Received: 14 September 2020 | Revised: 28 January 2021 | Accepted: 6 March 2021

DOI: 10.1002/aic.17257

RESEARCH ARTICLE

AIChE JOURNAL

Multivariate statistical data analysis of cell-free protein synthesis toward monitoring and control

Carlos A. Duran-Villalobos[1] | Olotu Ogonah[2] | Beatrice Melinek[2] | Daniel G. Bracewell[2] | Trevor Hallam[3] | Barry Lennox[1]

[1]Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, UK

[2]Department of Biochemical Engineering, University College London, London, UK

[3]Sutro Biopharma, Inc., South San Francisco, California, USA

Correspondence
Carlos A. Duran-Villalobos, Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK.
Email: carlos.duran@manchester.ac.uk

Funding information
UK Engineering and Physical Sciences Research Council (EPSRC), Grant/Award Number: EP/P006485/1

Abstract

The optimization and control of cell free protein synthesis (CFPS) presents an ongoing challenge due to the complex synergies and nonlinearities that cannot be fully explained in first principle models. This article explores the use of multivariate statistical tools for analyzing data sets collected from the CFPS of Cereulide monoclonal antibodies. During the collection of these data sets, several of the process parameters were modified to investigate their effect on the end-point product (yield). Through the application of principal component analysis and partial least squares (PLS), important correlations in the process could be identified. For example, yield had a positive correlation with pH and NH_3 and a negative correlation with CO_2 and dissolved oxygen. It was also found that PLS was able to provide a long-term prediction of product yield. The presented work illustrates that multivariate statistical techniques provide important insights that can help support the operation and control of CFPS processes.

KEYWORDS
biopharmaceutical manufacturing, cell-free synthesis, process control, process data analytics, process optimization

1 | INTRODUCTION

The integration of machine learning techniques into the operation and control of cell free protein synthesis (CFPS) systems[1] offers significant potential for improving productivity and the quality of materials manufactured using this relatively new processing technique. CFPS offers advantages over in-vivo protein production for applications that require more precise control of product physiochemical properties, such as bispecific antibodies, antibody drug conjugates, vaccines, and membrane proteins.[2] It also offers direct access and control of the synthesis environment, giving potential for the development of highly productive CFPS platforms, for the rapid and efficient production of recombinant proteins.[3]

Guidelines proposed by pharmaceutical regulatory authorities, such as those published by the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), emphasize the importance of identification, control, design space, and process validation in the development and manufacture of pharmaceutical products.[4] In particular, Q8, Q9, Q10, and Q11 from the ICH highlight the importance of process modeling as a tool to implement quality by design.

Models used to describe CFPS production processes and in particular its application to quality by design can only be useful in practice if sufficient process knowledge is available to explain the effect of critical process parameters on critical quality attributes. In this respect, mechanistic models offer great value in determining causality to support the optimization of CFPS processes[5,6]; however, the development of such models requires significant time and resources which typically make them impractical.[7] In addition, CFPS relies on a complex network of interacting reactions, reactants, and enzymatic catalysts, which are not yet fully understood. Although there are

continuous efforts to improve CFPS mechanistic models such as less data intensive models based on flux balance analysis,[8] models developed from first principles are still scarcely being used to optimize the processes[9] and hence empirical modeling techniques have been investigated as a possible alternative. Unfortunately, challenges associated with the multidimensionality of the data being analyzed, as well as variations introduced by disturbance factors such as experimental error and noise can lead to problems for many empirical modeling techniques. However, multivariate data analysis techniques have been applied to many other processes where studies have demonstrated them to be capable of overcoming these challenges, allowing models to be developed that can be used to support process optimization.[10]

One approach to modeling multidimensional data is to use artificial neural networks (ANNs). Multiple publications on bioprocess optimization, prediction, delivery and prognoses have shown the capabilities of ANNs, particularly their ability to predict behavior in CFPS processes.[11] The main advantage offered with ANNs is that they are better at capturing nonlinear relationships when compared with traditional statistical modeling techniques, such as partial least squares (PLS). However, they require greater computational resources and have a complex structure requiring model developers to explicitly identify possible causal relationships between process variables. In addition, model developers need to go through an empirical process of performing sensitivity analyses on parameters such as learning rates, momentum terms, and model structure.[12] As a result, ANNs models tend to be opaque and difficult to thoroughly validate and verify.

An alternative approach to modeling multidimensional data is to use multivariate statistical models, such as those commonly referred to as multivariate statistical data analysis (MSDA). MSDA techniques have been applied successfully to pharmaceutical production processes for optimization, monitoring, online control and detection of sensor faults in seed, batch, and fed-batch cultivations.[10,13-21] For example, the capabilities of MSDA in a cell-culture process for small scale (2 L) and large scale (2000 L) batches were published in Kirdar et al.[22] This work aimed to evaluate whether MSDA was able to characterize the process through the analysis of process parameters such as CO_2, O_2, glucose, pH, lactate, ammonium ions, purity, viable cell density, viability, and osmolality. The proposed methodology included analyzing various control charts to identify fault conditions, with their results demonstrating that MSDA could be used as a tool for extracting knowledge from such processes.

There has been very little work published that has focused on the analysis of cell free expression systems using MSDA techniques. In Reference 23, the authors assessed the differences in the metabolite profiles of four lysates by analyzing the data collected from the process using principal component analysis (PCA), with the aim being to standardize lysate activity and to design an improved cell free expression system. The lack of published research in this area indicates a clear knowledge gap regarding the potential benefits of using MSDA tools to optimize and control such processes.

The aim of the present work was, therefore, to apply MSDA techniques to experimental data collected from CFPS processes to determine which process parameters, including pH, temperature, and O_2 have the most significant effect on end-point qualities, such as yield and aggregation. In addition, the work also aimed to identify the suitability of using MSDA techniques to provide long-term predictions of end-point quality metrics during operation of the process. The specific case study for the analyses presented in this article is the scalable cell-free synthesis of monoclonal antibodies (mAb), using the cell-free lysate system developed by Sutro Biopharma.[24,25] Additionally, the present work aims to provide a generic methodology that allows rapid process characterization and optimization to the manufacturing of other therapeutic proteins and biopolymers by developing data-based relationships between process parameters and process outputs such as yield, which are likely to be highly dependent on the cell extract type, target protein, construct design, and so forth.[26]

Section 2 describes the methodology that was followed when applying the MSDA techniques, with Section 3 providing the methodology for operating the laboratory equipment and collecting data. Section 4 presents the results and provides a discussion and finally, Section 5 provides the conclusions and the authors' perspective on the use of MSDA techniques as tools for optimizing the performance of CFPS processes. In addition to the main body of the article, Appendix A provides a list of abbreviations and Appendix B describes the statistical tools used in this article. The codes have been shared on GitHub (https://github.com/CarlosADuranVIllalobos/CFPS-Multivariate-Statistical-Data-Analysis) to enable use of the same technique to other CFPS reactions.

2 | METHODOLOGY OF THE DATA ANALYSIS

2.1 | Data set organization

Five sets of experimental data were analyzed in this work and a summary of these date are provided in Table 1. Four of these data sets were collected from experiments completed at UCL, D1–D4, with the fifth, D5, collected from a laboratory operated by Sutro Biopharma, Inc. All experiments were undertaken by the same person, Experimenter A, with the exception of D4 which was completed by Experimenter B. Each data set consisted of the mean process measurements collected every minute from observations from each well on a 24-well plate. The data sets were averaged to hourly data to make them robust to instrumental noise and to enable simpler analysis. In some wells, the reaction failed and for these wells, the data were discarded. The observations column provides the number of wells that were used in each data set after failures were discarded. The values for each of the control parameters: temperature, run length and pH, for the five data sets (D1–D5) are shown in Table 1.

