Journal of Biotechnology 179 (2014) 17–25

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Data mining for rapid prediction of facility fit and debottlenecking of biomanufacturing facilities

Yang Yang[a], Suzanne S. Farid[b,*], Nina F. Thornhill[a,**]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK

ARTICLE INFO

Article history:
Received 20 September 2013
Received in revised form 27 February 2014
Accepted 4 March 2014
Available online 15 March 2014

Keywords:
Biopharmaceutical manufacture
Stochastic discrete-event simulation
Decision tree classification
Multivariate data analysis
Data mining

ABSTRACT

Higher titre processes can pose facility fit challenges in legacy biopharmaceutical purification suites with capacities originally matched to lower titre processes. Bottlenecks caused by mismatches in equipment sizes, combined with process fluctuations upon scale-up, can result in discarding expensive product. This paper describes a data mining decisional tool for rapid prediction of facility fit issues and debottlenecking of biomanufacturing facilities exposed to batch-to-batch variability and higher titres. The predictive tool comprised advanced multivariate analysis techniques to interrogate Monte Carlo stochastic simulation datasets that mimicked batch fluctuations in cell culture titres, step yields and chromatography eluate volumes. A decision tree classification method, CART (classification and regression tree) was introduced to explore the impact of these process fluctuations on product mass loss and reveal the root causes of bottlenecks. The resulting pictorial decision tree determined a series of if-then rules for the critical combinations of factors that lead to different mass loss levels. Three different debottlenecking strategies were investigated involving changes to equipment sizes, using higher capacity chromatography resins and elution buffer optimisation. The analysis compared the impact of each strategy on mass output, direct cost of goods per gram and processing time, as well as consideration of extra capital investment and space requirements.

1. Introduction

In recent years, cell culture titres of monoclonal antibodies (mAbs) have increased dramatically as a result of improvements to cell lines, media composition, and feeding strategies (Birch and Racher, 2006; Li et al., 2010). Furthermore, it is common for titres to increase by 50% or more as a product progresses from early to late process development (Kelley et al., 2009). Higher titre processes can pose facility fit challenges for downstream processing (DSP), particularly during tech transfer to legacy biopharmaceutical manufacturing facilities that were constructed with multiple large-volume bioreactors (>10,000 L) and DSP capacities matched to lower titre processes. Legacy facilities can struggle to cope with the resulting higher protein loads onto DSP due to bottlenecks reached in DSP unit operations (e.g. chromatography columns) or tank storage capacities (Aldington and Bonnerjea, 2007; Chang, 2011; Farid, 2008; Kamarck, 2006; Kelley, 2009; Stonier et al., 2012). Thus systematic and rigorous tools for facility fit analysis and debottlenecking are critical to gaining greater understanding of the root causes of suboptimal facility fit and identifying the most promising debottlenecking strategies.

Facility fit analysis and DSP debottlenecking efforts are complicated by the inherent batch-to-batch variability present in bioprocess unit operations (Farid, 2008; Stonier et al., 2013). Facility fit assessments that are based on single point expected values for key process parameters, and hence do not account for process fluctuations, may not identify the correct bottleneck. Certain combinations of worst case values can lead to volumes that exceed equipment capacities and result in having to discard expensive product. The likelihood and consequences of such scenarios would not be captured by facility fit assessments based solely on expected values. Furthermore, large scale facilities often have fixed stainless steel equipment and piping networks. This makes it harder to adopt debottlenecking strategies involving equipment changes in response to fit issues arising from process variability and higher titres.

* Corresponding author at: University College London, Biochemical Engineering, Torrington Place, London WC1E 7JE, UK. Tel.: +44 20 7679 4415.
** Corresponding author at: Imperial College London, Department of Chemical Engineering, London SW7 2AZ, UK Tel.: +44 (0)20 7594 6622.
E-mail addresses: s.farid@ucl.ac.uk (S.S. Farid), n.thornhill@imperial.ac.uk (N.F. Thornhill).

http://dx.doi.org/10.1016/j.jbiotec.2014.03.004

Effective facility fit assessments can benefit from advanced data mining of datasets generated from bioprocess models that can capture the dynamics of bioprocesses as well as the impact of resource constraints and process variability. Commercial bioprocess modelling packages (e.g. Superpro Designer (Intelligen, Inc., Scotch Plains, NJ) tend to be useful for equipment sizing and costing but are not typically designed to capture the consequences of resource delays (e.g. due to buffer storage tank availability) or uncertainties (e.g. titre). In this work, a discrete-event data-driven simulation platform developed by the Advanced Centre for Biochemical Engineering at UCL (Stonier et al., 2012, 2013) was used to model the performance of bioprocesses exposed to uncertainties and facility constraints. The model captures the mass balances, equipment sizing, dynamic resource allocation and process economics of purification sequences. Monte Carlo simulation methods have been used in this work to mimic a batch history record by accounting for key process fluctuations and generating the possible outcomes and their likelihood. These simulations enable predictions of the impact of process fluctuations on the possibility of product loss. Monte Carlo simulation has been used increasingly in various bioprocessing examples to capture the impact of technical, clinical or commercial uncertainties on unit operation models (Sin et al., 2009), whole bioprocess costs (Farid et al., 2005; Pollock et al., 2013) and on portfolio management and capacity planning decisions (George and Farid, 2008).

Data mining has been used in the biotech sector to identify trends in large datasets from historical batch records, often applied to fermentation data (Charaniya et al., 2010; Mercier et al., 2013; Rommel and Schuppert, 2004). Principal component analysis (PCA) is a common multivariate analysis method that uses an orthogonal transformation to convert a set of variables into a set of linearly uncorrelated variables. It has been applied in manufacturing process analysis to reveal the internal structure and pattern of historical data (Edwards-Parton et al., 2008; Pate et al., 1999; Thornhill et al., 2006) but cannot generate the potential rules hidden behind the data. Decision tree classification is an effective data mining method that has been applied in fermentation parameter identification (Buck et al., 2002; Ma et al., 2004) and fermentation process optimization (Coleman et al., 2003; Lam and Malik, 2001). In this work, the classification and regression tree (CART) was introduced to analyse the large complex downstream manufacturing bioprocess datasets generated by Monte Carlo simulations and to find the hidden root causes of bottlenecks in existing facilities exposed to batch-to-batch variability and higher titres. The data mining outputs can be used to support better process understanding through rigorous root cause analysis and continuous risk management and hence contribute to effective implementation of quality by design (QbD) principles throughout the lifecycle of a product.

