Reducing structural non-identifiabilities in upstream bioprocess models using profile-likelihood

Heiko Babel¹, Ola Omar¹, Albert Paul^{1*}, Joachim Bär¹

¹ Boehringer Ingelheim Pharma GmbH & Co.KG, Biopharmaceuticals Germany, Birkendorfer Strasse 65, D-88397 Biberach a. d. Riss, Germany;

*Corresponding author: Albert Paul albert.paul@boehringer-ingelheim.com

Abstract

Process models are increasingly used to support upstream process development in the biopharmaceutical industry for process optimization, scale-up and to reduce experimental effort. Parametric unstructured models based biological mechanisms are highly promising, since they do not require large amounts of data. The critical part in the application is the certainty of the parameter estimates, since uncertainty of the parameter estimates propagates to model predictions and can increase the risk associated with those predictions. Currently Fisher-Information-Matrix based approximations or Monte-Carlo approaches are used to estimate parameter confidence intervals and regularization approaches to decrease parameter uncertainty. Here we apply profile likelihood to determine parameter identifiability of a recent upstream process model. We have investigated the effect of data amount on identifiability and found out that addition of data reduces non-identifiability. The likelihood profiles of non-identifiable parameters were then used to uncover structural model changes. These changes effectively alleviate the remaining non-identifiabilities except for a single parameter out of 21 total parameters. We present the first application of profile likelihood to a complete upstream process model. Profile likelihood is a highly suitable method to determine parameter confidence intervals in upstream process models and provides reliable estimates even with non-linear models and limited data.

Keywords:

Parameter identifiability, Profile-likelihood, Modelling, Biopharmaceuticals

Introduction

Mathematical models find increasingly application in biopharmaceutical process development, which is encouraged by regulatory agencies through the PAT initiative of the FDA (FDA, 2004) and recommended for the use of process validation (FDA, 2011). In upstream processing mathematical models should increase the process understanding, reduce the experimental effort, increase the productivity and facilitate a Quality-by-Design approach (Narayanan et al., 2019; Varsakelis et al., 2020; Zobel-Roos et al., 2019). Currently there are different approaches used to model upstream bioprocesses. So-called hybrid models combining machine learning and mechanistic parts show promising results (Narayanan et al., 2019; Popp et al., 2016) for process optimization and process understanding, however parametrization or training of such hybrid models requires either larger amounts of data or expensive offline analytics of metabolites with a high time resolution.

In contrast macroscopic unstructured models use mechanistic insight and can be parametrized with single fermentation experiments. Unstructured models have been used for different approaches. Promising applications are process optimization (Möller et al., 2019a) and optimization of pH shifts (Paul et al., 2019). They have been used to uncover biological mechanisms that affect scale-up (Möller et al., 2019b), but also other upstream processes such as the seed-train have successfully been optimized with unstructured models (Hernández Rodríguez et al., 2019; Kern et al., 2016).

One major challenge in the use of macroscopic models is the determination of the model parameters. With non-linear models and limited data availability model parameters can be difficult to determine (Raue et al., 2009; Ulonska et al., 2018). Consequently the uncertainty in the parameter estimates propagates into uncertainty of the simulation (Anane et al., 2019; Kreutz et al., 2012; Liu and Gunawan, 2017b; Maiwald et al., 2016). It is therefore of high importance to determine not only the parameter values but also the confidence intervals of the parameters in a reliable manner, since this can critically affect the risk associated with the model predictions.

To this end several methods have been used (Rajamanickam et al., 2021). The classical approach uses numerical estimates of the Jacobi Matrix and approximates the confidence interval with a parabola (Meeker and Escobar, 1995), also known as Covariance approach or Fisher-Information Matrix (FIM) (Rischawy et al., 2019; Ulonska et al., 2018; Wieland et al., 2021). This estimate is true under the assumption of normal distributed errors and large number of data (Neale and Miller, 1997; Seber and Wild, 2003). This method is known to over or underestimate true

confidence intervals non-linear models with limited data (Moerbeek et al., 2004; Wieland et al., 2021). The method also relies on the numerical estimation of sensitivities or the Jacobi matrix which can be error-prone (Wieland et al., 2021). Another approach are Monte-Carlo approaches (Liu and Gunawan, 2017b; Möller et al., 2019b) which have the problem of handling unknown covariances and fine tuning of the sampling algorithm without autocorrelation (Hug et al., 2013).

In the field of systems biology an alternative approach to parameter confidence intervals has been established, the so-called profile-likelihood (Raue et al., 2009). In contrast to the FIM approach the profile-likelihood is an exact method and results in non-symmetric confidence intervals. Profile-likelihood is suited for non-linear models that are parametrized with a limited number of data-points (Moerbeek et al., 2004; Neale and Miller, 1997). In bioprocess modeling this approach has only been applied once by Kroll et. al. (Kroll et al., 2017). Here a model only for the viable cell density was parametrized with a single run with high quality data and high time resolution. It is currently unknown if the profile-likelihood method can be applied to complete upstream process models comprising all relevant state-variables, i.e. not only viable cell concentration but also dead cell, product and metabolite concentrations. To consider all available concentrations is of importance since metabolite concentrations and viability also influence product quality (Paul et al., 2018; Xu et al., 2016). Compared to Kroll et. al. (Kroll et al., 2017) process data is usually sampled daily with less precision. So far parameters of complete upstream processes have only been characterized by approximate confidence intervals but not exact confidence intervals (Ulonska et al., 2018) and exact confidence intervals for such process models with typical process data have not been determined.

