

Analysis of Transformed Upstream Bioprocess Data Provides Insights into Biological System Variation

Anne Richelle, Boung Wook Lee, Rui M. C. Portela, Jonathan Raley, and Moritz von Stosch*

In recent years, multivariate data analysis (MVDA) and modeling approaches have found increasing applications for upstream bioprocess studies (e.g., monitoring, development, optimization, scale-up, etc.). Many of these studies look at variations in the concentrations of metabolites and cell-based measurements. However, these measures are subject to system inherent variations (e.g., changes in metabolic activity) but also intentional operational changes. It is proposed to perform MVDA and modeling on data representative of the underlying biological system operation, that is, the specific rates, which are per se independent of the scale, operational strategy (e.g., batch, fed-batch), and biomass content. Two industrial case studies are highlighted to showcase the approach: one is HEK medium performance comparison study and the other is CHO scale-up/-down study. It is shown that analyzing processes in this way reveals insights into behavior of the underlying biological system, which cannot to the same degree be deduced from the analysis of concentrations.

MVDA and modeling approaches are often used in the context of upstream process development,^[9,11,17] likely as it is the key step for determining the maximum achievable product quality and quantity but also as a large amount of potentially useful data is generated at this stage. For instance, sensors available at upstream level produce data online (e.g., for temperature, pH, aeration, stirrer speed) that are traditionally employed to enable standard process control. However, besides the sensor data, an increasing number of applications uses concentration data in the process analysis.^[9,11,17,18] For example, MVDA was exploited for process monitoring (golden batch),^[18,19] development,^[9,11,20] scale-up,^[4] or reproducibility, and robustness studies together with more standard frequentist statistics.^[17]


1. Introduction

Multivariate data analysis (MVDA) has become a standard tool for process analysis in many industries.^[1–3] In the last ten years, the number of studies using MVDA for biopharmaceutical manufacturing and process development has significantly increased, as reflected in the number of research papers published.^[3–11] This increase is partially due to the quality by design paradigm which has found adoption in biopharmaceutical sector,^[8,12,13] but also due to increasing economic pressures stemming from biosimilars, increased R&D spending per successful drug, as well as shorter periods in which the commercialized drug is under patent protection.^[14,15,16]

While the application of MVDA for this type of analysis might allow to detect changes/variations in process performance, the relevance of the analysis is significantly influenced by intentional changes in process operation or small variations during inoculation of the reactors. For instance, Gnoth et al.^[21] described that small changes in the initial biomass concentration (after inoculation) can result in significant variations in the process performance, that is, the final titer concentrations, when the process is run under optimal production conditions. Thus, they conclude that cultivation processes that are run at maximum productivity cannot be robust in the classical sense as significant variations in final titer concentrations (or more generally concentrations) might occur. However, the underlying biological production system might not have changed its production mode, and in fact the observed changes in concentration could somewhat be “artifacts” related to process modifications, not stemming from biological variation that could impact product quality. While metabolic modeling (such as metabolic flux analysis, flux balance analysis, etc.) could elucidate changes in the underlying biological system, it typically requires a larger number of concentrations to be measured than those routinely assessed in industrial settings as well as the development of metabolic models, which might be cumbersome. In light of resource and time constraints, MVDA of concentrations seems an appealing practical tool that helps to highlight differences in the data. However, it is not clear whether those differences stem from the underlying biological system or process operation.

Dr. A. Richelle, Dr. R. M. C. Portela, Dr. M. von Stosch^[+]
Process Systems Biology and Engineering Center of Excellence
Technical Research and Development, GSK
Rixensart 1330, Belgium
E-mail: deq07002@fe.up.pt

Dr. B. W. Lee, J. Raley
Microbial and Cell Culture Development
Biopharm Product Development & Supply, GSK
King of Prussia, PA 19406, USA

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/biot.202000113>

^[+]Present address: DataHow AG, Zurich 8093, Switzerland

DOI: 10.1002/biot.202000113

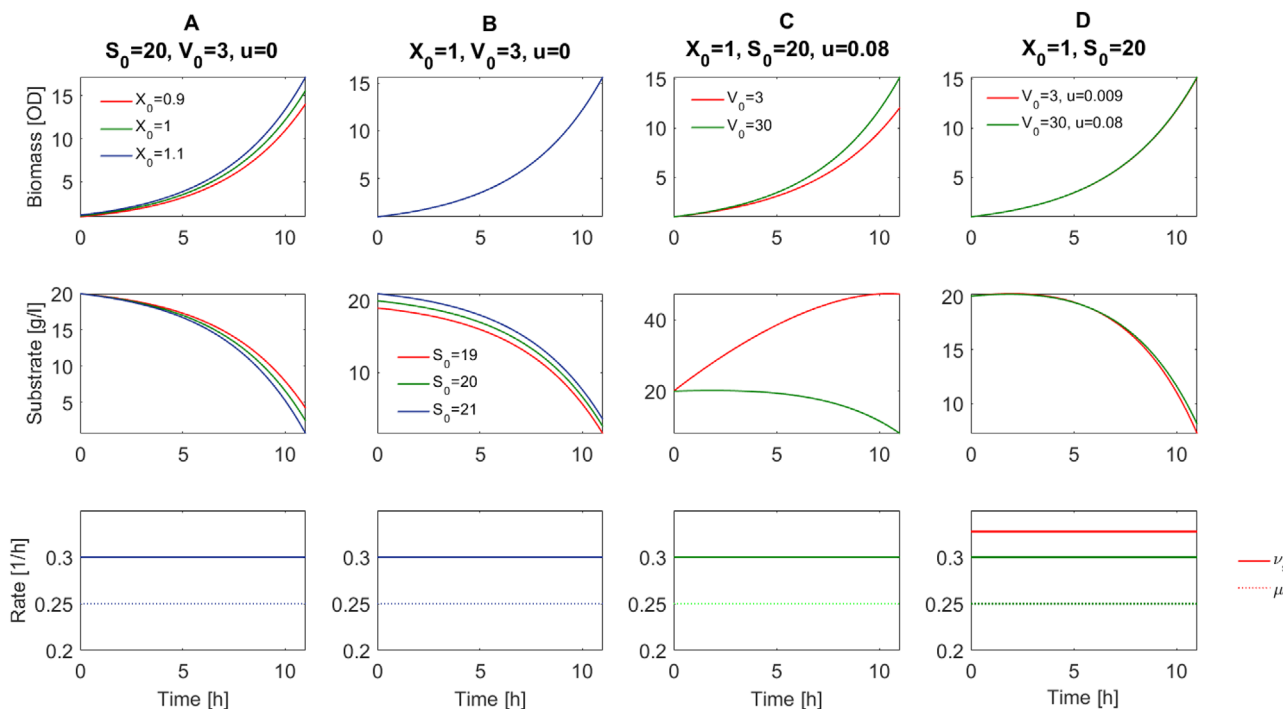


Figure 1. Simulated biomass and substrate concentrations over time and associated specific rates for four constructed cases representing typical upstream process development studies A) process comparability/process monitoring, B) medium development/optimization, C,D) scale-up and feeding optimization in which either the specific rates are identical (A–C) or there are different though the concentrations appear to be very similar (D).

