (48–51)
Product volume	Eqs. (15–19)	Eqs. (15), (16), (18), (19), (52)	Eqs. (15–19)	Eqs. (15), (16), (18), (19), (52)
Buffer volume	Eqs. (20–26)	Eqs. (20), (21), (23–25), (53–57)	Eqs. (20–26)	Eqs. (20), (21), (23–25), (53–57)
Processing time	Eqs. (27–34)	Eqs. (28), (32), (33), (58–78)	Eqs. (28), (32–34), (107–110)	Eqs. (28), (32), (33), (61–68), (75–78), (122–130)
Data-driven model	Eqs. (3), (4), (35–38)	Eqs. (3), (79–91)	Eqs. (3–5), (111–113)	Eqs. (3), (79), (131–142)
Cost and objective	Eqs. (39), (S1–S19)	Eqs. (92–104), (S1–S8), (S10–S19)	Eqs. (114), (S1–S12), (S14–S19)	Eqs. (92–104), (143), (S1–S8), (S10–S12), (S14–S19)

Table 3
Alternatives of chromatography column sizes in two cases.

Step	Case 1	Case 2
Diameter (cm)	50, 60, 70, 80, 90, 100, 120, 160, 180, 200	50 to 300 with a step size of 10
Bed height (cm)	15 to 25 with a step size of 1	10 to 30 by with a step size of 1

Table 4
Parameters of chromatographic operations.

Step	Affinity chromatography	Cation-exchange chromatography	Anion-exchange chromatography
Yield, yd_s	95%	95%	98%
Dynamic binding capacity, dbc_s (g·L^{-1})	15	50	100
Eluate volume to column volume ratio, ecv_s	2	3	0
Buffer volume to column volume ratio, bcv_s	15	22	18
Resin lifetime, l_s (cycle)	100	100	100
Resin price, rpc_s (GBP·L^{-1})	60 000	1 200	1 400

Table 5
Parameters of non-chromatographic unit operations.

Unit operation parameter	Value
Cell culture	
Titer, *titer* ($g \cdot L^{-1}$)	2
Bioreaction time, *brt* (d)	4
Seed train bioreaction time, *st* (d)	2
Number of bioreactors, *brn*	1
Bioreactor working volume ratio, α	75%
Bioreactor volume, *brv* (L)	20 000
Centrifugation 1	
Yield, yd_s	95%
Buffer volume ratio, bvr_s	0.9
Processing rate, pr_s ($L \cdot h^{-1}$)	250
Homogenization	
Yield, yd_s	100%
Processing rate, pr_s ($L \cdot h^{-1}$)	143
Centrifugation 2	
Yield, yd_s	95%
Buffer volume ratio, bvr_s	0.1
Processing rate, pr_s ($L \cdot h^{-1}$)	250
Filtration	
Yield, yd_s	98%
Processing rate, pr_s ($L \cdot h^{-1}$)	125
UF/DF 1	
Yield, yd_s	98%
Flush volume ratio, *fvr*	6
Processing rate, pr_s ($L \cdot h^{-1}$)	40
UF/DF 2	
Yield, yd_s	98%
Defiltration volume ratio, *dvr*	7
Processing rate, pr_s ($L \cdot h^{-1}$)	40
Bulk fill	
Yield, yd_s	98%
Filling time, T_s (h)	6
Final concentration, *fconc* ($g \cdot L^{-1}$)	10

ing MINLP models with the same column sizing and key operational decisions. They are about 1×10^{-4} lower than the optimal objective values of the MINLP models, being lower bounds as indicated by the theory of multiparametric disaggregation. The MINLP model A is able to achieve the optimum of 201.7 GBP·g^{-1} within 494 s. After linearization, the MILP model A* takes a slightly shorter time (381 s) to find the optimal solution, despite having an increased number of equations and variables in the model. Meanwhile, as shown in Table 6, the MINLP model B has the same continuous variables as the MINLP model A, but significantly reduces the number of equations and discrete variables by one order of magnitude. As a result, the model is able to find the optimal solution within 47 s. The MINLP model B*, which results from linearizing the MINLP model B, involves much fewer equations and variables than the MILP model A*, and takes only 5 s to find the optimal solution, saving CPU by two orders of magnitude. From the comparison, it can be concluded that among the four proposed models, the models B and B* show obvious computational advantages over the other two models. In particular, the MILP model B* requires the smallest computational effort and has the most potential for larger scale problems, which also becomes evident from the larger example (Case 2) discussed in the next subsection. Note that the predication errors of the resulting piecewise linear regression models have very minor effects on the optimal objective values of the optimization models; for example, a 17% shift of the estimated throughput output results in only a 0.1% difference in the optimal objective value. A similar observation is made for Case 2.

