



Article

Data Augmentation to Support Biopharmaceutical Process Development through Digital Models—A Proof of Concept

Andrea Botton, Gianmarco Barberi  and Pierantonio Facco * 

CAPE-Lab—Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, Via Marzolo 9, 35131 Padova, Italy

* Correspondence: pierantonio.facco@unipd.it

Abstract: In recent years, monoclonal antibodies (mAbs) are gaining a wide market share as the most impactful bioproducts. The development of mAbs requires extensive experimental campaigns which may last several years and cost billions of dollars. Following the paradigm of Industry 4.0 digitalization, data-driven methodologies are now used to accelerate the development of new biopharmaceutical products. For instance, predictive models can be built to forecast the productivity of the cell lines in the culture in such a way as to anticipate the identification of the cell lines to be progressed in the scale-up exercise. However, the number of experiments that can be performed decreases dramatically as the process scale increases, due to the resources required for each experimental run. This limits the availability of experimental data and, accordingly, the applicability of data-driven methodologies to support the process development. To address this issue in this work we propose the use of digital models to generate *in silico* data and augment the amount of data available from real (i.e., *in vivo*) experimental runs, accordingly. In particular, we propose two strategies for *in silico* data generation to estimate the endpoint product titer in mAbs manufacturing: one based on a first principles model and one on a hybrid semi-parametric model. As a proof of concept, the effect of *in silico* data generation was investigated on a simulated biopharmaceutical process for the production of mAbs. We obtained very promising results: the digital model effectively supports the identification of high-productive cell lines (i.e., high mAb titer) even when a very low number of real experimental batches (two or three) is available.

Keywords: data augmentation; monoclonal antibodies; bioprocess development; digitalization; machine learning; hybrid modeling; first principles modeling



Citation: Botton, A.; Barberi, G.; Facco, P. Data Augmentation to Support Biopharmaceutical Process Development through Digital Models—A Proof of Concept. *Processes* **2022**, *10*, 1796. <https://doi.org/10.3390/pr10091796>

Academic Editor: Alina Pyka-Pajak

Received: 28 July 2022

Accepted: 2 September 2022

Published: 6 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Monoclonal antibodies (mAbs) are a class of recombinant proteins utilized against human immunological and oncological diseases, which are typically produced at the industrial level in fed-batch cultures of mammalian cells, engineered to secrete the protein of interest [1]. In the last few years, mAbs are gaining a lot of interest: they comprise over one-half of the biopharmaceutical approvals by regulatory agencies, and their market passed the threshold of USD 120 billion in annual sales [2] expecting to reach USD 140 billion in 2024 [3]. However, the development of new monoclonal antibodies is a time-consuming and resource-intensive procedure [1,4], which usually requires many years and large investments from biopharmaceutical companies [5,6]. In fact, experiments on mammalian cells may last several weeks and cost tens of thousands of dollars each. For this reason, the number of performed experimental runs is often limited. Furthermore, while scaling up the process, the number of experiments gradually decreases because the cost of a single experimental run increases with the process volume. Hence, the number of experimental runs decreases from several dozens, if not hundreds, at the milliliter scales to 12–24 at a shake-flask scale, while a couple of runs only are typically performed at the pilot scale [4].

Following the wave of digitalization in Industry 4.0, large amounts of data (e.g., culture variables from high throughput technologies [7], and omics data such as transcriptomics [8] or metabolomics [9]) are usually collected from all the stages of the scale-up. The wealth of information contained in the experimental data can be extracted to support the mAbs development through machine learning [9,10]. In particular, different data-driven techniques were demonstrated to be effective to: (i) understand the similarity among bioreactors at different scales and improve the similarity between scales in the scaled-down [11]; (ii) predict the mAbs concentration at harvest allowing to identify the parameters that promote or suppress production [12]; (iii) estimate the mAbs quality and interpret the relationship between process and product when coupled with genetic algorithms [13]; and (iv) capture very complex biological relationships through neural networks coupled with first principles models of the culture environment and accurately predict the mAbs quality attributes [14]. Despite their efficacy, data-driven methods suffer when the number of available data is limited [15]. In this case, the main driving forces and correlations in the data cannot be reliably captured due to sample underrepresentation and the large biological variability. Furthermore, the estimation performance of data-driven models degrades with few data, and models become prone to overfitting and sensitive to outliers [16]. For this reason, the industrial practice is to switch to univariate modeling [16]. Since biopharmaceutical processes are intrinsically multivariate, univariate techniques provide only a poor representation of the system under investigation and may fail to understand the complex correlation among critical process parameters (CPPs) and critical quality attributes (CQAs) [17]. For this reason, elaborating alternative strategies to overcome the limitation of a restricted amount of data from few experiments is of paramount importance to accelerate the process/product development without increasing the experimental burden.

In this respect, the generation of *in silico* data is a possible solution to the limited data problem. For example, in the fields of artificial intelligence and image processing [18,19], data augmentation was successfully applied to industrial microelectronic and chemical processes [20,21]. *In silico* data may be generated artificially either by perturbing the available data points or by combining them if no prior knowledge of the process is available. The data augmentation by means of perturbation can be performed simply by adding Gaussian noise to the available data points [22,23]. Furthermore, the available data can be linearly combined to generate new artificial samples [24]. As an alternative, prior process knowledge can be exploited for the purpose of data augmentation and *in silico* batch generation by building a digital version of the process. For example, a hybrid mechanistic-empirical model was built to explore different settings and scenarios for a large-scale fed-batch mammalian cell culture producing a therapeutic antibody [25]. A Gaussian process state-space model coupled with a resampling from the high-frequency acquisition system was used to generate *in silico* samples and improve the multivariate monitoring of biopharmaceutical batch processes [26]. Moreover, generative adversarial neural networks were used to generate *in silico* single-cell RNA sequence data for biomedical research [27].

Despite considerable effort being made to solve the problem of limited data availability, the research and application of *in silico* model-based data generation in the biopharmaceutical industry is still an open issue. In this field, overcoming the limited availability of data in a digital manner can significantly reduce the experimental burden and development timelines, allowing for a reduction in the cost of life-saving drugs and making them available to patients earlier.

In this work, we show how, in the development of monoclonal antibodies, the application of different strategies for *in silico* batch generation can improve the identification of cell lines with the desired CQA (i.e., high mAb titer) in the scenario of limited available data. Specifically, we propose the use of two approaches based on the following digital models: a first principles model [28], and a hybrid semi-parametric model [29]. The proposed methods for data augmentation will be applied to the case study of a simulated process for mammalian cell culture [30] for the purpose of improving the estimation of mAb titer.

The rest of the paper is organized as follows: Section 2 describes the general framework of the proposed procedure for in silico data generation, the (simulated) process, the digital models used for in silico batch generation, and the multivariate modeling used for the estimation of mAb titer at harvest; Section 3 reports the mAb titer estimation performance in a data-poor scenario and the capability of understanding the process evaluated for both in silico data generation strategies; and Section 4 contains the final remarks and future perspectives of this study.

2. Materials and Methods

2.1. Methodological Procedure

The methodological procedure for the in silico data augmentation with digital models (Figure 1) goes through the following steps:

- Step 1—Experimental campaign on the mAbs production process: batch data are obtained from experiments performed on the development scale of the process under study according to the availability of resources. In this work, we consider a simulated process for the production of mAbs at the shake-flask scale (Section 2.2);
- Step 2—In silico batch generation from a digital model: data on real batches are utilized through digital models of the process to drive the generation of in silico batches with a wider variety of behaviors. In particular, two alternative modeling strategies are adopted in this work: a first principle digital model (Section 2.3) and a hybrid digital model (Section 2.4);
- Step 3—Multivariate data-based modeling: all the available data (both the ones from the process and the ones generated in silico) are fed to a data-based model to support the process development and scale-up. In this work, process and in silico generated batches are regressed to estimate a CQA (i.e., mAb titer at harvest) through multivariate latent variable modeling (Section 2.5). In this way, the multivariate models exploit the data of a few process batches and the additional process knowledge extracted from the in silico generated batches, to make estimations of cell behavior for new samples from the culture variable time trajectories. Such estimations, especially in the presence of biological variability in the batches, are not feasible with the digital models of the process, which can only estimate the culture variable trajectories when the inputs (i.e., process initial conditions, feed composition, and scheduling) are manipulated given the biological characteristics already hardcoded in the digital model parameters.

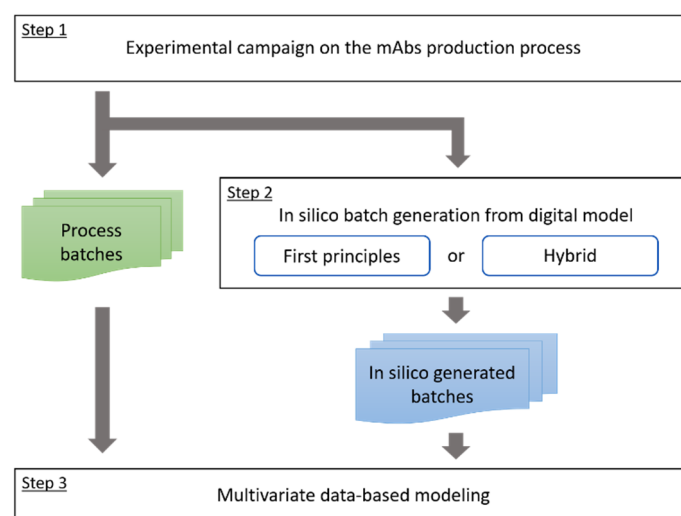


Figure 1. Methodological procedure for in silico data augmentation from digital models.

2.2. Process for the Production of Monoclonal Antibodies

We consider a simulated cell culture process for the production of mAbs in fed-batch mode at a shake-flask scale. The process is based on the well-established human embryonic

kidney (HEK) cell first principles model [30]. The available culture variables are: viable cell concentration (VCC); nutrients (i.e., glucose and glutamine); by-products concentrations (i.e., lactate and ammonia); and mAbs titer (i.e., antibody concentration).

