Check for updates

Received: 18 February 2025 | Revised: 15 May 2025 | Accepted: 19 May 2025

DOI: 10.1002/btpr.70046

RESEARCH ARTICLE

Process Sensing and Control

BIOTECHNOLOGY PROGRESS

A recurrent neural network for soft sensor development using CHO stable pools in fed-batch process for SARS-CoV-2 spike protein production as a vaccine antigen

Sebastian-Juan Reyes[1,2] | Robert Voyer[1] | Yves Durocher[1] | Olivier Henry[2] | Phuong Lan Pham[1]

[1]Human Health Therapeutics Research Centre, National Research Council Canada, Montréal, Quebec, Canada

[2]Department of Chemical Engineering, Polytechnique Montreal, Montreal, Quebec, Canada

Correspondence

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

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

Funding information

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

Abstract

Fed-batch recombinant therapeutic protein (RTP) production processes utilizing Chinese Hamster Ovary (CHO) cells can take a long period of time (>10 days). Within this period, not all critical features may be measured routinely, and in fact, some are only measured once the process is terminated, complicating decision making. As a consequence, utilizing routine current day bioreactor online data to aid in next day predictions is a promising strategy for model predictive control-based feeding strategies. The article details the development of a proposed soft sensor that merges current day bioreactor online data and offline historical sampling data to generate predictions about the next day of the production process. This approach demonstrated the ability to track product titer, cell growth, key metabolites, and cumulative glucose consumption across the 17-day process with low normalized root mean squared error (nRMSE = 0.24) and low normalized mean absolute error (nMAE = 0.18) as well as high linearity with respect to ground data (average $R^2 = 0.97$). It was also demonstrated that the same model architecture could effectively soft sense product titer and metabolic profiles (glucose, lactate, ammonia) without having sampling day's offline data as inputs to the model. This suggests that the proposed model could act as a true soft sensor of hard-to-determine variables such as the trimeric SARS-CoV-2 spike protein that relies on end-of-process measurements to acquire the data (labor-intensive semi-quantitative SDS-PAGE gels or ELISA assay). Instantaneous specific glucose consumption rates were also predicted and showed good agreement with experimental measurements, further offering opportunities for online glucose control.

KEYWORDS

CHO fed-batch bioreactor production, CHO stable pool, data driven sensor, process analytical technologies (PAT), recurrent neural network (RNN), SARS-CoV-2 spike protein, soft sensor

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 His Majesty the King in Right of Canada and The Author(s). *Biotechnology Progress* published by Wiley Periodicals LLC on behalf of American Institute of Chemical Engineers. Reproduced with the permission of the Minister of Innovation, Science and Industry.

1 | INTRODUCTION

A soft sensor is the concomitant use of software-implemented models (soft) and hardware devices (sensor) to gather and gain new information about the process.[1] This is key because without the use of soft sensors, that is to say just exclusively using a sensor, it would be impossible to derive the same information.[2–5] At its core, these soft sensors are used with the explicit purpose of leveraging on-line data in order to infer quantitative information about complex process variables that are impossible to measure directly in a sterile system or can be measured at a very low sampling frequency.[6] Consequently, soft sensors can become useful tools in terms of monitoring and control applications within the biopharmaceutical industry.[2–5] A well-developed soft sensor that considers the needs of its stakeholders should, in theory, result in a reduction of operational surveillance and maintenance work. Additionally, soft sensors should increase the interpretability of the results of culture runs given the capacity of the models to relate various key variables to each other.[1] Given this promise, soft sensors are perfect candidates for the PAT initiative to contribute toward automated control.[1,6] Soft sensors can be split into three categories: model-driven sensors, data-driven sensors, and hybrid models.[1,7]

Model-driven sensors involve mechanistic models that are based on engineering principles, like mass or energy balances. They can provide an understanding of the processes that is inherently ingrained in biological insights.[7] Such models are capable of introducing known culture conditions such as media composition and/or culture performance indicators (cell growth, titer) to set up tangible models. Because of these characteristics, model-driven sensors can exploit known kinetic equations that capture dynamic changes of relevant variables.[8,9] In essence, these soft sensors incorporate reaction kinetics, transport phenomena, and thermodynamic constraints into the model.[8] However, these types of soft sensors must go through rigorous phases of parameter identification, uncertainty, and sensitivity analysis to properly validate said models. It must be noted that in the case where the model is reproducible and reliable, biologically interpretable information is provided to the end user that can increase the understanding of the production process.[1,8] Model-driven soft sensors can be split into two distinct categories: dynamic models and steady-state models. Dynamic models depend on balances and kinetic presumptions to suitably express rate expressions as functions of the state variables.[10] On the other hand, steady-state models originate from mass and heat transfer laws. Good examples of such steady-state models can be flux balance analysis (FBA) or metabolic flux analysis (MFA) which are stoichiometric-rooted techniques routinely used to characterize cell metabolism.[11–13] They are also useful in estimating intracellular fluxes by leveraging known extracellular analyte consumption or compound production rates as model constraints. Given that the quasi-steady state presupposition for intracellular metabolites is key, such models are considered static in nature.[8] On the other hand, kinetic models are usually expressed as a series of ordinary differential equations (ODEs) that can describe dynamic changes in metabolite concentrations, cell density, and protein expression during the cell culture process.[2–5] Because of this, cell growth and cell death can be unambiguously linked to changes in concentration of relevant nutrients and metabolic by-products. In addition, protein expression has been linked to cell growth and amino acid metabolism.[8,14,15] As a direct consequence, dynamic models can be designed with varying levels of complexity conditioned on the assumptions made by the researcher regarding the culture system in the bioreactor. This diversified degree of complexity can be tuned by considering heterogeneity within the cell population or by acknowledging the existence of known cellular compartments and their respective behaviors. On the other hand, dynamic models can be simplified if reactions are lumped to rate limiting steps.[8,14,15] Because of such caveats, model-driven sensors can be very complex and time consuming to develop.[16]

Data driven soft sensors utilize multivariate data analysis (MVDA) techniques such as partial least squares (PLS), principal component regression (PCR), and non-linear regressions such as artificial neural networks (ANN) and support vector machine regression (SVMR) in order to relate input features to predict desired variables.[17–27] These nonlinear models are particularly useful in understanding cell cultures given the fact that a lot of the interactions between key metabolic and process variables remain unknown or are highly cell line specific.[1] PLS regression and ANN are notably used to analyze spectral data. Under such conditions, the spectral data is used as input and linked to outputs such as substrate concentrations, biomass, cellular viability, or product titer.[28,29] Thanks to such models, it is possible to predict critical process parameters (CPPs) that are not available through the spectral signals or multi-sensor data alone but arise from the deconvolution of the datasets generated from such sensors. This, in theory, is an advantage over mechanistic models given that online measurements (temperature, pH, DO, oxygen flowrate, base addition, dissolved carbon dioxide flow rate, oxygen uptake rates, biocapacitance signals, Raman spectral data, cell volume) are not directly coupled to cell counts. Metabolic parameters such as lactate production/consumption, ammonia production/accumulation, and glucose consumption can be utilized to help predict said variables (protein expression, cell growth, etc.). Alternative approaches applied in bioprocessing include multiple linear regression, k-nearest neighbors (KNN), regression trees, ensemble approaches (Gradient Boosting Machine, Extreme Gradient Boosting, Adaptive Boosting, Random Forest) and Gaussian process regression.[30] Since mammalian cell culture data is complex both in its time-dependent variation and multivariate nature, methods developed for sequence forecasting have been applied. Even though ANN can capture dynamics of non-linear systems like cell culture runs, RNN is a subclass of ANN that better captures the internal temporal dependencies of a system. These architectures are particularly useful for making t-step ahead predictions of relevant state variables.[31] They have recently been applied in predicting biomass growth before and after transfection of an rAAV production process.[32] This was done by utilizing cumulative oxygen sparged, dissolved oxygen values, and cumulative dissolved oxygen as features and relating their time-related variance to cellular growth. A subclass of RNN models denominated long short-term memory (LSTM) has been applied for multivariate estimation of mammalian cell culture

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

data (total cell density, viable cell density, viability, lactate, glucose, titer) and has also been developed.[33]

Hybrid models, known as gray box models, are another relevant class of soft sensors. These types of soft sensors can be considered to be a combination of data driven soft sensors and mechanistic model driven soft sensors. They have the capacity of utilizing the benefits of each method.[2–5,9,16,34–39] Various architectural techniques exist when developing hybrid models and they can fall in three general categories: (i) Calibration, (ii) Composition, and (iii) Transformation. Calibration architectures utilize black box models to reduce mechanistic model errors. Composition architectures utilize black box models to estimate unknown terms within a mechanistic model.[16,40,41] Lastly, a transformation approach utilizes mechanistic models to generate data rich environments from which training a black box model is possible.[42] Examples of these are state observers that integrate dynamic modeling (white box models) and data driven modeling (black box models). This is realized by updating state estimates derived from noisy measurements and gradually reducing the estimation error with each iteration.[1] This is usually done assuming linear dynamics within the process and a Gaussian distribution for the error terms. Under such assumptions, a Kalman filter can be used. However, given that the process dynamics within a bioprocess are non-linear in nature, the extended Kalman filter may be applied. This method realizes a piecewise linearization through a first order Taylor series expansion.[43] Another important version of the Kalman filter that is widely used for non-linear systems is the unscented Kalman filter. This method employs an unscented transform to avoid relying on a Taylor series expansion of the system of equations to linearize the model.[44] This method can be advantageous since the unscented transform allows non linearizable functions to be used as a state observer and thus black box techniques such as support vector machine regression (SVMR) can be utilized so as to relate an online sensor output to a non-online variable.[44] It must be noted that because the accuracy of a hybrid soft sensors (gray box models) can be significantly impacted by the accuracy of the mechanistic model embedded within the gray box model, the mechanistic model requires extensive validation to ensure it can successfully represent the process.[45] Extended Kalman filters have been applied to data generated (cell density, glucose concentration, glutamine concentration, lactate concentration, ammonia concentration, and rAAV viral titer) from HEK293 processes producing recombinant adeno-associated virus (rAAV), where online viable cell densities (measured through bio-capacitance signals) and an unstructured mechanistic model are used in conjunction with neural ordinary differential equations (ODEs) to estimate cell-specific rates. This results in a soft sensor that is able to estimate continuously other state variables that are measured at low frequency.[46] Likewise, hybrid extended Kalman filters which utilize partial least squares (PLS) in order to estimate specific rates in the model have been developed and applied to CHO fed batch culture datasets (measured variables: viable cell density, the concentration of glucose, lactate, glutamine, ammonium, osmolarity, pH, pO_2,pCO_2 and titer).[47] Hybrid models can also describe the biological system by way of a mechanistic framework but define the cell specific rates through statistical expressions.[2–5] The mechanistic framework inherently constrains the solution space of the model and thus, the statistical cell-specific rate expressions can be automated.[48] Within this structure, PLS or ANN prediction resulting from multi-wavelength spectra or multivariate parameters can be fed as inputs into a mechanistic model for CHO fed batch culture monitoring.[49,50] XGBoost regressors, random forest regressors and multilayer perceptron (MLP) regressors have been used to estimate CHO cell specific rates (specific growth rate, specific productivity, and specific cumulative glucose consumption) throughout a fed-batch process by utilizing VCD, titer, temperature, glucose concentration, glucose consumed, lactate concentration, and viability as inputs.[41] These updated specific rates are then utilized as parameters in a mechanistic model to predict relevant culture outcomes (viable cell density, titer, and cumulative glucose consumption).[41] Hybrid models have also been utilized for the prediction of key product quality attributes (impurity levels, charge variants species, intact mass, total low molecule weight (LMW), and N-glycan profiling) by coupling a propagation model that describes the time evolution of cell culture variables (viable cell density (VCD), glucose, glutamine, glutamate, lactate, ammonia, cell viability, and titer) with a PLS model that regresses quality attributes as a function of cell culture variables and process conditions.[51] Alternatively, to increase the overall data richness of a production process, the utilization of unstructured mathematical models would allow the user to generate cell line relevant data so as to create information-rich environments that can be purposed to train nonlinear deep learning regression techniques like recurrent neural networks (RNN).[52] It is worth noting that transformer architecture which has been highly successful in natural language processing has been shown to not outperform- RNN structures in terms of time series forecasting.[53,54]

In this article, the development of a multivariate long-term time series forecasting model is detailed. This model is capable of realizing one step ahead predictions of key state variables (titer, viable cell density, cumulative glucose consumption, lactate, ammonia) by relying on both offline sampling data and bioreactor online data. This is key given that the offline sampling data is unevenly spaced with respect to time (every other day or every two days; e.g., total of 17 process days but only 10 measurements) thus, reliance on online data (temperature, pH, base addition, DO, integral of DO, cumulative O_2 flow, cumulative CO_2 flow) is required to update predictions on days in which no offline sampling data was available. The model was developed on a dataset generated with multiple cumate inducible CHO-GS cell stable pools expressing variants of the SARS-CoV-2 spike protein with changes in process conditions (cell passage number, MSX supplementation at induction). The important advantage of this data driven method when compared to mechanistic or hybrid modeling is that the fully pre-trained model can be readily applied by non-experts as it needs no knowledge about boundary conditions or metabolic networks. This can thus be readily transferred to production processes without additional training on its operators. The model can qualitatively capture the dynamics of metabolic and protein expression profiles by relying exclusively on bioreactor online data and easily accessible viable cell counts throughout the whole

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

17-day process. Furthermore, this model advantageously soft senses hard-to-measure variables like the SARS-CoV-2 spike protein in a daily fashion.

