Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-3 (2021) 85–90

Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics

Gianmarco Barberi*, Antonio Benedetti*, Paloma Diaz-Fernandez**, Gary Finka**, Fabrizio Bezzo*, Massimiliano Barolo*, Pierantonio Facco***

*CAPE-Lab – Computer Aided Process Engineering Laboratory, Padova, 35138 Italy (Tel: +39.049.8275470; e-mail: pierantonio.facco@unipd.it)

** Biopharm Process Research, GlaxoSmithKline R&D, Stevenage, UK

*** Process Engineering & Analytics, GlaxoSmithKline R&D, Stevenage, UK

Abstract: The development of biopharmaceutical therapeutics, such as monoclonal antibodies, requires the testing of several cell lines at different development scales and the selection of the high performing cell lines which allow meeting the desired quality attributes of the product. In this context, data analytics, which is extremely useful for a better process understanding and a faster scale-up, can be used to understand the relation between biological information, such as cell metabolism, and process productivity.
This study shows that monoclonal antibodies end-point titer can be estimated in the early stages of the industrial product development for cell line selection using information on cell metabolism dynamics. This allows the anticipated identification of the high-performing cell lines, and a better understanding of the relationships between the time evolution of both the metabolic information and the product titer.

Keywords: biopharmaceutical development, scale-up, monoclonal antibodies, data analytics, multivariate statistics, metabolomics

1. INTRODUCTION

In recent years there has been an increasing interest in the development of biopharmaceutical drugs as a response to the increased cost of conventional drug development, patent expiration, and market erosion through generic drugs. This has resulted in a rapid growth of the biopharmaceutical market, which recently assesses over 7000 drugs on development and more than 450 billion $ annual sales. Among the biopharmaceutical products, recombinant proteins, such as monoclonal antibodies (mAbs), are the highest selling class with more than 1500 drugs in the development pipeline (Hong et al., 2018).

Monoclonal antibodies are an important class of therapeutic molecules used to treat immunological and oncological diseases in humans. Mammalian cell cultures, such as Chinese Hamster Ovary (CHO) cells, represent nowadays the preferred choice for mAbs production, since they guarantee enzymatic post-translational modification, such as glycosylation, which are essential for the activity and safety of mAbs (Karst et al., 2017). The successful development of a mAb, as well as of any other biopharmaceutical product, is a resource intensive, time consuming and multiple step process (Li et al., 2010). For these reasons, a major challenge in biopharmaceuticals development is the rapid development of a robust and scalable process, which allows accelerating the progress of several mAbs into clinical trials. Thus, biopharmaceutical companies are increasingly looking at innovative solutions to reduce the drug time to market while maintaining the desired quality attributes in order to improve the economics of drug development (Rameez et al., 2014).

The development pipeline of cell culture process for mAbs production typically begins selecting the best cell lines in small scale systems (e.g.: nanofluidic and optoelecto positioning systems, microwell plates, and shake flasks), which allow high throughput experiments. However, these systems lack the control of some important process parameters, such as agitation rate, dissolved oxygen, and pH. From this stage, the top cell-line candidates are isolated and further analyzed in fed-batch laboratory scale multi-parallel bioreactors, which mimic the performance of the industrial scale processes providing control of agitation, dissolved oxygen, and pH (Rameez et al., 2014). This stage enables the selection of the top cell lines, which will be further analyzed at bioreactor scale to identify the production and backup cell line.

Such a scale-up procedure is essential to obtain the best final production cell line, since performance is cell line dependent. For this reason, several quality attributes, such as specific and volumetric productivity, cell growth, and product quality, are considered for cell selection. In particular, the volumetric productivity (product titer) has an important role because it quantifies the product concentration obtained in the bioreactor. Product quality can be effectively predicted through first principles (Duvigneau et al., 2020), data-driven or hybrid approaches. Data-driven methods do not require fundamental knowledge, and can be effectively used for early estimation of productivity. This class of models allow understanding the impact of process variables on the biopharmaceutical product quality, which is important to speed up and better understand the process development (Facco et al., 2020; Gregersen and Jørgensen, 1999). Furthermore, methods for quality control and process understanding are encouraged by regulatory

