TIME TO INNOVATE

ON-DEMAND Webinars

L'Oréal: Design experiments to be robust to input variation

Victor Guiller
Scientific Expertise Engineer, L'Oréal

Video

JMP STATISTICAL DISCOVERY

Received: 10 May 2022 | Revised: 25 August 2022 | Accepted: 6 September 2022

DOI: 10.1002/biot.202200237

Biotechnology Journal

RESEARCH ARTICLE

Hierarchical time-series analysis of dynamic bioprocess systems

Masoumeh Alinaghi[1] | **Izabella Surowiec**[1] | **Steffi Scholze**[2] | **Chris McCready**[3] | **Christoph Zehe**[4] | **Erik Johansson**[5] | **Johan Trygg**[1,6] | **Olivier Cloarec**[1]

[1]Sartorius Corporate Research, Sartorius, Sartorius Stedim Data Analytics, Umeå, Sweden

[2]Sartorius Stedim Biotech GmbH, Göttingen, Germany

[3]Sartorius Corporate Research, Oakville, Ontario, Canada

[4]Sartorius Corporate Research, Sartorius Stedim Cellca GmbH, Ulm, Germany

[5]Sartorius Stedim Data Analytics, Umeå, Sweden

[6]Computational Life Science Cluster (CLiC), Umeå University, Sweden

Correspondence
Olivier Cloarec, Sartorius Corporate Research, Sartorius Stedim Cellca GmbH, 89081 Ulm, Germany.
Email: olivier.cloarec@sartorius.com

Abstract

Background: Monoclonal antibodies (mAbs) are leading types of 'blockbuster' biotherapeutics worldwide; they have been successfully used to treat various cancers and chronic inflammatory and autoimmune diseases. Biotherapeutics process development and manufacturing are complicated due to lack of understanding the factors that impact cell productivity and product quality attributes. Understanding complex interactions between cells, media, and process parameters on the molecular level is essential to bring biomanufacturing to the next level. This can be achieved by analyzing cell culture metabolic levels connected to vital process parameters like viable cell density (VCD). However, VCD and metabolic profiles are dynamic parameters and inherently correlated with time, leading to a significant correlation without actual causality. Many time-series methods deal with such issues. However, with metabolic profiling, the number of measured variables vastly exceeds the number of experiments, making most of existing methods ill-suited and hard to interpret.

Methods and Major Results: Here we propose an alternative workflow using hierarchical dimension reduction to visualize and interpret the relation between evolution of metabolic profiles and dynamic process parameters. The first step of proposed method is focused on finding predictive relation between metabolic profiles and process parameter at all time points using OPLS regression. For each time point, the p(corr) obtained from OPLS model is considered as a differential metabogram and is further assessed using principal components analysis (PCA).

Conclusions: Compared to traditional batch modeling, applying proposed methodology on metabolic data from Chinese Hamster Ovary (CHO) antibody production characterized the dynamic relation between metabolic profiles and critical process parameters.

Abbreviations: BEM, batch evolution model; BLM, batch level model; CCF, central composite face-centered; CHO, Chinese Hamster Ovary; CPPs, critical process parameters; CQAs, critical quality attributes; CV-ANOVA, ANOVA of cross-validated residuals; DOEs, design of experiments; ^{1}H-NMR, ^{1}H-nuclear magnetic resonance; IVCC, integral viable cell concentration; LOD, limit of detection; LOQ, limit of quantification; mAbs, monoclonal antibodies; MVDA, multivariate data analysis; OPLS, orthogonal projections to latent structures; OPLS-DA, orthogonal partial least squares-discriminant analysis; PCs, principal components; PCA, principal components analysis; PLS, partial least squares; QbD, quality by design; SUS-plot, shared and unique structure plot; UV, unit variance; VCC, viable cell concentration; VCD, viable cell density.

KEYWORDS
bioprocess data, dynamic system, hierarchical analysis, meta-analysis, metabolomics data, time-series analysis, viable cell density (VCD)

1 | INTRODUCTION

Biologics are pharmaceutical drug products manufactured from biological sources, mainly living cells. Of these, Chinese Hamster ovary (CHO) cells are the most used mammalian[1] system in industrial bioprocesses to produce therapeutic proteins[2] such as monoclonal antibodies.[3] Their development and production is an inherently complex process, not only because mammalian cells are extremely complicated systems, but also due to the lack of understanding of the factors impacting cell productivity and product quality.[4] CPPs are process parameters whose variability has an impact on the product's critical quality attributes (CQAs) or general process performance, like the amount of produced drug (its titer), cost, and time of the process, and so on. Therefore, CPP monitoring, and controlling is a prerequisite to ensure that the biopharmaceutical process will efficiently produce products with the desired quality.[5] Enhanced characterization of the process at the molecular level aligns with the quality by design (QbD) concept which requires understanding the influence and causality of factors and parameters, such as temperature and pH,[6] to identify CQAs of the product.[7,8] The small-scale bioreactor systems designed for parallel operation, like the Ambr technology (high throughput High Performance Single-Use Fully Automated Bioreactor System), combined with new types of design of experiments (DOEs) have the potential to enable the rapid process characterization and development at a fraction of cost with higher throughput compared to traditional bench top reactors, thereby drastically shortening both time and costs. Extraction of knowledge from complex data obtained from such systems is, however, still a challenge that calls for the development of new methods and pipelines.

Metabolomics together with process monitoring can be used to decipher complex interactions between cells, media, and process parameters, as it provides knowledge of biochemical reactions and metabolic pathways occurring within living cells. It can be used to characterize the impact of small molecules (i.e., amino acids and metabolites) present in cell culture media on CHO cell growth behavior and productivity to allow identifying, understanding, and manipulating the production and consumption of metabolites of interest to enhance cell productivity and final product quality.[9] In metabolomics studies, methods such as ^{1}H-nuclear magnetic resonance (^{1}H-NMR) spectroscopy can be used as a non-invasive and rapid technique for simultaneous detection and quantification of metabolites in a complex biological fluid. Metabolomics data are highly complex and multidimensional and thus require specific and tailored approaches to be analyzed properly.

Even though multidimensional time-series data obtained from the biopharmaceutical process contain valuable information on the dynamic characteristics of the bioprocess, and thus could potentially assist increasing process knowledge,[10] little has been accomplished to improve knowledge through efficient data analytical methods. Although several data analysis methods,[11–22] such as principal components analysis (PCA)[23] are discussed in the literature for multivariate data exploration,[24–29] development of a tailored workflow that is adjusted to bioprocess time-series of different biomanufacturing systems is required, especially with metabolomics data where the number of observations can be lower than the number of highly correlated variables.

PCA is frequently used to explore the main sources of variations in metabolomics data when the observations are assumed independent.[30] However, this is not the case with time-series data with additional correlation structure because of non-independent observations over process time. Indeed, PCA inspects for directions in the data space with maximum variation, and therefore time-related variation is involved as a confounding factor hindering finding metabolites significantly correlated to the process variables. Hence, proper models are required to model the bioprocess data and to extract vital information.

To properly analyze time-series metabolomics data, some PCA extensions have been proposed. The weighted PCA model[31] is one of the PCA-extended models that uses weights to explain variation due to repeated measurements. Moreover, the analysis-of-variance–simultaneous component analysis[32] combines analysis of variance and simultaneous components analysis to address the timely designed multivariate data which can also be applied for integrative analysis.[33] Local PCA models[34] were also proposed to be employed for each time point which subsequently linked to each other. The dynamic PCA model[35] employs a back-shift matrix to simultaneously analyze the data from multiple time points. However, the difficulty in uncertainty assessment of the fitted model estimates is the main limitation of such methods. Mixed-effects models,[36] for instance, linear mixed-effects model[37] have also been applied for feature selection in time-series metabolomics data, however, the correlation structure of the metabolomics data was not considered in the model. On the other hand, dimension reduction features of methods like PCA are still appealing.

The primary goal of bioprocess data models, such as batch evolution model (BEM) and batch level model (BLM),[38] is to reduce the dimensionality of data and discover the dynamics unfolding over time. BEM models can then be used to forecast future process performance and to evaluate how variations in each input parameter influence the batch trajectory. Whereas averaged trajectories of different process parameters (e.g. media composition) can be tracked well with time-BEM models, pinpointing significant metabolites that change with time, and have a significant connection to the specific process parameters (like viable cell concentration (VCC), viable cell density (VCD), titer,

etc.) is a more difficult task, mainly because all process parameters are to some extent correlated with time. Therefore, it is interesting to reduce the dimensionality of the data and at the same time highlight significant metabolites which change over time and are also correlated to the process parameters.