To determine exact confidence intervals of complete process models we use the process model of Ulonska et. al. (Ulonska et al., 2018) comprising viable cell, dead cell, glucose, glutamine, lactate, ammonia and mAb product concentration and thereby modeling relevant process parameters. We then studied the effect of data availability by comparing profile likelihoods obtained from a single run and a data set with differential feed rate and seeding cell density. Next, we employed a profile-likelihood guided analysis to change the model structure and alleviated remaining non-identifiabilities. Finally, we compared the different model structures with increased parameter identifiability with respect to their fit to the experimental data.

Material and Methods

Experimental Data

We used the experimental data from the fed-batch cultivation of a CHO-DG44 cell line cultivated in a chemically defined medium. The five fermentation experiments to produce a monoclonal antibody were all conducted in controlled automated miniaturized bioreactor (ambr250) systems (Sartorius AG, Germany) for 14 days. Beside the center point, the data of additional four runs varying in feed rate and seeding cell density were used. Compared to the 'center point' the seeding cell density was increased or decreased by 30 %. The feed rate in the were decreased by 25 % and increased by 12 % respectively compared to the 'center point'. The feed was started after 24h of the cultivation start time. Apart from these variations, the cultivation took place under the same conditions. The temperature and dissolved oxygen was held constant. Starting from day 2, glucose boluses were applied and until day 3 glutamine boluses were added.

The glucose, glutamine, lactate and ammonia concentrations were measured daily using the Prime60 system (Software: Konelab Prime 7.2.1, Co. Thermo Fisher Scientific, USA). The titer concentration was measured starting from day 8 via IGG method using the Prime60 system. The viable cell density and the total cell density were measured using the Cedex HiRes (Roche, Switzerland). Dead cell density was calculated from the difference between viable and total cell density.

Mathematical Model

The process model was adapted from Ulonska et. al. (Ulonska et al., 2018) and is described in detail there. The differential equations and kinetics are shown in the supplementary text S1. In brief the model contains viable cell density, dead cells (or total cell density) and cell lysis. Product formation is proportional to viable cell concentration. Cell growth depends on glucose and glutamine and a glutamine containing supplement. In addition, the metabolites lactate and ammonia are considered. The kinetics were modeled using zero and first-order rates, Monod-terms and rational inhibition terms.

Model Parametrization and Profile Likelihood Analysis

We use the weighted sum of squared residuals as objective function for the parametrization of the models:

$$\chi^2(\theta) = \sum_{k=1}^m \sum_{l=1}^d \left(\frac{y_{kl}^D - y_k(\theta, t_l)}{\sigma_{kl}^D} \right)^2 \quad (\text{equation 1})$$

here y_{kl}^D is the data-point of the k-th observable at the l-th measurement timepoint, $y(\theta, t_l)$ is the model prediction for the corresponding datapoint with the parameter set θ . σ_{kl}^D is the standard deviation of the measurement. Under the assumption of normal distributed observational noise the weighted sum of squared residuals corresponds to the maximum likelihood estimate (MLE) (Raue et al., 2009), since then $\chi^2(\theta) = \text{const} - 2\text{Log}(L(\theta))$.

Instead of ‘normal theory’ based confidence regions derived from the asymptotic distribution of the ‘Wald statistic’ we use parameter confidence intervals derived from likelihood ratio tests (Meeker and Escobar, 1995):

$$\{\theta \mid \chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta\alpha\} \text{ with } \Delta\alpha = \chi^2(\alpha, df) \quad (\text{equation 2})$$

Wherein the borders represent the confidence interval with confidence level α . The threshold $\Delta\alpha$ is the α quantile of the χ^2 - distribution. When $df=1$ the pointwise confidence interval are obtained, and $df = \#\theta$ give simultaneous confidence levels (Raue et al., 2009). For finite samples likelihood based confidence intervals are considered superior compared to the asymptotic distribution of the ‘Wald statistic’ (Moerbeek et al., 2004; Neale and Miller, 1997).

Parameter identifiability is determined by the profile likelihood given by:

$$PL(\theta_i) = \min[\chi^2(\theta)] \quad (\text{equation 3})$$

A parameter is classified as ‘identifiable’ if the profile likelihood reaches the confidence threshold for lower and higher parameter values then the optimal parameter value.