1.1. Changes in Concentration could be “Artifacts” Related to Process Modifications: An Illustrative Example

There exist at least four scenarios, where comparative analysis of concentrations profiles does not provide a representative insight into the observed differences in the behavior of the system, namely process monitoring, reproducibility analysis, medium development/optimization and scale-up (Figure 1). This becomes directly eminent, considering the following simple material balance-based example that describes the evolution of volume (V) as well as biomass (X) and substrate (S) concentrations over time for a fed-batch operated ideally-mixed bioreactor,

$$\frac{dX}{dt} = \mu \cdot X - u/V \cdot X \quad (1)$$

$$\frac{dS}{dt} = -\nu_s \cdot X - u/V \cdot (S - S_f) \quad (2)$$

$$\frac{dV}{dt} = u, \quad (3)$$

where u is the volumetric feeding rate, $S_f = 200$ is the substrate concentration in the feeding, μ is the specific biomass growth rate, and ν_s is the specific substrate uptake rate. Consider there is no variation in the underlying biological system (metabolism), that is, constant specific rates ($\mu = 0.25$ and $\nu_s = 0.25 \times 1.2$).

Three cases can be constructed, which are highlighted in Figure 1A–D, representing typical scenarios encountered during:

1.1.1. Process Monitoring/Reproducibility Analysis

During a typical exponential growth phase, small variations in the initial biomass concentration will be amplified which can lead to significant differences in biomass at the end of the phase resulting into different substrate profiles (Figure 1A). Does this mean that the process is not robust or falls out of specifications? While one can aim at reducing fluctuations in the initial biomass concentration, we argue that if there is no impact (caused by different biomass concentrations) on the process operations downstream to the cultivation (in particular the chromatographic purification steps as more biomass might increase the amount of impurities that need to be removed), all profiles are in fact comparable from a biological point of view as the specific rates exhibit no difference/variations.

1.1.2. Medium Development

The most simplified version of comparing different media is simulated by changing the initial substrate concentration for different experiments (Figure 1B). At the same time, we could assume that differences in the initial biomass concentration occur, but for the sake of simplicity those fluctuations are not included. Comparing the evolution of the substrate concentration over time for different experiments carried out with different media gives little insight into the underlying process performance, as the variations are only due to the variation in the media concentrations. In reality, several concentration profiles (i.e., feeding rate and concentration in the feeding medium) would have to be assessed.

However, the analysis of variations in the specific rates could already provide insights as to whether the metabolism has changed or whether some parts of the metabolism was boosted by the changes in the medium.

1.1.3. Scale-Up/-Down and Feeding Rate Optimization Studies

Scale-up/-down is one of the most significant challenges in bioprocessing. Typically, it is accomplished by keeping limiting factor(s) such as the mass transfer rate or dimensionless numbers (e.g., power per unit volume) identical across scales.^[22] Process performance indicators commonly used for scale-up/-down are concentration based (e.g., viable cell concentration (VCC) and titer) as shown in Equation (1). In the scale-up scenario presented in Figure 1C, the volume changes across scales but the adaptation of the feeding rate was considered unnecessary since substrate concentration was not considered as a limitation factor. However, the impacts on the evolution of the concentration profiles can be observed. This is particularly notable at high concentrations (i.e., the term $-u/V \cdot S$ will increase), where it results into very different trajectories. The observed variation can be attributed to the changes in dilution, but the underlying biological system behaves the same. Contrarily, the scale-up scenario presented in Figure 1D exhibits very small variations in the concentrations during the process, yet the underlying biological system behaves differently at the different scales.

1.1.4. General Conclusions

While the presented scenarios are somewhat extreme cases, they exemplify the importance of differentiating the effect of “intentional process variations” from the “true” inherent variations stemming from changes in the cell metabolism, as well as cell growth and production. This consideration considerably influences the interpretation of the experimental results and, therefore, might lead, for example, to the rejection of experiments as part of a reproducible lot (Figure 1A), the unnecessary generation of experiments (Figure 1B), the rejection of scale-down model or the validation of a feeding profile modification (Figure 1D). In what follows, we advocate for the application of MVDA and model development in the specific rate space (i.e., specific uptake and production rates) rather than in the concentration space. To this end, we used two industrial examples, a HEK medium comparison study and a CHO scale-up study, to showcase the impact of changing the focus to the rate space. More generally, this approach shows how data transformation—which could be interpreted as a form of “feature engineering”—can help to understand and model the system better.

2. Results

2.1. Medium Comparison Study

The aim of this study is to identify the metabolites that influence the viable cell density of HEK293 cell cultures. This analysis

Table 1. Summary of the Ambr conditions.

Group of conditions	Replication	Ambr ID	Family of medium and preculture pool
G1	triplicate	BR05, BR12, BR24	B
G2	duplicate	BR06, BR13	B
G3	duplicate	BR07, BR14	B
G4	duplicate	BR08, BR15	B
G5	duplicate	BR09, BR16	B
G6	duplicate	BR10, BR17	B
G7	duplicate	BR11, BR18	B
G8	single	BR19	C
G9	duplicate	BR20, BR21	C
G10	duplicate	BR22, BR23	C
G11	single	BR01	A
G12	duplicate	BR02, BR03	B
G13	single	BR04	B

focuses on investigating the performance of medium family B and C, in reference to medium A. In brief, we want to understand the origin of observed differences in the processes between the reference medium A and media B and C.

2.1.1. Data Generation

Batch cultivations of HEK293 in a 24 Ambr250 device were performed for 8 days. There are 13 groups of conditions associated with 3 medium families and preculture pools that were investigated, **Table 1**.

For each batch, samples were collected on days 0, 1, 2, 3, 5, and 8. The samples were analyzed for the 20 amino acids (UPLC), glucose, ammonium, glycerol, lactate, urea, calcium, and magnesium (Cedex-Bio). The quantitative performance indicator is the viable cell density.

2.1.2. Data Transformation

The measurement of metabolite concentrations is associated with measurement errors. As the estimation of the specific rates from concentration measurements is an ill-conditioned inverse problem, the measurement errors might be amplified resulting in large uncertainty bounds of the specific rates. To quantify to which degree this is the case, we used the following Monte-Carlo based sampling approach.