In this work, we provide a fast overview of the most used methods for modeling batch process data and then propose a new two-step hierarchical modeling approach specifically targeted to model important process parameters such as VCD. The ^{1}H-NMR based metabolomics data from the antibody production process by CHO cells were used to represent the performance of the proposed approach. We managed to successfully evaluate correlations between metabolite levels and important bioprocess parameters that change over time to understand the underlying dynamic behavior of biological systems.

2 | MATERIAL AND METHOD

2.1 | Cell line and culture conditions

Cultivations were performed in a 0.25-L bioreactor (Ambr 250 High Throughput bioreactors, Sartorius Stedim Biotech GmbH, Göttingen, Germany) and a CHO cell line (Sartorius) expressing a monoclonal antibody (mAb, IgG1) was used. All experiments were carried out using chemically defined media (Sartorius).[39]

To maximize the beneficial impact of the culture parameter change to control the cell proliferation, bioreactor cultures were maintained at 36.8 °C, pH = 7.1 until day 7 (~170 h) to reach a suitable cell density under the standard condition.[40,41] Different experimental conditions (pH = 6.9, 7.1, 7.3, and temperature = 31, 33.9, 36.8 °C) were implemented in bioreactors after day 7 to investigate the impact of temperature and pH by central composite face-centered (CCF) experimental design. The replicated condition was implemented for the standard temperature (36.8 °C, pH = 7.1) in the experimental design (Figure 1).

Feeding was started on day 3 at 14:00 (~80 h) and fed daily thereafter. The changes in pH and temperature were induced on day 7 before feeding (before the start of the stationary phase) at approximately 170 h. Samples were drawn from the bioreactors several times per day to measure the metabolites by NMR spectroscopy as well as the BioProfile FLEX2 and titer data by HPLC (SEC) (Figure 1).

2.2 | Sample preparation and offline measurement/analytics

Samples were taken automatically from the Ambr250 high throughput bioreactors to separate vials. VCC, viability, average cell diameter, pH, pO_2, pCO_2, as well as metabolite levels (glucose, lactate, ammonia, and glutamine) and osmolality were measured using BioProfile FLEX2. The remaining sample volume (1 mL) was centrifuged (16,600 $\times g$, RT = 23°C, 5 min) and the supernatant was stored at −20°C for subsequent analytics.

2.3 | Determination of the cell-specific productivity (Qp)

The specific growth rate (μ (1/d)) can be calculated by Eq. (1).[42]

$$\mu = \frac{\ln x_2 - \ln x_1}{t_2 - t_1} \quad (1)$$

Here, x_2 and x_1 are the VCCs at time points t_2 and t_1, respectively.

Cell specific productivity (pg cell^{-1} day^{-1}) (Qp) was calculated by Eq. (2).

$$Qp = \frac{\Delta z}{\Delta IVCC} \quad (2)$$

Where Δz is the differences of the titer and $\Delta IVCC$ is the corresponding integral viable cell concentration (IVCC). Growth phases (i.e., exponential, stationary and dead phases) and productivity phases can be determined based on the changes in the μ and Qp values, respectively, as a sharp change can be evaluated as a new phase.

2.4 | Experimental data (NMR metabolomics data)

Samples were analyzed for metabolite content by Eurofins Spinnovation Analytical BV (The Netherlands), using their proprietary protocols (Spedia-NMR). Obtained data were concentrations for 39 identified metabolites in mM. The identified metabolites were mainly amino acids, vitamins, saccharides, and organic acids.

2.5 | Data preprocessing

Before analysis, one variable, which had missing values higher than 75% of the number of samples, was removed from the analysis. The values lower than the limit of quantification (LOQ) were also included in the analysis. The values lower than the limit of detection (LOD) were not reported from the experimental quantification. Therefore, in order to include these values, they were simulated as log-normally distributed values below the LOD. Unit variance (UV) scaling has been applied to remove differences in the range between variables.

3 | THEORY AND METHODS

Multivariate data analysis techniques such as PCA, PLS, orthogonal projections to latent structures (OPLS), and orthogonal PLS-discriminant analysis (OPLS-DA) can be used for the monitoring of bioprocess behavior to distinguish between systematic information and unsystematic information.[11–22] In this section, first, we will shortly introduce the BEM and BLM models and discuss their properties and then explain the proposed hierarchical modeling for analysis of the bioprocess data. Multivariate data analysis (MVDA) using PCA and OPLS was carried out using SIMCA (Sartorius Stedim Data Analytics, vs. 16.0.2, Umeå, Sweden) software. All data were scaled to UV and

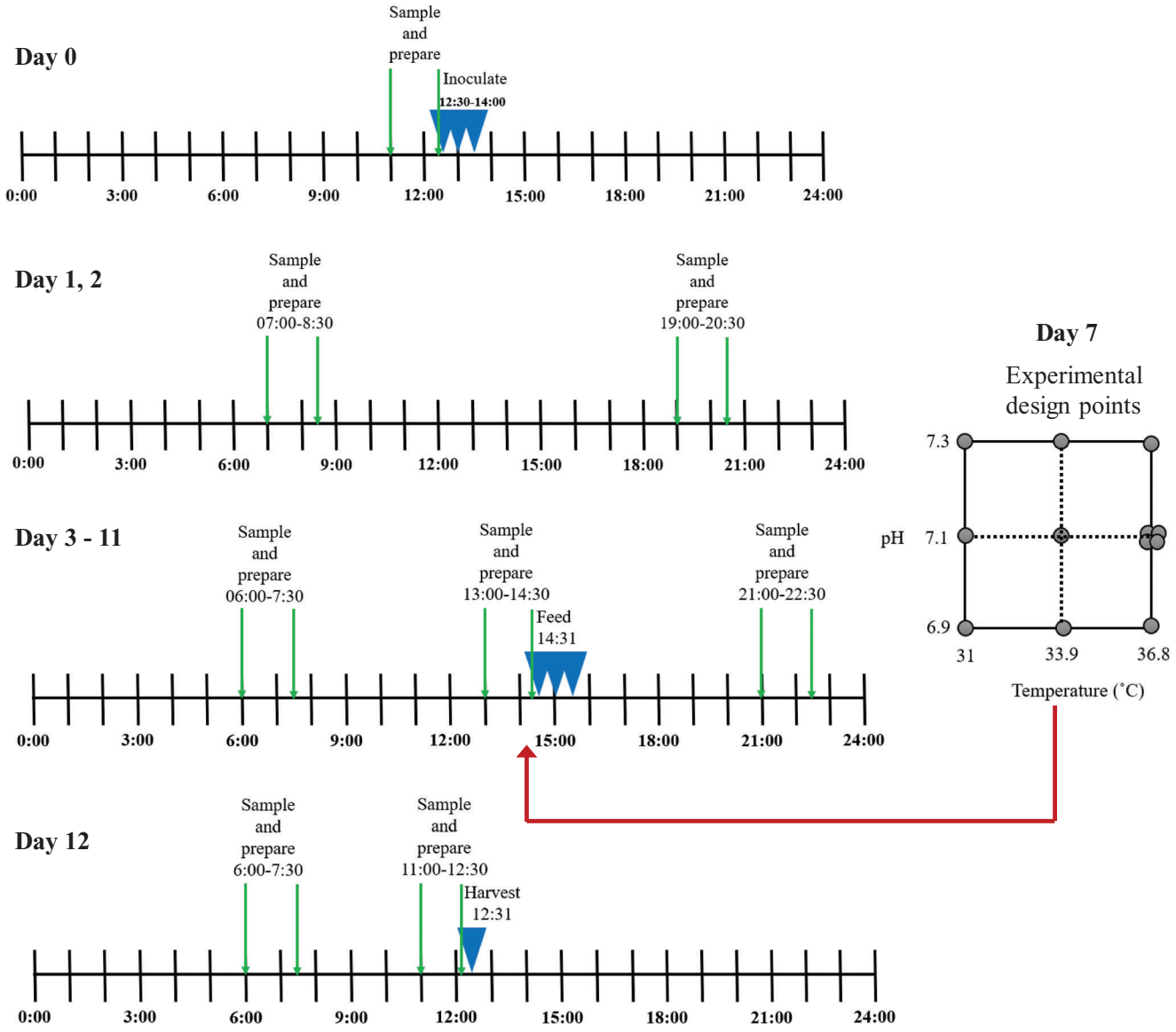

FIGURE 1 Illustration of the feeding and sampling timestamp of the fed-batch CHO cell cultivation process for 12 days (left figure), as well as the design points for 2-variable central composite face-centered (CCF) on day 7 of the process (right figure)

mean-centered prior to modeling. Data cleaning and visualization for hierarchical modeling were performed using Python libraries Pandas v.1.3.4 and Matplotlib v.3.4.3.[43]

3.1 | Batch evolution model (BEM)

The BEM is used to estimate the maturity of a process from its measured parameters and set points. It is also used to detect variations of CPPs from their ideal trajectory in relation to maturity.[38] To accomplish BEM, the three-way batch (i.e., *K* batches, *T* time points, and *J* variables) data should be unfolded to a two-way data such that the variable mode is preserved. It will result in a matrix with $K * T$ rows (observations) and *J* columns (variables). Batch age/time or any other monotonous batch maturity index can be included as a y-variable in the PLS/OPLS model. To establish the control charts, BEM scores can be rearranged in a way that the score values of each batch are as row vectors underneath each other. By computing the average and the standard deviation of score vectors of all batches at different time points, the control limits can be derived as the average score ± 3 SD. Loading plots represent the variables that are responsible for the metabolite changes over time.