The variability among batches lies in the different cultured cell lines, which are simulated to display different specific productivity, $Q_P = [mAb]_T / \int_0^T [X_v] dt$, where $[mAb]_T$ is the mAbs titer at harvest and $[X_v]$ is the viable cell concentration along the batch (cell/L). For this purpose, the HEK model parameters are sampled from normal distributions with mean and standard deviation reported in Appendix A Table A1. These values are adjusted from the reference parameters found in the HEK model reference [30] in such a way as to obtain a variability of the batch time trajectories that mimic the dynamic behavior of real experimental batches at that scale. Furthermore, measurement error is simulated by adding ~6% white noise to the culture variables' profiles, accounting for the typical measurement uncertainty of analytical equipment.

An experimental campaign is carried out in 0.2 L cultures with an inoculation seed density of 2×10^8 cell/L. The initial media composition is set to 25.1 mM of glucose and 5.1 mM of glutamine. Feeding is performed every 20 h starting from 10 h after cell seeding by feeding 0.00875 L in 10 min. The feed composition is set to 50 mM of glucose and 10 mM of glutamine. The measurement sampling is performed prior to the feeding through a 0.0015 L withdrawal from the culture in 10 min, resulting in 10 measurement sampling points during the batch.

All the considered experimental batches satisfy the following conditions: (i) the final mAbs titer is below 5000 mg/L; (ii) the peak of VCC is reached after 50 h; and (iii) the specific productivity is in the range 0–20 pg/(cell·day).

The available data are concerned with 100 batches, which are organized in: matrix $\mathbf{X}_{PC} = [100 \text{ batches} \times 5 \text{ variables} \times 10 \text{ time points}]$ that contains the time profiles of all the culture variables; and vector $\mathbf{y}_{PC} = [100 \times 1]$ that contains the mAbs titer at harvest (time point 10). These data are used in different ways to calibrate digital and multivariate models. Similarly, 10 validation batches are available and organized in the matrix $\mathbf{X}_{PV} = [10 \times 5 \times 10]$ for process data and vector $\mathbf{y}_{PV} = [10 \times 1]$ for mAbs titer. These validation batches are used to test the estimation performance of the multivariate models.

In this study, a simulated process is selected, not only because it reduces the time and cost of the experimental campaign, but also because it allows a full knowledge of the relationship between CPPs and CQAs, and better control of both the process behavior and the biological diversity in the experiments.

2.3. Modeling Strategy 1: First Principles Digital Model

The first principles digital model (FPDM) is a modified version of the simplified mathematical model proposed by del Val et al. (2016) describing a fed-batch mAbs production process [28]. The culture variables described by the FPDM are VCC, glucose, lactate, and mAbs titer. The FPDM is modified with respect to the original model to better resemble the process. In fact, in the original model [28] cells grow until glucose is available in the culture and this causes a substantial difference between the behavior of the model and the process. This makes the original model unusable for the generation of batches that conform to the ones of the process. Accordingly, we added a simplified material balance for glutamine, and introduced growth limitation at low glutamine concentration and glucose consumption limitation at reduced cell growth.

The simplified material balance for glutamine is defined as:

$$\frac{d(V_c c_{GLN})}{dt} = - \left(\frac{\mu_g}{Y_{x,glN}} \right) X_v V_c \quad (1)$$

where V_c is the liquid volume in the culture (L), c_{GLN} is the glutamine concentration (mM), μ_g is the specific growth rate (h^{-1}), and $Y_{x,glN}$ is the cell yield on glutamine (cell/mmol).

In order to account for the effect of glutamine on cell growth, a limiting factor f_{lim} is added to the specific growth rate expression:

$$\mu_g = \mu_{g,max} \left(\frac{c_{GLC}}{K_{m,glc} + c_{GLC}} \right) - \frac{X_v}{\alpha_x} f_{lim} \quad (2)$$

where $\mu_{g,max}$ is the maximum specific growth rate (h^{-1}), $K_{m,glc}$ is the Monod constant for the growth on glucose (mM), α_x is the cellular carrying capacity (cell/mmol), and c_{GLN} is the glutamine concentration (mM). The limiting factor f_{lim} is defined as:

$$f_{lim} = \frac{c_{GLN}}{c_{GLN} + k_{gln}} \quad (3)$$

where k_{gln} is the Monod constant for glutamine (mM). The limiting factor f_{lim} decreases with the glutamine concentration, reducing cell growth when the glutamine decreases.

To limit the glucose consumption with reduced cell growth, the glucose material balance is modified as:

$$\frac{d(V_c c_{GLC})}{dt} = F_{in} c_{GLC,in} - F_{out} c_{GLC} - q_{glc} [X_v] V_c (f_{lim} + K_{glc}) \quad (4)$$

where $c_{GLC,in}$ is the glucose concentration in the feeding stream, F_{in} and F_{out} are the inlet and outlet flow rates of the bioreactor (L/h), respectively, q_{glc} is the specific glucose consumption rate (mmol/(cell·h)) and K_{glc} (–) is the glucose maintenance constant.

In Silico Batch Generation through First Principles Digital Model

The FPDM is used for in silico batch generation. The reference parameters for FPDM are estimated from the reference process batch (i.e., obtained using the reference process parameters from Kontoravdi et al., 2010 [30]). In silico batches are generated by sampling the parameter values from a normal distribution with mean and standard deviation reported in Appendix B Table A2. These distributional parameters are heuristically determined to generate batches with a variability slightly larger than the one observed in the process batches. An example of in silico generated batches is reported in Figure 2: Figure 2a shows the time profile along the entire batch duration for viable cells concentration, and Figure 2b shows the time profile along the entire batch duration for mAbs concentration.

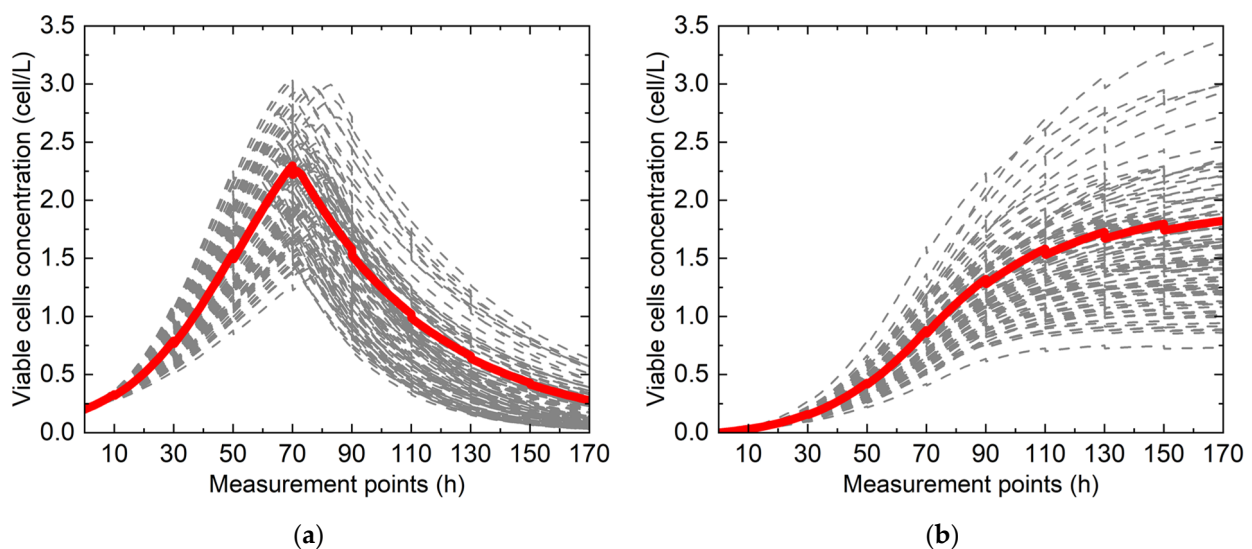


Figure 2. Example of batches generated in silico through the FPDM: (a) VCC profiles and (b) mAbs titer profiles for 100 batches. The thick red continuous lines represent the reference batch estimated from the process while the grey dashed lines represent the simulated ones.

This strategy is used to generate 100 in silico batches. The generated variables profiles are subsampled in the same 10 time points in which the process measurements are available. The resulting data are organized in matrix $\mathbf{X}_{\text{FPDM}} = [100 \times 4 \cdot 10]$, which contains the time profiles of the culture variables, and vector $\mathbf{y}_{\text{FPDM}} = [100 \times 1]$, which contains the mAbs titer at harvest.

2.4. Modeling Strategy 2: Hybrid Digital Model

The hybrid digital model (HDM) is a hybrid semi-parametric model [29,31–33], whose considered culture variables are VCC, glucose, glutamine, lactate, ammonia, and mAbs titer.

The HDM has the serial structure [34,35] reported in Figure 3, with a mechanistic section describing the material balances of the chemical species and an artificial neural network (ANN) [36] to estimate the complex and unknown kinetic expressions from cell culture experimental data.

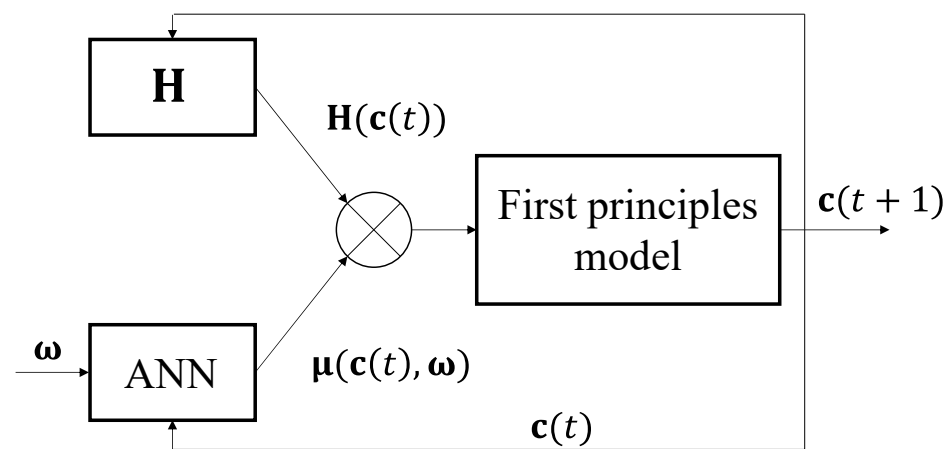


Figure 3. Structure of the hybrid digital model to generate in silico batches.