- 1) Draw a random value from a normal or uniform distribution^[23] with the measurement's standard deviation for each time-point at which a concentration was measured. Add these values to the original measured concentration values.
- 2) Perform the same for biomass concentrations, as these will be used to compute the specific rates.
- 3) Follow the approach described in the Experimental Section for the rate estimation from the modified concentration values

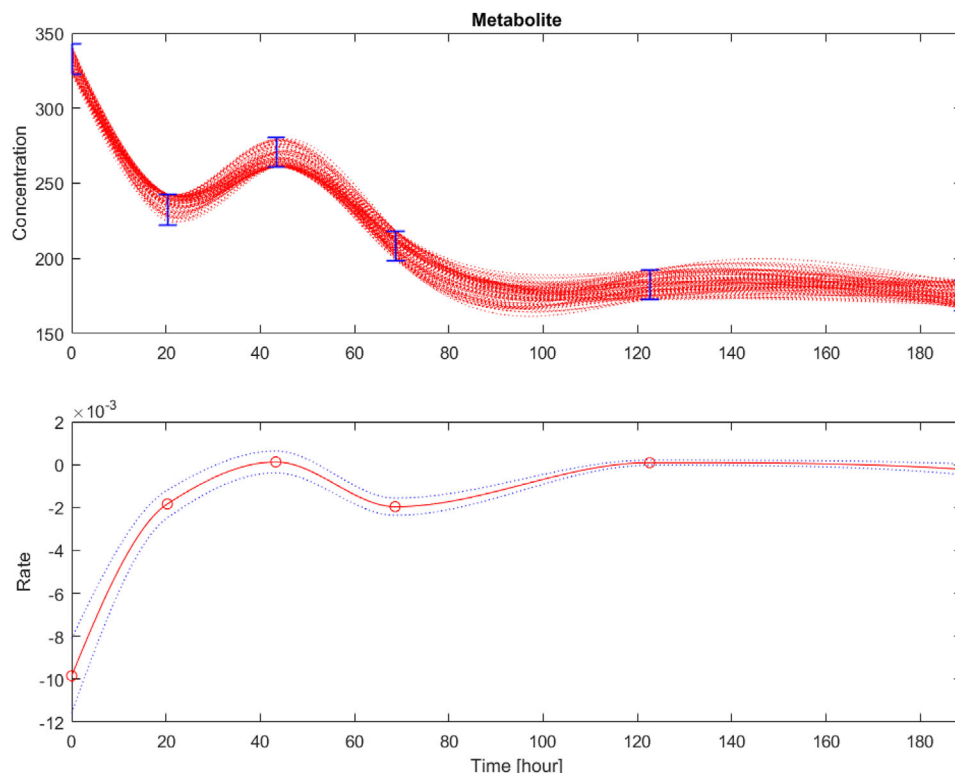


Figure 2. Top—concentration over time—Representative example of 100 spline approximations (red continuous lines) and the measured points with confidence intervals (blue lines); Bottom—specific rate over time—specific rate estimations obtained from the concentration splines.

using a piecewise cubic interpolation (function `csaps`, Matlab 2016a) to approximate the measured quantities (i.e., $f(t, w)$ is a smooth piecewise cubic spline).

- 4) Assemble the rate values with previous ones and repeat steps (1) to (4) until the standard deviation estimates of the rates computed on the growing rate data set converge.

Corresponding to typical measurement errors, for each concentration data, we assumed a coefficient of variation of 10%, except for the biomass measurement for which an error of only 5% is assumed. Note that we assumed that the minimal experimental error is 0.1 mg L^{-1} and that the maximal one is 10 mg L^{-1} . For each measurement, we generated 100 random measurements uniformly distributed across the experimental confidence interval associated to this point. We therefore obtained 100 different evolutions of concentration over time for each measured metabolite that consider the experimental measurement error. For each of these randomly generated data, we constructed a spline. The spline of the concentration data has been defined using the function `csaps` of MATLAB and by identifying an adequate smoothing parameter p . The default smoothing parameter p was defined with the following formula: $p = 1/(1+(h^3/60))$ where h is the average sampling time interval. Using this default smoothing parameter, we checked that the spline was comprised in the confidence interval of the measurement. If not, the smoothing parameter was increased by increment of 0.0001 till the spline went through all the confidence interval. The rates were computed using the methodology presented above for each of the 100 random splines. The figure below presents the mean of all the computed

rates and the associated confidence intervals (\pm standard deviation error).

Figure 2 is a representative example of the spline approximations and rates showing that the concentrations are uniformly distributed over the measured confidence interval, as expected. The estimated specific rates show slightly larger confidence intervals in regions where the concentration profile is changing, whereas they are tighter elsewhere as the methodology forces the variation of the curvatures of the set of splines constructed during the MC sampling to be minimal while taking into account experimental measurement errors. Interestingly, the specific rate data tends to be normal distributed (as can be seen in the violin plots—**Figure 3**) though the concentration samples were drawn from a uniform distribution. Whether this observation is case specific or general is not clear.

2.1.3. Insights from Principal Component Analysis for Concentration versus Rate Data

Two Principal Component Analysis (PCA) were performed on **batch-wise unfolded concentration** and **rate (flux) data**, that is, $((n_c \cdot n_e) \times n_t)$ and $((n_r \cdot n_e) \times n_t)$, respectively. The data was **auto-scaled** along the variable dimension and **three latent components** were enough to describe the variation in the data, $66.15\% \pm 1\%$ and $52.42\% \pm 1.73\%$, respectively for concentration and rate. **Figure 4** shows the score plots for the **first three latent components obtained** for the concentration and rate data. Comparing