3.2 | Batch level model (BLM)

In contrast to BEM data, where each row is corresponding to one time point in a specific batch, each row of the BLM data is obtained by

unfolding the three-way batch along the time axis, one row represents the whole batch including all variables at all time points. Multivariate methods such as PCA, PLS/OPLS can then be used to provide the overall pattern in all the batches in relation to experimental design or any other batch related time-independent parameters.[38]

3.3 | Shared and unique structure plot (SUS-plot)

Comparison between two models can be performed using a shared and unique structure plot (SUS-plot), in which the p(corr) of the predictive part for two OPLS models are plotted against each other. It can reveal differences and similarities in the metabolite contribution between models.[44] The SUS-plot can be interpreted based upon variable location in the plot. For identical profiles, the SUS-plot should have the points on the diagonal line, while off-diagonal points show variables with different effects in two models. Variables with unique effects can be found as points close to the X- or Y-axis away from the origin.

3.4 | Hierarchical modeling

A hierarchical data analytic workflow has been developed to solve the confounding influence of time on the relation between dynamic process parameters and metabolic profiles. The first step of the hierarchical model is to capture the relation between the metabolic profile and the parameter of interest such as VCD at each time point using OPLS regression. The second step, that is, PCA, here can be considered as an extension of the SUS methodology when more than two profiles can be compared (Figure 2).

3.4.1 | Base level OPLS model

OPLS is an extension to the PLS regression method that divides the systematic variation in the **X** data matrix into two parts which makes the results easier to interpret. One part models the correlation between **X** and **Y** and the other part expresses the systematic variation in **X** that is orthogonal to **Y**. In the OPLS model, the component related to **Y** are called predictive which are appointed with the subscript P. Components unrelated to **Y** are called orthogonal and are appointed with the subscript O. Suppose $\mathbf{X}_t$ holds the number of rows K and the number of columns J, then OPLS can decompose the data set at each time point t according to Eq. (3). For this step of the model, each line of the **X** matrix is the metabolite profile corresponding to one cultivation and the **y**-variable is the corresponding VCD profile.

$$\mathbf{X}_t = \mathbf{T}_{Pt}\mathbf{P}_{Pt}^{\mathrm{T}} + \mathbf{T}_{Ot}\mathbf{P}_{Ot}^{\mathrm{T}} + \mathbf{E}_t \tag{3}$$

Here, $\mathbf{X}_t \in \mathrm{R}^{K\times J_t}$, where $t = 1, \ldots, T$ corresponds to the various multivariate data sets at different time points, whose rows contain the measurement of objects $k = 1, \ldots, K$ and columns contain the measurement of variables $j_t = 1, \ldots, J_t$. $\mathbf{T}_{Pt} \in \mathrm{R}^{K\times R}$ is the score matrix related to **Y** with $r = 1$ number of significant components and $\mathbf{P}_{Pt} \in \mathrm{R}^{J_t\times R}$contains the relevant loadings related to **Y** for each data $\mathbf{X}_t$. $\mathbf{T}_{Ot} \in \mathrm{R}^{K\times R}$ is the score matrix unrelated to **Y** with $r = 1, \ldots, R$ number of significant components and $\mathbf{P}_{Ot} \in \mathrm{R}^{J_t\times R}$contains the relevant loadings unrelated to **Y** for each data $\mathbf{X}_t$. $\mathbf{E}_t \in \mathrm{R}^{K\times J_t}$ is the residual matrix.

With a single y-variable, only one predictive component is feasible, but possibly several orthogonal components. The p(corr) is desirable as the input of the second step in the model, that is, top level PCA model. The p(corr) is a correlation loading profile of the OPLS model. The X loading weight w, which combines the X variables to form the scores t, can also be used as the input of the second step in the model.

In this model, the p(corr) from each OPLS model of $t = 1, \ldots, T$ is called differential metabogram which can be augmented column-wise to build the **Z** data matrix according to Eq. (4).

$$\mathbf{Z} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ . \\ . \\ . \\ \mathbf{s}_t \end{bmatrix} \tag{4}$$

Here, $\mathbf{s}_t \in \mathrm{R}^{K\times R}$ is the p(corr) and $\mathbf{Z} \in \mathrm{R}^{K\times(t\times J)}$ is the matrix of differential metabograms, where $t = 1, \ldots, T$ corresponds to the various multivariate data sets at different time points, whose rows contain the measurement of objects $k = 1, \ldots, K$ and columns contain the measurement of variables $j_t = 1, \ldots, J_t$.

To validate the models against overfitting, the number of model components was evaluated by the 7-fold cross-validation parameter Q^2 (goodness of prediction). Moreover, the performance of models was also evaluated by R^2X and R^2Y (goodness of fit). The statistical significance of the OPLS models was assessed by permutation tests with 10^3 repetitions to compare the fit of the original model with the randomly permuted models.[45]

3.4.2 | Top level PCA model

The second step of hierarchical modeling is an analysis by PCA on the matrix of differential metabograms calculated from the OPLS models. PCA can reduce multivariate data to a smaller number of uncorrelated variables, that is, the scores of the PCs, and provide an overview of the bioprocess data from all cultivations at all time points. Suppose **Z** contains the number of rows T and the number of columns J, then PCA can decompose the data set according to Eq. (5).

$$\mathbf{Z} = \mathbf{T}\mathbf{P}^T + \mathbf{F} \tag{5}$$

Here $\mathbf{Z} \in \mathrm{R}^{T\times J}$, whose rows contain the OPLS p(corr) for different time points $t = 1, \ldots, T$ and columns contain the measurement of variables $j = 1, \ldots, J$. $\mathbf{T} \in \mathrm{R}^{T\times R}$ is the PCA score matrix with $r = 1, \ldots, R$

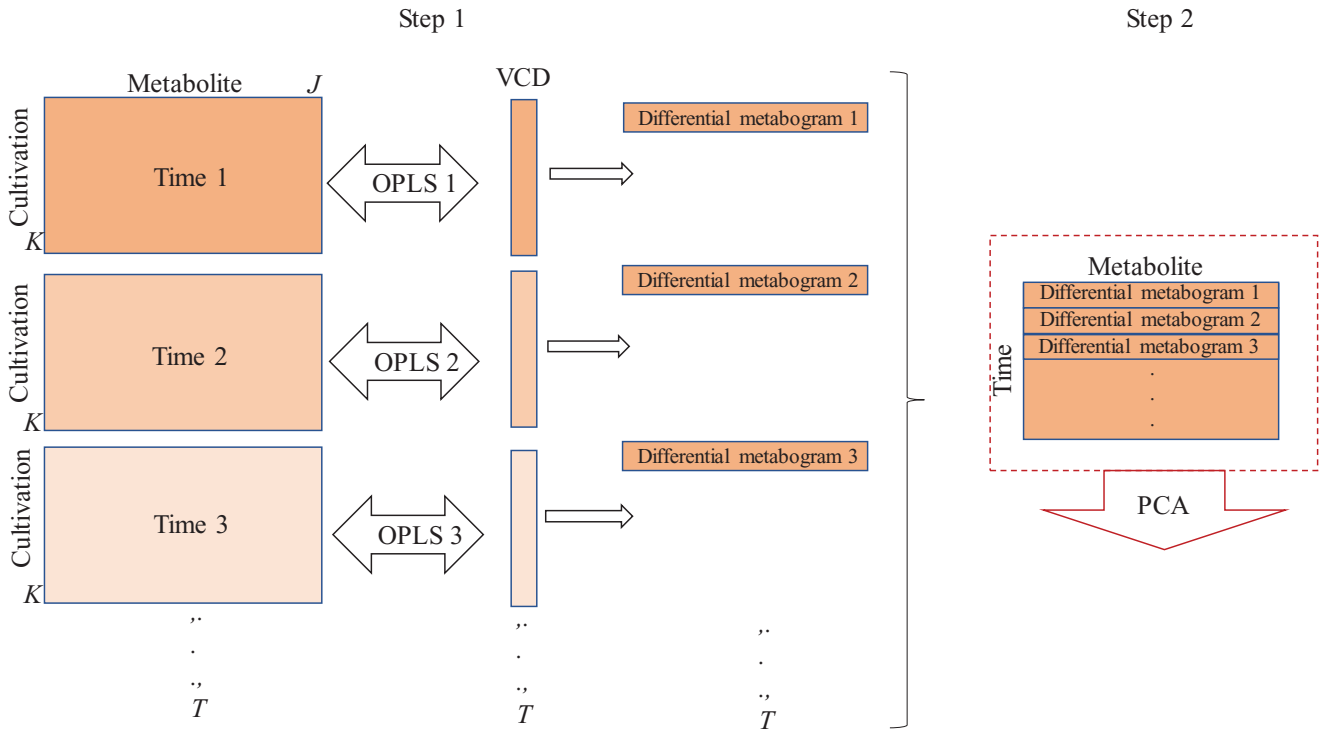

FIGURE 2 Schematic representation of the proposed hierarchical modeling. The OPLS regression model in the first step of the hierarchical model captures the relation between the metabolic profile and the parameter of interest (here: VCD) at each time point. PCA in the second step reduces the multivariate data to a smaller number of uncorrelated variables and provides an overview of bioprocess data from all cultivations at all time points

number of significant components and $\mathbf{P} \in R^{J \times R}$ contains the relevant loadings for each data **Z**. $\mathbf{F} \in R^{T \times J}$ is the residual matrix.