This paper is organized as follows. First, downstream bioprocess facilities used in the case study are described. Second, the methods applied in the case study including stochastic discrete-event simulation, correlation coefficients analysis and CART decision trees are briefly introduced. In Section 4, the Monte Carlo simulation datasets are analysed to identify mismatches in pool volumes resulting in product losses. The key process fluctuations leading to mass loss and threshold values for those process fluctuations are derived using CART decision trees. This work demonstrated that the decision tree classification method can be applied to explore not only the impact of process fluctuations on product mass loss but also the critical combinations of parameter values that lead to mass loss. Furthermore, the pictorial CART tree result with its series of if-then rules of the critical combinations of factors that lead to different mass loss levels can be used to identify debottlenecking solutions worth pursuing. Finally, three different debottlenecking solutions are compared in relation to their impact on three key metrics: mass output, direct cost of goods per gram (COG/g) and processing time. The solutions explored spanned changes to equipment sizing, using more efficient purification resins and reducing the eluate volume fluctuations expected through buffer optimisation.

Table 1
Facility specification for the chromatography and filtration downstream processing steps.

Parameter		Step		
Chromatography		Protein A	AEX	CEX
Column diameter (m)		1	1	1
Bed height (m)		0.20	0.25	0.15
Bed volume (L)		157	196	118
Load capacity (g/L)		25	50	15
Linear velocity (cm/h)		450	450	140
Expected number of cycles		9	3	15
Expected pool volume (*CV*/cycle)		2	3	2.5
Pool tank volume (L)		5000	5000	5000
Expected step yield (%)		88	88	88
Filtration	Post Protein A UF/DF	Post AEX UF	Final UF/DF	VRF
Retentate tank volume (L)	5000	5000	5000	5000
Expected average flux rate (L/m^2 h)	100/55	110/60	140/80	N/A
Target concentration (g/L)	25	25	38	
Diafiltration volumes (CV)	3	0	10	0
Expected step yield (%)	95	95	95	99

Note: Pool volume refers to the volume of the product stream. In Protein A and CEX steps operated in bind-and-elute mode this refers to the eluate volume collected. In AEX operated in flow-through mode this refers to the load and post wash volumes collected.

2. Problem domain

An existing standard monoclonal antibody (mAb) manufacturing process was considered in this work, as shown in Fig. 1. The volume of bioreactor broth generated during each batch was 10,000 L. Biomass and other debris were removed using centrifugation and depth filtration with step recovery yields of 95%. The mAb downstream processing sequence was defined as: Protein A affinity chromatography capture step, low pH virus inactivation, ultrafiltration/diafiltration (UF/DF), anion exchange chromatography (AEX), ultrafiltration (UF), cation exchange chromatography (CEX), virus reduction filtration (VRF) and a final UF/DF.

The downstream process was originally built to handle titres up to 2 g/L but it now needed to cope with average titres of 4 g/L and hence a harvest kg/batch value of 40 kg rather than 20 kg. The impact of the higher titre feed on the number of cycles required for each DSP step and hence the expected pool volumes from each step were calculated and used to allocate larger product collection tanks where appropriate. The specification of the downstream process equipment sizes and process parameters (e.g. resin binding capacities) is presented in Table 1. This facility configuration, modified to cope with 4 g/L titre feeds, was identified as the base case facility. The aim was to investigate the impact of batch-to-batch variability on its performance and predict facility fit issues.

Facility fit assessments are carried out often with information from a limited number of batches at scale, particularly for new processes or new drug candidates. In the absence of a significant number of batch history records such assessments are typically based on expected or worst case values which do not capture the full range of possible outcomes or their likelihood of occurrence. Hence in this paper stochastic simulation datasets were generated as a mimic of batch record data and then analysed using data mining techniques. The simulation datasets capture typical batch-to-batch variability expected at large scale which can be useful to companies

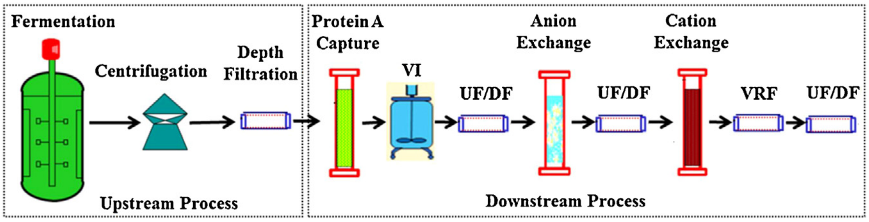

Fig. 1. A standard monoclonal antibody (mAb) manufacturing process. VI = virus inactivation, UF/DF = ultrafiltration/diafiltration, VRF = virus retention filtration.

for predicting facility fit issues and bottlenecks when new processes are introduced into an existing facility. This work integrates stochastic simulation with data mining to identify the hidden reasons behind bottlenecks in the facility that result in product being discarded and then uses these insights to propose and evaluate possible debottlenecking solutions.

3. Methodology

3.1. Stochastic discrete-event simulation

A database-driven discrete-event simulation tool (Stonier et al., 2012) developed in ExtendSim 8 (Imagine That! Inc, San Jose, USA) was used to simulate the manufacturing process under uncertainty using Monte Carlo simulations. In this work, the simulation takes input variables such as facility data (e.g. availability of different equipment sizes), process data (e.g. process sequence, process parameters) and economic data (e.g. chromatography resin and buffer costs). Examples of key input variables are shown in Table 1. The output variables are key scheduling parameters (e.g. batch durations, delays), technical performance metrics (e.g. mass output and product loss) and financial metrics (e.g. COG/g).

In order to mimic batch-to-batch variability caused by fluctuations in real manufacturing conditions, representative triangular distributions indicated in Table 2 were assigned to four key parameter types: product titre, step yields for the purification steps, chromatography eluate volumes and filter flux rates. Triangular distributions were used since typical minimum, maximum and most likely values for each of the parameters could be derived through discussions with industrial experts in the user consortium of the EPSRC Centre for Innovative Manufacturing in Emergent Macromolecular Therapies as well as literature sources (Amanullah et al., 2010; Abu-Absi et al., 2010; Legmann et al., 2009). For example, a ±10% variation in product titres was captured which affects the mass load of product onto the chromatography steps and hence the number of cycles in each chromatography step. In this work, the chromatography eluate volume refers to the pool volume per cycle which is the number of column volumes collected per chromatography cycle in bind-and-elute mode or the number of load and post wash column volumes collected per chromatography cycle in flow-through mode. In this case study, a value of ±50% variation in eluate volumes was considered to reflect the challenges in predicting the position and shape of the product peak on the UV trace and hence the collection criteria for the eluate upon transfer of a process to a new facility. This is particularly the case when significant leading or tailing on the elution peaks is observed with steps that are highly sensitive to pH and conductivity of the elution buffer (Stonier et al., 2013). Whilst sensible ranges in process variability were sought for each of the parameters, the primary aim of the paper was to demonstrate the application of the proposed data mining methodology to perform more rigorous and predictive facility fit assessments. Hence, the actual inputs and answers should not be seen as definitive but an illustration of how to approach such an assessment.

Table 2
Variable distribution ranges.

