Universität für Bodenkultur Wien

Hybrid Modelling and Quality by Design Implementation in Upstream Processing

Dissertation

zur Erlangung des Doktorgrades

an der Universität für Bodenkultur Wien

Department für Biotechnologie

Institut für Bioverfahrenstechnik

Vorstand: Reingard Grabherr, Univ. Prof. DI Dr.

Betreuer: Gerald Striedner, Assoc. Prof. DI Dr.

Eingereicht von

Benjamin Bayer, MSc.

Wien, Juli 2020

Acknowledgements

First, I want to thank Associate Professor Doctor Gerald Striedner, who was my supervisor during my master's thesis at the University of Natural Resources and Life Sciences, Vienna, (BOKU) and who offered me the opportunity to pursue this doctoral thesis early on. I consider myself very fortunate that you were my working group leader and I have good memories of our collaboration. In addition, I also want to thank all other members of the working group for their support and all the fun we had, inside and outside of the laboratory. My special thanks go to Florian Strobl, who dedicated much time to training me at the beginning and who later, most importantly, taught me how to stay calm and focused, even when something went wrong or almost exploded.

Special thanks also to my project team at Novasign GmbH, Mark Dürkop, Maximilian Krippl, Roger Dalmau Diaz and Armin Khodaei. We faced many challenges but it was always a pleasure to work with you. This project would have been far more difficult and stressful without the great atmosphere within the team. Moreover, many thanks to one of my best friends, Lina Vranitzky, who I additionally supervised during her master's thesis. She taught me to be patient, take things with humour and to stay calm in tricky situations. I hope she had as much fun in the laboratory and enjoyed this time as much as I did. Big thanks also to my advisory team during this doctoral thesis, Moritz von Stosch and Oliver Spadiut, who gave me input and critical review, whenever it was necessary.

Thanks also to my closest friends. Some of them I have known since we started school together and others I met later, either starting as a '+1' or during my bachelor or master's studies. All of you brought fun into my life through game nights, booze-related activities, journeys, spa visits, or simply by spending time together. Thank you for that; I value this kind of support highly. Of course, I also want to mention my handball club, my team and all former teammates. Although you all thought I was a farmer for the first two years of my doctoral studies, understanding 'tractor' instead of 'reactor', you supported me anyway. Your distractions, when I was working too much, including several events such as our boys' nights just hanging out, the Presslufthammer trips and tournaments were always amazing.

Last but not least, thanks to every single member of my family. You were always there, encouraging me, showing interest in my studies and sharing pieces of advice, of which some were actually helpful. Herein, special thanks to my parents, Chantal Rodgarkia-Dara and Wolfgang Bayer for their unconditional support, interest and honest opinions in every situation. Thanks also to my siblings, Johanna, Henrik, Alexander and Elinja, for a lot of distraction. They always managed to change the subject to something unrelated to work, since beating the high score at Just Dance, playing football or UNO was always way more important. I would never trade a single one of these experiences.

Kurzfassung

Die Charakterisierung und Optimierung von Bioprozessen zur Herstellung von Therapeutika ist eine zeit- und resourcenintensive, jedoch wichtige Aufgabe in der biopharmazeutischen Industrie und hierfür notwendige Technologien wurden bereits im Rahmen der “Quality by Design“ Initiative der FDA vorgeschlagen. Die prominentesten Ansätze für ein besseres Prozessverständnis sind statistische Versuchsplanung kombiniert mit Prozessmodellierung. Diese wurden jedoch bis heute unzureichend implementiert, weshalb eine hohe Anzahl an benötigten Experimenten gepaart mit veralteten unzureichenden Modellierungstechniken noch immer Standard in der Industrie ist.

Um diese Schwachpunkte zu adressieren, wurde zur Datengenerierung eine Versuchsplanung mit *Escherichia coli* Fed-Batch Kultivierungen (20L) in einem dreidimensionalen Designspace erstellt. Dieser Datenraum wurde einmal mit statischen und ein zweites Mal mit intra-experimentellen Sollwertänderungen der zu charakterisierenden Prozessgrößen experimentell erarbeitet.

Die in dieser Arbeit entwickelten Tools zur “Quality by Design“ Implementierung umfassen eine Off-Line Technik zur präzisen Bestimmung von spezifischen Raten, einen On Line Softsensor zum Biomassemonitoring, sowie ein Hybridmodel, welches zeitgleich die Biomassekonzentration und den löslichen Produkttiter in Fermentationen vorhersagen kann. Um lange Entwicklungszeiten zu adressieren, wurde das Konzept intensivierter Kultivierungen entwickelt (intra experimentelle Parametershifts). Mit diesen Daten wurde ein intensiviertes Hybridmodel erstellt.