D1–D4 contained online measurements of pH, dissolved oxygen (DO) and temperature and end-point yields. D5 contained measurements of pH, DO, temperature, volumetric O_2, volumetric CO_2 and volumetric NH_3, and end-point yields of the quality variables: mAb, monomer percentage and aggregate percentage.

Set-points for the temperature and pH controllers were held constant for D1, D2, and D5. However, after analyzing the data from these experiments and observing findings in relevant literature,[27]- pH

TABLE 1 Control parameters of experimental data sets

Data set	Location	Experimenter	Temp. (°C)	Run length (h)	pH	Observations
D1	UCL	A	27, 28.5, and 30	4, 6, and 8	6.7, 6.9, 7, and 7.3	22
D2	UCL	A	29, 30, and 32	2, 4, and 6	6.7, 7, and 7.3	24
D3	UCL	A	30	8	Variable (6.6–7.8)	24
D4	UCL	B	30	12	Variable (6.4–7.5)	23
D5	Sutro	A	28	10, 12, and 14	6.4, 6.7, and 7	19

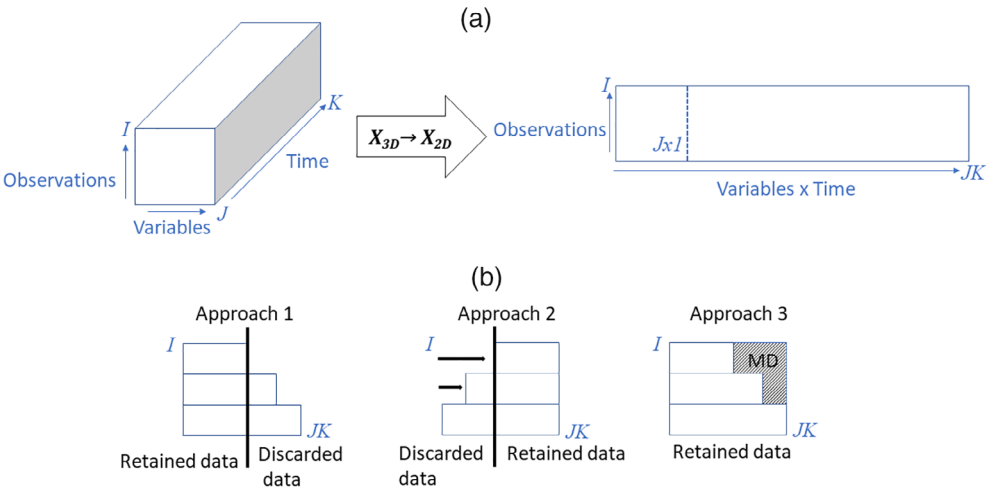

FIGURE 1 Preprocessing for the matrices of process measurements **X**. (A) Unfolding into a two-dimensional array. (B) Approaches used to address uneven vectors [Color figure can be viewed at wileyonlinelibrary.com]

was seen to be a key factor impacting the reaction yield and to analyze its effect further the set-point was varied through experiments D3 by introducing low-frequency pseudo random binary signals (PRBS). PRBS are often used in system identification to excite the process at different times and frequencies of the input variables trajectories such that it produces small changes in the output. Finally, the set-points for pH during D4 were specified so that the effect that a predefined pH trajectory had on the reaction could be compared to the case were pH was maintained at a fixed level in reactions operating over a longer duration (12 h).

2.2 | Data preprocessing

Data were arranged into matrices of process measurements, ***X***, and matrices of end-point qualities, ***Y***. The matrices were then normalized prior to model identification by subtracting the mean and dividing by the standard deviation of each variable. For data analysis and predictions that considered temporal information, the matrix ***X***, which contained three-dimensional information (variables of size *J*, time intervals of size *K*, and observations or repeats of size *I*) was unfolded into a two-dimensional array using the technique referred to as multiway unfolding[29] as shown in Figure 1(A). This technique has been used in several previous studies for monitoring batch processes.[30] In this article, the letter "M" will be used as a prefix of the model to indicate multiway models (e.g., MPLS will refer to multiway PLS).

In some cases, the total reaction length, and therefore the number of observations recorded during the run, or batch, varied, and as a result, the unfolded data matrix ***X*** was incomplete because of the difference in vector sizes. Uneven vector lengths creates difficulties when using multiway unfolding as it requires all vectors to be of equal length. To address this issue three approaches were used: Approach 1 compiled the unfolded X matrix using only those measurements that were recorded up to the shortest reaction length. Approach 2 compiled the unfolded X matrix using only the measurements that were recorded in the last few hours of each reaction, which all the observations had in common and Approach 3 compiled the unfolded X matrix using missing data (MD) techniques. Specifically, the technique of Projection to the Modal Plane[31] was used to estimate the progress of each run had the reaction been allowed to continue. Figure 1(B) shows the three approaches used to address uneven vectors.

2.3 | Methodology of the exploratory data analysis

The initial data analysis was provided using the relatively standard techniques of PCA and PLS and their multiway counterparts (MPCA and MPLS). Further details of these techniques can be found in Appendix B.

PCA was initially applied to the five data sets to help in the identification of critical process parameters that could potentially be used in the future to control the cell-free synthesis reaction. In addition, PCA was used to identify similarities and differences between the

observations collected in the experiments conducted by different laboratory experimenters and in different locations. For the PCA models, the matrix $\boldsymbol{X}$ included the initial and final measurements from the process variables and the end-point, quality variable(s). PCA models were then identified for this matrix using the NIPALS algorithm.[29]

PLS and MPLS were also used for exploratory data analysis to observe the effect that each process variable had on the quality variable(s). The data used to construct the $\boldsymbol{X}$ matrix for PLS and MPLS were the process variable measurements through each batch. For the PLS model, the quality variable(s) was used to construct the $\boldsymbol{Y}$ matrix. The PLS and MPLS models were identified using the SIMPLS algorithm[32] and the number of latent variables used in the models was chosen as being that which minimized the root mean square error of cross validation when using 10-fold cross validation.

2.4 | Methodology of the prediction assessment

A variety of process control techniques based on MSDA have been developed and a key factor when demonstrating the ability of these control techniques to increase the efficiency and the robustness of CFPS processes is to determine the accuracy that MSDA models can predict key quality variables. The first part of this analysis provides a comparison of the ability of different MSDA modeling techniques to predict end-point quality variable(s). The models that were applied in this work are listed below, with further details regarding each technique provided in Appendix B:

- Ordinary least squares (OLS) and multiway OLS (MOLS) regression.
- PLS and MPLS regressions.
- Quadratic PLS (QPLS) and multiway QPLS (MQPLS) regression.

2.5 | Methodology of monitoring using MSDA control charts

The most frequently used control charts for monitoring purposes are based on two statistics: the Hotelling's statistic, T^2, which provides a measure of the deviation of an observation from the region covered by the identification data set and the squared prediction error (SPE) of $\boldsymbol{x}$, SPE_x, which is the error between the data vector of an observation, $\boldsymbol{x}$, and its reconstructed value obtained using the MSDA model. The confidence limits for the charts were calculated using the bootstrap-resampling technique used in Duran-Villalobos et al,[16] which infers confidence intervals from an empirical distribution function.