Variable	Min (%)	Most	Max (%)
Product titre	−10	Base case	10
Eluate volumes	−50	Base case	50
Filter flux rates	−10	Base case	10
Step yield			
Chromatography steps	83	88%	93
Virus inactivation	98	99%	100
Ultrafiltration/diafiltration	90	95%	99
Virus retention filtration	90	95%	99

In Monte Carlo simulation methods, convergence diagnostics is an important topic because lack of convergence would affect the reliability of the simulation result (Cassettari et al., 2012). In this work, convergence was tested on all dynamic variables. Four hundred iterations were sufficient to reach convergence and hence provided reliable probability distributions of possible outcomes for analysis. The Monte Carlo simulation was set up to run for 400 iterations to generate the stochastic dataset for data analysis. The values of the uncertain input parameters (product titre, step yields, chromatography eluate volumes and filter flux rates) vary under the triangular distributions described in Table 2 from run to run. All input variables as well as the key outputs such as mass loss were recorded for each iteration. After the simulation experiments, the results were used to generate frequency distribution plots of mass loss and pool volumes at each step such as those presented later in Fig. 2 and to enable the root causes of unwanted events to be investigated using data mining techniques.

3.2. Correlation coefficients analysis

Correlation coefficients usually known as Pearson's product moment correlation coefficients provide a measure of the strength of the linear relationship between two variables (Rodgers and Nicewander, 1988). Correlation coefficients between two N-dimensional vectors x and y is defined by

$$\rho_{xy} = \frac{\sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k-\bar{x})^2}\sqrt{\sum_{k=1}^{N}(y_k-\bar{y})^2}} \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are defined as the mean of x and y, respectively.

Correlation coefficients only measure the linear relationship between two variables. So when more than one input factor is under consideration, partial correlation coefficients can be used to characterize the strength of the linear relationship between two variables when all linear effects of other variables are removed. $\rho_{X_1,Y|Z}$ is the partial correlation coefficient of X_1 (e.g. titre) and Y (unexpected mass loss caused by process fluctuations in this work) holding $Z = X_2, \ldots, X_k$ (e.g. step yields, eluate volumes and other process parameters except titre):

$$\rho_{X_1,Y|Z} = \frac{\rho_{X_1,Y} - \rho_{X_1,Z}\times\rho_{Y,Z}}{\sqrt{\left(1-\rho^2_{X_1,Z}\right)\left(1-\rho^2_{Y,Z}\right)}} \qquad (2)$$

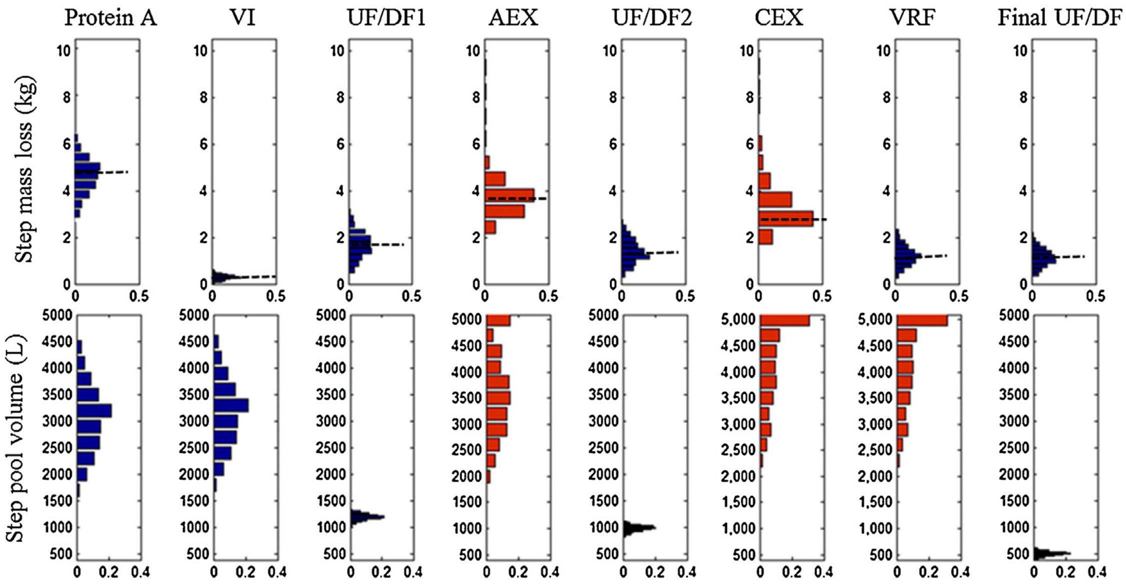

Fig. 2. Distribution of mass loss and pool volumes for each process step. Mass loss can occur due to step yield losses as well as discarding product due to volume mismatches. The mass loss at each step is defined as the difference of product mass before and after each step processing. The dotted line indicates the anticipated mean loss due to step yield alone. The volume of the bioreactor is 10,000 L. Fluctuations assumed for cell culture titre (4 g/L ± 10%), step yields, eluate volumes, and filtration flux rates (Table 2). Monte Carlo simulation iterations, $n = 400$.

In this work, partial correlation coefficients have been used to measure the parameter importance.

3.3. Decision tree classification

Decision tree algorithms are well-established machine learning techniques that have been used for a wide range of applications, especially for classification problems (Grajski et al., 1986; Quinlan, 1996). Decision trees were chosen for this case study given their ability to convert large complex datasets into easy-to-understand and yet information-rich graphical displays. More specifically, the resulting pictorial tree representation was considered a useful tool for rapid elucidation of the critical combinations of parameter values that lead to unacceptable product loss which could then be converted into a set of rules. Further advantages of using decision tree algorithms include minimal requirements for data preparation and robust performance on large datasets.

CART (classification and regression tree) is a nonparametric procedure that uses a stepwise method to establish splitting rules (Breiman et al., 1984; Grajski et al., 1986). CART divides the data into homogenous subsets using binary recursive partitions. The most discriminative variable is first selected as the root node to partition the data set into branch nodes. The root nodes and branch nodes in this study represent critical process parameters driving product loss. The partitioning is repeated until the nodes are homogenous enough to be terminal nodes which are called leaves. The terminal nodes represent critical ranges for the output metric of interest (e.g. unexpected mass loss in this case study). So in a tree structure, leaves represent class labels (e.g. <5% unexpected mass loss) and branches represent conjunctions of features (e.g. critical values for titre and eluate volumes) that lead to those class labels.

3.3.1. Gini impurity index for splitting

There are different splitting criteria for CART such as Gini impurity index (Breiman et al., 1984; Sadras and Bongiovanni, 2004) and Twoing (Piccarreta, 2008). The Gini index was used in this study. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it were randomly labelled according to the distribution of class labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single category or class label.