10.1016/j.ifacol.2021.08.223

agencies (Food and Drug Administration, 2004). Data-based method have rarely been applied to study the relationship between quality attributes and biological information (Zürcher et al., 2020), and studies on the early estimation of product titer through the biological information on cell metabolism are missing. Cell metabolism information can be obtained through untargeted metabolomics (Zhou et al., 2012), which consists in the collection of all the metabolites of a biological system with liquid chromatography mass spectrometry (LC-MS). Through untargeted metabolomics, information on the dynamics of metabolic profiles (i.e., how the metabolic profile of cell lines varies during the culture) can be linked to product titer. In this way, cells exhibiting high product titer can be identified observing the temporal variations in cell metabolism.

In this study, two main objectives are pursued. First, information on the dynamics of cell metabolic profiles is used to early estimate product end-point titer in CHO cell cultures. The early identification of product titer allows: *i*) a reduction of the time required for the experimentation at the laboratory scale, and *ii*) the early identification and selection of best-performing cell lines. The second objective is the study of the correlation between the product titer time profile during culture and the dynamics of cell metabolic profiles, because a better understanding of the relationship between cell metabolism and product titer allows the cell selection based on the desired metabolic traits.

2. METHODS

2.1 Dataset

Cell culture data collected in the production of a human mAb are considered in this study. In particular, 48 CHO cell lines were cultured for approximatively 2 weeks in an Ambr®15 Cell Culture apparatus (Sartorius Lab Instruments GmbH & Co. KG, Goettingen, Germany). Several culture process variables were measured at 7 time points along the experimental batch.

The product titer (in mg/L), i.e., the concentration of the mAb in the culture, is the target quality attribute and is measured with Cedex Bio HT analyzer (Roche Diagnostic Corporation, Indianapolis, US). The measurement uncertainty is estimated in 6% of the measured value.

Metabolomics data of the culture supernatant were collected from LC-MS measurements performed at the same 7 time points as product titer in 2 replicates. Samples were analyzed in negative ionization mode under a scan range of mass over charge (*m/z*) 50-1000 (Fuhrer et al., 2011). LC-MS measurements were pre-processed prior the statistical analysis following a standard procedure based on scan alignment, peak detection and modeling, peak alignment, baseline correction and removal of spurious peaks (Frederick et al., 2020). Then, metabolite identification and confirmation were performed prior to normalization and log-10 transformation. Several quality control samples were used to remove batch effects across different instrument runs.

Pre-processed metabolomics data consist in log-10 transformed intensity of $M = 4489$ detected ions, which are characterized by their *m/z*. Pre-processed data are organized in a four-dimensional array $\underline{\mathbf{X}}'$ $[N \times M \times I \times K]$, where $N = 48$ cell cultures are located along the first dimension, M ions *m/z* along the second dimension, $I = 7$ time points along the third dimension, and $K = 2$ measurement replicates along the fourth dimension.

2.2 Multiway Partial Least-Squares

In this study, multiway partial least-squares (MW-PLS; Nomikos and MacGregor, 1995) is considered to investigate the relation among cell metabolism and process performance and deal with multidimensional matrices. MW-PLS consists in a proper unfolding of multi-way data followed by a partial least-squares (PLS; Wold et al., 2001) modeling.

Prior to the analysis, metabolomics data are mean centered and Pareto-scaled (Eriksson et al., 2006; i.e., each ion's intensity divided by the square root of its standard deviation).