Next, the detailed optimal solutions of Case 1 are discussed. The optimal chromatography column-sizing strategies are given in Fig. 3, where the diameters of the chromatography columns are proportional to the widths of the shapes plotted, while the bed heights are proportional to the shapes' heights. The dashed-line shapes represent the cycles needed in each batch. At each chromatography step, only one column is utilized. The column at the affinity chromatography step has a diameter of 180 cm and a bed height of 15 cm, while the cation-exchange chromatography step uses a smaller column with a diameter of 90 cm and a bed height of 21 cm. The chromatography column at the anion-exchange chromatography step has the smallest diameter, 80 cm, but the largest bed height, 25 cm.

Here, the performance of the throughput regression models and operational decisions are examined. In the optimal solution, the throughput output of the metamodel is 1869.8 $g \cdot h^{-1}$ at the affinity chromatography step and 1823.9 $g \cdot h^{-1}$ at the cation-exchange chromatography step. As shown in Fig. 3, five and four cycles per batch are required at these two chromatography steps, respectively. The resulting product masses loaded to each column in each cycle are 5306.7 and 6301.7 g, receptively. After converting to the values for a 1 m-diameter column, the loaded mass falls into the first interval in the piecewise linear regression model at the affinity chromatography step, and the corresponding function is used to estimate the throughput, as given below:

$$TP_{af} = 0.1914 \cdot LM_{af} + 1.8^2 \cdot (0.3570 \cdot V_{af} - 12.0477 \cdot H_{af} + 230.131) \quad (146)$$

where 1.8^2 is added to convert the performance of the 1 m-diameter column to that of the selected 1.8 m-diameter column, and $LM_{af} = 1.8^2 \cdot LM^1_{af}$, according to Eq. (37). At the cation-exchange chromatography step, the first interval in the regression model is also selected. Similarly, the regression model used is as follows:

$$TP_{ce} = 0.1287 \cdot LM_{ce} + 0.9^2 \cdot (2.3940 \cdot V_{ce} - 51.4883 \cdot H_{ce} + 895.2814) \quad (147)$$

In both of the above functions, the linear velocity variable contributes to the throughput positively; therefore, the linear velocity of the flows at both steps reaches its upper bound, 600 $cm \cdot h^{-1}$.

At last, from the optimal cost distribution shown in Fig. 4, the consumables cost—that is, the resin cost in this problem—represents the largest portion, over 30%, due to the high price of the affinity resin used. Also, the capital cost of the equipment, chemical reagents (buffer and media) cost, and labor cost all make a significant contribution to the total cost.

6.2. Case 2

In Case 2, a larger number of alternative diameter and bed height are considered, resulting in larger scale optimization models, as indicated in Table 7. Compared with Case 1, MINLP model B has slightly more discrete variables, while the other models involve an increased number of both equations and variables. It is worth noting that models A and A* fail to terminate before the computational time limit of 3600 s, although the gaps to the best possible

Table 6
Statistics and computational performance of the proposed models for Case 1.

Model	No. of equations	No. of continuous variables	No. of discrete variables	Optimal objective (GBP·g^{-1})	CPU (s)
MINLP A	423	85	670	201.7	494
MILP A*	13 646	11 092	12 097	201.7	381
MINLP B	93	85	46	201.7	47
MILP B*	1 508	700	595	201.7	5

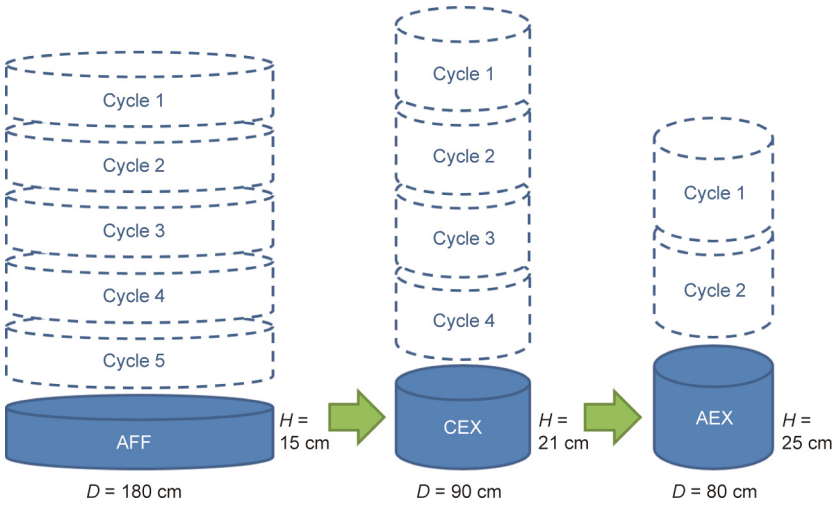

Fig. 3. Optimal chromatography column-sizing strategies of Case 1. AFF: affinity chromatography; CEX: cation-exchange chromatography; AEX: anion-exchange chromatography; *D*: diameter; *H*: bed height.