To compute Gini impurity for a set of items f, Eq. (3) is used where the class label i takes on values in $\{1, 2,\ldots,m\}$, and f_i is the fraction of items labelled as category i in the set:

$$I_G(f)=\sum_{i=1}^{m} f_i(1-f_i)=\sum_{i=1}^{m} f_i-\sum_{i=1}^{m} f_i^2=1-\sum_{i=1}^{m} f_i^2 \qquad (3)$$

3.3.2. Resubstitution estimation

Resubstitution estimate is the proportion of cases that are misclassified by the classifier constructed from the entire learning set. For a learning set consisting of (x_i, ω_i), $i = 1, 2,\ldots,N$, where $d(x)$ is the classifier, the resubstitution estimate is computed in the following manner:

$$R(d)=\frac{1}{N}\sum_{i=1}^{N} I_{\{d(x_i)\neq\omega_i\}} \qquad (4)$$

where $I_{(\bullet)}$ is the indicator of event $\{\bullet\}$; $I_{\{d(x_i)\neq\omega_i\}}=1$, if the event $d(x_i)\neq\omega_i$ is true while $I_{\{d(x_i)\neq\omega_i\}}=0$ if the event is false. In this case study, x_i is the ith case of learning set and its class label, ω_i, is 0% unexpected mass loss, for example. The classifier d is the CART tree model built in Fig. 4 or 5 $d(x_1)$ is the predicted class label for x_1 using classifier d. If the predicted class label $d(x_1)$ is 0–5% unexpected mass loss which is not the same as the class label ω_1, then $I_{\{d(x_i)\neq\omega_i\}}=1$, so the ith case is a misclassified case by the CART tree classifier.

3.3.3. k-Fold cross-validation

k-Fold cross-validation is a widely used technique for assessing the robustness of a model. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once

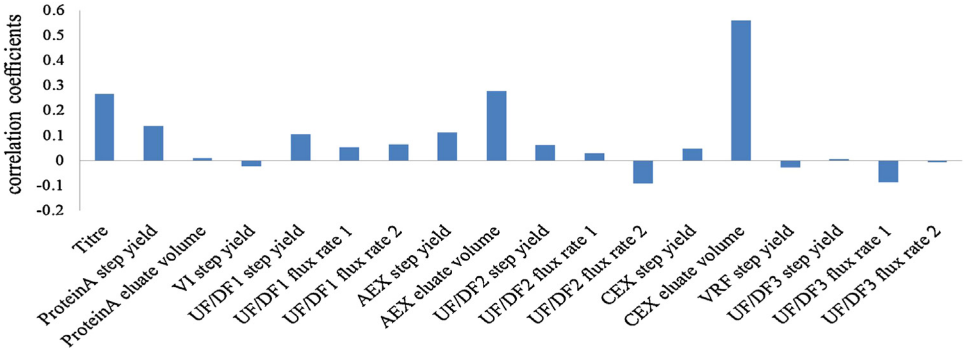

Fig. 3. Correlation coefficients between each input factor and overall unexpected product mass loss. Positive correlation coefficients represent positive linear relationships while negative correlation coefficients represent negative linear relationships. The absolute correlation coefficient values can indicate the strength of linear dependence between process variables and unexpected product mass loss.

as the validation data. The *k* results from the folds can be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. In this work, CART tool in Matlab Statistics Toolbox was used and 10-fold cross-validation (Efron, 1983; Moreno-Torres et al., 2012) was applied for estimating the prediction error.

4. Results

4.1. Bottleneck Identification

The mass output of the facility (kg/batch) is one of the key performance metrics. The stochastic tool was used to predict the frequency distribution of mass output and the likelihood of mass loss using Monte Carlo simulation, given the expected fluctuations in key performance indicators and purification operating parameters indicated in Table 2. Fig. 7(a) shows the predicted batch mass output for the processes running in the base case facility. Based on deterministic values (product titre = 4 g/L, overall process yield = 53%, fermenter scale = 10,000 L), the expected mass output of the base case process should be 21 kg/batch. The values predicted by the simulation fall well short of this value. A very small proportion of batches meet the expected output. This is suggestive of facility fit issues and prompts further investigation.

In order to identify the location of the equipment limitations causing the facility fit issues, the product mass output and product volume output of each process step have been examined. Fig. 2 shows the distribution of the product mass loss and pool volumes at each step. The product mass loss at each step is defined as the difference of product mass before and after each processing step. Usually, the product mass loss will be caused by step yield and should follow the step yield distribution described in Table 2. The step mass loss distribution plot in Fig. 2 identifies abnormal mass loss distributions in the AEX and CEX steps which mean unexpected mass loss.

Furthermore, the step pool volume distribution plot in Fig. 2 reveals the bottleneck location. The vertical histogram plotted alongside each column of data points shows a spike in the distributions at 5000 L for the AEX, CEX and VRF pool tanks. This is due to the fact that the largest volume that can be stored in these tanks was 5000 L. Surplus volume was diverted to waste and the product was lost impacting throughput on a large number of batches. This facility fit issue was caused by the process fluctuations and hence is unexpected and hard to predict at the early process design stage. In the following sections, the term *unexpected mass loss* means the mass loss caused by the process fluctuations only excluding the loss caused by step yields.

4.2. Partial correlation coefficients analysis for variable importance

Partial correlation coefficients analysis was used to find out which of the uncertain input parameters (product titre, step yields, chromatography eluate volumes and filter flux rates) were more important in determining the uncertainty in the key output of interest, the overall unexpected product mass loss. Fig. 3 shows the correlation coefficients between the 18 input factors and overall unexpected product mass loss. A positive correlation coefficient value means a positive relationship with the overall unexpected product mass loss while a negative value means a negative relationship. Fig. 3 shows that, at 0.6, the CEX eluate volume has the highest positive correlation coefficient followed by the AEX eluate volume and titre. These three input parameters have the strongest influence on the unexpected process mass loss. This reinforces observations in Fig. 2 that the unexpected mass loss happened due to tank volume limitations in CEX and AEX where the eluate volumes were collected prior to further concentration. Although partial correlation coefficients can identify the most important drivers of unexpected mass loss, they cannot offer detailed information about how the most important variables impact the unexpected product mass loss. In order to obtain an understanding of the critical parameter values that combine to result in unexpected mass loss, a decision tree analysis was performed to explore the base case simulation data.

4.3. Decision tree analysis for debottlenecking

The Monte Carlo simulation dataset generated by the discrete-event simulation tool has 400 data records. Each data record represents one manufacturing batch. Before using the Monte Carlo simulation dataset as a training dataset for the decision tree analysis, each data record was allocated a class label (as described in Section 3.3.1) since decision tree classification is a supervised learning method (Grajski et al., 1986). The expected mass load from the 10,000 L mAb bioreactor was 40 kg. Unexpected mass loss per batch exceeding 5% (2 kg) of the expected mass load was considered a heavy unexpected mass loss. According to the quantity of product loss, each data record in the Monte Carlo simulation dataset was classified into one of three groups: 0% unexpected mass loss, 0–5% unexpected mass loss and ≥5% unexpected mass loss. The summary of the training dataset is shown in Table 3.

According to the 10-fold cross validation method, the training dataset was randomly divided into 10 disjoint subsets. Each subset

Table 3
Summary of training dataset for CART classification.