The HDM comprises the material balances for the culture variables of interest $\mathbf{c} [V \times 1] = [c_{X_V}, c_{\text{GLC}}, c_{\text{GLN}}, c_{\text{LAC}}, c_{\text{AMM}}, c_{\text{mAb}}]$ as:

$$\frac{d}{dt} \begin{bmatrix} X_V \\ c_{\text{GLC}} \\ c_{\text{GLN}} \\ c_{\text{LAC}} \\ c_{\text{AMM}} \\ c_{\text{mAb}} \end{bmatrix} = \mu_{\max} \begin{bmatrix} X_V & 0 & 0 & 0 & 0 & 0 \\ 0 & -X_V & 0 & 0 & 0 & 0 \\ 0 & 0 & -X_V & 0 & 0 & 0 \\ 0 & 0 & 0 & X_V & 0 & 0 \\ 0 & 0 & 0 & 0 & X_V & 0 \\ 0 & 0 & 0 & 0 & 0 & X_V \end{bmatrix} \begin{bmatrix} \mu_{X_V} \\ \mu_{\text{GLC}} \\ \mu_{\text{GLN}} \\ \mu_{\text{LAC}} \\ \mu_{\text{AMM}} \\ \mu_{\text{mAb}} \end{bmatrix} \quad (5)$$

$$= \mu_{\max} \mathbf{H}(\mathbf{c}) \boldsymbol{\mu}(\mathbf{c}^*, \boldsymbol{\Omega}) \mathbf{x}$$

where μ_{\max} [37] is the vector of the maximum specific rates of production/consumption for each culture variable (reported in Appendix B Table A3), $\boldsymbol{\mu}(\mathbf{c}^*, \boldsymbol{\Omega}) = [\mu_{X_V}, \mu_{\text{GLC}}, \mu_{\text{GLN}}, \mu_{\text{LAC}}, \mu_{\text{AMM}}, \mu_{\text{mAb}}]$ is the vector of the specific production/consumption rates estimated by the ANN, $\mathbf{H}(\mathbf{c}) [V \times V]$ contains the known kinetic expressions, and $\mathbf{c}^* = [X_V, c_{\text{GLC}}, c_{\text{GLN}}, c_{\text{LAC}}, c_{\text{AMM}}]$ is the reduced concentration vector used as input for the ANN.

The matrix $\mathbf{H}(\mathbf{c})$ contains all the known mechanistic information for the calculation of the reaction rates in Equation (5). In this work, the known mechanistic part of the reaction rates, $\mathbf{H}(\mathbf{c})$, has no fitted parameters. $\mathbf{H}(\mathbf{c})$ accounts for the dependence of the reaction rates on the cell concentration X_V and the apparent stoichiometric coefficient, which indicates if a metabolite is produced or consumed. The maximum specific rates of production/consumption μ_{\max} are constant parameters heuristically set in preliminary studies to appropriately scale the ANN outputs in the desired experimental ranges.

The vector of specific production/consumption rates is modeled through an artificial neural network. In particular, a two-layer ANN is used to estimate the specific production and consumption rates from the reduced concentration vector \mathbf{c}^* . The selected ANN has a 10-neurons hidden layer with a hyperbolic-tangent activation function and a linear output layer with 6 neurons (i.e., given by the dimension of $\boldsymbol{\mu}$). The mathematical expression of the ANN is:

$$\boldsymbol{\mu}(\mathbf{c}^*, \boldsymbol{\Omega}) = \boldsymbol{\omega}^{(2)} \tanh(\boldsymbol{\omega}^{(1)} \mathbf{c}^* + \boldsymbol{\omega}_0^{(1)}) + \boldsymbol{\omega}_0^{(2)} \quad (6)$$

where $\boldsymbol{\omega}$ is the weight vector, $\boldsymbol{\omega}_0$ is the bias vector and the superscript (1) and (2) refer to the hidden and output layer, respectively. In this case, it is assumed that the specific production/consumption rates do not depend on the mAbs titer while depending on the number of cells, nutrients, and by-products in the culture.

The hybrid model identification is performed through the sensitivity method [31,35], by backpropagating the errors in the concentration space through the model. In this work, the normalized sum of squared errors (SSE) between the measured concentrations c_v and the ones calculated by the HDM, \hat{c}_v , is directly minimized as:

$$\operatorname{argmin}(\text{SSE}) = \operatorname{argmin} \left(\sum_{t=1}^T \sum_{v=1}^V \frac{(\hat{c}_v(t) - c_v(t))^2}{\sigma_v} + \lambda_{\text{reg}} \|\boldsymbol{\mu}\| \right) \quad (7)$$

where σ_v is the standard deviation of the v -th process variable calculated over the training data, c_v is the measured concentration of the v -th culture variable at the t -th time instant, \hat{c}_v is the calculated concentration of the v -th culture variable at the t -th time instant, and λ_{reg} is a regularization term [38], which is added to aid training convergence. In this work, the error backpropagation is performed by calculating the gradient of the concentration errors (i.e., SSE) with respect to the ANN weights, because the hybrid model does not contain any mechanistic parameter to be fitted and the only adjustable parameters are the ANN weights. The gradient of the concentration errors (i.e., SSE) with respect to the ANN weights is calculated as:

$$\frac{\partial \text{SSE}}{\partial \boldsymbol{\Omega}} = \sum_{t=1}^T \sum_{v=1}^V (\hat{c}_v(t) - c_v(t))^2 \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\Omega}} \right)_t \quad (8)$$

where $(\partial \mathbf{c} / \partial \boldsymbol{\Omega})_t$ is the gradient of the concentrations with respect to the ANN weights, calculated with the sensitivity method [31].

An Adam optimizer [39] is then used to adjust the ANN parameters according to the calculated gradient. The hybrid model is trained for 400 iterations with a learning rate $\eta = 10^{-3}$, and subsequent 300 iterations with a learning rate $\eta = 10^{-4}$. Prior to the training, the ANN weights are initialized by sampling from a normal distribution $N(0, \sigma)$ where $\sigma = 0.01$.

The integration of the HDM is performed stepwise between the feeding time points. A bolus feeding of glucose and glutamine (consistent with the training batches) is simulated by updating the initial concentration after the feeding according to [29]:

$$c_f(t^+) = c_f(t^-) + \Delta c_f(t) \quad (9)$$

where $c_f(t^+)$ is the concentration of nutrient f after the feeding, $c_f(t^-)$ is the concentration of the nutrient f before the feeding and $\Delta c_f(t)$ is the change in concentration of the nutrient f (i.e., GLC or GLN) due to the feeding at time instant t .

2.4.1. In Silico Batch Generation through Hybrid Digital Model

The HDM is used for in silico batch generation as an alternative to FPDM. First, the HDM is trained on 10 batches [40], which are selected from \mathbf{X}_{PC} to cover a sufficient range of process variability. Then, in silico batches are generated by changing the maximum specific rate of production/consumption $\boldsymbol{\mu}_{\text{max}}$, which is kept constant during training.

The values of μ_{\max} are sampled from a normal distribution with mean and standard deviation reported in Appendix B Table A3. These values are heuristically selected, based on preliminary tests, to cover a sufficiently large variability around the process batch profiles, while preserving similarity with them. An example of the batches generated by the HDM is shown in Figure 4.

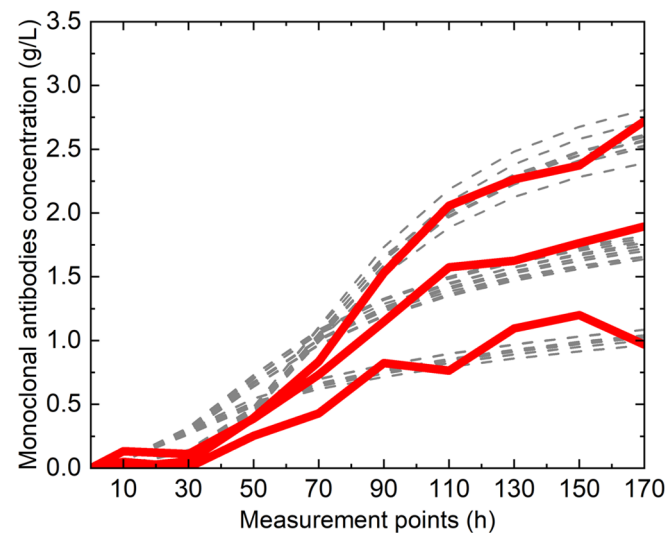


Figure 4. Example of batches generated through HDM: mAbs titer profiles. In this example, 10 batches are generated from 3 training batches taken from the process. The thick red continuous lines represent the training batches while the grey dashed lines represent the simulated ones.

This strategy is used to generate 100 in silico batches, 10 from each batch used to train the HDM. The generated variables profiles are subsampled in the same 10 time points in which the process measurements are available. The resulting data are organized in $\mathbf{X}_{\text{HDM}} = [100 \times 5 \times 10]$, which contains the time profiles of the culture variables, and vector $\mathbf{y}_{\text{HDM}} = [100 \times 1]$, which contains the mAbs titer at harvest.

2.5. Multivariate Predictive Modeling

In this study, multi-way partial least squares regression (MPLS) [41] is used to estimate the CQA, namely, the mAbs titer at harvest, from the multi-dimensional datasets of the correlated culture variables (both real and generated in silico) time trajectories.

