Check for updates

Received: 5 January 2024 | Revised: 25 March 2024 | Accepted: 28 March 2024

DOI: 10.1002/btpr.3467

RESEARCH ARTICLE

Cell Culture and Tissue Engineering

BIOTECHNOLOGY PROGRESS

Multivariate data analysis of process parameters affecting the growth and productivity of stable Chinese hamster ovary cell pools expressing SARS-CoV-2 spike protein as vaccine antigen in early process development

Sebastian-Juan Reyes[1,2] | Lucas Lemire[1,2] | Raul-Santiago Molina[3] | Marjolaine Roy[2] | Helene L'Ecuyer-Coelho[2] | Yuliya Martynova[2] | Brian Cass[2] | Robert Voyer[2] | Yves Durocher[2] | Olivier Henry[1] | Phuong Lan Pham[2]

[1]Department of Chemical Engineering, Polytechnique Montreal, Montreal, Canada

[2]Human Health Therapeutics Research Centre, National Research Council Canada, Canada

[3]Proelium S.A.S, Bogotá, Colombia

Correspondence
Phuong Lan Pham, Human Health Therapeutics Research Centre, National Research Council Canada, 6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada.
Email: phuonglan.pham@nrc-cnrc.gc.ca

Olivier Henry, Department of Chemical Engineering, Polytechnique Montreal, Montreal, Quebec H3T 1J4, Canada.
Email: olivier.henry@polymtl.ca

Present address
Helene L'Ecuyer-Coelho, Biodextris, 525 Bd Cartier W, Laval, H7V 3S8, Laval, Quebec, Canada.

Funding information
Natural Sciences and Engineering Research Council of Canada, NSERC-CREATE PrEEmiuM, Grant/Award Number: RGPIN/4048-2021; National Research Council Canada, Grant/Award Number: PR-023-1

Abstract

The recent COVID-19 pandemic revealed an urgent need to develop robust cell culture platforms which can react rapidly to respond to this kind of global health issue. Chinese hamster ovary (CHO) stable pools can be a vital alternative to quickly provide gram amounts of recombinant proteins required for early-phase clinical assays. In this study, we analyze early process development data of recombinant trimeric spike protein Cumate-inducible manufacturing platform utilizing CHO stable pool as a preferred production host across three different stirred-tank bioreactor scales (0.75, 1, and 10 L). The impact of cell passage number as an indicator of cell age, methionine sulfoximine (MSX) concentration as a selection pressure, and cell seeding density was investigated using stable pools expressing three variants of concern. Multivariate data analysis with principal component analysis and batch-wise unfolding technique was applied to evaluate the effect of critical process parameters on production variability and a random forest (RF) model was developed to forecast protein production. In order to further improve process understanding, the RF model was analyzed with Shapley value dependency plots so as to determine what ranges of variables were most associated with increased protein production. Increasing longevity, controlling lactate build-up, and altering pH deadband are considered promising approaches to improve overall culture outcomes. The results also demonstrated that these pools are in general stable expressing similar level of spike proteins up to cell passage 11 (~31 cell generations). This enables to expand enough cells required to seed large volume of 200–2000 L bioreactor.

KEYWORDS

CHO stable pool, Cumate induction, fed-batch bioreactor production, MVDA, random forest, SARS-CoV-2 trimeric spike protein

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2024 National Research Council Canada. *Biotechnology Progress* published by Wiley-VCH GmbH on behalf of American Institute of Chemical Engineers. Reproduced with the permission of the Minister of Innovation, Science, and Economic Development.

1 | INTRODUCTION

Chinese hamster ovary (CHO) cells are currently the industry standard when it comes to producing recombinant therapeutic proteins. This is because CHO cells are able to produce human-like proteins that have adequate glycosylation profiles. CHO cells can come in two flavors, cell lines derived from single clones and stable pools. Cell line development requires numerous stringent screening procedures that select the best performing cell clone.[1] These tests are generally carried out at the micro-liter scale to evaluate growth, protein yield, and critical quality attributes (CQAs). Such screening procedures can take months to complete and consequently these strategies are sub-optimal when reacting to rapidly evolving public health crisis, such as the global pandemic caused by the SARS-CoV-2.[2,3] To fast-track the cell line creation to large-scale protein production, mammalian cell stable pools have been considered an attractive alternative.[4] This is because even though the cell pools are less homogeneous than their cell line counterparts, they can still be utilized to produce recombinant proteins at a sufficiently large-scale. It is noted however that cell pools can be susceptible to cell age effects and thus for stable pools to be a viable alternative to cell lines to provide enough materials for toxicology study and Phase 1 clinical trials, the process must be understood and optimized.[5] Optimizing pool process intensification can rely on design-of-experiment to identify process conditions that enhance culture performance in a given production scale. Once sufficient data are generated across various production systems and with different pools, exploratory multivariate data analysis (MVDA) can be utilized to sort through the data to understand what parameters promote good performance or impact adversely the production process.[6] Additionally, soft sensors that predict process outcomes can also be developed not only to predict future yields based on monitored variables but to also understand the variable importance within the model.[7] When variable impact on model is combined with intimate process knowledge, insights can be gained to create representative models or improve process performance.

In terms of large dataset analysis, principal component analysis (PCA) and partial least square (PLS) are the two tools frequently used in biomanufacturing. PCA is generally employed as the first step of MVDA to explore the large data structure and provide process dynamics. PLS can be applied as the second part of MVDA focusing on process optimization and forecasting. In the first part of this article, we will focus on the use of PCA to reduce data dimensionality thus facilitate data visualization and comparison.[8] PCA can uncover correlations among variables or show relationships between outcomes and variables to ultimately detect trends and outliers. PCA is a process that transforms a large dataset with collinearity into a low dimensional space of new uncorrelated variables so-called principal components (PCs) in such a way that each PC explains a certain portion of the overall variance within the dataset.[8,9] In biomanufacturing, PCA has been readily used to evaluate the success of scale-up/scale-down and process impacts on glycosylation profiles.[10–14] Bioprocessing data can be considered as a three-dimensional dataset that consists of various experiments where all the variables vary through culture time (lot number, process parameters, sampling time points). The first step in PCA is data unfolding. There are numerous approaches such as batch-wise, time-wise, and variable-wise unfolding. Collectively, these methods are called multiway PCA (MPCA).[15,16] Batch-wise unfolding allows the direct comparison between different batches although granularity along the time dimension is lost.[9,17] However, by comparing among batches, different experimental clusters can be visualized and then understood depending on how various process parameters were changed between the conditions. Although PCA is generally used on late process development or manufacturing datasets to evaluate outliers and improve process robustness, studies have shown that applying similar techniques to early process development datasets can be beneficial revealing information such as uncontrolled variance and experimental flaws.[18,19] As mentioned above, PLS is a similar tool that transforms the original dataset to latent variables, thus reducing the original dimensionality.[8,20] However, a key difference is that PLS relates a feature vector (X) to a response vector (Y) which can then be utilized in regression to predict outcomes. This tool has also been applied widely in bioprocessing scale-up.[21] For instance, PLS was used as a regression technique to predict monoclonal antibody (mAb) production and compare a 3 L scale process with a 2000 L scale process.[22] Through interpreting the PLS loading plots and altering the process conditions of the 3 L bioreactor, a comparable process was created at small-scale that displayed similar behaviors and outcomes when compared to the 2000 L process. Thus, the 3 L bioreactor can be a scale-down model to predict the 2000 L performance. Even though the PLS technique is highly interpretable (carry out dimensionality reduction on feature vector X and response vector Y and then realize a linear regression with the non-correlated latent variables), it has the downside of utilizing a linearity assumption.[8] Importantly, when PLS was coupled with amino acid stochiometric balances an approach to rapidly optimize amino acid concentrations in chemically defined media additives was developed. Here, PLS was deployed to comprehend the relationship between the dynamics of time-dependent stochiometric balances and critical response variables like cell growth or mAb productivity.[23] Consequently, important nutrients that impacted specific response variables or cellular phenotypical states could be detected and further translated into experimental designs for validation studies.[23] Given that biological processes are nonlinear in nature, less interpretable but better performing techniques can be utilized through machine learning (ML) models (such as support vector machine [SVM] regression), especially if the goal is to develop a soft sensor.[24] Such technique has been applied in large-scale bioreactors to predict titer based on features.[25] Lactate metabolism was found to be a key process indicator (KPI) in terms of predicting final protein yield and suggested the importance of controlling glycolytic fluxes during seed train inoculation at large-scale. Tree-based models are also powerful tools that can accurately model nonlinearity among biological systems.[8] Here, data can be segmented into ranges and a decision tree can represent the outcome of a response variable depending on what ranges various input variables are observed at. Since this approach can lead to overfitting, which is an undesirable ML characteristic, random forest (RF) can be realized. RF

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

techniques are able to run a collection of decision trees in parallel and then return the average prediction of each decision tree. Since it averages out multiple regression trees, RF can overcome the overfitting drawbacks that are observed with single decision trees.[26] Such approaches have also been utilized in the modeling of Raman spectroscopy and shown to be an alternative to the PLS gold standard.[27] Extreme Gradient Boosting algorithms (XGBoost) are also decision tree-based methods and differ with RF in the sense that trees are constructed sequentially rather than in parallel. This allows for a gradual improvement of the decision tree by continuously re-weighting trees that correctly predict outcomes that were previously poorly modeled by previous decision trees.[8] Since this objective function can be optimized with a learning rate, it can be more prone to overfitting and as such, care must be taken when building and tunning the model. Alternative approaches applied in bioprocessing include multiple linear regression, *k*-nearest neighbors, Gaussian process regression, classification, and regression tree, and ensemble approaches (Gradient Boosting Machine, Adaptive Boosting).[8,26,28]

MVDA tools can help improve our understanding of a process which is key to develop a robust production platform capable of being transferred to different cell types for various target proteins. It is also known that ML models can play a role in modeling the interaction between variables and outcomes. This, in turn, provides a pipeline for transferring knowledge gained during the early process development to the manufacturing stage where soft sensor prediction capability will be greatly increased given that the process will be locked in to a specific range within the design space (DS); the latter is needed to assure predefined CQA.[6] Such tools are part of the process analytical technology initiative and an important aspect of Biopharma 4.0 manufacturing which aims to improve process understanding.[6]

The current article focuses on applying MPCA to examine early process development data with the purpose of expanding process knowledge and improving conditions that maximize SARS-CoV-2 trimeric spike protein production using stable pools instead of stable clones to accelerate development timelines. Further modeling with RF demonstrated that endpoint product titer can be predicted utilizing key cumulative and endpoint process values. In-depth analysis of the model demonstrated that improving overall longevity of the cell culture as well as limiting lactate build-up are key variables that if tunned appropriately with process related changes could improve spike protein yields. It is worth mentioning that SARS-CoV-2 spike protein is a difficult-to-express protein due to its structural complexity. Any upstream process development to improve its yield would be valuable to help manufacture this potential vaccine antigen.[3]

2 | MATERIALS AND METHODS

2.1 | Stable CHO cell pool and small-scale cell culture conditions

Four stable CHO cell pools expressing SmT1 trimeric spike proteins namely Wuhan (Wu), Wuhan Tagless (WuTL), Delta (De), and Beta (Be) variant were generated as described previously.[3] Pool cells were thawed and grown in BalanCD CHO Growth A medium (Fujifilm/Irvine Scientific) supplemented with 50 μM MSX (L-Methionine sulfoximine, Sigma-Aldrich) and 0.1% (w/v) Kolliphor P188 surfactant. A total of 125 mL (20 mL working volume) shake flasks without baffles (Corning) were used for cell maintenance and passage. The flasks were shaken at 120 rpm (25 mm orbital diameter) in an incubator regulated at 37°C, 5% CO_2, and 75% relative humidity. Cells were passaged every 2 or 3 days to keep a maximum viable cell density (VCD) between 2 and 3×10^6 cells/mL.

2.2 | Cell culture analytical methods

Cell density, viability, main metabolites (glucose, lactate, ammonia) were measured utilizing the previously reported methodology.[2,29] Briefly, cell counts were realized with Innovatis Cedex (Roche) or ViCell Blue (Beckman Coulter) automated cell counter using trypan blue dye exclusion assay. Key metabolites such as glucose, lactate, and ammonia were determined using the Vitros 350 Chemistry System (Orthoclinical Diagnostics). Volumetric protein titers were estimated using TGX Stain-free SDS-PAGE gels (Bio-Rad) quantification method.