3 | METHODOLOGY OF CFPS

A simple mAb system was chosen to demonstrate proof of principle for the use of control and optimization in cell-free synthesis. The aim of this proof of concept study was to determine whether therapeutic dose levels of a mAb could be achieved within 24 h by modifying the operating conditions of the process. The focus of the optimization was on the controllable parameters of the upstream production, i.e. the reactions producing protein from a pDNA plasmid provided by Sutro Biopharma, Inc., the exact sequences of the pDNA plasmid are confidential.

Raw materials for the CFPS reaction were kindly provided by Sutro Biopharma, Inc. and this included a cell-free extract (XtractCF), a reaction mix containing amino acids, nucleic acids, salts and energy source (2x Supermix), plasmids for the heavy chain and light chain of a mAb and T7 RNA polymerase (prepared in-house and in the form of an *Escherichia coli* lysate). None of these components are commercially available. All materials were stored frozen at −80°C and sufficient quantities for each experiment thawed at room temperature immediately prior to use.

The laboratory procedures that were followed for the CFPS reaction were as follows[24,25]:

- 30 ml of thawed XtractCF was pretreated with 22.5 μl of 100 mM iodoacetamide (IAM) (A3221-10VL, Sigma Aldrich) for 30 min at room temperature (with the aim to stabilize redox potential to facility disulfide bond formation).
- The IAM-treated extract was added to 50 ml of 2x supermix, 0.5 ml of T7 RNA polymerase produced in-house by Sutro Biopharma, Inc. to a standard unit activity level, 0.375 and 0.125 ml of the heavy and light chain plasmids, respectively (to a final concentration 0.005 mg/ml), and the volume made up to 100 ml with milliq water.
- The mixture of reactants and plasmids was mixed by inversion, and 3.5 ml added to each well of a 24-well PERC plate (MRT-PRC-2I, Pall). The plate was sealed with a breathable AeraSeal membrane (Excel Scientific) and placed in a Micro-24 micro-bioreactor (MicroReactor Technologies, Pall, Port Washington), where the temperature was controlled by heat plates, the pH was controlled by the addition of CO_2 or NH_3 gases and DO controlled by addition of an oxygen-air mix. The controllers on the Micro-24 were tuned using the Ziegler-Nichols open-loop method, followed by trial and error, fine-tuning. The agitation rate in the Micro-24 was set at 600 rpm to ensure the reactants remained well mixed.
- Reaction yields of cereulide mAb were calculated from the UV absorption at 280 nm of the eluate from a single step protein A purification and an extinction coefficient of 1.474 mg/ml/AU. Then, 400 μl of the final reaction mix was loaded on to a 200 μl MabSelect SuRe robocolumn (28986107, GE Healthcare Lifesciences), pre-equilibrated with 800 μl of 25 mM Tris, 100 mM NaCl at pH 7.4, using a tecan liquid handling unit. The robocolumn was washed with 1000 μl of 25 mM Tris, 100 mM NaCl at pH 7.4, eluted with 600 μl of 50 mM acetate at pH 3.7 and stripped with 600 μl of 100 mM phosphoric acid. With yields of the order of 0.5 mg/ml the resin loading was 1.5 mg/L resin, well within the quoted binding capacity of the resin (35 mg hIgG/ml resin). The flow rate at all stages was set to give a residence time of approximately 5 min. The UV 260, UV 280, and UV 900 of the wash, eluate, and strip were measured with a tecan microplate reader (Infinite M200PRO) and recorded in EVOware.

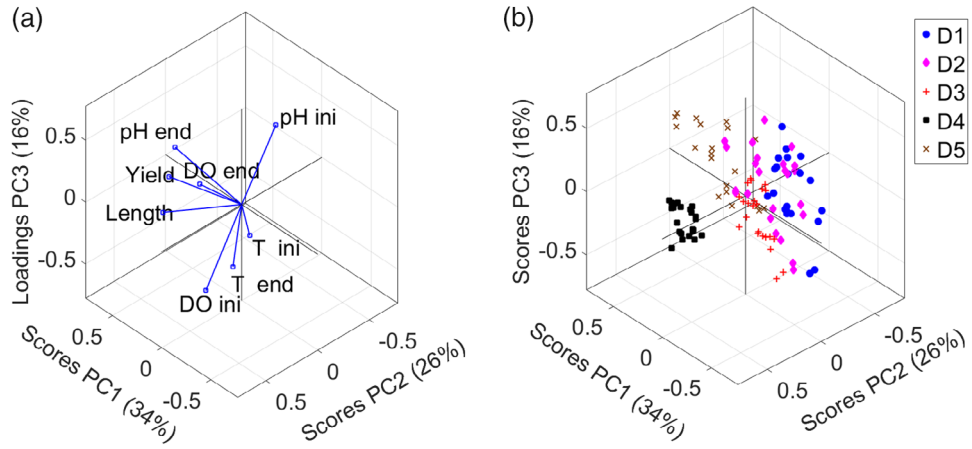

FIGURE 2 Principal component analysis (PCA) of D1–D5 over the first three PCs with axis showing in brackets the variance explained by each PC. (A) PCA loadings plot. (B) PCA scores plot [Color figure can be viewed at wileyonlinelibrary.com]

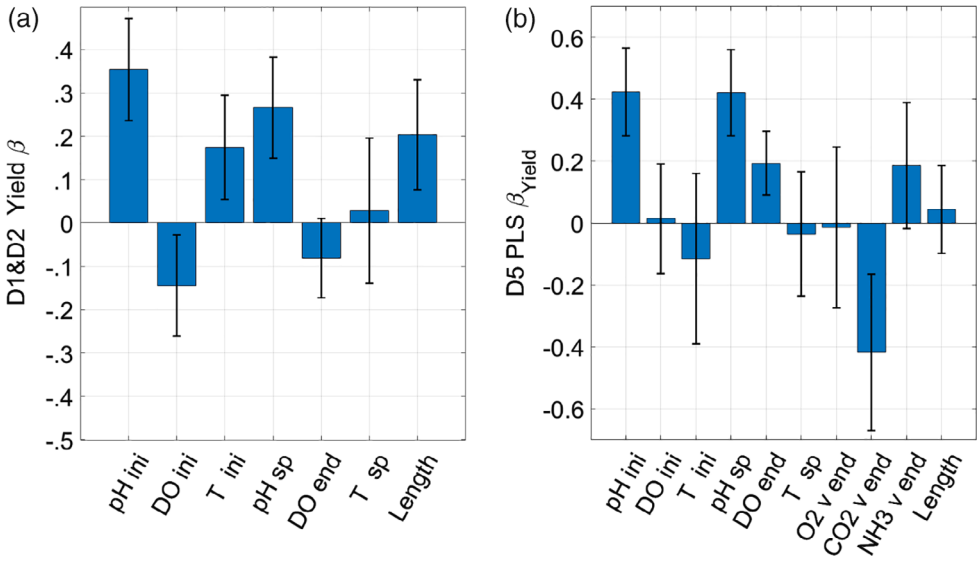

FIGURE 3 Comparison of partial least squares (PLS) regression coefficients between shorter and longer reactions, with error bars indicating twice their standard uncertainty. (A) PLS regression coefficients for yield estimation of a model identified using D1 and D2. (B) PLS regression coefficients for yield estimation of a model identified using D5 [Color figure can be viewed at wileyonlinelibrary.com]

- The columns were reused so additional steps were conducted to restore the columns to their initial condition: a 200 μl flush with 25 mM Tris, 100 mM NaCl at pH 7.4 over 2 min, a clean in place with 800 μl of 0.5 M NaOH (with a slower flow rate to give a 15-min residence time) and a re-equilibration with 2000 μl of 25 mM Tris, 100 mM NaCl at pH 7.4 over 10 min.