MPLS consists of a proper unfolding of the data followed by standard PLS modeling.

Batch-wise unfolding is performed in this study to capture the correlation between the culture variables' time profiles and the response together with the cross-correlation between culture variables at different time points. In batch-wise unfolding, the two-dimensional slices at each time point $k = 1, 2, \dots, K$ of the matrix $\mathbf{X} [N \times V \times K]$, $\mathbf{X}_k [N \times V]$, where N is the number of batches and V is the number of variables, are horizontally concatenated, resulting in two-dimensional matrix $\mathbf{X} = [N \times V \cdot K] = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k, \dots, \mathbf{X}_K]$. Accordingly, the matrices \mathbf{X}_{PC} , \mathbf{X}_{PV} , \mathbf{X}_{FPDM} and \mathbf{X}_{HDM} are unfolded in the bidimensional matrices: $\mathbf{X}_{\text{PC}} = [100 \times 5 \cdot 10]$, $\mathbf{X}_{\text{PV}} = [10 \times 5 \cdot 10]$, $\mathbf{X}_{\text{FPDM}} = [100 \times 4 \cdot 10]$, $\mathbf{X}_{\text{HDM}} = [100 \times 5 \cdot 10]$.

Partial least squares regression (PLS) [42] is a multivariate statistical linear regression technique that identifies the directions of maximum covariance between a regressor matrix $\mathbf{X} = [N \times V \cdot K]$ and a response matrix $\mathbf{Y} = [N \times M]$, where M is the number of response variables. PLS decomposes both the regressor and response matrices into a common latent space of orthogonal latent variables (LVs). In this study, \mathbf{X} and \mathbf{Y} are auto-scaled to zero mean and unit variance (i.e., by subtracting to each column its mean value and dividing each column by its standard deviation). PLS decomposes the auto-scaled matrices \mathbf{X} and \mathbf{Y} as:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}, \quad (10)$$

and

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F}, \quad (11)$$

with

$$\mathbf{T} = \mathbf{X}\mathbf{W}^*, \quad (12)$$

where $\mathbf{T} = [\mathbf{N} \times \mathbf{A}]$ is the scores matrix that captures the relationships among batches according to the features of the covariance between \mathbf{X} and \mathbf{Y} ; $\mathbf{P} = [\mathbf{V} \cdot \mathbf{K} \times \mathbf{A}]$ and $\mathbf{Q} = [\mathbf{M} \times \mathbf{A}]$ are the loadings matrices which capture the relationships among the variables' dynamics in \mathbf{X} and variables in \mathbf{Y} , respectively; $\mathbf{E} = [\mathbf{N} \times \mathbf{V} \cdot \mathbf{K}]$ and $\mathbf{F} = [\mathbf{N} \times \mathbf{M}]$ are the residual matrices for \mathbf{X} and \mathbf{Y} , respectively, which contain the information that is not described by the model; \mathbf{W}^* is the weights matrix, which directs the scores to be the most predictive for the response \mathbf{Y} ; \mathbf{A} is the number of selected LVs; and the superscript \mathbf{T} represents the transpose operation. In this work, the selected number of latent variables is $\mathbf{A} = 2$ which minimizes the estimation error of the responses in cross-validation [43].

PLS is used to estimate the response variable $\hat{\mathbf{Y}}$ for a set of \mathbf{I} new batches, whose predictors $\mathbf{X}_{\text{new}} = [\mathbf{I} \times \mathbf{V} \cdot \mathbf{K}]$ are known, from:

$$\hat{\mathbf{Y}} = \mathbf{X}_{\text{new}}\mathbf{W}^*\mathbf{Q}^T, \quad (13)$$

To improve PLS estimations, variable selection [9,44] is used, in such a way as to identify and retain in the model only the variables with the largest information content on the mAbs titer and exclude the other variables. Variable importance is assessed through the variable importance in projection (VIP) [45] index:

$$\text{VIP}_v = \frac{\sqrt{\mathbf{V} \cdot \mathbf{K} \cdot \sum_{a=1}^{\mathbf{A}} \mathbf{R}_{\mathbf{Y},a}^2 \mathbf{w}_{v,a}^2}}{\sqrt{\sum_{a=1}^{\mathbf{A}} \mathbf{R}_{\mathbf{Y},a}^2}} \quad (14)$$

where $\mathbf{R}_{\mathbf{Y},a}^2$ is the \mathbf{Y} variance captured by the a -th latent variable and $\mathbf{w}_{v,a}$ is the weight corresponding to the a -th LV and v -th \mathbf{X} variable. In this work, the selection of variables with $\text{VIP} > 1$ is performed over a 100-iteration Monte Carlo cross-validation; only variables with high selection frequency (i.e., 80% of the iterations with $\text{VIP} > 1$) are considered informative for the estimation and used to recalibrate the model.

The mAbs titer estimation performances are evaluated through the mean absolute estimation error (MAPE):

$$\text{MAPE}_m = \frac{\sum_{i=1}^{\mathbf{I}} |y_{m,i} - \hat{y}_{m,i}|}{\mathbf{I}} \quad (15)$$

where $\hat{y}_{m,i}$ is the estimation of the m -th response variable for the i -th batch and $y_{m,i}$ is the measured value.

When process data only are utilized for the titer estimation the model calibration matrices \mathbf{X} and \mathbf{Y} are obtained from \mathbf{X}_{PC} and \mathbf{y}_{PC} . As an alternative, when few data are available from process data and in silico generated batches are used in PLS modeling to augment the calibration dataset, data from the digital model \mathbf{X}_{FPDM} and \mathbf{y}_{FPDM} (or \mathbf{X}_{HDM} and \mathbf{y}_{HDM}) are vertically concatenated to the available process data in \mathbf{X}_{PC} and \mathbf{y}_{PC} to create augmented matrices \mathbf{X} and \mathbf{Y} . Hence, the number of batches that is used for the model calibration is much larger than the number of batches available from the process. Autoscaling is applied as a data normalization preprocessing directly on the augmented matrices \mathbf{X} and \mathbf{Y} . Note that in this study the process and the digital models have very similar statistical characteristics and separate preprocessing did not improve model performance. However, if in silico generated data showed different statistical characteristics with respect to process data, separate and specific preprocessing would be required.

3. Results and Discussion

The results are organized as follows. First of all, the mAbs titer at harvest estimation performance is presented when only the process batches are available and then compared to the performance when the in silico generated batches are present. Furthermore, the ability to identify the most influential CPPs for mAbs productivity is discussed critically for both the model on the process batches and the improved models with augmented data.

3.1. Monoclonal Antibodies Titer Estimation

In this section, we analyze the performance of an MPLS that estimates the mAbs titer at harvest when only the process batches are used (i.e., base case). Then, this model is compared to the one in which process data are augmented with the in silico batch data generated through the digital models.

3.1.1. Titer Estimation Performance and Sensitivity to the Available Number of Process Calibration Batches

Here, we analyze the estimation performance of the MPLS model and assess the sensitivity of its estimation performance to the number of process batches available for calibration.

For this purpose, we iteratively increase the number of calibration batches from 3 to 50, by randomly extracting them from X_{PC} and y_{PC} . This extraction is repeated 20 times for each number of calibration batches. At each step, a 2 LVs MPLS model is built with the available calibration batches and validated with X_{PV} and y_{PV} . The titer estimation performance for the validation dataset and its sensitivity to the number of calibration batches are examined in terms of MAPE (averaged over the 20 iterations) as a function of the number of batches used to calibrate the MPLS model (Figure 5). As expected, MAPE (black dashed line in Figure 5) decreases with an increasing number of calibration batches. In particular, with more than 20–25 calibration batches, the MAPE average stabilizes around 210 mg/L, which is a good estimation performance, because it is comparable to the measurement error of ~150 mg/L. The MAPE increases when less than 20 batches are used for calibration and reaches large values when the number of batches is lower than 10 (MAPE > 230 mg/L). Note that a substantial increase of the estimation error is observed exactly in the range of experimental runs typically performed at the shake-flask scale, which spans between 12 and 24. Due to these inaccurate estimations, the identification of cell lines meeting the target mAb titer to be progressed in the scale-up becomes much more difficult, especially when the number of available experimental runs approaches 10. Furthermore, it should be highlighted that with less than 10 batches the model performance is inaccurate.

Furthermore, the MAPE variability (in terms of the 95% confidence region of the Gaussian MAPE distribution over different iterations, grey shaded area in Figure 5) increases with a low number of available batches. This indicates that the lower the number of calibration batches, the more the estimation performances are erratic and depend on the batches included in the model. In fact, if the model is calibrated on a small number of batches, the limited portion of the wide process variability captured by the model is insufficient to correctly describe new batches whose operating conditions may be far from the ones of a limited calibration dataset.

For this reason, the generation of in silico batches could be valuable to widen the variability in calibration data and cover new portions of variability which cannot be included in a limited set of calibration batches. This will eventually improve the estimation performance, providing an invaluable benefit to the selection of high-productive cell lines, especially when the number of available batches is lower than 10. Since this case is often encountered in the biopharmaceutical industry at the scales of shake flasks and stirred bioreactors, where the typical number of available batches ranges between 1 and 8, in the following, we will focus on this range of batches available from the process.