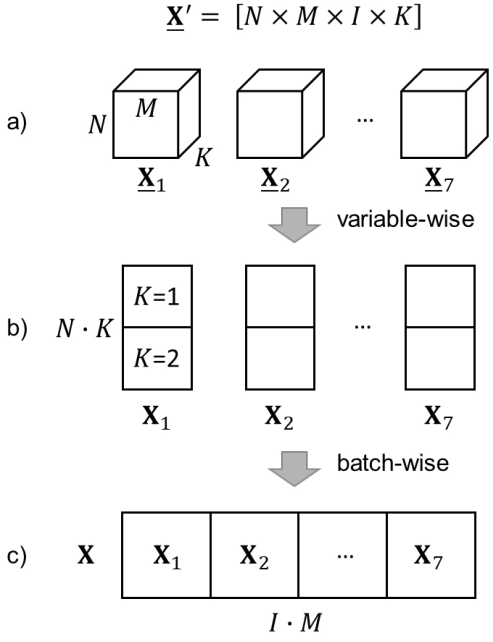

Fig. 1. Data unfolding.

Batch-wise unfolding (Nomikos and MacGregor, 1995) is then applied to $\underline{\mathbf{X}}'$ to take into consideration the dynamics of metabolomics data and the correlation structure among ion intensities collected at different time points during the culture evolution. Data collected at different time points, $\underline{\mathbf{X}}_i$ $[N \times M \times K]$ with $i = 1, 2, \dots, 7$, were isolated (Figure 1a) and variable-wise unfolded by vertically concatenating measurement replicates (Figure 1b) to generate $\mathbf{X}_i$ $[N \cdot K \times M]$. Then, data collected at different time points, $\mathbf{X}_i$ $(i = 1, 2, \dots, 7)$, are horizontally concatenated (Figure 1c) to generate matrix $\mathbf{X} = [N \cdot K \times I \cdot M] = [96 \times 31423]$, which is the batch-wise unfolded version of $\underline{\mathbf{X}}'$. Similarly, the response matrix $\underline{\mathbf{Y}}'$ $[N \times 1 \times I]$ (i.e., product titer) is variable-wise unfolded by vertically concatenating two copies of $\underline{\mathbf{Y}}'$ in $\underline{\mathbf{Y}}''$ $[N \cdot K \times 1 \times$

I], in such a way as to match measurement replicates in $\mathbf{X}$. Then, $\underline{\mathbf{Y}}''$ is batch-wise unfolded and product titers measured along culture evolution are horizontally concatenated to generate matrix $\mathbf{Y} = [N \cdot K \times I] = [96 \times 7]$.

PLS is then applied. PLS is a linear multivariate regression model which relates the matrix $\mathbf{X}$ $[N \cdot K \times I \cdot M]$ of $I \cdot M$ regressors (i.e., the time profiles of collected ions) for $N \cdot K$ observations (i.e., experimental batches with replicates) to a matrix $\mathbf{Y}$ $[N \cdot K \times I]$ of I responses (i.e., the time profile of the titer) for the same observations. PLS decomposes both matrices $\mathbf{X}$ and $\mathbf{Y}$ into a reduced space of A orthogonal latent variables (LVs) according to:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}, \quad (2)$$

where $\mathbf{P}^{\mathrm{T}}$ $[A \times I \cdot M]$ and $\mathbf{Q}^{\mathrm{T}}$ $[A \times I]$ are the transpose of the loading matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\mathbf{T}$ $[N \cdot K \times A]$ and $\mathbf{U}$ $[N \cdot K \times A]$ are the score matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathbf{E}$ $[N \cdot K \times I \cdot M]$ and $\mathbf{F}$ $[N \cdot K \times I]$ are the residual matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, which are minimized in a least square sense. The loadings describe how the metabolites time profiles are correlated and combined to generate the subspace of LVs, while the scores describe how the experimental batches are related with respect to how $\mathbf{X}$ and $\mathbf{Y}$ covary, namely with respect to the covariance structure among both the metabolites time profiles and the titer time profile.