2.3 | Bioreactor fed-batch process

The bioreactors were seeded at 0.2×10^6 cells/mL (low-seed) or 0.4×10^6 cells/mL (high-seed) and cultivated for 17 days in the fed-batch mode. Temperature downshift (from 37 to 32°C) was realized 3 days after seeding. A pH shift (from 7.05 ± 0.05 to 6.95 ± 0.05) was performed on all batches 2 days after seeding. Induction was conducted with 4-Isopropylbenzenecarboxylate (Cumate, ArkPham), concomitantly with the temperature downshift. Cultures were fed with BalanCD CHO Feed 4 (Fujifilm/Irvine Scientific) and supplemented with glucose to maintain the concentration above 17 mM (3 g/L) for the next sampling point. Samples were taken from the bioreactors on days −3, −2, −1, 0, 3, 5, 7, 10, 12, and 14 days post-induction (dpi) for off-line analysis, while feeding was realized in a bolus dosage from 0 dpi onward. Table 1 shows a summary of studied process conditions. The impact of seeding density (low vs. high), cell age through cell passage number (passage 5, 8, 11), MSX concentration (50 vs. 125 μM) on process outputs was examined. Volumetric power input (P/V) indicating the relationship between agitation speed and culture volume was set in a range between 40 and 30 W/m^3 for the Multifors 0.75 L (Infors) and BioFlo 1 L (Eppendorf) systems. The final P/V value was decreased due to volume increase with feed events while keeping a same agitation speed. For the BioFlo 10 L bioreactor (Eppendorf), a P/V range between 20 and 80 W/m^3 was explored. A dissolved oxygen (DO) setpoint of 40% (of air saturation) was chosen for the Multifors 0.75 L and BioFlo 1 L systems while for the BioFlo 10 L bioreactor, DO of 40% and 60% were studied. A Kolliphor P188 surfactant concentration of 0.2% (w/v) was used for the Multifors 0.75 L

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Bioreactor production process conditions.

Bioreactor System	Pool	Seeding Density (10^6 cells/mL)	Cell passage number	MSX (μM)	P/V range (W/m^3)	DO (%)	Kolliphor P188 (%, w/v)	Sparger	Aeration strategy	Number of impellers
Multifors 0.75 L (Infors)	Wuhan (Wu) Delta (De) Beta (Be) WuhanTL (WuTL)	Low: 0.2 High: 0.4	5 8 11	50 125	40–30	40	0.2	Micro	Air cap (AC) with Air/O_2 cascade	2
BioFlo 1 L (Eppendorf)	Wuhan (Wu)	High: 0.4	5	50	40–30	40	0.2	Macro Micro Dual	Air cap (AC) No air cap	1
BioFlo 10 L (Eppendorf)	Wuhan (Wu)	High: 0.4	5	50	20–80	40 60	0.2 0.6	Macro Micro Dual	Air cap (AC) No air cap	1 2

and BioFlo 1 L whereas a variation from 0.2% to 0.6% (w/v) Kolliphor P188 was investigated in the BioFlo 10 L system. In the Multifors 0.75 L, micro-spargers (10 μm pore diameter) with an air cap (AC) were used in a cascade air/oxygen strategy. Cell passage number and MSX concentration were varied across batches. For the BioFlo 1 L, micro-sparger, macro-sparger, and dual sparger composed of a micro-sparger and a macro-sparger were compared. A mix between air caped and no air caped strategies were also studied within this sub dataset. The BioFlo 10 L studies encompassed a variety of process conditions such as ACs, sparger type, agitation rate, and number of impellers. It must be noted that for bioreactor runs that employed dual sparger configuration, the macro-sparger sent only air (AC) while the micro-sparger injected flow of both oxygen and carbon dioxide as needed.

2.4 | Dataset structure and batch-wise unfolding method

The dataset is made up of 59 batches (productions). It is worth mentioning we kept the terminology "batch" to indicate a production run as this term is widely used in data science. All productions (batches) were performed in fed-batch mode. Of the total 59 productions, 38 were conducted in the Multifors 0.75 L, 13 batches in the BioFlo 1 L, and 8 batches were realized in the BioFlo 10 L. For the Multifors 0.75 L parallel benchtop bioreactor platform, 14 batches were run with Wuhan pool, 12 batches were realized with Delta pool, 6 batches with Beta pool, and 6 batches with Wuhan Tagless pool. The other two systems (BioFlo 1 and 10 L) used exclusively the Wuhan pool. Table 2 shows the variables considered in the MVDA. Product titer was not listed given that the Wuhan pool sub dataset has only endpoint titers while titer evolution profiles were available for Delta, Beta, and Wuhan Tag-less pools. It was decided to exclude titer as a variable to keep analysis consistent across pools. However, during the analysis of variable relationship, endpoint titers were considered for result discussion. Viability, VCD, residual glucose, lactate, and ammonia were measured following the same schedule as mentioned in the process conditions section above. Calculated values such as integral viable cell concentration (IVCC), glucose consumed per day, and cell specific metabolic rates (glucose, lactate, ammonia) were estimated using the same procedure across all batches. Online data from the bioreactor runs were also added into the dataset. pH, cumulative volumetric base addition, cumulative oxygen flow, and cumulative carbon dioxide flow were calculated into daily averages such that direct comparison with the sampling day's data could be made. The cumulative gas flows were normalized with respect to the bioreactor volume so as it represents the cumulative gas volume per liquid volume per minute (VVM) for each gas. Batch-wise unfolding of the data (Figure 1) was realized in such a manner, so each row represents a

TABLE 2 Variables considered in batch-wise multiway PCA (MPCA).

Offline measurements	Cell growth	IVCC, cells*day/mL
		VCD, cells/mL
		Viability, %
	Metabolites	Residual glucose, mM
		Lactate, mM
		Ammonia, mM
		Total consumed glucose (TCG), mM
	Cell specific metabolite rates	Glucose (qGlc), pmol/cell/day
		Lactate (qLac), pmol/cell/day
		Ammonia (qAmon), pmol/cell/day
Online environmental continuous measurements	Gas sparging	Oxygen, mL/min
		Air, mL/min
	pH control	pH profile
		Base addition volume, mL
		Carbon dioxide, mL/min

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

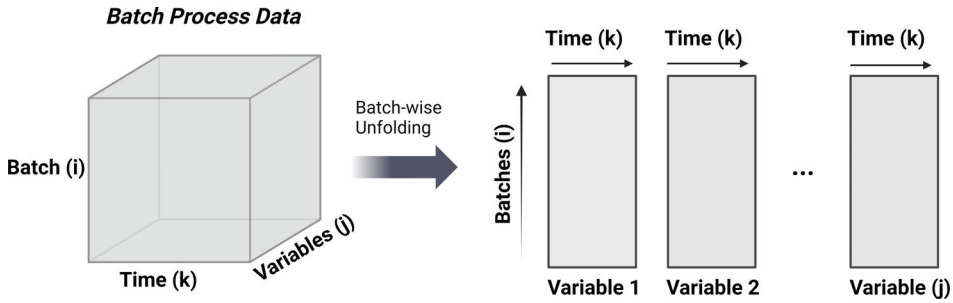

FIGURE 1 Batch-wise unfolding from 3D dataset to 2D data arrangement. Each row represents a batch (*i*). Time (*k*) and variables (*j*) are presented as columns and arranged in a cyclical mode such that every variable is a block from time 0 to *k*.

batch, and each column represents a variable at a given sampling time-point (−3 to 14 dpi, Day post-induction).

For this unfolding method, each experiment (batch) becomes a score in the PC plot effectively collapsing the time dimension while there is a loading value for each sampling point for every variable. Loadings (presented in Supporting Information) provide information about how each variable contributes to the formation of the PCs and can help interpret the data structure and relationships among variables within the dataset. Consequently, loading plots can help understand what variables are driving the position of a given score (batch experiment) in the PC axis. Loadings can be positive or negative representing the direction of the relationship (positive loadings indicate positive correlations between the variable and the PC while negative loadings represent negative correlations). Magnitude of the loading values are also relevant as they represent the strength of the relationship. Larger values suggest more significant influences of the variable on the PC. It must be noted that since all the runs were performed to investigate specific process parameters (cell passage number, seeding density, MSX concentration, pool type), the objective of the MVDA is to infer global trends within the dataset and determine KPI. This is especially important since there is a lack of published information regarding inducible pools and much less large dataset analysis of such production platform. To the best of our knowledge, this is the first report demonstrating the benefit of MVDA to help interpret the data obtained during the early bioreactor process development of an inducible stable CHO pool.

2.5 | MVDA approach, software, and package

Pre-processing of online data and analysis was carried out using **R** programming language. The **mdatools**[30] package was utilized for PCA while the **caret**[31] package was employed to build the ML models. Multi-steps were used to treat the data. First, Savitzky–Golay filtering[32] was carried out on pH data before calculating daily averages for bioreactor online data (pH, oxygen flow, carbon dioxide flow, base addition) so as to filter out noise sensor data. For the remaining variables (base, oxygen, and carbon dioxide sparging), root cause analysis was carried out to determine if abnormal behavior could be explained by sensor faults and thus excluded from analysis. Second, daily averages and sampled variables were gathered in Excel spreadsheets (Microsoft) and arranged such that they matched in the time dimension. Daily averages of the online data were taken and for the sparge rates, since they are highly variable (on/off flow of pure oxygen and carbon dioxide), cumulative of said variables (oxygen and carbon dioxide flow) were taken so as to compare trends. Lastly, of the daily online bioreactor values, only values that matched with sampling days were taken so as to not bias the dataset (in the 17-day fed-batch process, there would be 10 offline sampling data values for each variable while there would be 17 values for each online variable). For the online variables, values starting from −2 dpi (culture Day 1) were included since base addition and oxygen sparging are nonexistent in the first day of culture (−3 dpi = culture Day 0). Input data (IVCC, VCD, viability, lactate, ammonia, cumulative glucose consumed, residual glucose, base addition, cumulative oxygen sparged, cumulative carbon dioxide sparged, pH, qLac, qGlc, qAmon) were arranged in batch-wise unfolding so as to carry out PCA where score values and PC axis generation can be considered the output for the purpose of data visualization. Loading plots, presented in supplemental materials (see Figures S1–S18) were utilized to determine the driving factors behind score plot distributions.

Once correlations among variables were better understood through MPCA, key parameters describing the behavior of each batch (IVCC, endpoint viability, max lactate, endpoint lactate, endpoint ammonia, cumulative glucose consumed, endpoint residual glucose, total base addition, cumulative oxygen sparged, cumulative carbon dioxide sparged, endpoint pH, pool type, passage number) were calculated and used as input features to predict endpoint titers (output) using four modeling methods as described below. MSX concentration was excluded from the model because the extra MSX supplementation did not show impact as demonstrated in the Results and Discussion section below.

SVM, PLS, RF, and XGBoost were utilized as regression methods to relate the input variables (features) to the output variable (titer). The total dataset used to generate the regression models was made up of 50 batches. It was split into training (80%) and test (20%) sub datasets. Model metrics were obtained for both training and test datasets. Each model was subjected to the same hyperparameter tuning strategy. Adaptive resampling of the tuning parameter grid was realized in such a way that the random search of hyperparameters is concentrated on values that are in the neighborhood of the optimal parameters by discarding settings judged sub-optimal. To assess the performance of the training regression models, bootstrapping was conducted using the training and test dataset separately. For each

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

iteration ($i = 100$), a bootstrap sample was created by resampling the imputed sub dataset (training or test) with replacement. The regression model was then used to predict the target variable for the bootstrap sample. Key evaluation metrics including root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) were evaluated at each iteration. By repeating this process multiple times, a distribution of these metrics was obtained, providing insights into the performance and variability of the pre-trained regression models on both the training and test datasets. Best performing model based on RMSE, MAE, and R^2 was then analyzed with Shapely value dependency plots to obtain more information about the impact each feature has on the prediction outcome and at what range of values said features had a positive or negative impact on prediction. The **R caret** package was utilized to build the models, **boot** package[33,34] was used to get bootstrapped statistics, and **fastshap** package[35] was employed to construct the Shapely value dependency plots.

3 | RESULTS AND DISCUSSION

3.1 | Seeding cell density impact

To investigate the impact of seeding cell densities, two densities (low: 0.2×10^6 cells/mL and high: 0.4×10^6 cells/mL) were evaluated within 14 batches. Our previous data (not shown) demonstrated that higher density at induction contributes to accelerate production pace thus shortens process duration without affecting final titers. However, the superior seeding density bound needs to be controlled mainly because the existing feed regimen has been designed to support a specific range of VCD. Post-induction cell overgrowth due to high induction cell density may lead to premature cellular decline probably due to nutrient limitation occurred at high cell biomass.

From Figure 2a, it is clear that two distinct clusters are created. The orange cluster represents productions seeded at high density while bioreactors seeded at low density are distributed within the green cluster. When evaluating what variables drive this phenomenon in the principal component 1 (PC1) axis, it is found that batches located on the positive PC1 axis have higher lactate accumulation (27.8 mM peak lactate vs. 23.7 mM peak lactate), increased peak VCD during the 3-day growth phase (4.35×10^6 c/mL vs. 2.98×10^6 c/mL), and increased oxygen requirements (16.77 vs. 10.52 mL/min). These variables are intimately related with increasing biomass which explains the segmentation based in seeding density. It is worth mentioning when the endpoint titers and viabilities are compared, no clear relationship with respect to seeding density can be attributed (Table 3).