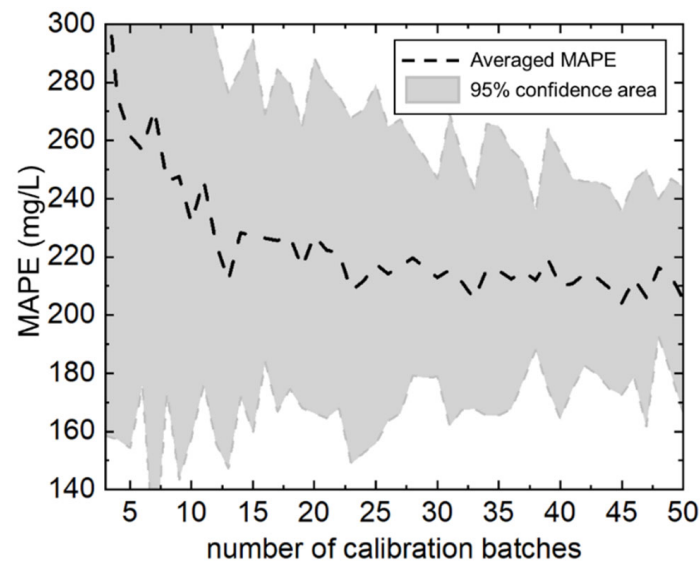


Figure 5. MPLS performance sensitivity to the available process calibration batches in the estimation of mAbs titer at harvest. Black dashed line—average validation MAPE (averaged over the 20 random selections of the calibration batches from the set of process batches) as a function of the number of calibration batches; grey area—95% confidence area of the distributions.

3.1.2. Effect of Data Augmentation on the Estimation Performance

In this section, we assess the sensitivity of the MPLS estimation performance to the number of calibration process batches when data are augmented through *in silico* batches.

For this purpose, we iteratively increase the number of process calibration batches from 1 to 8, randomly extracted from a subset of 10 batches contained in X_{PC} and y_{PC} (the same 10 batches used for the training of the HDM, Section 2.4.1) to inspect the range of available process batches in which unsatisfactory performances were observed in the base case (Section 3.1.1). The extraction is repeated 20 times for each number of process batches. At each step, a 2 LVs MPLS model is built with the available process batches concatenated either: (i) with 30 FPDM *in silico* generated batches randomly extracted from X_{FPDM} and y_{FPDM} ; or (ii) with the 10 HDM *in silico* batches from X_{HDM} corresponding to each process batch used in MPLS calibration. The number of *in silico* batches is selected to increase the variability in batch behavior without overwhelming the information provided by process data. At each repetition, the MPLS models are then validated with X_{PV} and y_{PV} . In all the cases, only the most important variables for the estimation are included in the models. Details about the selected variables will be given in the next section.

We compare the MAPE distributions in the estimation of the mAbs titer at harvest obtained through MPLS models built on: (i) process batches; (ii) process batches plus FPDM *in silico* generated batches; and (iii) process batches plus HDM *in silico* generated batches.

The MAPE distributions in the 20 repetitions are reported in Figure 6 as a function of the number of calibration process batches through boxplots. The boxes represent the 25° and 75° percentile with the median value; the dots represent the mean value of the MAPE; the error bars represent the 95% confidence intervals; and the diamonds represent errors outside the 95% confidence intervals. In Figure 6, green boxes represent the error distribution of the base case; red boxes represent the error distribution of the FPDM data augmentation strategy; and blue boxes represent the error distribution of the HDM data augmentation strategy.

In the base case, MAPE decreases with the number of available process batches, reaching ~180 mg/L when 8 process batches are used for model calibration (note that this value differs from Section 3.1.1 because the variable selection is applied here, indicating that variable selection improves the estimation performance).

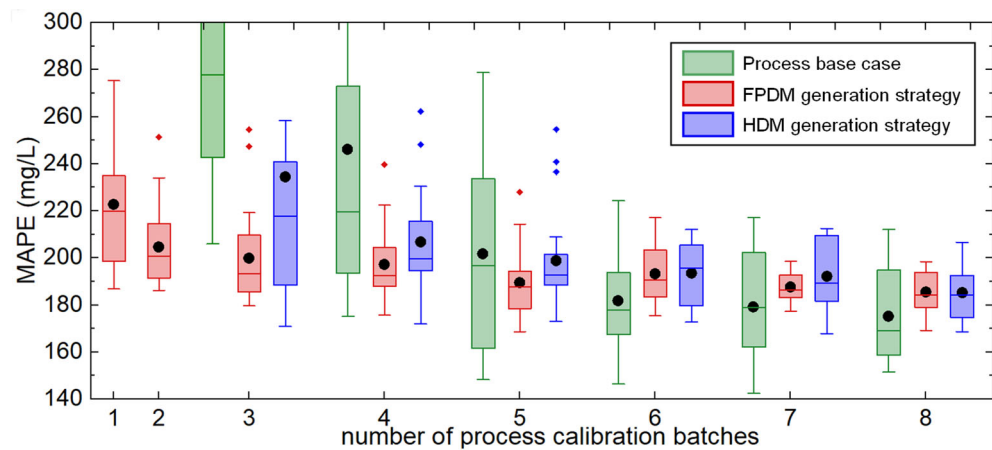


Figure 6. Validation estimation performance comparison: MAPE distribution profiles from a 20-repetitions validation in the estimation of mAbs titer at harvest through MPLS. Green boxes—process base case; red boxes—FPDM data augmentation strategy; blue boxes—HDM data augmentation strategy. The boxes represent the 25° and 75° percentile and the median value, the dots represent the mean value of the MAPE, the error bars represent the 95% confidence intervals, and the diamonds are errors outside the 95% confidence intervals.

When more than 5 process batches are available, both data augmentation strategies show similar performance ($170 < \text{MAPE} < 200$ mg/L), even if the lowest average error values are obtained in the process base case (down to ~ 150 mg/L). The addition of *in silico* batches considerably reduces the variability of the estimation error with respect to the process base case, independently of the augmentation strategy. This indicates that the augmented number of batches helps to increase the estimation robustness and reduces the sensitivity of the performance to the specific calibration batches. However, the average estimation error slightly increases because the *in silico* batches present some differences from the process and add variability to the dataset.

By contrast, when 4 or 5 process batches are available, the addition of the simulated batches is highly beneficial. In fact, both FPDM and HDM augmentation strategies improve the estimation performance and reduce error variability ($170 < \text{MAPE} < 220$ vs. $150 < \text{MAPE} < 300$ mg/L). In this case, the FPDM augmentation strategy provides the largest improvement. When even less than 4 process batches are available, the FPDM augmentation strategy is very helpful for the mAbs titer estimation, because it allows better performances than both the base case and the HDM generation strategy. Good models can even be built when a very reduced number of process batches is available, namely fewer than 3 ($190 < \text{MAPE} < 250$ mg/L). In this case, the HDM augmentation strategy does not improve the estimation performance (results not shown) and provides errors that are similar to the ones of the base case ($300 < \text{MAPE} < 500$ mg/L). This is due to the high similarity between the process batches and the ones generated *in silico* through HDM.

These results show that the FPDM generation strategy allows to properly mimic the behavior of the process batches and identify the batches with high mAb titer to be progressed in the scale-up. This is because it improves the multivariate regression model estimation performance and increases the captured variability independently on process batches availability. The HDM augmentation strategy provides very good estimation performance when more than 4 or 5 process batches are available and allows to represent the behavior of the process batches more accurately than the FPDM, which makes the HDM unhelpful when the number of calibration batches is extremely small.

3.2. Process Understanding for mAbs Titer Estimation

In this section, we analyze the most important CPPs (i.e., culture variables) for the estimation of mAbs titer at harvest when the data from the process are used alone and when they are combined with the batches generated through the digital models.

3.2.1. Process Understanding with Process Batches Only

In this section, we compare the identification of the most important culture variables for the estimation of mAbs titer at harvest in two scenarios: Scenario 1, rich in available data from the process (i.e., a high number of available batches, $N_P = 80$ batches), and Scenario 2 with only limited data (i.e., number of available batches $N_P = 8$).

For this purpose, two MPLS models are built on 2 LVs to estimate the mAbs titer at harvest, one for each scenario. The models are built 100 times using batches randomly extracted from X_{PC} and y_{PC} . At each iteration, the VIP index is calculated for the model variables, and the importance of each culture variable is assessed by selection absolute frequency, namely the number of iterations in which a variable has $VIP > 1$.

The importance of the culture variables at each time point is shown in Figure 7 through a heatmap of the selection absolute frequency: the green color represents variables that are important with high frequency (>75 – 80) for the estimation of the mAbs titer at harvest, while the red one represents variables which are important only in few iterations (<20 – 25).

