Journal of Biotechnology 162 (2012) 210–223

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

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

Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator

Huong Le[a,1], Santosh Kabbur[b,1], Luciano Pollastrini[c], Ziran Sun[c], Keri Mills[c], Kevin Johnson[c], George Karypis[b], Wei-Shou Hu[a,*]

[a] Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, USA
[b] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[c] Genentech, Inc., Vacaville, CA 95688, USA

ARTICLE INFO

Article history:
Received 29 April 2012
Received in revised form 31 July 2012
Accepted 30 August 2012
Available online xxx

Keywords:
Bioprocess data mining
Multivariate data analysis
Support vector regression
Partial least square regression
Lactate consumption
Cell culture
Chinese hamster ovary (CHO) cells

ABSTRACT

Multivariate analysis of cell culture bioprocess data has the potential of unveiling hidden process characteristics and providing new insights into factors affecting process performance. This study investigated the time-series data of 134 process parameters acquired throughout the inoculum train and the production bioreactors of 243 runs at the Genentech's Vacaville manufacturing facility. Two multivariate methods, kernel-based support vector regression (SVR) and partial least square regression (PLSR), were used to predict the final antibody concentration and the final lactate concentration. Both product titer and the final lactate level were shown to be predicted accurately when data from the early stages of the production scale were employed. Using only process data from the inoculum train, the prediction accuracy of the final process outcome was lower; the results nevertheless suggested that the history of the culture may exert significant influence on the final process outcome. The parameters contributing most significantly to the prediction accuracy were related to lactate metabolism and cell viability in both the production scale and the inoculum train. Lactate consumption, which occurred rather independently of the residual glucose and lactate concentrations, was shown to be a prominent factor in determining the final outcome of production-scale cultures. The results suggest possible opportunities to intervene in metabolism, steering it towards the type with a strong propensity towards high productivity. Such intervention could occur in the inoculum stage or in the early stage of the production-scale reactors. Overall, this study presents pattern recognition as an important process analytical technology (PAT). Furthermore, the high correlation between lactate consumption and high productivity can provide a guide to apply quality by design (QbD) principles to enhance process robustness.

1. Introduction

In recent years, cell culture bioprocessing has seen a tremendous growth in data generation and collection. In modern manufacturing facilities, it is not uncommon to encounter hundreds of process parameters being monitored and acquired automatically every few seconds throughout the entire production train. This enormous volume of data further accumulates across multiple campaigns and at multiple manufacturing sites. Mining these historical data holds promise to gain insights into fluctuations in process performance, uncover hidden characteristics of high-performing cultures, and discern process parameters with pivotal contributions to the overall process performance.

Cell culture bioprocess data, however, pose significant challenges to mining practices due to the inherent heterogeneities in time scale and data type (Charaniya et al., 2008). Yet many have successfully applied an array of classification and prediction techniques to investigate hidden process patterns. Principal component analysis (PCA), partial least square regression (PLSR), and other unsupervised techniques, which have the advantage of capturing the interactions among process parameters, have been used for detecting state transitions related to product and lactate formation, online monitoring, fault detection and diagnosis, scale-up assessment, process characterization, and root cause analysis (Bachinger et al., 2000; Gunther et al., 2007; Kirdar et al., 2008; Ündey, 2004). In other studies, powerful supervised approaches such as decision tree (DT), artificial neural network (ANN), and support vector regression (SVR) were used to optimize a control scheme incorporating time-course data, predict the final process

* Corresponding author at: 421 Washington Avenue SE, Minneapolis, MN 55455-0132, USA. Tel.: +1 612 626 7630; fax: +1 612 626 7246.
E-mail address: wshu@umn.edu (W.-S. Hu).
[1] These authors contributed equally to this work.

0168-1656/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbiotec.2012.08.021

outcome, and reveal key parameters (Buck et al., 2002; Charaniya et al., 2010; Coleman and Block, 2006). Among these multivariate analysis approaches, PLSR and SVR appear to be well-suited to handle the various challenges associated with bioprocess data, namely high-dimensionality and co-linearity between various parameters.

Among the important contributors to differentiating between high- and low-productivity runs of a cell culture process are parameters related to lactate metabolism, including pH, base addition, osmolarity, dissolved CO_2, and lactate concentration (Charaniya et al., 2010). Excessive lactate accumulation has long been known to be an impediment to achieving high cell concentration and superior productivity (Glacken et al., 1986; Hu et al., 1987). Introducing metabolic shifts (i.e., controlling lactate production at low levels or, to a further extent, inducing lactate consumption) has been achieved through various strategies. These approaches include dynamic feeding to control glucose at low levels (Cruz et al., 1999; Zhou et al., 1997), using alternative carbon sources (Altamirano et al., 2006; Wlaschin and Hu, 2007), knocking down LDH-A (Chen et al., 2001; Kim and Lee, 2007a), and enhancing glucose carbon flux into the TCA cycle (Irani et al., 1999; Kim and Lee, 2007b). Understanding the linkage between lactate metabolism and high productivity thus offers the opportunity to discover the metabolic signatures of these high-performing processes.

In this study, we employed support vector regression (SVR) and partial least square regression (PLSR) methods to predict the final process outcome using process data from 243 production runs at a Genentech manufacturing facility. This dataset comprises 134 temporal parameters acquired online and offline throughout the inoculum train (80 L, 400 L, and 2000 L) and the production-scale bioreactors (12,000 L). Parameters pivotal to prediction accuracy were assessed based on two criteria: the frequency of occurrence (f) in the best parameter sets for SVR models and the magnitude of the regression coefficient (β) in the optimal PLSR models. Among these pivotal parameters, various aspects of the lactate consumption phenomenon at the production scale in high-titer runs were further investigated.

2. Methods

2.1. Data pre-processing and organization

Process data from 243 production runs of a recombinant IgG molecule, produced using the same Chinese hamster ovary (CHO) cell line, were used for analysis. The same batch process was applied for all seed cultures (80 L, 400 L, and 2000 L). At the production scale (12,000 L), a fed-batch mode with glucose and medium feeding was used. Across these scales, temperature, pH, and dissolved oxygen were maintained at 37 °C, 7.0, and 30% of air saturation, respectively. A temperature shift to 33 °C at approximately 70 h post-inoculation was performed at the 12,000 L scale.

The data were pre-processed as described previously (Charaniya et al., 2010) with minor modifications. Briefly, online data acquired at each of the four scales (80 L, 400 L, 2000 L, and 12,000 L) were smoothed using a moving window average method with a time window of 100 min. Offline data were linearly interpolated and/or extrapolated every 20 h. Furthermore, specific rates of lactate production, glucose consumption, and cell growth were derived from these measured parameters and smoothed using third-order polynomials. In total, time-series data of 134 temporal process parameters across all scales, including 33 parameters at each of the three inoculum scales and 35 parameters at the production scale, were used (Table 1).

Process data from all scales were organized into eight individual and seven cumulative datasets as shown in Table 2. The first dataset comprised process data from the 80 L scale bioreactors. The second dataset contained data from the next scale of 400 L, and so on. Since the run time at the production scale (260 h) was much longer compared to that at each of the inoculum scales (70 h), it was segregated into several stages: up to 70 h, 120 h, 170 h, 220 h, and 260 h. In addition to these eight individual datasets, process data were also accumulated across scales with the largest dataset compiling data from 80 L, 400 L, 2000 L, and up to 260 h of the 12,000 L scale.

Table 1
Temporal process parameters used in the analysis: 33 parameters at each of the inoculum scales (80 L, 400 L, and 2000 L), and 35 parameters at the production scale (12,000 L).

Offline parameters	Online parameters
Ammonium ion concentration	Air sparge rate
Dissolved CO_2 (pCO_2)	Air sparge set point
Dissolved O_2 (pO_2)	Backpressure (12,000 L only)
Glucose concentration	CO_2 sparge rate
Integrated packed cell volume (IntvPCV) (12,000 L only)	Dissolved oxygen (DO) controller output
Lactate concentration	DO (primary)
Osmolarity	DO (secondary)
Packed cell volume (PCV)	Flowrate overlay
pH (offline)	Jacket temperature
Sodium ion concentration	O_2 sparge rate
Viability	pH controller output
Viable cell density (VCD)	pH (online)
Derived parameters	Pressure exhaust valve
Specific cell growth rate (μ)	Reactor weight
Specific glucose consumption rate (q_{Lac})	Total air sparged
Specific lactate consumption rate (q_{Glc})	Total base added
	Total CO_2 sparged
	Total O_2 sparged
	Total gas sparged
	Vessel temperature

2.2. Model training and evaluation using 10-fold cross-validation

A 10-fold cross-validation scheme as shown in Fig. 1 was used for training and evaluation of both support vector regression (SVR) and partial least square regression (PLSR) models. Process data from 243 runs in each of the 15 datasets described above were randomly divided into ten subsets of approximately equal sizes. During each round of cross-validation, nine of the ten subsets were used as the training set on which model optimization was performed. The best performing model on each training set was used to predict process outcome of runs in the corresponding, unseen test set (the 10th subset). This process was repeated 10 times on different pairs of training and test subsets. Model performance was evaluated using the Pearson's correlation coefficient (r) r and the root mean square error (ε) between the predicted and the actual final process outcome:

$$r = \frac{\sum_{i=1}^{n} y_i f(x_i) - (\sum_{i=1}^{n} y_i \sum_{i=1}^{n} f(x_i))/n}{\sqrt{(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n)(\sum_{i=1}^{n} f(x_i)^2 - (\sum_{i=1}^{n} f(x_i))^2/n)}} \quad (1)$$

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}} \quad (2)$$

where n is the number of runs, and y_i and $f(x_i)$ are the actual and the predicted titer values of run i, respectively. The model performance was averaged across the 10 folds. As a baseline for evaluating model performance, a random predictor with one million simulations of randomized final process outcome was generated.