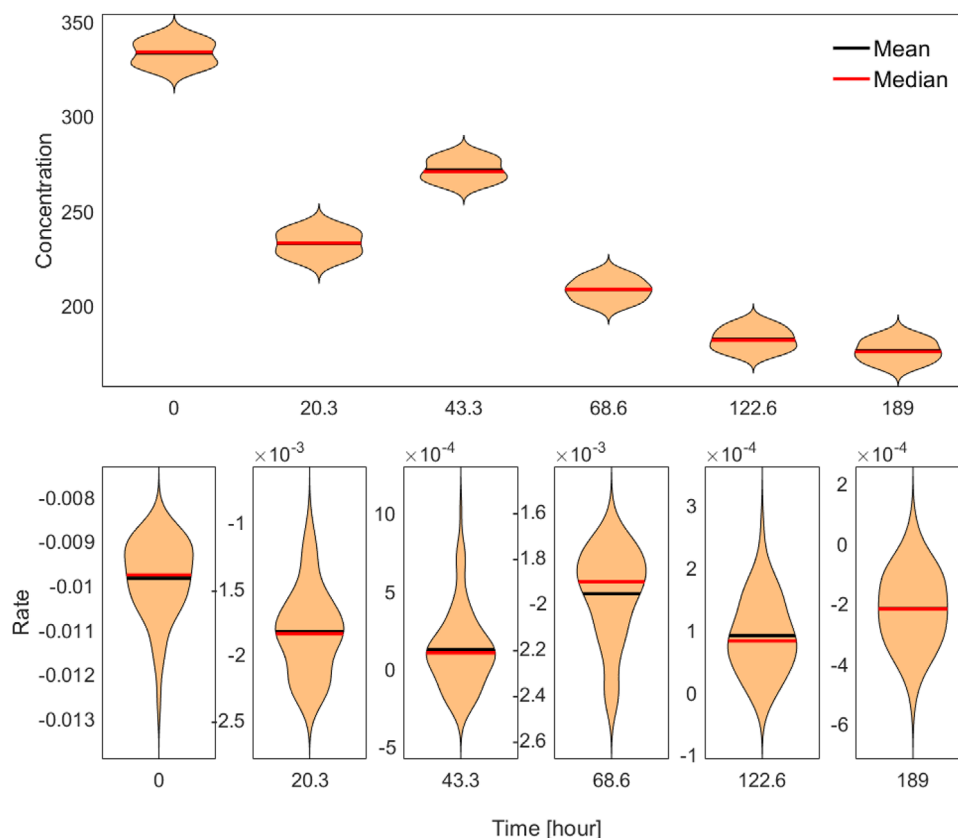


Figure 3. Top—Distribution of concentration samples for each time point. Bottom—Respective distributions of the estimated specific rates.

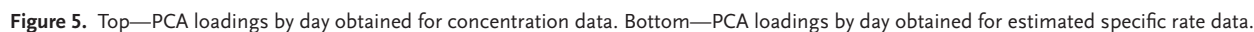
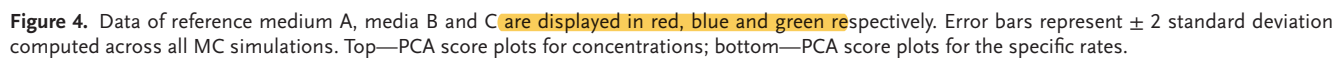
the different media in the concentrations' latent space it seems that the A, B, and C give rise to different cultivations, as visually media B and C are fairly distinct from the reference medium A. In contrast in the rates' latent space the cultivation with the reference medium A, seems fairly close to the cultivations with the other media B and C for certain conditions. The different behaviors observed for the concentrations' and rates' latent spaces seem to indicate that though the media are different in terms of concentrations, the cellular metabolism is similar. Thus, it can be expected that product quality is similar, though not necessarily quantity (which would be related to the concentrations).

The loadings of the concentrations provide a direct insight into the main differences in concentrations between experiments for each day (Figure 5). It can be seen that over the days the positions of the loadings are more spread. This highlights how the intentional changes in concentration of day 1 influence the evolution of the culture giving rise to differences in many more concentrations while they were not originally differing to the same degree (see the decrease in variance explained by the first two latent components). Note that the evolution of loadings for specific compounds along the days should not be analyzed since the position of the loading of each day is influenced by those of other compounds. Though a similar behavior can be observed for the rates' loadings, the loadings indicate a difference in uptake/secretion of compounds and are such indicative of differences in metabolism between experiments. As the cells come from the same pre-culture pool, their metabolism at the begin-

ning is rather similar and just changes in certain rates can be observed. This indicates a first response of the cells to intentionally added compounds. Over the days the changes in rates, indicative of metabolism, become more pronounced, which reflects the adaption of the cells to the exposed environment, medium. Interestingly, lactate and glutamine at days 3 and 5 take a much more prominent role in the rates' loadings than in those of the concentrations. The importance given to the pathways related to lactate and glutamine uptake/secretion, which are influenced by the changes made to the conditions and media, might thus differ in the analysis of the concentration and rate space. To get more insight into the metabolism of each experiment the data could be time wise unfolded or analyzed with methods that take the autocorrelation into account.

2.2. Process Scale-Up/-Down Case

In the pharmaceutical industry, "an important requirement for successfully characterizing a process is the availability of a scale down model that is representative of the manufacturing process,"^[22] as this allows for direct translation of extensive knowledge gained at bench scales to full scale commercial facilities. At bench scale, recent advances in automated and high-throughput mini-bioreactor systems (e.g., Ambr 15 and Ambr 250) offer significant advantages in terms of data amount and



quality (e.g., in Manahan et al.^[24]). However, care must be given on interpretation of raw data from these systems, especially as it relates to scale down model development. In part, this is inherently due to the proportion of sample volume to culture volume at bench scale. As is common practice within process development, a daily sample volume is required to assess growth rate, metabolite concentrations, and total yield. This sample volume is required regardless of culture/reactor volume, and thus contributes significantly to process variability at bench scale as it skews the proportion of feed volume to culture volume.

Use of dimensionless numbers (e.g., Reynolds, Flow, and Froude numbers) is a common engineering practice for scale-up/-down, as mentioned before. For cell culture processes, some of the successful approaches reported in the literature also include keeping scale-independent parameters identical^[22] (e.g., $k_L a$ —mass transfer coefficient times interphase area, P/V —power per unit volume, and OUR—cell specific Oxygen Uptake Rate). The most common way to qualify scale down model is to perform statistical tests such as F-test and TOST (two one-sided t -test), on end-of-run or peak process performance indicators, namely VCC, titer, and viability. Consideration of holistic batch evolution behavior via means of MVDA techniques is gaining more popularity, but the fundamental question remains: are we looking at the data correctly, given the difference in process scale and inherent changes due to reactors, sampling strategy, analytical instruments, labs, operators, and process scientists?

2.2.1. Data Generation

Data were generated at different scales using the same CHO cell line. For all practical purposes, process parameters were identical across all batches. These include process set points (pH, DO), methods to control them (pH control, DO control), media, and instruments and methods used for cell count and titer measurements. Antifoam was added as needed.