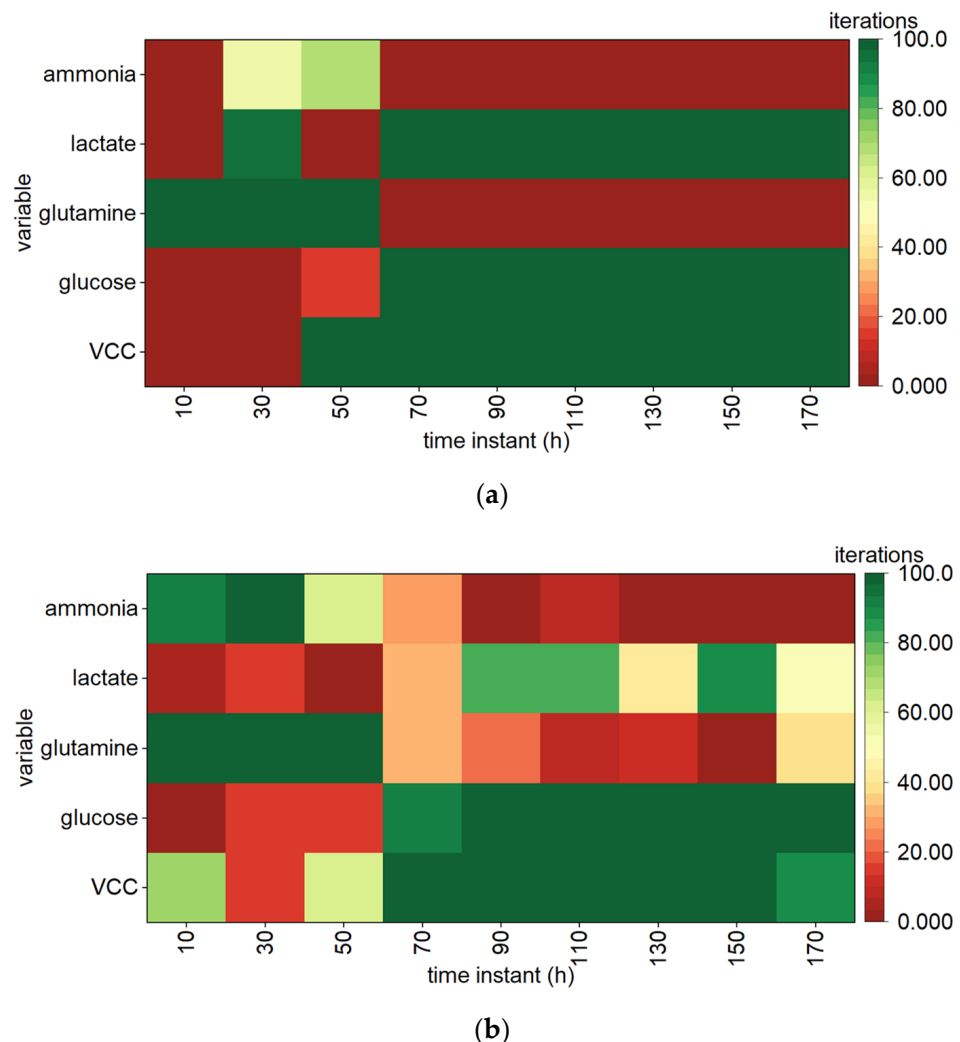


Figure 7. Process understanding for mAbs titer estimation through MPLS variable importance at each time point selection frequency: (a) MPLS model calibrated with 80 process batches, and (b) MPLS model calibrated with 8 process batches, randomly extracted from X_{PC} and y_{PC} .

In Scenario 1, when MPLS is calibrated with $N = N_P = 80$ batches (Figure 7a), glucose, VCC, and lactate show high importance for mAbs titer estimation in the second half of the batch (70 to 170 h), while glutamine shows high importance in the first half (10 to 50 h). Other variables at other time points have a very low selection frequency, except for ammonia

on the second and third day of culture. As expected, the most important factors for the estimation are the concentration of viable cells (VCC) at later culture stages, which represents the number of cells that can produce mAbs, and the available glucose, which represents the available nutrient for growth and mAbs production. Similarly, glutamine, which is the limiting nutrient in the initial part of the batch and remains constant after the initial few days, is identified as particularly important within the first 50 h of the experimental batches. Lactate, instead, significantly limits cell growth only above a certain concentration, confirming its importance only in the second half of the batch. Moreover, a high concentration of ammonia in the initial part of the batch increases cell death causing a reduction in the number of producing cells, hence limiting mAbs production. Accordingly, ammonia shows moderate importance only in the first few days of culture.

The variable importance obtained in Scenario 2, the model calibrated with 8 batches (Figure 7b), indicates that glucose and VCC are important in the second half of the batch, while glutamine is important in the first half. However, this model fails in the identification of the importance of lactate and ammonia. In fact, their importance is not always significant as in the previous case. Furthermore, the model identifies as mildly important variables that were completely unimportant in the previous case (see ammonia, glutamine, and VCC).

According to these results, the limited availability of batches does not allow completely reliable identification of the CCPs that are most related to the CQAs. This spoils the process understanding that can be achieved through the multivariate latent variable model. For this reason, the generation of *in silico* batches could be a valuable strategy to improve process understanding and performance.

3.2.2. Process Understanding Supported by FPDM *In Silico* Data Augmentation

Here, we study the impact that the number of available batches has on the identification of the most important process factors for the estimation of the mAbs titer at harvest when *in silico* batch generation is performed by means of the FPDM.

The procedure utilized here is similar to the one used in Section 3.2.1. We build a 2 LVs MPLS with $N_P = 8$ process batches plus $N_{FP} = 80$ FPDM generated *in silico* to estimate the mAbs titer at harvest. The model building is repeated 100 times randomly selecting the process calibration batches from a subset of 10 batches contained in X_{PC} and the FPDM calibration batches from X_{FP} . The importance of each culture variable is assessed similarly to Section 3.2.1. In this case, it is worth noticing that, since the ammonia is not modeled by the FPDM, it is not present in this MPLS model.

The importance of the culture variables at each time point is shown in Figure 8 in terms of selection frequency. VCC, glucose, and lactate are important for the estimation of mAbs titer from the second day of culture. This is coherent with important variables identified in Scenario 1, when a large number of process batches are available. Differently from the previous case, the importance of the glutamine in the first half of the batch is not identified. This is due to the simplified nature of the glutamine balance, which has not a relevant impact on the first principles model.

This result shows that the generation of *in silico* batches through the FPDM model provides an improved identification of the important variables, even if a limited number of process batches is available. In fact, the addition of *in silico* batches allows identifying more clearly the variables that are important for the estimation than process Scenario 2 (Section 3.2.1), having the same availability of process batches. However, this improved understanding strongly relies on the effectiveness of the model used for batch generation. In fact, the *in silico* batches do not allow correct identification of the glutamine importance, due to the simplified nature of its equations. Despite that, in absence of additional process information, the FPDM *in silico* batch generation is helpful to improve process understanding, even when a simplified model is available.

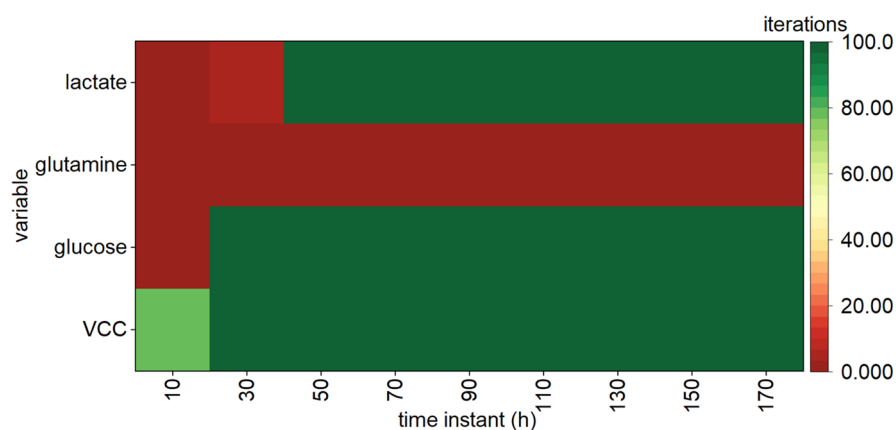


Figure 8. Process understanding for mAbs titer estimation through MPLS variable importance at each time point selection frequency: MPLS model calibrated with 8 process batches, randomly extracted from a subset of 10 batches contained in X_{PC} and y_{PC} , and 80 FPDM in silico generated batches from X_{FPD} and y_{FP} .

3.2.3. Process Understanding Supported by HDM In Silico Data Augmentation

In this section, we study the impact of the HDM in silico batch generation on the identification of the most predictive variables for the mAb titer at harvest. The procedure is analogous to the one presented in the previous section, but here HDM in silico batches are combined with process ones. The 10 HDM batches corresponding to each training process batch are used for the augmentation.

The importance of the culture variables for the titer estimation at each time point is shown in Figure 9. VCC, glucose, and lactate in the second half of the batch (70–170 h) are identified to be the most important variables for mAbs titer estimation. This result is in accordance with the important variables identified in Scenario 1, when a large number of process batches are available. However, lactate shows an average selection frequency (~60), meaning that the identified relationship between lactate and mAbs titer is not as strong as it appears from the process batches. Furthermore, similarly to Scenario 1, glutamine is correctly identified as important in the first half of the batch (10–50 h) and irrelevant in the second half, while ammonia as mildly important only in the first half of the batch. However, glutamine at 10 h has a relatively low selection frequency (~40), indicating that its importance is not correctly identified. Finally, several variables that result to be uninfluential from the process data (Scenario 1) show an average selection frequency (~50).

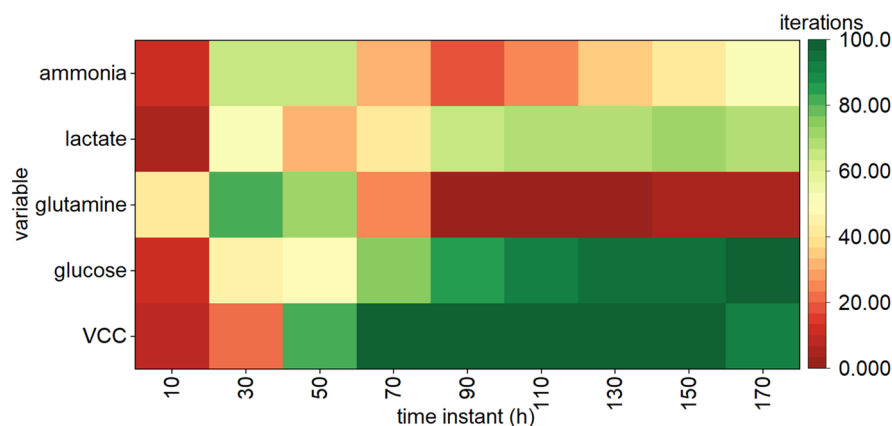


Figure 9. Process understanding for mAbs titer estimation through MPLS variable importance at each time point selection frequency: MPLS model calibrated with 8 process batches, randomly extracted from a subset of 10 batches contained in X_{PC} and y_{PC} , and 80 HDM in silico generated batches extracted from X_H and y_H , 10 for each of the corresponding calibration batches. MPLS variable importance at each time point.

This result shows that HDM in silico batch generation does not identify the important process factors which are completely faithful to the one provided by process batches. In fact, the identification performance is not better than process Scenario 2 when only the reduced number of process batches is used. This is probably due to the high representation accuracy of the HDM, resulting in in silico batches very similar to the training ones. For this reason, the HDM does not increase the amount of information contained in the augmented data, providing less accurate identification of the important factors.