To get a better estimate of the generalization error of the constructed models, model optimization (i.e. selection of model and

Table 2
Prediction accuracy of PLSR and SVR models using process data acquired at different stages, evaluated as the Pearson's correlation coefficient (r) and the root mean square error (ε) between the predicted and the actual final process outcome.

Dataset	Final antibody concentration (titer)				Final lactate concentration			
	PLSR		SVR		PLSR		SVR	
	r	ε	r	ε	r	ε	r	ε
80 L	0.42	0.10	0.40	0.10	0.44	4.18	0.43	4.15
400 L	0.21	0.12	0.43	0.10	0.33	4.79	0.43	4.15
2000 L	0.28	0.11	0.35	0.10	0.28	4.63	0.37	4.30
12,000 L up to 70 h	0.73	0.07	0.73	0.07	0.72	3.18	0.77	2.93
12,000 L up to 120 h	0.80	0.06	0.77	0.07	0.85	2.41	0.78	2.84
12,000 L up to 170 h	0.88	0.05	0.86	0.06	0.95	1.47	0.92	1.98
12,000 L up to 220 h	0.92	0.04	0.91	0.04	0.97	1.09	0.96	1.56
12,000 L up to 260 h	0.92	0.04	0.92	0.04	0.97	1.13	0.98	1.33
80 L+400 L	0.41	0.10	0.47	0.09	0.50	3.97	0.48	4.00
80 L+400 L+2000 L	0.45	0.09	0.48	0.09	0.45	4.10	0.50	3.94
80 L+400 L+2000 L+12,000 L up to 70 h	0.71	0.07	0.68	0.08	0.73	3.13	0.68	3.39
80 L+400 L+2000 L+12,000 L up to 120 h	0.77	0.07	0.72	0.08	0.76	2.94	0.73	3.24
80 L+400 L+2000 L+12,000 L up to 170 h	0.88	0.05	0.83	0.06	0.90	1.97	0.87	2.65
80 L+400 L+2000 L+12,000 L up to 220 h	0.91	0.04	0.91	0.05	0.97	1.18	0.95	2.00
80 L+400 L+2000 L+12,000 L up to 260 h	0.92	0.04	0.92	0.05	0.97	1.14	0.96	1.82

For a random predictor: $r = 0.00$, $\varepsilon = 0.17$ and $r = 0.00$, $\varepsilon = 7.79$ when the final titer and the final lactate concentration was used as the objective function, respectively.

process parameters) was further performed on each training set, also using 10-fold cross-validation, as shown in the shaded box in Fig. 1. Model optimization was performed for each round of the 10-fold cross-validation. This involved further partitioning of the training set randomly into 10 smaller groups of about equal sizes. The model was trained on nine groups using a different set of parameters for SVR approach or PLS factors for PLSR approach. The performance of the resulting model was subsequently tested

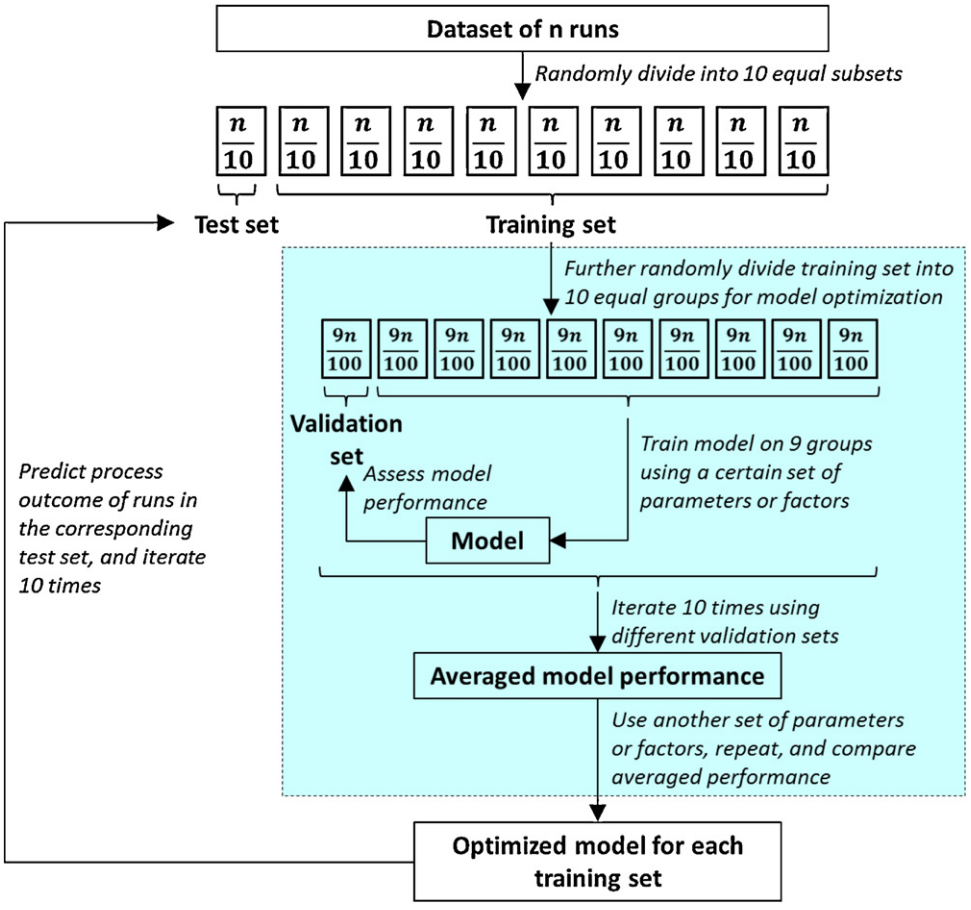

Fig. 1. Scheme of 10-fold cross-validation with model optimization which was used for both multivariate approaches: SVR and PLSR. All n process runs ($n = 243$ in this study) were randomly separated into 10 equal subsets. Nine were used as the training set on which model optimization was performed, and the optimized model was used to predict the outcome of runs in the 10th subset (test set). Model optimization involved further random separation of the training set into 10 equal groups. Again, nine were used to train a model with a certain set of parameters (for SVR approach) or PLS factors (for PLSR approach), and the performance of this model was tested on the 10th group (validation set). This process was repeated 10 times to obtain the average performance for each set of parameters/factors, which was later compared to identify the parameter/factor set that resulted in the best predictive (optimized) model. The shaded box contains all steps in model optimization.

in the 10th group, called the validation set. This procedure was repeated 10 times for each set of parameters or PLS factors. The average performance of the model over these inner 10 folds was used to determine the optimal set of parameters or PLS factors for each round of the outer 10-fold cross-validation. Subsequently, the best model was selected and used to predict the outcome of runs in the corresponding, unseen test set.

2.3. Construction of partial least square regression (PLSR) models

Partial least square regression (PLSR) models were constructed using the SIMPLS algorithm (Chong and Jun, 2005; de Jong, 1993). Time-series data for each process parameter were extracted every 10 h, resulting in multiple discrete "variables" originating from the same parameter. These variables were concatenated over the run time of each scale into a data matrix (**X**). Data in each column of this matrix were further autoscaled to a mean of zero and a standard deviation of one to give a new matrix **X0**. A similar transformation was also performed on the response vector (**y**) to obtain the autoscaled final process outcome (**y0**) (either antibody titer or lactate concentration at the end of the 12,000 L cultures).

The autoscaled data matrix (**X0**) was projected onto mutually orthogonal PLS factors (**XS**), each of which is a weighted linear combination of the original variables in **X0**. A set of these PLS factors can be used to construct a regression function to predict the autoscaled final process outcome in **y0**. The SIMPLS algorithm for a univariate response in **y0** can be simplified in the following equations:

$$\mathbf{XS}_{n\times a} = \mathbf{X0}_{n\times p} \cdot \mathbf{W}_{p\times a} \quad (3)$$

$$\mathbf{X0}_{n\times p} = \mathbf{XS}_{n\times a} \cdot \mathbf{XL}^{\mathrm{T}}_{p\times a} + \mathbf{XE}_{n\times p} \quad (4)$$

$$y0_{n\times 1} = \mathbf{XS}_{n\times a} \cdot \boldsymbol{\beta}_{a\times 1} + \mathbf{ye}_{n\times 1} \quad (5)$$

such that the covariance between **X0** and **y0** is maximized.

In these equations, **XS**, **X0**, and **W** are the matrix of orthogonal PLS factors, the autoscaled data matrix, and the matrix of PLS weights, respectively. The matrices **XL** and **XE** contain loadings of the PLS factors and the residuals when factorizing **X0** into a product of **XS** and $\mathbf{XL}^{\mathrm{T}}$, respectively. The vectors **y0**, **β**, and **ye** comprise the autoscaled response, the regression coefficients of **y0** using **XS**, and the residuals when regressing **y0** using **XS**, respectively. The variables n, p, and a are the number of process runs, the number of variables (in this case, a product between the number of process parameters m and the number of time points t), and the number of PLS factors used for regression, respectively.

The *plsregress* subroutine, an implementation of the SIMPLS algorithm in the Matlab's statistics toolbox, was used for constructing the PLSR models. For each of the 15 datasets, a PLSR model was constructed and optimized as described in model training and evaluation using 10-fold cross-validation. The number of PLS factors in each model was varied from one to the maximum possible (which is the rank of the data matrix **X0**). For each fold of the outer 10-fold cross-validation, an optimal set of PLS factors, and thus variables, could be identified. Furthermore, as each original parameter was discretized into multiple variables, the average magnitude of the regression coefficients of all variables which originated from the same parameter was used to assess the importance of that parameter.