2.2.2. Data Transformation

The specific rates were estimated as described in the methodology section. The R mgcv package was used to obtain a cubic spline approximation, minimizing the sum of residual squares while penalizing on curvature, that is

$$\sum_{i=1}^n (y_i - s(t))^2 + \rho \cdot \int s''(t)^2 dt \quad (4)$$

where $s(t) = \sum_{j=1}^q b_j(t) \cdot \alpha_j$ is the cubic regression spline, with b_j the j^{th} basis function, q the basis function dimension and α_j the weight for the j^{th} basis function. Basis dimension was chosen as four based on cell growth phases: lag, exponential growth, stationary, and death phases. Further increasing basis dimension did not improve model performance. The package also predicts standard error based on the Bayesian posterior covariance matrix of the parameters,^[25] which can be used to estimate sample mean and variance. This is particularly interesting in case of

early phase bioprocess development because of the lack of data at larger scales. For smaller bench scale reactors such as 0.25 and 3 L reactors, it is relatively easy to collect “replicate” experimental data and make statistical inferences about the process, cell behaviors, and robustness. However, for larger scale reactors, for example, 200, 1000, and 2000 L, data generation is very expensive. Therefore, “replicate” batch data is sometimes not possible to obtain. Yet, even without them, scientists may be asked to make inferences and decisions on whether a process is scalable or not. To make things even more challenging, data from these runs may (or even typically) have different sampling times, missing data, without proper quantitative knowledge of process, sampling, and/or intrinsic variability.

Figure 6 demonstrates this concept pictorially for a larger scale batch, 1000 L, where replicate data have not yet been collected. The resulting regression via cubic regression spline consists of estimated mean VCC and 95% confidence interval, that is, estimate of VCC profiles if we could run multiple batches, which in turn can be used to give more (or less) insights and confidence on process/cell scalability.

2.2.3. Analysis of Specific Rates and Concentrations across Scales

Top figures from Figure 7 show cell growth and production profiles for practically identical processes at different scales. It seems that at 3 L scale, end-of-the-batch and peak VCC concentrations are slightly lower; whereas for titer, 3 and 200 L scale results lay on top of each other, with 1000 L scale seemingly having slightly higher values. Considering all the variations stemming from analytical measurements, initial VCC, and process scales, it is difficult to conclude whether the differences and/or similarities stem from the biological system.

Shown at the bottom of Figure 7 is the data transformed with cubic spline approximation, and the mean growth and production rates calculated based on estimated mean response profiles. It can be seen that the specific growth rate at the 1000 L scale is greater during the first 4 to 5 days than at smaller scales and VCC is comparable after 4 to 5 days, despite starting at lower initial VCC compared to 3 L scale. The cell specific productivity does not seem to be impacted by the faster growth at 1000 L and is above that observed at 3 L from day 3 on. The slightly increased production rate throughout the run, in combination with slightly greater VCC values around 6.5 elapsed days, results in slightly higher titer values at the time the 1000 L batch is completed (≈ 7.5 elapsed days) compared to other scales. The rate data analysis thus allows to understand that the higher VCC at ≈ 6.5 elapsed days is not the sole reason for the increased final titer, but that higher specific productivity could also be observed. For 200 L scale, the higher VCC values do not result in higher titers as the specific production rate is low. Also, the “catch up” of the VCC at 200 L seems slightly delayed as growth rate is shown to be only slightly higher during days 4 to 7. The added value of data transformation also becomes apparent for the 3 L results as complete opposite conclusions are drawn depending on whether the concentration or rate data are analyzed. Indeed, the specific rate data suggest that most of the variability associated with the overall reactor performance (i.e., end of the batch titer) is due to the

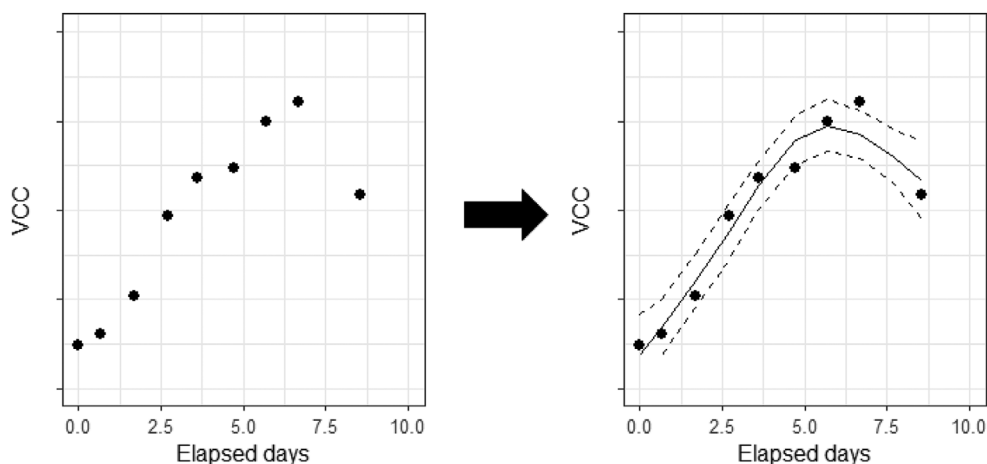


Figure 6. Using cubic regression spline to estimate the mean and variance of data collected for a 1000 L batch with no replicates.