2 | MATERIALS AND METHODS

2.1 | Stable CHO Cell Pools and Small-Scale Cell Culture Conditions

Three stable CHO-GS cell pools expressing SmT1 trimeric spike proteins, namely Wuhan Tagless (WuTL), Delta (De), and Beta (Be) variant, were generated as described previously.[55–58] Stable 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 (Sigma-Aldrich). 125-mL (20 mL working volume) shake flasks without baffles (Corning) were used for cell maintenance. 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 between 2×10^6 and 3×10^6 cells/mL.

2.2 | Cell Culture Analytical Methods

Viable and total cell density, cell viability, main metabolites (glucose, lactate, ammonia) were measured utilizing the previously reported methodology.[55–58] Briefly, cell counts were performed with Innovatis Cedex (Roche) or ViCell Blue (Beckman Coulter) automated cell counters 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. Table 1 summarizes online and offline measurements as well as the respective relative standard deviations for analytical measurements, which were measured from replicate measurements using platform data.

2.3 | Fed-batch Cell Culture Process Conditions

All productions were conducted in parallel benchtop bioreactors 0.75 L Multifors 2 (Infors) under the conditions detailed in Table 2. Corning shake flasks were used to generate the seed trains. The bioreactors were seeded at 0.4×10^6 cells/mL and cultivated for 17 days. Temperature downshift (37°C to 32°C) was realized 3 days after seeding. A pH shift was conducted 2 days post-seeding (from 7.05 ± 0.05 to 6.95 ± 0.05). A dissolved oxygen (DO) set point of 40% (of air saturation) was chosen. Micro-spargers with a 0.0033 vvm (volume of gas per initial working volume per minute) air cap were implemented in a cascade air/oxygen strategy. Air flow rate was automatically increased to a selected maximum value (air cap) then remained constant to the end. Pure oxygen was injected as needed to maintain the DO setpoint. CO_2 and an in-house mix of $NaHCO_3$/NaOH were used to maintain pH within its selected deadband. Production induction was initiated with the addition of 4-Isopropylbenzenecarboxylate (Cumate, ArkPham). Cultures were fed with BalanCD CHO Feed 4 (Fujifilm/Irvine Scientific) and supplemented with glucose as needed to maintain glucose concentration above 17 mM (3.06 g/L) for the next sampling point. Glucose supplementation relied on estimating glucose consumption between sampling days and extrapolating the observed glucose consumption for the next sampling period. Consequently, enough glucose is added to counteract the expected glucose consumption such that the residual glucose is maintained above 17 mM. Samples were taken from the bioreactors on days −3, −2, −1, 0, 3, 5, 7, 10, 12, and 14 dpi (days post-induction) for off-line analysis, while feeding was realized in a bolus dosage from 0 dpi (induction day) onward. Cell passage number was varied across different batches (passage 5, 8, and 11) to study the impact of cell age on pool expression stability. It is known that the cell pool is heterogeneous, composing of different cells with

TABLE 1 Process variables and parameters (pH, temperature and DO) considered in the model and relative standard deviations in analytical measurements.

Offline Measurements	Cell Growth	VCD (cells/mL)	7%
	Metabolites	Lactate (mM)	2%
		Ammonia (mM)	3%
		Cumulative Glucose Consumed (mM)	3%
	Protein Production	Titer (mg/L)	12%
Online Continuous Measurements	pH Control	pH profile	-
		Base Addition Volume (mL)	-
		Total Carbon Dioxide sparged (mL)	-
	Oxygen Requirements	Total Oxygen sparged (mL)	-
		DO (% of air saturation)	-
		Integral DO (%*Day)	-
	Temperature Control	Temperature (°C)	-

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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 2 Bioreactor production process conditions.

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	Feed 4 Supplementation (% based on initial volume)
Delta	0.4	5 8 11	50 125	40–30	40	0.2	Micro	Air cap (AC) with Air/O_2 cascade	2	0 dpi (5%), 3 dpi (5%), 5 dpi (5%), 7 dpi (7.5%), 10 dpi (5%), 12 dpi (5%)
Beta	0.4	5 8 11	50 125	40–30	40	0.2	Macro	Air cap (AC) with Air/O_2 cascade	2	0 dpi (5%), 3 dpi (5%), 5 dpi (5%), 7 dpi (7.5%), 10 dpi (5%), 12 dpi (5%)
WuhanTL (WuTL)	0.4	5 8 11	50 125	40–30	40	0.2	Macro	Air cap (AC) with Air/O_2 cascade	2	0 dpi (5%), 3 dpi (5%), 5 dpi (5%), 7 dpi (7.5%), 10 dpi (5%), 12 dpi (5%)

different expression levels in contrast to a stable clone.[59] Therefore, it is critical to determine the cell age operation window to avoid a significant expression loss when the cell passage number increases.[56] Additionally, MSX supplementation (75 μM) at induction (0 dpi) was also investigated to evaluate the impact of high MSX concentration on production performance. High MSX concentration has enabled increased protein expression observed in our previous unpublished data; this effect was also shown with another group.[60] 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 as it decreased with every feed bolus addition. Raw data of all the culture conditions can be visualized in Figures S1–S5 of Annex.

2.4 | Dataset and data handling methodology

The dataset is made up of 21 production runs. 10 runs were performed with the Delta pool (De), 6 runs with Beta pool (Be), and 5 runs with the Wuhan Tag-less pool (Wu-TL). For all the productions, viable cell density, cumulative glucose consumption, lactate, ammonia, and titer were measured or calculated. Cumulative glucose consumption was estimated by adding up the glucose consumed between sampling days (difference between media glucose concentration after feed and measured glucose concentration in next sampling day) thus keeping track of total glucose consumption. Online data underwent a structured preprocessing workflow. First, Savitzky–Golay filtering was applied to the pH data to reduce sensor noise.[59] Next, daily averages were calculated for online pH, DO, and temperature measurements to smooth temporal fluctuations.[59] For other variables, such as base addition, oxygen, and carbon dioxide sparging, root cause analysis was employed to identify and exclude data points influenced by sensor faults or anomalies. As a result, key variables including DO, the integral of DO (DO_{int}), total oxygen sparged, pH, base addition, total carbon dioxide sparged, and temperature were compiled into Excel spreadsheets (Microsoft), allowing for direct comparison with sampling day data. Integral of the DO curve (DO_{int}) was chosen as a variable to aid in giving the model information about possible changes in the DO profile between monitoring days. Integral of DO, O_2 sparge rates, and CO_2 sparge rates were estimated by calculating the area under the curve of each signal through trapezoidal rule for numerical integration:

$$\mathrm{DO_{int}} = \sum_{i=1}^{n} \frac{(\mathrm{time}_{i+1} - \mathrm{time}_i) \times (DO_{i+1} + DO_i)}{2}$$

$$\text{Total Oxygen sparged} = \sum_{i=1}^{n} \frac{(\mathrm{time}_{i+1} - \mathrm{time}_i) \times (\mathrm{O}_{2_i+1} + \mathrm{O}_{2_i})}{2}$$

$$\text{Total Carbon Dioxide sparged} = \sum_{i=1}^{n} \frac{(\mathrm{time}_{i+1} - \mathrm{time}_i) \times (\mathrm{CO}_{2_i+1} + \mathrm{CO}_{2_i})}{2}$$

where i is the counter that ranges from 1 to n (from the first data point until the end of the data in the time series); time_i represents the time associated with the ith data point; $time_{i+1}$ is the time associated with the $(i+1)$th data point; DO_i represents the value of DO at the ith data point; DO_{i+1} represents the value of DO at the $(i+1)$th. Similarly, O_{2_i} and CO_{2_i} represent the respective gas flow value at the ith data point, while O_{2_i+1} and CO_{2_i+1} signify the $(i+1)$th data point of the respective gas flows.

2.5 | Details of the RNN methodology

Recurrent neural networks (RNN) are a family of neural networks for processing sequential data. There is a shared similarity with multilayer perceptron (MLP). The general network structure is represented by an input layer, one (or numerous) hidden layers, and an output layer. However, the RNN structure (Figure 1) allows the model to carry over latent information from time step to time step, thus capturing time-varying profiles.[61] RNNs and MLPs have distinct differences that have significant implications for sequence learning. While MLPs can only map from input to output vectors, RNNs have the ability to map from the entire history of previous inputs to each output. This means that an RNN can leverage a memory of past inputs stored in its internal state to influence the network's output.[62] In fact, the universal approximation theory applies to both RNNs and MLPs, but with different implications. For MLPs, it states that with enough hidden units,

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

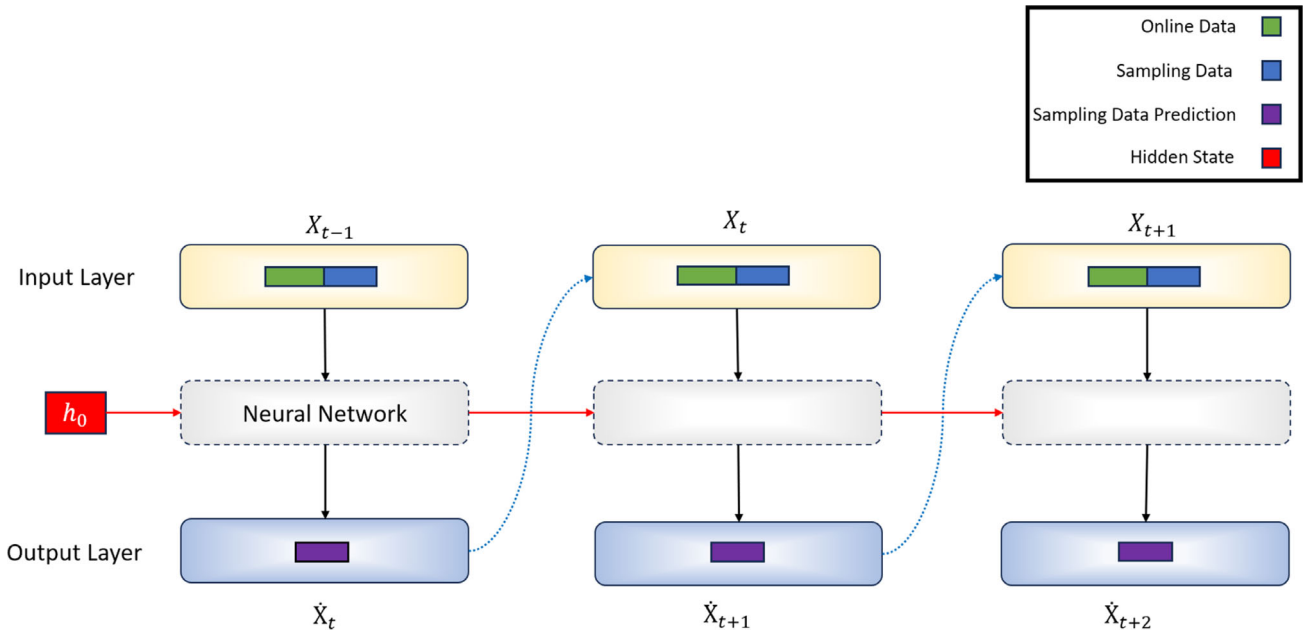

FIGURE 1 Soft sensor architecture for predicting next day sampling data. Online bioreactor data (total oxygen sparged, total carbon dioxide sparged, pH, base addition, DO, integral of DO, temperature) and measured sampling data (lactate, ammonia, cumulative consumed glucose, viable cell density, titer) along with an initial hidden state are received as inputs to a neural network. This neural network outputs the next discrete time prediction of each sampled data along with an updated hidden state to be used in the next iteration. This process is repeated for all days during the production process (17 days). If there is no available sampling data, then the bioreactor online data and the previous output predictions (rather than the ground sampling data) will be used as inputs in the next iteration.

an MLP can approximate any measurable mapping from input to output. On the other hand, for RNNs, the equivalent result is that with a sufficient number of hidden units, an RNN can approximate any measurable sequence-to-sequence mapping.[62] RNNs are specifically designed for processing sequential data, just as convolutional neural networks (CNNs) are specialized for processing grid-like data such as images.[63] RNNs excel at handling sequences of values, and they can scale to longer sequences compared to networks without sequence-based specialization. Additionally, most RNNs are capable of processing sequences of variable length, providing flexibility in handling diverse data inputs.[63]