2.4. Construction of support vector regression (SVR) models

LIBSVM (Chang and Lin, 2001), an implementation of the SVR algorithm in C, was used to construct ν-SVR models as described previously (Charaniya et al., 2010) with several modifications. For each individual parameter, the Euclidean distance between any two runs i and j was computed and scaled to a range from 0 to 1. This scaled distance (d_{ij}) was converted into a similarity value ($s_{ij} = 1 - d_{ij}$) and organized into a matrix for all pairwise comparisons of runs ($n \times n$, where n is the number of runs). The similarity matrices of all parameters were linearly combined to form a final similarity matrix, which was used as a pre-defined kernel in the ν-SVR algorithm. Upon combination, each parameter was either given equal weights of $1/m$ (where m is the number of process parameters) or weighted according to how well it correlates to the final process outcome as described previously (Charaniya et al., 2010). Thus all entries in the final similarity matrix were maintained between 0 and 1. The objective function (y), either the final titer or the final lactate concentration, was also scaled to the same range of 0–1.

ν-SVR models were constructed and optimized for each of the eight individual datasets as described in model training and evaluation using 10-fold cross-validation and in Section 2.5. For the seven cumulative datasets, due to computational constraints imposed by the large number of parameters, SVR models were built using the best performing sets of parameters obtained for the corresponding individual datasets. In addition, a simple grid search within the range of 0–1 with 0.1 intervals was performed on the cost function. The best value was used in the subsequent step of model optimization to identify pivotal process parameters.

2.5. Identification of pivotal process parameters using SVR approach

A greedy parameter selection approach based on the wrapper feature-selection method (Liu and Hiroshi, 1998) was used to find the best performing set of parameters for the SVR models. This approach determined the suitability of a set of features (i.e., process parameters) by first building an SVR model using these features and then assessing its performance on a subset of the data that was not used for training (i.e., validation set). The set of features whose model achieved the best performance on the validation set became the set of selected parameters. Since each of the eight individual datasets contains either 33 parameters at the inoculum scales or 35 parameters at the production scale, a direct application of the wrapper feature-selection method will require an evaluation of $2^{33} - 1$ or $2^{35} - 1$ (excluding the null set) possible parameter subsets, which is prohibitively large. For this reason, we employed a greedy strategy that only considers a substantially smaller number of parameter subsets.

In this approach, the different parameter subsets were organized into a lattice structure, whose ith level contained all the subsets of size $(m - i)$ where m is the number of parameters. All nodes at each level were connected to the nodes of the preceding level that were its supersets. The algorithm started by evaluating the performance of the subsets at levels 0 and 1 (i.e., the entire set of m parameters and the m subsets that were obtained by removing one parameter, respectively). Among the m subsets at level 1, N subsets whose models achieved the best performance on the validation set were retained. The algorithm then proceeded to evaluate the performance of subsets at level 2 that are descendants of at least one of the N nodes retained at level 1. Among those subsets, it also retained the N best performing ones. This process continued until the last level of the lattice. Note that, by setting N to a small value (in our experiments, $N = \{5, 15, 25, 35\}$) and by considering only subsets whose supersets were among the N best performing subsets of the previous level, the total number of subsets being considered became computationally feasible. In addition, since the subsets that were pruned are those that did not perform well, this approach could still identify good performing parameter subsets.

Since 10-fold cross-validation was performed, each fold generated an optimal set of parameters. Thus the occurrence frequency

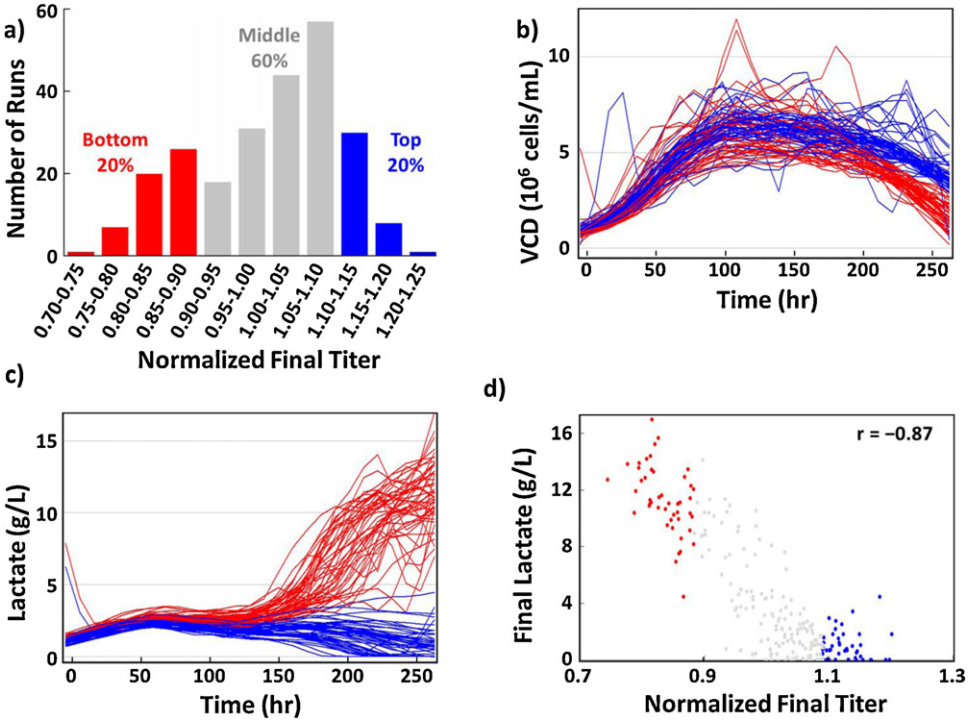

Fig. 2. Differences in process performance as indicated by the final antibody concentration (titer), viable cell density (VCD), and lactate concentration across 243 production runs. (a) Distribution of the final titer (normalized such that the average across all runs is 1.00). Roughly 20% of runs have final titers greater than 1.10 (top 20%—in blue); 20% of runs have titers less than 0.90 (bottom 20%—in red); and 60% of runs have titers between 0.90 and 1.10 (middle 60%—in gray). (b) Variation in viable cell density at 12,000 L scale between runs in the top 20% (blue) and the bottom 20% (red). (c) Variation in lactate concentration at 12,000 L scale between runs in the top 20% (blue) and the bottom 20% (red). (d) High correlation (−0.87) between the final lactate concentration and the final titer across all runs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(f) of each parameter over all 10 folds can be used as an estimate of its contribution to the overall model performance.

3. Results

3.1. High- and low-performing runs exhibit distinct process characteristics

The 243 production runs investigated in this study exhibited considerable variation in a number of process parameters and outcome as shown in Fig. 2. The pre-harvest recombinant antibody concentration (final titer), previously normalized to an average of 1.00, varied across a wide range from 0.70 to 1.25 (Fig. 2a). These runs were categorized into three classes: top 20% (in blue), middle 60% (in gray), and bottom 20% (in red), with their final titer approximately over 1.10, between 1.10 and 0.90, and below 0.90, respectively. Because of measurement error of recombinant antibody concentration, it is possible that runs within the middle 60% class have a high degree of similarity. In contrast, comparison of the top 20% and the bottom 20% runs should reveal distinct characteristics of high-titer cultures.

As shown in Fig. 2b, both the top and bottom 20% cultures started with a similar range of cell concentration in the production-scale bioreactors. There was a substantial spread of cell concentration at peak growth and at the end of the culture, even among runs within the top or bottom 20% class. In general, the top 20% runs reached higher peak cell concentrations (between 100 and 150 h) although the range was rather wide. It is apparent that more runs of the bottom 20% class had lower peak cell concentrations, and all bottom 20% runs had lower viable cell concentrations at the end of the production run.

The lactate concentration profiles also showed profound differences between the top 20% and bottom 20% runs (Fig. 2c). Although lactate concentrations were in similar ranges in all cultures initially, by the time cell concentration reached the peak, they had become higher in the bottom 20% runs. Despite a period between 100 and 130 h during which lactate production subsided, all bottom 20% runs proceeded to return to the lactate production state whereas nearly all top 20% runs switched to the lactate consumption state. Many of these top 20% runs resulted in complete exhaustion of lactate previously produced during the exponential growth stage. As expected, the final lactate concentration was found to be highly correlated to the product yield in all runs with a Pearson's correlation coefficient of −0.87 (Fig. 2d), indicating a close connection between cellular metabolic activities and product titer.

3.2. Process outcome is predicted accurately using multivariate models

Two multivariate regression approaches, support vector regression (SVR) and partial least square regression (PLSR), were employed. Time-series process data were acquired for 134 online, offline, and derived parameters throughout the inoculum train (80 L, 400 L, and 2000 L) and the production-scale bioreactors (12,000 L). To investigate the importance of these parameters at each scale, process data were organized into 15 different datasets as shown in Table 2. In addition to the final titer, the final lactate concentration was also used as an objective function because of the indication of its being an attribute of a culture's performance. The prediction accuracy of these models was assessed based on the Pearson's correlation coefficient (r) and the root mean square error (ε) as described in model training and evaluation using 10-fold cross-validation.