4 | RESULTS AND DISCUSSION

4.1 | Bioprocess parameters

The data derived from twelve fed-batch cultivations were used for the analysis. Figure 3A – D shows the time trajectories for the process variables of VCD (10^6 cell.mL^{-1}), viability (%), average live cell diameter (μm), IgG titer (g.L^{-1}) in all analyzed batches. Figure 3E shows the final titer and average Qp for each batch. The average Qp was calculated as the ratio of final titer (titer at last process time point) to the final IVCC for each batch. The differences in the trajectories, final titer and average Qp related to temperature variation can be seen in the plots. No clear differences related to the pH could be observed in the plots. The VCD increased during the first 175 h and reached 20×10^6 cell/ml within 7 days. The temperature and pH changes were applied to the bioreactors on day 7 (~170 h). After 230 h, the VCD started to decrease, and its decline was faster in the bioreactor with the lowest temperature (31°C). The cell diameter started to increase after reaching a minimum of 20 μm and increased over time up to a maximum of 26 μm, while the cell growth was lower at 31°C. Final IgG titers of about 3 and 2.8 and 2 g L^{-1} were obtained at temperatures 36.8, 33.9, and 31°C, respectively (Figure 3E), which is equivalent to the cell specific productivity of about 19, 18,14 (pg cell^{-1} day^{-1}).

4.2 | Batch modeling

The BEM was implemented on the processed metabolite data from all cultivations. OPLS models can be fitted by correlating metabolomic data with either time or maturity related critical process variables. However, maturity related critical process variables should be monotonic to be implemented in the BEM. Here, an OPLS model was built with time as maturity y-variable (y). Cross-validation indicated the significance of one predictive and one orthogonal OPLS components corresponding to the R^2X for predictive component = 0.369, R^2X for orthogonal component = 0.408, R^2 for predictive component = 0.976, and Q^2 = 0.972. Scores and loadings from the BEM model can be seen in Figure 4-A, which illustrates that there was a strong similarity among the different batches and none of the batches deviated from the overall trend in the bioprocess. Slightly higher variability in the batch trajectories could be seen after imposing temperature and pH changes on day 7 (~170 h). The score plot shows the overall time trend in the predictive score, while the orthogonal score represents the variation regarding the temperature parameter. In the predictive loading, we could see the metabolites that are strongly correlated to the batch time with increased levels of alanine, glycine, serine, tyrosine, choline,

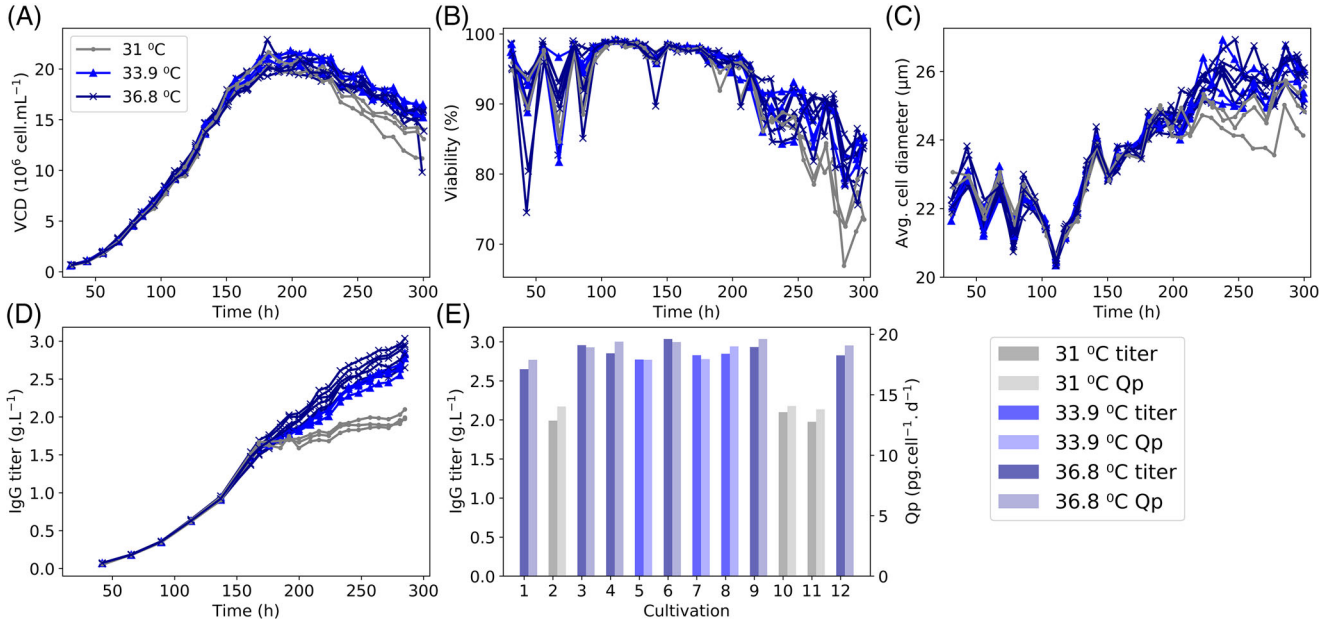

FIGURE 3 Offline measurements of all CHO cultivations in the Ambr250: (A) viable cell density (VCD) (10^6 cell.mL^{-1}), (B) viability (%), (C) average live cell diameter (μm), (D) IgG titer (g.L^{-1}) over 12 days, (E) bar plot of final titer and average Qp for all cultivations

cyanocobalamin, pantothenic acid, acetic acid, butyrate, citrate, formate, fumaric acid, and isovaleric acid, as well as decreased levels of succinic acid, pyroglutamic acid, folic acid, nicotinamide, and several amino acids such as aspartic acid, asparagine, cysteine, glutamic acid, glutamine, histidine, hydroxyproline, isoleucine, leucine, lysine, threonine, and valine over batch time. Whereas connecting each metabolite change with time to specific biochemical pathways would require more effort, including additional experiments, which is outside the scope of presented work, the majority of observed changes in metabolite levels could be connected to gradual consumption of amino acids present in the media and accumulation with time of cell by-products. Therefore, BEM is a proper model to understand if there is any deviation in the process trajectory and find variables that are responsible for the overall time trend. Even though we can observe the impact of the temperature in the OPLS orthogonal components in the score plots, further analysis can also be implemented by applying BLM to find the overall pattern in relation to the experimental design or any other batch related time-independent parameters (see text below).

Finding connections between metabolite levels and specific process parameters, such as VCD, can provide valuable insights into the biological mechanisms involved in the bioprocess. Analysis of variation of metabolic profile of the cell culture in correlation to VCD or titer can contribute to a better understanding of the studied biological system. Here, a proper BEM model with the VCD could not be built in a strict BEM context, due to the drop of the VCD at the last phase of the process. To find the correlation of metabolite changes with VCD, an OPLS model of the metabolomics data was built with VCD as **y**-variable (Figure 4-B). Similar to the model in Figure 4-A, the score plot shows the overall time trend in the predictive score and the temperature variation in the orthogonal score.

These two models (BEM with y = time and OPLS model with y = VCD) had the same dimension in terms of weights, loadings, and regression coefficients. Hence a comparison of the two models was possible by using the SUS-plot methodology using the p(corr) values of BEM with y = time and p(corr) values of the OPLS model with y = VCD. The SUS-plot demonstrated a strong correlation of the metabolite patterns between the two models (Figure 4-C), meaning that no major differences could be seen between the metabolic profiles connected to time and VCD. However, we could not conclude that the loadings and regression coefficients of the VCD OPLS model showed the actual VCD relation with metabolic composition. Even though VCD does not grow monotonically with time, its correlation with time is still important and thus time acts as a confounding factor when comparing metabolic profiles with VCD. This can lead to the strong correlation of results between the two models.