Work conducted at Sutro Biopharma, Inc. (South San Francisco, CA) used comparable equipment, techniques and materials. Work at Sutro Biopharma, Inc. also included measurement of % aggregate and % monomer using a high-performance liquid chromatography and capillary zone electrophoresis system.

4 | RESULTS AND DISCUSSION

4.1 | Exploratory data analysis

Figure 2(A) shows a PCA Loadings plot of the five data sets (D1–D5). This figure shows the first three principal components (PCs) which primarily captured the information contained in the measurements of yield, length of the reaction and initial (ini) and final (end) process variables. This type of chart can be used to identify relationships within the measured data, such as the interdependence of different variables and their relative impact. The blue lines, which begin at the origin, show the relationship between variables: two lines close to each other indicate a strong correlation, two lines at 90° indicates no correlation and two lines at 180° indicates negative correlation. Furthermore, the further away from the origin a variable lies, the stronger the impact that variable has on the model.

Additionally, Figure 2(B) shows the PCA scores plot for the same five data sets (D1–D5). The colored dots in this figure represent the value in the score of the objects in the PC space: If the observation markers lie in the same quadrant of a blue line in the loadings plot, it suggests a strong association with that variable. The score plot can also be used to assess the data structure of the observations and detect clusters, outliers, and trends.

The blue lines in Figure 2(A) suggest that, as might be expected, yield is strongly correlated to the length of reaction and the end-point pH, whereas it has little correlation with the end-point DO.

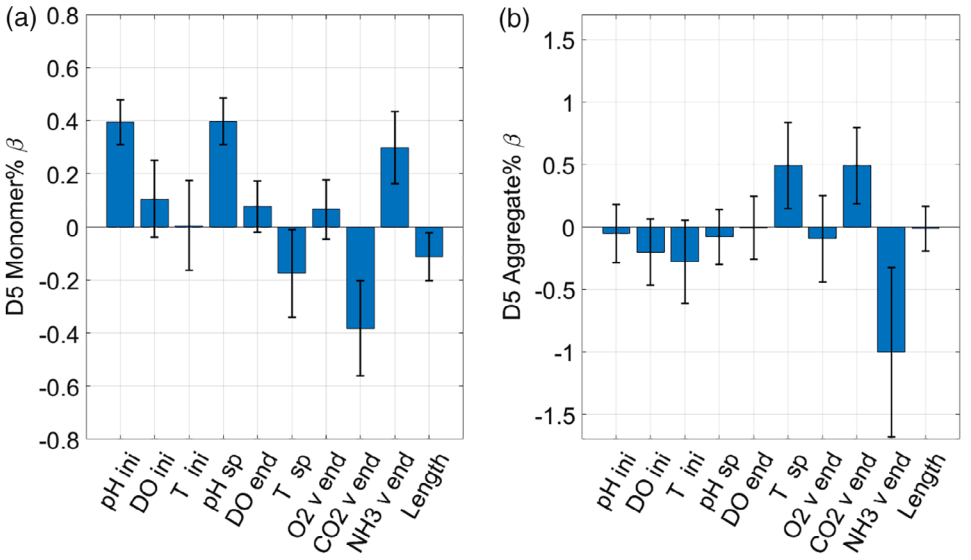

FIGURE 4 Partial least squares (PLS) regression coefficients of a model identified using D5, with error bars indicating twice their standard uncertainty. (A) PLS regression coefficients for monomer % estimation. (B) PLS regression coefficients for aggregate % estimation [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Main correlations of process variables

	Strong correlations with yield identified from PCA model	
Data sets	**Positive correlation**	**Negative correlation**
All data sets	Length of reaction, pH (end), DO (end)	None captured by the model
D1 and D2 (2–8 h)	pH (ini), pH (sp), Length of reaction, T (ini)	DO (ini)
D5 (10–14 h)	pH (ini), pH (sp), O_2 vol (end), CO_2 vol (end), NH_3 vol (end)	T (ini)
	Strong correlations with end-point qualities identified from PLS model	
Data sets	**Positive coefficients**	**Negative coefficients**
D1 and D2 (yield)	pH (ini), pH (sp), Length of reaction, T (ini)	DO (ini)
D5 (yield)	pH (ini), pH (sp), DO (end)	CO_2 vol (end)
D5 (monomer)	pH (ini), pH (sp), NH_3 vol (end)	CO_2 vol (end)
D5 (aggregate)	T (sp), CO_2 vol (end)	NH_3 vol (end)

Abbreviations: DO, dissolved oxygen; PCA, principal components analysis; PLS, partial least squares.

The scatter plot in Figure 2(B) also indicates that the length of reaction has an impact on yield since the data sets with the longer experiments, D4 and D5, are grouped closer to the yield variable. In addition, this plot shows that the data from D4 are separated from the other data sets. The reason for this could be that the length of the reaction is substantially longer and that the initial pH was substantially lower than most of the observations in the other data sets. This variability could have been partly caused by how the experiments were designed but it could also have been caused by the differences in how laboratory experimenters, A and B conducted the experiments. Furthermore, it is possible that the extract did not freeze uniformly, giving another possible source of variation between the experiments, which could have been alleviated by aliquoting the extract before freezing. Figure 2 provides an overview of the interactions between process variables, yet the set-points for pH were not constant for D3 and D4 which could provide a wrong representation of the correlation between pH and yield.

The PLS prediction coefficients offer an additional interpretation of the effect of process variables on yield. In a PLS coefficients plot, the prediction coefficients, β, provide a measure of the effect that each variable has on the predicted dependent variable(s). However, these effects need to be interpreted with caution since such interpretations assume that the model has accurately captured the cause to effect relationships within the process.

Figure 3 shows a comparison of PLS regression coefficients, with error bars indicating twice their standard uncertainty[33] that were obtained using bootstrap[34] replications, for the models identified on the shorter (D1 and D2) and longer (D5) reaction lengths. Process parameters whose error bars cross the axis indicate that either the parameters were not significant or were unreliable for yield prediction. Additionally, the PLS regression coefficients for monomer % and aggregate % estimation are shown in Figure 4. In these experiments (D1, D2, and D5), the PI controller set-points for pH and temperature were kept constant using different experimental configurations; hence, "sp" was used instead of "end" following the process variable names.

A summary of the main correlation identified from the PCA and PLS models are shown in Table 2.

PCA for shorter reactions (D1 and D2) and PCA for the longer reactions (D5) shows a high correlation between yield and pH for both data sets. On the other hand, yield had a slightly higher correlation with short duration batches (D1 and D2), as with the longer reaction lengths (D5) the reaction may have completed before the end of the batch. From a biochemical point of view, it is logical that as CFPS reactions are time limited, due to exhaustion of resources and accumulation of inhibitors, and so after a certain period of time, further increases in reaction time would lose significance as the reaction begins to slow.