In constructing SVR models, a grid search of the cost function in the range of 0–1 with 0.1 intervals yielded an optimal value of 1, which was used for constructing subsequent SVR models. Both

differential and equal weighting schemes as described in Section 2.4 were employed to combine all similarity matrices. Since the equal weighting scheme resulted in slightly better model performance (data not shown), it was used for the subsequent step of feature selection. A wrapper-based feature selection algorithm as described in Section 2.5 was further employed to identify the optimal combination of parameters that result in the lowest root mean square error (ε). The top 35 nodes (i.e., $N = 35$) were expanded at each level. Three additional values of $N = \{25, 15, 5\}$ were also evaluated, and resulted in similar performances for the 8th dataset

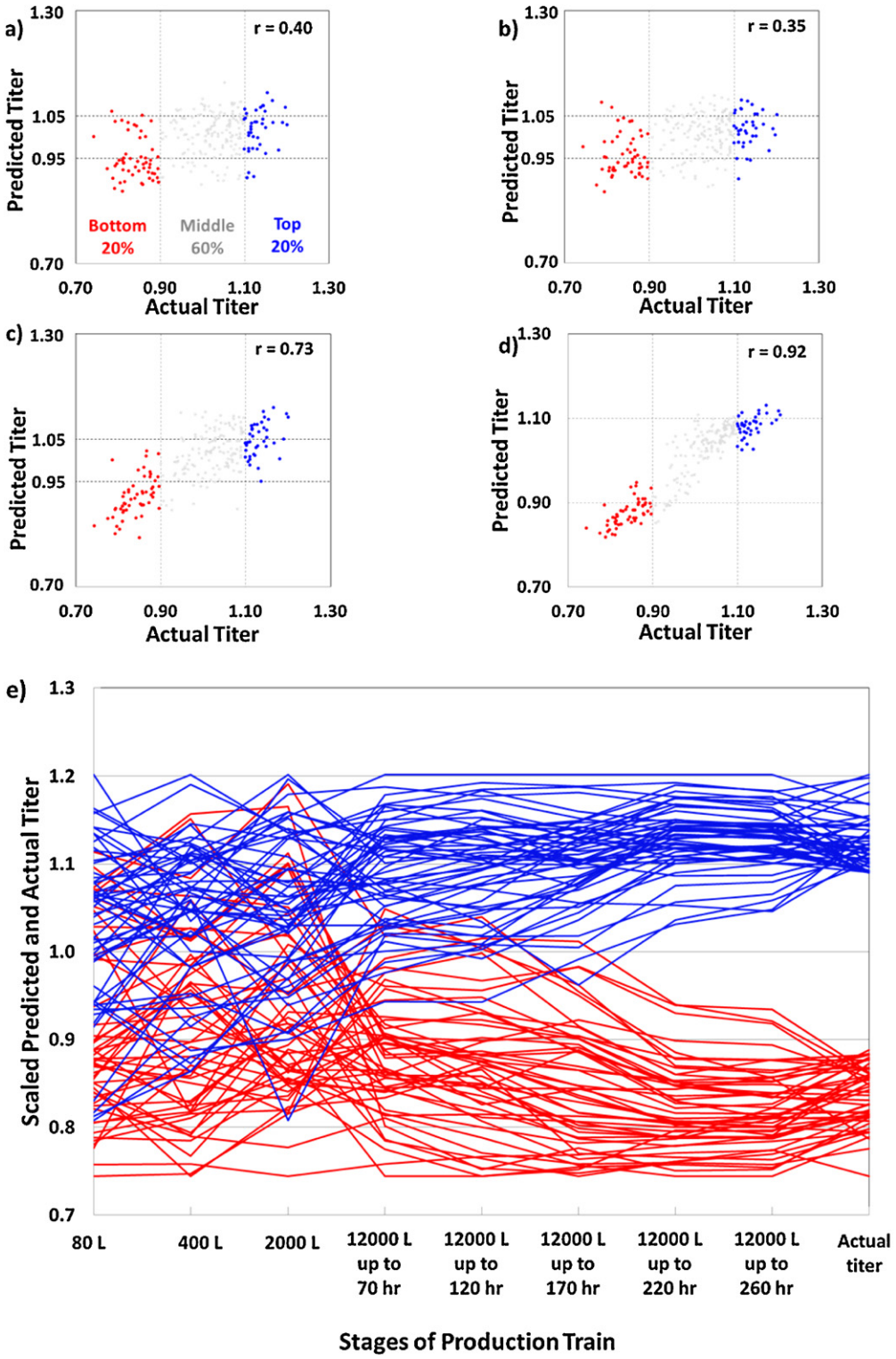

Fig. 3. SVR models' prediction accuracy of the final titer using different datasets. The correlation coefficient (r) between the predicted and the actual titer is shown for each case. The dashed lines indicate the separation of the top 20%, middle 60%, and bottom 20% of runs based on the predicted titer (y-axis) or the actual titer (x-axis). Runs in the top 20% class based on the actual titer are colored in blue; runs in the middle 20% class are colored in gray; and runs in the bottom 20% class are colored in red. (a) 80 L scale. (b) 2000 L scale. (c) Up to 70 h of 12,000 L scale. (d) Up to 260 h of 12,000 L scale. (e) The progression of predicted titer is shown over the course of the cultures for the top and the bottom 20% runs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(12,000 L up to 260 h). Thus, for the other individual datasets, the maximum number of nodes to be expanded at each level was fixed at 35.

Similarly, in constructing PLSR models, a 10-fold cross-validation scheme (Fig. 1) was used to find the optimal number of PLS factors to be incorporated in each model that gives the best predicted final process outcome. As described in model training and evaluation using 10-fold cross-validation, the number of PLS factors was varied from one to the rank of the data matrix **X0**. The optimized number of PLS factors in each training set was used to construct a PLSR model for the corresponding test set.

It is interesting that prediction trends across different datasets were considerably similar irrespective of the multivariate approach as shown in Table 2. Overall, PLSR approach appeared to result in slightly better models than those constructed using SVR approach. However, when the input data were noisy (400 L and 2000 L), these PLSR models failed to maintain good performances whereas SVR models built using the same datasets were still robust. Furthermore, similar correlations between the predicted and the actual final process outcome were observed across all datasets regardless of whether the final titer or the final lactate concentration was used as the objective function. This result indicates that product yield and cellular metabolic activities are indeed closely interconnected, confirming the high correlation between these two characteristics as previously shown in Fig. 2d. Due to this considerable similarity in prediction accuracy, results are presented for SVR models predicting the final titer herein. It is noteworthy that a random predictor generates a root mean square error of 0.17 and 7.80 for the final titer and the final lactate concentration, respectively, and a Pearson's correlation coefficient of zero in both cases.

Data acquired at the smallest scale of the inoculum train (80 L) were moderately indicative of the final titer with a correlation coefficient (r) of 0.40 and a root mean square error (ε) of 0.10 (Fig. 3a). The SVR model constructed using data from the next scale of 400 L performed slightly better with $r = 0.43$ and $\varepsilon = 0.10$. Data from 2000 L scale bioreactors, surprisingly, were less informative than data from the two smaller scales. The correlation coefficient dropped to 0.35 and the error remained at 0.10 as shown in Fig. 3b. This reduced performance appeared to be circumvented by concatenating data across these scales. The SVR model built upon data concatenated from 80 L and 400 L scales exhibited a slight improvement compared to those built using data from each individual scale ($r = 0.47$, $\varepsilon = 0.09$). Similarly, cumulative data across all three scales of the inoculum train resulted in a correlation coefficient of 0.48 and an error of 0.09.

When data from the first 70 h of the production scale was used, the prediction accuracy increased sharply to 0.73 with a root mean square error of 0.07 (Fig. 3c). By the time most runs reached peak growth at 120 h, the performance improved to $r = 0.77$ and $\varepsilon = 0.07$. As the runs approached the end, the final titer could be predicted with higher correlation coefficients of 0.86 ($\varepsilon = 0.06$) by 170 h, and 0.92 ($\varepsilon = 0.04$) upon completion at 260 h (Fig. 3d). Interestingly, the regression models built upon data acquired at the production scale alone were slightly more predictive compared to those with the addition of data from the inoculum train. Concatenating data from the inoculum train to the first 70 h of the production scale actually reduced the prediction accuracy from $r = 0.73$ and $\varepsilon = 0.07$ to $r = 0.68$ and $\varepsilon = 0.08$. At around peak growth (~120 h), addition of inoculum data did not result in a better model ($r = 0.72$, $\varepsilon = 0.08$ compared to $r = 0.77$, $\varepsilon = 0.07$). Similarly, concatenating inoculum data with data from the late stage of the production scale also did not improve prediction accuracy. The model built upon concatenating all data showed little to no improvement ($r = 0.92$, $\varepsilon = 0.05$) over the model built on data from the production scale only ($r = 0.92$, $\varepsilon = 0.04$). This result suggests that the inoculum data are rather noisy relative to the production-scale data and incorporation of these data may not help increase model prediction accuracy.

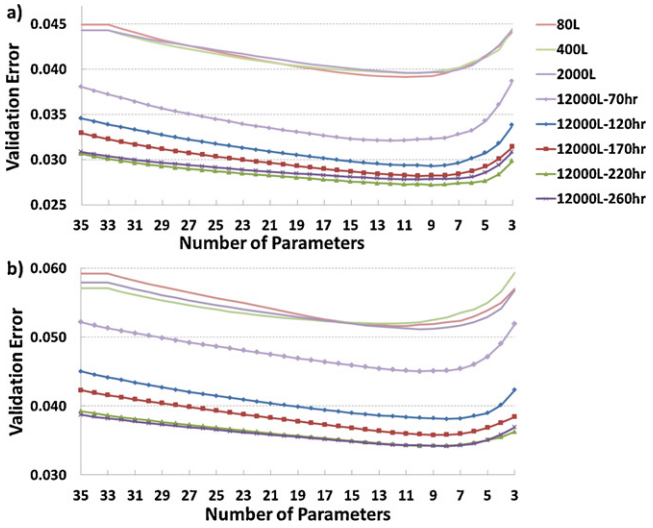

Fig. 4. Variation of validation error as a function of the number of parameters. (a) Final titer as the objective function. (b) Final lactate concentration as the objective function.