To find the correlation of metabolite changes with titer, the BEM of the metabolomics data was also built with titer as y-variable (Figure S1). However, it should be considered that titer is a cumulative parameter and thus cannot be modeled independently of time as such, and thus it might not be a proper variable for correlating to the metabolite changes. In this case, the correlation with time is even greater as the titer grows monotonically with time. In general, metabolic profiles as well as VCD and titer are dynamic parameters and hence are inherently correlated with time, a correlation that does not need to mean actual causality.

A BLM was also implemented on the processed metabolite data from all cultivations sets (Figure S2). The score plot showed that the variation related to the separation of groups by temperature differences could be explained by the first PC in BLM. The batches with lower temperatures (Temp. = 31) were separated more from the two other

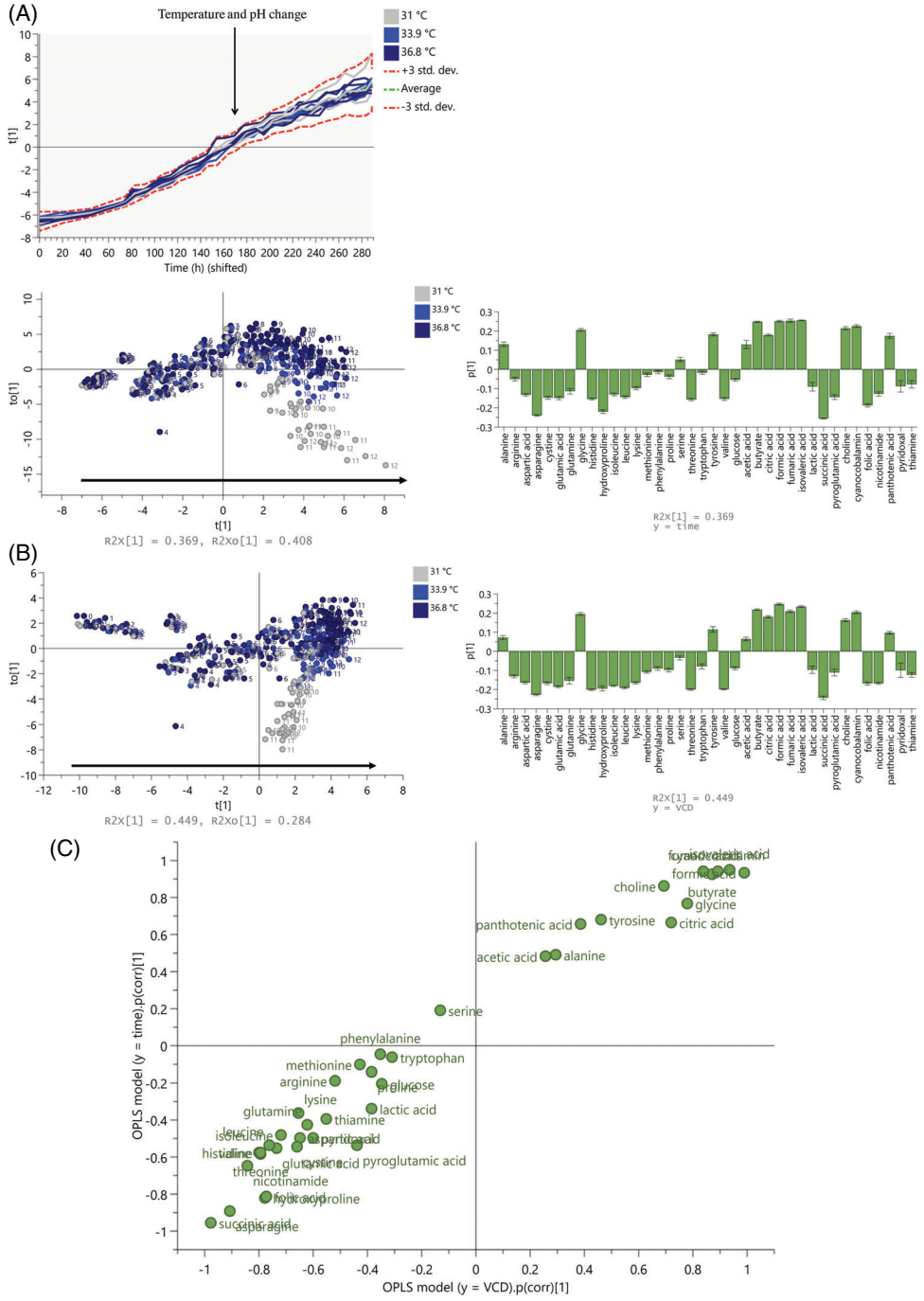

FIGURE 4 BEM (OPLS model with two components) on the metabolomics (A) data with 372 observations and 38 variables for 12 cultivations, BR1 – BR12 and time as y-variable. OPLS model parameters are as follows, R^2X for predictive component = 0.369, R^2X for orthogonal component = 0.408, R^2 for predictive component = 0.976, and Q^2 = 0.972. (B) data with 372 observations and 38 variables for 12 cultivations, BR1 – BR12 and VCD as y-variable. OPLS model parameters are as follows, R^2X for predictive component = 0.449, R^2X for orthogonal component = 0.284, R^2 for predictive component = 0.827, and Q^2 = 0.823. The left figures represent the first OPLS score over time and the score plots of t_2 versus t_1, colored based on the temperature imposed after day 7 (~170 h) (the arrows in the score plots of t_2 versus t_1 illustrate the time direction and the points are labeled by the bioprocess day). The right figures represent the loading plot. (C) SUS-plot of p(corr) vectors for two OPLS models with y = VCD (x-axis) and y = time (y-axis)

groups. The impact of pH on the grouping could not be observed. The effect of temperature and pH can be further investigated by PLS-DA or OPLS-DA model with the temperature or pH as y-variable. Inspection of the loading plots (Figure S2), which contain the information on the metabolite levels in relation to the studied y-variable, revealed that the levels of some metabolites such as isovaleric acid, butyric acid, glycine, and butyric acid decreased in the last phase of the process, while the level of most metabolites increased. BLM is informative on the experimental design observed in the row space/mode of the data as the data is arranged to account for all variables at all time points of completed batches and therefore is proper for showing the quality or yield parameters of finished batches. However, the data arrangement is not proper for the analysis of process evolution in correlation to the VCD profile. The dynamic nature of process parameters such as VCD needs to be accounted for in the modeling. Thus, it is not proper to regress the dynamic parameters versus the BLM data as time will be a confounding factor in the column mode of the data.

Exploring the metabolite differences in correlation to the process parameters, OPLS-DA models can also be built for each growth phase and be compared using SUS-plots (Figure S3). Growth phases are determined based on the process parameters (Theory section) and can reveal metabolite alternations in connection to these parameters. Even though we cannot assess the dynamic nature of the data and find the metabolite evolution over all process time points by applying discriminant analysis, we still divide the data into three time intervals of exponential, stationary and dead phases, and therefore it can provide a picture of metabolite shifts between the specific time intervals of bioprocess, in another word, the dynamic of the system in lower resolution.

We applied OPLS-DA model to evaluate metabolic differences between various growth phases (i.e., exponential vs. stationary, stationary vs. dead, and also exponential vs. dead phases). Therefore, three OPLS-DA models for the three growth phases were built. Table S1 provides the parameters of these three studied models. To compare the similarity of metabolic profiles between different growth phases, we further investigated SUS-plots with p(corr) vectors from the OPLS-DA models of exponential versus stationary and stationary versus dead (Figure S3-D). Obtaining common and unique metabolic profiles between different phases could give relevant information about perturbations in metabolite levels between phases.

4.3 | Hierarchical modeling

The proposed methodology combines OPLS and PCA to model and visualize metabolic profiles in correlation to the dynamic process parameters such as VCD while accounting for the high dimensionality of metabolomics observations (Figure 2).

The method has two steps. At first, it looks at the predictive variation between metabolite levels and VCD. For that, OPLS models are built between metabolite data (X) and VCD (y) at each time point. The OPLS model parameters (number of OPLS components, R^2, Q^2, and ANOVA of cross-validated residuals (CV-ANOVA) *p*-values) can be seen in Table S2 and Figure 5. Growth phases and productivity reported in the table are determined based on Qp and μ values as explained in the theory section.