The latent information, which functions as a network memory, is captured within the hidden state (h_t), which is updated at each iteration as described in the equations below:

$$\mathrm{MLP}_{\mathrm{hidden}} : h_t = W_{\mathrm{hidden}} \times \mathbf{\Phi}(W_{in} \times [h_{t-1}, O_t, X_t])$$

$$\mathrm{MLP}^i_{\mathrm{sampled}} : \dot{X}^i_{t+1} = W^i_{\mathrm{out}} \times \mathbf{\Phi}\left(W^i_{h_in} \times h_t\right)$$

$\mathrm{MLP}_{\mathrm{hidden}}$ is multilayer perceptron that receives sampled data and online data, t – discrete time index, h_t – hidden state at time step t, W_{hidden} – hidden state weight matrix, $\mathbf{\Phi}$ – standard hidden layer activation function in ANN (logistic, hyperbolic, tangent, sigmoidal, etc.), W_{in} – input weight matrix, h_{t-1} – hidden state at time step t-1, O_t – online variable at time t ,X_t – sampled input vector at time t; $\mathrm{MLP}^i_{\mathrm{sampled}}$ – multilayer perceptron that projects the resulting hidden state to a sampled variable space and predicts the updated time series prediction for all i sampled variables (titer, viable cell density, lactate, ammonia, cumulative glucose consumption), $\dot{X}^i_{t+1}$ – state vector predictions at a future time discrete index for all i sampled inputs (titer, viable cell density, lactate, ammonia, cumulative glucose consumption); W^i_{out} – intermediate hidden state to predicted variable weight matrix for all i sampled variables, $\mathbf{\Phi}$ – standard hidden layer activation function in ANN (logistic, hyperbolic, tangent, sigmoidal, etc.), $W^i_{h_in}$ – weight matrix to transition from a global hidden state to a variable specific hidden state for all i sampled variables, h_t – hidden state at time step t.[31,33]

For the purpose of this paper, the rectified linear unit (ReLu) function was utilized. The hidden state serves as an internal representation of the network and captures information about previous inputs in the sequence. The hidden state is updated at each time step and serves as a way for the network to maintain information about the context or history of the input sequence (Figure 2). The initial hidden state $h0$ was initialized with a zero vector, which can be interpreted as carrying over no information from the past, as the process had not yet been initialized.

Data preprocessing was done with Pandas and Numpy which are important libraries of Python (version 3.9.13) programming.[64,65] Architecture design and training of the soft sensor were done with PyTorch.[66] The following hyperparameters were used: The MLPs were all conformed by 256 hidden units. All MLPs utilized a rectified

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

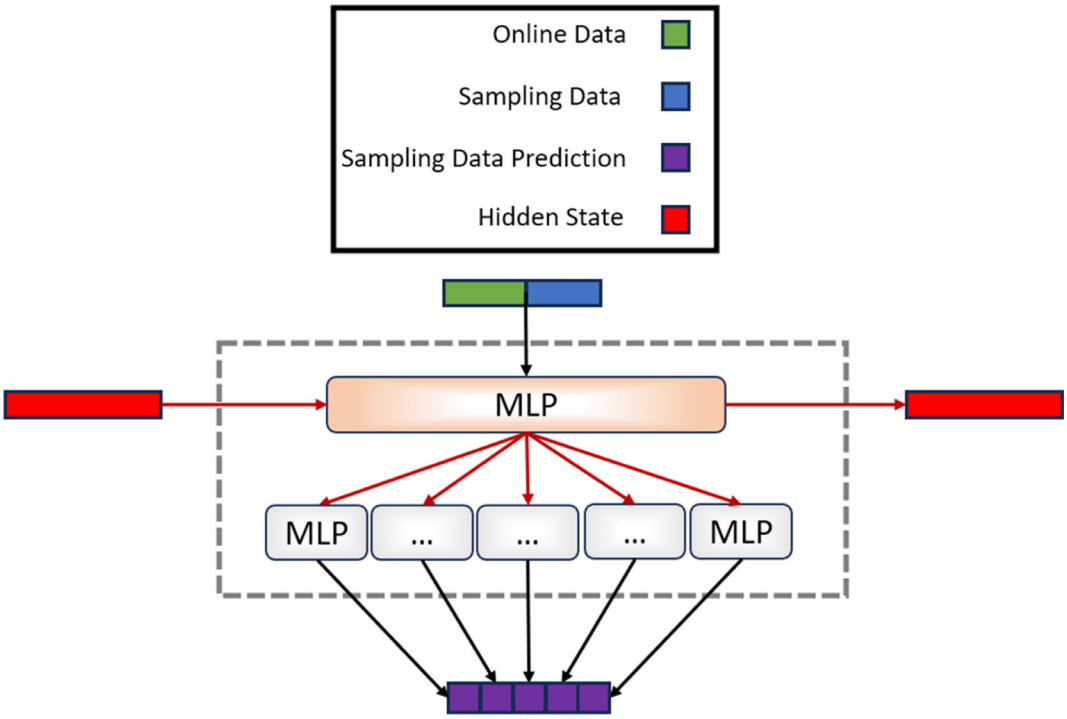

FIGURE 2 Internal neural network architecture. Offline sampling data, bioreactor online data, and hidden state values are received as inputs to an MLP. The MLP outputs an updated hidden state which serves as input for several MLPs (an MLP for each measured variable). Each MLP projects the resulting hidden state to a sampled variable space and predicts the updated time series prediction for each variable. The hidden state is stored for the next iteration.

linear unit activation function (ReLu). Learning rate was set to 1.6e-4. All input data were mean centered and standardized for each variable. Stochastic gradient descent (SGD) was employed as the optimizer with momentum of 0.97 and weight decay of 0.125 to avoid over fitting in the training phase. The training phase included 5000 epochs. An 80% training and 20% test split on the dataset was approximated. Training-test split was randomly realized such that 8 Delta cultures are used to train the network while 2 are left for testing. Similarly, 5 Beta cultures were used to train and 1 was left over to test. Lastly, 4 Wuhan Tag-less cultures were used to train and 1 culture was left over to test. Regarding model validation, root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) were used as validation metrics to evaluate model performance for each predicted feature (e.g., glucose consumed per day, lactate, ammonia, viable cell density, titer) across individual culture run. RMSE, MAE, and R^2 were first calculated for each batch prediction then averaged out across batches. To facilitate comparisons across features, RMSE and MAE values for each feature in every batch were normalized by the corresponding standard deviation of the feature in that batch. This normalization yields unbiased estimates, normalized RMSE (nRMSE) and normalized MAE (nMAE). Having normalized errors (nMAE and nRMSE) below 1 indicates that the model's prediction errors are smaller than the inherent variability in the experimental data, as represented by the feature's standard deviation within the experimental runs.[32,33]

$$\mathrm{RMSE}_{\mathrm{global}} = \frac{\sum \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} \left(y_{\mathrm{ground_truth}} - y_{\mathrm{pred}}\right)^2}}{\text{number of batches}}$$

$$\mathrm{MAE}_{\mathrm{global}} = \frac{\sum \frac{1}{n} * \sum_{i=1}^{n} abs\left(y_{\mathrm{ground_truth}} - y_{\mathrm{pred}}\right)}{\text{number of batches}}$$

$$\mathrm{nRMSE}_{\mathrm{global}} = \frac{\sum \frac{\sqrt{\frac{1}{n} * \sum_{i=1}^{n} \left(y_{\mathrm{ground_truth}} - y_{\mathrm{pred}}\right)^2}}{\text{Standard_Deviation}}}{\text{number of batches}}$$

$$\mathrm{nMAE}_{\mathrm{global}} = \frac{\sum \frac{\frac{1}{n} * \sum_{i=1}^{n} \mathrm{abs}\left(y_{\mathrm{ground_truth}} - y_{\mathrm{pred}}\right)}{\text{Standard_Deviation}}}{\text{number of batches}}$$

Here, y_{ground_truth} is the real sampled data of a feature, y_{pred} is the prediction of a feature from the model, i represents the ith value of a given feature in a batch, and n represents the total number of values of a feature in a batch, Standard_Deviation is the standard deviation of a batch for a given feature. Calculating nRMSE and nMAE for the test set is particularly important; values significantly below 1 demonstrate that the model achieves higher precision than randomly selecting parameter values from the experimental ensemble distribution.[33] Additionally, estimating error metrics for each unique batch allowed for the calculation of statistical measures such as the mean and

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

standard deviation of model errors, providing a comprehensive view of model performance.

3 | RESULTS AND DISCUSSION

The prediction versus ground truth sampled data scatter plot for each of the 17 training batches is shown in Figure 3. A clear linear relationship (R^2 values above 0.95 as indicated in Table 3) is seen in all feature predictions (titer, cumulative consumed glucose, ammonia, lactate, and VCD). The scatter plot results can center around the diagonal line which represents the $R^2 = 1$ ideal model. It must be noted that for lactate and VCD predictions (Figure 3d,e), there exists a slight deviation from the $R^2 = 1$ at higher concentrations. This dataset focused on evaluating the effects of increased cell age on protein production and determining whether higher MSX concentrations (from 50 μM to 125 μM) at induction could enhance protein production outcomes.[59] Both the WuTL and Beta pools showed stable protein production across different passage numbers. In contrast, for the Delta pool, passage number influenced the variability of key metrics, with higher passage numbers resulting in greater spread in the data. Average protein concentrations decreased with higher passage numbers (P5: 600 mg/L, P8: 361 mg/L, P11: 398 mg/L), while ammonia accumulation increased with cell age, particularly at P8 and P11.[59] No significant differences were observed between the base (50 μM MSX) and higher MSX (125 μM MSX) conditions, and accounting for passage number revealed that MSX supplementation did not significantly affect cell behavior.[59] Importantly, each CHO pool exhibited distinct growth patterns and protein production characteristics, with the Delta pool showing the highest VCD accumulation and the WuTL pool achieving higher average spike protein yield (Figures S1–S5). The Delta pool dataset exhibited the most variability, particularly in terms of lactate and ammonia accumulation (Figures S1–S5). which again could be related to its pool age related impacts.

The time series progression of each feature in the training set (Figures S6–S10) demonstrates that the span of predictions closely aligns with the span of the true data for all pool types. This alignment indicates that the training phase effectively captured the distinct growth and metabolic characteristics unique to each cell pool. Additionally, an examination of the percentage error time profiles reveals a clear pattern: errors decrease significantly after the initial two predictions and subsequently stabilize at values below 20% across most features. This consistency suggests that model accuracy remains robust in the training phase throughout the 17-day prediction process without noticeable degradation over time.

Next, to determine if the model can generalize, it was tested on 4 production runs that were not utilized in the training process. These hold-over cultures consisted of 2 Delta cultures, 1 Beta culture, and 1 WuTL culture (Figure 4). The WuTL production was terminated 2 sampling days earlier than normal (harvest at 12 dpi rather than 14 dpi). As it can be seen from Figure 4, time profile tracking of key features is possible even in non-sampling days. This is key given the fact that the manufacturing process lasts 17 days in which it is impractical to have every day sampled values, especially for the long-to-measure outputs such as protein determined by SDS-PAGE, which is the case for SARS-CoV-2 spike protein. For these non-sampling days, predictions are carried out by relying on previous day model predictions and the online data of the bioreactor (base addition, oxygen flow, DO, integral of DO, pH, temperature). In Figure 4a, it is clear that titer time series tracking is possible despite differences in pool protein production behavior. For example, the Beta and WuTL pools demonstrate vastly different endpoint titer results despite having identical kinetics from 0 to 7 dpi (Figure 4a, left side). Both Delta pools demonstrate similar protein production behavior until 9 dpi, from which Delta 1 culture outperforms Delta 2 culture by close to 100 mg/L (Figure 4a, right side). This subtle increase in the end phase of the culture is also captured by the model, although global RMSE overlaps between both cultures. From Figure 4b, ammonia profiles can be detailed. It is clear that the Beta and WuTL pools exhibit different ammonia accumulation kinetics after the onset of protein production (0 dpi). Increased ammonia accumulation is evident in the WuTL culture following the onset of induction while the Beta pool has an accumulation lag between 0 and 7 dpi, followed by rapid ammonia accumulation after 7 dpi. These changes in kinetics are predicted by the model despite having no measured sampling values at the points in which the kinetics begin to diverge. Similarly, for the Delta cultures, both productions show a plateau in ammonia accumulation from 0 to 7 dpi and then an increase accumulation until 14 dpi (Figure 4b, right side). This behavior is tracked by the model even between 7 to 10 dpi, in which no sampling data exist. From Figure 4c, it can be seen how the cumulative consumed glucose profile changes with culture time, namely the deceleration in cumulative consumption following the temperature shift from 37°C to 32°C. This reduction in glucose consumption can be linked to the cell growth slowdown that occurred at a lower temperature. Delta 1 pool predictions overestimate (from 5dpi onwards) the total profile such that endpoint cumulative glucose consumed is overestimated by 20.5 mM. This may be caused by the fact that Delta 1 culture has high lactate accumulation (above 30 mM) (Figure 4d) and thus the model may be predicting the cumulative glucose consumption to be representatively higher. The Beta culture glucose consumed is underestimated from 10 dpi until the end of the process, such that the measured values do not overlap with the global RMSE of model predictions. Figure 4d shows how lactate profiles can be tracked with the proposed model. For the Delta cultures, large differences in peak lactate values are observed which, in turn, impacts the lactate re-absorption profile (5–14 dpi). These differences in kinetics for the same pool in two different culture runs are captured by the model. Moreover, it is worth mentioning that Delta 1 culture was performed with cells having a passage number 11 while Delta 2 culture was done with cell passage number 8. The large deviation in lactate concentration could be related to their different cell passage numbers, as previously mentioned.[59] The latter work shows that Delta pools demonstrated increased variability with increased culture age. The lactate time courses of the Beta and WuTL pools are tracked throughout culture, even in the case where a sudden lactate re-production occurs after 10 dpi (WuTL) (Figure 4d). This increase lags the real value