Furthermore, as evident from Fig. 3a and b, using data from the 80 L and 2000 L scales, only a few runs predicted to be in the top 20% class (above the horizontal grid line of $y = 1.05$) actually fell to the bottom 20% class of the actual titer (on the left of the vertical grid line of $x = 0.90$). Similarly, the number of runs predicted to be in the bottom 20% class (below the horizontal grid line of $y = 0.95$) that ended up in the top 20% class (on the right of the vertical grid line of $x = 1.10$) was also small. Once data from the production scale, even as early as the first 70 h, was used, this class switch was not observed in any runs (Fig. 3c and d). This result indicates that process characteristics at the early stage of the production scale are already indicative of the final outcome, and no runs are inclined to switch between the top and the bottom classes after this stage.

We next examined those few runs which switched classes by tracking their performance over the course of the run. The performance as judged by titer predicted using each of the eight individual datasets is shown in Fig. 3e. For better visualization, the titer values predicted using each dataset were linearly scaled such that they were in the same range as the actual titer values throughout. Again red and blue colors indicate the bottom and the top 20% of runs, respectively. It is interesting to note that class switch occurred relatively gradually over different stages of the inoculum train (80 L, 400 L, and 2000 L). By 70 h of the production scale, switching has virtually completed. The results suggest that intervention may be carried out prior to that time point of the production scale to influence the outcome.

3.3. *Majority of pivotal parameters are related to cell growth and lactate metabolism*

The contribution of each parameter to the prediction of the final process outcome was assessed using two criteria: the magnitude of the regression coefficient (β) in the optimized PLSR models and the frequency of occurrence (f) in the selected parameter sets for SVR models. As described in Section 2.5, a *wrapper-based feature selection* algorithm was employed to identify the minimum combination of parameters that results in an SVR model with the lowest validation error. Shown in Fig. 4 is this error as a function of the number of parameters incorporated into SVR models at each

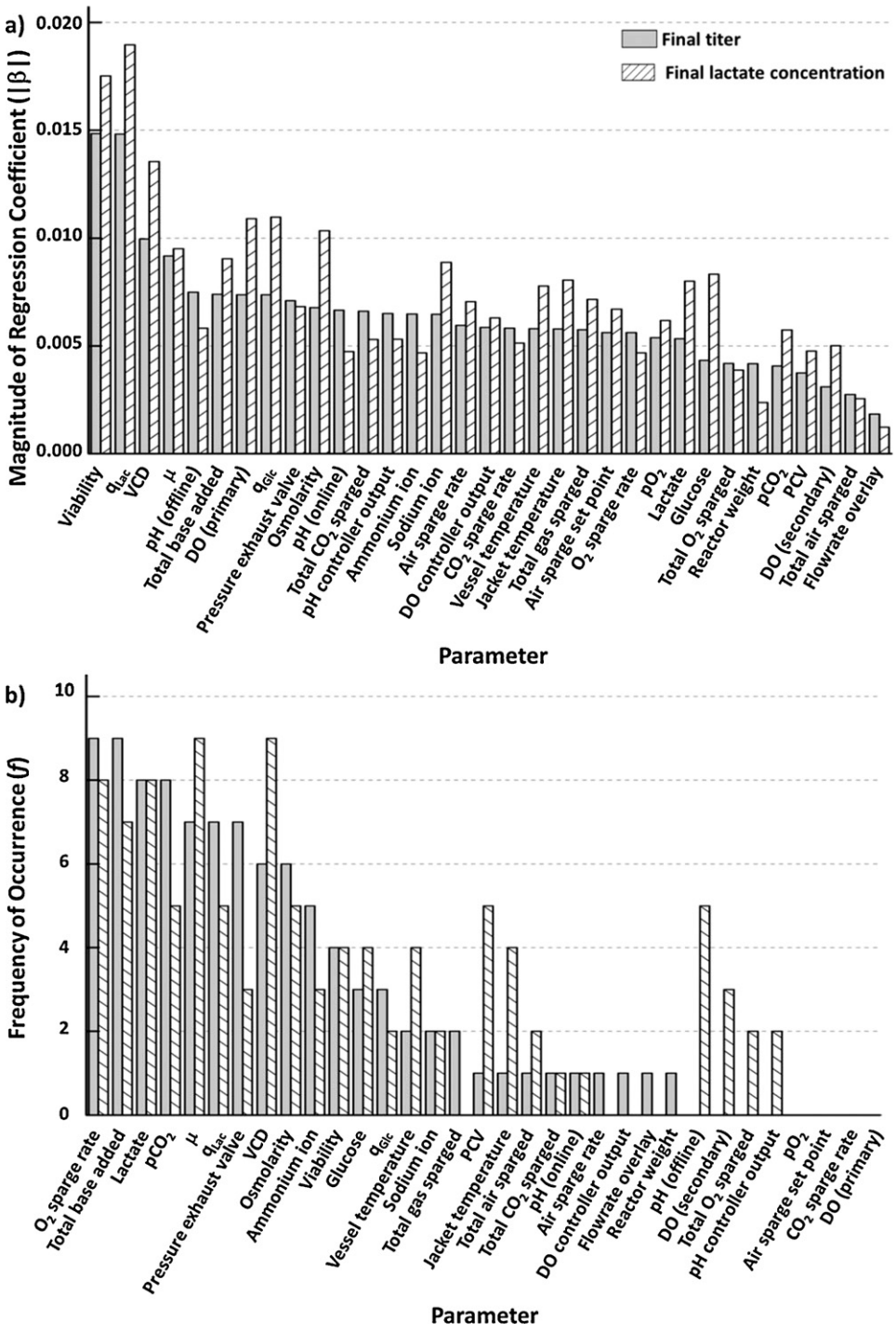

Fig. 5. Contribution of process parameters to prediction accuracy of the final titer ( ) and the final lactate concentration ( ) using data acquired at 80 L scale bioreactors as evaluated using: (a) magnitude of regression coefficient ($|\beta|$) of each parameter in optimized PLSR models. (b) Frequency of occurrence (f) of each parameter in optimized SVR models.

scale. Initially, the SVR models appeared to perform better with the gradual removal of parameters, indicating that most of these parameters are indeed redundant or even noisy. In most cases, the best model was constructed using a set of six to eight parameters as indicated by the valley in the validation error profile. The immediate, sharp rise of error following the removal of parameters in this selected set from the model suggests that they play a pivotal role in model prediction accuracy. Thus the occurrence frequency (f) of each parameter in all selected sets represents its relative contribution to the SVR models' performance.

As shown in Section 3.2, class switch appeared to occur either during the inoculum train or by 70 h of the production scale. We thus focused on identifying pivotal parameters at these two stages to search for possible hints of intervention. Fig. 5 shows the relative importance of 33 process parameters acquired at the smallest scale of the inoculum train (80 L) using PLSR and SVR approaches. Both criteria of β and f led to a common conclusion that the majority of parameters pivotal to prediction of the final titer appeared to be related to cell growth and lactate metabolism by different degrees. These parameters include viable cell density, viability, specific cell

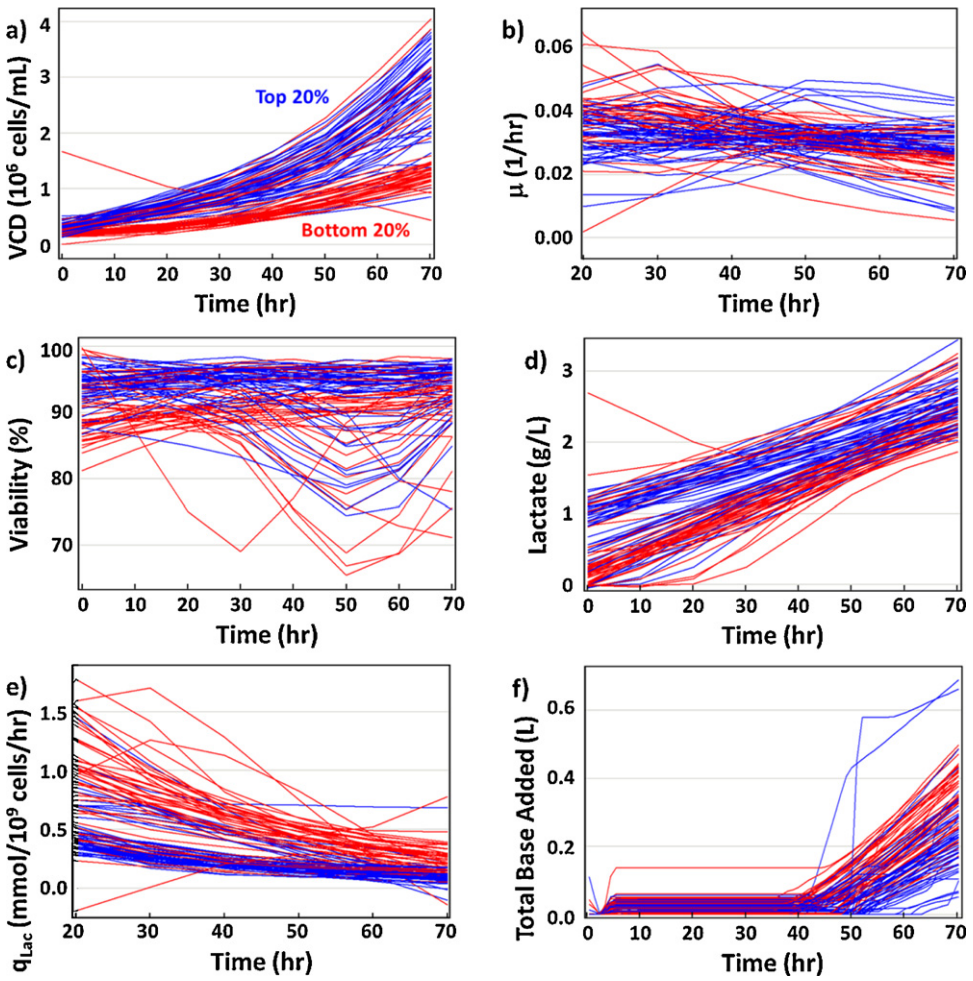

Fig. 6. Time profiles of several pivotal parameters at 80 L scale. Runs in the top 20% are colored in blue; those in the bottom 20% are in red. (a) Viable cell density (VCD). (b) Specific cell growth rate (μ). (c) Viability. (d) Lactate concentration. (e) Specific lactate production rate (q_{Lac}). (f) Total base added. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

growth rate, specific lactate production rate, total base added, lactate, and osmolarity. It is noteworthy that when the final lactate concentration was used as the objective function, similar parameters were identified as pivotal, supporting the notion that product yield and cellular metabolism are indeed strongly correlated.