In PLS, weights are introduced to preserve the orthogonality among LVs scores and estimate the response $\hat{\mathbf{Y}}$ form the observations:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{W}\,(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\,\mathbf{Q}^{\mathrm{T}} \quad , \quad (3)$$

where $\mathbf{W}$ $[I \cdot M \times A]$ is the weight matrix. Additionally, weights are used to calculate the model scores $\mathbf{T}_{new}$ of new observations $\mathbf{X}_{new}$ as:

$$\mathbf{T}_{new} = \mathbf{X}_{new}\mathbf{W}\,(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \quad , \quad (4)$$

and predict the response associated to the new observations:

$$\hat{\mathbf{Y}}_{new} = \mathbf{X}_{new}\,\mathbf{W}\,(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\,\mathbf{Q}^{\mathrm{T}} \quad . \quad (5)$$

2.3 Evolving Partial Least-Squares

An evolving PLS model (E-PLS; Ramaker et al., 2005) is used for the real-time estimation of the product end-point titer. E-PLS is a multi-model strategy (Figure 2) that at the i-th time point builds a PLS model on matrix $\mathbf{X}^i = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_i]$, the batch-wise unfolded dataset of dimension $[N \cdot K \times i \cdot M]$ of the metabolomics data up to i-th time point, and the end-point titer, the response variable $\mathbf{Y}_E$ $[N \cdot K \times 1]$ that contains the product titer at 7th time point.

For both MW-PLS and E-PLS, the number of LVs was selected through a 9-fold cross-validation (Geladi and Kowalski, 1986). Model performance is evaluated through a 250-iteration Monte Carlo cross-validation, which consists in a random division of the dataset in calibration and validation samples (88% and 12% of the dataset, respectively). In cross-validation both measurement replicates of a sample are included either in the calibration dataset or in the validation one.

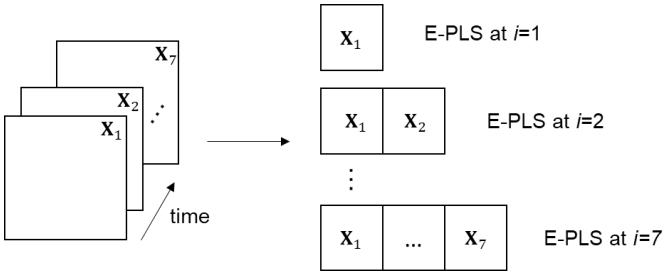

Fig. 2. E-PLS building procedure.

2.4 Variable selection

The selection of the most relevant variables (i.e., ions) for the PLS model improves performance, removes redundant and noisy variables, and allows a better interpretation and understanding of the system (Mehmood et al., 2012). In this study, variable selection was performed through a bootstrap procedure (Afanador et al., 2013) on the variable importance in prediction (VIP; Eriksson et al., 2006) index. The VIP score of an ion at a specific time point, v, is defined as:

$$\mathrm{VIP}_v = \frac{\sqrt{I \cdot M \sum_{a=1}^{A} R_{Y,a}^2 w_{v,a}^2}}{\sqrt{\sum_{a=1}^{A} R_{Y,a}^2}} \quad , \quad (6)$$

where $R_{Y,a}^2$ is the variance of the response explained by the a-th LV of the model, and $w_{v,a}$ is the weight of the v-th ion and a-th LV.

The bootstrap procedure allows assessing the variability in variables' importance in order to perform a robust selection of the most influential variables for the titer prediction. This methodology retains only those variables whose VIP scores remain high independently of the available subset of samples in the validatory iterations. In particular, the bootstrap procedure was performed through p = 250 iterations. At each iteration, a PLS model was built by excluding a randomly selected 12% of the available samples, and following 3 steps over the results of the p iterations:

1. calculation of VIP index standard deviation of each ion;
2. calculation of the 5th percentile ($\alpha = 0.1$) of the VIP-index distribution for each ion, under the assumption that that VIPs are distributed according to a Student's t distribution. Then, the 5th percentile is calculated through: $\hat{\sigma}_{\mathrm{VIP}_{im}} t_{1-\alpha/2,p-1}$, where $t_{1-\alpha/2,p-1}$ identifies the 5% confidence threshold of a t-distribution with $(p-1)$ degrees of freedom and variance $\hat{\sigma}_{\mathrm{VIP}_{im}}$ determined from the values of the VIP index for each ion im over p iterations;
3. selection of the 5% top ranked variables according to the 5th percentile. This percentage of selected ions provides good model performance without retaining an

excessive number of variables, thus allowing an easier interpretation of the results.