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

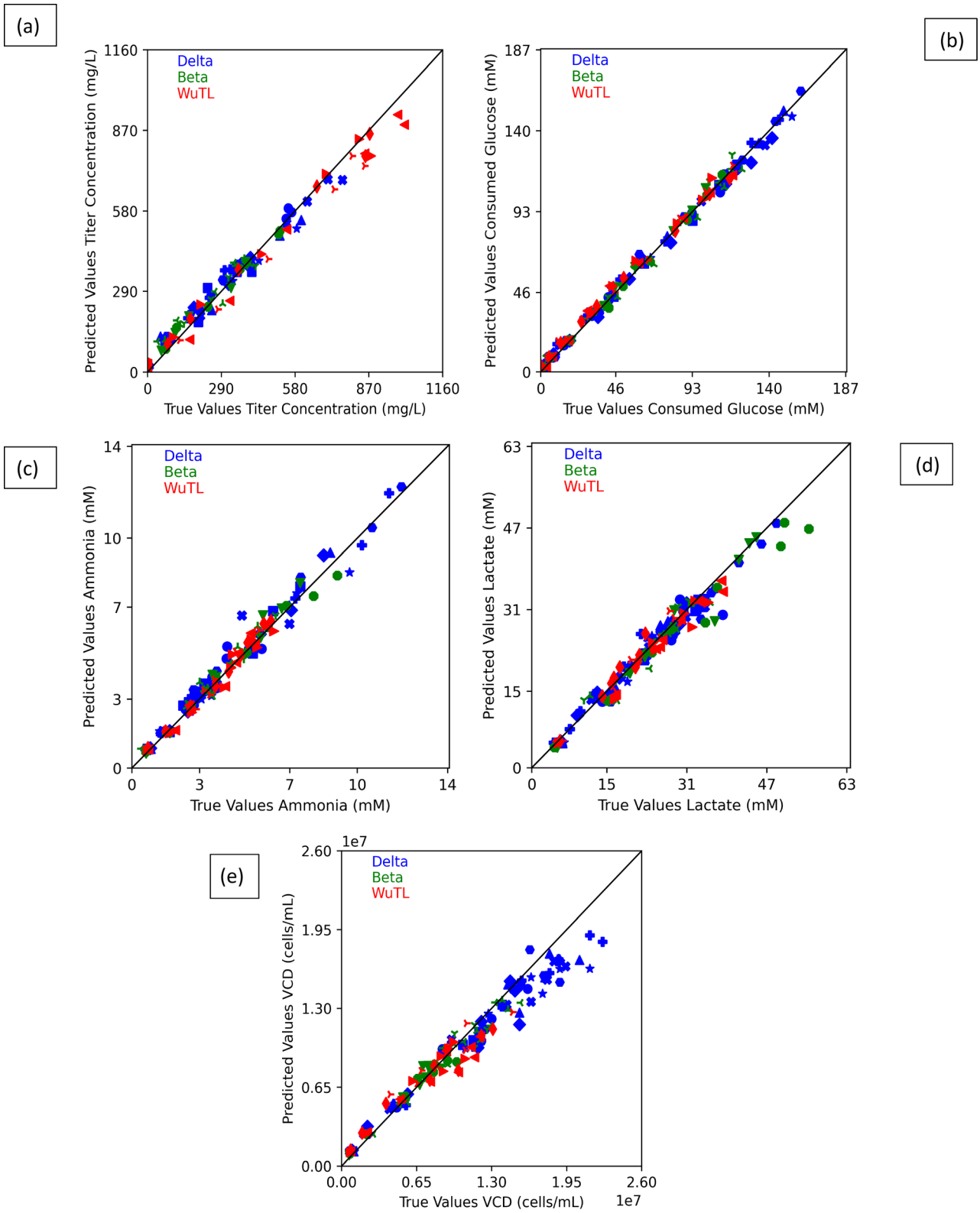

FIGURE 3 Model results for trained features in the training dataset comprised of 17 cultures (8 Delta pool batches, 5 Beta pool batches and 4 WuTL pool batches). Scatter plots of every batch prediction versus every batch sampled data for the whole training set. Diagonal line represents the $R^2 = 1$ ideal model. (a) Titer. (b) Cumulative glucose consumed. (c) Ammonia. (d) Lactate. (e) Viable cell density (VCD). Pools are color coded; Beta is blue, delta is green and WuTL is red. Each unique symbol represents an individual batch culture. There are 9 comparable data points for each culture which results in 153 points (17×9) on each graph.

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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 3 Global RMSE, MAE, nRMSE, nMAE, R^2 metrics for titer, ammonia, cGC (cumulative glucose consumed), lactate, VCD (viable cell density) for train and test datasets and their respective standard deviations.

	$RMSE_{train}$	MAE_{train}	R^2_{train}	$nMAE_{train}$	$nRMSE_{train}$	$\sigma RMSE_{train}$	σMAE_{train}	$RMSE_{test}$	MAE_{test}	R^2_{test}	$nMAE_{test}$	$nRMSE_{test}$	$\sigma RMSE_{test}$	σMAE_{test}
Titer	33.5	27.7	0.99	0.16	0.13	10.3	8.3	51.5	38.8	0.96	0.18	0.24	23.0	15.6
Ammonia	0.4	0.3	0.97	0.18	0.14	0.1	0.1	0.4	0.3	0.98	0.12	0.18	0.1	0.1
cGC	3.3	2.8	1.00	0.08	0.06	0.7	0.7	7.9	6.2	0.99	0.16	0.20	4.4	3.3
Lactate	2.0	1.5	0.97	0.23	0.17	0.8	0.5	2.8	2.3	0.94	0.25	0.31	1.1	0.9
VCD	1.24e6	9.74e5	0.97	0.24	0.19	5.02e5	3.68e5	1.36e6.	1.03e6	0.96	0.20	0.26	7.03e5	5.49e5

presumably because, in the absence of measured lactate values, online signals like pH and base (that are known to relate to lactate changes) did not show a marked change in response to the lactate accumulation (pH went from 6.96 at 10 dpi to 6.99 at 11 dpi to 6.97 at 12 dpi and base addition was never triggered as pH values stayed within the allowed deadband). Figure 4e shows that cell growth profiles can be tracked across the 17-day culture run. It is noteworthy that even though WuTL and Beta cultures had identical viable cell density profiles until 3 dpi, the Beta culture undergoes a secondary growth spurt which is consequently tracked by the model while the WuTL culture remains in a plateau phase. Similarly, for the Delta cultures, identical viable cell density profiles are observed until 3 dpi, from which Delta culture 2 enters a secondary growth phase while Delta culture 1 does not. Interestingly, both distinct profiles are captured by the model. This is important because, when taking into account final titers, it suggests that cultures that underwent secondary growth phases had lower endpoint titers. Consequently, monitoring that cell densities remain within a plateau phase during the protein production phase would be of interest. This argument can be supported by the development of a biphasic process strategy in which cells will be allowed to grow to a certain high density; then a process trigger such as lower temperature (31°C to 34°C from 37°C) or chemical addition (sodium butyrate, valeric acid)[67,68] will be introduced to keep cells in a biomass steady state while boosting the production.

As it can be seen from Table 3, both train and test model predictions have strong linearity with respect to measured values ($R^2 \geq 0.9$). The lowest R^2 values in the test set correspond to lactate, titer and VCD. These time series profiles displayed sharp variations (sudden lactate re-absorption or secondary growth phases) causing the model predictions to lag measured values before converging again. When detailing the normalized RMSE and MAE, which take into account the standard deviation of the datasets, it can be seen that the predicted features have normalized error metrics significantly below unity (<1). This demonstrates the applicability for the model to predict the time series variation of multiple features with highly nonlinear kinetics like cellular growth and lactate formation which can increase rapidly and then decline depending on the culture phase. The model's nRMSE and nMAE metrics for VCD, titer, and lactate are in line with similar state-of-the-art applications of data driven soft sensors for time series tracking.[32,33] The standard deviation of the error metrics was observed to increase between the training and test datasets across all predicted variables (Table 3). Figure 5 presents box plots of the RMSE and MAE distributions, which show that while both error metrics and their distributions increase, the interquartile ranges generally overlap between the training and test sets for all predicted variables. This suggests that the model is not subject to overfitting. The only notable outlier in the test set occurs in the lactate prediction errors for the WuTL culture (Figure 5d). This outlier can be attributed to the culture's lactate reproduction behavior, which was not adequately forecasted on day 12. An additional observation is the increased variability in the error metrics for consumed glucose in the test set (Figure 5c). This increase is due to the overestimation of glucose consumption in the Delta 1 culture and the underestimation in the Beta

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

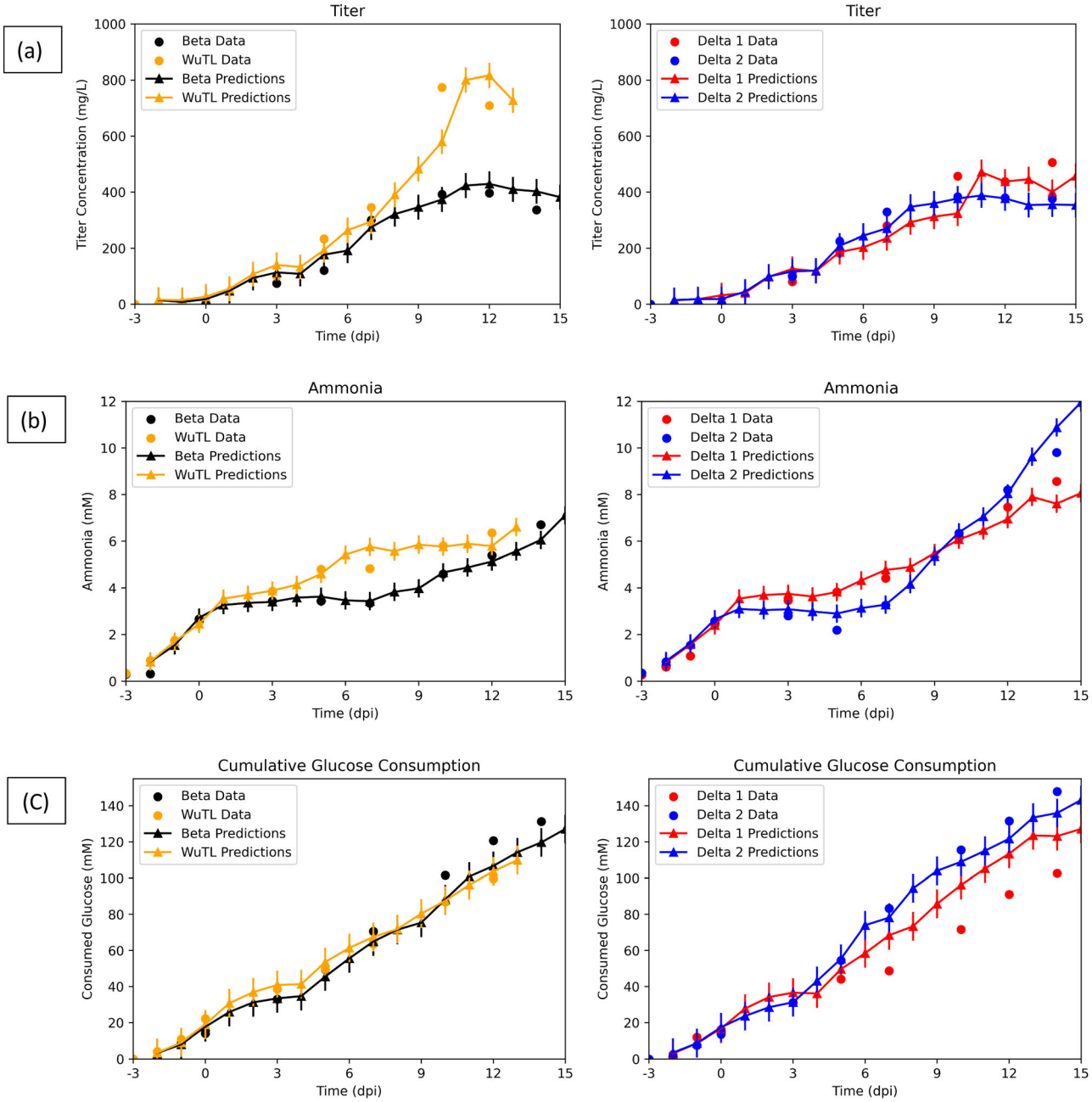

FIGURE 4 Model results for key features in the test dataset comprised of 4 cultures (1 Beta pool batch, 1 WuTL pool batch and 2 Delta pool batches). Continuous lines with triangle dots are everyday predictions, while circular dots are measured values. (a) Titer profile. (b) Ammonia concentration. (c) Cumulative glucose consumption. (d) Lactate concentration. (e) Viable cell density. The error bars represent the global root-mean-square errors of the RNN predictions based on the test dataset and the corresponding target variable.

culture. This suggests that incorporating information about feed addition and glucose supplementation into the model could help reduce the spread in these error metrics. While the titer error does increase (Figure 5a) in the test set, it should be noted that due to its estimation with semi-quantitative gels, the model error remains below the method's variability. This suggests that the proposed framework can aid in monitoring recombinant protein production. Furthermore, it can be observed that the distributions of error metrics for VCD (Figure 5e) are comparable between the training and test sets further supporting the notion of low overfitting.