The time profiles of several pivotal parameters in the top 20% and bottom 20% runs at the 80 L scale are shown in Fig. 6. Although the differences between runs in these two classes were rather modest, a general trend could still be discerned. Runs in both classes appeared to be inoculated at similar cell concentrations, yet cells in most top 20% runs grew at relatively faster rates, giving rise to consistently higher viable cell density in these runs (Fig. 6a and b). Cell viability also largely remained high (>90%) in these cultures (Fig. 6c). Surprisingly, the majority of the top 20% runs experienced somewhat higher lactate concentration at this scale as shown in Fig. 6d. However, the specific lactate production rate was lower (Fig. 6e) and less base was added to maintain a constant pH (Fig. 6f).

It is interesting that these pivotal parameters identified at the beginning of the inoculum train continued to be critical during the early stages of the production scale as evident from Fig. 7. Furthermore, the subtle differences between runs in the top and bottom 20% classes observed at the 80 L scale were significantly magnified at the production scale (Fig. 8). As early as 60 h, a number of high-titer runs already had a metabolic shift to lactate consumption as indicated by negative specific lactate consumption rates, whereas most low-titer runs continued to produce lactate (Fig. 8a). The majority of runs in the top 20% class eventually shifted to the lactate-consuming state. In contrast, runs in the bottom 20% class produced lactate at elevated rates, resulting in substantially high lactate concentrations in most cultures (Fig. 2c).

Specific glucose consumption rates also differed significantly between the two classes (Fig. 8b). High-titer runs consumed glucose at much reduced levels throughout the cultures compared to those with low titer. Thus, high-titer runs did not appear to require multiple additions of glucose after the main feed at 70 h (Fig. 8c).

The low lactate concentration in the top 20% runs reduced or even eliminated the need for base addition whereas large amounts of base were added to the bottom 20% runs (Fig. 8d). This base addition in turn led to accumulation of sodium ion to significantly higher concentrations (Fig. 8e), and therefore osmolarity (data not shown), in these low-titer runs. The difference in lactate concentration between the two classes was also reflected in the pH controller output as shown in Fig. 8f. The opposing behaviors of parameters related to lactate metabolism in the two classes further strengthened the findings that this set of parameters played an important role in predicting the final process outcome.

3.4. Lactate consumption at production scale emerges as process indicator

The analysis presented so far indicates a high correlation of cell growth and lactate metabolism to the final titer. The majority

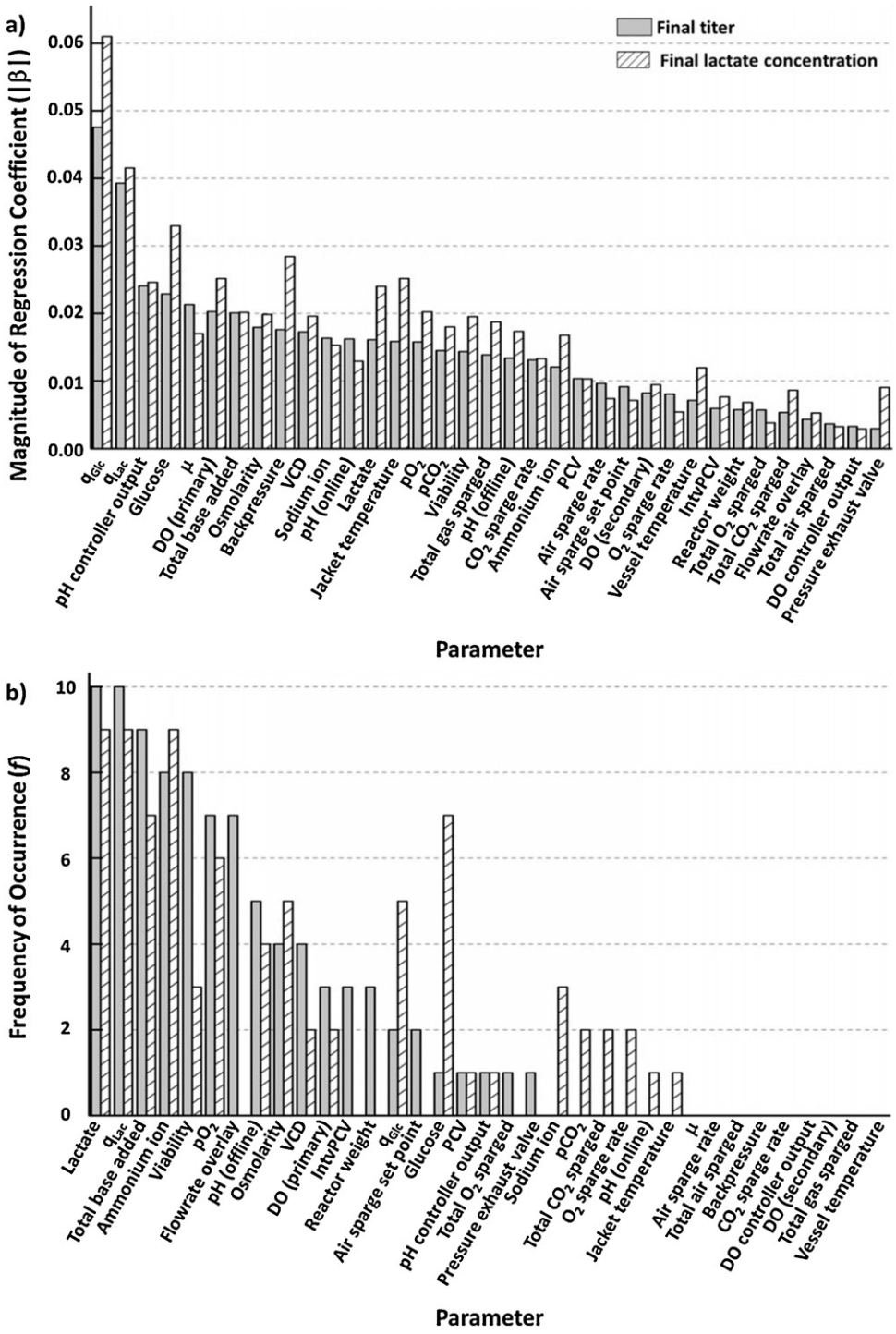

Fig. 7. Contribution of process parameters to prediction accuracy of the final titer ( ) and the final lactate concentration ( ) using data acquired up to 70 h of 12,000 L scale bioreactors as evaluated using: (a) Magnitude of regression coefficient ($|\beta|$) of each parameter in optimized PLSR models. (b) Frequency of occurrence (f) of each parameter in optimized SVR models.

of parameters identified as pivotal for prediction of the final process outcome, using data from the inoculum train or the early stage of the production scale, are related to cell growth and lactate metabolism (Figs. 5 and 7). Runs with high viable cell concentration and low final lactate concentrations or consumed lactate at the production scale yielded high levels of recombinant antibody (Figs. 2c and 8a). Runs with low lactate production rates and high cell growth rates at the beginning of the inoculum train often had high final titer (Fig. 6). Indeed, when specific lactate production rate was plotted against viable cell concentration or specific glucose consumption rate at 80 L scale (Fig. 9), two clusters of the top and the bottom 20% runs could be seen, albeit with a high degree of overlap. These metabolic indicators of the final process outcome thus hint at possible means to intervene with the process as early as the inoculum stage.

To gain more insights into the metabolic shift occurring at the production scale, which is highly correlated to hyper-productivity, the specific rates of lactate production, glucose consumption, and