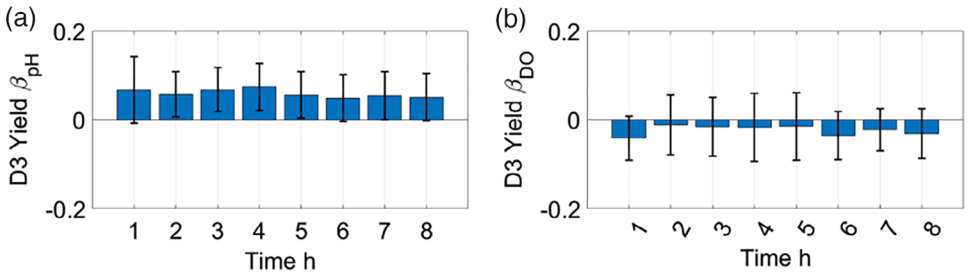

FIGURE 5 MPLS regression coefficients of (A) pH and (B) dissolved oxygen (DO) of a model identified using D3 designed to estimate yield, with error bars indicating twice their standard uncertainty [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Comparison of percent prediction error (SMAPE) using various modeling techniques [Color table can be viewed at wileyonlinelibrary.com]

		2D (initial and final measurements)			3D (time-variant measurements)		
		OLS	PLS	QPLS	MOLS	MPLS	MQPLS
Yield	D1 + D2	23.7	24.4	30.8	34.8	21.1	37.9
	D3 + D4	15.9	16.2	17.4	28.2	17.8	21.6
	D5	29.0	30.0	51.5	56.2	30.8	30
Monomer %	D5	4.4	4.5	6.0	5.4	3.4	6.5
Aggregate %	D5	30.8	27.5	33.4	48.5	28.4	38.1
Multiple	D5 yield	n/a	22.9	53.4	n/a	26.4	45.6
	D5 agg.	n/a	5.1	10.5	n/a	4.1	6.3
	D5 mon.	n/a	28.1	40.2	n/a	29.4	31.5

Abbreviations: MOLS, multiway OLS; MQPLS, multiway QPLS; OLS, ordinary least squares; PCA, principal components analysis; PLS, partial least squares; QPLS, quadratic PLS; SMAPE, symmetric mean absolute percentage error.

The PLS regression coefficients of the model identified using D1 and D2 shows similarities with the loadings of the PCA model identified on these same date-sets as shown in Table 2. Similarly, the PLS regression coefficients of the model identified using D5 are consistent with the loadings of the PCA model identified on the same data set, except for those coefficients corresponding to the volumetric values of O_2 and CO_2. This difference suggests two possibilities: that the volumetric O_2 and CO_2 values in the PC space are correlated with NH_3 and/or pH rather than with yield; or that either the PCA or the PLS model was not able to capture the correlations of these process variables. Since CO_2 and NH_3 were used to control pH as the reaction proceeded, it is logical that there was a relationship between these three factors.

The prediction coefficients of the PLS model can be extended over time using data sets with experimental conditions designed to capture this variability over time (D3) and through the identification of an MPLS model, as described in Sections 2.1 and 2.2. The coefficients from this model are shown in Figure 5. Since the PI controller of pH in D3 was designed to change process conditions over time, the regression coefficients of the MPLS model identified using D3, shows the effect that the different process variables have on estimated yield as the reaction proceeds and as Figure 5 shows the coefficients related to pH and DO vary with time. The coefficients of temperature were not included in Figure 5 since they did not provide any significant effect on yield. A possible cause for this was that the temperature set-points were kept constant and hence the model did have enough information to capture the time-varying behavior of temperature. The temperature range was selected to give a high and robust reaction, we would speculate that outside this temperature range the rate of reaction and thus titer would decrease.

Analysis of time varying MPLS coefficients, such as those shown in Figure 5, provides a useful tool for the optimization of process parameter trajectories during the reaction. These trajectories, also known as manipulated variable trajectories, provide the targets to be implemented by the feedback controllers that are in operation. Using this MPLS model, the process could be optimized by identifying an optimal trajectory for pH and DO that provides the most cost-effective solution for specific manufacturing requirements, or end-point qualities. Despite the potential benefits of MPLS models in CFPS optimization, its effectiveness relies on the accuracy of the model predictions and confirmation of the cause–effect relationships between process parameters and yield.

The correlations shown in this article must be interpreted with caution as the impact of pH and temperature is expected to vary to some degree from one target product to another and extract type to another. Simplistically, the ideal temperature and pH will be that which allows the optimal conformation of the protein or proteins which catalyze the rate limiting reactions. Additionally, the impact of pH on aggregation will likely vary depending on the product isoelectric point (pI).[35] Some degree of optimum pH and temperature variation with time is expected, as different reaction steps/sets (transcription, translation, metabolic) and their attendant enzymes become limiting. For example, at the beginning of a CFPS reaction transcription is expected to be limiting for at least 30 min to 1 h, as

mRNA accumulates.[36] Furthermore, there is evidence that the optimum conditions for transcription and translation are different.[37] Added to this, exhaustion of resources, in terms of macromolecular building blocks will differ from one product or DNA construct to another, while with a change in reaction mix composition and/or extract the rate at which inhibitors accumulate will also differ, and thus so may the impact of reaction length. In this context, the use of the MSDA tools proposed in this work provides a generic methodology to help identify optimal conditions.

4.2 | Prediction assessment

The accuracy with which the various MSDA modeling techniques described in Section 2.4 were able to predict the end-point quality variables from the CFPS process was compared using the symmetric mean absolute percentage error[38] between the estimated and the measured end-point quality using leave-one-out predictions. Table 3 shows the results of the comparison of the prediction errors from different modeling techniques, with the lowest errors highlighted by intensity in green and the highest errors highlighted by intensity in red for each identification data set.

The results for the two-dimensional data in Table 3, which contained the least complex data sets, show that the performance of the OLS models for the prediction of a single variable was better than any other compared technique followed closely by that of the PLS models. On the other hand, the results for three-dimensional data, where the measurements collected throughout the reaction process are considered, show that MOLS models had the largest prediction errors. The reason for this is that the additional measurements included in the $\boldsymbol{X}$ matrix are highly correlated and this has an adverse impact on the accuracy of OLS models.[39]

With respect to the computational speed of the studied modeling techniques, the difference in computing time was negligible relative to the long reaction times encountered in CFPS so the use of any of these techniques for predictive control should be feasible.

Table 4 shows the results of the comparison of the prediction errors using the approaches to address uneven vectors mentioned in Section 2.2, with the lowest errors highlighted in green and the highest errors highlighted in red for each identification data set. The best results in Table 4 were obtained using the first approach, which consisted on using measurements that were recorded up to the shortest reaction length, whereas the worst results were obtained using the third approach which consisted on using MD algorithms.

Models that use the length of reaction as an input, such as those shown in Tables 3 and 4, can be used for optimization but they are not useful for online prediction of end-point parameters (yield). In contrast, the process measurements used to identify the models from Table 5 did not need to consider the time of reaction as an input to the model; hence, the model can be used for online monitoring and control.

Table 5 shows a comparison of the prediction errors when MPLS models were used to provide long-term predictions of yield during the reaction, with the lowest errors highlighted by intensity in green and the highest errors highlighted by intensity in red for each identification data set. In each case, the errors that are reported are the errors when the end-point quality is predicted after the reaction had proceeded for a length of time, for example, after 1h, 2 h, and so forth. As discussed in Section 2.2, this introduces problems for some of the predictions later in the runs, as the lengths of the vectors for each run will vary depending on the length of time of each reaction run. Two techniques were applied to address this. The first technique identified multiple models as the reaction progressed. During the first hour of reaction, the model was identified using only measurements collected during the first hour of the reaction; during the second hour, measurements collected during the first and second hour of reaction

TABLE 4 Comparison of percent prediction error (SMAPE) with uneven vectors [Color table can be viewed at wileyonlinelibrary.com]

		Approach 1 (first hour)	Approach 2 (last hours)	Approach 3 (missing data)
Yield D1 + D2	OLS	33.8	34.8	42.1
	PLS	21.8	21.1	22.4
	QPLS	37.5	37.9	30.7
Yield D3 + D4	OLS	19.9	28.2	61.6
	PLS	18.5	17.8	19.5
	QPLS	20.5	21.6	22.1

Abbreviations: OLS, ordinary least squares; PLS, partial least squares; QPLS, quadratic PLS; SMAPE, symmetric mean absolute percentage error.