The normalized root-mean squared error was calculated for each observable by:

$$NRMSE = \frac{\sqrt{\frac{\sum^N y_l^D - y(\theta, t_l)}{N}}}{y_{\max} - y_{\min}} \quad (\text{equation 4})$$

All calculations were performed in Matlab R2019a and profile-likelihoods were calculated using the data2dynamics toolbox (Raue et al., 2015) with the lsqnonlin optimizer of Matlab. Initial model parameters were estimated using a global optimization approach with latin-hypercube sampling within specified parameter bounds. Convergence was checked using the arLocalLHS function of data2dynamics.

Results

Parameter identifiability depends on the data

The minimal dataset for an upstream process models consists of a single fermentation. We fitted the state-of-the-art process model described by Ulonska et. al. (Ulonska et al., 2018) to a single center point run of a process development study (black curve in Figure 1). The model captures the dynamics of cell-growth and death (VCC and DCC) well, going from growth to death phase. The model fit agrees with the experimental data of glucose, lactate and glutamine and captures the product formation. It performs poorly for the ammonia formation. The experimental data showed for ammonia an initial increase until day 3, a slight decrease and then further increase starting at day 7 until the end of the process. Here, the model only increased ammonia values until day 4 and then stays constant with a slight decline until the end of the fermentation.

We next assessed if the model parameters that were determined from the single center point run are identifiable. The identifiability was determined by calculation of the profile-likelihood for each fit parameter. Of the 21 parameters 8 parameters could not be identified with the single run. Figure 2 shows that several profile likelihoods of the parametrization with a single run (black lines) do not hit the 95 % confidence threshold and do not significantly increase in the observed parameter range. The parameters biomass yields from glucose and glutamine, $Y_{BM,Glc}$ and $Y_{BM,Gln}$ respectively, have both a lower bound but no upper boundary. For the Monod-constant of lactate uptake k_{Lac} , the maximal Lactate uptake rate $q_{Lac,m}$ and the glucose maintenance m_{Glc} we do not observe any curvature in the profile likelihood, but only a flat profile. The inhibition constants $v_{Lac,Gln}$ and $v_{Lac,Prod}$ have no lower and upper boundary, respectively. The maximal glutamine uptake rate $q_{Gln,m}$ appears to mirror the biomass/glutamine yield coefficient with a defined minimum but no lower boundary. In the cases of e.g. the biomass/glutamine yield $Y_{BM,Gln}$ but also the ammonia formation rate $r_{NH4,Gln}$ the approximation with a parabola significantly underestimates the true parameter confidence interval (supplementary table 1).

Next, we wondered if these non-identifiabilities could be alleviated by parametrization of the model with additional datasets. To this end four additional runs with variations in seeding-cell density and feed rate were added and the model was reparametrized. The reparametrized model does not show differences in its simulation of the center point (blue line in Figure 1). We observe small deviations at day 8 compared to the previous parametrization in VCC and DCC. Also, the simulation for DCC deviates from the data at the final days of the process. The simulation of the

lactate concentration is slightly shifted in the reparametrized model and the lactate peak in the data appears to correspond better with the simulation.

As shown in figure 2 most of the optimal parameter values and therefore also the profile likelihoods are shifted by the addition of 4 complementary runs. For the two biomass yield coefficients $Y_{BM,Glc}$ and $Y_{BM,Gln}$, the inhibition constants $v_{Lac,Gln}$ and $v_{Lac,Prod}$, the glucose maintenance m_{Glc} and maximal Glutamine uptake rate $q_{Gln,m}$ the non-identifiabilities were resolved by the addition of new data, since the likelihood profiles now hit the limit for the 95 % confidence interval. The non-identifiability of the lactate uptake parameters k_{Lac} and $q_{Lac,m}$ could not be resolved by the additional experiments. Unexpectedly the addition of new data introduced new non-identifiability in the cell death kinetic. Both parameters of the cell-death kinetic, the maximal cell death rate $k_{D,m}$ and the inhibition constant by cell growth k_{μ} cannot be identified anymore. Data addition alleviated most non-identifiabilities, however some non-identifiabilities remained.

Structural non-identifiability of Lactate uptake kinetic and cell death kinetic

We next examined the remaining non-identifiabilities in detail. To this end we plotted the observed parameter changes in the profile-likelihood calculation of k_{Lac} and $q_{Lac,m}$ of the lactate kinetic and $k_{D,m}$ and k_{μ} of the cell death kinetic (Figure 3). For all those profiles the X^2 remained below 95 % confidence threshold, except for k_{Lac} where the profile reaches the 95 % confidence threshold for lower values than the optimal value.

Interestingly, for all profile-likelihood calculations we observe a linear correlation of the two parameter combinations. In the case of the lactate kinetic, k_{Lac} and $q_{Lac,m}$ exhibit a positive linear correlation with a slope of 0.99. The relation we observe is approximately $\log_{10} q_{Lac,m} \approx \log_{10} k_{Lac} + C$ which means that:

$$\frac{q_{Lac,m}}{k_{Lac}} \approx \text{constant} \text{ (equation 5)}$$

The range of optimal parameter values for k_{Lac} in the parametrization extends from 10^2 to $10^{3.5}$ g/L, this is highly above the observed lactate concentrations in CHO based bioprocesses. In this case the Monod-term in the lactate uptake kinetic can be approximated with the limit of $c_{Lac} \ll k_{Lac}$ and becomes:

$$q_{lac,con} = q_{Lac,m} * \frac{c_{Lac}}{k_{Lac} + c_{Lac}} * Inhibition \approx \frac{q_{Lac,m}}{k_{Lac}} * c_{Lac} * Inhibition \text{ (equation 6)}$$

To alleviate this non-identifiability the lactate uptake kinetic could be replaced by this approximation. This yields a new effective parameter $q_{Lac,lin} = \frac{q_{Lac,m}}{k_{Lac}}$ that replaces $q_{Lac,m}$ and k_{Lac} in the model.