4. Conclusions

In this work, we investigated the utility of in silico data augmentation through digital models to support the development of monoclonal antibodies in scenarios when only a few experiments can be carried out at a given scale. In particular, we investigated two strategies for in silico data generation: a first principles digital model and a hybrid digital model. We applied these strategies to increase the number of available data used in multivariate regression models to estimate the antibody titer at harvest in a simulated process for the production of monoclonal antibodies on a shake-flask scale.

Both in silico data generation strategies were demonstrated to be very effective. In particular, the first principles digital model augmentation strategy allowed a significant improvement in the estimation performance especially when the number of available process batches is extremely limited (1-5), providing a low estimation error of the antibody titer at harvest, comparable with the typical measurement errors (~150–200 mg/L). Furthermore, the first principles digital model improved process understanding. In fact, it allowed to clearly provide process understanding and identify the most important CPPs for the CQA (namely, the mAbs titer at harvest), even when the availability of process batches is limited (<10). The hybrid digital model generation strategy, instead, did not allow an equivalent identification of the important CPPs. Nonetheless, it improved the estimation performance when the number of available process batches is greater than 4. It should be highlighted that the success of in silico data generation relies on the quality of the digital model and its representativeness of the process.

In silico data generation could provide great advantages at different scales of the product and process development, especially at the stirred bioreactor scales, where the number of available batches is typically between 2 and 10.

This study is a proof of concept for the use of in silico data generation in the biopharmaceutical field and further studies will be oriented to adapt the investigated strategies to in vivo applications. Specifically, different ways of combining and preprocessing process and in silico data will be studied. Furthermore, strategies to estimate the parameters for in silico data generation from the experimental batches will be developed.

Author Contributions: Conceptualization, G.B. and P.F.; methodology, A.B., G.B., and P.F.; software, A.B. and G.B.; validation, A.B. and G.B.; formal analysis, A.B.; resources, P.F.; data curation, A.B.; writing—original draft preparation, A.B. and G.B.; writing—review and editing, G.B. and P.F.; visualization, A.B.; supervision, G.B. and P.F.; project administration, P.F.; funding acquisition, P.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data and software used in this study can be replicated from the cited references.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The HEK model [30] used in this work is a first principles mathematical model which simulates batches for the production of mAbs. It is composed of 3 main parts: cell growth and death, cell metabolism, and mAbs synthesis and secretion which are described by 28 equations and 31 parameters in total.

The overall culture material balance is given by:

$$\frac{dV_c}{dt} = F_{in} - F_{out} \quad (A1)$$

The growth and death of the cells part models the life of the cells influenced by nutrients (i.e., glucose and glutamine) and by-products (i.e., lactate and ammonia). It is described by:

$$\frac{d(V_c X_v)}{dt} = \mu V_c X_v - \mu_d V_c X_v - F_{out} X_v \quad (A2)$$

$$\frac{d(V_c X_t)}{dt} = \mu V_c X_v - F_{out} X_t \quad (A3)$$

$$\mu = \mu_{max} f_{lim} f_{inh} \quad (A4)$$

$$f_{lim} = \left(\frac{c_{GLC}}{K_{glc} + c_{GLC}} \right) \left(\frac{c_{GLN}}{K_{gln} + c_{GLN}} \right) \quad (A5)$$

$$f_{inh} = \left(\frac{K_{I_{lac}}}{K_{I_{lac}} + c_{LAC}} \right) \left(\frac{K_{I_{amm}}}{K_{I_{amm}} + c_{AMM}} \right) \quad (A6)$$

$$\mu_d = \frac{\mu_{d,max}}{1 + (K_{d,amm}/c_{AMM})^n} \quad n > 1 \quad (A7)$$

The cell metabolism part models the consumption of nutrients and their conversion into by-products. It is described by:

$$\frac{d(V_c c_{GLC})}{dt} = -Q_{glc} V_c X_v + F_{in} c_{GLC,in} - F_{out} c_{GLC} \quad (A8)$$

$$Q_{glc} = \frac{\mu}{Y_{x,glc}} + m_{glc} \quad (A9)$$

$$\frac{d(V_c c_{GLN})}{dt} = -Q_{gln} V_c X_v - K_{d,gln} V_c c_{GLN} + F_{in} c_{GLN,in} - F_{out} c_{GLN} \quad (A10)$$

$$Q_{gln} = \frac{\mu}{Y_{x,gln}} + m_{gln} \quad (A11)$$

$$m_{gln} = \frac{\alpha_1 c_{GLN}}{\alpha_2 + c_{GLN}} \quad (A12)$$

$$\frac{d(V_c c_{LAC})}{dt} = Q_{lac} V_c X_v - F_{out} c_{LAC} \quad (A13)$$

$$Q_{lac} = Y_{lac,glc} Q_{glc} \quad (A14)$$

$$\frac{d(V_c c_{AMM})}{dt} = Q_{amm} V_c X_v - K_{d,gln} V_c c_{GLN} - F_{out} c_{AMM} \quad (A15)$$

$$Q_{amm} = Y_{amm,gln} Q_{gln} \quad (A16)$$

Finally, the synthesis and secretion of mAbs part is a structured one that models the kinetics of the amino acid chains assembly to create mAbs. It is described by:

$$\frac{dm_H}{dt} = N_H S_H - K m_H \quad (A17)$$

$$\frac{dm_L}{dt} = N_L S_L - K m_L \quad (A18)$$

$$\frac{d[H]}{dt} = T_H m_H - R_H \quad (A19)$$

$$\frac{d[L]}{dt} = T_L m_L - R_L \quad (A20)$$

$$R_H = \frac{2}{3}K_A C_H^2 \quad (A21)$$

$$R_L = 2K_A C_{H_2} C_L + K_A C_{H_2} L \quad (A22)$$

$$\frac{dC_{H_2}}{dt} = \frac{1}{3}K_A C_H^2 - 2K_A C_{H_2} C_L \quad (A23)$$

$$\frac{dC_{H_2} L}{dt} = 2K_A C_{H_2} C_L - K_A C_{H_2} L C_L \quad (A24)$$

$$\frac{dC_{H_2}^{ER} L_2}{dt} = K_A C_{H_2} L C_L - K_{ER} C_{H_2}^{ER} L_2 \quad (A25)$$

$$\frac{dC_{H_2}^G L_2}{dt} = \varepsilon_1 K_{ER} C_{H_2}^{ER} L_2 - K_G C_{H_2}^G L_2 \quad (A26)$$

$$\frac{d(V_c C_{mAb})}{dt} = (\gamma_2 - \gamma_1 \mu) Q_{MAB} V X_v - F_{out} C_{mAb} \quad (A27)$$

$$Q_{MAB} = \varepsilon_2 \lambda_{mAb} K_G C_{H_2}^G L_2 \quad (A28)$$

Table A1 reports the list of the parameters with the corresponding mean and standard deviations used for process batch generation.

Table A1. Mean (reference) and standard deviation values of the parameters used in process batch generation. Missing standard deviations represent that the parameter is kept constant at the reference value.

Parameter	Kontoravdi et al. [30] (Mean)	Standard Deviation
μ_{max} (h^{-1})	0.05800	0.0068
$\mu_{d,max}$ (h^{-1})	0.03000	0.0025
k_{glc} (mM)	0.75000	-
k_{gln} (mM)	0.07500	-
$k_{i,lac}$ (mM)	171.76000	-
$k_{i,amm}$ (mM)	28.48000	-
$K_{d,amm}$ (mM)	1.76000	0.4253
N (-)	2.00000	-
$Y_{x,glc}$ (Cell/mmol)	2.60×10^8	3.10×10^7
m_{glc} (mmol/cell h)	4.9×10^{-14}	-
$Y_{x,gln}$ (Cell/mmol)	8.00×10^8	1.6×10^8
α_1 (mmol L/cell h)	3.4×10^{-13}	-
α_2 (mM)	4.00000	-
$Y_{lac,glc}$ (mmol/mmol)	2.00000	-
$Y_{amm,gln}$ (mmol/mmol)	0.45000	0.0825
$k_{d,gln}$ (h^{-1})	9.6×10^{-3}	0.0030
N_H (gene/cell)	100.00000	-
S_H (mRNA/gene h)	3000.00000	-
K (h^{-1})	0.10000	-
N_L (gene/cell)	100.00000	-
S_L (mRNA/gene h)	4500.00000	-
K_A (cell/molecule L)	1.0×10^{-6}	-
T_H (chain/mRNA h)	17.00000	-
T_L (chain/mRNA h)	11.5.00000	-
K_{ER} (h^{-1})	0.69000	-
ε_1 (-)	0.99500	0.1492
K_G (h^{-1})	0.14000	-
γ_1 (-)	0.10000	-
γ_2 (h)	2.0000	0.333
ε_2 (-)	1.0000	0.15
λ_{mAb} (g/mol)	2.5×10^{-16}	-

Appendix B

The lists of the parameters for in silico batch generation in the first principles digital model and in the hybrid digital model are shown in Tables A2 and A3, respectively.

Table A2. Reference, minimum and maximum values of the parameters used for first principles in silico batches generation. Missing ranges represent that the parameter is kept constant at the reference value.

Variable	Reference	Minimum	Maximum
$\mu_{g,max}$ (h^{-1})	0.073	0.058	0.090
$k_{m,glc}$ (mM)	0.010	-	-
α_x (10^5 Cell/mmole)	44,704.000	-	-
$k_{d,max}$ (h^{-1})	0.020	0.015	0.041
$k_{d,\mu}$ (h^{-1})	0.635	-	-
$Y_{x,glc}$ (10^5 Cell/mmole)	65,341.000	47,700.000	80,700.000
$Y_{lac,glc}$ (mmole/mmole)	1.700	-	-
$Y_{x,lac}$ (10^5 Cell/mmole)	182,050.000	-	-
$k_{m,lac}$ (mM)	3.908	-	-
$Y_{mab,glc}$ (mg/mmole)	150.000	100.000	180.000
k_{gln} (mM)	0.020	0.020	0.050
$Y_{x,gln}$ (10^5 Cell/mmole)	8000.000	7000.000	11,000.000
K_{glc} (-)	0.200	-	-
$\mu_{g,max}$ (h^{-1})	0.073	0.058	0.090
$k_{m,glc}$ (mM)	0.010	-	-
α_x (10^5 Cell/mmole)	44,704.000	-	-

Table A3. Mean (training) and standard deviation values of the parameters used for hybrid in silico batches generation.