The selected variables are then organized in a matrix $\mathbf{X}_s$ $[N \cdot K \times V]$ of much reduced dimension ($V \ll I \cdot M$), which is used to train updated versions of the abovementioned models (similarly to Sections 2.2 and 2.3) which show improved prediction performance.

3. RESULTS AND DISCUSSIONS

3.1 Early estimation of product titer

The early estimation of product end-point titer $\mathbf{Y}_E$ is performed using E-PLS model from the dynamics of cell metabolic profiles. The estimation performance with variable selection is shown in Table 1, where the estimation accuracy is reported in terms of: determination coefficients in calibration R_Y^2 and in validation Q^2, average absolute estimation error $\bar{\varepsilon}$ in validation and ratio between the average absolute error in validation and the variability of the calibration data $\bar{\varepsilon}/\sigma_{cal}$.

Table 1. Validation estimation performance of E-PLS with variable selection for the early estimation of end-point titer

Time point	R_Y^2 [%]	Q^2 [%]	$\bar{\varepsilon}$ [mg/L]	$\bar{\varepsilon}/\sigma_{cal}$ [%]
1	95.9	43.4	412.0	45.6
2	98.3	66.0	329.2	36.5
3	96.8	64.1	346.9	38.5
4	96.5	63.4	348.0	38.5
5	94.8	65.8	320.3	35.3
6	92.8	63.5	323.3	35.6
7	93.1	65.3	320.3	35.2

Apart from good fitting performance in calibration ($R_Y^2 >$ 90% for all the time points), E-PLS with variable selection provides satisfactory estimation in validation with Q^2>40% and $\bar{\varepsilon}/\sigma_{cal}$<50%, especially after time point 2. The estimated values are close or within the instrumental measurement uncertainty and always smaller than σ_{cal}. Accordingly, cell lines exhibiting high end-point titer can be identified through the proposed methodology very early. In fact, even at time point 2 (i.e., few days after the beginning of the experiment), the end-point titer is estimated with high accuracy ($\bar{\varepsilon}$ < 330 mg/L, with an error ε > 800 mg/L, namely ε/σ_{cal} > 87% only in 6.6% of the predictions). According to these results, the dynamics of cell metabolic profiles represents a good indicator of cell performances and could be exploited for the early screening of cell lines behavior.

Since E-PLS with variable selection provides a sufficiently accurate estimation of end-point titer, it can be used to select cell lines from the early stages of the cell culture. An example of early estimation of end-point titer is presented for two cell lines exhibiting high and low product titer. Figure 3 reports the real end-point titer (dashed line) compared to the estimated one (lines with circles/squares) at all time points for the two cell lines (in different colors). A clear difference in the expected end-point titer of the two cells is visible at all time points, even when only time point 1 is considered. In fact, an error ~400 mg/L is found at that time point, which still provides an indication of the expected end-point titer. A more accurate estimation ($\bar{\varepsilon}$ < 200 mg/L) is obtained for both cell lines after time points 2. Based on these results, the proposed model allows an accurate early estimation of the end-point titer since the first week of culture. In this way, high-performing cell lines can be identified in few days after the beginning of the culture, while the low-performing cells can be discarded. Furthermore, the model suggests a considerable reduction (~50% - 80%) in the experiment duration because, even with a reduced experiment length the screening capability is satisfactory.

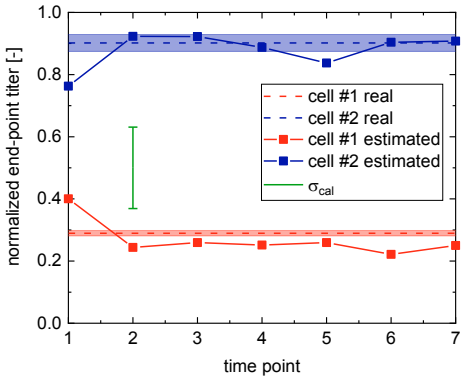

Fig. 3. Example of E-PLS estimation with variable selection for two cell lines: real and estimated end-point titers are compared at all time points. Shaded areas report the 6% measurement uncertainty. For confidentiality reasons the y-axis scale is normalized between 0 and 1.