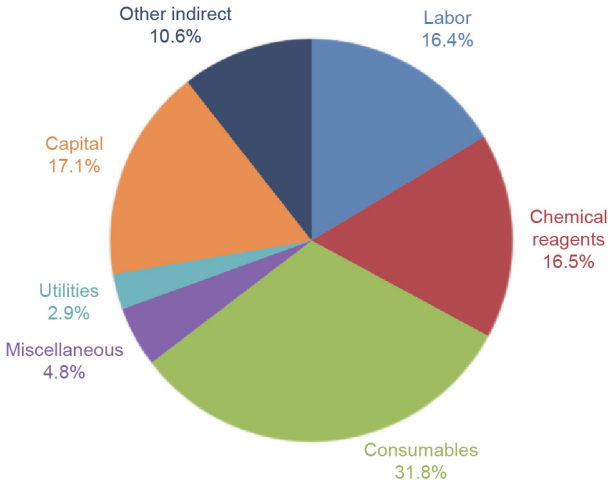

Fig. 4. Optimal cost distribution of Case 1.

solutions are only 0.6% and 0.4%, respectively. According to the solution process of the two models shown in Fig. 5, the MILP model A* finds a good feasible solution at around 220 s, and actually achieves the optimum within 10 min. However, the lower bound of the solution in the branch and bound process converges so slowly that the optimum of the obtained objective, 200.3 GBP·g^{-1}cannot be proven within the given time limit. Meanwhile, the MINLP model A is relatively much slower, only reaches the first feasible solution after around 1000 s, and obtains a good feasible solution at nearly 30 min. Compared with the models A and A*, models B and B* show significantly improved computational performance and achieve the optimal solutions within 4 min. The MINLP model B takes around 1 min for a close feasible solution and 192 s for the optimum. The MILP model B* achieves a CPU saving of one order of magnitude, with only 24 s for its optimum, which is also a very close lower bound of the MINLP model B* due to multiparametric disaggregation. For a good feasible solution within 1% of the optimum, only 6 s is needed for the model B*. Thus, the computational advantage of the MILP model B* is demonstrated.

Regarding the column-sizing decisions, Case 2 has more possible column size options. Fig. 6 shows that in comparison with the optimal decisions of Case 1, a column with a larger diameter but smaller bed height is installed at the first two chromatography steps. The selected bed heights (11 and 14 cm) are beyond the bed height range allowed in Case 1 (15–25 cm). In addition, the selected diameters (190 and 110 cm) are not available in Case 1. With more possible alternatives, the optimal solution of Case 1 is only a feasible solution of Case 2, and there is an improvement of 1.4 GBP·g^{-1} in the optimal COG per gram of Case 2. Meanwhile, the same column (80 cm diameter and 25 cm bed height) is selected at the anion-exchange chromatography step. The selected larger diameter columns result in higher cost in equipment investment, while the smaller bed heights lead to less resin and relevant cost. The differences in these costs are relatively very small; therefore, the cost distribution is very similar to that of Case 1, which is not discussed further here.

At both the affinity and cation chromatography steps, the highest allowed flow velocity, 600 cm·h^{-1}, is implemented. The actual throughput regression models at the two steps are slightly different from those in Case 1, due to the differences in the selected diameter sizes:

$$TP_{af} = 0.1914 \cdot LM_{af} + 1.9^{2} \cdot (0.3570 \cdot V_{af} - 12.0477 \cdot H_{af} + 230.131) \quad (148)$$

Table 7
Statistics and computational performance of the proposed models for Case 2.

Model	No. of equations	No. of continuous variables	No. of discrete variables	Optimal objective (GBP·g^{-1})	CPU (s)
MINLP A	1 731	85	3 286	200.5[a]	3600[a]
MILP A*	63 350	52 948	59 185	200.3[b]	3600[b]
MINLP B	93	85	94	200.3	192
MILP B*	1 972	828	1 027	200.3	24

[a] Obtained solution has an optimility gap of 0.6% when the CPU limit is reached.
[b] Obtained solution has an optimility gap of 0.4% when the CPU limit is reached.

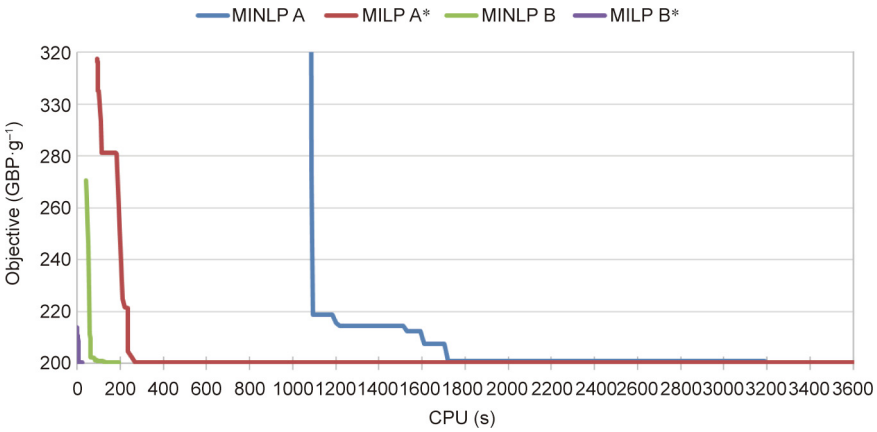

Fig. 5. Solution process of the four proposed models for Case 2.