Figure 2b shows that some productions differ in the principal component 3 (PC3) axis. This component is primarily driven by addition of base to regulate pH (negative PC3) and increasing oxygen sparging flowrate (positive PC3). Given that links between oxygen consumption and protein production have been investigated,[36] the final protein production of the batches at opposite ends of the PC3 axis was evaluated. It was determined that the points lying on near the origin and the right-hand side of the PC3 axis had an average protein expression of 1064 ± 137 mg/L ($n = 10$) while the points lying on the left side of the PC3 axis had a final protein expression of 765 ± 83 mg/L ($n = 4$). Since the variation in protein expression happened in both high and low seeding density conditions, a deep dive into the time series data was required to explore the root cause of this variation. As such, time profiles were plotted (Figure 3). High performers (blue) were defined as batches that were above average in terms of final protein expression whereas low performers (red) were assigned as batches that had below-average protein expression.

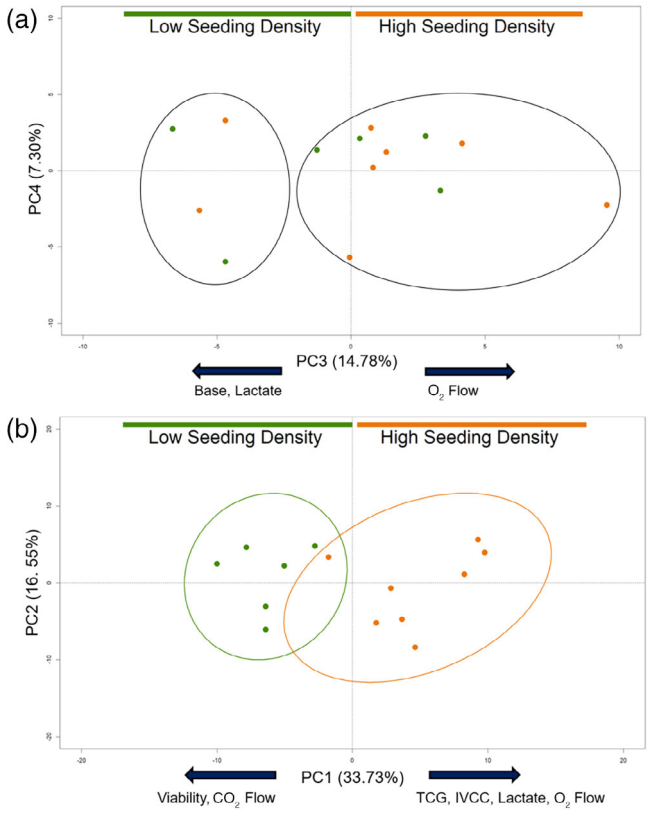

FIGURE 2 Principal component (PC) scatter plots for Wu pool in Multifors 0.75 L bioreactors. (a) Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot showing seeding density impact on the distribution of batch experiment scores. High seeding density batches are colored in orange while green indicates low seeding density batches, (b) Principal component 3 (PC3) versus principal component 4 (PC4) scatter plot showing high and low titer batches spread between both seeding densities. Ellipses represent 95% confidence interval. Left ellipse indicates low yield productions while right ellipse relates to high yield runs.

TABLE 3 Endpoint product titers and viabilities.

	Product titer, mg/L	Viability, %
High seeding density ($n = 8$)	1010 ± 216.2	91.19 ± 3.42
Low seeding density ($n = 6$)	937 ± 141.4	94.7 ± 1.65

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

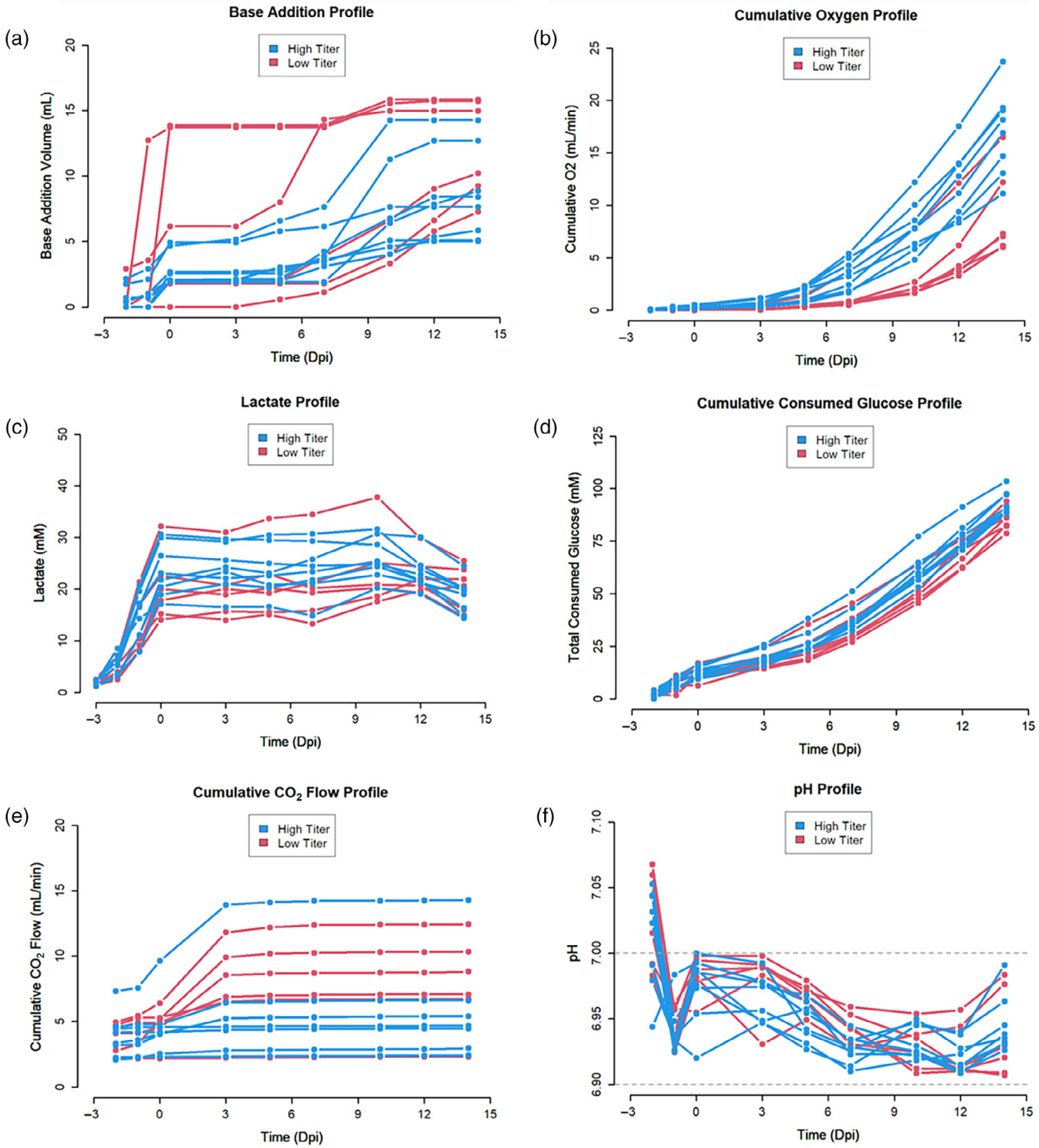

FIGURE 3 Color coded time profiles for the Wuhan pool experiments ($n = 14$) performed in the Multifors 0.75 L system. (a) Base addition volume, (b) cumulative oxygen sparging, (c) lactate profile, (d) total consumed glucose, (e) cumulative carbon dioxide sparging, and (f) daily pH profile.

It is clear that high performing batches (blue) required generally lower base addition (Figure 3a) and higher oxygen sparging (Figure 3b) compared to low performing batches. It is unclear why two low yield batches displayed elevated base volume at the beginning of culture. The large sudden addition of base for these two cultures may have been caused by an error in priming the base lines for proper pH control at the culture start. This, in turn, could lead to an overly aggressive response from the proportional integral derivative controller. It was estimated that the total volume inside the base line is only 2 mL and thus total base volume uncertainty is near 12%. This low amount of

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

base volume addition uncertainty was not considered to be impactful enough to merit excluding the two batch runs from further modeling. When evaluating the corresponding lactate profiles (Figure 3c), it is hard to conclude that lactate metabolism alone is enough to explain the variation in protein expression even though high titer cultures displayed less lactate production variability. Figure 3d shows that high performers had higher glucose consumption. It can be postulated that high glucose demand relates to higher levels of protein expression through tricarboxylic acid cycle (TCA) cellular respiration activity as evidenced by the increased oxygen requirements (Figure 3b). Higher overall CO_2 sparging rates were observed with low performing batches (Figure 3e). All the cultures were performed with pH setpoint ± deadband of 7.05 ± 0.05 (covered range: 7.0–7.1) until −1 dpi and pH was downshifted to 6.95 ± 0.05 (range: 6.9–7.0) from −1dpi to 14 dpi. When looking at the daily average pH profile (Figure 3f), it can be discerned that the worst performing cultures had the largest deviations within the deadband, with values near the 6.9 or the 7.0 threshold. Increasing the pH deadband to cover a larger pH range (e.g., 6.8–7.2) could potentially diminish addition of carbon dioxide or base thus eventually improve process robustness.

3.2 | Impact of bioreactor culture system

The Wu pool in the Multifors 0.75 L system was compared to a dataset resulted from the BioFlo 1 L system (Figure 4). The BioFlo 1 L dataset centered around exploring impact of aeration conditions. The main goal was to find the appropriate aeration strategies that would re-create the results observed in the established Multifors 0.75 L process as part of a technology transfer project.

Figure 4 shows that the negative PC1 axis is strongly influenced by high levels of total consumed glucose (TCG), base addition, and lactate accumulation. Alternatively, the positive PC2 axis is driven by increased IVCC, oxygen sparging, and ammonia. For example, when comparing the left most culture with the right most culture, it is found that the left most culture shows a 2.54-fold increase in cumulative glucose consumption before temperature shift, large lactate accumulation (2.52-fold increase in endpoint lactate), 5.2-fold increase in base addition, and initial large cell growth (2.69-fold increase in IVCC before temperature shift). Conversely, the positive PC1 axis is influenced strongly by high CO_2 sparging, high growth phase pH, and high viability specifically towards the end of the batch (24% higher endpoint viability for the culture in the bottom right when compared to the culture in the bottom left). Importantly, the far left-bottom batch which was a dual sparger culture had a base addition of 42.8 mL which was 2.8-fold higher than the average base addition in the BioFlo 1 L dataset (14 batches). Interestingly, ammonia accumulation in this left-bottom batch was lower (4.0 mM) when compared to the average ammonia accumulation of the BioFlo 1 L dataset (5.5 mM). This points at the idea that batches towards the right side of the graph had less lactate accumulation and better longevity. This viability dependence explains why the negative PC1 axis is driven by large initial cell growth and high glycolysis/lactate metabolism as the cultures are unable to sustain the large biomass increase and are followed by a premature decrease in viability. It is clear that the BioFlo 1 L system has a wider spread in the score plot when compared to the Multifors 0.75 L system (evidenced by the spread in the confidence intervals [CIs] of each reactor system). This makes sense given that the BioFlo 1 L dataset focused on testing a variety of sparging and aeration strategies whereas the Multifors 0.75 L dataset studied the impact of

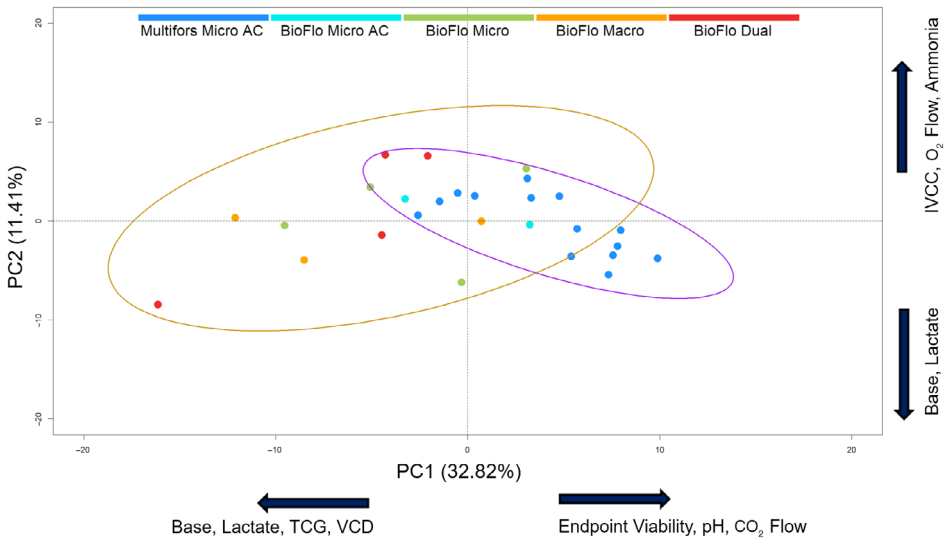

FIGURE 4 Principal component (PC) scatter plots of Wu pool in two bioreactor systems (Multifors 0.75 L and BioFlo 1 L). Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot showing the same pool (Wuhan) with different aeration strategies. Orange batches (BioFlo Macro) employ macro-sparger in BioFlo 1 L, green batches (BioFlo Micro) use micro-sparger in BioFlo 1 L, red batches (BioFlo Dual) utilize dual sparger configuration in BioFlo 1 L, blue batches (Multifors Micro AC [Air Cap]) employ micro-sparger with AC for the Multifors 0.75 L system, and ocean blue batches (BioFlo Micro AC) deploy air caped micro-sparger in the BioFlo 1 L system. Brown ellipse represents the BioFlo 1 L system (13 batches), and purple is assigned to the Multifors 0.75 L system (14 batches). Ellipses represent 95% confidence interval. TCG, total consumed glucose; VCD, viable cell density.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

seeding densities, MSX concentration, and cell passage number under the same hydrodynamics conditions in regard to mixing (volume, agitation, and aeration). Since cell passage number's effect was determined to not be an important factor for the Wu pool nor MSX increased concentration after induction, seeding density is the main factor driving the variation of the overall spread for the Wu pool cluster. Importantly, this implies that aeration strategies have a strong impact on process outcomes given that the batch spread in the scatter plot from BioFlo 1 L cultures was larger when compared to the Multifors 0.75 L batch distribution extent.