3.2 Study of the correlation between product titer time profile and the dynamics of metabolic profiles

An in-depth understanding of the relation among the cell metabolism and the process performance can be achieved by studying the correlation between the product titer time profile and the dynamics of metabolic profiles. To this purpose, a MW-PLS model was built to estimate the product titer time profile (**Y**) from the metabolic dynamic profiles (**X**).

The performance of MW-PLS with variable selection is shown in Figure 4. A four-LV model, capturing 87.1% of **Y** variability by means of 30.5% of **X** variability, was built. Even if this model does not provide good estimations of the product titer at time points 1 and 2 (Q^2 < 30% and $\bar{\varepsilon}/\sigma_{cal}$> 55%), proving that in the early stages of the culture the dynamics of cell metabolic profiles does not contain a fingerprint of product titer (because the metabolic information regarding product titer is partially hidden by measurement noise and unsystematic variability related to other phenotypes or inherent cell variability), from time point 3 on MW-PLS with variable selection estimates product titer with relatively small error (Q^2 > 50% and $\bar{\varepsilon}/\sigma_{cal}$< 45%). This means that, from time point 3, product titer leaves a fingerprint on the dynamics of cell metabolic profiles. As a consequence, this allows a better understanding of the metabolic traits which are typical of the most promising cell lines to be progressed in the scale-up.

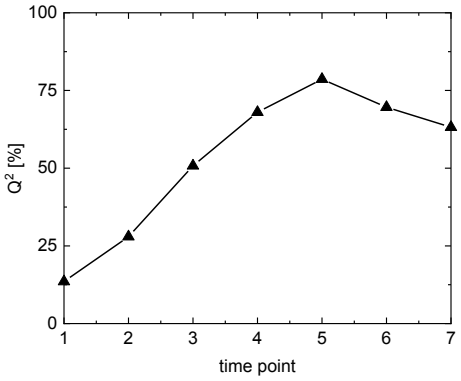

Fig. 4. Validation estimation performance of the MW-PLS with variable selection for the product titer time profile from the dynamics of metabolic profiles.

The typical **Y** loading plot of the model in one iteration of the cross-validation (Figure 5) shows both the auto-correlation between product titer at different time points and its relationship with the variability of the dynamics of metabolic profiles captured by model LVs. The first LV (plain blue bars) shows high positive values from time point 4 on, indicating a positive auto-correlation in titer. Differently, **Y** loadings of LV 2 (striped gray bars) show high positive values between time points 1 and 4, indicating a positive auto-correlation in titer in the first week of culture. The fact that the model captures product titer variability in the first part of the culture (time point 1 to 4) and in the second part of the culture (time point 4 to 7) with 2 LVs, which are orthogonal by definition, means that distinct and independent metabolic phenomena are related to product titer in these two phases of the culture. Since these phenomena are independent, cell lines can show titer below the average in the initial part of the culture, while showing titer above the average in the final part of the culture (and the opposite occurs, as well).

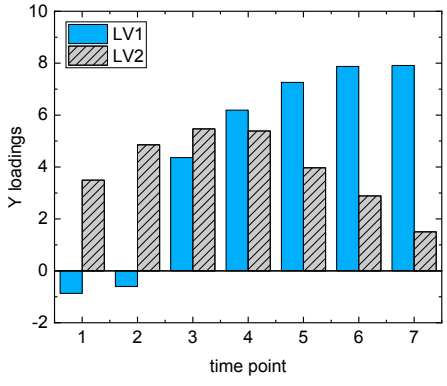

Fig. 5. MW-PLS with variable selection for the estimation of product titer time profile from the dynamics of metabolic profiles: **Y** loading plot of one cross-validation iteration.

MW-PLS scores show in a single point the entire dynamics of metabolic profiles, and together with **Y** loadings allow to understand the relationship between the dynamics of metabolic profiles and the product titer at each time point. Figure 6 shows that the first LV relates a large portion of **X** variability (21.2%) to 52.9% of **Y** variability, capturing the largest part of metabolic variability related to product titer. The second LV relates a very small portion of **X** variability (3.5%) to one fifth of **Y** variability (21.1%).