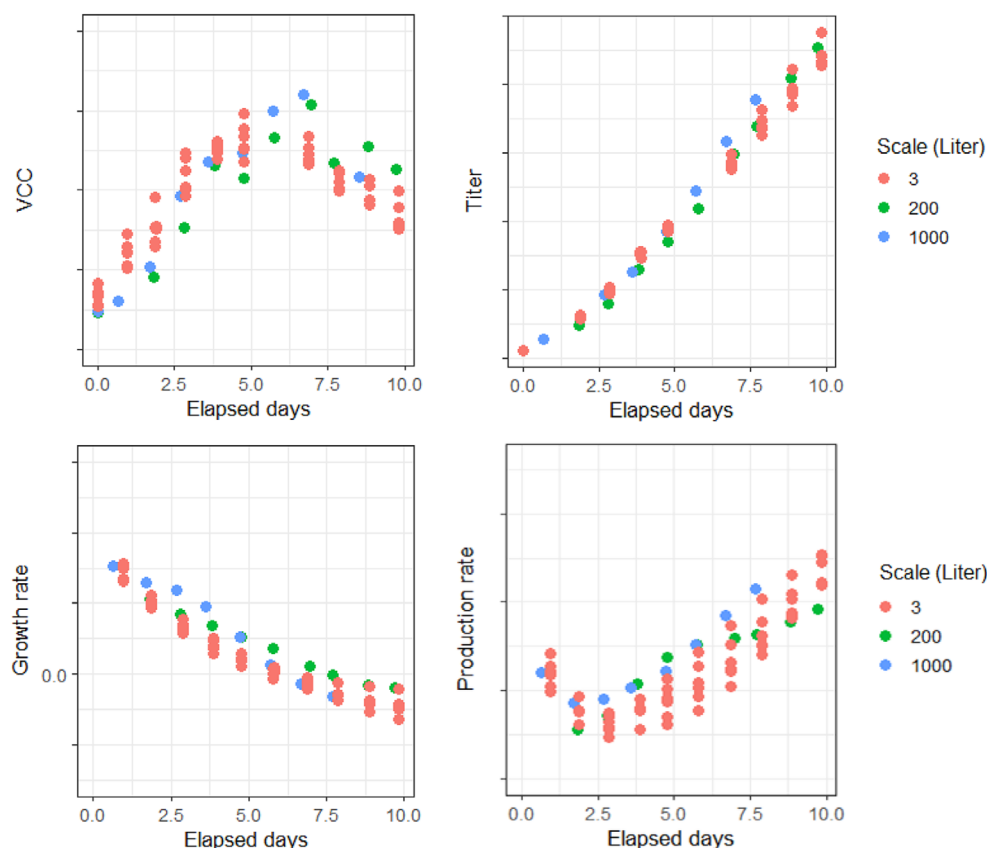


Figure 7. Top row—VCC and titer for vessel volume change; bottom rows—the respective estimated specific cell growth and specific production rates. Red, green, and blue colors represent experiments at 3, 200, and 100 L scale data, respectively.

protein production mechanisms as opposed to cell growth while the concentration data suggest that VCC presents the higher degree of variability.

In the context of scale-up studies, traditionally more experimental data are required at the larger scales (i.e., 200 and 1000 L) to perform ad-hoc statistical test, such as F-test on VCC and

titer measurements. Collecting such data is difficult in short time frames. However, typically concentrations are measured several times during the runs at larger scale. As the entire concentration profile, that is, all concentration measurements of a run, is used for fitting arbitrary time dependent functions the impact of the measurement error is reduced for comparison of the data at

different scales. The transformation of the data could be understood as a “normalization” that computationally separates the variation of the biological system from that of “small” variations in the operation of the process (e.g., variations in the initial biomass concentration). Therefore, comparing estimated cell specific growth and production rates (Figure 7 bottom row) of the different scales seems to be a viable option, indicative of changes in the biological system. Though requiring a more in-depth study, potentially including “-omic” data, to show that indeed conclusions about the reproducibility can be based on the analysis of transformed data of few runs, a significant reduction the number of runs at larger scales as compared to traditional reproducibility studies seems possible. While obviously reducing costs, the potential reduction in the timelines seems even more attractive, for example, when responding to a pandemic, even more though as it would not compromise on the safety or quality.

3. Discussion

(M)VDA of concentration data is a widely used approach in industrial upstream bioprocess studies.^[1–3] While it allows for an analysis of changes/variations in process performance, it does not per se allow to differentiate between “intentional” design/operational variations and inherent biological variations. However, the differentiation of the sources of variation is critical in the biopharmaceutical context as the metabolic states of the host cells can be expected to influence the drug production machinery and therefore might impact the critical quality attributes of the drug substance. In contrast, variations stemming from operational changes, for example, in feedings could be neglected if they do not change the effectiveness of operations downstream to the cultivations.

The transformation of process data and their subsequent analysis with standard (M)VDA methods provides insights into the variations of the biological system. We showed for two upstream bioprocess studies that the biological system was performing much more comparable for the different media and scales, as could be concluded from the concentration data. The analysis of the concentration data alone would have resulted into different conclusions and follow-up actions. For instance, the relative distance of the PCA scores obtained for the concentration data in the media study indicated that cultivations on the tested media were not similar to the reference media, whereas the PCA scores obtained for the specific rate data indicated that the cultivations behaved similarly. This insight turned the focus to finding compounds that change the specific rates, as a proxy for changing the metabolism.

Models in which the specific rates are approximated by parametric (the structure of the function is fixed by knowledge, e.g. mechanistic or empirical models^[26,27]) or nonparametric (the structure of the model is derived from data, e.g. artificial neural networks or polynomial functions^[28–32]) functions could provide additional insights into the factors that drive cell behavior while at the same time allowing to differentiate between inherent biological and intended operational variation. However, the development of such models requires data in which the specific rates vary in function of the process conditions.^[28] Even though data might have been generated with design of experiment approaches, sufficient variation in the specific rate data might not

be present. This becomes eminent considering, for example, the use of experiments for which you vary the substrate amount, S , in the range of 10 to 15 concentration unit to study the influence of substrate availability on the specific biomass growth that can be represented by the following Monod model: $\mu = \mu_{\max} \cdot S/(S + 0.1)$, that is, $\mu_{\max} \cdot 0.9901$ versus $\mu_{\max} \cdot 0.9934$, respectively. Furthermore, data generated in the context of process robustness studies will also likely exhibit only small variations in the specific rates and are not suitable for the development of specific rate models. Data transformation can be applied in any of these cases to overcome these problems but will require frequent measurements of the concentrations during the experiment. However, in light of the current trend toward adopting more and more online in process analytical technology^[34] this requirement seems to be more often met.

The analysis of transformed upstream bioprocess data (i.e., data of cultivations/fermentations) can potentially add value across the drug product lifecycle. During process development it can provide a better insight into what, for example, is metabolically different between clones or process conditions. Indeed, the analysis helps to shed light into the interplay of the biological system with the process system. In development as well as manufacturing conditions, an additional monitoring of the rates can help to pinpoint to the source of process variation as the differentiation between operational and inherent changes is more apparent. Similarly, the analysis of specific rate data in scale-up studies can provide an indication whether the cell behave similarly across scales. Also for process control purposes, the analysis of specific rate data can help the design of the control equations as they represent the underlying dynamic nature of the system.^[34–38] This also brings the development of dynamic design spaces in upstream bioprocess into reach, a topic that at present seems to be mostly found in drug formulation.^[39,40] Hence, we believe that a systematic analysis of specific rate data in upstream process development, monitoring, scale-up, or reproducibility studies is of added value. Even the more, as it can be assumed that variation in the biological system seems much more indicative of product quality issues than process variability.