From Figure 4, it is also worth noting that the five Multifors 0.75 L cultures that do not overlap (and a sixth batch that exists at the boundary) within the 95% CI ellipse of the BioFlo 1 L system are low density cultures. These 6 batches are the only low-density cultures of the Multifors 0.75 L dataset. Since seeding density was observed to have an important impact on culture outcome (Figure 2a), these six cultures are shown different from the rest of high-density batches. Interestingly, within the 95% CI ellipse of the Multifors 0.75 L cluster, seven batches from the BioFlo 1 L system are found. These cultures use a micro-sparger to provide oxygen to the cells. This may suggest that sparging oxygen with micro-sparger best recreates the hydrodynamics environment of the Multifors 0.75 L cultures; the latter use uniquely a micro-sparger. In regard to aeration strategy (with or without AC), large variation was found in non-air caped batches, both with micro- and macro-sparger. Dual sparger also shows variability that was mostly driven by fluctuations in lactate metabolism and its impacts (base addition). It can be postulated from the process development data that multivariate tools can be deployed to help pick aeration strategies that diminish variability and translate process conditions across bioreactors to facilitate scale-up.

Scale-up from the small-scale bioreactors (Multifors 0.75 L and BioFlo 1 L) to a BioFlo 10 L system was next explored as shown in Figure 5. From Figure 5a, it can be observed that the spread of the BioFlo 10 L system is significantly higher than the BioFlo 1 L or the Multifors 0.75 L. This is due to the fact that data was collected under varying experimental conditions such as differences in DO conditions (40% vs. 60%), sparger (macro, dual, micro), agitation speed, AC, initial volume (5 L vs. 7.5 L), and impeller configuration (1 impeller

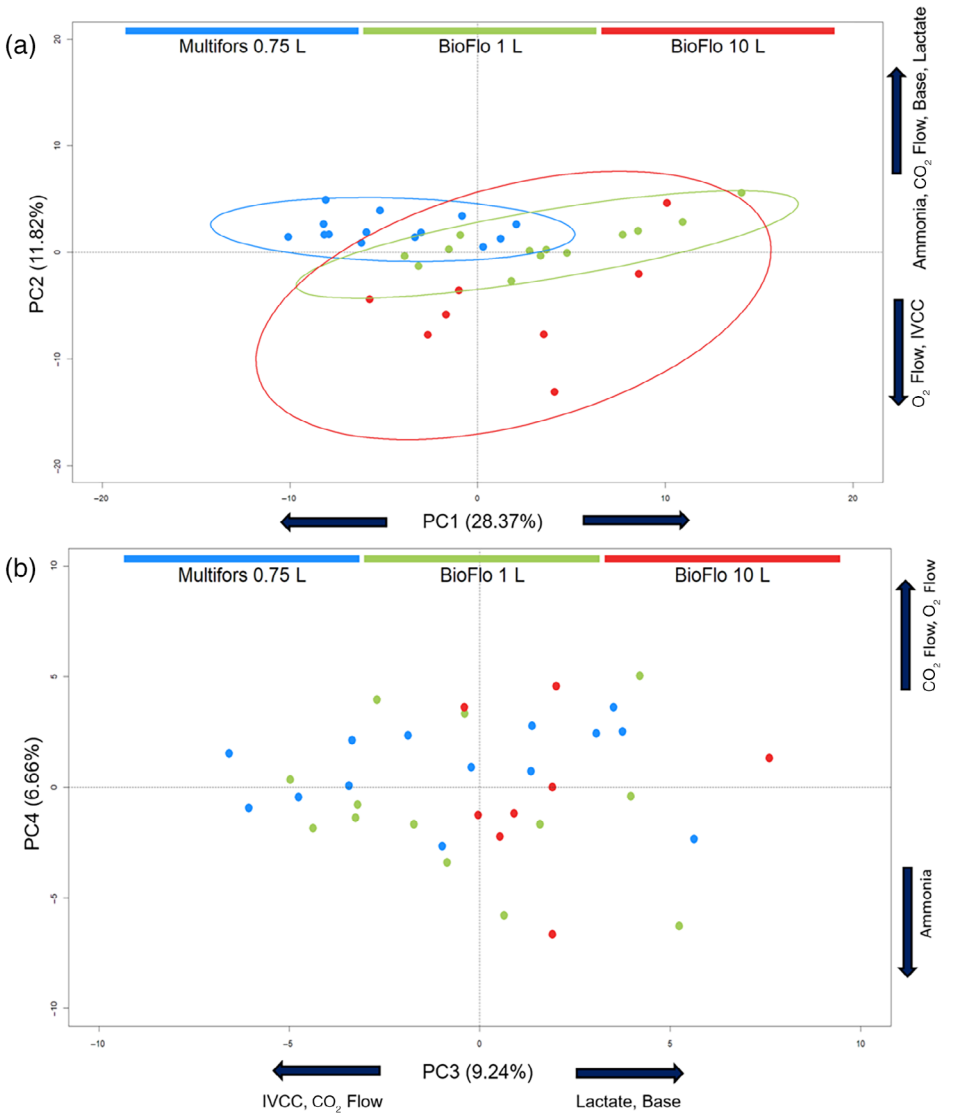

FIGURE 5 Principal component (PC) scatter plots of three bioreactor systems (Multifors 0.75 L, BioFlo 1 L, and BioFlo 10 L). (a) Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot and 95% confidence ellipses of Multifors 0.75 L, BioFlo 1 L, and BioFlo 10 L. Ellipses represent 95% confidence interval, (b) principal component 3 (PC3) versus principal component 4 (PC4) scatter plot of Multifors 0.75 L, BioFlo 1 L and BioFlo 10 L. Blue, green, red batches indicate the Multifors 0.75 L, BioFlo 1 L, and BioFlo 10 L, respectively. IVCC, integral viable cell concentration.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

vs. 2 impellers) as described in Table 1. The PC1 axis is driven by IVCC, lactate, TCG, and base addition in the positive direction while carbon dioxide and viability are the main driving factors in the negative direction. The low-density Multifors 0.75 L cultures are centered in the left most negative axis whereas the high-density cultures are near the origin. BioFlo 1 L and BioFlo 10 L have equal seeding densities but varying hydrodynamics conditions. This difference in spread emphasizes the idea that the impact of hydrodynamics is much more important than other initial variables such as seeding densities (the spread of the Multifors 0.75 L dataset is driven by variance in seeding density, Figure 2). The negative PC2 axis is driven by oxygen sparging and IVCC while the positive PC2 axis is strongly driven by ammonia, carbon dioxide, base addition, and lactate accumulation. There was a strong difference in sparging strategies between the 10 L system (different sparger types, different ACs and different DO set points were utilized) and the small-scale systems (BioFlo 1 L and Multifors 0.75 L) that can explain the differences between the 10 L cluster and the BioFlo 1 L and Multifors 0.75 L clusters along the negative PC2 dimension despite gas flows being normalized with respect to reactor volume. Four BioFlo 1 L batches fall within the 95% CI of the Multifors 0.75 L data which again underscores the idea that batches with similar hydrodynamics (micro-sparger) across different bioreactors can be evaluated with MVDA. When examining the PC3 versus PC4 scatter plot (Figure 5b), one outlier (right-most batch) from the 10 L system is evident. The positive PC3 axis is driven by increased lactate accumulation and base addition while the negative PC3 axis is strongly driven by IVCC and carbon dioxide sparging. The right-most batch outlier utilized a single impeller with a macro-sparger which probably reduces oxygen transfer capability, and as such, large amounts of oxygen sparging was required while at the same time having issues maintaining its 40% DO setpoint. This suboptimal DO control could have impacted cell metabolism as it produced unusually high amounts of lactate (peak lactate was 98 mM) and consequently, required a lot of base addition (313.6 mL). This unusual behavior concluded in a 390 mg/L titer yield which is below the 732 mg/L average that the 10 L bioreactors had. In the literature, similar adverse behaviors induced by inadequate DO controls have been reported.[37]

3.3 | Cell age effect

As it was mentioned above, the expression stability of pools needs to be evaluated to ensure a commercial-scale production which requires an expanded seed train. Three pools stability was evaluated in the Multifors 0.75 L system as part of an early process development objective (Figure 6). For the Wuhan pool, final spike protein yield remains high even at increasing passage number (P5 = 11 generations: 908 ± 161.91 mg/L [$n = 4$]; P8 = 20 generations: 1252 ± 158.39 mg/L [$n = 2$]; P11 = 31 generations: 971 ± 250.31 mg/L [$n = 2$]). Two-tailed t-tests (Table S1) for the Wuhan pool comparing passage number impact show no significant difference between the three cell passage numbers (p-value <0.05). The Beta pool's final endpoint protein yield displays however a gradient behavior such that P5 (455 ± 91.21 mg/L, $n = 2$) > P8 (370 ± 18.66 mg/L, $n = 2$) > P11 (341 ± 7.41 mg/L, $n = 2$). Maximum lactate was observed to be different such that P5 had a higher peak lactate (50.15 mM) when compared to P8 (33.9 mM) and P11 (32.3 mM). Endpoint ammonia was higher in P5 (8.28 mM) compared to P8 (5.97 mM) and P11 (6.12 mM). Endpoint IVCC also demonstrated cell age variance such that P5 (1.14×10^8 cell*day/mL) < P8 (1.56×10^8 cell*day/mL) < P11 (1.69×10^8 cell*day/mL). Interestingly, when detailing total oxygen sparged, it is clear that the cumulative average flow rates follow the IVCC trend such that P5 (18 mL/min) < P8 (26.58 mL/min) < P11 (27.27 mL/min). On the other hand, even WuTL endpoint protein production did not seem to be negatively impacted (P5 = 838.5 ± 48.8 mg/L [$n = 2$], P8 = 751.5 ± 60.1 mg/L [$n = 2$], P11 = 908 ± 106.06 mg/L [$n = 2$]), an inverse relationship between base addition and oxygen sparging was found. Passages 5 and 11 which had the higher protein production also had higher oxygen sparging and less base addition when compared to

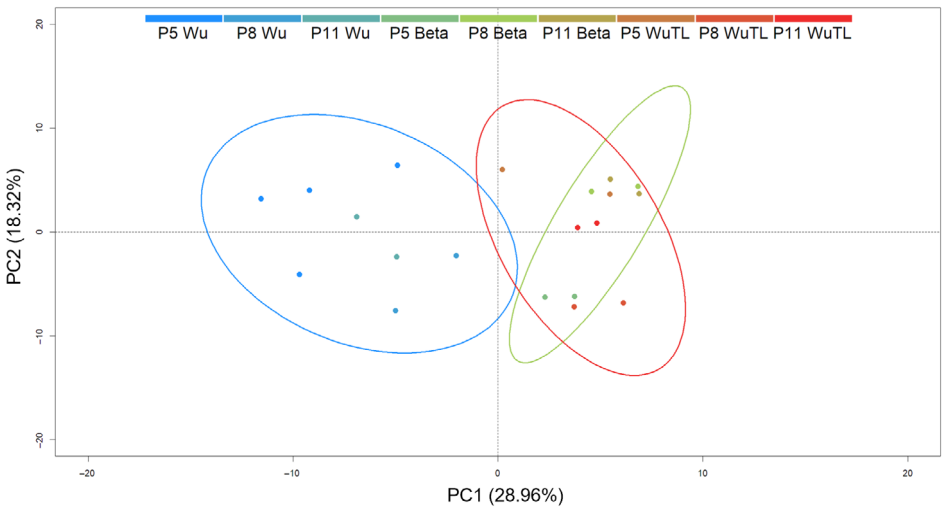

FIGURE 6 Scatter plot of principal component 1 (PC1) versus principal component 2 (PC2) of different cell passage numbers used for productions of three pool variants in Multifors 0.75 L bioreactor. Batches are color-coded based on their passage number and pool type (Wu, Beta, WuTL). Ellipses represent 95% confidence interval.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

passage 8. A similar conclusion is reached when evaluating above-average batches and below-average batches in terms of final yield. The above-average (832 mg/L) batches required more oxygen sparging and needed less base addition. The difference was driven by increased lactate accumulation in the low performing batches. The best performing batches also had on average higher total glucose consumption demonstrating that the high metabolic activity was shared across glycolysis and other metabolic pathways that do not end in lactate accumulation.