Given that the model relies on previous day predictions and current day online data to estimate next day features and that it was shown to work reasonably well during the non-sampling periods of the 17-day production process, it was hypothesized that the same model without alternative training for parameter tuning could be utilized to predict metabolic data and titer profiles in the absence of such

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

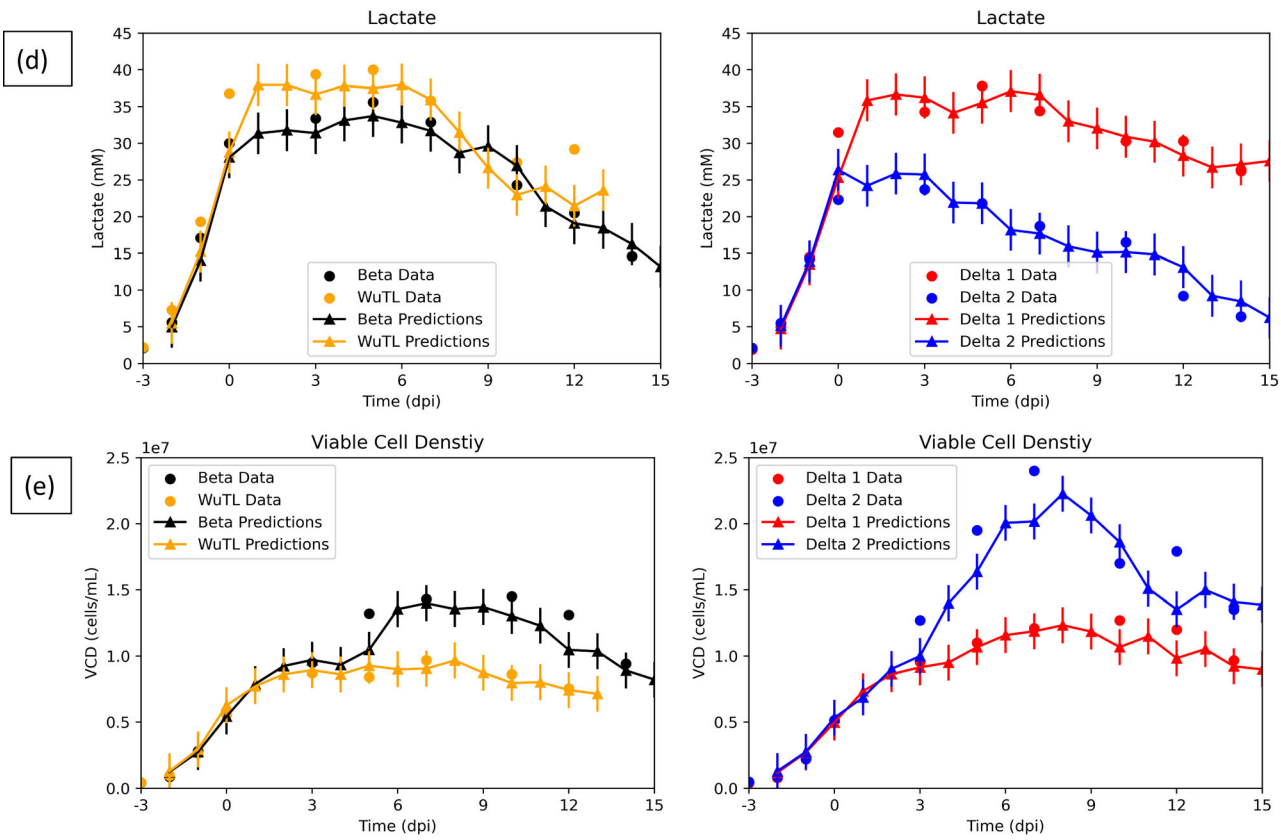

FIGURE 4 (Continued)

data. Consequently, the model would rely on initial feature values which function as initial guesses, sampling day cell counts, and bioreactor online data to generate predictions of ammonia, cumulative consumed glucose, lactate, and titer throughout the 17-day process. The objective then was to evaluate if the current model has suitable potential for true soft sensing capabilities which would be of interest for hard-to-detect variables such as non-antibody recombinant proteins. As it can be seen from Figure 6a,b, qualitative trends of titer and ammonia are tracked despite having no sampling day values of said features to aid in the updated prediction. Furthermore, glucose consumption (Figure 6c) is also tracked throughout the culture with only Delta 2 culture being overestimated at 14 dpi by 20 mM. Alternatively, lactate profiles (Figure 6d) for the Beta and WuTL pool are underestimated from induction until the end of the process. For the WuTL process, the lack of sampling lactate data did not allow the model to predict the sudden lactate re-production phase (10–12 dpi). One of the reasons for this underestimation may be due to the fact that since key online data values like base addition are governed by a PID control, which is implemented with a pH deadband, indirect estimation of lactate may lag or be underestimated. One such example is the case of the WuTL and Delta 1 cultures. WuTL produces a higher peak lactate (40 to >37.8 mM) while having lower total base addition (12 to <13.3 mL). This pH deadband activation is then very connected to the lactate absorption which pushes the pH values away from the activating deadband limit. Additionally, predicting lactate re-production in the end phase of the culture within the context of pH deadbands also becomes difficult as the previous lactate consumption coupled with the base addition during the lactate production phase necessarily drives pH values closer to the upper edges of the pH deadband. Thus, any low amounts of lactate production will not be enough to activate base addition and will generally result in pH remaining closer to the upper edges of the pH deadband. For example, in the case of the WuTL culture from 7 dpi until 12 dpi, daily average pH values are [6.96, 6.98, 6.99, 6.96, 6.99, 6.99] and total carbon dioxide sparged values increased every day from 7 to 12 dpi [4603, 4650, 4680, 5211, 5866, 6023 mL] to control the pH in the deadband (6.9–7.0). It must also be noted that global RMSE did increase for lactate prediction when compared to operating the model with sampling day values for metabolites. Similarly, although cell counts were added to the model every sampling day for prediction, nRMSE and nMAE did increase when compared to the previous results (Table 3) in which all sampling data was available to aid in feature predictions. It suggests that the measured quantities regarding lactate accumulation, glucose consumption, ammonia accumulation, and protein production aided in next day prediction of cell culture growth dynamics.

As shown in Table 4, only VCD shows a reduction in R^2 indicating slight deviation in linearity from predictions. An increase in nRMSE and nMAE can be observed for lactate, ammonia, and VCD

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

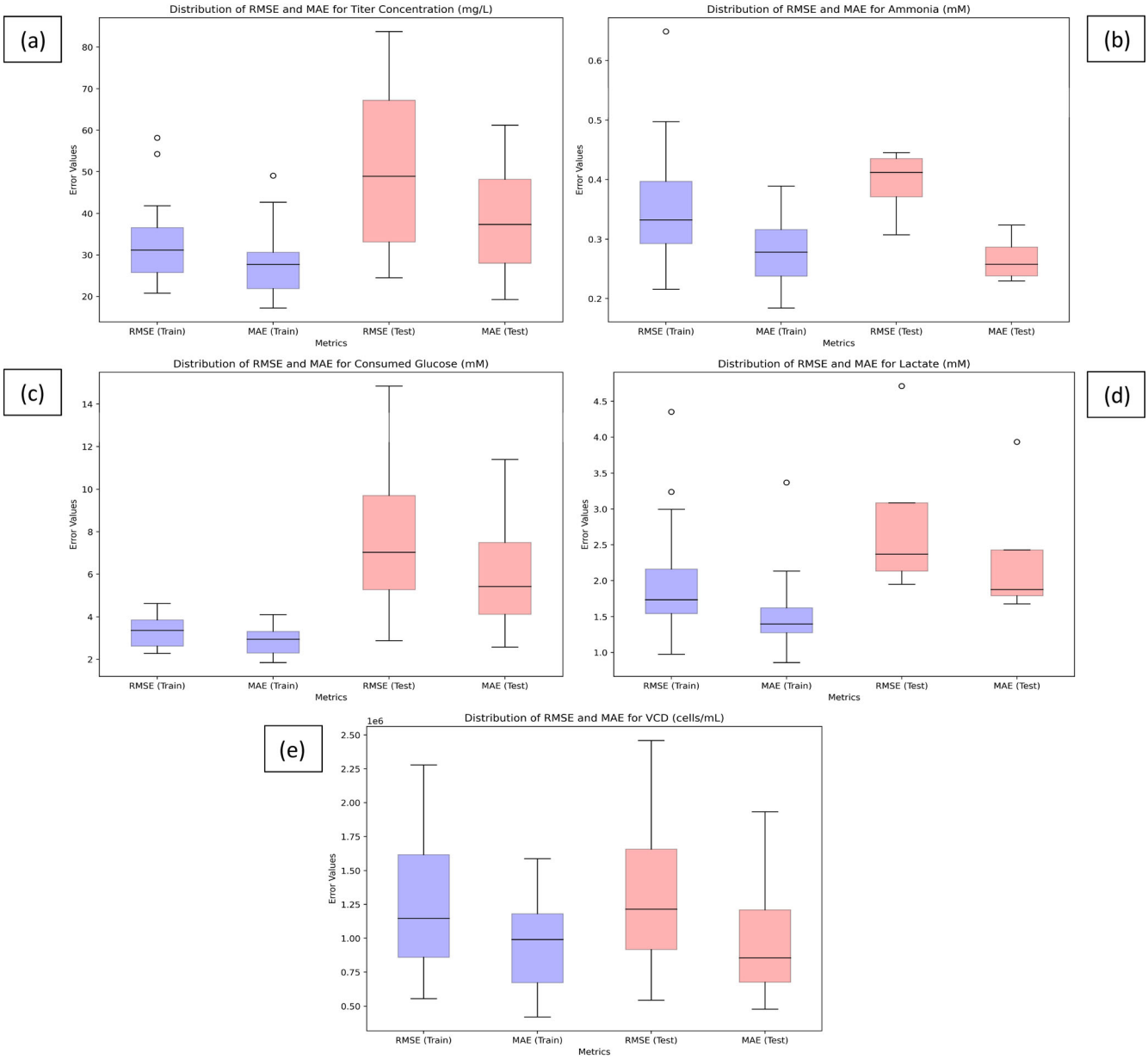

FIGURE 5 Box plot of RMSE and MAE distribution in train and test dataset for (a) Titer. (b) Ammonia concentration. (c) Cumulative glucose consumption. (d) Lactate concentration. (e) Viable cell density.

predictions indicating slight increase in error predictions. However, despite the increase in error metrics, qualitative tracking is still possible. This suggests that the model architecture with cell count measurements (which are easy to obtain manually let alone with dedicated cell counting machines) and online data from the bioreactor is robust enough to soft sense hard-to-detect variables (in this case metabolic variables and protein production). This is especially interesting within the context of difficult-to-quantify proteins such as the SARS-CoV-2 spike protein since quantification is generally done through ELISA assays or semi-quantitative SDS-PAGE once the process is finished.[56,58] Consequently, throughout the process, no knowledge regarding the trajectory of the titer values is known, hampering decision making and slowing down proposals for process improvement since retrospective analysis of titers needs to be realized before conclusions can be drawn.