TABLE 5 Comparison of percent prediction error (SMAPE) with incomplete measurements [Color table can be viewed at wileyonlinelibrary.com]

		Time during reaction when end-point yield is predicted											
		1 h	2 h	3 h	4 h	5 h	6 h	7 h	8 h	9 h	10 h	11 h	12 h
MPLS yield	D1 and D2	26	25.5	24.3	23.8	22.2	22						
	D1 and D2 (MD)	68.4	57.4	28.1	28.4	23.5	23	23.2	23.2				
	D3 and D4	23.5	22.5	20.8	20.4	21	21.6	19.2	19.3				
	D3 and D4 (MD)	94.8	40.6	26.3	22.2	21.4	20.3	19.8	19.47	19.53	19.6	19.7	19.7
	D5	24.8	25.4	26.8	26.2	36	33.8	30.6	31.5	30.4	29.1	28.7	29
	D5 VIP	19.8	19.8	19.5	19.3	19.1	19.1	19	18.99	18.96	19.3	19	19.5
	D5 VIP (MD)	28	24.1	20.4	18.9	18.9	19.1	18.8	18.9	19	19.1	18	18.1

Abbreviations: SMAPE, symmetric mean absolute percentage error; VIP, variable of importance in projection.

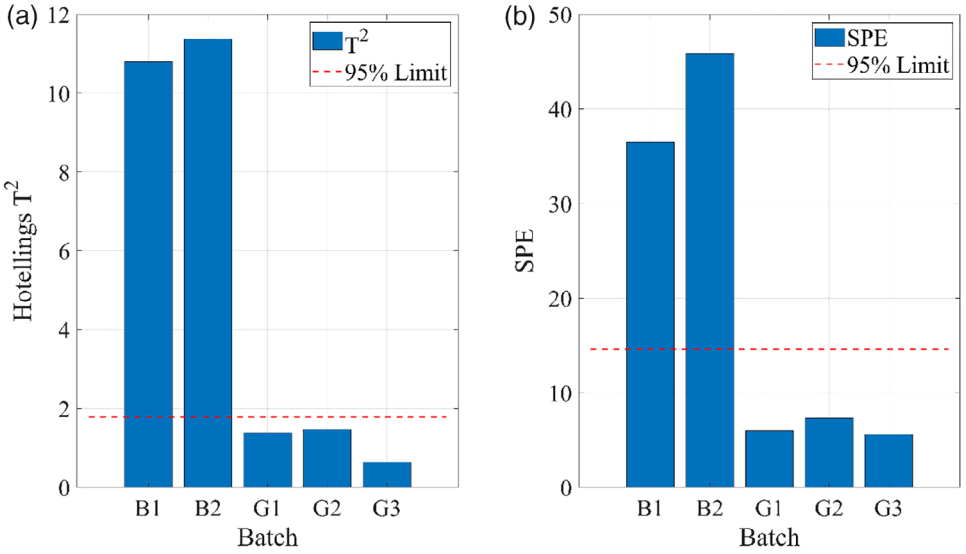

FIGURE 6 MPLS post batch charts of five observations from D3 and D4. (A) Hotelling's statistic, T^2. (B) Squared prediction error (SPE) statistic [Color figure can be viewed at wileyonlinelibrary.com]

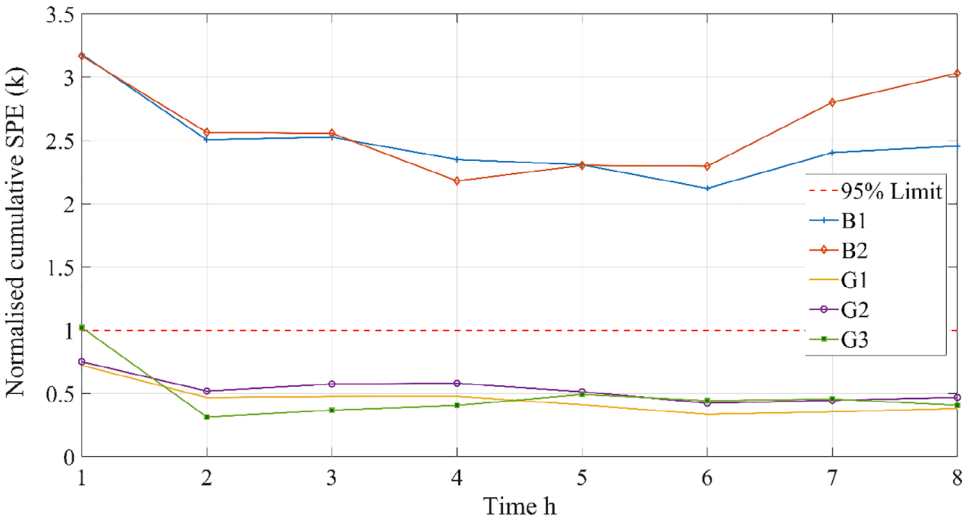

FIGURE 7 Normalized cumulative squared prediction error (SPE) batch charts [Color figure can be viewed at wileyonlinelibrary.com]

were used to identify the model and so on. Data collected from the runs with shorter reaction lengths was then discarded as their reaction lengths were exceeded. Hence, this approach utilized a different model to predict end-point yield for each hour of reaction.

The second approach calculated a single model with all the available measurements from all the observations in the data set and used MD techniques to estimate the "missing" future values of measurements for the shorter reactions and at each new "online" observation, as described in Section 2.2 (these are labeled MD in Table 5).

The lowest prediction errors in Table 5 were observed later in the reaction, rather than near the start. This is to be expected as later in the run, more information is available regarding how the reaction has proceeded and hence the prediction of end-point yield should be improved. The exception to this was with D5, where the lowest prediction error was at the start of the reaction. A possible cause of this inconsistency could be that many of the variables used to identify the model were not necessary for the prediction of yield and hence the developed model was overparameterized. To test if this was the case, a technique used to identify the most appropriate variables in the model, known as variable of importance in projection (VIP)[40] resulted in reduced prediction errors, particularly toward the middle and end of the batch. High VIP values were automatically selected to be included in the model, while lower VIP values were discarded. Lower VIP values were found to be usually of variables measurements which did not show a significant effect on yield prediction such as temperature in D5.

Another important observation from Table 5 is that predictions obtained using only data from the first hour of the reaction provided useful predictions of final yield using the multiple model approach. In contrast, predictions using MD algorithms seemed to be accurate only after about 3 h of reaction. However, the advantage of the MD procedure is that only one model was required.