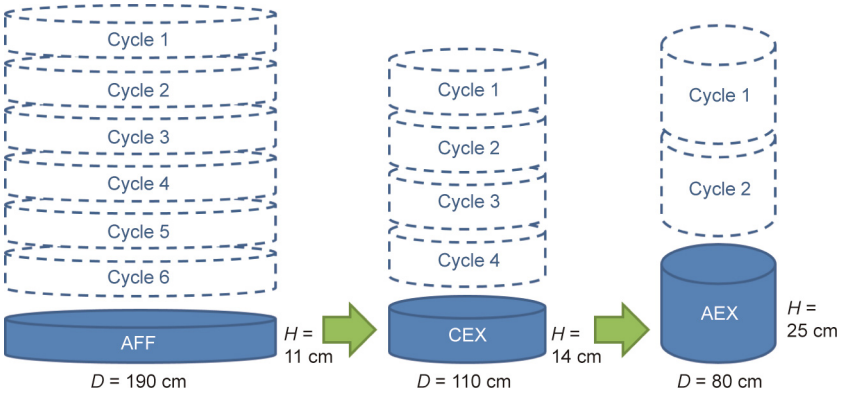

Fig. 6. Optimal chromatography column-sizing strategies for Case 2.

$$TP_{ce} = 0.1287 \cdot LM_{ce} + 1.1^{2} \cdot (2.3940 \cdot V_{ce} - 51.4883 \cdot H_{ce} + 895.2814) \quad (149)$$

As shown in Fig. 6, one more cycle is used at the affinity chromatography step than in Case 1, and therefore less mass (4422.3 g) is loaded in each cycle. However, due to the smaller bed height and larger diameter, the throughput increases to 1972.0 $g \cdot h^{-1}$. For the cation-exchange chromatography step, although the same number of cycles and loaded mass as Case 1 are obtained, a higher throughput of 2760.2 $g \cdot h^{-1}$ is achieved, due to the chosen larger diameter and smaller bed height.

7. Concluding remarks

In this work, the multiscale optimization of an antibody manufacturing process has been investigated. At the operational level, to mimic the complex behavior of the chromatography process, data-driven models were developed to estimate the chromatography throughput, using manufacturing-scale simulated datasets based on microscale experimental data. Through a comparison of a number of methods for metamodeling, piecewise linear regression models were developed based on the simulated datasets.

At the process design level, in order to determine the optimal chromatography column-sizing strategies, two alterative MINLP models were proposed to minimize the COG per gram. Adopting linearization techniques, two MILP models were developed. The throughput regression models were incorporated into the optimization models to determine the optimal operational decisions—that is, the flow velocity and the number of cycles per batch.

By studying two industrially relevant cases with different column size alternatives, the computational performance of the four proposed optimization models were compared. In conclusion, models B and B* demonstrated more efficient computational performances. In particular, the second MILP model was shown to be the most computationally efficient, and can thus be recommended for large-scale optimization studies. In addition, optimal solution details were discussed, and the data-driven models were shown to work well to achieve optimal throughput.

A future research direction of this work could be the development of data-driven metamodels for anion-exchange chromatography, and for more chromatography parameters, such as yield and binding capacity, with more input variables, such as pH value and temperature. In addition, other performance criteria of the purification process, such as impurity removal capacity, could be taken into account for multi-objective optimization [9,21,22]. The uncertainty of the parameters, such as yield and titer, could also be considered [19,22]. Finally, single-use chromatography could be considered in the purification process as an important direction in smart biopharmaceutical manufacturing [48].

Acknowledgements

Funding from the UK Engineering & Physical Sciences Research Council (EP/I033270/1 and EP/M027856/1) is gratefully acknowledged. The authors would like to thank Dr. Spyridon Gerontas for providing data and useful discussions, and Dr. Lingjian Yang for implementing a preliminary analysis of regression models.

Compliance with ethics guidelines

Songsong Liu and Lazaros G. Papageorgiou declare that they have no conflicts of interest or financial conflicts to disclose.