For the cell death kinetic the correlation between $k_{D,m}$ and k_{μ} is negative with a slope of -0.99. In contrast to the lactate uptake parameters here the product of $k_{D,m}$ and k_{μ} is constant:

$$k_{D,m} * k_{\mu} \approx constant \text{ (equation 7)}$$

In this parametrization the values of k_{μ} are very small and the cell death kinetic can be approximated for the limit that $k_{\mu} \ll \mu$ to:

$$k_D = \frac{k_{D,m} * k_{\mu}}{k_{\mu} + \mu} \approx \frac{k_{D,m} * k_{\mu}}{\mu} \text{ (equation 8)}$$

Here it is not advisable to replace the current term with the approximated version since this could result in divisions by zero, however to alleviate the non-identifiability of $k_{D,m}$ we suggest to remove k_{μ} from the numerator.

Model structure adaptation alleviate non-identifiability

Based on the profile-likelihood calculations we could identify structural changes in the model that might alleviate the observed parameter non-identifiabilities. In a next step we implemented those structural model changes to see if they would indeed result in identifiable parameter values. We implemented single structural changes first and then combined them together in a fully adapted model.

As shown in figure 4 the remaining lactate-kinetic parameter $q_{Lac,lin}$ is identifiable in the model where only the lactate uptake kinetic is changed and the fully adapted model. Similarly we observe that the death rate parameter $k_{D,m}$ can be identified in the model where only the cell death rate is adapted and the fully adapted model. As expected, the cell death parameter k_{μ} can still not be identified as it tends to 0. We observe that the optimal parameter values are shifted slightly to higher values in the fully adapted model compared to the models with single model term adaptations.

Finally, we wanted to know how those adaptations in the model structure affect the overall fit of the model. To this end we calculated the normalized root mean square error (NRMSE) for all observables and the different models. We observe only slight differences in the model performance for all single model adaptations compared to the original model (figure 5). The

strongest deviations can be found for the fully adapted model. Here the NRMSE is increased by a value of 0.01 for both the dead cell concentration DCC and product concentration and increased by a value of 0.003 for VCC. However, the error is decreased by a value of 0.01 for both glucose and lactate concentration and decreased by values of 0.004 and 0.001 for glutamine and ammonia concentration respectively.

A different picture emerges if we look at the negative Log-Likelihood value of each model, which can be derived from the objective function (equation 1). The negative log-likelihood decreases when the lactate kinetic or the cell death kinetic are adapted by values of $-2 \log(L)$ of 5.9 and 1.7, respectively. However, when both adaptations are implemented in the model, the negative log-likelihood strongly increases by a $-2 \log(L)$ value of 97.

Discussion

The process model used in this work and previously described by Ulonska et. al. (Ulonska et al., 2018) describes fed-batch fermentation processes accurately, which we could confirm in this study with new data. Ulonska et. al. also described only a reasonable model fit to the ammonia data. The simulations showed an increase in ammonia concentration and then a plateau, however in the data clearly two periods of concentration increases can be observed. In this macroscopic model glutamine and a glutamine containing supplement are considered as sources for ammonia. Defined media used in current fed-batch processes also contain amino acids in the feed medium (Xing et al., 2011). Amino acids can also be a source for ammonia but are not described in this model. Addition of these additional ammonia sources might increase the descriptive capability of this model.

We could show that 6 parameters could be identified after the addition of new runs. The additional runs had variations in seeding cell density and in feed rate. Both experiments effectively change the nutrient availability. Variations in feed-rate change the available glucose that is provided in the feed, which can lead to the identification of parameters associated with glucose consumption, i.e. the yield and the maintenance metabolism. When the seeding cell density is shifted, this changes both the glutamine and glucose nutrient availability. Consequently, the yield coefficient for biomass from glutamine and the maximal glutamine uptake rate can be identified. The changes in nutrient availability also triggers changes in the metabolism, and therefore additional variations in lactate concentrations are introduced with the new dataset which enable the identification of the inhibition constants of the lactate production and consumption.

Since we used the same model as Ulonska et al. (Ulonska et al., 2018) we can directly compare the estimated parameter values. 9 of the 19 parameter values published there are in the same range. There are three major differences. Due to the non-identifiabilities determined in our study we have deviating parameter values for the lactate uptake parameters and the cell-death kinetic. This is to be expected since non-identifiabilities result in large confidence intervals for the parameter values. The biomass yield for glutamine and glucose have higher parameter values in our model, this is however compensated by lower uptake rates. Finally, our data suggests a higher production rate. The deviating parameter values are highly dependent on the clone and cell line (Ulonska et al., 2018) but also the process scale (Möller et al., 2019b) and media composition (Robitaille et al., 2015) can affect those parameters.