4.3 | Postbatch analysis and monitoring using PLS charts

This section provides an example of CFPS batch analysis and online monitoring with MPLS statistical charts, with the intention being to

TABLE 6 Summary of MS data analysis techniques

Technique	Capabilities	Potential usage in CFPS
PCA loadings plot	Illustrate variables correlation	Process characterization and optimization, scale-up
PCA scores plot	Illustrate observation clustering and outliers	Process characterization and optimization, scale-up, comparability between experiments
PLS loadings	Illustrate variables correlation and importance to product quality variable(s)	Process characterization and optimization, scale-up
PLS pred. coefficients	Illustrate importance of variables to estimate product quality variable(s)	Process characterization and optimization
MPLS pred. coefficients	Illustrate importance over time of variables to estimate product quality variable(s)	Process characterization and optimization
VIP scores	Variable selection to discover the more relevant features of a model	Process characterization and optimization
T^2 and SPE charts	Detect outlying observations with respect to control limits	Offline batch analysis
$T^2(k)$ and SPE (k) charts	Detect outlying deviations of the process over time from control limits	Online monitoring

Abbreviations: CFPS, cell free protein synthesis; PCA, principal components analysis; PLS, partial least squares; SPE, squared prediction error; VIP, variable of importance in projection.

determine whether MSDA techniques could be used to monitor the process and provide early warning of abnormal operation. The MPLS model used for these charts was identified with the measurements from 37 observations from D3 and D4 where the yield was consistently higher than 0.3 mg/ml. Another set of 3 "good" observations (G1, G2, and G3) with yields closer to the mean value of the 37 observations were used to validate the procedure. In addition, measurements of two observations (B1 and B2) from D3 and D4 with yield lower than 0.2 mg/ml were used to represent poor quality or "faulty" reactions.

Figure 6 shows the post batch (a) T^2 and (b) SPE statistics charts for the faulty and the validation observations. In both charts, the "faulty" batches exceeded the 95% confidence limit, whereas the validation observations stayed within the 95% confidence interval limit, suggesting that the technique could be used to characterize different batches. These charts also show that the confidence limit in T^2 is lower relative to the confidence limit for SPE. This difference was also observed in another study,[41] where the author suggested that limits on SPE may reduce Type I and Type II errors compared with limits on T^2.

Figure 6 shows that B1 and B2 are outside the 95% confidence limit. After further analysis of the individual contributions of each variable to the SPE in B1 and B2, it was observed that faulty observations had larger values corresponding to measurements of pH and Temperature along the entire reaction. These large individual SPE values corresponded to measurements of unusually high temperature and lower than average pH compared to those in the identification data set. B1 and B2 were obtained from D3; hence, it is likely that the PRBS in the DOE caused these two observations to have a low pH during critical times of the reaction.

A useful approach for online monitoring of product quality is the use of cumulative contributions of process variables over time to SPE and T^2. These charts allow engineers to set operating windows for process measurement deviations, which if violated would suggest an adverse effect on CFPS performance. Figure 7 shows the cumulative SPE of faulty and validation observations through the batch, normalized to the magnitude of the confidence limits at each hour.

From the results in the chart, the SPE from B1 and B2 were consistently outside the confidence limit. On the other hand, the SPE of the validation observations stayed within the confidence limits, except the first hour of G3. A way to reduce false positives is to use more relaxed confidence limits (e.g., 99%), particularly in the early stages of the reaction. It is also important to consider that the observations used in this example are from experimental data in a research environment, and that the variability encountered among the observations is likely to be larger than that encountered in observations from standardized production.

MSDA could be extremely useful for quality control purposes. For example, if a fault is detected the operator could induce a response in the process parameters during the remainder of the reaction to bring the process back within the control limits. An automated way to achieve this objective has already been presented in literature for other applications. For example, in,[16] the authors use the same MPLS charts statistics and confidence limits within a cost function to provide a model predictive control strategy aimed at reducing quality variability in a fermentation process.

Table 6 summarizes the capabilities and potential use of MSDA in improving CFPS operations.

5 | CONCLUSIONS

This article provides a summary of a comprehensive exploratory data analysis that was carried out on measurements collected from a CFPS reaction system with the aim of discovering/confirming correlations between process conditions and final product properties. The ease with which high throughput data can be derived from CFPS reactions lends itself particularly well to big data analysis techniques. The aim of this article therefore was to analyze MSDA techniques available, identify those which are most suitable for CFPS analysis, and provide tools to allow CFPS practitioners to make use of these techniques. The

experimental results showed that when temperature and pH were held constant throughout the reaction using a feedback controller, pH and NH_3 had a positive correlation with yield and CO_2 and DO had a negative correlation. This was not necessarily the case for experiments where the controller set-points varied through the reaction, where the effect of process variables on yield was shown to vary as the reaction proceeded. In addition, the length of reaction was found to have a significant, positive impact on yield; although this impact and the impact of most of the other process variables were found to only be relevant in the first 10 h of reaction. Temperature, in the range of values used in the experiments, did not have a statistically significant impact on yield. However, temperature was found to have a strong relationship with monomer % and aggregates % estimation.

Multivariate regression models, and in particular PLS, were shown to be able to provide accurate prediction of end-point product quality during the CFPS process. This offers future potential for the development of feedback predictive control systems able to regulate end-point quality metrics by manipulating process variables, such as temperature and pH and is the subject of ongoing work.

Overall, the results presented in this article demonstrate that MSDA techniques offer a useful tool for extracting information from experimental data sets in CFPS and can be used for process characterization and identifying optimal experimental conditions. Moreover, it has been shown that these techniques can be effective in predicting and monitoring end-point quality parameters from measurements of process variables, even within the first hour of the synthesis reaction, which could be useful to improve large-scale manufacturing.

The applications for MSDA tools in CFPS range from process characterization to online automated multivariable control, regardless of the mode of operation (fed-batch or continuous). However, to effectively apply this technology, the implementation should consider which data from different unit operations are collected and analyzed to provide a robust support tool. In addition, if modeling techniques are to be successfully applied, automation and standardization are necessary since it is suspected that the largest sources of variation found in this study and others[42,43] were caused by the change of the operators, equipment and operating environment.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) with reference number: EP/P006485/1 and a consortium of industrial users and sector organizations in the Future Targeted Healthcare Manufacturing Hub hosted by UCL Biochemical Engineering in collaboration with UK universities. CFPS is envisioned as a potential solution for the simplified, robust, flexible, and local production of stratified or personalized biotherapeutic medicines, as might be necessary in a future pharmacy or hospital setting. The authors thank Noelle Colant and Stephen Morris from the Department of Biochemical Engineering at UCL for their help in the experiments. Special thanks to Nina Abi Carlos and Rishard Chen from Sutro for their collaboration and support, also to Gang Yin and James Zawada from Sutro for their insightful comments on the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://github.com/CarlosADuranVIllalobos/MSDA-Viral-vector and released version V1.0.0 at http://doi.org/10.5281/zenodo.4469875.