The main conclusion from the applied analysis is that metabolic profiles have different relations to VCD over time. This can be seen firstly by the different significance of the corresponding OPLS models, that is, some of the models in the initial phase are not significant, which can be interpreted as a result of lack of correlation of metabolite profiles to VCD in the first phase of the bioprocess. It might be due to the fact that in this phase cells are in low number and slowly start to grow and hence the effect of their metabolism was still not visible in the composition of the media. A low correlation between metabolite levels and VCD could also be seen in the late stage of the stationary phase, after imposing the temperature and pH change on the system. At this stage, some of the cultivations have disturbed metabolite equilibrium as a result of changes imposed in the bioreactor. This could be again connected to the fact that cells did not have time to establish a new correlation structure and provide their metabolic imprint into the cell media after rearrangement imposed by changes in the bioreactor conditions. The presented approach for modeling metabolic data can be hence helpful in the diagnosis of the instability and changing correlation structure between metabolite profiles and VCD as a result of the disturbance in the system.

Further confirmation of dynamic relation of metabolites to VCD can be obtained by comparison of metabolic profiles related to VCD at different time points, as investigated by SUS-plots with p(corr) vectors coming from the OPLS models for different time points/growth phases (Figure 6). Analysis of SUS-plots from VCD models for time points from different growth phases clearly shows that transitions between phases are due to alterations in culture conditions such as nutrient depletion and/or waste accumulation. At the same time, two consecutive time points from the same phase (with approximately 8 to 9 h time differences) show that metabolic profiles in the same phase are comparable (diagonal plots in Figure 6). However, the similarities of the metabolite profiles within the dead phase (i.e., SUS-plot of t = 274 h and t = 283 h) are more than in other phases and the weakest similarity can be observed for the profiles within the exponential phase (i.e., SUS-plot of t = 82 h and t = 90 h).

As mentioned above, the differences between the VCD-related metabolite profiles between different growth phases were higher than those for the same phase. However, more similarities could still be observed between VCD related metabolic profiles of stationary and dead phases (i.e., SUS-plot of t = 178 h and t = 274 h) in comparison to the exponential and stationary phases (i.e., SUS-plot of t = 82 h and t = 178 h). Discussion on the main cause of phase transition differences is outside the scope of present work. Analysis of SUS-plots is a good visualization method for comparison of the VCD-related metabolic profiles from different process times in the bioreactor. However, with a higher number of time points, the number of plots will increase, and the proper evaluation of the changes can be difficult.

For the second step of hierarchical modeling, the p(corr) vectors (differential metabograms) obtained from OPLS models of each time point could be compared using an additional MVDA step, here PCA,

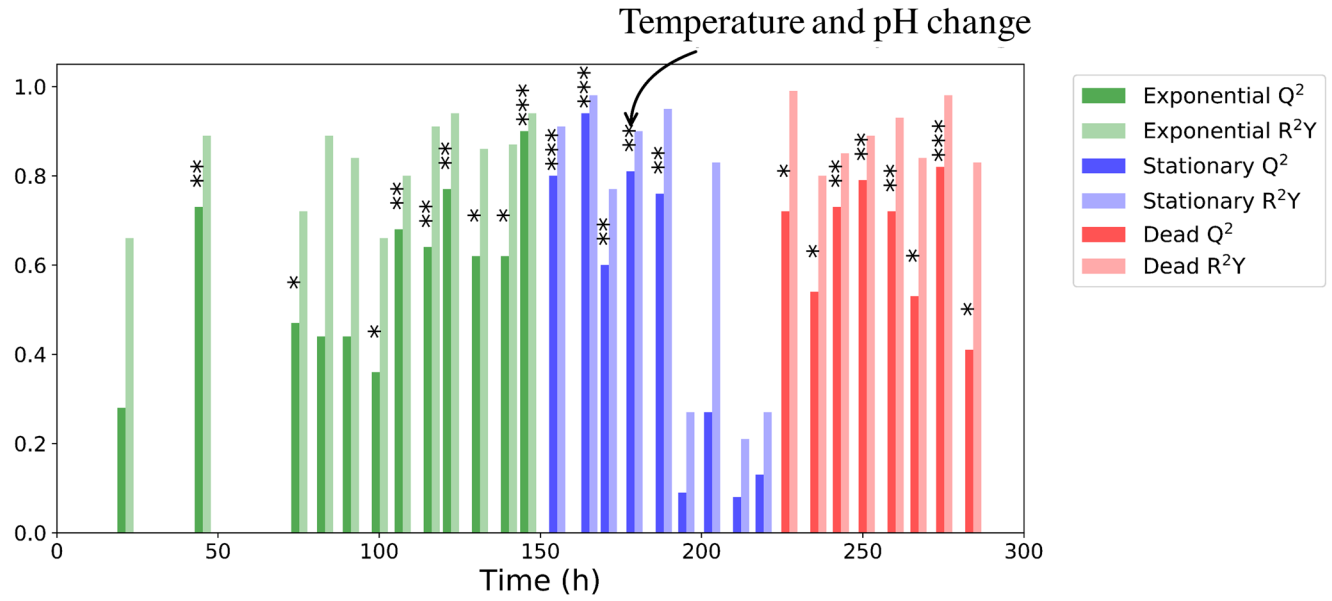

FIGURE 5 Plot of OPLS model parameters R^2Y, and Q^2 on the metabolomics from BR1 – BR12 cultivations at different process times. The Q^2 values are illustrated as bar plots with dark colors according to the growth phase. The R^2Y values are illustrated as bar plots with light colors according to the growth phase

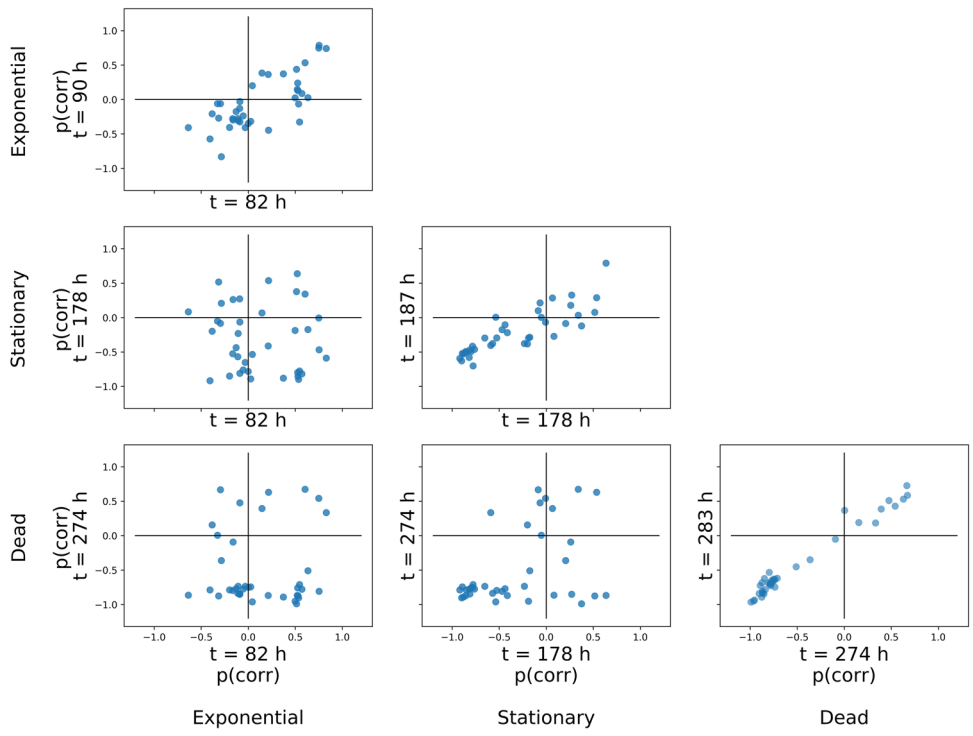

FIGURE 6 SUS-plot of p(corr) vectors from the two OPLS models of some selected time points from various growth phases. The OPLS models are performed on the metabolomics from BR1 – BR12 cultivations. SUS-plots p(corr) vectors from the two OPLS models from different growth phases (off-diagonal) and models from the same growth phase (diagonal)

highlighting the dynamic relation between metabolic profile and the process parameter of interest.