When evaluating two-tailed *t*-tests (Table S2) on key metrics, it can be said that although the passage number did not have statistical impact on final titer concentration, there was passage number related variation with respect to total glucose consumption for the WuTL pool. Similarly, although passage-related statistical significance was found in endpoint lactate and endpoint IVCC, no evidence for statistical impact on endpoint titers was determined in the Beta pools. In stark contrast, when evaluating cell age impact on the Delta pool (Figure 7), passage number seems to have an impact on the spread of the scores such that higher passage number batches have more spread. Interestingly, it was determined from the loadings (Figure S13) that oxygen sparging and base addition are inversely related, while lactate and base are directly correlated. This is because spread in the PC1 axis is driven by base and lactate in the negative direction while glucose consumption, IVCC, and oxygen flow increase in the positive direction. Additionally, the positive PC2 axis is mostly driven by carbon dioxide sparging whereas the negative direction is driven by base addition. Since the scores did demonstrate a cell passage number's dependence, the average protein concentration for each passage was calculated. Titer of 600 ± 112 mg/L, 361 ± 31 mg/L, and 398 ± 92 mg/L were estimated for P5 ($n = 4$), P8 ($n = 4$), and P11 ($n = 4$), respectively. Additionally, endpoint ammonia showed cell age related behavior since P5 (6.88 mM) had significantly lower ammonia accumulation when compared to P8 (8.79 mM) and P11 (10.82 mM). Table 4 shows a two-tailed t-tests comparing passage number impact. Only endpoint titers are statistically different. Taken together, it can be concluded that culture age does indeed play a role in culture outcomes for the Delta pool. This pool is likely to become unstable over time, so considerations will be necessary when scaling up, such as limiting the cell passage to five in order to preserve high titers.

3.4 | MSX concentration impact

The studied CHO pools in this work express Glutamine Synthetase (GS) gene implying MSX is required during the cell line generation process. Tian[38] showed that increasing the MSX concentration to certain level can lead to an improved overall protein yield. In this study, MSX supplementation up to 125 μM was investigated across the Wu, Delta, WuTL pools to determine if the increased selection pressure at the moment of induction has a positive impact on final protein yield. As it can be seen from Figure 8, no discernible clustering is evident when analyzed based on MSX concentration. This suggests that extra

TABLE 4 Two-tailed *t*-test of cell age's impact on different key variables for Delta pool.

Delta pool	P5 versus P8	P5 versus P11
Base volume	0.37	0.17
Endpoint lactate	0.83	0.33
Total oxygen sparging	0.11	0.82
Endpoint titer	0.01*	0.03*
Total glucose consumption	0.66	0.63
Endpoint IVCC	0.48	0.56
Max lactate	0.53	0.27
Endpoint ammonia	0.24	0.06

Note: Values with asterisk * represent conditions in which statistical significance was found (*p*-value <0.05).

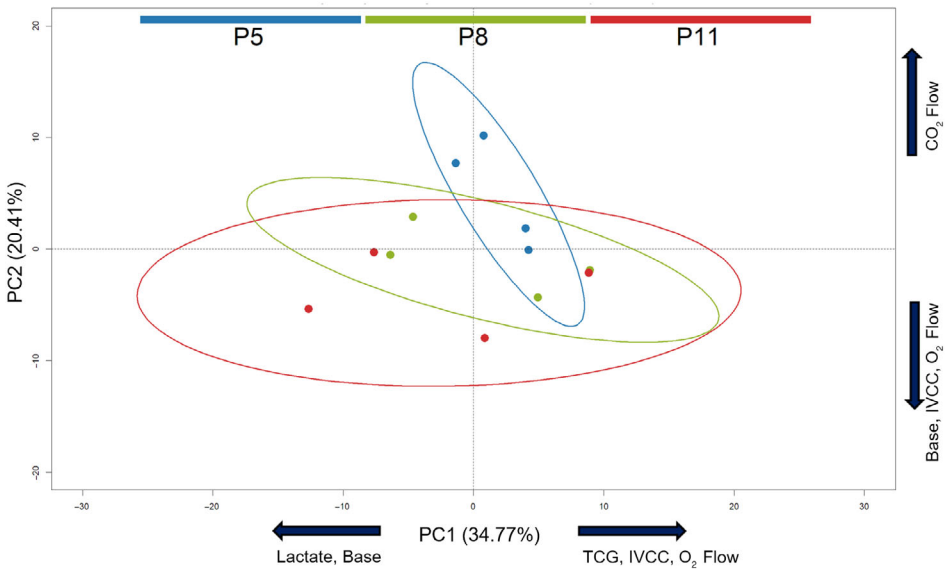

FIGURE 7 Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot showing cell age impact on Delta pool. The blue, green, and red color represent batches performed with cell passage number P5, P8, and P11, respectively. Ellipses represent 95% Confidence Interval.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

addition of 75 μM MSX during induction (Day 3 post-seeding) does not provide a clear impact. The visual clustering that occurs in the scores plot as evidenced by the 95% confidence ellipses is driven by differences in pool behavior. To further determine the impact of extra 75 μM MSX addition, two-tailed *t*-tests (Table S3) were realized. No statistical differences were found in almost all the key variables between the base level MSX+ (50 μM MSX) and extra MSX ++ (125 μM MSX) conditions at the exception of endpoint ammonia ($p = 0.01 < 0.05$). The Delta cluster has the most spread profile probably due to the fact that this specific pool seemed to be significantly impacted with cell increasing age (Figure 7). Given that the Delta pool dataset had enough data points, an additional two-tailed *t*-test (Table S4) was conducted. It was determined that even if passage number is accounted for, MSX has no statistical impact on general cell culture variables except for the passage 11 batches; 35 mL of base was added with 50 μM MSX productions (MSX+) while only 8.9 mL of base was needed for 125 μM MSX bioreactors (MSX++). Taken together, it can be postulated that additional MSX supplementation at induction has no added value and thus can be avoided in the future when working with such pools.

3.5 | Global pool analysis

Once every pool was analyzed separately, the four pools were then examined together to better understand pool related clustering and behavior that may not be self-evident when investigating individually. From Figure 9, it is possible to discern that the pools tested in the

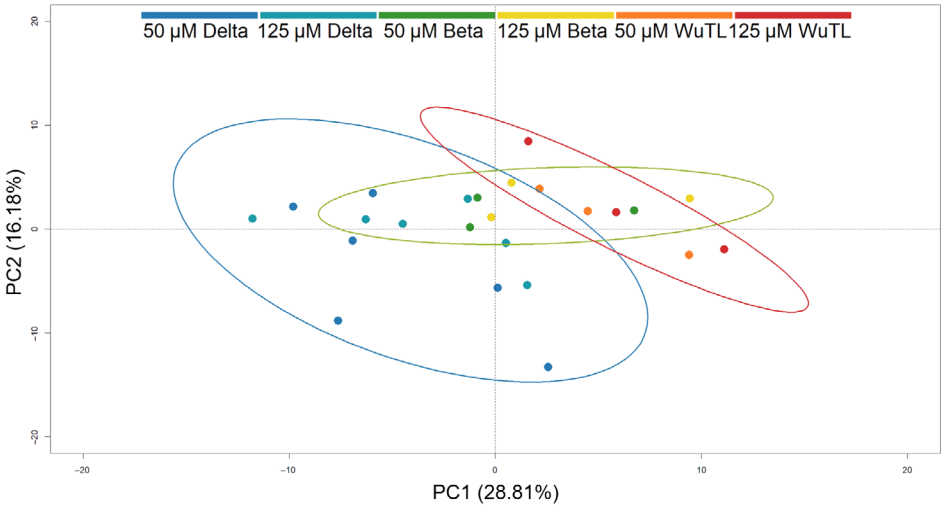

FIGURE 8 Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot showing the impact of methionine sulfoximine (MSX) addition during induction on various pools. Batches are colored with respect to pool and MSX addition. Blue and ocean green batches represent Delta pools with no extra MSX addition (50 μM Delta) or with addition of 75 μM MSX at induction (125 μM Delta in total), respectively. For Beta pools, green color represents the productions without extra MSX addition at induction (50 μM Beta) while yellow batches are assigned to Beta pools with 125 μM MSX. Wuhan Tag-less (WuTL) productions were conducted without extra MSX addition at induction (50 μM WuTL, orange color) compared to red batches with 125 μM MSX. Colorful ellipses show 95% Confidence Interval for respective studied conditions.

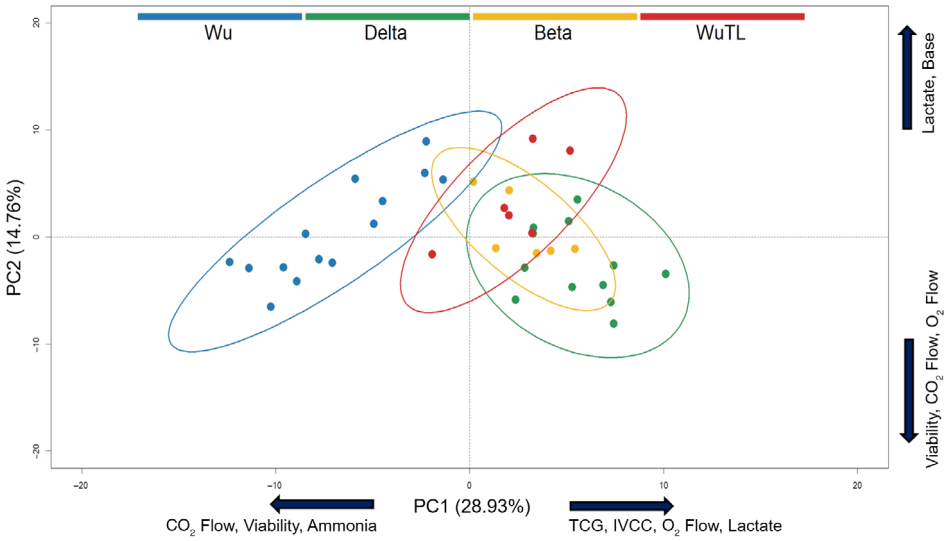

FIGURE 9 Principal component 1 (PC1) versus principal component 2 (PC2) scatter plot of all the Multifors 0.75 L experiments encompassing 4 pools (37 batches). Wu pool colored in blue, Delta pool colored in green, Beta pool colored in yellow, and WuTL pool colored in red. Colorful ellipses represent 95% Confidence Interval (CI) for each respective pool.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Multifors 0.75 L system exhibit distinct behavior as evident by the extent in the distribution and spread variation in the PC1 versus PC2 graph. The Wuhan variant pool has a large spread in the PC1 axis that is driven by its high density versus low density seeding batches. This is evident from the loading plots (Figure S17) in which it is clear that the driving factors are IVCC, lactate, total glucose consumption, and oxygen sparging requirements. It is worth mentioning that the Wu pool dataset is split into high-density and low-density seeding batches. Such a segmentation based on biomass is captured. On average, Delta pools had 1.7-fold higher IVCC when compared to the Wu pool which again explains how the Delta pool is centered in the positive PC1 axis while the Wu pool is spread out in the negative direction. This further demonstrates that every pool had different growth patterns and, in consequence, had different specific protein production characteristics. Delta pool had the most overall growth but provided the second poorest endpoint titer. The PC2 axis is heavily driven by lactate and base addition in the positive direction while oxygen sparging, carbon dioxide sparging, and viability are in the negative direction. Even though most pools demonstrated an inverse relationship between base addition and oxygen sparging individually, it is not possible to carry out this analysis across pools. However, between batches within the same pool, analyzing these critical attributes does seem to hold as a predictor.

TABLE 5 Variables used for protein prediction modeling.