Although data addition alleviated parameter identifiabilities, one non-identifiability persisted, and an additional non-identifiability occurred. The lactate-uptake kinetic parameters could not be identified with the available dataset. The likelihood-profile showed a distinct lower limit of k_{Lac} making it a practical non-identifiability. In the used dataset no lactate concentrations are in the concentration range of k_{Lac} where the optimal parameter value is 10^3 g/L. In principle this non-identifiability could be removed by adding data where cells are treated with lactate concentrations $>10^3$ g/L. However, this is not feasible due to practical reasons such as the solubility of lactate and that additional effects might be observed by high lactate concentrations like growth inhibition (Lao and Toth, 1997). Therefore, the model structure needs to be adapted. Here we replaced the Michaelis-Menten uptake kinetic with its approximation for small substrate concentrations, i.e. a linear lactate uptake rate. We could reduce the number of model parameters from 21 to 20 and the new effective parameter $q_{Lac,lin}$ could successfully be identified from the available dataset. Moreover, introduction of this model structure change increased the fit significantly ($\Delta X^2=5.98$ $p=0.0145$ with $df=1$). These types of non-identifiabilities are a common problem in biological models and readily resolved after identification with the profile likelihood (Maiwald et al., 2016).

A more complex situation arose for the cell death kinetic. Here, we observed identifiability of the cell death parameters for the single center point run, but the parameters became non-identifiable after additional runs were added. We also observed that the model fit to the center point becomes worse if more data is added at the end of the process for the dead cell concentration (Figure 1 B). The profile-likelihood analysis revealed that the product of $k_{D,m}$ and k_{μ} is in fact constant and replacing this product in the cell-death kinetic with a new parameter rendered the new parameter identifiable. But the non-identifiability of k_{μ} could not be resolved. We suggest that the non-

identifiability arose due to a biological phenomenon that is observed in the data but not covered by the model structure. The experimental data in figure 1 showed that the viable cell concentration is constant after 288 hours, at the same time the dead cell concentration increases. Ignoring dilution effects, the viable cell concentration is described by:

$$\frac{dVCC}{dt} = \mu_{eff}VCC = (\mu - \mu_D)VCC$$

since we observe a positive death rate (increasing DCC) $\mu_D > 0$ and stationary viable cell concentration $\mu_{eff} = 0$ it follows that cells still replicate although they are at the same time dying and that $\mu > 0$. The cell death kinetic is formulated such that cell growth $\mu > 0$ and cell death $\mu_D > 0$ are mutually exclusive. Here we suggest that for our dataset a different mechanism plays a role that allows for this type of stationary viable cell concentration. Cell death might depend on limiting substrate concentrations or accumulation of toxic compounds as e.g. in proposed in the model of Pörtner et. al. (Möller et al., 2019b).

The likelihood profiles obtained in this study are often skewed and show the non-linearity of the parameter estimation problem for upstream process models. Especially for model parametrization with only one control run the profiles initially resemble parabolas but then have abrupt jumps. Jumps in profile-likelihood have been observed previously (Raue et al., 2009) and originate from local likelihood minima that are in close proximity to the global optimum. Here, approaches based on the Fisher-Information Matrix (FIM) or covariance (Ulonska et al., 2018), i.e. based on the Jacobi matrix at the optimal parameter point strongly over or underestimate the true parameter confidence intervals (see supplementary table 1). Figure 2 shows some symmetric confidence intervals; however, the parameter values are plotted on a log scale. Therefore, an exponential transformation would reveal that most parameters confidence intervals indeed do have asymmetric confidence interval on a linear scale. Methods which assume symmetric confidence intervals such as the FIM will in this case still over or underestimate the true confidence intervals. The likelihood profiles calculated from the full dataset are rather noisy. This can be due to the roughness of the optimization problem at hand, which might be introduced by non-steady events such as feed-starts and bolus addition in the simulation. Also, the profile-likelihood depends critically on the optimizer used. Only if the optimizer reaches an optimum a valid profile-likelihood can be obtained. Here adjustment of the optimizer parameters or different optimizers might reduce the noise currently observed in the likelihood profiles.

The uncertainty in the data of upstream bioprocesses and in the predictions of models have been identified as a critical issue in seed train predictions (Hernández Rodríguez et al., 2019), scale-

up prediction (Möller et al., 2019b) and process optimization (Liu and Gunawan, 2017a) . It has also been shown previously that uncertainty in the model parameters propagates to uncertainty of model predictions (Anane et al., 2019). Since uncertain predictions of process performance are from little value, only models with small uncertainties are of value for process development and commercial model use. It is therefore of high interest to correctly quantify parameter uncertainty and potentially decrease uncertainty of parameters. The profile-likelihood approach has clear advantages over the FIM approach in determining parameter confidence intervals, as it correctly predicts confidence intervals even for small amounts of data and non-linear models. Therefore, we suggest to use the profile-likelihood based confidence intervals as standard approach for mechanistic upstream process models. In addition to the exact determination of parameter confidence intervals, the profile-likelihood analysis is a guide to changes in the model structure that reduce non-identifiability and should lead to certain predictions that can be used for with high confidence in bioprocess development.