Nomenclature

Indices

d	Position in multiparametric disaggregation = p, . . ., max$p$
i	Column volume size
j	Column number = 1, . . .,maxCN_s
k	Cycle number = 1, . . .,maxCYN_s
m	Diameter size
n	Digit of the binary representation = 1, . . .,$\lceil \log_2 \max BN \rceil$
q	Integer number in multiparametric disaggregation
r	Interval in piecewise regression function
s	Downstream step = ct_1 (centrifugation 1), ho (homogenization), ct_2 (centrifugation 2), fi (filtration), af (affinity chromatography), ce (cation-exchange chromatography), uf_1 (UF/DF 1), ae (anion-exchange chromatography), uf_2 (UF/DF 2), bf (bulk fill)

Sets

CS	Set of chromatography steps = {af, ce, ae}

Parameters

a, b, c	Utilities cost coefficients
aot	Annual operating time, d
bcv_s	Buffer volume to column volume ratio at chromatography step s
$bp_{s,r}$	Breakpoint of loaded mass between intervals r and (r + 1) at chromatography step s, g
bpc	Buffer price, GBP·L^{-1}
brc	Bioreactor cost, GBP
brn	Number of bioreactors
brt	Bioreaction time, d
brv	Bioreactor volume, L
bvr_s	Buffer volume ratio at centrifugation step s
$cc_{s,i}$	Column cost of size i at chromatography step s, GBP
$\widetilde{cc}_{s,m}$	Column cost of diameter size m at chromatography step s, GBP
$cv_{s,i}$	Volume of column size i at chromatography step s, L
dbc_s	Dynamic binding capacity at chromatography step s, g·L^{-1}
$dm_{s,i}$	Diameter of column size i at chromatography step s, cm
$\widetilde{dm}_m$	Diameter of size m at chromatography step s, cm
don	Number of operators for downstream processing
dvr	Diafiltration volume to column volume ratio at the second UF/DF step
ecv_s	Elute volume to column volume ratio at chromatography step s
el	Equipment lifetime, yeara
$fconc$	Final concentration of product, g·L^{-1}
fvr	Flush volume ratio at the first UF/DF step
gef	General equipment factor
gu	General utility unit cost, GBP·L^{-1}
$h_{s,i}$	Bed height of column size i at chromatography step s, cm
ir	Interest rate
$i\lambda$	Ratio of insurance cost to fixed capital investment
l_s	Lifetime of resin at chromatography step s, cycle
$lang$	Lang factor
maxBBV	Maximum buffer volume per batch, L
maxBN	Maximum number of batches
maxCN_s	Maximum number of columns at chromatography step s
maxCOG	Maximum COG per gram, GBP·L^{-1}
maxCYN_s	Maximum number of cycles at chromatography step s
maxH_s	Maximum column bed height size at chromatography step s, cm
maxLM_s	Maximum product mass loaded at chromatography step s, g
max$p$	Maximum position in multiparametric disaggregation
maxT_s	Maximum processing time per batch at chromatography step s, h
maxTCV_s	Maximum total column volume at chromatography step s, L
maxTP_s	Maximum throughput at chromatography step s, g·L^{-1}
$ma\lambda$	Maintenance cost ratio to the fixed capital investment
$mepc$	Media price, GBP·L^{-1}
$mi\lambda$	Miscellaneous material cost ratio to chemical reagent and consumable costs
$m\lambda$	Management cost ratio to direct labor cost
$oe\lambda$	Other equipment cost ratio to the bioreactor cost
of	Resin overpacking factor
pr_s	Processing rate of step s, L·h^{-1}
$q\lambda$	Ratio of QCQA cost to direct labor cost
rpc_s	Resin price at chromatography step s, GBP·L^{-1}
refCC	Reference cost of a chromatography column, GBP
refDM	Reference diameter of a chromatography column, cm
sfd	Duration per shift, h
sfn	Number of shifts per day, d^{-1}
st	Seed train bioreaction time, d
$s\lambda$	Supervisors cost ratio to direct labor cost
$titer$	Upstream product titer, g·L^{-1}
$t\lambda$	Tax cost ratio to the fixed capital investment
uon	Number of operators per bioreactor in USP
uot	USP operating time per day
vel	Linear velocity of flow at the anion-exchange chromatography step, cm·h^{-1}
w	Wage of an operator, GBP·L^{-1}
yd_s	Product yield at step s
α	Bioreactor working volume ratio
$\beta^0_{s,r}$	Constant coefficient in interval r at chromatography step s
$\beta^H_{s,r}$	Coefficient for bed height in interval r at chromatography step s
$\beta^{LM}_{s,r}$	Coefficient for loaded mass in interval r at chromatography step s
$\beta^V_{s,r}$	Coefficient for velocity in interval r at chromatography step s
ε	A small number to separate two successive intervals at the breakpoint s
θ	Media overfill allowance
μ	Chromatography resin utilization factor
σ	Batch success rate