4. Experimental Section

Methodology for Rate Estimation: It was proposed to analyze the underlying biological system by applying MVDA to the specific rates, instead of performing the upstream process performance analysis in the concentration space. The method was based on the extracellular material balances (for an ideally mixed reactor).

$$\frac{dc_{\text{ex}} \cdot V}{dt} = q \cdot x \cdot V + u \quad (5)$$

With c_{ex} the vector of extracellular concentrations that also comprises the biomass concentration x , V the volume of the broth, q a vector of specific rates (also comprising the specific biomass growth rate μ), and u a vector of compound specific feeding rates (in case of fed-batch or continuous operation). These material balances were used to estimate the specific rates from measurements of concentrations, volumes, and feeding. The estimation of the rates can be accomplished in two different ways, a differential or integral way.^[41–44] Since the interest was in analyzing changes in the

specific rates over time, the differential way was adopted here, following the best practice,^[45,46,28] that is,

- 1) Starting from the integrate version of the material balance (Equation (5)), the rate related terms were isolated on the right-hand side since they cannot be measured:

$$c_{\text{ex}}(t_i) \cdot V(t_i) - c_{\text{ex}}(t_0) \cdot V(t_0) - \int_{t_0}^{t_i} u \cdot dt = \int_{t_0}^{t_i} q \cdot x \cdot V \cdot dt \quad (6)$$

- 2) Fit arbitrary time dependent functions (e.g., cubic smoothing splines, gaussian process models, polynomials or others), $f(t, w)$, to approximate the measured quantities, $\gamma_m(t_i) = c_{\text{ex},m}(t_i) \cdot V_m(t_i) - c_{\text{ex},m}(t_0) \cdot V_m(t_0) - \int_{t_0}^{t_i} u_m \cdot dt$, such that the residual ε was small, though the function also does not overfit the data.

$$\gamma_m = f(t, w) + \varepsilon \quad (7)$$

- 3) Build the derivative of $f(t, w)$ analytically with respect to time, that is, $\frac{df(t, w)}{dt}$
- 4) Evaluate the derivative at the time instance t_i at which the concentrations have been measured (assuming that the concentrations have the lowest measurement frequency) and divide by the approximated biomass ($x_m(t) \cdot V_m(t) = g(t, w) + \varepsilon$):

$$\left. \frac{df(t, w)}{dt} \right|_{t_i} / g(t_i, w) = q_m(t_i) \quad (8)$$

- 5) In this way a data matrix of the estimated specific rates, Q , was obtained for each cultivation with dimensions ($n_c \times n_t$), number of concentrations, and number of time points, respectively.

Data of the estimated specific rates, Q , for several experiments, n_e , were then combined for MVDA, $Q_{1..n_e}$. Depending on the objective of the analysis, the typical unfolding and normalization methods can be adopted.

It should be apparent from the transformation, that the estimated specific rates are per se independent of the process scale, biomass content, and operation strategies (batch, fed-batch, perfusion or continuous). However, changes in any of these factors might impact on the cellular metabolism and thus will reveal itself in the specific rate estimates.

Underlying Assumptions and Limitations of the Approach: The first assumption was that the reactor content was ideally-mixed with no macroscopic concentration gradients (i.e., the material balance was used for a well stirred tank). While some may challenge this assumption for larger scale (in the order of thousand liters), the assumption was reasonable as the processing times for bioreactors of interest range from days to weeks, compared to mixing times in the order of 10 min to 2 h. This assumption leads to all measured concentration variations being representative of the system, process, and cells themselves.

The second assumption was that the rates in Equation (5) can be represented by the multiplication of the biomass concentration, x , with the specific rates, q (i.e., the uptake and secretion reactions were “catalyzed” by the cells). In general, the order of magnitude of the reaction values of the cell catalyzed reactions were significantly greater than that of spontaneous reactions, except for the spontaneous decomposition of glutamine into pyrrolidone-carboxylic acid and ammonia^[47] (which occurs under certain conditions and can be corrected for^[48]).

The estimation of the specific rates in the differential way is a mathematically ill conditioned inverse problem (i.e., uncertainties in the measurement of the concentrations were propagated and potentially amplified). However, the errors in the estimated rates were acceptable, in case that the frequency and number of measurements were high as well as the uncertainty of the measurements were low.^[45] Ideally, the analytical error of the measurements should be significantly lower than the variations in the signal (signal to noise ratio); the number of measurements should be such that the parameters of the approximating function can be sufficiently accurately identified; and the frequency of measurements would be greater

than the one over the characteristic time of the system. This seems particularly important with respect to recent findings that oscillations might occur^[49] that were otherwise not observed.

Acknowledgements

The development of this article was sponsored by GlaxoSmithKline Biologicals SA which was involved in all stages of the study conduct and analysis.

Author Contributions

A.R., B.W.L., and M.v.S., were involved in the conception and design of the study. All authors analyzed and interpreted the results. All authors were involved in drafting the manuscript or critically revising it for important intellectual content. All authors had full access to the data and approved the final draft of the manuscript.

Conflict of Interest

All authors are, or were at the time of the study and the development of this article, employees of the GSK group of companies. M.v.S. was a co-founder of Novasign GmbH at the time of the study. A.R. and R.P. report ownership of shares and/or restricted shares in GSK.

Keywords

cell cultivation, development, monitoring, multivariate data analysis, optimization, scale-down, scale-up

Received: March 15, 2020

Revised: May 30, 2020

Published online:

- [1] J. F. MacGregor, M. J. Bruwer, I. Miletic, M. Cardin, Z. Liu, *IFAC-PapersOnLine* **2015**, 48, 520.
- [2] C. Duchesne, J. J. Liu, J. F. MacGregor, *Chemom. Intell. Lab. Syst.* **2012**, 117, 116.
- [3] A. S. Rathore, S. Mittal, M. Pathak, A. Arora, *Biotechnol. Prog.* **2014**, 30, 967.
- [4] S. M. Mercier, B. Diepenbroek, M. C. F. Dalm, R. H. Wijffels, M. Streefland, *J. Biotechnol.* **2013**, 167, 262.
- [5] C. Ündey, S. Ertunç, T. Mistretta, B. Looze, *J. Process Control* **2010**, 20, 1009.
- [6] A. Tulsyan, C. Garvin, C. Ündey, *Biotechnol. Bioeng.* **2018**, 115, 1915.
- [7] A. Tulsyan, C. Garvin, C. Ündey, *J. Process Control* **2019**, 77, 114.
- [8] M. J. T. Carrondo, P. M. Alves, N. Carinhas, J. Glassey, F. Hesse, O.-W. Merten, M. Micheletti, T. Noll, R. Oliveira, U. Reichl, A. Staby, A. P. Teixeira, H. Weichert, C.-F. Mandenius, *Biotechnol. J.* **2012**, 7, 1522.
- [9] M. Sokolov, J. Ritscher, N. MacKinnon, J. Souquet, H. Broly, M. Morbidelli, A. Butté, *Biotechnol. Prog.* **2017**, 33, 1368.
- [10] D. Brühlmann, M. Sokolov, A. Butté, M. Sauer, J. Hemberger, J. Souquet, H. Broly, M. Jordan, *Biotechnol. Bioeng.* **2017**, 114, 1448.
- [11] M. Sokolov, J. Ritscher, N. MacKinnon, J.-M. Bielser, D. Brühlmann, D. Rothenhäusler, G. Thanei, M. Soos, M. Stettler, J. Souquet, H. Broly, M. Morbidelli, A. Butté, *Biotechnol. Prog.* **2017**, 33, 181.
- [12] M. Jenzsch, C. Bell, S. Buziol, F. Kepert, H. Wegele, C. Hakemeyer, in *New Bioprocessing Strategies: Development and Manufacturing of*