Class labels	Description	No. of records
0% unexpected mass loss	Batches with no unexpected mass loss at all	284
0–5% unexpected mass loss	Batches with unexpected mass loss less than 2 kg	75
≥5% unexpected mass loss	Batches with unexpected mass loss equal to or more than 2 kg	41

had roughly equal size and roughly the same class proportions as in the training set. Using nine of the subsets, all possible combinations of trees were developed and these were then tested on the 10th subset. This result provides a cross-validation error rate, which gives an equitable evaluation of the predictive precision of tree models of different sizes. Resubstitution error describes how well the decision tree fits the training dataset while cross-validation error describes the prediction ability of the decision tree. A larger tree has a smaller resubstitution error but can cause over-fitting. The optimal tree should have the minimum cross-validation error and tolerance to resubstitution error.

In order to identify the key process fluctuations driving whether unexpected mass loss occurred or not, two CART trees have been built separately. In Fig. 4(a), 0–5% and ≥5% unexpected mass loss classes have been merged into a >0% unexpected mass loss class versus the 0% unexpected mass loss class. The two-class tree model in Fig. 4(a) reveals that the CEX eluate volume and AEX eluate volume are the key process fluctuations driving the level of unexpected product mass loss occurring. Furthermore, from top to bottom along the branch to each leaf node of the tree, the "if-then" rules can be generated to describe and predict whether there is unexpected mass loss or not caused by critical combinations of key process fluctuations. For example, the left branch of the tree indicates that if CEX eluate volume <2.8 CV and AEX eluate volume <4 CV, then there is no unexpected mass loss in the process with prediction accuracy as high as 95%.

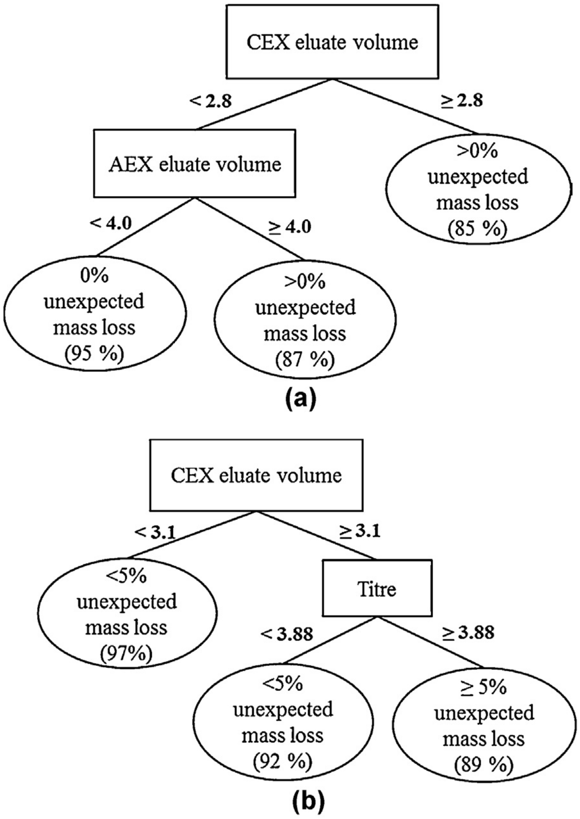
CEX eluate volume
< 2.8
≥ 2.8
AEX eluate volume
>0% unexpected mass loss (85 %)
< 4.0
≥ 4.0
0% unexpected mass loss (95 %)
>0% unexpected mass loss (87 %)
(a)
CEX eluate volume
< 3.1
≥ 3.1
<5% unexpected mass loss (97%)
Titre
< 3.88
≥ 3.88
<5% unexpected mass loss (92 %)
≥ 5% unexpected mass loss (89 %)
(b)

Fig. 4. The optimal CART decision trees based on different class definitions. (a) CART tree for 0% or >0% unexpected mass loss classes, (b) CART tree for <5% or ≥5% unexpected mass loss classes. Numerical values are the threshold levels of the split points for the corresponding split conditions. Rectangle nodes are branch nodes (e.g. eluate volume, titre) which represent the process parameters leading to split. Circle nodes are leaves representing subsets with different class labels for unexpected mass loss levels (e.g. <5% unexpected mass loss). The number in brackets in each leaf represents the percentage of observations in the leaf as an indicator of the confidence level in the predictions.

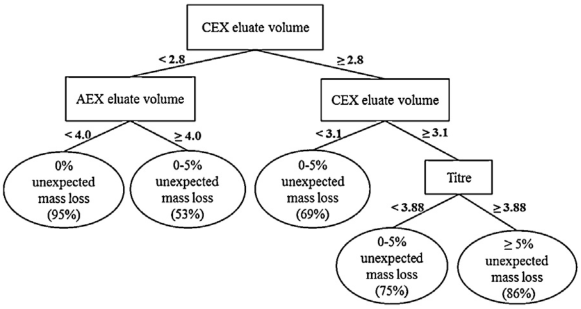
CEX eluate volume
< 2.8
≥ 2.8
AEX eluate volume
CEX eluate volume
< 4.0
≥ 4.0
< 3.1
≥ 3.1
0% unexpected mass loss (95%)
0-5% unexpected mass loss (53%)
0-5% unexpected mass loss (69%)
Titre
< 3.88
≥ 3.88
0-5% unexpected mass loss (75%)
≥ 5% unexpected mass loss (86%)

Fig. 5. The optimal CART decision tree based on a Monte Carlo simulation dataset. Numerical values are the threshold levels of the split points for the corresponding split conditions. Rectangle nodes are branch nodes which represent the process parameters leading to split. Circle nodes are leaves representing subsets with different class labels for 0% unexpected mass loss, 0–5% unexpected mass loss or ≥5% unexpected mass loss. The number in each leaf represents the percentage of observations in the leaf as an indicator of the confidence level in the predictions.

In Fig. 4(b), the 0% and 0–5% unexpected mass loss classes have been merged to form <5% unexpected mass loss class versus the ≥5% unexpected mass loss class. The two-class tree model in Fig. 4(b) reveals that the key process fluctuations leading to the different unexpected mass loss outcomes in the base case process are CEX eluate volume and product titre. As in Fig. 4(a), the "if-then" rules can be generated to describe and predict what critical combinations of key process fluctuations result in <5% or ≥5% unexpected mass loss. The right branch of the tree indicates that if the CEX eluate volume ≥3.1 CV and titre ≥3.88 g/L, then there is ≥5% unexpected mass loss in the process with prediction accuracy as high as 89%.

In order to explore the root causes for three categories of mass levels (0% unexpected mass loss, 0–5% unexpected mass loss and ≥5% unexpected mass loss), rather than two, a three-class CART tree has been built as shown in Fig. 5. This three-class tree model reveals that the key process fluctuations leading to different unexpected product mass loss levels in the base case process are product titre, CEX eluate volume and AEX eluate volume. This reinforces the partial correlation coefficients analysis results in Fig. 3.