Given that estimation of cell growth and cumulative glucose consumption rates were accurate even when no glucose consumption data was utilized in the test set, estimation of specific glucose consumption rates (qGluc) was performed. Since the model prediction represents cumulative glucose consumption up to any given day, the derivative of this output gives the glucose consumed every day (since every prediction is equally spaced in the time dimension, the derivative is calculated using the difference between consecutive datapoints such that Consumed_Glucose $= cGC_{(i+1)} - cGC_{(i)}$. Given the VCD

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

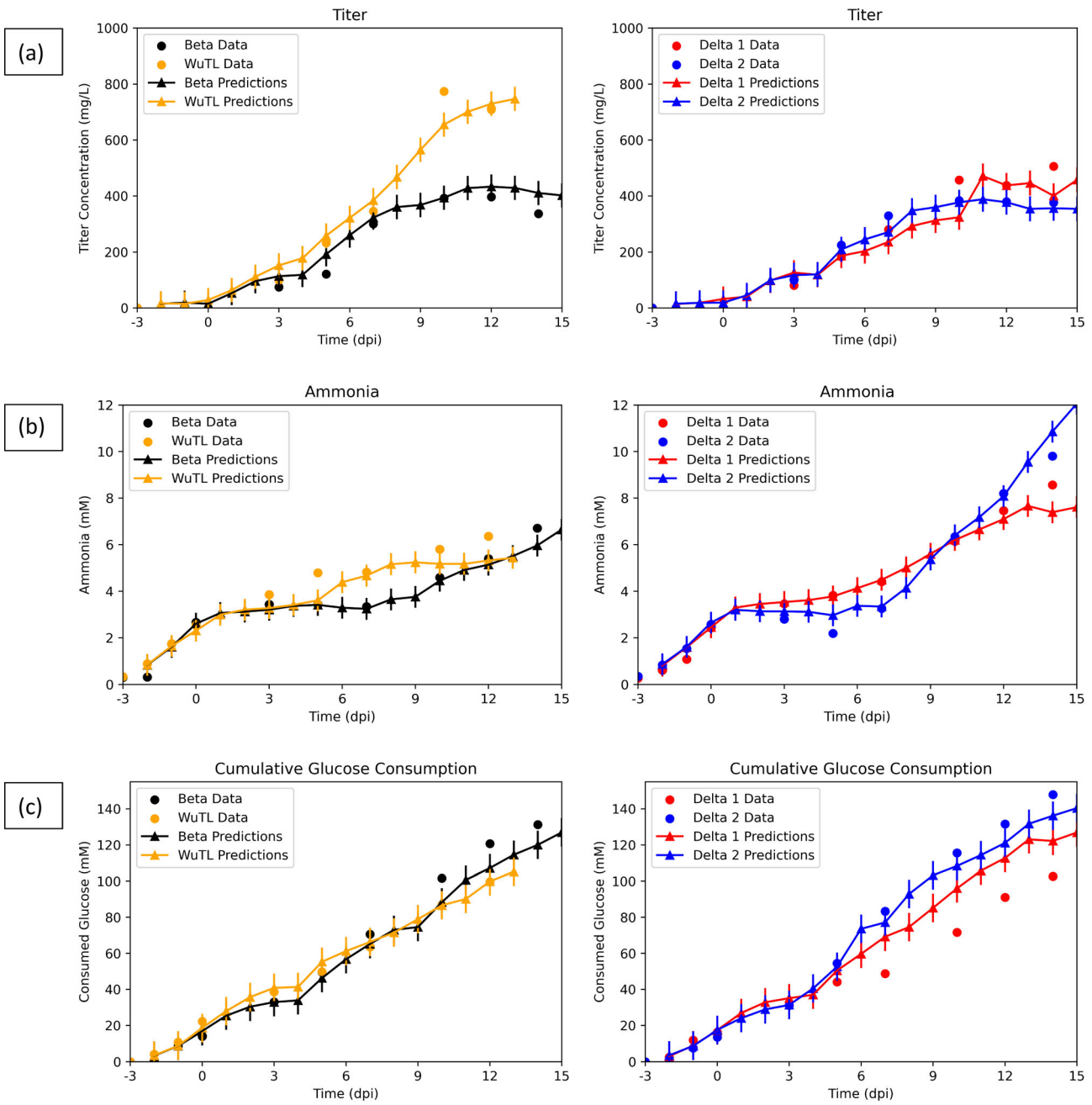

FIGURE 6 Model results for key features in the test dataset comprised four cultures (one Beta pool batch, one WuTL pool batch, and two Delta pool batches) without metabolic and titer sampling data. Continuous lines with triangle dots are everyday predictions, while circular dots are measured values. (a) Titer. (b) Ammonia. (c) Cumulative glucose consumed. (d) Lactate. (e) Viable cell density (VCD). The error bars represent the global root-mean-square errors of the RNN predictions based on the test dataset and the corresponding target variable.

predictions, the integral of viable cell concentration (IVCC) can be estimated through the trapezoid rule of numerical integration. Consequently, differentiating these IVCC values allows for the estimation of ΔIVCC ($\Delta IVCC = IVCC_{(i+1)} - IVCC_{(i)}$). If the consumed glucose estimated at every time point is then divided by the ΔIVCC at every point in time ($\Delta IVCC_i$), predicted specific glucose consumption rates ($qGluc_{pred}$) can be calculated.

$$qGluc_{pred} = \frac{Consumed_Glucose_i}{\Delta IVCC_i}$$

The ground data of glucose consumed between sampling days ($Consumed_Glucose_k$) can be divided by the $\Delta IVCC_k$ (change in IVCC between sampling days), then specific glucose consumption rates can be estimated. The subscript k indicates the ground data at sampling day k.

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

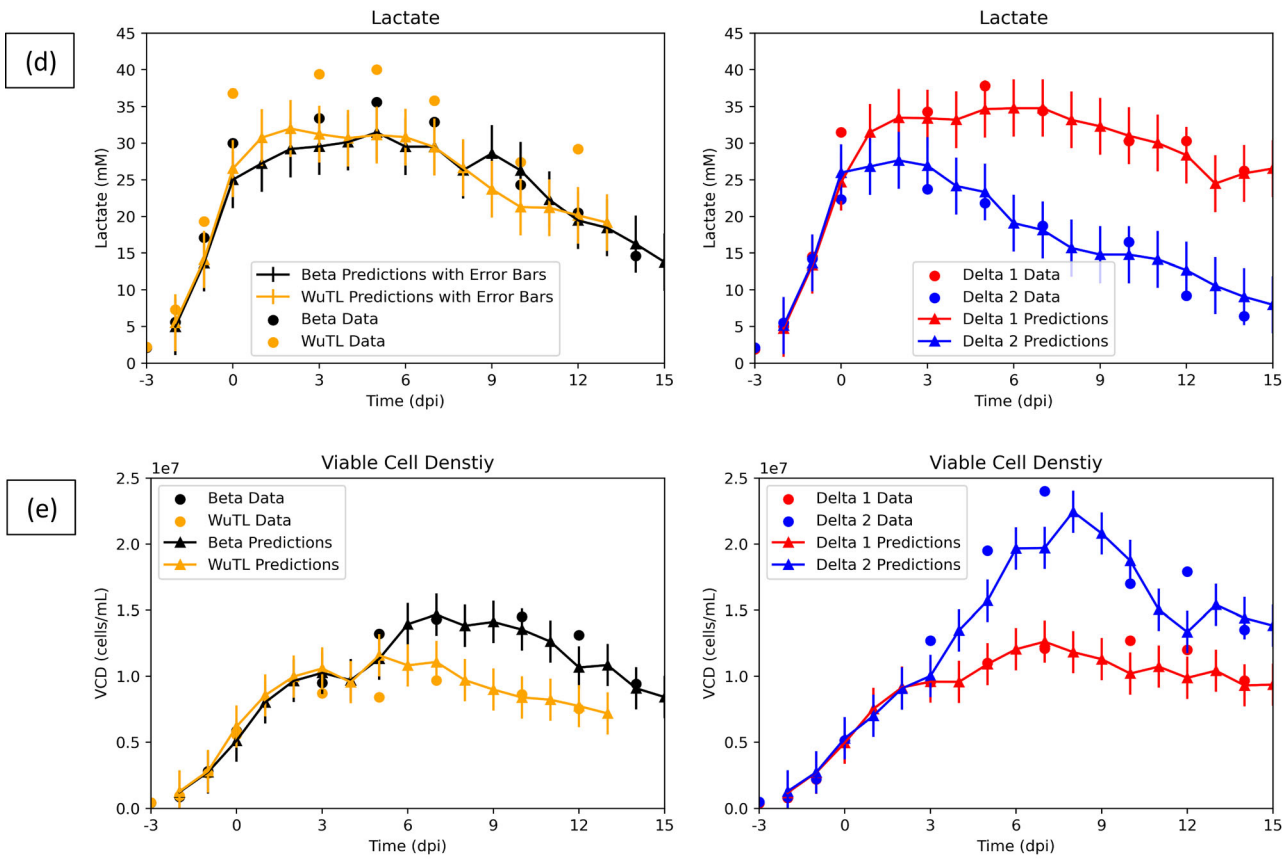

FIGURE 6 (Continued)

TABLE 4 Global RMSE, MAE, nMAE, nRMSE, R^2 metrics and their respective standard deviations for titer, ammonia, cGC (cumulative glucose consumed), lactate, VCD (viable cell density) for the test dataset without utilizing sampling day metabolic and titer data in model predictions.

	$RMSE_{test}$	MAE_{test}	R^2_{test}	$nMAE_{test}$	$nRMSE_{test}$	$\sigma RMSE_{test}$	σMAE_{test}
Titer	48.3	38.1	0.98	0.19	0.24	18.2	13
Ammonia	0.5	0.3	0.98	0.15	0.22	0.1	0.1
cGC	7.9	6.0	0.99	0.15	0.20	4.3	3.3
Lactate	3.9	3.4	0.95	0.36	0.41	2.1	2.1
VCD	1.60e6	1.19e6	0.94	0.24	0.33	6.4e5	5.46e6

$$qGluc_{Ground_Truth} = \frac{Consumed_Glucose_k}{\Delta IVCC_k}$$

In Figure 7, it is possible to see an overlay between the estimated specific glucose consumption rates and the predicted specific glucose consumption rates. For the 4 test cases, qualitative tracking of specific glucose consumption rates is possible, showing a fast decrease from −1 to 0 dpi, which represents the end of the growth phase and then near constant qGluc after temperature downshift. This qualitative tracking of specific glucose consumption rates (qGluc) is important, as knowledge regarding the specific consumption represents knowledge regarding its internal metabolic state.[69,70] Additionally, having 1 day ahead predictions of specific glucose consumption rates represents a pseudo-online opportunity for nutrient control which has been tried utilizing various approaches (Raman spectroscopy, optical probes, oxygen monitoring).[71–74] Here, glucose concentrations inside the bioreactor can be adjusted based on the prediction of specific glucose consumption rates such that glucose levels within the bioreactor are maintained within a given set point. Such monitoring and control strategies have shown promise in reducing product glycation.[75] Given the qualitative accuracy of tracking specific glucose consumption rates, a similar approach may be taken to track specific lactate production/consumption rates, specific ammonia accumulation, growth rates, and specific protein production rate. Additionally, since it was observed that indirect measurements could be learned (qGluc), it can be postulated that by adding total cell density (TCD) prediction into the model, viability (viability = VCD/TCD) can be indirectly monitored throughout the manufacturing process.

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

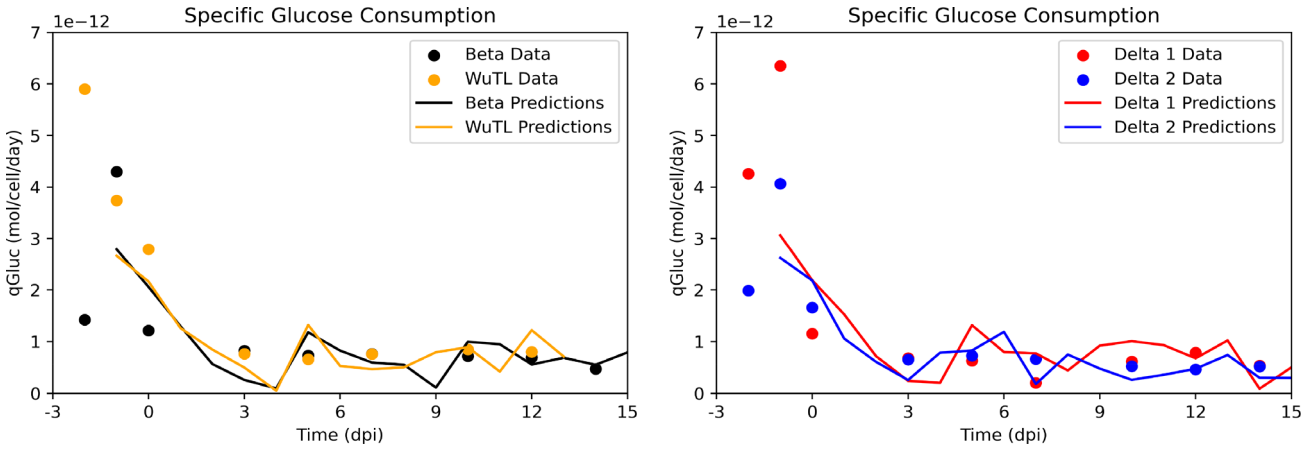

FIGURE 7 Specific glucose consumption rate (qGluc) estimation for Test dataset comprised of 4 cultures (1 Beta pool batch, 1 WuTL pool batch, and 2 Delta pool batches). Continuous lines are everyday predictions while dots are measured values.