Variable	Training (Mean)	Standard Deviation
$\mu_{max,VC}$	2.00	0.13
$\mu_{max,glucose}$	8.00	0.27
$\mu_{max,glutamine}$	3.00	0.10
$\mu_{max,lactate}$	8.00	0.53
$\mu_{max,ammonia}$	2.00	0.13
$\mu_{max,mAb}$	2.00	0.13

References

1. Tripathi, N.K.; Shrivastava, A. Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development. *Front. Bioeng. Biotechnol.* **2019**, *7*, 420. [CrossRef] [PubMed]
2. Walsh, G. Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* **2018**, *36*, 1136–1145. [CrossRef] [PubMed]
3. Yang, O.; Qadan, M.; Ierapetritou, M. Economic Analysis of Batch and Continuous Biopharmaceutical Antibody Production: A Review. *J. Pharm. Innov.* **2020**, *15*, 182–200. [CrossRef] [PubMed]
4. Li, F.; Vijayasankaran, N.; Shen, A.; Kiss, R.; Amanullah, A. Cell culture processes for monoclonal antibody production. *MAbs* **2010**, *2*, 466–479. [CrossRef]
5. Farid, S.S.; Baron, M.; Stamatis, C.; Nie, W.; Coffman, J. Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *MAbs* **2020**, *12*, 1754999. [CrossRef]
6. Epifa, The Pharmaceutical Industry in Figures—Key Data 2021. Available online: <https://www.efpia.eu/publications/downloads/efpia/the-pharmaceutical-industry-in-figures-2021/> (accessed on 28 July 2022).
7. Rameez, S.; Mostafa, S.S.; Miller, C.; Shukla, A.A. High-throughput miniaturized bioreactors for cell culture process development: Reproducibility, scalability, and control. *Biotechnol. Prog.* **2014**, *30*, 718–727. [CrossRef]
8. Clarke, C.; Doolan, P.; Barron, N.; Meleady, P.; O’Sullivan, F.; Gammell, P.; Melville, M.; Leonard, M.; Clynes, M. Predicting cell-specific productivity from CHO gene expression. *J. Biotechnol.* **2011**, *151*, 159–165. [CrossRef]
9. Barberi, G.; Benedetti, A.; Diaz-Fernandez, P.; Sévin, D.C.; Vappiani, J.; Finka, G.; Bezzo, F.; Barolo, M.; Facco, P. Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metab. Eng.* **2022**, *72*, 353–364. [CrossRef]

10. Facco, P.; Zomer, S.; Rowland-Jones, R.C.; Marsh, D.; Diaz-Fernandez, P.; Finka, G.; Bezzo, F.; Barolo, M. Using data analytics to accelerate biopharmaceutical process scale-up. *Biochem. Eng. J.* **2020**, *164*, 107791. [[CrossRef](#)]
11. Ahuja, S.; Jain, S.; Ram, K. Application of multivariate analysis and mass transfer principles for refinement of a 3-L bioreactor scale-down model-when shake flasks mimic 15,000-L bioreactors better. *Biotechnol. Prog.* **2015**, *31*, 1370–1380. [[CrossRef](#)]
12. Goldrick, S.; Holmes, W.; Bond, N.J.; Lewis, G.; Kuiper, M.; Turner, R.; Farid, S.S. Advanced multivariate data analysis to determine the root cause of trisulfide bond formation in a novel antibody-peptide fusion. *Biotechnol. Bioeng.* **2017**, *114*, 2222–2234. [[CrossRef](#)] [[PubMed](#)]
13. Sokolov, M.; Ritscher, J.; MacKinnon, N.; Bielser, J.-M.; Brühlmann, D.; Rothenhäusler, D.; Thanei, G.; Soos, M.; Stettler, M.; Souquet, J.; et al. Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnol. Prog.* **2017**, *33*, 181–191. [[CrossRef](#)] [[PubMed](#)]
14. Kotidis, P.; Kontoravdi, C. Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab. Eng. Commun.* **2020**, *10*, e00131. [[CrossRef](#)] [[PubMed](#)]
15. Kjeldahl, K.; Bro, R. Some common misunderstandings in chemometrics. *J. Chemom.* **2010**, *24*, 558–564. [[CrossRef](#)]
16. Tulsyan, A.; Garvin, C.; Undey, C. Industrial batch process monitoring with limited data. *J. Process Control.* **2019**, *77*, 114–133. [[CrossRef](#)]
17. Mercier, S.M.; Diepenbroek, B.; Wijffels, R.H.; Streefland, M. Multivariate PAT solutions for biopharmaceutical cultivation: Current progress and limitations. *Trends Biotechnol.* **2014**, *32*, 329–336. [[CrossRef](#)]
18. Maharana, K.; Mondal, S.; Nemade, B. A Review: Data Pre-Processing and Data Augmentation Techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [[CrossRef](#)]
19. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
20. Rato, T.J.; Delgado, P.; Martins, C.; Reis, M.S. First Principles Statistical Process Monitoring of High-Dimensional Industrial Microelectronics Assembly Processes. *Processes* **2020**, *8*, 1520. [[CrossRef](#)]
21. Chen, Z.S.; Zhu, B.; He, Y.L.; Yu, L.A. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets. *Eng. Appl. Artif. Intell.* **2017**, *59*, 236–243. [[CrossRef](#)]
22. Lee, S.S. Noisy replication in skewed binary classification. *Comput. Stat. Data Anal.* **2000**, *34*, 165–191. [[CrossRef](#)]
23. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.; Le, Q.V. Unsupervised Data Augmentation for Consistency Training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6256–6268.
24. Chawla, N.V.; Bowyer, K.W.; Hall, L.O. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
25. O'Brien, C.M.; Zhang, Q.; Daoutidis, P.; Hu, W.S. A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. *Metab. Eng.* **2021**, *66*, 31–40. [[CrossRef](#)] [[PubMed](#)]
26. Tulsyan, A.; Garvin, C.; Undey, C. Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. *Biotechnol. Bioeng.* **2018**, *115*, 1915–1924. [[CrossRef](#)] [[PubMed](#)]
27. Marouf, M.; Machart, P.; Bansal, V.; Kilian, C.; Magruder, D.S.; Krebs, C.F.; Bonn, S. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **2020**, *11*, 166. [[CrossRef](#)]
28. Jimenez del Val, I.; Fan, Y.; Weilguny, D. Dynamics of immature mAb glycoform secretion during CHO cell culture: An integrated modelling framework. *Biotechnol. J.* **2016**, *11*, 610–623. [[CrossRef](#)]
29. 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*, 2540–2549. [[CrossRef](#)]
30. Kontoravdi, C.; Pistikopoulos, E.N.; Mantalaris, A. Systematic development of predictive mathematical models for animal cell cultures. *Comput. Chem. Eng.* **2010**, *34*, 1192–1198. [[CrossRef](#)]
31. Oliveira, R. Combining first principles modelling and artificial neural networks: A general framework. *Comput. Chem. Eng.* **2004**, *28*, 755–766. [[CrossRef](#)]
32. Teixeira, A.; Cunha, A.E.; Clemente, J.J.; Moreira, J.L.; Cruz, H.J.; Alves, P.M.; Carrondo, M.J.T.; Oliveira, R. Modelling and optimization of a recombinant BHK-21 cultivation process using hybrid grey-box systems. *J. Biotechnol.* **2005**, *118*, 290–303. [[CrossRef](#)]
33. Von Stosch, M.; Oliveira, R.; Peres, J.; Feyo de Azevedo, S. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Comput. Chem. Eng.* **2014**, *60*, 86–101. [[CrossRef](#)]
34. Yang, S.; Navarathna, P.; Ghosh, S.; Bequette, B.W. Hybrid Modeling in the Era of Smart Manufacturing. *Comput. Chem. Eng.* **2020**, *140*, 106874. [[CrossRef](#)]
35. Sansana, J.; Joswiak, M.N.; Castillo, I.; Wang, Z.; Rendall, R.; Chiang, L.H.; Reis, M.S. Recent trends on hybrid modeling for Industry 4.0. *Comput. Chem. Eng.* **2021**, *151*, 107365. [[CrossRef](#)]
36. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [[CrossRef](#)]
37. Teixeira, A.P.; Alves, C.; Alves, P.M.; Carrondo, M.J.T.; Oliveira, R. Hybrid elementary flux analysis/nonparametric modeling: Application for bioprocess control. *BMC Bioinform.* **2007**, *8*, 30. [[CrossRef](#)] [[PubMed](#)]
38. Yang, A.; Martin, E.; Morris, J. Identification of semi-parametric hybrid process models. *Comput. Chem. Eng.* **2011**, *35*, 63–70. [[CrossRef](#)]
39. Kingma, D.P.; Ba, J.L. ADAM: A method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980.

40. Narayanan, H.; Luna, M.; Sokolov, M.; Arosio, P.; Butté, A.; Morbidelli, M. Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Capture Chromatographic Step. *Ind. Eng. Chem. Res.* **2021**, *60*, 10466–10478. [[CrossRef](#)]
41. Nomikos, P.; MacGregor, J.F. Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97–108. [[CrossRef](#)]
42. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [[CrossRef](#)]
43. Valle, S.; Li, W.; Qin, S.J. Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res.* **1999**, *38*, 4389–4401. [[CrossRef](#)]
44. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69. [[CrossRef](#)]
45. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. *Multi-and Megavariate Data Analysis*; Umetrics Ab: Umea, Sweden, 2006.