Comparing Figs. 4 and 5, the three-class tree in Fig. 5 is the combination of the two-class trees in Fig. 4(a) and (b) but with the subdivision of the >0% unexpected mass loss class of Fig. 4(a) and <5% unexpected mass loss class of Fig. 4(b) so that it could reveal more specific combinations of process fluctuations leading to 0–5% unexpected mass loss. However, the three-class tree in Fig. 5 has lower prediction reliability than two-class trees due to the subdivision of 0–5% unexpected mass loss and 5% unexpected mass loss since they are both minority classes in the training dataset. Nevertheless, even with the lower prediction reliability the three-class tree of Fig. 5 gives interesting insights. Similar classification results and prediction reliability have been reinforced by using a hyper-box approach (Xu and Papageorgiou, 2009).

In order to clearly display the relationship between unexpected mass loss distribution and the key process parameters identified by the decision tree predictive model, windows of operation of CEX eluate volume vs. AEX eluate volume under different titre

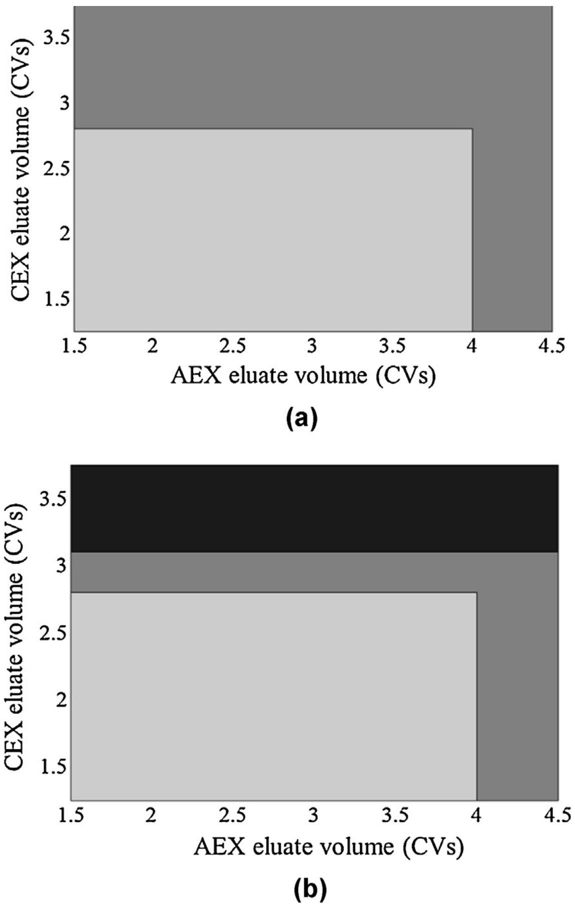

Fig. 6. Windows of operation indicating critical combinations of AEX chromatography eluate volume and CEX eluate volumes that drive mass loss levels for (a) titre < 3.88 g/L and (b) titre ≥ 3.88 g/L. The black area represents batches with ≥5% unexpected mass loss, dark grey areas represent batches with 0–5% unexpected mass loss and light grey areas represent batches with 0% unexpected mass loss.

fluctuation ranges were generated in Fig. 6. Key observations deduced from analysis of Fig. 6 are highlighted below:

- When the AEX eluate volume is below 4.0 CV and CEX eluate volume is below 2.8 CV, there is no unexpected mass loss at all irrespective of the titre fluctuations as illustrated in Fig. 6(a) and (b). When the CEX eluate volume lies within 2.8–3.1 CV and AEX eluate volume lies within 4.0–4.5 CV, 0–5% unexpected mass loss occurs irrespective of the titre fluctuations. These observations reveal the required eluate volume ranges for the installed 5000 L pool tanks to handle the titre fluctuations whilst accepting unexpected mass losses up to the 5% threshold.
- When the CEX eluate volume is in the range of 3.1–3.75 CV, the level of unexpected mass loss depends on whether the titre is higher than 3.88 g/L or not. The analysis shows that the unexpected mass loss will exceed the threshold of 5% if the titre is above 3.88 g/L. Higher titres can increase the probability of needing more cycles to process a batch. This combined with the higher number of column volumes collected as eluate leads to higher pool volumes than can exceed the installed capacity.
- Eluate volumes in AEX and CEX are the dominant factors at 0–5% and 0% unexpected mass loss levels while titre and CEX eluate volume are the dominant factors at ≥5% unexpected mass loss level as indicated in the decision tree in Fig. 5. Furthermore, unexpected mass loss levels in the CEX step are more sensitive to titre fluctuations than in the AEX step. This observation can be attributed to the lower dynamic binding capacity of the CEX resin (15 g/L versus 50 g/L), which results in higher cycle numbers that amplify the effect of eluate volume fluctuations on tank volume limitations.

The results highlight the greater level of information that can be derived through uncertainty analysis combined with the decision tree analysis compared to the traditional approach in industry based on calculations using expected or worst-case values. The uncertainty analysis provides more information as it simulates all possible combinations of variability and indicates the likelihood of different levels of unexpected mass loss. In contrast, facility fit using worst case values alone can lead to equipment being oversized and batch costs rising to cope with events that have a low likelihood of occurrence. The decision tree analysis adds to the insights by providing a series of rules for the critical combinations of parameter values that lead to different mass loss levels.

4.4. *Debottlenecking solutions comparison*

Having identified the critical combinations of parameter values leading to loss, it was possible to propose debottlenecking solutions and evaluate their impact on three key performance metrics: mass output, direct COG/g and time. In this work, the direct COG/g captures the key direct costs incurred when running a batch such as the consumable costs (e.g. resins), buffer costs (e.g. elution buffer), and operator costs. Changes in indirect costs such as the potential increase in operating overhead costs due to purchasing new equipment were not accounted for in this analysis. The effect of different solutions on time was also translated into changes in plant throughput and hence impacted the cost of goods per gram. Three debottlenecking solutions were explored relating to purchasing larger tanks to accommodate the eluate volume fluctuations, narrowing the eluate volume fluctuation through buffer optimization and purchasing higher capacity resins that require fewer cycles. These are discussed in more detail below.

4.4.1. *Debottlenecking solution 1—New vessel*

Introducing 40% larger volume pool tanks to AEX, CEX and VI steps (using 7000 L to replace 5000 L) to handle the predicted peak product volumes results in no product being discarded in any of the process fluctuation scenarios. In Fig. 7(a), the mass throughput result of this solution (dotted line) is much improved compared to the base case facilities (solid line) with most batches meeting the expected throughput of 21 kg/batch. The improvement on the direct COG/g can also be seen in Fig. 7(b) (dotted line). However, this solution needs an extra 6 h processing time per batch than the base case facility as shown in Fig. 8 since larger output volumes from AEX and CEX need to be processed.

Introducing larger vessels is a natural and simple way to solve the bottlenecks caused by tank size mismatching. However, this change can be an expensive solution not only because of the cost of larger vessels but also due to downtime and physical limitations such as space which may incur retrofitting costs during installation. A further shortcoming of this solution is unsustainability. When titre becomes higher, larger vessels are needed again.