Variables	Indicator
Endpoint viability	Batch longevity
Peak lactate accumulation	Maximum glycolysis activity
Endpoint lactate concentration	Lactate consumption or production
Residual glucose	Metabolic activity
IVCC	Accumulative cell biomass
Endpoint ammonia concentration	Metabolic activity for glutamine synthesis and waste accumulation through amino acid deamination
Total glucose consumed	Metabolic activity
Total base added	Lactate build-up
Total average oxygen sparged per day	Cellular respiration
Total average carbon dioxide sparged per day	pH controlling profile for upper bound of pH deadband
Endpoint pH	Secondary indicator for lactate consumption (low pH indicates lactate accumulation while high pH relates to lactate consumption status)
Cell passage number	Pool stability
Pool variant	Indicator of product specific nature

3.6 | Endpoint titer modeling

Since many of the process conditions and measured variables were observed to have impact on culture behavior and a close relationship to final titer, the possibility of creating four ML models based on widely used linear (PLS) and nonlinear regressors (RF, SVM, Extreme Gradient Boosting [XGB]) was explored. In order to build regression models, key variables were utilized. Table 5 summarizes the important variables used to predict batches not utilized during the training process. Endpoint viability can be considered an indicator of batch longevity whereas endpoint IVCC can be an indicator of accumulated biomass which has been observed to be a strong determinant in terms of total protein production given the strong link between the two variables (Protein yield = qP * IVCC). Endpoint residual glucose is a relevant parameter as it gauges metabolic activity at the end of the culture. Low endpoint viability indicates that the culture suffered a culture decline and thus was probably impacted in terms of its capacity to remaining metabolically active enough to produce protein in the end stage of the process. Cumulative sparged oxygen can also be understood as proxy parameter for metabolic activity. Given the important relationship between oxygen requirement and TCA cycle activity, it is coherent that oxygen sparging should be included in the model.[36] Peak lactate can be interpreted as a proxy measurement for maximum glycolytic activity while endpoint lactate can be understood as an estimation of lactate absorption which has been observed to be a good process indicator in CHO cell culture processes.[25] It can be observed as well that different pools have different importance on the protein yields. Endpoint ammonia should also be included within the model as high ammonia accumulation cultures maybe negatively impact cell culture longevity and consequently productivity. TCG can be accepted as a proxy for overall metabolic activity (glucose can be consumed through glycolysis to yield lactate or it can be transformed to pyruvate to link with the Krebs [TCA] cycle; Figure 10) and as such, it should also be represented within the model. As the Delta pool was observed to have a clear impact on protein production with increasing passage number, this information was also included in the modeling process. Total base addition can be understood as an indirect measurement of total lactate build-up and an indicator of pH acidic profile. Total carbon dioxide sparged can be reasoned as a pH control indicator that contains information about the pH upper deadband and also indicates whether cells switch to lactate consumption phase. Lastly, endpoint pH which can be interpreted as a clear indicator of lactate consumption was also included within the model development process.

A dataset comprising 50 batches, encompassing both Multifors 0.75 L and BioFlo 1 L experimental runs, was split into training (84%) and test (16%) sub datasets. The BioFlo 10 L cultures were excluded from the modeling phase on the basis of high experimental variability without replicates (varying impeller configuration, varying sparger type, varying sparging strategy, varying DO set point). Furthermore, one culture from the BioFlo 1 L dataset was excluded from the modeling phase on the basis that respiratory tests were realized throughout the culture, thus potentially impacting the online values that are utilized as features in the model. From the considered data, a split was chosen randomly with the condition that it contains a high/low production batch performance. The features were regressed in function

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

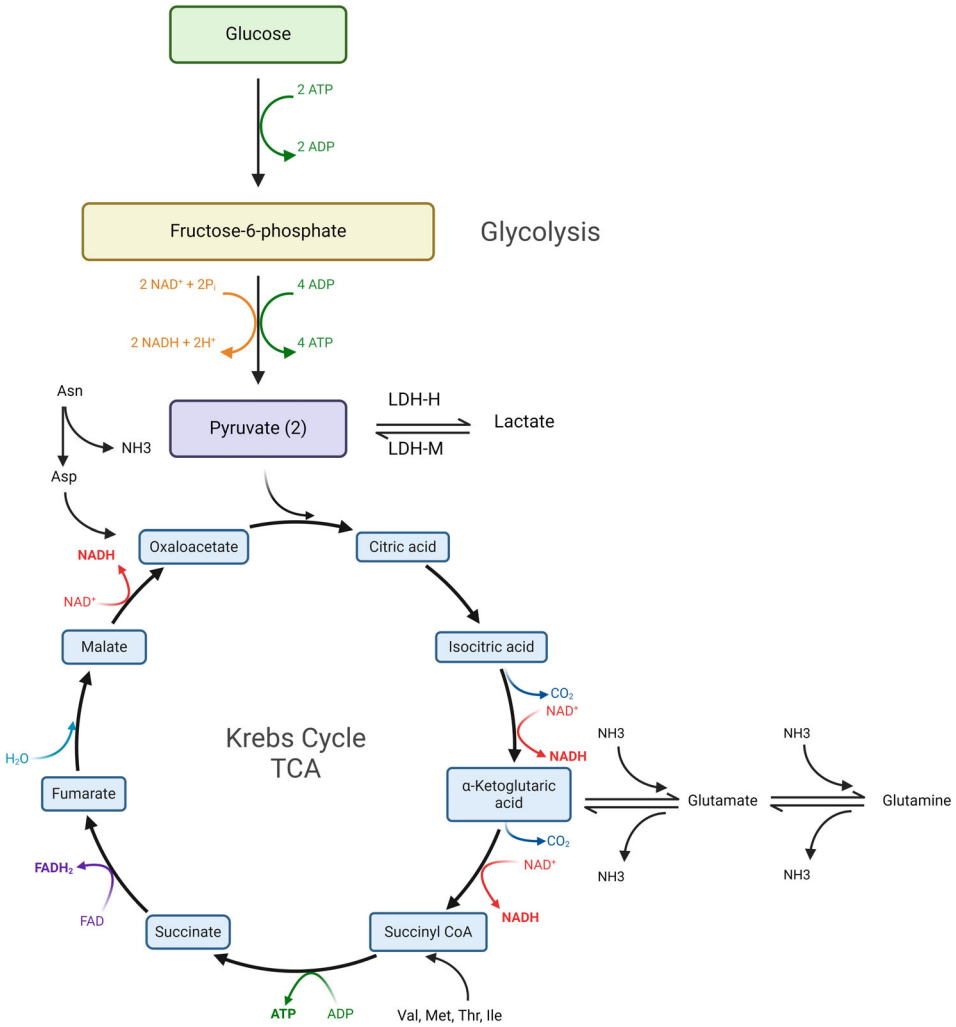

FIGURE 10 Schematic of the glycolysis pathway and TCA cycle. 2 ATPs are formed in the glycolytic pathway while 36 ATPs are generated through the oxidative phosphorylation pathway.

TABLE 6 Mean and confidence intervals for training and test results for respective metric after bootstrapping.

Training dataset	Mean RMSE	Mean MAE	Mean R^2	95% CI RMSE	95% CI MAE	95% CI R^2
SVM	95.92	66.93	0.88	61.56–125.98	43.84–82.58	0.82–0.96
RF	62.15	49.13	0.96	47.10–75.27	33.98–59.80	0.94–0.98
XGB	45.48	28.36	0.98	27.00–69.41	17.33–38,45	0.96–0.98
PLS	217.97	184.88	0.41	189.20–260.70	151.20–219.00	0.26–0.72
Test dataset	**Mean RMSE**	**Mean MAE**	**Mean R^2**	**95% CI RMSE**	**95% CI MAE**	**95% CI R^2**
SVM	80.35	74.33	0.92	61.80–97.42	52.39–89.30	0.84–0.98
RF	65.96	58.15	0.94	51.62–88.24	41.44–80.20	0.90–1.00
XGB	99.51	89.71	0.83	73.17–122.90	52.56–111.02	0.73–1.00
PLS	158.36	120.00	0.72	115.00–260.30	59.70–199.70	0.50–0.98

of endpoint titer to generate a model capable of predicting final yield given key process outcomes. As to give a fair chance to each model, the same strategy for tunning the hyperparameters (parameters used to control the learning process in ML). Here, adaptive resampling of the tuning parameter grid was realized in such a way that the random search of hyperparameters is concentrated on values that are in the neighborhood of the optimal parameters. This is done by discarding settings that are clearly sub-optimal. This approach has been observed to reduce training time.[39]

As it can be seen from Table 6 (bootstrapping results), the RF model was able to outperform SVM and PLS models in all the tested metrics (RMSE, MAE, and R^2), for the test and training datasets.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

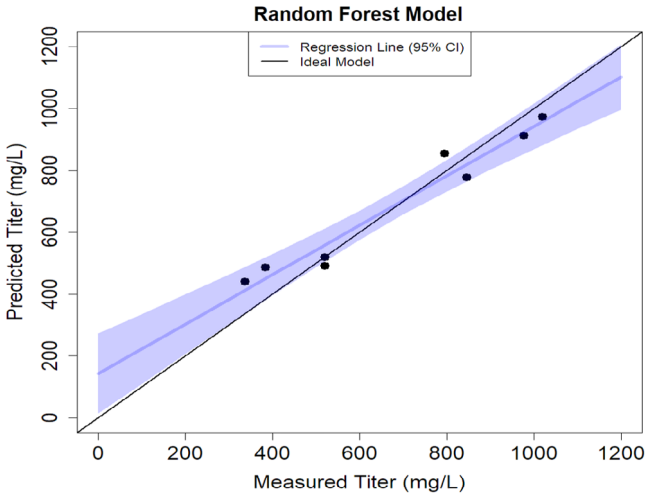

FIGURE 11 Measured versus predicted titer scatter plot. $R^2 = 1$ line in the diagonal indicates ideal model and blue line shows regression line with 95% confidence interval.

Overall metrics of the RF model performed similar in the training and test datasets suggesting that the model was generalizable across the dataset, this may be due to the fact that RF algorithms are known to be robust to outliers and noise within datasets.[40] Importantly, XGB outperforms RF in the training dataset but performs worse-off in the test dataset possibly indicating a lack of generalization with the available data for this particular model. When detailing the 95% CIs, it can be observed that SVM and RF are statistically different than PLS in terms of RMSE metric. Additionally, the RF model has the narrowest intervals (training and test for RMSE, MAE, and R^2), when compared to the other three models.

As it can be observed from Figure 11, good predictive capability is attained with the RF model, not only is the R^2 value high (0.95) but the majority (6/8) of the resulting predictions fall along the $R^2 = 1$ line which represents an ideal model of perfect prediction. When taking into account the CIs of the regression line, we can see that for the total span of the data, the 95% CIs contain the ideal model suggesting that despite the low data amount to test the model, predictions are statistically in line with an ideal model. Interestingly, outside the span of the data, CI widens and begins to stop overlapping with the ideal model specifically within the 0–300 mg/L range. This is to be expected as there are no batches spanning this range thus no model prediction falling within this zone. Consequently, extrapolation of a linear model (regression line) outside the training and test range of the RF model is not an appropriate indicator for performance along these ranges.

Recent approaches have centered around improving the interpretability of ML models.[41–43] This is especially important for industries where process understanding is a key requirement for regulatory approval, as is the case in the biotherapeutic industry. SHapley Additive exPlanation (SHAP) was developed using the idea of the Shapley value which is a notion in game theory that helps determine fair profit allocation to various stakeholders by evaluating their respective contribution to the outcome.[44] In the context of ML, each stakeholder can be understood as a feature and the payout is the outcome of the model itself. In summary, the Shapley value for each feature represents each feature contribution to the model's prediction of a particular datapoint. This is estimated by calculating the average marginal contribution of a feature considering every possible combination. Consistent results have been shown with SHAP values, and SHAP dependency plots offer a helpful model summary.[7,42,45] From the generated dependency plots, various conclusions can be discerned. Features with positive magnitude SHAP values have a positive impact on the prediction, while negative magnitude values represent a negative impact. The bigger the magnitude of the SHAP values, the stronger the effect.

In Figure 12a, there seems to exist an optimal IVCC value for which good protein production is obtained. When coloring based on cell pools, the resulting behavior seems to be cell pool dependent given that the Beta and Delta pools in general reached higher total IVCC but also had less cell specific protein productivity (qP). For the Wu pool, two clusters (one below 0 in the Y axis and another above 0 reaching 35) are formed. These two clusters correspond to high and low seeding densities, respectively. It can be concluded that for the two Wu pools (and presumably WuTL since the two high density clusters overlap for both pools), higher seeding density was concomitant with better protein yields. It may also suggest that the current feed regimen developed is unable to sustain higher cell densities and since Beta and Delta pools exhibited large biomass growth, an optimization of the feeding regimen could allow these pools to reach protein expression levels comparable to Wu and WuTL. As detailed in Figure 12b, maximum lactate accumulation of 35 mM and beyond has a negative impact on endpoint protein yield. This observation holds across all pools given that the fast decrease in SHAP values is observed for Wu, Beta, and Delta pools. It is paramount to reduce lactate accumulation in a given culture to avoid increased osmolality due to base addition. One simple strategy that could be implemented to control lactate accumulation is to replace bolus feeding with slow continuous feeding rate as it has been observed to diminish metabolic waste build-up by decreasing the variations in nutrient availability which might alter the metabolic behavior of the cell culture run.[46] From Figure 12c, it can be observed that high endpoint viability has a positive impact on final protein production. This impact rapidly turns negative once viability is below 85%. This suggests that increasing culture longevity and thus maintaining metabolic activity is critical. In order to avoid early cell culture decline, appropriate measures must be taken such as lowering osmolality impact (through less base addition by using only sparging gasses for effective pH control), decreasing hydrodynamic stress caused by shear damage and/or optimizing feeding strategies.[47–49] Feeding based on oxygen consumption rates or bio-capacitance measurements may be an attractive starting point to develop on-demand feeding strategies given the strong relationship that these signals have with viable cell volume and consequently, with metabolic activity, given that larger cells have consumed more oxygen.[50–55] Interestingly, when noting the SHAP endpoint pH dependency plot (Figure 12d), it is clear that endpoint pH of 6.93–