Conclusion

We show in this study that profile-likelihood can be applied to determine parameter confidence intervals and non-identifiabilities of state-of-the-art macroscopic upstream process models. Currently parameter non-identifiabilities have been tackled by regularization (Anane et al., 2019; Ulonska et al., 2018) and Bayesian approaches (Hernández Rodríguez et al., 2019), here we instead adopt an approach where the model structure is adapted to alleviate non-identifiabilities guided by the profile likelihood. This approach has successfully been applied in systems biology (Maiwald et al., 2016) and has the advantage that no priors or regularization parameters need to be defined. However structural changes might decrease the model fit and then models need to be selected carefully.

The profile-likelihood approach gives true confidence intervals even with small amounts of data (i.e. a single run) and has therefore clear advantages over FIM-based approaches. Since parameter confidence intervals directly influence the confidence intervals of predictions, it is necessary to *exactly* determine those confidence intervals. When models are used for predictions that are relevant for e.g. the biologics license application (BLA) of regulatory authorities, parameter identifiability is a key step in the characterization of the risk associated with the model predictions. Hence, it is an essential prerequisite for the commercialization of modeling approaches.

Acknowledgment

We would like to thank Alireza Ehsani, Sabine Arnold and Jens Smiatek for critical discussion of the results and Boehringer Ingelheim BPAD and IU for provision of the experimental data.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Nomenclature

c_{glc} – glucose concentration [g/L]

c_{gln} – glutamine concentration [g/L]

c_{lac} – lactate concentration [g/L]

c_{NH_4} – ammonia concentration [g/L]

c_{P} – product concentration [g/L]

c_{target} – target glucose or glutamine concentration [g/L]

DCC – dead cell concentration [10^9 cells/L]

F_{feed} – feed rate [L/h]

k_{d} – death rate [1/h]

$k_{\text{d,max}}$ – maximal death rate [1/h]

k_{glc} – affinity constant to glucose concentration [g/L]

k_{gln} – affinity constant to glutamine concentration [g/L]

k_{lac} – affinity constant to lactate concentration [g/L]

k_{lysis} – lysis rate of dead cells [1/h]

k_{μ} – inhibitory effect of growth rate on death rate [1/h]

m_{glc} – glucose maintenance [g/ 10^9 cells*h]

q_{glc} – specific glucose consumption rate [g/ 10^9 cells*h]

- 383 $q_{\text{glc,max}}$ – maximal specific glucose uptake rate [g/10⁹ cells*h]
- 384 q_{gln} – specific glutamine uptake rate [g/10⁹ cells*h]
- 385 $q_{\text{gln,max}}$ – maximal specific glutamine uptake rate [g/10⁹ cells*h]
- 386 q_{lac} – net rate of specific lactate metabolism [g/10⁹ cells*h]
- 387 $q_{\text{lac,con}}$ – specific lactate consumption rate [g/10⁹ cells*h]
- 388 $q_{\text{lac,max}}$ – maximal specific lactate uptake rate [g/10⁹ cells*h]
- 389 $q_{\text{lac,prod}}$ – specific lactate production rate [g/10⁹ cells*h]
- 390 q_{NH_4} – specific ammonia production rate [g/10⁹ cells*h]
- 391 q_p – specific production rate [g/10⁹ cells*h]
- 392 $r_{\text{NH}_4,\text{gln}}$ – rate of glutamine decay [1/h]
- 393 VCC – viable cell concentration [10⁹ cells/L]
- 394 $u_{\text{lac,glc}}$ – inhibitory effect of glucose consumption on lactate consumption [g/10⁹ cells*h]
- 395 $u_{\text{lac,gln}}$ – inhibitory effect of glutamine consumption on lactate consumption [g/10⁹ cells*h]
- 396 $u_{\text{lac,prod}}$ – inhibitory effect of lactate consumption on lactate production [g/10⁹ cells*h]
- 397 $Y_{\text{BM}/\text{glc}}$ – cells formed on glucose [10⁹ cells/g]
- 398 $Y_{\text{BM}/\text{gln}}$ – cells formed on glutamine [10⁹ cells/g]
- 399 $Y_{\text{BM}/\text{lac}}$ – cells formed on lactate [10⁹ cells/g]
- 400 $Y_{\text{gln}/\text{dGG}}$ – glutamine yield on media supplement [g/g]
- 401 $Y_{\text{lac,glc}}$ – lactate yield on glucose [g/g]
- 402 $Y_{\text{NH}_4,\text{gln}}$ – ammonia yield on glutamine [g/g]
- 403 μ – growth rate [1/h]

404

405 **References**

- 406 Anane E, López C DC, Barz T, Sin G, Gernaey K V., Neubauer P, Cruz Bournazou MN. 2019.
- 407 Output uncertainty of dynamic growth models: Effect of uncertain parameter estimates on

model reliability. *Biochem. Eng. J.* **150**:107247. <https://doi.org/10.1016/j.bej.2019.107247>.