Es wurde gezeigt, dass Hybridmodelierung gegenüber dem derzeitigen Stand der Technik überlegen ist. Das intensivierte Hybridmodel zeigt eine vergleichbare Leistung gegenüber dem voll faktoriellen Design, benötigt jedoch nur ein Drittel der Daten, was zu >66 % Zeitersparnis führt. Diese beeindruckenden Leistungsdaten demonstrieren das hohe Potential dieses kombinierten Ansatzes zur Beschleunigung der Bioprozesscharakterisierung.

Abstract

Upstream bioprocess characterization and optimization are time and resource-intensive but essential tasks in the biopharmaceutical industry and necessary technologies to promote this concept have already been suggested in the FDA's quality by design initiative. Within this initiative, prominent approaches to generate process understanding are design of experiments studies in combination with process modelling. However, due to insufficient implementation of these methodologies, large numbers of required experiments paired with outdated and inadequate modelling techniques are still state-of-the-art in the industry.

To eliminate these major issues, design of experiments studies for *Escherichia coli* fed-batch fermentations (20L) in a three-dimensional design space were performed. This design space was completely characterized twice, with cultivations operated either with static or dynamic process parameter settings.

The tools for implementation of quality by design concepts in upstream bioprocessing developed in this work comprise an off-line applicable method for accurate specific rate calculations, a soft sensor for biomass monitoring and an advanced hybrid model for simultaneously predicting the biomass concentration and soluble product titre. To address long development times, the concept of intensified design of experiments was introduced, i.e., intra-experimental setpoint shifts of critical process parameters. Based on these data, an intensified hybrid model was developed.

The results demonstrated that the developed tools have superior performance compared with state-of-the-art modelling techniques. The hybrid model, based on the intensified cultivations, performed comparably but only required one-third of the data for model-training compared to the static hybrid model, resulting in >66 % fewer experiments. This demonstrates the high value of this combinatorial approach for accelerating bioprocess characterization and advancing the quality by design initiative.

Keywords: hybrid modelling, machine learning, process control, process monitoring, quality by design

Table of Contents

1 Introduction

1.1 Recombinant Protein Production

Recombinant protein production for therapeutic use is a tremendously important market with an increasing value every year [1]. The used organisms for expressing the desired products are versatile and depend on the characteristics of the product. Microorganisms, e.g., bacteria, yeasts or microalgae are a cost-efficient and fast choice for the expression of relatively simple, non-glycosylated proteins such as insulin. Otherwise, larger molecules, such as antibodies also need post-translational modifications, e.g., glycosylation and have to be expressed in higher organisms, i.e., mammalian cells such as Chinese hamster ovary cells (CHO) [2]. However, regardless of the chosen host for protein production, the state-of-the-art operating procedure for recombinant protein production applies to all these systems, i.e., to operate a process according to a fixed protocol, ensuring process performance consistency. However, the state-of-the-art procedure contains major weaknesses, e.g., the inputs to the process contain unavoidable variabilities and occurring deviations are only investigated after the process has ended. The resulting fluctuations in product quality are currently often only tested after the process. This means that no corrective actions to avoid possible batch rejection due to the addressed incidents are possible in real-time [3]. Moreover, to guarantee higher quality standards, the specifications for releasing a batch are narrowed repeatedly. Since the process input variability is unaffected, only testing the output specifications in a tighter acceptance range increases the number of rejected batches. This inevitably leads to heavy losses with respect to time, money and material goods [4]. Counteracting these issues requires generating process knowledge, which until now has only been considered during the initial development stage of an upstream process.

1.2 Upstream Process Development

The development of an upstream process demands a resource-intense investigation of multiple factors to find the most influential impacts on the process performance. This laborious characterization requires an enormous amount of time. The most prominent approaches for process development are to investigate: the media composition [5], [6], the cell line used [7], [8], the expression vector [9], metabolic engineering [10], [11] and the addition of a single substance [12], [13]. Different process parameter settings also need investigating e.g., stirrer speed, dissolved oxygen, cultivation duration, temperature, pH, induction strength or the feeding strategy [14]. Up to now, plenty of research has addressed the development stage of

processes in various organisms, putting high effort into this initial upstream process development.

However, once the final optimal process conditions are found, e.g., to obtain the highest product yield, the process is always carried out with this setup. This results in a long-lasting and expensive procedure, investigating many factors while simultaneously only generating little understanding of the process. Consequently, this long development time from discovery until a product enters the market and the return on investment begins leads to high costs per dose, which are eventually covered by the patient and insurers.