Dynamic changes in metabolite composition in correlation to VCD obtained from hierarchical modeling are visualized in the score plot in Figure 7. The differential metabograms obtained from the OPLS models with Q^2 values lower than 0.3 are illustrated with diamond symbols, as a valid correlation between the metabolite profiles and VCD values could be questioned at these time points. As mentioned

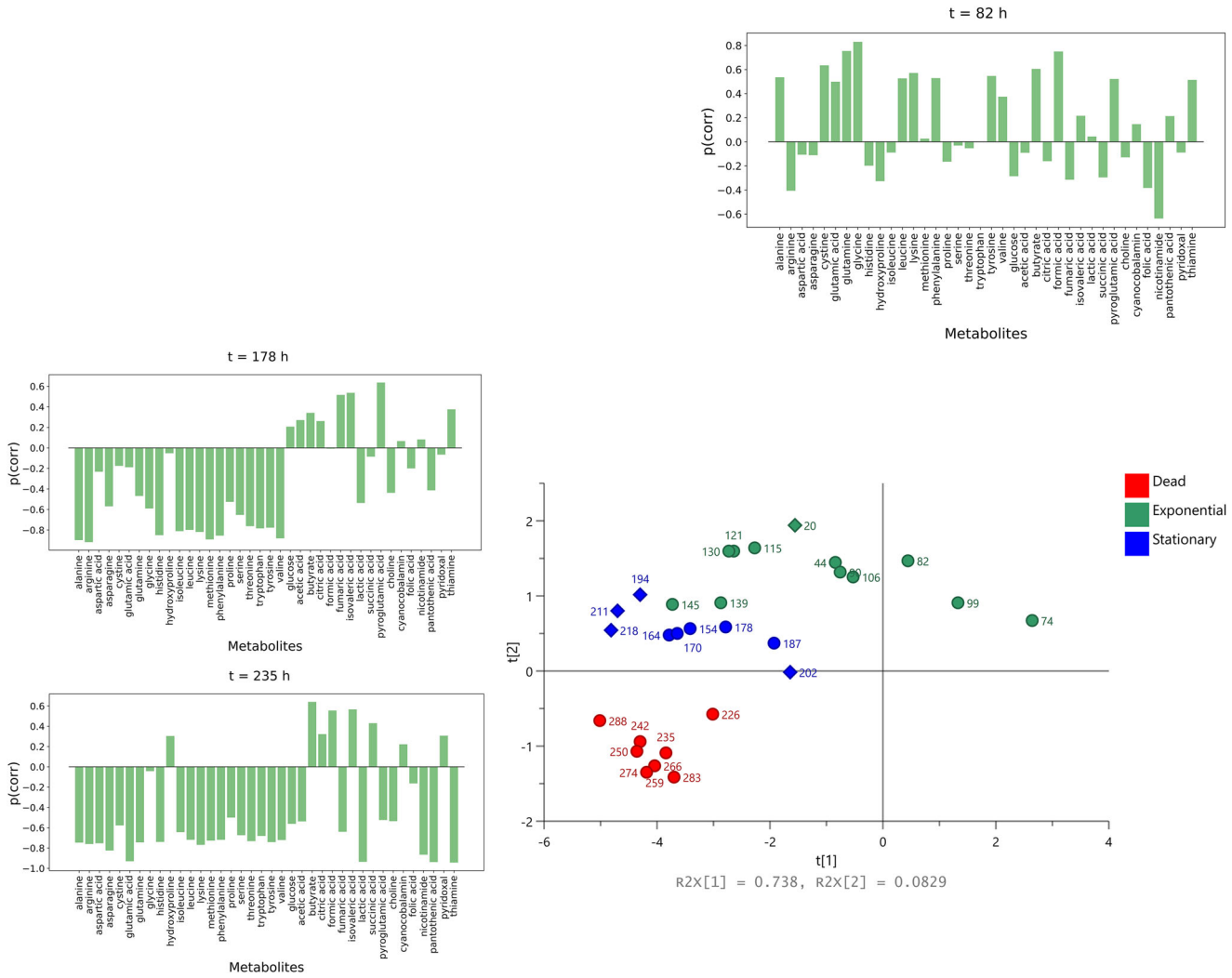

FIGURE 7 Hierarchical modeling on the metabolomics data from BR1 – BR12 cultivations. PCA score plot colored according to growth phase, differential metabograms obtained from the OPLS models with Q2 < 0.3 illustrated with diamond symbols (t[1] and t[2] explain 74% and 8% variation of the data) and p(corr) plots of three time points (i.e., time = 82, 178, and 235 h) are illustrated in the plot

before, these points are positioned at the edges of the stationary phase and are the last points before the dead phase. Therefore rearrangement of the metabolites in the cells might be the reason for lack of a valid correlation at these time points. Metabolite compositions were more comparable to each other over stationary and dead phases, while higher variation could be seen in the exponential phase, which could be the result of general intensive rearrangement of the cell metabolism in this phase of cell growth. The p(corr) plots of three time points (i.e., time = 82, 178, and 235 h) are illustrated in Figure 7 to show metabolite profiles related to VCD at different growth phases. These three p(corr) plots are chosen as an example and p(corr) plots of all time points are presented in Figure S4.

In the PCA score plot, the last two points of the exponential phase (points at 139 and 145 h process time) were closer to the stationary points. We could conclude that the metabolite composition of these two time points, which were determined as exponential phase based on the specific growth rate values, were more similar to the metabolite composition of the later stage in the bioprocess, which could suggest that specific growth rate calculation, in this case, may not be the most accurate approach for determination of the break between the phases. Hence, the hierarchical modeling approach presented here, taking into account real information of metabolite composition correlated to the VCD, can potentially reveal more accurate information about the cell metabolic status.

Even though the traditional batch process analysis such as BEM, BLM, and also OPLS-DA models could provide valuable knowledge about the bioprocess overall trend and design, the confounding influence of time on the relation between dynamic process parameters and metabolic profiles could not be resolved with such models. Therefore, there is a need to properly model and assess the dynamic nature of the data. The proposed model successfully addressed the challenges of time-series data regarding dimensionality as well as the correlation of CPPs with time. It has been shown that hierarchical modeling could visualize the metabolic trajectories related to the CPPs and highlight the metabolites that are correlated to the VCD and evolve over process time. Understanding the cellular metabolism based on

metabolic trajectories and enhanced characterization of the bioprocess allow improvement of the culture conditions, that is, diminishing toxic metabolites or providing nutritional demands of cells over process time. This could lead to enhanced cellular productivity, longer culture durations, and higher titer production.

5 | CONCLUSIONS

The practical approach for maintaining minimum product variability is by running the processes on pre-defined and well-established trajectories. This requires increased bioprocess understanding together with advanced process monitoring and application of comprehensive data mining models to be able to extract relevant information from the obtained data. While it is currently possible to effectively monitor and control bioprocesses, understanding complex interactions between cells and process parameters that influence final process outcome and product quality is still not fully achieved, even if high-quality biological data such as metabolomics data are acquired. Analyzing time-series data from bioprocess studies is challenging due to dimensionality of the data, the correlation structure in the metabolite data and the correlation structure due to repeated measurements over time. Many currently existing approaches to analyzing such data sets either have the limitation of confounding treatment variation with variability due to the dynamic nature of the data or they ignore the fact that metabolites do not work independently of each other.

Here, a hierarchical approach, based on OPLS and PCA is proposed to achieve dimension reduction while appropriately modeling the correlation of metabolite levels to the CPPs such as VCD. We demonstrated the application of the proposed methodology on metabolic data from CHO based antibody production and compared obtained results to traditional approaches used in modeling bioprocess data (i.e., batch modeling) to characterize the dynamic relation between metabolic profiles and CPPs. The proposed model successfully addressed the aims of the bioprocess study, that is, it enabled visualization of the metabolomic trajectories through time, presenting the time effect on the metabograms, and highlighting metabolites that evolve over time. Many areas of further research naturally arise from the proposed model such as investigating the application of the model to different process parameters. It could be also interesting to assess how this type of model can be applied to other processes.

AUTHOR CONTRIBUTIONS

Masoumeh Alinaghi: Formal analysis; Investigation; Methodology; Validation; Visualization; Writing – original draft; Writing – review & editing. Izabella Surowiec: Data curation; Investigation; Methodology; Project administration; Supervision; Writing – review & editing. Steffi Scholze: Data curation; Investigation; Methodology; Writing – review & editing. Chris McCready: Conceptualization; Investigation; Methodology. Christoph Zehe: Conceptualization; Investigation; Methodology. Erik Johansson: Investigation; Methodology; Writing – review & editing. Johan Trygg: Conceptualization; Investigation; Methodology. Olivier Cloarec: Conceptualization; Investigation; Methodology; Supervision; Writing – review & editing.

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

The datasets generated during the current study are not publicly available but are available from the corresponding author on reasonable request.