ORCID

Carlos A. Duran-Villalobos https://orcid.org/0000-0002-5350-761X

REFERENCES

1. Liu WQ, Zhang L, Chen M, Li J. Cell-free protein synthesis: recent advances in bacterial extract sources and expanded applications. *Biochem Eng J*. 2019;141:182-189.
2. Perez JG, Stark JC, Jewett MC. Cell-free synthetic biology: engineering beyond the cell. *Cold Spring Harb Perspect Biol*. 2016;8(12):1-25.
3. Ogonah OW, Polizzi KM, Bracewell DG. Cell free protein synthesis: a viable option for stratified medicines manufacturing? *Curr Opin Chem Eng*. 2017;18:77-83.
4. ICH. *ICH Official Web Site: ICH*. Geneva, Switzerland: *IchOrg*; 2015 https://www.ich.org/page/quality-guidelines.
5. Martin RW, Majewska NI, Chen CX, et al. Development of a CHO-based cell-free platform for synthesis of active monoclonal antibodies. *ACS Synth Biol*. 2017;6(7):1370-1379.
6. Murakami S, Matsumoto R, Kanamori T. Constructive approach for synthesis of a functional IgG using a reconstituted cell-free protein synthesis system. *Sci Rep*. 2019;9(671):1-13.
7. Glassey J, Gernaey KV, Clemens C, et al. Process analytical technology (PAT) for biopharmaceuticals. *Biotechnol J*. 2011;6(4):369-377.
8. Vilkhovoy M, Horvath N, Shih CH, et al. Sequence specific modeling of *E. coli* cell-free protein synthesis. *ACS Synth Biol*. 2018;7(8):1844-1857.
9. Caschera F, Bedau MA, Buchanan A, et al. Coping with complexity: machine learning optimization of cell-free protein synthesis. *Biotechnol Bioeng*. 2011;108(9):2218-2228.
10. Martin EB, Morris AJ. Enhanced bio-manufacturing through advanced multivariate statistical technologies. *J Biotechnol*. 2002;99(3):223-235.
11. Puri M, Solanki A, Padawer T, Tipparaju SM, Moreno WA, Pathak Y. Introduction to artificial neural network (ANN) as a predictive tool for drug design, discovery, delivery, and disposition: basic concepts and modeling. Basic concepts and modeling. *Artificial Neural Network for Drug Design, Delivery and Disposition*. Amsterdam, The Netherlands: Elsevier; 2016:3-13.
12. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol*. 1996;49(11):1225-1231.
13. Gabrielsson J, Lindberg NO, Lundstedt T. Multivariate methods in pharmaceutical applications. *J Chemometr*. 2002;16(3):141-160.
14. Liesum L, Kummli DS, Peinado A, McDowall N. The role of multivariate statistical process control in the pharma industry. *Multivariate Analysis in the Pharmaceutical Industry*. Amsterdam, The Netherlands: Elsevier; 2018:357-384.
15. Kourti T, Nomikos P, MacGregor JF. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J Process Control*. 1995;5(4):277-284.
16. Duran-Villalobos CA, Goldrick S, Lennox B. Multivariate statistical process control of an industrial-scale fed-batch simulator. *Comput Chem Eng*. 2020;132:106620.
17. Tomba E, de Martin M, Facco P, et al. General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling-an industrial case study. *Int J Pharm*. 2013;444(1–2):25-39.
18. Severson K, VanAntwerp JG, Natarajan V, Antoniou C, Thömmes J, Braatz RD. Elastic net with Monte Carlo sampling for data-based

modeling in biopharmaceutical manufacturing facilities. *Comput Chem Eng*. 2015;80:30-36.
19. Gunther JC, Seborg DE, Conner JS. Fault detection and diagnosis in industrial fed-batch cell culture. *IFAC Proc Vol*. 2006;39(2): 203-208.
20. Edwards-Parton S, Thornhill NF, Bracewell DG, Liddell JM, Titchener-Hooker NJ. Principal component score modeling for the rapid description of chromatographic separations. *Biotechnol Prog*. 2008;24 (1):202-208.
21. Gnoth S, Jenzsch M, Simutis R, Lübbert A. Control of cultivation processes for recombinant protein production: a review. *Bioprocess Biosyst Eng*. 2008;31(1):21-39.
22. Kirdar AO, Conner JS, Baclaski J, Rathore AS. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnol Prog*. 2007;23(1):61-67.
23. Miguez AM, McNerney MP, Styczynski MP. Metabolic profiling of *Escherichia coli*-based cell-free expression systems for process optimization. *Ind Eng Chem Res*. 2019;58(50):22472-22482.
24. Zawada JF, Yin G, Steiner AR, et al. Microscale to manufacturing scale-up of cell-free cytokine production-a new approach for shortening protein production development timelines. *Biotechnol Bioeng*. 2011;108(7):1570-1578.
25. Yin G, Garces ED, Yang J, et al. Aglycosylated antibodies and antibody fragments produced in a scalable in vitro transcription-translation system. *MAbs*. 2012;4(2):217-225.
26. Colant N, Melinek B, Teneb J, et al. A rational approach to improving titer in *Escherichia coli*-based cell-free protein synthesis reactions. *Biotechnol Prog*. 2020;37:e3062(1):1-16.
27. Franco R, Daniela G, Fabrizio M, Ilaria G, Detlev H. Influence of osmolarity and pH increase to achieve a reduction of monoclonal antibodies aggregates in a production process. *Cytotechnology*. 1999;29 (1):11-25.
28. Kim TW, Kim HC, Oh IS, Kim DM. A highly efficient and economical cell-free protein synthesis system using the S12 extract of *Escherichia coli*. *Biotechnol Bioprocess Eng*. 2008;13(4):464-469.
29. Wold S, Geladi P, Esbensen K, Öhman J. Multi-way principal components-and PLS-analysis. *J Chemometr*. 1987;1(1):41-56.
30. MacGregor J, Cinar A. Monitoring, fault diagnosis, fault-tolerant control and optimization: data driven methods. *Comput Chem Eng*. 2012; 47:111-120.
31. Nelson P, Taylor P. Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemom Intel Lab Syst*. 1996; 35:45-65.
32. de Jong S, de Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemom Intel Lab Syst*. 1993;18(3):251-263.
33. Martens H, Høy M, Westad F, Folkenberg D, Martens M. Analysis of designed experiments by stabilised PLS regression and jack-knifing. *Chemometrics and Intelligent Laboratory Systems*. 2001;58(2): 151-170.
34. Duchesne C, MacGregor JF. Jackknife and bootstrap methods in the identification of dynamic models. *J Process Control*. 2001;11(5): 553-564.
35. Li R, Wu Z, Wangb Y, Ding L, Wang Y. Role of pH-induced structural change in protein aggregation in foam fractionation of bovine serum albumin. *Biotechnol Rep*. 2016;9:46-52.
36. Marshall R, Noireaux V. Quantitative modeling of transcription and translation of an all-*E. coli* cell-free system. *Sci Rep*. 2019;9(1):1-12.
37. Georgi V, Georgi L, Blechert M, et al. On-chip automation of cell-free protein synthesis: new opportunities due to a novel reaction mode. *Lab Chip*. 2016;16(2):269-281.
38. Kreinovich V, Nguyen HT, Ouncharoen R. How to Estimate Forecasting Quality: A System-Motivated Derivation of Symmetric Mean Absolute Percentage Error (SMAPE) and Other Similar Characteristics; 2014. http://digitalcommons.utep.edu/cs_techrephttp://digitalcommons.utep.edu/cs_techrep/865.
39. Yeniay O, Goktas A. A comparison of partial least squares regression with other prediction methods. *Hacettepe J Math Stat*. 2002;31:99-111. http://www.mat.hacettepe.edu.tr/hjms/english/issues/vol31/full-text/99-111.pdf.
40. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemom Intel Lab Syst*. 2012;118:62-69.
41. Qin SJ. Statistical process monitoring: basics and beyond. *J Chemom J Chemom*. 2003;17:480-502.
42. Romantseva E, Strychalski EA. CELL-FREE (Comparable Engineered Living Lysates for Research Education and Entrepreneurship) Workshop Report. La Jolla, CA; 2019. https://www.nist.gov/system/files/documents/2019/06/19/nist_cell-free_workshop_report.pdf.
43. Dopp JL, Jo YR, Reuel NF. Methods to reduce variability in *E. coli*-based cell-free protein expression experiments. *Synth Syst Biotechnol*. 2019;4(4):204-211.
44. Hotelling H. Analysis of a complex of statistical variables into principal components. *Br J Educ Psychol*. 1932;24:417-520.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Duran-Villalobos CA, Ogonah O, Melinek B, Bracewell DG, Hallam T, Lennox B. Multivariate statistical data analysis of cell-free protein synthesis toward monitoring and control. *AIChE J*. 2021;67:e17257. https://doi.org/10.1002/aic.17257