An additional shortcoming to this approach is the inevitable variance from the many process inputs and the process operation itself. These factors all influence the final product quality of every single process, which must be determined before a product batch can be released. The state-of-the-art procedure to test the quality of a product batch is called quality by testing (QbT) and is only performed after the process ends. Only the final product quality is determined, therefore and according to these results and the required quality specification, a tested batch is either released or rejected.

1.3 Quality by Design and Process Analytical Technology

To guarantee higher quality standards through QbT, the specifications for post-process testing of the product quality can be raised. However, a variable input to a process that is always performed in the same way will always lead to a variable output. Following this approach leads to an increased number of batches being rejected, due to the steadily increasing standards, while the operating procedure of the process itself stays unchanged. This QbT approach is therefore disadvantageous and is also economically impractical because of the long development times [15]. To counteract issues for quality assurance, e.g., insufficient process understanding, batch-to-batch variability, lack of process knowledge and inadequate on-line monitoring, the U.S. Food and Drug Administration (FDA) proposed the process analytical technology (PAT) guidance, for science and risk-based technologies in the biopharmaceutical industry. Foremost, these guidelines emphasize technical risk assessment to identify the critical process parameters (CPP), which have a strong and direct impact on the product's critical quality attributes (CQA), e.g., correct folding and post-translational modifications. This risk assessment and subsequent risk management should already be implemented during the development phase of a process [16] and continually integrated until the end of the life cycle [17] to permanently guarantee a stable and controlled output quality [18]. This process characterization, investigating the multidimensional impacts of the CPPs at different levels on the CQAs can rapidly lead to an unfeasible number of required experiments. However, planning and setting up a statistical design of experiments (DoE) offers a manageable solution

for process characterization, i.e., to systematically investigate the relationship between the defined input factors and the process response of interest. Hereby, the main effects and interactions of different identified process parameter combinations and levels can be investigated, e.g., different pH setpoints or cultivation temperatures, on the process output, e.g., the product [19]. To keep the number of experiments for such DoE studies to a manageable quantity, different designs can be applied, with the maximum number of combinations being investigated using a full factorial design [20]. By choosing multiple levels of each CPP, a better process understanding of the respective impacts is achieved. Typically, reduced screening designs, e.g., definite screening, fractional factorial, Box-Behnken, or Doehlert, are applied in the initial phase to exclude negligible factors and find important factors, including the levels at which further experiments should be performed [6]. As a result, the area of interest in the original DoE can be identified and extended to derive robust process settings and to describe their outcomes [10]. However, conclusions drawn from a reduced design have to be handled with care, since different amounts of information are generated [21]. For optimization purposes, in a later phase a full factorial design investigating all CPP combination settings can be performed [22].

The effect of parameter changes during cultivation is also one of the most discussed and crucial issues [23], [24]. In addition, an on-line monitoring strategy, ensuring real-time monitoring of the variables of interest, typically biomass and product titre, needs to be defined for this design space. PAT is recommended as the key to establish such in-process monitoring and control through improved understanding and to fulfil the regulatory quality requirement to the extent that the requested product quality is already assured by the process itself. By following these guidelines, quality does not have to be tested afterwards [25]. Meeting these guidelines achieves a paradigm shift from the state-of-the-art QbT approach, which follows a ’recipe‘ and tests the quality of the end product, to a quality by design (QbD) concept. This QbD concept aims to gain deeper process insights, improve process understanding and as a result successfully counteract the ever present batch-to-batch-variability, e.g., fluctuations in the cultivation temperature or temporary pump malfunctions [26]. The main objective of this proposed guidance is on-line monitoring, testing and control [27]. Besides DoE, many built-in tools are recommended that can be applied to contribute to this guidance for the industry, e.g., multivariate data analysis (MVDA), soft sensor development and advanced process control (APC) as well as the use of process models or even model predictive control (MPC) [28]. This allows the manufacturing process and the monitored process variables to be well-understood and to operate as expected, since deviations are noticed promptly. This allows input variations to be addressed, ensuring a more constant and uniform product output to the extent that the quality standards and batch-release are guaranteed [29], as outlined in **Fig. 1**.