Continuous variables

ABV	Annual buffer volume, L
AP	Annual product output, g

AT Annual downstream processing time, d
BAT Time for adding buffer per batch at anion-exchange chromatography step, h
BBV Buffer volume added per batch, L
BC Buffer cost, GBP
BRC Bioreactor cost, GBP
BT Downstream processing time per batch, d
BV_s Buffer volume per batch in chromatography step s, L
CAC Capital cost, GBP
CC Consumables cost, GBP
COG Annual cost of goods, GBP
CRC Chemical reagents cost, GBP
CAP_q Continuous variable for annual production in multiple disaggregation at digit q
$CTP_{s,q}$ Continuous variable for throughput in multiple disaggregation at digit q, step s
DLC Direct labor cost, GBP
FCI Fixed capital investment, GBP
GUC General utility cost, GBP
IC Insurance cost, GBP
LC Labor cost, GBP
LM_s Mass loaded to single column at chromatography step s, g
LM_s^1 Mass loaded to single 1 m-diameter column at chromatography step s, g
M_0 Initial product mass entering downstream processes per batch, g
M_s Product mass per batch after step s, g
MAC Maintenance cost, GBP
MC Management cost, GBP
MEC Media cost, GBP
MIC Miscellaneous material cost, GBP
OIC Other indirect costs, GBP
PLT Time for loading product per batch at anion-exchange chromatography step, h
PV_0 Initial product volume entering downstream processes per batch, L
PV_s Product volume per batch after step s, L
QC QCQA cost, GBP
RV_s Resin volume required at chromatography step s, L
SC Supervisors cost, GBP
T_s Processing time per batch of step s, h
TC Tax cost, GBP
TCV_s Total column volume at chromatography step s, L
TP_s Throughput of single column at chromatography step s, g·h^{-1}
TP_s^1 Throughput of single 1 m-diameter column at chromatography step s, g·h^{-1}
UC Utilities cost, GBP
V_s Linear velocity of flow at chromatography step s, cm·h^{-1}
VFR Volumetric flow rate at anion-exchange chromatography step, L·h^{-1}

Binary variables

$BAP_{d,q}$ 1 if digit q for power d is selected for annual production output; 0 otherwise
$BTP_{s,d,q}$ 1 if digit q for power d is selected for throughput at chromatography step s; 0 otherwise
$E_{s,m}$ 1 if diameter size m is selected at chromatography step s; 0 otherwise
$F_{s,j}$ 1 if there are j columns at chromatography step s; 0 otherwise
$O_{s,r}$ 1 if the function at interval r is selected at chromatography step s; 0 otherwise
$W_{s,i,j}$ 1 if there are j columns of size i at chromatography step s; 0 otherwise
$X_{s,i}$ 1 if column size i is selected at chromatography step s; 0 otherwise
$Y_{s,k}$ 1 if there are k cycles at chromatography step s; 0 otherwise
Z_n 1 if the nth digit of the binary representation of variable BN is equal to 1; 0 otherwise

Integer variables

BN *Number of completed batches*
$CN_{s,}$ Number of columns of size i at chromatography step s
$\widetilde{CN}_s$ Number of columns at chromatography step s
CYN_s Number of cycles per batch at chromatography step s
H_s Bed height of column at chromatography step s, cm

Auxiliary variables

$\overline{BCOGAP}_{d,q} \equiv COG \cdot BAP_{d,q}$
$\overline{BTTP}_{s,d,q} \equiv T_s \cdot BTP_{s,d,q}$
$\overline{CCOGAP}_q \equiv COG \cdot CAP_q$
$\overline{COGAP} \equiv COG \cdot AP$
$\overline{CTTP}_{s,q} \equiv T_s \cdot CTP_{s,q}$
$\overline{EF}_{s,m,j} \equiv E_{s,m} \cdot F_{s,j}$
$\overline{EFH}_{s,m,j} \equiv E_{s,m} \cdot F_{s,j} \cdot H_s$
$\overline{EFT}_{s,m,j} \equiv E_{s,m} \cdot F_{s,j} \cdot PLT$
$\overline{ELM}_{s,m} \equiv E_{s,m} \cdot LM_s^1$
$\overline{ETP}_{s,m} \equiv E_{s,m} \cdot TP_s^1$
$\overline{FTTP}_{s,j} \equiv F_{s,j} \cdot T_s \cdot TP_s$
$\overline{FY}_{s,j,k} \equiv F_{s,j} \cdot Y_{s,k}$
$\overline{FYLM}_{s,j,k} \equiv F_{s,j} \cdot Y_{s,k} \cdot LM_s$
$\overline{TTP}_s \equiv T_s \cdot TP_s$
$\overline{WT}_{s,i,j} \equiv W_{s,i,j} \cdot PLT$
$\overline{WTTP}_{s,i,j} \equiv W_{s,i,j} \cdot T_s \cdot TP_s$
$\overline{WY}_{s,i,j,k} \equiv W_{s,i,j} \cdot Y_{s,k}$
$\overline{WYLM}_{s,i,j,k} \equiv W_{s,i,j} \cdot Y_{s,k} \cdot LM_s$
$\overline{XLM}_{s,i} \equiv X_{s,i} \cdot LM_s^1$
$\overline{XTP}_{s,i} \equiv X_{s,i} \cdot TP_s^1$
$\overline{XY}_{s,i,k} \equiv X_{s,i} \cdot Y_{s,k}$
$\overline{YH}_{s,k} \equiv Y_{s,k} \cdot H_s$
$\overline{YV}_{s,k} \equiv Y_{s,k} \cdot TCV_s$
$\overline{ZM}_{s,n} \equiv Z_n \cdot M_s$
$\overline{ZT}_n \equiv Z_n \cdot BT$
$\overline{ZV}_n \equiv Z_n \cdot BBV$
$\overline{ZYV}_{s,k,n} \equiv Z_n \cdot Y_{s,k} \cdot TCV_s$

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2019.10.011.