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

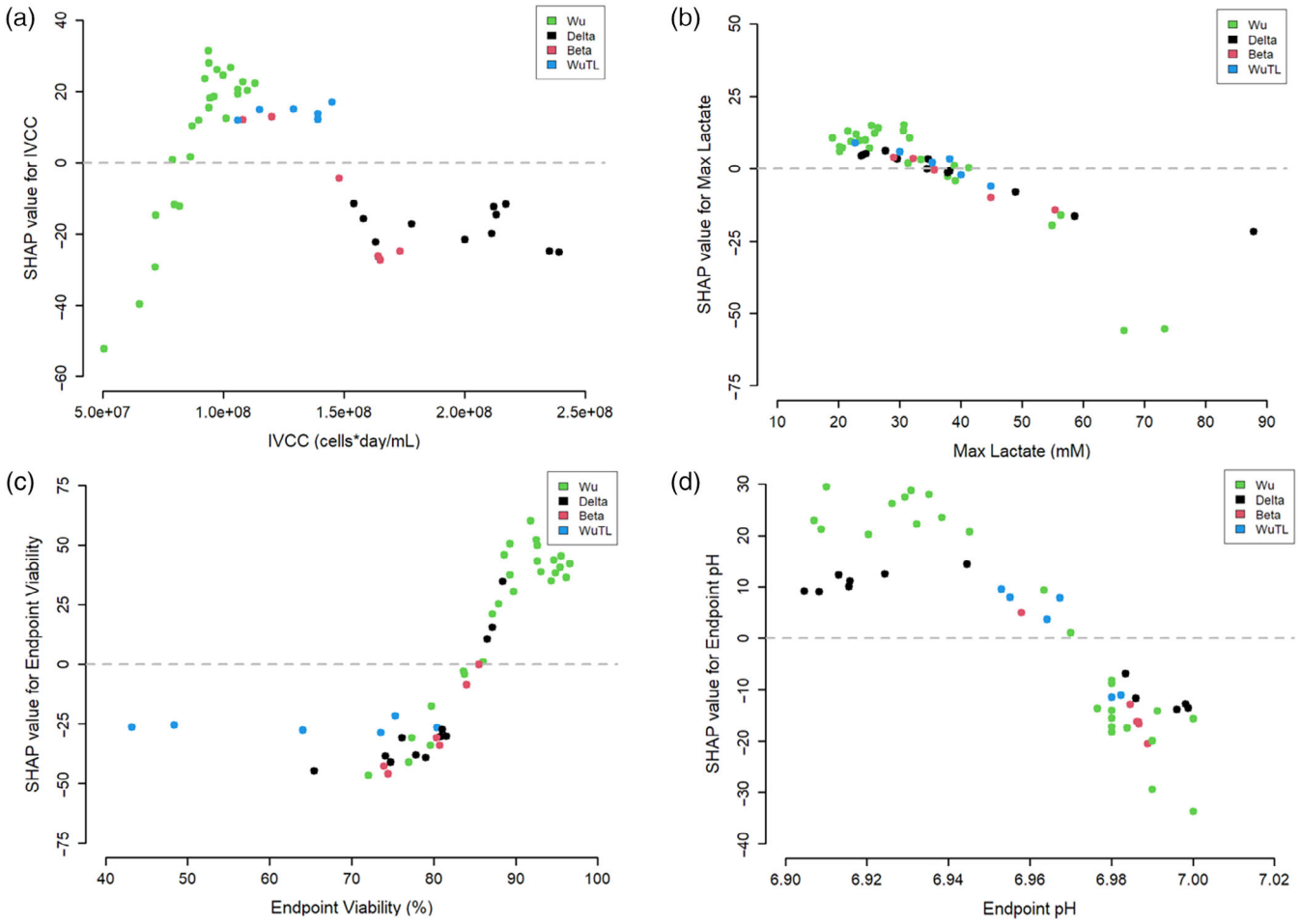

FIGURE 12 Shapley value dependency plot. (a) Integral viable cell concentration (IVCC), (b) max lactate, (c) endpoint viability, (d) endpoint pH, (e) cumulative oxygen sparged, (f) total glucose consumption, (g) endpoint ammonia, (h) endpoint residual glucose.

6.95 had the most positive impact on final titers. It is worth mentioning that all the studied cultures existed within a deadband of ±0.05 between 6.9 and 7. Thus, cultures with endpoint pH of 6.93–6.95 represent processes in which no base addition or carbon dioxide was added in the last day of the fed-batch process. This could imply that unnecessary action upon cultures may be a net negative and thus pH deadband can be increased to ±0.2 around the 7.0 setpoint so as to avoid base addition and carbon dioxide sparging.

From Figure 13a, it is discernible that even if there seems to be a pool dependence in terms of cumulative oxygen sparging (Delta and Beta cluster differently from Wu and WuTL pools), there are signs of an optimal total oxygen sparging that is concomitant with high yields given the overlaps among pools and observing the fact that beyond 0.035 cumulative VVM, no further increase in SHAP values can be detailed. The Wu cluster near the −50 (*Y* axis) represents low density seeding cultures. This demonstrates the close link culture oxygen requirements have with VCD and viable cell volume.[56,57] For the SHAP cumulative glucose consumption dependency plot (Figure 13b), it can be observed that lower glucose consumption has a negative impact on protein production. This could be explained by either lower VCD or lower metabolic activity, both of which directly impact the culture capacity to achieve high titers. Alternatively, very high cumulative glucose consumption also had a negative impact on protein yields. This may be explained by the fact that cultures that consume glucose at high rates tend to have high lactate productions and thus adverse culture outcomes. This again points towards the idea that regulating the glucose intake of cells may be beneficial in terms of avoiding high lactate accumulation. From the endpoint ammonia dependency plot (13C), it is noticeable that ammonia concentrations beyond 8 mM, for the process studied, were generally associated with negative protein production prediction. This makes sense given that ammonia is another relevant by-product that can be a direct result of amino acid metabolism (Figure 10). Lastly, from Figure 13d, it can be observed how residual glucose serves as a proxy indicator of endpoint metabolic activity. Most notably, it can be understood as an inverse of the final viability measurement given that lower residual glucose values represent higher metabolic activity while high residual glucose values demonstrate cell cultures with little glucose metabolism and thus low overall metabolic activity.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

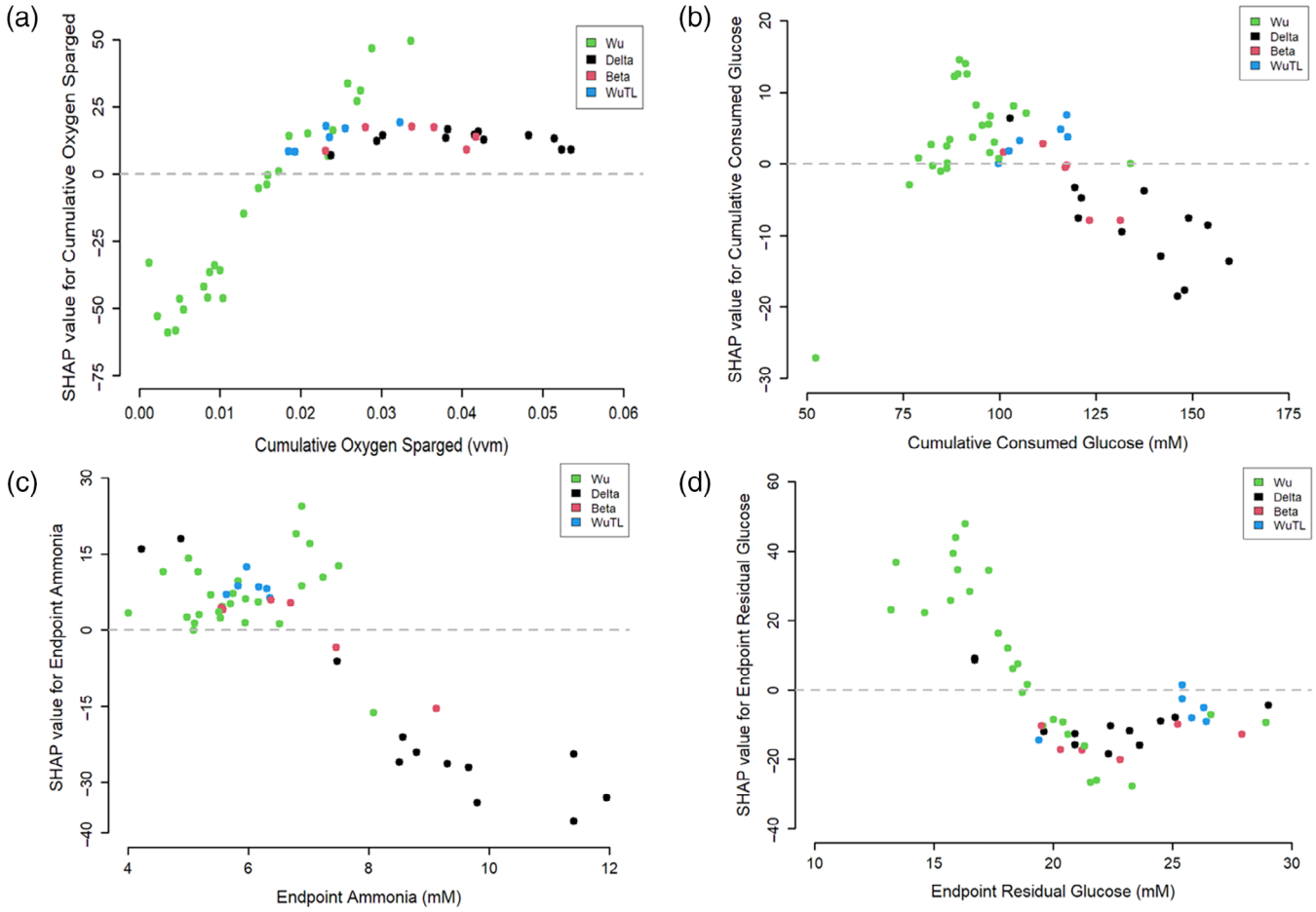

FIGURE 13 Shapley value dependency plot. (a) Cumulative oxygen sparged, (b) total glucose consumption, (c) endpoint ammonia, (d) endpoint residual glucose.

4 | CONCLUSION

MVDA tools can be utilized in early process development to generate wider insights that might otherwise be difficult to conclude through standard univariate analysis. In the context of this article, it was possible to determine that although seeding density constraints the overall cell density available during the production phase, impacts on viability mean that there is a relative parity in terms on protein outcome by the end of the 17-day fed-batch process. Additionally, it was found that oxygen sparge requirements can be used as a process performance indicator to evaluate culture outcomes. It was observed that cultures with high sparge rates and low base addition were concomitant with high endpoint titers. On the other hand, by comparing the same pool across different systems, it was possible to determine which culture conditions reproduced similar behaviors such that representative processes are created. In this case, it was possible to determine which aeration conditions in the BioFlo 1 L system best re-create the hydrodynamics conditions of the Multifors 0.75 L system. Cell passage dependence on different pools was also readily observed through clustering in the PCA plots. This is key as it allows to determine which pools may be viable for larger scale production.

Difference in pool behavior was also evaluated through clustering and careful analysis of the loading plots. It was observed that difference in lactate metabolism and cell growth (and the consequent derivatives such as base added volume, oxygen requirements, and longevity) were the main drivers in differentiating the culture outcomes. It was also demonstrated that a RF model, utilizing these key features, is able to capture the nonlinear relationships between the measured variables and the final protein yield in order to generalize its predictive capabilities. These ML models can then be analyzed through SHAP dependency plots to recognize the interactions and given early process development goals, improve process understanding. To the best of the knowledge of the authors, this is the first time SHAP dependency plots have been applied for the purpose of CHO cell culture process performance analysis. This may prove to be a worthwhile strategy, given that the analysis of early process development datasets served to gain insights that can aid further process optimization. For example, it was concluded that increasing pH deadband may be beneficial so as to limit unnecessary base and carbon dioxide additions. This strategy may be used in tandem with slow continuous feeding to diminish sudden metabolic by-product build-up. Importantly, longevity was also determined to be a relevant factor, and as such, finding strategies that improve said longevity may also be

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

adequate more than just increasing cell density but in detriment of viability. Such strategies could center around diminishing shear force, osmolarity stress, and improved feedings since these processing parameters (high shear, high osmolarity, inadequate feeding) have been observed to be important drivers of cellular apoptosis.[47,58,59]

AUTHOR CONTRIBUTIONS

Olivier Henry: Funding acquisition; writing – review and editing; conceptualization; formal analysis; supervision; resources. **Sebastian-Juan Reyes:** Conceptualization; methodology; data curation; investigation; validation; writing – original draft. **Lucas Lemire:** Investigation; data curation; formal analysis. **Raul-Santiago Molina:** Formal analysis; software. **Marjolaine Roy:** Methodology. **Helene L'Ecuyer-Coelho:** Methodology. **Yuliya Martynova:** Methodology. **Brian Cass:** Methodology. **Robert Voyer:** Resources; supervision. **Yves Durocher:** Supervision; resources; formal analysis. **Phuong Lan Pham:** Conceptualization; methodology; investigation; supervision; formal analysis; validation; funding acquisition; writing – review and editing; resources; project administration.