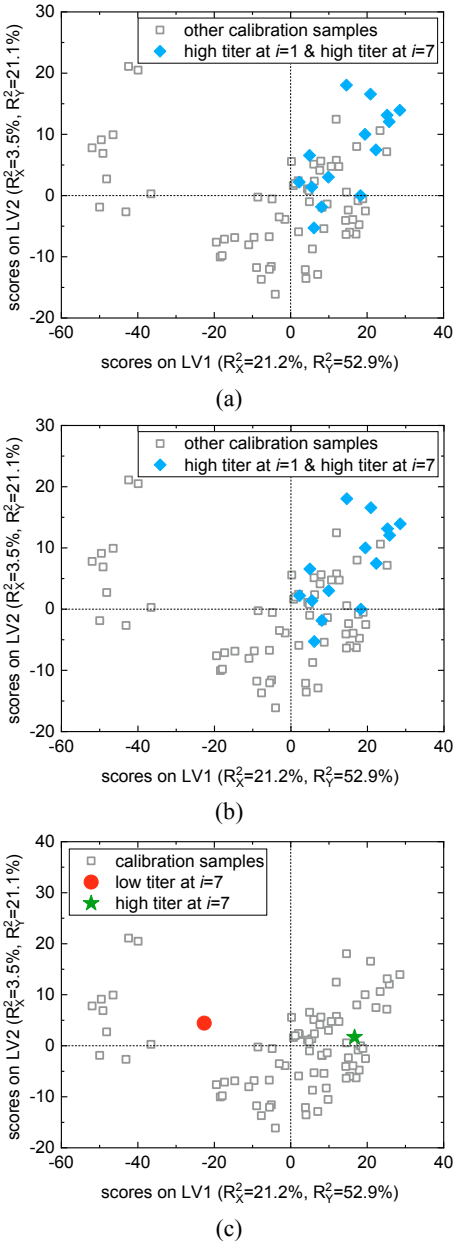

Fig. 6. MW-PLS with variable selection for the estimation of product titer time profile from the dynamics of metabolic profiles, score space of the first two LVs: (a) calibration and validation samples, (b) mapping of cell lines with high titer at time points 1 and 7, and (c) mapping of cell lines with high titer at time point 1 and low titer at time point 7.

The score space can be effectively used to map the cell lines according to product titer time profile. Figure 6a shows that the cell lines with high product titer at time points 1 and 7 are typically located in the first quadrant. These samples generate a compact cluster (blue diamonds), indicating high similarity in the metabolic profiles dynamics. Similarly, Figure 6b shows that the cell lines with low product titer at time point 1, but high product titer at time point 7 are typically located in the fourth quadrant. This group of cells (orange triangles) achieves

high product titer at the end of the culture, despite beginning to produce later during the culture.

Finally, it should be highlighted that the proposed model can correctly map the validation samples, namely, new unknown cell lines that are not included into the model. As an example, two validation cell lines are projected onto the score space (Figure 6c). The cell line exhibiting high product titer at time point 1 and low product titer at time point 7 (red dot) is correctly mapped in the third quadrant. Similarly, the cell line exhibiting average titer at time point 1 and high product titer at time point 7 (green star) is correctly mapped in the first quadrant.

4. CONCLUSIONS

In this work we proposed a multivariate multi-way approach to deal with metabolomics time profiles to allow an informed screening and an early selection of the cell lines that are good candidates to be progressed in the scale-up because they exhibit high productivity at the Ambr® 15 scale. The proposed methodology aids to significantly speed up the screening process, since low performing cultures can be aborted before their natural end, while the early estimation of product end-point titer allows a 50% reduction of the experiment duration. Furthermore, the estimation performance of the product titer time profile from the dynamics of metabolic profiles guaranteed a better understanding of the relationship between biological information (i.e., the evolution of cell metabolism) and process performance (product titer in this case). The proposed approach is general and can be easily extended to other biopharmaceutical processes.