References

[1] Davis J, Edgar T, Porter J, Bernaden J, Sarli M. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. Comput Chem Eng 2012;47:145–56.

[2] Thoben KD, Wiesner S, Wuest T. "Industrie 4.0" and smart manufacturing—a review of research issues and application examples. Int J Automat Technol 2017;11(1):4–16.
[3] Kang HS, Lee JY, Choi SS, Kim H, Park JH, Son JY, et al. Smart manufacturing: past research, present findings, and future directions. Int J Pr Eng Man-GT 2016;3(1):111–28.
[4] Bogle IDL. A perspective on smart process manufacturing research challenges for process systems engineers. Engineering 2017;3(2):161–5.
[5] Simaria AS, Turner R, Farid SS. A multi-level meta-heuristic algorithm for the optimisation of antibody purification processes. Biochem Eng J 2012;69:144–54.
[6] Brunet R, Guillén-Gosálbez G, Pérez-Correa JR, Caballero JA, Jiménez L. Hybrid simulation-optimization based approach for the optimal design of single-product biotechnological processes. Comput Chem Eng 2012;37:125–35.
[7] Allmendinger R, Simaria AS, Farid SS. Efficient discovery of chromatography equipment sizing strategies for antibody purification processes using evolutionary computing. In: Proceedings of the 12th International Conference on Parallel Problem Solving from Nature-Volume Part II; 2012 Sep 1–5; Taormina, Italy. Berlin: Springer; 2012. p. 468–77.
[8] Allmendinger R, Simaria AS, Turner R, Farid SS. Closed-loop optimization of chromatography column sizing strategies in biopharmaceutical manufacture. J Chem Technol Biotechnol 2014;89(10):1481–90.
[9] Allmendinger R, Simaria AS, Farid SS. Multiobjective evolutionary optimization in antibody purification process design. Biochem Eng J 2014;91:250–64.
[10] Martagan T, Krishnamurthy A, Leland PA, Maravelias CT. Performance guarantees and optimal purification decisions for engineered proteins. Oper Res 2018;66(1):18–41.
[11] Asenjo JA, Montagna JM, Vecchietti AR, Iribarren OA, Pinto JM. Strategies for the simultaneous optimization of the structure and the process variables of a protein production plant. Comput Chem Eng 2000;24(9–10):2277–90.
[12] Montagna JM, Vecchietti AR, Iribarren OA, Pinto JM, Asenjo JA. Optimal design of protein production plants with time and size factor process models. Biotechnol Prog 2000;16(2):228–37.
[13] Pinto JM, Montagna JM, Vecchietti AR, Iribarren OA, Asenjo JA. Process performance models in the optimization of multiproduct protein production plants. Biotechnol Bioeng 2001;74(6):451–65.
[14] Simeonidis E, Pinto JM, Lienqueo ME, Tsoka S, Papageorgiou LG. MINLP models for the synthesis of optimal peptide tags and downstream protein processing. Biotechnol Prog 2005;21(3):875–84.
[15] Liu S, Simaria AS, Farid SS, Papageorgiou LG. Mixed integer optimisation of antibody purification processes. In: Kraslawski A, Turunen I, editors. Computer aided chemical engineering. Proceedings of the 23rd European Symposium on Computer Aided Process Engineering; 2013 June 9–12; Lappeenranta, Finland. Amsterdam: Elsevier; 2013. p. 157–62.
[16] Liu S, Simaria AS, Farid SS, Papageorgiou LG. Designing cost-effective biopharmaceutical facilities using mixed-integer optimization. Biotechnol Prog 2013;29(6):1472–83.
[17] Liu S, Simaria AS, Farid SS, Papageorgiou LG. Optimising chromatography strategies of antibody purification processes by mixed integer fractional programming techniques. Comput Chem Eng 2014;68:151–64.
[18] Liu S, Simaria AS, Farid SS, Papageorgiou LG. Mathematical programming approaches for downstream processing optimisation of biopharmaceuticals. Chem Eng Res Des 2015;94:18–31.
[19] Liu S, Farid SS, Papageorgiou LG. Integrated optimization of upstream and downstream processing in biopharmaceutical manufacturing under uncertainty: a chance constrained programming approach. Ind Eng Chem Res 2016;55(16):4599–612.
[20] Liu S, Gerontas S, Gruber D, Turner R, Titchener-Hooker NJ, Papageorgiou LG. Optimization-based framework for resin selection strategies in biopharmaceutical purification process development. Biotechnol Prog 2017;33(4):1116–26.
[21] Liu S, Papageorgiou LG. Optimal production of biopharmaceutical manufacturing. In: Yuan Z, Singh R, editors. Process systems engineering for pharmaceutical manufacturing. Computer aided chemical engineering. Amsterdam: Elsevier; 2018. p. 569–95.