ACKNOWLEDGMENTS

This work was supported by the Pandemic Response Challenge Program (PRCP) of the National Research Council of Canada (NRC).

FUNDING INFORMATION

This work was funded by the National Research Council of Canada (grant PR-023-1) and by the Natural Sciences and Engineering Research Council of Canada (grant RGPIN/4048-2021 and stipend allocated to Sebastian-Juan Reyes via the NSERC-CREATE PrEEmiuM program).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Olivier Henry https://orcid.org/0000-0003-2106-1331

REFERENCES

1. Hacker DL, Balasubramanian S. Recombinant protein production from stable mammalian cell lines and pools. *Curr Opin Struct Biol*. 2016;38:129-136. doi:10.1016/j.sbi.2016.06.005
2. Stuible M, Gervais C, Lord-Dufour S, et al. Rapid, high-yield production of full-length SARS-CoV-2 spike ectodomain by transient gene expression in CHO cells. *J Biotechnol*. 2021;326:21-27. doi:10.1016/j.jbiotec.2020.12.005
3. Joubert S, Stuible M, Lord-Dufour S, et al. A CHO stable pool production platform for rapid clinical development of trimeric SARS-CoV-2 spike subunit vaccine antigens. *Biotechnol Bioeng*. 2023;120(7):1746-1761.
4. Stuible M, van Lier F, Croughan MS, Durocher Y. Beyond preclinical research: production of CHO-derived biotherapeutics for toxicology and early-phase trials by transient gene expression or stable pools. *Curr Opin Chem Eng*. 2018;22:145-151. doi:10.1016/j.coche.2018.09.010
5. Ye J, Alvin K, Latif H, et al. Rapid protein production using CHO stable transfection pools. *Biotechnol Prog*. 2010;26(5):1431-1437. doi:10.1002/btpr.469
6. Reyes SJ, Durocher Y, Pham PL, Henry O. Modern sensor tools and techniques for monitoring, controlling, and improving cell culture processes. *Processes*. 2022;10(2):189-225. doi:10.3390/pr10020189
7. Molina RS, Molina-Rodríguez MA, Rincón FM, Maldonado JD. Cardiac operative risk in Latin America: a comparison of machine learning models vs EuroSCORE-II. *Ann Thorac Surg*. 2022;113(1):92-99.
8. Alavijeh MK, Baker I, Lee YY, Gras SL. Digitally enabled approaches for the scale up of mammalian cell bioreactors. *Digital Chem Eng*. 2022;4:100040.
9. Irfan K. Carbon Dioxide Control in Bioreactors and the Application of Principal Component Analysis to Cell Culture Process Data. Dissertation. Newcastle University. 2017.
10. Tescione L, Lambropoulos J, Paranandi MR, Makagiansar H, Ryll T. Application of bioreactor design principles and multivariate analysis for development of cell culture scale down models. *Biotechnol Bioeng*. 2015;112(1):84-97.
11. Rathore AS, Mittal S, Pathak M, Mahalingam V. Chemometrics application in biotech processes: assessing comparability across processes and scales. *J Chem Technol Biotechnol*. 2014;89(9):1311-1316. doi:10.1002/jctb.4428
12. Goldrick S, Sandner V, Cheeks M, et al. Multivariate data analysis methodology to solve data challenges related to scale-up model validation and missing data on a micro-bioreactor system. *Biotechnol J*. 2020;15(3):1800684. doi:10.1002/biot.201800684
13. Facco P, Zomer S, Rowland-Jones RC, et al. Using data analytics to accelerate biopharmaceutical process scale-up. *Biochem Eng J*. 2020;164:107791. doi:10.1016/j.bej.2020.107791
14. Powers DN, Trunfio N, Velugula-Yellela SR, Angart P, Faustino A, Agarabi C. Multivariate data analysis of growth medium trends affecting antibody glycosylation. *Biotechnol Prog*. 2020;36(1):e2903.
15. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J*. 1994;40(8):1361-1375. doi:10.1002/aic.690400809
16. Wold S, Kettaneh N, Fridén H, Holmberg A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometric Intell Lab Syst*. 1998;44(1–2):331-340.
17. Glassey J. Multivariate data analysis for advancing the interpretation of bioprocess measurement and monitoring data. *Adv Biochem Eng Biotechnol*. 2013;132:167-191. doi:10.1007/10_2012_171
18. Mercier SM, Diepenbroek B, Dalm MC, Wijffels RH, Streefland M. Multivariate data analysis as a PAT tool for early bioprocess development data. *J Biotechnol*. 2013;167(3):262-270.
19. Suarez-Zuluaga DA, Borchert D, Driessen NN, Bakker WAM, Thomassen YE. Accelerating bioprocess development by analysis of all available data: a USP case study. *Vaccine*. 2019;37(47):7081-7089. doi:10.1016/j.vaccine.2019.07.026
20. Roychoudhury P, O'Kennedy R, Faulkner J, McNeil B, Harvey LM. Implementing multivariate data analysis to monitor mammalian cell culture processes. *Eur Pharmac Rev*. 2013;18(3):15-20.
21. 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.
22. Tsang VL, Wang AX, Yusuf-Makagiansar H, Ryll T. Development of a scale down cell culture model using multivariate analysis as a qualification tool. *Biotechnol Prog*. 2014;30(1):152-160.
23. Salim T, Chauhan G, Templeton N, Ling WLW. Using MVDA with stoichiometric balances to optimize amino acid concentrations in chemically defined CHO cell culture medium for improved culture performance. *Biotechnol Bioeng*. 2022;119(2):452-469.

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

24. Ding X, Liu J, Yang F, Cao J. Random radial basis function kernel-based support vector machine. *J Franklin Inst*. 2021;358(18):10121-10140. doi:10.1016/j.jfranklin.2021.10.005
25. Le H, Kabbur S, Pollastrini L, et al. Multivariate analysis of cell culture bioprocess data—lactate consumption as process indicator. *J Biotechnol*. 2012;162(2–3):210-223.
26. Hassan SS, Farhan M, Mangayil R, Huttunen H, Aho T. Bioprocess data mining using regularized regression and random forests. *BMC Syst Biol*. 2013;7(1):1-7.
27. Rafferty C, Johnson K, O'Mahony J, Burgoyne B, Rea R, Balss KM. Analysis of chemometric models applied to Raman spectroscopy for monitoring key metabolites of cell culture. *Biotechnol Prog*. 2020;36(4):e2977.
28. Park S-Y, Kim S-J, Park C-H, Kim J, Lee D-Y. Data-driven prediction models for forecasting multistep ahead profiles of mammalian cell culture toward bioprocess digital twins. *Biotechnol Bioeng*. 2023;120:2494-2508. doi:10.1002/bit.28405
29. Poulain A, Perret S, Malenfant F, Mullick A, Massie B, Durocher Y. Rapid protein production from stable CHO cell pools using plasmid vector and the cumate gene-switch. *J Biotechnol*. 2017;255:16-27. doi:10.1016/j.jbiotec.2017.06.009
30. Kucheryavskiy S. Mdatools – R package for chemometrics. *Chemom Intel Lab Syst*. 2020;198:103937. doi:10.1016/j.chemolab.2020.103937
31. Kuhn M, Wing J, Weston S, et al. Package 'caret' 2020;223:7.
32. Signal: Signal processing. 2014 http://r-forge.r-project.org/projects/signal/
33. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge university press; 1997.
34. Canty A. Ripley BJCsRfhcr-Powpb. Boot: Bootstrap R (S-Plus) Functions. R package version 1.3-18. 2016.
35. Greenwell B, Greenwell MB. Package 'fastshap'. 2020.
36. Lin J, Takagi M, Qu Y, Yoshida T. Possible strategy for on-line monitoring and control of hybridoma cell culture. *Biochem Eng J*. 2002;11(2–3):205-209.
37. Hippach M, Schwartz I, Pei J, Huynh J, Kawai Y, Zhu M. Fluctuations in dissolved oxygen concentration during a CHO cell culture process affects monoclonal antibody productivity and the sulfhydryl-drug conjugation process. *Biotechnol Prog*. 2018;34(6):1427-1437.
38. Tian J, He Q, Oliveira C, et al. Increased MSX level improves biological productivity and production stability in multiple recombinant GS CHO cell lines. *Eng Life Sci*. 2020;20(3–4):112-125.
39. Kuhn M. Futility analysis in the cross-validation of machine learning models. arXiv preprint arXiv:14056974. 2014.
40. Aman F, Rauf A, Ali R, Hussain J, Ahmed I. Balancing complex signals for robust predictive modeling. *Sensors*. 2021;21(24):8465-8483. doi:10.3390/s21248465
41. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed*. 2022;214:106584. doi:10.1016/j.cmpb.2021.106584
42. Yang R. Who dies from COVID-19? Post-hoc explanations of mortality prediction models using coalitional game theory, surrogate trees, and partial dependence plots. *MedRxiv*. 2020;2020(6):07.20124933. doi:10.1101/2020.06.07.20124933
43. Scapin D, Cisotto G, Gindullina E, Badia L. Shapley value as an aid to biomedical machine learning: a heart disease dataset analysis. *IEEE*. 2022;933-939.
44. Rozemberczki B, Watson L, Bayer P, et al. The shapley value in machine learning. arXiv preprint arXiv:220205594. 2022.
45. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *Adv Neur Inform Process Syst*. 2017;30:4768-4777.
46. Xiao S, Ahmed W, Mohsin A, Guo M. Continuous feeding reduces the generation of metabolic byproducts and increases antibodies expression in Chinese hamster ovary-K1 cells. *Life*. 2021;11(9):945.
47. Grilo AL, Mantalaris A. Apoptosis: a mammalian cell bioprocessing perspective. *Biotechnol Adv*. 2019;37(3):459-475.
48. Hoshan L, Jiang R, Moroney J, et al. Effective bioreactor pH control using only sparging gases. *Biotechnol Prog*. 2019;35(1):e2743. doi:10.1002/btpr.2743
49. Betts J, Warr S, Finka G, et al. Impact of aeration strategies on fed-batch cell culture kinetics in a single-use 24-well miniature bioreactor. *Biochem Eng J*. 2014;82:105-116.
50. Wagner BA, Venkataraman S, Buettner GR. The rate of oxygen utilization by cells. *Free Radic Biol Med*. 2011;51(3):700-712.
51. Downey BJ, Graham LJ, Breit JF, Glutting NK. A novel approach for using dielectric spectroscopy to predict viable cell volume (VCV) in early process development. *Biotechnol Prog*. 2014;30(2):479-487. doi:10.1002/btpr.1845
52. Opel CF, Li J, Amanullah A. Quantitative modeling of viable cell density, cell size, intracellular conductivity, and membrane capacitance in batch and fed-batch CHO processes using dielectric spectroscopy. *Biotechnol Prog*. 2010;26(4):1187-1199. doi:10.1002/btpr.425
53. Pan X, Dalm C, Wijffels RH, Martens DE. Metabolic characterization of a CHO cell size increase phase in fed-batch cultures. *Appl Microbiol Biotechnol*. 2017;101(22):8101-8113. doi:10.1007/s00253-017-8531-y
54. Lloyd DR, Holmes P, Jackson LP, Emery AN, Al-Rubeai M. Relationship between cell size, cell cycle and specific recombinant protein productivity. *Cytotechnology*. 2000;34(1–2):59-70. doi:10.1023/a:1008103730027
55. Huang YM, Hu W, Rustandi E, Chang K, Yusuf-Makagiansar H, Ryll T. Maximizing productivity of CHO cell-based fed-batch culture using chemically defined media conditions and typical manufacturing equipment. *Biotechnol Prog*. 2010;26(5):1400-1410.
56. Pappenreiter M, Sissolak B, Sommeregger W, Striedner G. Oxygen uptake rate soft-sensing via dynamic kLa computation: cell volume and metabolic transition prediction in mammalian bioprocesses. *Front Bioeng Biotechnol Adv*. 2019;7:195.
57. Wallocha T, Popp O. Off-gas-based soft sensor for real-time monitoring of biomass and metabolism in Chinese hamster ovary cell continuous processes in single-use bioreactors. *Processes*. 2021;9(11):2073.
58. Kuystermans D, Al-Rubeai M. Bioreactor systems for producing antibody from mammalian cells. *Antibody Exp Produc*. 2011; 7:25-52.
59. Krampe B, Al-Rubeai M. Cell death in mammalian cell culture: molecular mechanisms and cell line engineering strategies. *Cytotechnology*. 2010;62(3):175-188. doi:10.1007/s10616-010-9274-0

SUPPORTING INFORMATION

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

How to cite this article: Reyes S-J, Lemire L, Molina R-S, et al. Multivariate data analysis of process parameters affecting the growth and productivity of stable Chinese hamster ovary cell pools expressing SARS-CoV-2 spike protein as vaccine antigen in early process development. *Biotechnol. Prog*. 2024;e3467. doi:10.1002/btpr.3467

15206033, 0, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3467 by Ecole Polytech De Montreal, Wiley Online Library on [17/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License