FDA. 2004. Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. *FDA website*.

FDA. 2011. Guidance for Industry Process Validation : General Principles and Practices Guidance for Industry Process Validation : General Principles and Practices. *FDA website*.

Hernández Rodríguez T, Posch C, Schmutzhard J, Stettner J, Weihs C, Pörtner R, Frahm B. 2019. Predicting industrial-scale cell culture seed trains—A Bayesian framework for model fitting and parameter estimation, dealing with uncertainty in measurements and model parameters, applied to a nonlinear kinetic cell culture model, using an MCMC method. *Biotechnol. Bioeng.* **116**:2944–2959.

Hug S, Raue A, Hasenauer J, Bachmann J, Klingmüller U, Timmer J, Theis FJ. 2013. High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Math. Biosci.* **246**:293–304. <http://dx.doi.org/10.1016/j.mbs.2013.04.002>.

Kern S, Platas-Barradas O, Pörtner R, Frahm B. 2016. Model-based strategy for cell culture seed train layout verified at lab scale. *Cytotechnology* **68**:1019–1032.

Kreutz C, Raue A, Timmer J. 2012. Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Syst. Biol.* **6**.

Kroll P, Hofer A, Stelzer I V., Herwig C. 2017. Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. *Process Biochem.* **62**:24–36. <http://dx.doi.org/10.1016/j.procbio.2017.07.017>.

Lao MS, Toth D. 1997. Effects of ammonium and lactate on growth and metabolism of a recombinant Chinese hamster ovary cell culture. *Biotechnol. Prog.* **13**.

Liu Y, Gunawan R. 2017a. Bioprocess optimization under uncertainty using ensemble modeling. *J. Biotechnol.* **244**:34–44. <http://dx.doi.org/10.1016/j.jbiotec.2017.01.013>.

Liu Y, Gunawan R. 2017b. Bioprocess optimization under uncertainty using ensemble modeling. *J. Biotechnol.* **244**:34–44.

Maiwald T, Hass H, Steiert B, Vanlier J, Engesser R, Raue A, Kipkeew F, Bock HH, Kaschek D, Kreutz C, Timmer J. 2016. Driving the model to its limit: Profile likelihood based model reduction. *PLoS One* **11**:1–18.

437 Meeker WQ, Escobar LA. 1995. Teaching about approximate confidence regions based on
438 maximum likelihood estimation. *Am. Stat.* **49**.

439 Moerbeek M, Piersma AH, Slob W. 2004. A Comparison of Three Methods for Calculating
440 Confidence Intervals for the Benchmark Dose. *Risk Anal.* **24**:31–40.

441 Möller J, Kuchemüller KB, Steinmetz T, Koopmann KS, Pörtner R. 2019a. Model-assisted Design
442 of Experiments as a concept for knowledge-based bioprocess development. *Bioprocess*
443 *Biosyst. Eng.* **42**.

444 Möller J, Rodríguez TH, Müller J, Arndt L, Kuchemüller KB, Frahm B, Eibl R, Eibl D, Pörtner R.
445 2019b. Model uncertainty-based evaluation of process strategies during scale-up of
446 biopharmaceutical processes. *Comput. Chem. Eng.*:106693.

447 Narayanan H, Sokolov M, Morbidelli M, Butté A. 2019. A new generation of predictive models:
448 The added value of hybrid models for manufacturing processes of therapeutic proteins.
449 *Biotechnol. Bioeng.* **116**:2540–2549. <http://dx.doi.org/10.1002/bit.27097>.

450 Neale MC, Miller MB. 1997. The use of likelihood-based confidence intervals in genetic models.
451 *Behav. Genet.* **27**.

452 Paul AJ, Handrick R, Ebert S, Hesse F. 2018. Identification of process conditions influencing
453 protein aggregation in Chinese hamster ovary cell culture. *Biotechnol. Bioeng.* **115**.

454 Paul K, Rajamanickam V, Herwig C. 2019. Model-based optimization of temperature and pH shift
455 to increase volumetric productivity of a Chinese hamster ovary fed-batch process. *J. Biosci.*
456 *Bioeng.* **128**:710–715. <https://doi.org/10.1016/j.jbiosc.2019.06.004>.

457 Popp O, M€ D, Didzus K, Paul W, Lipsmeier F, Kirchner F, Niklas J, Mauch K, Beaucamp N. 2016.
458 A Hybrid Approach Identifies Metabolic Signatures of High-Producers for Chinese Hamster
459 Ovary Clone Selection and Process Optimization. *Biotechnol. Bioeng.* **113**:2005–2019.
460 <http://onlinelibrary.wiley.com/doi/10.1002/bit.25958/abstract>.