[22] Liu S, Papageorgiou LG. Multi-objective optimisation for biopharmaceutical manufacturing under uncertainty. Comput Chem Eng 2018;119:383–93.
[23] Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. Comput Chem Eng 2018;108:250–67.
[24] Wang GG, Shan S. Review of metamodeling techniques in support of engineering design optimization. J Mech Des 2007;129(4):370–80.
[25] Forrester A, Sóbester A, Keane A. Engineering Design via Surrogate Modelling: A Practical Guide. Chichester: John Wiley & Sons; 2008.
[26] Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S, et al. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. J Chem Inf Comput Sci 2002;42(6):1347–57.
[27] Mandenius CF, Brundin A. Bioprocess optimization using design-of-experiments methodology. Biotechnol Prog 2008;24(6):1191–203.
[28] Ghose S, Zhang J, Conley L, Caple R, Williams KP, Cecchini D. Maximizing binding capacity for protein A chromatography. Biotechnol Prog 2014;30(6):1335–40.
[29] Wang G, Briskot T, Hahn T, Baumann P, Hubbuch J. Root cause investigation of deviations in protein chromatography based on mechanistic models and artificial neural networks. J Chromatogr A 2017;1515:146–53.
[30] Nagrath D, Messac A, Bequette BW, Cramer SM. A hybrid model framework for the optimization of preparative chromatographic processes. Biotechnol Prog 2004;20(1):162–78.
[31] Pirrung SM, van der Wielen LAM, van Beckhoven RFWC, van de Sandt EJAX, Eppink MHM, Ottens M. Optimization of biopharmaceutical downstream processes supported by mechanistic models and artificial neural networks. Biotechnol Prog 2017;33(3):696–707.
[32] Nascimento A, Pinto IF, Chu V, Aires-Barros MR, Conde JP, Azevedo AM. Studies on the purification of antibody fragments. Sep Purif Technol 2018;195:388–97.
[33] Scanlan C, Shumway J, Castano J, Wagner M, Waghmare R. Challenges and strategies for the downstream purification and formulation of Fab antibody fragments. Biopharm Int 2014;27(1):42–4.
[34] Gerontas S, Asplund M, Hjorth R, Bracewell DG. Integration of scale-down experimentation and general rate modelling to predict manufacturing scale chromatographic separations. J Chromatogr A 2010;1217(44):6917–26.
[35] Boushaba R, Baldascini H, Gerontas S, Titchener-Hooker NJ, Bracewell DG. Demonstration of the use of windows of operation to visualize the effects of fouling on the performance of a chromatographic step. Biotechnol Prog 2011;27(4):1009–17.
[36] Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KK. Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Netw 2000;11(5):1188–93.
[37] MacKay DJC. Introduction to Gaussian processes. In: Bishop CM, editor. Neutral networks and machine learning, NATO ASI series: Ser. F: computer and systems science. Berlin: Springer; 1998. p. 133–66.
[38] Wang Y, Witten IH. Modeling for optimal probability prediction. In: Proceedings of the Nineteenth International Conference on Machine Learning; 2002 July 8–12; Sydney, NSW, Australia; 2002. p. 650–7.
[39] Box GEP, Draper NR. Response surfaces, mixtures, and ridge analyses. 2nd ed. Hoboken: John Wiley & Sons; 2007.
[40] Yang L, Liu S, Tsoka S, Papageorgiou LG. Mathematical programming for piecewise linear regression analysis. Expert Syst Appl 2016;44:156–67.
[41] Frank E, Hall MA, Witten IH. Data mining: practical machine learning tools and techniques. 4th ed. San Francisco: Morgan Kaufmann; 2016.
[42] GAMS Development Corporation. GAMS: a user's guide. Washington, DC: GAMS Development Corporation; 2016.
[43] Glover F. Improved linear integer programming formulations of nonlinear integer problems. Manage Sci 1975;22(4):455–60.
[44] Kolodziej S, Castro PM, Grossmann IE. Global optimization of bilinear programs with a multiparametric disaggregation technique. J Glob Optim 2013;57(4):1039–63.
[45] Castro PM. Normalized multiparametric disaggregation: an efficient relaxation for mixed-integer bilinear problems. J Glob Optim 2016;64(4):765–84.
[46] Koleva MN, Styan CA, Papageorgiou LG. Optimisation approaches for the synthesis of water treatment plants. Comput Chem Eng 2017;106:849–71.
[47] Lin MH, Carlsson JG, Ge D, Shi J, Tsai JF. A review of piecewise linearization methods. Math Probl Eng 2013;2013:101376.
[48] Langer ES, Rader RA. Single-use technologies in biopharmaceutical manufacturing: a 10-year review of trends and the future. Eng Life Sci 2014;14(3):238–43.