It is also worth noting that the online data available in this case study is routine online data such as temperature, pH, DO, oxygen injection, and carbon dioxide addition. Thus, it stands to reason that as more biologically relevant online sensor data is added to the processes and consequently the model, better predictive capabilities can be achieved. Such sensor data that could be utilized is oxygen uptake rates (OUR) and bio-capacitance signals.[1] OUR has been known to relate both to cellular growth and metabolic activity as peak antibody production has been associated with increased oxidative metabolism which, in turn, means an increased oxygen demand.[76–82] Alternatively, bio-capacitance signals have been observed to not only relate to total viable cells within the culture but also to total biovolume as the capacitance signal is dependent on the volume of each cell.[83–85] This is key as cellular diameter has been observed to increase during a culture run[86] and consequently changes in cellular volume are expected. Cell volume also encodes information regarding cellular phase and recombinant protein production activity.[87,88] Given the close relationship between bio-capacitance and cellular volume, further improvements to the model's predictive capabilities could be made. Notably, sensor fusion (biocapacities, OUR and fluorescence data) coupled with machine learning techniques have demonstrated strong monitoring capacity of relevant features such as cellular viability, cellular density, metabolites, and viral yield[89] further underscoring the potential for improved process forecasting with the proposed model when coupled with process analytical technologies (PAT). Unlike mechanistic approaches, the proposed RNN modeling strategy can incorporate dynamically evolving parameters like temperature, pH, dissolved oxygen content, base addition, and gassing profiles to improve feature prediction. These bioreactor values, which are not easily modeled with simple dynamic equations, hold valuable information about cell culture behavior, as the control loops are directly linked to cell culture activity (e.g., lactate accumulation driving base addition through pH drops and oxygen requirements influencing oxygen supplementation via DO drops). Another major advantage of this data-driven method, compared to mechanistic or hybrid models, is that the pre-trained RNN can be readily applied by non-experts, as it doesn't require knowledge of boundary conditions or metabolic networks. This makes it easily transferable to production processes without additional training for operators. Moreover, the RNN's ability to capture internal temporal dependencies within the system enhances feature predictions, offering an advantage over models that assume independent observations for each prediction. However, a key downside of this approach is its reliance on high-quality process data. Its application to new, previously unseen processes or parameter variations not represented in the training set may lead to poor performance. Furthermore, RNN models can be computationally intensive, requiring large datasets and significant training time, which may limit their use in data-scarce environments. Additionally, unlike mechanistic models, RNNs are less interpretable, which can complicate understanding of the underlying processes and limit their use in regulated environments. In contrast, hybrid models, which combine mechanistic insights with data-driven techniques, may offer a more robust and interpretable solution for handling a broader range of process variations. To address the generalization challenge, one approach is to couple the RNN model with a system of differential equations that describe the dynamics of the process. For example, since specific glucose consumption rates can be reasonably predicted, it is plausible that other specific rates, such as specific ammonia production, specific lactate production, and specific protein production, could also be estimated. These specific rates can then be used as parameters in dynamic equations to construct a more robust prediction ensemble. This approach has been explored with data-driven regressors that assume independent observations,[41] and it stands to reason that utilizing models that leverage the temporal dependencies in the data for specific rate predictions could prove to be a worthwhile strategy.

It is important to note that this method requires historical data to train the model effectively. This data can typically be acquired during the process characterization phase of late-stage cell culture processes, particularly when producing novel therapeutics for approval.[90] Scale-down models, which demonstrate behaviors comparable to their GMP manufacturing counterparts, provide a valuable platform for generating large datasets that can be used to train the base model before a

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

fine tuning phase using large-scale GMP bioreactor data; such a transfer learning approach could prove feasible within the biomanufacturing industry framework. Furthermore, within platform processes, it is reasonable to assume that initial model base development can leverage data from similar cell lines or manufacturing processes. This approach allows for the construction of a base model framework, which can then be iteratively finetuned with relevant data as it becomes available. Importantly, to avoid pH sensor drift issues, the model could incorporate both online and offline pH sensor measurements as well as the delta offset between them to build redundancy and avoid the negative impact of possible sensor drift. In data-scarce environments, the use of synthetic data has also shown promise.[52] Mechanistic models can generate culture profiles, which can serve as an initial data source for model training. However, it is worth mentioning that, with the presented approach, online data from gassing profiles, pH, and base values have proven to be valuable in feature prediction. These real-time measurements may be more challenging to simulate alongside cell culture kinetics in mechanistic models, further underscoring the advantage of incorporating actual process data in this data-driven approach.

4 | CONCLUSION

The proposed soft sensor architecture can accurately (nRMSE, nMAE are below unity and $R^2 \geq 0.9$ for all features) predict product titer, total glucose consumption, ammonia, lactate, and viable cell densities, and for a long-term process (17 days) with ground sampling day data only available every other day or every two days to aid in next iteration predictions. To counter the lack of everyday sampling data, daily online data was also utilized into the model. The online data, although harder to directly interpret, still contains relevant information about culture phase (temperature downshifts denote process induced decrease in cellular growth to prime the process for protein production). In terms of PID controllers, it indirectly contains information about oxygen demand (DO, total oxygen sparge) and lactate metabolism (pH, base addition, total carbon dioxide addition). Interestingly, once the same model was applied in a test case where no ground sampling day data was given for titers, glucose consumption, ammonia, and lactate throughout the 17-day culture process, it was determined that the model was still effectively able to soft sense these hard-to-measure features (nRMSE and nMAE below unity and $R^2 \geq 0.9$ for all features). This is especially interesting in processes where the recombinant protein in question can be difficult to measure as is the case of the trimeric SARS-CoV-2 spike protein. In such cases, having a qualitative tracking of feature evolution can be of value. This was possible by considering the bioreactor data that is routinely available in all commercial bioreactor systems. Additionally, qualitative tracking of specific glucose consumption rates (qGluc) was enabled with the proposed method allowing for the possibilities of tight glucose control inside bioreactors by relying on the one day ahead specific glucose consumption rates predictions and adjusting glucose addition to keep overall glucose concentration near a given setpoint. With this knowledge, it stands to reason that the proposed soft sensor can gain from further use of process analytical technologies (PAT) such as off-gas analyzers, bio-capacitance, and Raman spectroscopy signals from which biologically relevant signals can be related to the discretely measured features.

AUTHOR CONTRIBUTIONS

Sebastian Juan Reyes: Data curation; data analysis; methodology; writing-original draft preparation. **Robert Voyer, Yves Durocher**: Supervision; writing–review. **Olivier Henry**: Supervision; writing–review and editing. **Phuong Lan Pham**: Experimental conceptualization; Supervision; writing–review and editing.

ACKNOWLEDGMENTS

This work was funded by the National Research Council of Canada (Pandemic Response Challenge Program (PRCP); 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). The authors gratefully acknowledge Helene L'Ecuyer-Coelho, Yuliya Martynova, and Marjolaine Roy for their contribution in conducting bioreactor production runs. We are also grateful to Simon Lord-Dufour and other members of Mammalian Cell Expression Section of the Human Health Therapeutics Research Center for the generation of the stable cell pools used in this work. The expertise of Raul-Santiago Molina (Proelium S.A.S, Carrera 16 C #153, Bogotá, 110131, Bogotá DC, Colombia) in data analysis and modeling was greatly recognized.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict 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. 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. doi:10.3390/pr10020189
2. Luttmann R, Bracewell DG, Cornelissen G, et al. Soft sensors in bioprocessing: a status report and recommendations. *Biotechnol J*. 2012;7(8):1040-1048. doi:10.1002/biot.201100506
3. Carrondo MJ, Alves PM, Carinhas N, et al. How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology? *Biotechnol J*. 2012;7(12):1522-1529. doi:10.1002/biot.201200226
4. Sommeregger W, Sissolak B, Kandra K, von Stosch M, Mayer M, Striedner G. Quality by control: towards model predictive control of mammalian cell culture bioprocesses. *Biotechnol J*. 2017;12(7):1600546. doi:10.1002/biot.201600546
5. Luo Y, Kurian V, Ogunnaike BA. Bioprocess systems analysis, modeling, estimation, and control. *Curr Opin Chem Eng*. 2021;33:100705. doi:10.1016/j.coche.2021.100705

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

6. Mandenius C-F, Gustavsson R. Mini-review: soft sensors as means for PAT in the manufacture of bio-therapeutics. *J Chem Technol Biotechnol*. 2015;90(2):215-227. doi:10.1002/jctb.4477
7. Randek J, Mandenius CF. On-line soft sensing in upstream bioprocessing. *Crit Rev Biotechnol*. 2018;38(1):106-121. doi:10.1080/07388551.2017.1312271
8. Sha S, Huang Z, Wang Z, Yoon S. Mechanistic modeling and applications for CHO cell culture development and production. *Curr Opin Chem Eng*. 2018;22:54-61. doi:10.1016/j.coche.2018.08.010
9. Bayer B, Duerkop M, Pörtner R, Möller J. Comparison of mechanistic and hybrid modeling approaches for characterization of a CHO cultivation process: requirements, pitfalls and solution paths. *Biotechnol J*. 2023;18(1):2200381. doi:10.1002/biot.202200381
10. Rathore AS, Nikita S, Jesubalan NG. Digitization in bioprocessing: the role of soft sensors in monitoring and control of downstream processing for production of biotherapeutic products. *Biosens Bioelectr*. 2022;12:100263. doi:10.1016/j.biosx.2022.100263
11. Martínez VS, Dietmair S, Quek L-E, Hodson MP, Gray P, Nielsen LK. Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnol Bioeng*. 2013;110(2):660-666. doi:10.1002/bit.24728
12. de Falco B, Giannino F, Carteni F, Mazzoleni S, Kim DH. Metabolic flux analysis: a comprehensive review on sample preparation, analytical techniques, data analysis, computational modelling, and main application areas. *RSC Adv*. 2022;12(39):25528-25548. doi:10.1039/d2ra03326g
13. Martínez-Monge I, Albiol J, Lecina M, et al. Metabolic flux balance analysis during lactate and glucose concomitant consumption in HEK293 cell cultures. *Biotechnol Bioeng*. 2019;116(2):388-404. doi:10.1002/bit.26858
14. Tang P, Xu J, Louey A, et al. Kinetic modeling of Chinese hamster ovary cell culture: factors and principles. *Crit Rev Biotechnol*. 2020;40(2):265-281.
15. Kyriakopoulos S, Ang KS, Lakshmanan M, et al. Kinetic modeling of mammalian cell culture bioprocessing: the quest to advance biomanufacturing. *Biotechnol J*. 2018;13(3):1700229. doi:10.1002/biot.201700229
16. Mahanty B. Hybrid modeling in bioprocess dynamics: structural variabilities, implementation strategies, and practical challenges. *Biotechnol Bioeng*. 2023;120(8):2072-2091. doi:10.1002/bit.28503
17. Zavala-Ortiz DA, Denner A, Aguilar-Uscanga MG, Marc A, Ebel B, Guedon E. Comparison of partial least square, artificial neural network, and support vector regressions for real-time monitoring of CHO cell culture processes using in situ near-infrared spectroscopy. *Biotechnol Bioeng*. 2022;119(2):535-549. doi:10.1002/bit.27997
18. 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.
19. Aehle M, Simutis R, Lübbert A. Comparison of viable cell concentration estimation methods for a mammalian cell cultivation process. *Cytotechnology*. 2010;62(5):413-422. doi:10.1007/s10616-010-9291-z
20. Schmidberger T, Posch C, Sasse A, Gülch C, Huber R. Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling. *Biotechnol Prog*. 2015;31(4):1119-1127. doi:10.1002/btpr.2105
21. Charaniya S, Le H, Rangwala H, et al. Mining manufacturing data for discovery of high productivity process characteristics. *J Biotechnol*. 2010;147(3):186-197. doi:10.1016/j.jbiotec.2010.04.005
22. Mondal PP, Galodha A, Verma VK, et al. Review on machine learning-based bioprocess optimization, monitoring, and control systems. *Bioresour Technol*. 2023;370:128523. doi:10.1016/j.biortech.2022.128523
23. Rathore AS, Nikita S, Thakur G, Mishra S. Artificial intelligence and machine learning applications in biopharmaceutical manufacturing. *Trends Biotechnol*. 2023;41(4):497-510. doi:10.1016/j.tibtech.2022.08.007
24. Kotidis P, Kontoravdi C. Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab Eng Commun*. 2020;10:e00131. doi:10.1016/j.mec.2020.e00131
25. Baako T-MD, Kulkarni SK, McClendon JL, Harcum SW, Gilmore J. Machine learning and deep learning strategies for Chinese hamster ovary cell bioprocess optimization. *Fermentation*. 2024;10(5):234.
26. Walsh I, Myint M, Nguyen-Khuong T, Ho YS, Ng SK, Lakshmanan M. Harnessing the potential of machine learning for advancing "quality by design" in biomanufacturing. *MAbs*. 2022;14(1):2013593.
27. Shiri TJ, Viau C, Gu X, Xu L, Lu Y, Xia JJM. The native microbiome member *Chryseobacterium* sp. CHNTR56 MYb120 induces Trehalose production via a shift in central carbon metabolism during early life in *C. elegans*. 2023;13(8):953.
28. Poth M, Magill G, Filgertshofer A, Popp O, Großkopf T. Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing. *J Raman Spectrosc*. 2022;53(9):1580-1591. doi:10.1002/jrs.6402
29. Konakovsky V, Yagtu AC, Clemens C, et al. Universal capacitance model for real-time biomass in cell culture. *Sensors (Basel)*. 2015;15(9):22128-22150. doi:10.3390/s150922128
30. 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(9):2494-2508. doi:10.1002/bit.28405
31. Wong WC, Chee E, Li J, Wang X. Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. *Mathematics*. 2018;6(11):242.
32. Iglesias JC, Mehta V, Venereo-Sanchez A, et al. Handling massive proportion of missing labels in multivariate long-term time series forecasting. *IOP Publishing*. 2021;2090(1):012170. doi:10.1088/1742-6596/2090/1/012170
33. Smiatek J, Clemens C, Herrera LM, et al. Generic and specific recurrent neural network models: applications for large and small scale biopharmaceutical upstream processes. *Biotechnol Rep*. 2021;31:e00640.
34. Rogers AW, Song Z, Ramon FV, Jing K, Zhang D. Investigating 'greyness' of hybrid model for bioprocess predictive modelling. *Biochem Eng J*. 2023;190:108761. doi:10.1016/j.bej.2022.108761
35. Mowbray MR, Wu C, Rogers AW, Rio-Chanona EAD, Zhang D. A reinforcement learning-based hybrid modeling framework for bioprocess kinetics identification. *Biotechnol Bioeng*. 2023;120(1):154-168. doi:10.1002/bit.28262
36. Cui T, Bertalan T, Ndahiro N, et al. Data-driven and physics informed modelling of Chinese hamster ovary cell bioreactors. *Comput Chem Eng*. 2024;183:108594. doi:10.1016/j.compchemeng.2024.108594
37. Pinto J, Mestre M, Ramos J, Costa RS, Striedner G, Oliveira R. A general deep hybrid model for bioreactor systems: combining first principles with deep neural networks. *Comput Chem Eng*. 2022;165:107952. doi:10.1016/j.compchemeng.2022.107952
38. Agharafeie R, Oliveira R, Ramos JRC, Mendes JM. Application of hybrid neural models to bioprocesses: a systematic literature review. *Authorea Preprints*. 2023;9(10):922.
39. Maton M, Bogaerts P, Vande Wouwer A. Hybrid dynamic models of bioprocesses based on elementary flux modes and multilayer Perceptrons. *Processes*. 2022;10(10):2084.
40. Narayanan H, von Stosch M, Feidl F, Sokolov M, Morbidelli M, Butté A. Hybrid modeling for biopharmaceutical processes: advantages, opportunities, and implementation. *Front Chem Eng*. 2023;5:1157889. doi:10.3389/fceng.2023.1157889
41. Yatipanthalawa BS, Fitzsimons SEW, Horning T, Lee YY, Gras SL. Development and validation of a hybrid model for prediction of viable cell density, titer and cumulative glucose consumption in a