REFERENCES

1. Kim, J. Y., Kim, Y.-G., & Lee, G. M. (2012). CHO CHO cells in biotechnology for production of recombinant proteins: Current state and further potential. *Applied Microbiology and Biotechnology*, *93*(3), 917–930.
2. Noh, S. M., Sathyamurthy, M., & Lee, G. M. (2013). Development of recombinant Chinese hamster ovary cell lines for therapeutic protein production. *Current Opinion in Chemical Engineering*, *2*(4), 391–397.
3. Jagschies, G., Lindskog, E., Lacki, K., & Galliher, P. M. (2018). *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes*, Elsevier.
4. Richelle, A., & Lewis, N. E. (2017). Improvements in protein production in mammalian cells from targeted metabolic engineering. *Current Opinion in Systems Biology*, *6*, 1–6.
5. Guideline, I. H. T. (2009). Pharmaceutical Development. *Q8 (2R)*. As revised in August.
6. Trummer, E., Fauland, K., Seidinger, S., Schriebl, K., Lattenmayer, C., Kunert, R., Vorauer-Uhl, K., Weik, R., Borth, N., & Katinger, H. (2006). Process parameter shifting: Part I. Effect of DOT, PH, and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. *Biotechnology & Bioengineering*, *94*(6), 1033–1044.
7. Rathore, A. S., & Winkle, H. (2009). Quality by design for biopharmaceuticals. *Nature Biotechnology*, *27*(1), 26–34.
8. N Politis, S., Colombo, P., Colombo, G., & M Rekkas, D. (2017). Design of experiments (DoE) in pharmaceutical development. *Drug Development and Industrial Pharmacy*, *43*(6), 889–901.
9. Alden, N., Raju, R., McElearney, K., Lambropoulos, J., Kshirsagar, R., Gilbert, A., & Lee, K. (2020). Using metabolomics to identify cell line-independent indicators of growth inhibition for Chinese hamster ovary cell-based bioprocesses. *Metabolites*, *10*(5), 199.
10. Craven, S., & Becken, U. (2014). A quality-by-design approach to upstream bioprocess interrogation and intensification. *Engineering Journal*.
11. Nomikos, P., & MacGregor, J. F. (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 97–108.
12. Wold, S., Kettaneh, N., Fridén, H., & Holmberg, A. (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory Systems*, *44*(1–2), 331–340.
13. Dahl, K. S., Piovoso, M. J., & Kosanovich, K. A. (1999). Translating third-order data analysis methods to chemical batch processes. *Chemometrics and Intelligent Laboratory Systems*, *46*(2), 161–180.
14. Undey, C., & Cinar, A. (2002). Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Systems Magazine*, *22*(5), 40–52.
15. Louwerse, D., Tates, A. A., Smilde, A. K., Koot, G. L., & Berndt, H. (1999). PLS discriminant analysis with contribution plots to determine

differences between parallel batch reactors in the process industry. *Chemometrics and Intelligent Laboratory Systems*, *46*(2), 197–206.
16. Singhal, A., & Seborg, D. E. (2002). Pattern matching in historical batch data using PCA. *IEEE Control Systems Magazine*, *22*(5), 53–63.
17. Andersen, S. W., & Runger, G. C. (2011). Partitioned partial least squares regression with application to a batch fermentation process. *Journal of Chemometrics*, *25*(4), 159–168.
18. Aguado, D., Ferrer, A., Ferrer, J., & Seco, A. (2007). Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems*, *85*(1), 82–93.
19. Kourti, T. (2003). Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics*, *17*(1), 93–109.
20. Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, *19*(4), 213–246.
21. García-Muñoz, S., Kourti, T., & MacGregor, J. F. (2004). Model predictive monitoring for batch processes. *Industrial & Engineering Chemistry Research*, *43*(18), 5929–5941.
22. Duchesne, C., & MacGregor, J. F. (2000). Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics and Intelligent Laboratory Systems*, *51*(1), 125–137.
23. Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.
24. Walsh, M. C., Brennan, L., Pujos-Guillot, E., Sébédio, J.-L., Scalbert, A., Fagan, A., Higgins, D. G., & Gibney, M. J. (2007). Influence of acute phytochemical intake on human urinary metabolomic profiles. *The American Journal of Clinical Nutrition*, *86*(6), 1687–1693.
25. Sachse, D., Sletner, L., Mørkrid, K., Jenum, A. K., Birkeland, K. I., Rise, F., Piehler, A. P., & Berg, J. P. (2012). Metabolic changes in urine during and after pregnancy in a large, multiethnic population-based cohort study of gestational diabetes. *Plos One*, *7*(12), e52399.
26. Carvalho, E., Franceschi, P., Feller, A., Palmieri, L., Wehrens, R., & Martens, S. (2013). A targeted metabolomics approach to understand differences in flavonoid biosynthesis in red and yellow raspberries. *Plant Physiology and Biochemistry*, *72*, 79–86.
27. Smolinska, A., Blanchet, L., Buydens, L. M., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*, *750*, 82–97.
28. Cassol, E., Misra, V., Holman, A., Kamat, A., Morgello, S., & Gabuzda, D. (2013). Plasma metabolomics identifies lipid abnormalities linked to markers of inflammation, microbial translocation, and hepatic function in HIV patients receiving protease inhibitors. *Bmc Infectious Diseases [Electronic Resource]*, *13*(1), 1–17.
29. Bathen, T. F., Geurts, B., Sitter, B., Fjøsne, H. E., Lundgren, S., Buydens, L. M., Gribbestad, I. S., Postma, G., & Giskeødegård, G. F. (2013). Feasibility of MR metabolomics for immediate analysis of resection margins during breast cancer surgery. *Plos One*, *8*(4), e61578.
30. Choi, Y. H., Kim, H. K., Linthorst, H. J., Hollander, J. G., Lefeber, A. W., Erkelens, C., Nuzillard, J.-M., & Verpoorte, R. (2006). NMR metabolomics to revisit the tobacco mosaic virus infection in nicotiana t abacum leaves. *Journal of Natural Products*, *69*(5), 742–748.
31. Jansen, J. J., Hoefsloot, H. C., Boelens, H. F., Van Der Greef, J., & Smilde, A. K. (2004). Analysis of longitudinal metabolomics data. *Bioinformatics*, *20*(15), 2438–2446.
32. Smilde, A. K., Jansen, J. J., Hoefsloot, H. C., Lamers, R.-J. A., Van Der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics*, *21*(13), 3043–3048.
33. Alinaghi, M., Bertram, H. C., Brunse, A., Smilde, A. K., & Westerhuis, J. A. (2020). Common and distinct variation in data fusion of designed experimental data. *Metabolomics*, *16*(1), 1–11.
34. Jansen, J. J., van Dam, N. M., Hoefsloot, H. C., & Smilde, A. K. (2009). Crossfit analysis: A novel method to characterize the dynamics of induced plant responses. *Bmc Bioinformatics [Electronic Resource]*, *10*(1), 1–15.
35. Smilde, A., Westerhuis, J., Hoefsloot, H., Bijlsma, S., Rubingh, C., Vis, D., Jellema, R., Pijl, H., Roelfsema, F., & Van Der Greef, J. (2010). Dynamic metabolomic data analysis: A tutorial review. *Metabolomics*, *6*(1), 3–17.
36. Berk, M., Ebbels, T., & Montana, G. (2011). A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics*, *27*(14), 1979–1985.
37. Mei, Y., Kim, S. B., & Tsui, K.-L. (2009). Linear-mixed effects models for feature selection in high-dimensional NMR spectra. *Expert Systems with Applications*, *36*(3), 4703–4708.
38. Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). *Multi-and Megavariate Data Analysis Basic Principles and Applications*, Umetrics Academy, Vol. 1.
39. Metze, S., Ruhl, S., Greller, G., Grimm, C., & Scholz, J. (2020). Monitoring online biomass with a capacitance sensor during scale-up of industrially relevant CHO cell culture fed-batch processes in single-use bioreactors. *Bioprocess and Biosystems Engineering*, *43*(2), 193–205.
40. Hendrick, V., Winnepenninckx, P., Abdelkafi, C., Vandeputte, O., Cherlet, M., Marique, T., Renemann, G., Loa, A., Kretzmer, G., & Werenne, J. (2001). Increased productivity of recombinant tissular plasminogen activator (t-PA) by butyrate and shift of temperature: A cell cycle phases analysis. *Cytotechnology*, *36*(1), 71–83.
41. Schatz, S. M., Kerschbaumer, R. J., Gerstenbauer, G., Kral, M., Dorner, F., & Scheiflinger, F. (2003). Higher expression of fab antibody fragments in a CHO cell line at reduced temperature. *Biotechnology and Bioengineering*, *84*(4), 433–438.
42. Chusainow, J., Yang, Y. S., Yeo, J. H., Toh, P. C., Asvadi, P., Wong, N. S., & Yap, M. G. (2009). A study of monoclonal antibody-producing CHO cell lines: What makes a stable high producer? *Biotechnology and Bioengineering*, *102*(4), 1182–1196.
43. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(03), 90–95.
44. Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T., & Trygg, J. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, *80*(1), 115–122.
45. Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., & Eriksson, L. (1996). Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics*, *10*(5-6), 521–532.

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: Alinaghi, M., Surowiec, I., Scholze, S., McCready, C., Zehe, C., Johansson, E., Trygg, J., & Cloarec, O. (2022). Hierarchical time-series analysis of dynamic bioprocess systems. *Biotechnology Journal*, *17*, e2200237. https://doi.org/10.1002/biot.202200237