Future developments will be oriented to the identification of the metabolic pathways and the networks of cellular reactions that characterize a product of desired quality attributes. In particular, the goal will be that of providing a methodology for interpreting the parameters of the data-based model in order to understand the relation between process behavior and biological functions.

REFERENCES

Afanador, N.L., Tran, T.N. and Buydens, L.M.C. (2013). Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression. *Analytica Chimica Acta*, 768(1), 49–56.

Duvigneau, S., Dürr, R., Laske, T., Bachmann, M., Dostert, M. and Kienle, A. (2020). Model-based approach for predicting the impact of genetic modifications on product yield in biopharmaceutical manufacturing—Application to influenza vaccine production. *PLoS computational biology*, 16(6), e1007810

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. and Wold, S. (2006). *Multi-and megavariate data analysis*. Umetrics Ab, Umea.

Facco, P., Zomer, S., Rowland-Jones, R.C., Marsh, D., Diaz-Fernandez, P., Finka, G., Bezzo, F. and Barolo, M. (2020). Using data analytics to accelerate biopharmaceutical process scale-up. *Biochemical Engineering Journal*, 164(April), 107791.

Food and Drug Administration. (2004). *Guidance for Industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance*.

Frederick, D.W., McDougal, A.V., Semenas, M., Vappiani, J., Nuzzo, A., Ulrich, J.C., Becherer, J. D., Preugschat, F., Stewart, E.L., Sévin, D.C. and Kramer, H.F. (2020). Complementary NAD+ replacement strategies fail to functionally protect dystrophin-deficient muscle, *Skeletal Muscle*, 10, 30.

Fuhrer, T., Heer, D., Begemann, B. and Zamboni, N. (2011). High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry, *Analytical Chemistry*, 83(18), 7074-7080.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1–17.

Gregersen, L. and Jørgensen, S.B. (1999). Supervision of fed-batch fermentations, *Chemical Engineering Journal*, 75(1), 69-76.

Hong, M.S., Severson, K.A., Jiang, M., Lu, A.E., Love, J.C. and Braatz, R.D. (2018). Challenges and opportunities in biopharmaceutical manufacturing control. *Computers and Chemical Engineering*, 110, 106–114.

Karst, D.J., Scibona, E., Serra, E., Bielser, J.M., Souquet, J., Stettler, M., Broly, H., Soos, M., Morbidelli, M. and Villiger, T. K. (2017). Modulation and modeling of monoclonal antibody N-linked glycosylation in mammalian cell perfusion reactors. *Biotechnology and Bioengineering*, 114(9), 1978–1990.

Li, F., Vijayasankaran, N., Shen, A.Y., Kiss, R. and Amanullah, A. (2010) Cell culture processes for monoclonal antibody production, *mAbs*, 2(5), 466-479.

Mehmood, T., Liland, K.H., Snipen, L. and Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chem. Int. Lab. Sys.*, 118, 62-69.

Nomikos, P. and MacGregor, J. F. (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, 30(1), 97–108.

Ramaker, H.J., Van Sprang, E.N.M., Westerhuis, J.A. and Smilde, A.K. (2005). Fault detection properties of global, local and time evolving models for batch process monitoring. *Journal of Process Control*, 15(7), 799–805.

Rameez, S., Mostafa, S.S., Miller, C. and Shukla, A.A. (2014). High-throughput miniaturized bioreactors for cell culture process development: Reproducibility, scalability, and control. *Biotechnology Progress*, 30(3), 718–727.

Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.

Zhou, B., Xiao, J.F., Tuli, L. and Ressom, H.W. (2012). LC-MS-based metabolomics. *Molecular BioSystems*, 8(2), 470–481.

Zürcher, P., Sokolov, M., Brühlmann, D., Ducommun, R., Stettler, M., Souquet, J., Jordan, M., Broly, H., Morbidelli, M. and Butté, A. (2020). Cell culture process metabolomics together with multivariate data analysis tools opens new routes for bioprocess development and glycosylation prediction. *Biotechnology Progress*, (December 2019), 1–11.