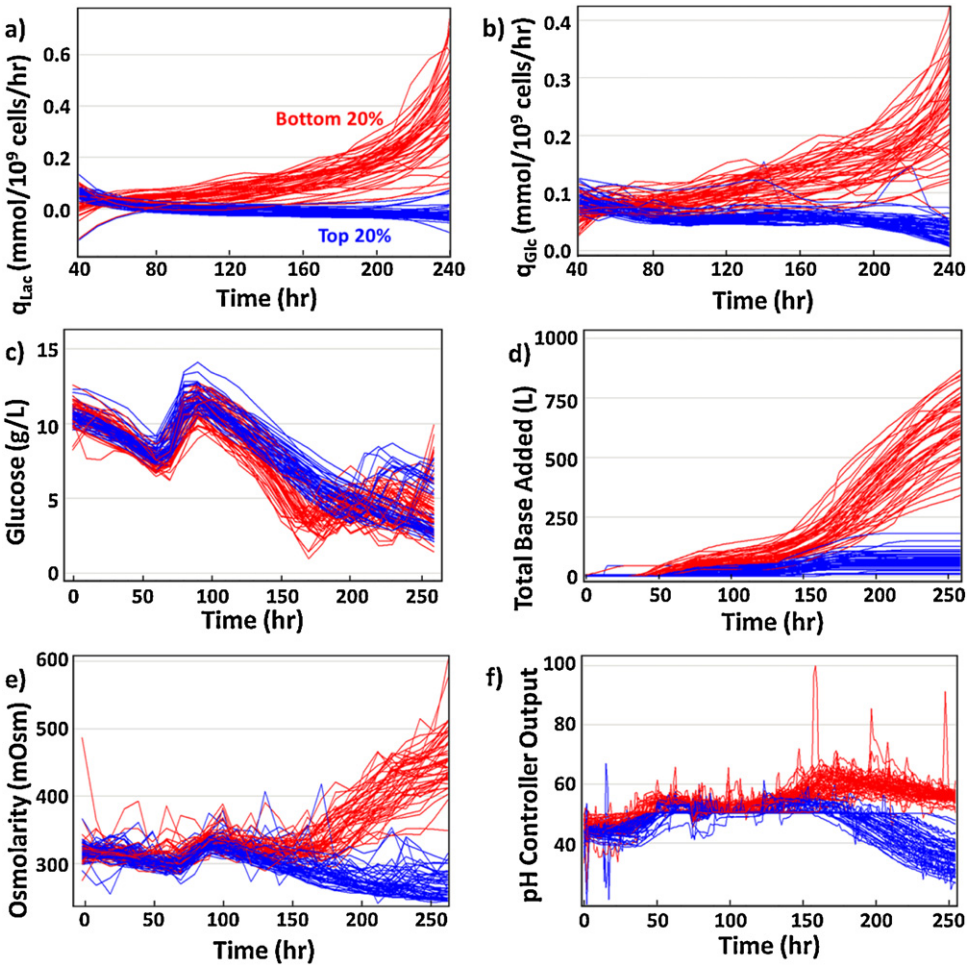

Fig. 8. Time profiles of several pivotal parameters at 12,000 L scale. Runs in the top 20% are colored in blue; those in the bottom 20% are in red. (a) Specific lactate production rate (q_{Lac}). (b) Specific glucose consumption rate (q_{Glc}). (c) Glucose concentration. (d) Total base added. (e) Osmolarity. (f) pH controller output. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cell growth in the top and bottom 20% of runs at the late stage of the production scale (from 120 to 240 h with 10 h intervals) were further analyzed as shown in Fig. 10. In low-titer runs, specific lactate production rate spanned over a wide range from a very low value to as high as 0.6 mmol/10^9 cells/h (Fig. 10a). In contrast, specific lactate production rate in high-titer runs spanned a much narrower range from 0.05 to −0.05 mmol/10^9 cells/h (consumption). Lactate consumption at the production scale, strikingly, occurred even when lactate was almost depleted in the cultures. This suggests that once cells start to consume lactate, they have a propensity to continue consuming it regardless of the low level of this metabolite. Likewise, cells in a lactate-producing culture appeared to remain in that state despite the extensive accumulation of lactate. In other words, high concentration of lactate alone is not sufficient to trigger lactate consumption, nor does it completely inhibit lactate production.

Interestingly, glucose concentration does not dictate lactate metabolism; both lactate production and consumption can occur over the same wide range of glucose concentration (Fig. 10b). In other words, the abundant presence of glucose does not deter

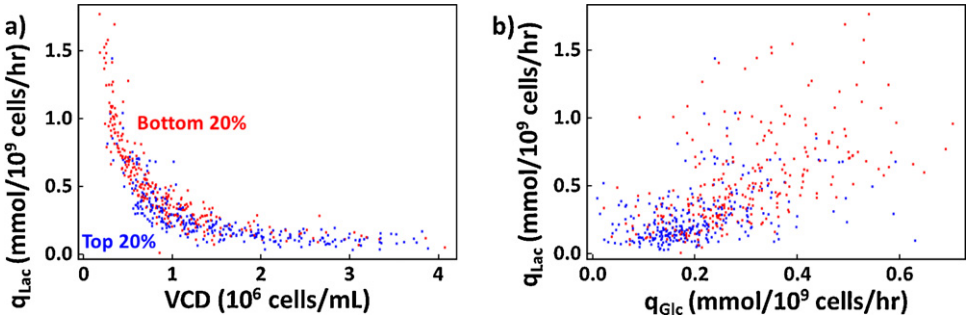

Fig. 9. Relationship between several parameters related to cell growth and lactate metabolism for runs in the top 20% (blue) and the bottom 20% (red) classes at 80 L scale. Each data point represents one time point from 20 h to 70 h of 80 L cultures with 10 h intervals. (a) Specific lactate production rate (q_{Lac}) vs. viable cell density (VCD). (b) q_{Lac} vs. specific glucose consumption rate (q_{Glc}). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

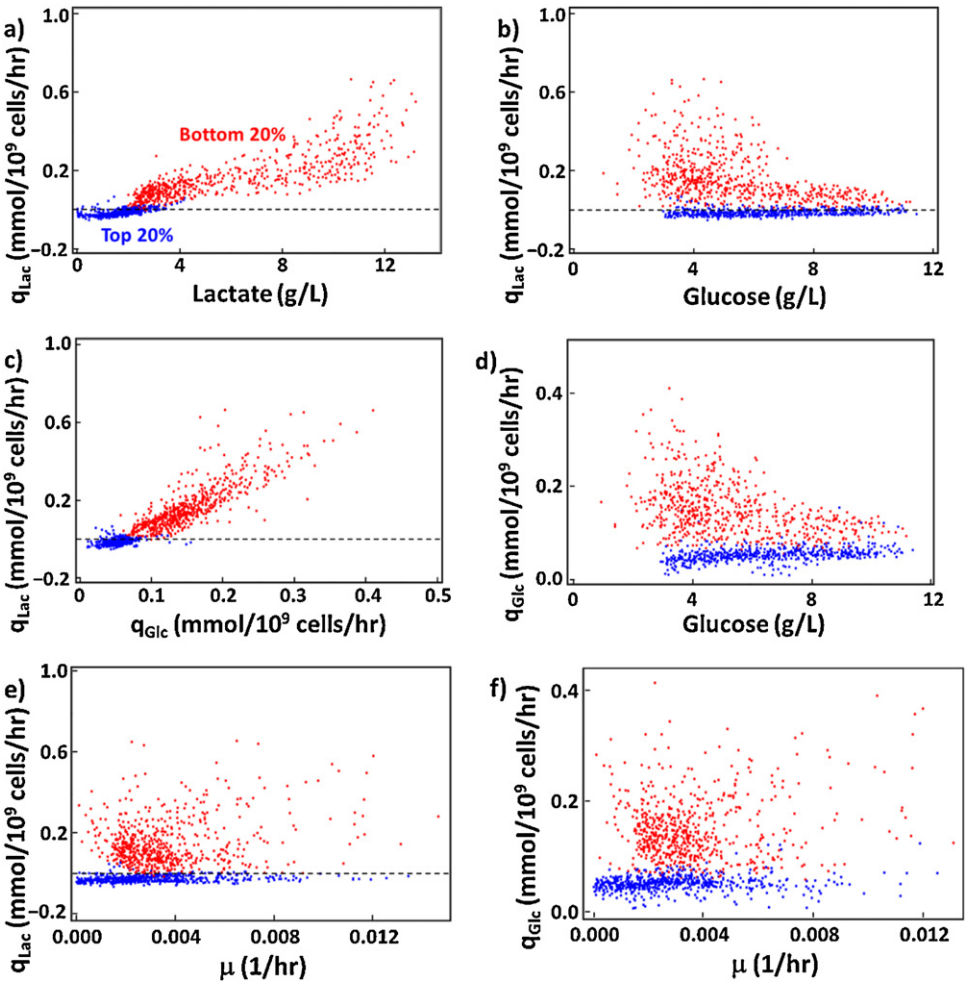

Fig. 10. Relationship among several parameters related to cell growth and lactate metabolism for runs in the top 20% (blue) and the bottom 20% (red) classes in the late stage of the production scale. Each data point represents one time point from 120 h to 240 h of 12,000 L cultures with 10 h intervals. The dashed line represents $q_{Lac} = 0$. (a) Specific lactate production rate (q_{Lac}) vs. lactate concentration. (b) q_{Lac} vs. glucose concentration. (c) q_{Lac} vs. specific glucose consumption rate (q_{Glc}). (d) q_{Glc} vs. glucose concentration. (e) q_{Lac} vs. specific cell growth rate (μ). (f) q_{Glc} vs. μ. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lactate consumption. It is evident that lactate consumption occurs only when the specific glucose consumption rate is low (below 0.07–0.1 mmol/10^9 cells/h) (Fig. 10c). There also seems to be a minimum specific glucose consumption rate that cells sustain, as the value never reaches zero. Furthermore, glucose concentration alone does not determine glucose consumption, as can be seen in Fig. 10d. There is virtually no difference in the range of glucose concentration between metabolically shifted cultures ($q_{Lac} \leq 0$ and low q_{Glc}) and "typical" cultures (high q_{Lac} and high q_{Glc}). It should be noted that the glucose concentrations in all cultures were maintained at more than 3 g/L, substantially higher than the reported K_m of the GLUT1 transporter for glucose (approximately 0.18 g/L).

It is interesting to observe that lactate consumption is not strongly dependent on how much cell growth slowed down during the late stage of the production scale (Fig. 10e). The specific growth rate spans over a wide and similar range for both lactate-consuming and lactate-producing cultures, although somewhat more frequent occurrence of slower growth rates is seen in lactate-consuming cultures. Likewise, the glucose consumption rate can vary greatly regardless of specific cell growth rate (Fig. 10f). Taken together, these observations indicate that the potential of cells to consume lactate in the late stage is largely a function of reduced glycolytic flux rather than of glucose or lactate concentration or cell growth.

4. Discussion

The immense volume of cell culture bioprocess data in historical archives certainly holds valuable insights into manufacturing processes and product characteristics. This resource has begun to be explored to generate process insights using multivariate data analysis tools. This study employed two such tools, SVR and PLSR, to investigate process data from more than two hundred production-scale cultures. Both methods could predict process performance with similar high accuracies if data from the production bioreactors were used with the objective function being either the final titer or the final lactate concentration.