4.4.2. *Debottlenecking solution 2—New buffer*

Based on the decision tree result in Fig. 5, reducing the eluate volumes of the CEX and AEX steps to 2.8 CV and 4.0 CV accordingly can avoid product mass loss using the base case facility. Tightening the fluctuations in eluate volumes of AEX and CEX from 50% to 10% would require design of experiment (DoE) studies to be conducted that focus on optimising the buffer components and their pH and conductivity. In Fig. 7(a), the mass output result of solution 2 (dashed line) is as good as solution 1 (dotted line) with most batches meeting the expected throughput of 21 kg/batch. The improvement

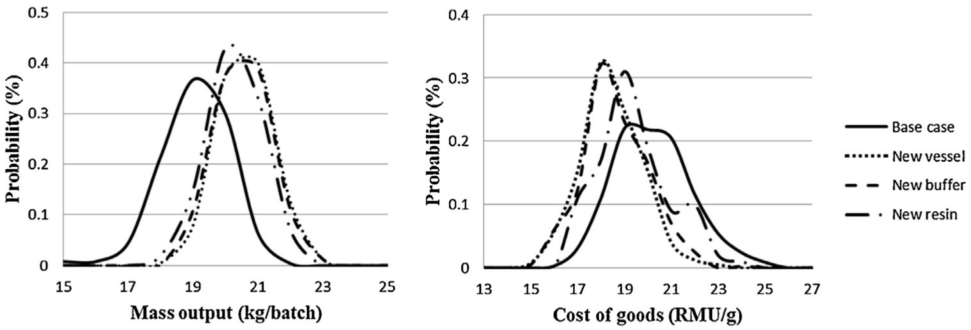

Fig. 7. Probability distributions for the (a) mass output and (b) direct COG/g from the base case facility and three different debottlenecking solutions: new vessel (dotted line), new buffer (dashed line) and new resin (dot–dashed line). RMU = relative monetary units.

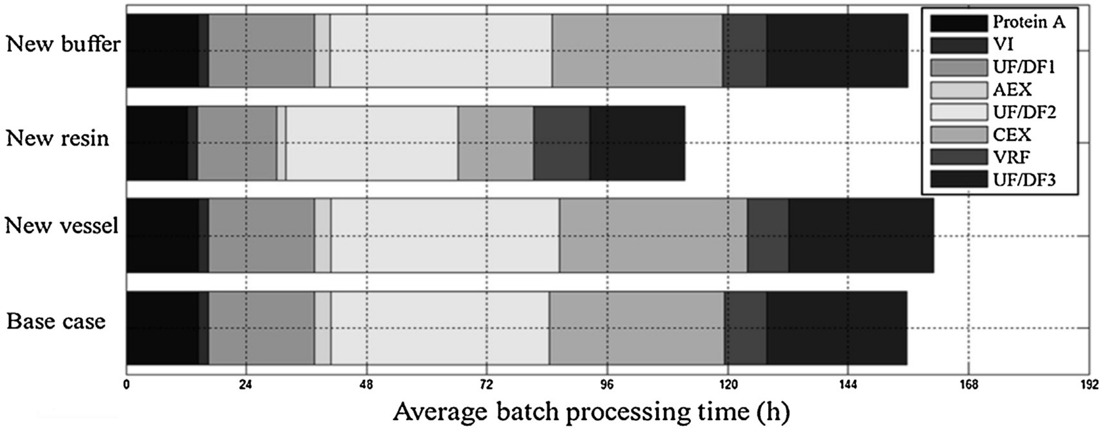

Fig. 8. Average batch processing time for base case facility and different debottlenecking solutions.

in the direct COG/g can also be seen in Fig. 7(b) (dashed line). The average batch processing time is the same as base case facility as shown in Fig. 8.

Generally speaking, solution 2 can give the same performance improvement as solution 1 but does not require larger vessels or space considerations. The cost of solution 2 is buffer optimisation studies which can potentially be cheaper than solution 1, although this depends on the development effort required.

4.4.3. Debottlenecking solution 3—New resin

Using higher capacity resin for AEX and CEX steps can reduce the number of cycles so that the total output volumes of purification steps can be reduced. Newer resins with a dynamic binding capacity of 100 g/L and 40 g/L for AEX and CEX steps, respectively were used in the analysis. In Fig. 7(a), the mass output result of solution 3 (dot–dashed line) is better than the base case but not as good as solutions 1 and 2. The direct COG/g of solution 3 is also higher than solution 2 and 3 in Fig. 7(b) due to the higher price of the newer resins.

Compared to other solutions, the most attractive advantage of solution 3 is the saving in the average batch processing time of two days per batch as shown in Fig. 8. However, if there is already slack in the schedule to meet annual demands, then the saving offered by solution 3 becomes less important.

The above discussion shows how the facility fit analysis can inform a facility manager about the bottlenecks in the process and help to suggest solutions. Three solutions were proposed for the mAb facility; however, the final choice would depend on considerations of likely future constraints such as further anticipated increases in titre or increases in production.

5. Conclusion

This work introduced the CART decision tree method to explore the impact of process fluctuations on product mass loss and to extract rules on the critical combinations of parameter values that lead to mass loss. A series of if-then rules generated by the decision tree method can be used to better understand the fluctuations in key process parameters leading to mass loss, to find out where the critical process constraints are and to predict the product loss. The case study in this work demonstrated that the decision tree results can provide ideas for debottlenecking solutions with different impacts on space requirements, extra expense and processing time. The analysis suggested that narrowing the eluate volume fluctuations expected through buffer optimisation would be an attractive sustainable solution, where possible. Combining this with the new higher capacity resins investigated in the paper would mean that the titre limit of the base case facility could increase from the original 2 g/L to 5 g/L without processes experiencing unexpected mass loss.

The work reported in this paper has examined the impact of new processes with process fluctuations on the mass output of an existing facility. The same methods have potential for other key performance metrics in commercial manufacturing processes such as facility run rate and batch processing time. In addition, such methods can be applied to examine facility bottlenecks that occur not only in the process steps but also in ancillary operations such as utility generation (e.g. water-for-injection, WFI) or buffer preparation.

Acknowledgments

Funding from the UK Engineering & Physical Sciences Research Council (EPSRC) for the EPSRC Centre for Innovative Manufacturing in Emergent Macromolecular Therapies is gratefully acknowledged. Financial support from the consortium of industrial and governmental users is also acknowledged.

References

Abu-Absi, S.F., Yang, L., Thompson, P., Jiang, C., Kandula, S., Schilling, B., Shukla, A.A., 2010. Defining process design space for monoclonal antibody cell culture. Biotechnol. Bioeng. 106, 894–905.

Aldington, S., Bonnerjea, J., 2007. Scale-up of monoclonal antibody purification processes. J. Chromatogr. B 848 (1), 64–78.