- Recombinant Antibodies and Proteins* (Eds: B. Kiss, U. Gottschalk, M. Pohlscheidt), Springer, Cham **2018**, pp. 211–252
- [13] J. Glassey, K. V. Gernaey, C. Clemens, T. W. Schulz, R. Oliveira, G. Striedner, C.-F. Mandenius, *Biotechnol. J.* **2011**, *6*, 369.
- [14] E. A. Blackstone, P. F. Joseph, *Am. Health Drug Benefits* **2013**, *6*, 469.
- [15] C. H. Song, J.-W. Han, *SpringerPlus* **2016**, *5*, 692.
- [16] I. O. Medicine, *The Changing Economics of Medical Technology*, The National Academies Press, Washington, DC **1991**
- [17] S. Goldrick, V. Sandner, M. Cheeks, R. Turner, S. S. Farid, G. McCreath, J. Glassey, *Biotechnol. J.* **2019**, *15*, 1800684.
- [18] L. Mears, R. Nørregård, S. M. Stocks, M. O. Albaek, G. Sin, K. V. Gernaey, K. Villez, in *Computer Aided Chemical Engineering* (Eds: K. V. Gernaey, J. K. Huusom, R. Gani), Elsevier, Amsterdam **2015**, pp. 1667–1672.
- [19] R. Luttmann, D. G. Bracewell, G. Cornelissen, K. V. Gernaey, J. Glassey, V. C. Hass, C. Kaiser, C. Preusse, G. Striedner, C.-F. Mandenius, *Biotechnol. J.* **2012**, *7*, 1040.
- [20] D. A. Suarez-Zuluaga, D. Borchert, N. N. Driessen, W. A. M. Bakker, Y. E. Thomassen, *Vaccine* **2019**, *37*, 7081.
- [21] S. Gnath, M. Jenzsch, R. Simutis, A. Lübbert, *J. Biotechnol.* **2007**, *132*, 180.
- [22] L. Tescione, J. Lambropoulos, M. R. Paranandi, H. Makagiansar, T. Ryll, *Biotechnol. Bioeng.* **2015**, *112*, 84.
- [23] Whether a sample is drawn from a normal or uniform distribution depends on the availability of experimental or analytical replicates. In case of a predefined error related to the example of the analytical precision, sampling from a uniform distribution seems more appropriate as experimental values within the interval of experimental error have the same probability to be drawn. On the contrary, if replicates exist the “real” experimental error can be estimated, therefore normal distribution is a better fit.
- [24] M. Manahan, M. Nelson, J. J. Cacciatore, J. Weng, S. Xu, J. Pollard, *Biotechnol. Prog.* **2019**, *35*, e2870.
- [25] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed., CRC Press, Boca Raton, FL **2017**.
- [26] Z. Xing, N. Bishop, K. Leister, Z. J. Li, *Biotechnol. Prog.* **2010**, *26*, 208.
- [27] S. Craven, N. Shirsat, J. Whelan, B. Glennon, *Biotechnol. Prog.* **2013**, *29*, 186.
- [28] M. J. Willis, M. von Stosch, *Comput. Chem. Eng.* **2017**, *104*, 366.
- [29] M. von Stosch, J.-M. Hamelink, R. Oliveira, *Bioprocess Biosyst. Eng.* **2016**, *39*, 773.
- [30] B. Bayer, M. von Stosch, G. Striedner, M. Duerkop, *Biotechnol. J.* **2020**, *15*, 1900551.
- [31] A. P. Teixeira, C. Alves, P. M. Alves, M. J. T. Carrondo, R. Oliveira, *BMC Bioinformatics* **2007**, *8*, 30.
- [32] H. Narayanan, M. Sokolov, M. Morbidelli, A. Butté, *Biotechnol. Bioeng.* **2019**, *116*, 2540.
- [33] A. Guerra, M. von Stosch, J. Glassey, *Crit. Rev. Biotechnol.* **2019**, *39*, 289.
- [34] Z. I. T. A. Soons, G. van Straten, L. A. van der Pol, A. J. B. van Boxel, *Bioprocess Biosyst. Eng.* **2008**, *31*, 453.
- [35] Z. I. T. A. Soons, J. A. Voogt, G. van Straten, A. J. B. van Boxel, *J. Biotechnol.* **2006**, *125*, 252.
- [36] R. Oliveira, Ferreira, E. C., Fayo de Azevedo, S., Stability, *J. Process Control* **2002**, *12*, 311.
- [37] D. Levisauskas, R. Simutis, D. Borvitz, A. Lübbert, *Bioprocess Eng.* **1996**, *15*, 145.
- [38] G. Bastin, D. Dochain, *Automatica* **1986**, *22*, 705.
- [39] S. T. F. C. Mortier, P.-J. Van Bockstal, J. Corver, I. Nopens, K. V. Gernaey, T. De Beer, *Eur. J. Pharm. Biopharm.* **2016**, *103*, 71.
- [40] B. Igne, Z. Shi, S. Talwar, J. K. Drennen, C. A. Anderson, *J. Pharm. Innovation* **2012**, *7*, 119.
- [41] C. N. Satterfield, *AIChE J.* **1973**, *19*, 206.
- [42] S. S. Ozturk, B. Ø. Palsson, *J. Biotechnol.* **1990**, *16*, 259.
- [43] R. M. C. Portela, A. Richelle, P. Dumas, M. von Stosch, *Microorganisms* **2019**, *7*, 620.
- [44] B. Bayer, B. Sissolak, M. Duerkop, M. von Stosch, G. Striedner, *Bioprocess Biosyst. Eng.* **2019**, *43*, 169.
- [45] M. Brendel, D. Bonvin, W. Marquardt, *Chem. Eng. Sci.* **2006**, *61*, 5404.
- [46] P. S. Swain, K. Stevenson, A. Leary, L. F. Montano-Gutierrez, I. B. N. Clark, J. Vogel, T. Pilizota, *Nat. Commun.* **2016**, *7*, 13766.
- [47] G. L. Tritsch, G. E. Moore, *Exp. Cell Res.* **1962**, *28*, 360.
- [48] Y. Sidorenko, A. Wahl, M. Dauner, Y. Genzel, U. Reichl, *Biotechnol. Prog.* **2008**, *24*, 311.
- [49] J. Möller, K. Bhat, K. Riecken, R. Pörtner, A.-P. Zeng, U. Jandt, *Biotechnol. Bioeng.* **2019**, *116*, 2931.