Data acquired at the inoculum train alone (80, 400 and 2000 L reactors) were somewhat less predictive of the final process outcome compared to the production-scale data. The difference in prediction accuracy between these two sets of data is largely due to the difference in their culturing mode. The duration in each seed train reactor is shorter (3–4 d) and the cell concentration achieved is lower when compared to the production culture. The values of process parameters which can be used for prediction are lower, and so are the differences in parameter values between high- and low-performing runs. Thus, data acquired from the seed train are not as accurate in predicting the final titer as the production data. Although the seed run data do not provide a highly accurate

Table 3
Number of runs predicted to be in one of the three classes using data at 80 L scale: top 20%, middle 60%, and bottom 20%. The actual class each of these runs belonged to at the end of the 12,000 L scale was also shown.

Predicted at 80 L		Actual titer at production scale		
	Total	Top 20%	Middle 60%	Bottom 20%
Top 20%	43	13	28	2
Middle 60%	148	33	95	20
Bottom 20%	52	3	22	27

prediction of the final titer, they do provide valuable information on the "class", i.e., high productivity and low productivity runs. This is illustrated in Table 3. Among the 52 runs predicted to be in the bottom 20% class using data obtained from the 80 L cultures, 52% actually turned out to be in the bottom 20% class regarding the final titer. Another 42% became middle titer runs. It is worth noting that any intervention would be targeted towards those low-titer runs, for which there is indeed a good class prediction using the seed culture data. Even if only those runs can be rectified by employing some remedial procedure, the overall increase in productivity is substantial. The key issue will be which intervention procedure can be applied. Although data mining can only yield correlations and rarely reveals causal relationships, one can gain insights from the key factors that contributed to the prediction.

The pivotal parameters identified both at the inoculum train and at the production scale are mostly associated with cell growth and lactate metabolism, indicating the prominent role of cellular metabolism in determining product titer. Previous analysis of a subset of runs used in this study (Charaniya et al., 2010) and data from another manufacturing process (Kirdar et al., 2008) also led to a similar conclusion. The results from this study further indicate that lactate consumption at the production scale serves as an indicator of high productivity. However, the conditions that induce lactate consumption in the high titer runs at this scale are still unknown.

The observed implication that the inoculum train possibly imparts a longer lasting effect on the process outcome reiterates our previous findings (Charaniya et al., 2010) and the results from another study (Ündey et al., 2010). It may hint at possible intervention during inoculum train operation to steer the low titer runs to higher productivity. A possible approach of intervention is the selective use of 80 L runs for subsequent inoculation into 400 L runs, although this will certain impose major constrains in reactor scheduling and increased cost of operation. A more acceptable approach might be identifying the cause and taking remedial actions. The prominence of glucose consumption and lactate production as important factors at the 80 L scale points to the possibility of reducing lactate production by metabolic intervention or by other means of lactate removal during the inoculum train.

A key correlated factor of low lactate production and lactate consumption at the production scale appears to be low specific glucose consumption. Thus, controlling glycolytic flux seems to be the key to modulating lactate metabolism and therefore the final product yield. Such a conclusion has also been reached through a metabolic study in conjunction with modeling (Mulukutla et al., 2012), which showed that the switch to the lactate consumption mode could be attributable to a moderate attenuation of glycolytic genes' expression and differential activities of the Akt and p53 signaling pathways. Indeed, inhibition of the Akt pathway by addition of its inhibitors in the late growth stage was shown to facilitate lactate consumption.

Remedial corrective measures at the production scale will need to focus on manipulating cell metabolism prior to 70 h, the point at which the correlation between predicted and actual titer still hints at some flexibility in the outcome. Many possible approaches of suppressing glucose metabolism and eliciting metabolic shift to lower lactate production or lactate consumption have been reported, including reducing glucose concentration (Cruz et al., 1999; Zhou et al., 1997), employing alternative sugars (Altamirano et al., 2006; Wlaschin and Hu, 2007), supplementing copper ion (Qian et al., 2011), and adding inhibitors of the Akt pathway (Mulukutla et al., 2012). Conceivably interventive measures can be taken by then if necessary. Whether those possible interventions will be effective can only be answered by experimentation. Whether the intervention methods, if proven effective, should be implemented in a manufacturing setting will largely depend on the operating protocols of each individual plant and the nature of the interventions.

With the increasing emphasis on the concept of Quality by Design (QbD) in the production of therapeutic biologics, we foresee such practices of mining bio-manufacturing data being extended to analyze Critical Quality Attributes (CQAs). Recently, clustering of glycosylation profiles of an antibody product has revealed a high correlation between product quality attributes and process characteristics (Le et al., 2011). As both process and product quality data continue to accumulate, the likelihood of identifying process characteristics which affect product quality will also increase. Harnessing the power of data mining will greatly strengthen our capability to produce high quality products through high-productivity processes.

Acknowledgments

The authors would like to thank the Minnesota Supercomputing Institute (MSI) for computational support. H.L. was supported in part by the Vietnam Education Foundation (VEF). The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of VEF. The authors declare no conflict of interest.

References

Altamirano, C., Illanes, A., Becerra, S., Cairó, J.J., Gòdia, F., 2006. Considerations on the lactate consumption by CHO cells in the presence of galactose. Journal of Biotechnology 125 (4), 547–556.
Bachinger, T., Riese, U., Eriksson, R., Mandenius, C.-F., 2000. Monitoring cellular state transitions in a production-scale CHO-cell process using an electronic nose. Journal of Biotechnology 76 (1), 61–71.
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. Biotechnology Progress 18 (6), 1366–1376.
Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A Library for Support Vector Machines.
Charaniya, S., Hu, W., Karypis, G., 2008. Mining bioprocess data: opportunities and challenges. Trends in Biotechnology 26 (12), 690–699.
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. Journal of Biotechnology 147 (3–4), 186–197.
Chen, K., Liu, Q., Xie, L., Sharp, P.A., Wang, D.I.C., 2001. Engineering of a mammalian cell line for reduction of lactate formation and high monoclonal antibody production. Biotechnology and Bioengineering 72 (1), 55–61.
Chong, I.-G, Jun, C.-H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78 (1–2), 103–112.
Coleman, M.C., Block, D.E., 2006. Retrospective optimization of time-dependent fermentation control strategies using time-independent historical data. Biotechnology and Bioengineering 95 (3), 412–423.
Cruz, H.J., Moreira, J.L., Carrondo, M.J.T., 1999. Metabolic shifts by nutrient manipulation in continuous cultures of BHK cells. Biotechnology and Bioengineering 66 (2), 104–113.
de Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18 (3), 251–263.
Glacken, M.W., Fleischaker, R.J., Sinskey, A.J., 1986. Reduction of waste product excretion via nutrient control: possible strategies for maximizing product and cell yields on serum in cultures of mammalian cells. Biotechnology and Bioengineering 28 (9), 1376–1389.
Gunther, J.C., Conner, J.S., Seborg, D.E., 2007. Fault detection and diagnosis in an industrial fed-batch cell culture process. Biotechnology Progress 23 (4), 851–857.
Hu, W., Dodge, T., Frame, K., Himes, V., 1987. Effect of glucose on the cultivation of mammalian cells. Developments in Biological Standardization 66, 279–290.

Irani, N., Wirth, M., van den Heuvel, J., Wagner, R., 1999. Improvement of the primary metabolism of cell cultures by introducing a new cytoplasmic pyruvate carboxylase reaction. Biotechnology and Bioengineering 66 (4), 238–246.

Kim, S., Lee, G., 2007a. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. Applied Microbiology and Biotechnology 74 (1), 152–159.

Kim, S., Lee, G., 2007b. Functional expression of human pyruvate carboxylase for reduced lactic acid formation of Chinese hamster ovary cells (DG44). Applied Microbiology and Biotechnology 76 (3), 659–665.

Kirdar, A.O., Green, K.D., Rathore, A.S., 2008. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. Biotechnology Progress 24 (3), 720–726.

Le, H., Castro-Melchor, M., Hakemeyer, C., Jung, C., Szperalski, B., Karypis, G., Hu, W.-S., 2011. Mining bioprocess data for discovery of key parameters influencing high productivity and quality. In: Proceedings of the 22nd Annual Meeting of the European Society for Animal Cell Technology (ESACT), Vienna, Austria, May 15–18. Springer.

Liu, H., Hiroshi, M., 1998. Feature Selection for Knowledge Discovery and Data Mining. Springer, Norwell, 244 pp.

Mulukutla, B.C., Gramer, M., Hu, W.-S., 2012. On metabolic shift to lactate consumption in fed-batch culture of mammalian cells. Metabolic Engineering 14 (2), 138–149.

Qian, Y., Khattak, S.F., Xing, Z., He, A., Kayne, P.S., Qian, N.-X., Pan, S.-H., Li, Z.J., 2011. Cell culture and gene transcription effects of copper sulfate on Chinese hamster ovary cells. Biotechnology Progress 27 (4), 1190–1194.

Ündey, C., 2004. Intelligent real-time performance monitoring and quality prediction for batch/fed-batch cultivations. Journal of Biotechnology 108 (1), 61–77.

Ündey, C., Ertunç, S., Mistretta, T., Looze, B., 2010. Applied advanced process analytics in biopharmaceutical manufacturing: challenges and prospects in real-time monitoring and control. Journal of Process Control 20 (9), 1009–1018.

Wlaschin, K., Hu, W., 2007. Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. Journal of Biotechnology 131 (2), 168–176.

Zhou, W., Chen, C.-C., Buckland, B., Aunins, J., 1997. Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. Biotechnology and Bioengineering 55 (5), 783–792.