Amanullah, A., Otero, J.M., Mikola, M., Hsu, A., Zhang, J., Aunins, J., Schreyer, H.B., Hope, J.A., Russo, A.P., 2010. Novel micro-bioreactor high throughput technology for cell culture process development: reproducibility and scalability assessment of fed-batch CHO cultures. Biotechnol. Bioeng. 106, 57–67.

Birch, J.R., Racher, A.J., 2006. Antibody production. Adv. Drug Deliver. Rev. 58, 671–685.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall, New York, NY.

Buck, K.K.S., Subramanian, V., Block, D.E., 2002. Identification of critical batch operating parameters in fed-batch recombinant *E. coli* fermentations using decision tree analysis. Biotechnol. Progr. 18 (6), 1366–1376.

Cassettari, L., Mosca, R., Revetria, R., 2012. Monte Carlo simulation models evolving in replicated runs: a methodology to choose the optimal experimental sample size. Math. Prob. Eng. 2012, 17, Article ID 463873.

Chang, J., 2011. Process validation challenges for tech transfer. Pharm. Outsourcing 12 (4).

Charaniya, S., Le, H., Rangwala, H., Mills, K., Johnson, K., Karypis, G., Hu, W.S., 2010. Mining manufacturing data for discovery of high productivity process characteristics. J. Biotechnol. 147 (3–4), 186–197.

Coleman, M.C., Buck, K.K.S., Block, D.E., 2003. An integrated approach to optimization of *Escherichia coli* fermentations using historical data. Biotechnol. Bioeng. 84 (3), 274–285.

Edwards-Parton, S., Thornhill, N.F., Bracewell, D.G., Liddell, J.M., Titchener-Hooker, N.J., 2008. Principal component score modeling for the rapid description of chromatographic separations. Biotechnol. Progr. 24 (1), 202–208.

Efron, B., 1983. Estimating the error rate of a prediction rule—improvement on cross-validation. J. Am. Stat. Assoc. 78 (382), 316–331.

Farid, S.S., 2008. Process economic drivers in industrial monoclonal antibody manufacture. In: Gottschalk, U. (Ed.), Process Scale Purification of Antibodies. John Wiley & Sons Inc., Hoboken, NJ, pp. 239–261.

Farid, S.S., Washbrook, J., Titchener-Hooker, N.J., 2005. Decision-support tool for assessing biomanufacturing strategies under uncertainty: stainless steel versus disposable equipment for clinical trial material preparation. Biotechnol. Progr. 21 (2), 486–497.

George, E.D., Farid, S.S., 2008. Strategic biopharmaceutical portfolio development: an analysis of constraint-induced implications. Biotechnol. Progr. 24 (3), 698–713.

Grajski, K.A., Breiman, L., Diprisco, G.V., Freeman, W.J., 1986. Classification of egg spatial patterns with a tree-structured methodology—CART. IEEE Trans. Biomed. Eng. 33 (12), 1076–1086.

Kamarck, M.E., 2006. Building biomanufacturing capacity—the chapter and verse. Nat. Biotechnol. 24 (5), 503–505.

Kelley, B., 2009. Industrialization of mAb production technology: the bioprocessing industry at a crossroads. Mabs 1 (5), 443–452.

Kelley, B., Blank, G., Lee, A., 2009. Downstream processing of monoclonal antibodies: current practices and future opportunities. In: Gottschalk, U. (Ed.), Process Scale Purification of Antibodies. John Wiley & Sons Inc., Hoboken, NJ, pp. 1–23.

Lam, H., Malik, K., 2001. Monitoring and modeling of batch fermentation processes using decision tree analysis, PCA, and PLS. Abstr. Pap. Am. Chem. Soc. 221 (1–2), 52-BIOT 52.

Legmann, R., Schreyer, H.B., Combs, R.G., McCormick, E.L., Russo, A.P., Rodgers, S.T., 2009. A predictive high-throughput scale-down model of monoclonal antibody production in CHO cells. Biotechnol. Bioeng. 104, 1107–1120.

Li, F., Vijayasankaran, N., Shen, A., Kiss, R., Amanullah, A., 2010. Cell culture processes for monoclonal antibody production. Mabs 2 (5), 1–14.

Ma, Y., Peng, Y.Z., Wang, S.Y., Wang, X.L., 2004. Nitrogen removal influence factors in A/O process and decision trees for nitrification/denitrification system. J. Environ. Sci. Chin. 16 (6), 901–907.

Mercier, S.M., Diepenbroek, B., Dalm, M.C.F., Wijffels, R.H., Streefland, M., 2013. Multivariate data analysis as a PAT tool for early bioprocess development data. J. Biotechnol. 167 (3), 262–270.

Moreno-Torres, J.G., Saez, J.A., Herrera, F., 2012. Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Trans. Neural Networks Learn. Syst. 23 (8), 1304–1312.

Pate, M.E., Turner, M.K., Thornhill, N.F., Titchener-Hooker, N.J., 1999. The use of principal component analysis for the modelling of high performance liquid chromatography. Bioprocess Eng. 21 (3), 261–272.

Piccarreta, R., 2008. Classification trees for ordinal variables. Comput. Stat. 23 (3), 407–427.

Pollock, J., Ho, S.V., Farid, S.S., 2013. Fed-batch and perfusion culture processes: operational, economic and environmental feasibility under uncertainty. Biotechnol. Bioeng. 110 (1), 206–219.

Quinlan, J.R., 1996. Learning decision tree classifiers. ACM Comput. Surv. 28 (1), 71–72.

Rodgers, J.L., Nicewander, W.A., 1988. 13 ways to look at the correlation-coefficient. Am. Statistician 42 (1), 59–66.

Rommel, S., Schuppert, A., 2004. Data mining for bioprocess optimization. Eng. Life Sci. 4 (3), 266–270.

Sadras, V., Bongiovanni, R., 2004. Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks. Field Crops Res. 90 (2–3), 303–310.

Sin, G., Gernaey, K.V., Eliasson Lantz, A., 2009. Good modelling practice (GMoP) for PAT applications: propagation of input uncertainty and sensitivity analysis. Biotechnol. Progr. 25, 1043–1053.

Stonier, A., Pain, D., Westlake, A., Hutchinson, N., Thornhill, N.F., Farid, S.S., 2013. Integration of stochastic simulation with multivariate analysis: short-term facility fit prediction. Biotechnol. Progr. 29 (2), 368–377.

Stonier, A., Simaria, A.S., Smith, M., Farid, S.S., 2012. Decisional tool to assess current and future process robustness in an antibody purification facility. Biotechnol. Progr. 28 (4), 1019–1028.

Thornhill, N.F., Melbo, H., Wiik, J., 2006. Multidimensional visualization and clustering of historical process data. Ind. Eng. Chem. Res. 45 (17), 5971–5985.

Xu, G., Papageorgiou, L.G., 2009. A mixed integer optimisation model for data classification. Comput. Ind. Eng. 56 (4), 1205–1215.