461 Rajamanickam V, Babel H, Montano-herrera L, Ehsani A, Stiefel F. 2021. About Model Validation
462 in Bioprocessing. *Processes* **9**:1–16. <https://www.mdpi.com/2227-9717/9/6/961>.

463 Raue A, Steiert B, Schelker M, Kreutz C, Maiwald T, Hass H, Vanlier J, Tönsing C, Adlung L,
464 Engesser R, Mader W, Heinemann T, Hasenauer J, Schilling M, Höfer T, Klipp E, Theis F,
465 Klingmüller U, Schöberl B, Timmer J. 2015. Data2Dynamics: A modeling environment
466 tailored to parameter estimation in dynamical systems. *Bioinformatics* **31**:3558–3560.

- 467 Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009. Structural
468 and practical identifiability analysis of partially observed dynamical models by exploiting the
469 profile likelihood. *Bioinformatics* **25**:1923–1929.
- 470 Rischawy F, Saleh D, Hahn T, Oelmeier S, Spitz J, Kluters S. 2019. Good modeling practice for
471 industrial chromatography: Mechanistic modeling of ion exchange chromatography of a
472 bispecific antibody. *Comput. Chem. Eng.* **130**:106532.
473 <https://doi.org/10.1016/j.compchemeng.2019.106532>.
- 474 Robitaille J, Chen J, Jolicoeur M. 2015. A single dynamic metabolic model can describe mAb
475 producing CHO cell batch and fed-batch cultures on different culture media. *PLoS One* **10**.
- 476 Seber GAF, Wild CJ. 2003. Nonlinear Regression. *John Wiley Sons*.
- 477 Ulonska S, Kroll P, Fricke J, Clemens C, Voges R, Müller MM, Herwig C. 2018. Workflow for
478 Target-Oriented Parametrization of an Enhanced Mechanistic Cell Culture Model. *Biotechnol.*
479 *J.* **13**.
- 480 Varsakelis C, von Stosch M, Dessoy S, Pysik A. 2020. A mathematical model of decision making
481 for investment in process modeling & simulation. In: . *AIP Conf. Proc.*, Vol. 2293, p.
- 482 Wieland F-G, Hauber AL, Rosenblatt M, Tönsing C, Timmer J. 2021. On structural and practical
483 identifiability. <http://arxiv.org/abs/2102.05100>.
- 484 Xing Z, Kenty B, Koyrakh I, Borys M, Pan SH, Li ZJ. 2011. Optimizing amino acid composition of
485 CHO cell culture media for a fusion protein production. *Process Biochem.* **46**.
- 486 Xu S, Hoshan L, Chen H. 2016. Improving lactate metabolism in an intensified CHO culture
487 process: productivity and product quality considerations. *Bioprocess Biosyst. Eng.* **39**.
- 488 Zobel-Roos S, Schmidt A, Mestmäcker F, Mouellef M, Huter M, Uhlenbrock L, Kornecki M,
489 Lohmann L, Ditz R, Strube J. 2019. Accelerating biologics manufacturing by modeling or: Is
490 Approval under the QbD and PAT approaches demanded by authorities acceptable without
491 a digital-twin? *Processes*. MDPI AG.

Figure 1: Original model fit to center point or full dataset. All curves are normalized to the maximal data-value. Experimental data is marked with square-symbols and the error-bar correspond to the standard deviation from the analytical measurement method. Lines denote the model simulation of the center-point run of a model parametrized only with the center-point (black line) or with the center point and 4 additional runs (blue line).

Figure 2: Profile-Likelihood of parameters from a model fit to center point and full dataset. Profile likelihood from the model parametrized with a single center point run are shown in black, the model parametrized with the full dataset is shown in blue. The dashed line shows the point-wise 95 % confidence level. The minimum value of the objective function χ^2 for each parametrization was subtracted from the profiles for comparability

Figure 3: Non-Identifiabilities in model with full dataset. (A-B) Profile likelihood of lactate uptake kinetic parameters and cell death kinetic parameters as a function of k_{Lac} and k_{μ} respectively. Dashed line shows 95 % confidence threshold. (C) Optimized parameter values of $q_{Lac,m}$ and k_{Lac} of the profile-likelihoods for $q_{Lac,m}$ and k_{Lac} . (D) Optimized parameter values of $k_{D,m}$ and k_{μ} of the profile-likelihoods for $k_{D,m}$ and k_{μ} .

Figure 4: Profile Likelihood of Parameters of Adapted Models (A) Profile likelihood of $q_{Lac,lin}$ for a model where only the lactate kinetic is adapted and both lactate and cell death kinetic are adapted. (B) Profile likelihood of k_{μ} for a model where only the cell death kinetic is adapted and both lactate and cell death kinetic are adapted. (C) Profile likelihood of $k_{D,m}$ for a model where only the cell death kinetic is adapted and both lactate and cell death kinetic are adapted.

Figure 5: Model fit of different model adaptations (A) Normalized root mean squared error (NRMSE) for the observables with the original model, the lactate kinetic adapted model, the cell death kinetic adapted model and the fully adapted model. (B) Negative Log-Likelihood of the different models.