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

mammalian cell culture system. *Comput Chem Eng*. 2024;184:108648. doi:10.1016/j.compchemeng.2024.108648
42. Solle D, Hitzmann B, Herwig C, et al. Between the poles of data-driven and mechanistic modeling for process operation. *Chem Ing Tech*. 2017;89(5):542-561. doi:10.1002/cite.201600175
43. Tuveri A, Pérez-García F, Lira-Parada PA, Imsland L, Bar N. Sensor fusion based on extended and unscented Kalman filter for bioprocess monitoring. *J Process Control*. 2021;106:195-207. doi:10.1016/j.jprocont.2021.09.005
44. Simutis R, Lübbert A. Hybrid approach to state estimation for bioprocess control. *Bioengineering*. 2017;4(1):21.
45. Khodarahmi M, Maihami V. A Review on Kalman Filter Models. *Arch Comput Methods Eng*. 2023;30(1):727-747. doi:10.1007/s11831-022-09815-7
46. Iglesias CF Jr, Xu X, Mehta V, et al. Monitoring the recombinant adeno-associated virus production using extended Kalman filter. *Processes*. 2022;10(11):2180.
47. Narayanan H, Behle L, Luna MF, et al. Hybrid-ekf: hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. *Biotechnol Bioeng*. 2020;117(9):2703-2714.
48. Tsopanoglou A, del Val IJ. Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. *Curr Opin Chem Eng*. 2021;32:100691.
49. Narayanan H, Sokolov M, Morbidelli M, Butté A. A new generation of predictive models: the added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnol Bioeng*. 2019;116(10):2540-2549.
50. Ohadi K, Legge RL, Budman HM. Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures. *Biotechnol Bioeng*. 2015;112(1):197-208.
51. Polak J, Huang Z, Sokolov M, et al. An innovative hybrid modeling approach for simultaneous prediction of cell culture process dynamics and product quality. *Biotechnol J*. 2024;19(3):2300473.
52. Iglesias CF Jr, Ristovski M, Bolic M, Cuperlovic-Culf M. rAAV manufacturing: the challenges of soft sensing during upstream processing. *Bioengineering*. 2023;10(2):229.
53. Zeng A, Chen M, Zhang L, Xu Q. Are transformers effective for time series forecasting? *Proc AAAI Conf Artif Intell*. 2023;37(9):11121-11128. doi:10.1609/aaai.v37i9.26317
54. Bilokon P, Qiu YJ. Transformers versus LSTMs for electronic trading. *arXiv*. 2023.
55. 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
56. 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.
57. 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
58. Reyes SJ, Pham PL, Durocher Y, Henry O. CHO stable pool fed-batch process development of SARS-CoV-2 spike protein production: impact of aeration conditions and feeding strategies. *Biotechnol Prog*. 2025;41(1):e3507. doi:10.1002/btpr.3507
59. Reyes SJ, Lemire L, Molina RS, 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;40:e3467.
60. 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.
61. DiPietro R, Hager GD. Chapter 21 - deep learning: RNNs and LSTM. In: Zhou SK, Rueckert D, Fichtinger G, eds. *Handbook of Medical Image Computing and Computer Assisted Intervention*. Academic Press; 2020:503-519.
62. Graves A, Graves A. *Supervised Sequence Labelling*. Springer; 2012.
63. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT press; 2016.
64. McKinney W. *Data Structures for Statistical Computing in Python*. Austin; 2010:51-56.
65. Harris CR, Millman KJ, Van Der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362.
66. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Proces Syst*. 2019;32.
67. Jiang Z, Sharfstein ST. Sodium butyrate stimulates monoclonal antibody over-expression in CHO cells by improving gene accessibility. *Biotechnol Bioeng*. 2008;100(1):189-194. doi:10.1002/bit.21726
68. Park JH, Noh SM, Woo JR, Kim JW, Lee GM. Valeric acid induces cell cycle arrest at G1 phase in CHO cell cultures and improves recombinant antibody productivity. *Biotechnol J*. 2016;11(4):487-496. doi:10.1002/biot.201500327
69. López-Meza J, Araíz-Hernández D, Carrillo-Cocom LM, López-Pacheco F, Rocha-Pizaña MR, Alvarez MM. Using simple models to describe the kinetics of growth, glucose consumption, and monoclonal antibody formation in naive and infliximab producer CHO cells. *Cytotechnology*. 2016;68(4):1287-1300. doi:10.1007/s10616-015-9889-2
70. Meuwly F, Papp F, Ruffieux PA, Bernard AR, Kadouri A, von Stockar U. Use of glucose consumption rate (GCR) as a tool to monitor and control animal cell production processes in packed-bed bioreactors. *J Biotechnol*. 2006;122(1):122-129. doi:10.1016/j.jbiotec.2005.08.005
71. Goldrick S, Lee K, Spencer C, et al. On-line control of glucose concentration in high-yielding mammalian cell cultures enabled through oxygen transfer rate measurements. *Biotechnol J*. 2018;13(4):e1700607. doi:10.1002/biot.201700607
72. Kozma B, Hirsch E, Gergely S, Párta L, Pataki H, Salgó A. On-line prediction of the glucose concentration of CHO cell cultivations by NIR and Raman spectroscopy: comparative scalability test with a shake flask model system. *J Pharm Biomed Anal*. 2017;145:346-355. doi:10.1016/j.jpba.2017.06.070
73. Lederle M, Tric M, Roth T, et al. Continuous optical in-line glucose monitoring and control in CHO cultures contributes to enhanced metabolic efficiency while maintaining darbepoetin alfa product quality. *Biotechnol J*. 2021;16(8):e2100088. doi:10.1002/biot.202100088
74. Rashedi M, Rafiei M, Demers M, et al. Machine learning-based model predictive controller design for cell culture processes. *Biotechnol Bioeng*. 2023;120(8):2144-2159. doi:10.1002/bit.28486
75. Gibbons L, Maslanka F, Le N, et al. An assessment of the impact of Raman based glucose feedback control on CHO cell bioreactor process development. *Biotechnol Prog*. 2023;39(5):e3371. doi:10.1002/btpr.3371
76. Templeton N, Dean J, Reddy P, Young JD. Peak antibody production is associated with increased oxidative metabolism in an industrially relevant fed-batch CHO cell culture. *Biotechnol Bioeng*. 2013;110(7):2013-2024. doi:10.1002/bit.24858
77. Zalai D, Hevér H, Lovász K, et al. A control strategy to investigate the relationship between specific productivity and high-mannose glycoforms in CHO cells. *Appl Microbiol Biotechnol*. 2016;100(16):7011-7024. doi:10.1007/s00253-016-7380-4
78. Huang Y-M, 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

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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

equipment. *Biotechnol Prog*. 2010;26(5):1400-1410. doi:10.1002/btpr.436
79. 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.
80. Reyes S-J, Lemire L, Durocher Y, Voyer R, Henry O, Pham PL. Investigating the metabolic load of monoclonal antibody production conveyed to an inducible CHO cell line using a transfer-rate online monitoring system. *J Biotechnol*. 2025;399:47-62. doi:10.1016/j.jbiotec.2025.01.008
81. Lemire L, Reyes SJ, Durocher Y, Voyer R, Henry O, Pham PLJBP. N-1 semi-continuous transient perfusion in shake flask for ultra-high density seeding of CHO cell cultures in benchtop bioreactors. *Biotechnol Prog*. 2025;e70029.
82. Sebastian Reyes Davila J, Lan Pham P, Durocher Y, Henry O. CHO stable pool fed-batch process development of SARS-CoV-2 spike protein production. 2023.
83. 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
84. 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
85. Moore B, Sanford R, Zhang A. Case study: the characterization and implementation of dielectric spectroscopy (biocapacitance) for process control in a commercial GMP CHO manufacturing process. *Biotechnol Prog*. 2019;35(3):e2782. doi:10.1002/btpr.2782
86. Wang X, Zhou G, Liang L, et al. Deep learning-based image analysis for in situ microscopic imaging of cell culture process. *Eng Appl Artif Intell*. 2024;129:107621. doi:10.1016/j.engappai.2023.107621
87. 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
88. 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
89. Xu X, Farnós O, Paes BC, Nesdoly S, Kamen AA. Multivariate data analysis on multisensor measurement for inline process monitoring of adenovirus production in HEK293 cells. *Biotechnol Bioeng*. 2024;121:2175–2192.
90. Xu J, Ou J, McHugh KP, Borys MC, Khetan A. Upstream cell culture process characterization and in-process control strategy development at pandemic speed. *MAbs*. 2022;14(1):2060724. doi:10.1080/19420862.2022.2060724

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, Voyer R, Durocher Y, Henry O, Pham PL. A recurrent neural network for soft sensor development using CHO stable pools in fed-batch process for SARS-CoV-2 spike protein production as a vaccine antigen. *Biotechnol. Prog*. 2025;41(5):e70046. doi:10.1002/btpr.70046

15206033, 2025, 5, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.70046 by Ecole Polytech De Montreal, Wiley Online Library on [17/11/2025]. 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