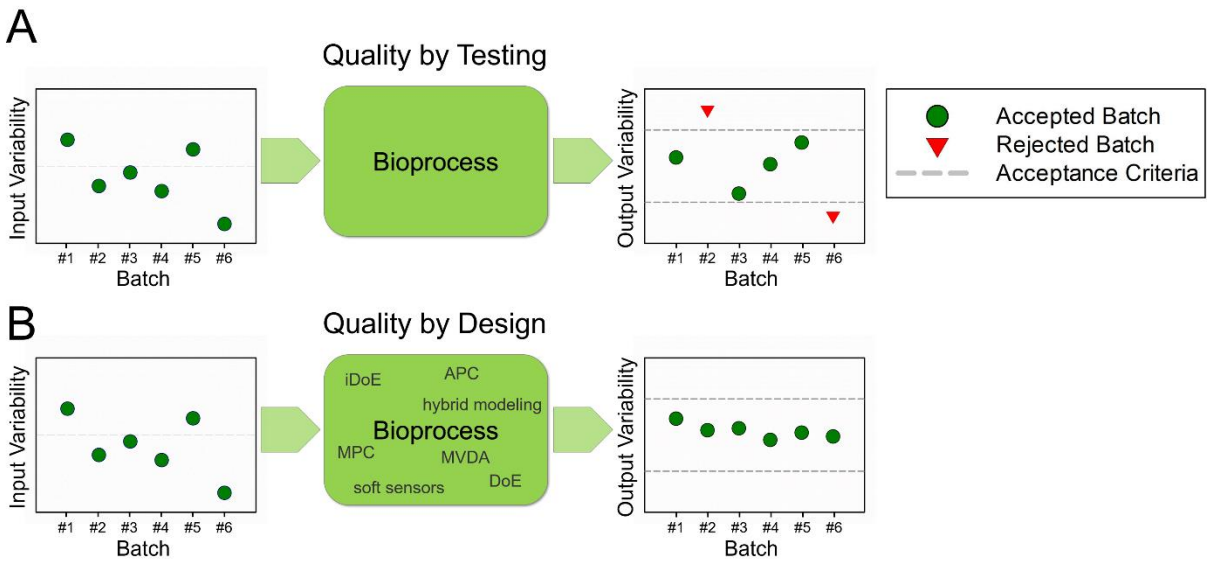

Figure 1. The working principles of QbT and QbD. The input quality to a bioprocess is inevitably variable. In the QbT approach (A), the process settings are fixed without any intervention and the quality of the batch is tested afterwards. Depending on the quality, the batch is rejected or released. In the QbD approach (B) built-in tools and strategies enable real-time monitoring and control of the process, always ensuring a robust process performance and uniform quality output to avoid batch rejection.

PAT focusses mainly on on-line monitoring of process variables, which cannot usually be accessed in real-time, e.g., the current biomass and the product titre. Accessing these variables in real-time is possible with PAT, providing not only highly valuable information about the current process state but also APC. This enables monitoring and control not only of the standard process parameters, e.g., pH, temperature, feed and inducer addition, but also the addressed variables [30]. The PAT guidelines specify the development of advanced on-line monitoring tools to achieve this goal [31]. The number of available sensor systems for on-line and at-line measurements is enormous as comprehensively reported elsewhere [32] and well-established sensors for physical and chemical variables (e.g., pH, dissolved oxygen, optical density, temperature and pressure) have already been part of the standard equipment for up- and downstream processes for a long time [33]. Furthermore, advanced sensor systems are already applied in the industry to measure more specific process variables and the number of systems is increasing [34]. By now, such advanced systems include, e.g., devices that can analyse the composition of volatile organic compounds in the cultivations' off-gas [35], other advanced sensors that can measure the broth itself, sensors to measure permittivity and conductivity [36], [37] and a group of non-invasive optical systems [38]. Amongst these, the most prominent representatives are Raman spectroscopy [39], [40], infrared spectroscopy [41], [42], 2D-fluorescence spectroscopy [43]–[46] and *in-situ* microscopy [47], [48]. The advantages of these methods are that their measurements are non-invasive, continuous and non-destructive, i.e., no sample has to be drawn, eliminating the risk of contaminations. A

detailed overview of these and further techniques including their respective functionalities, advantages and disadvantages is given by Rowland-Jones et al. [49].

QbD applications, advanced on-line monitoring tools and possible implementations have already been investigated and published extensively. Nevertheless, in contrast to other industries and related fields of biotechnology, the absence of suitable on-line monitoring tools to directly measure the variables of interest presents an obstacle. Especially in upstream processes, there is still a long way to go to implement QbD as the new state-of-the-art approach [50]. This further highlights the need for process characterization and process modelling to advance the QbD concept.

1.4 Process Characterization

Statistical DoE has already been established as the state-of-the-art approach for upstream bioprocess characterization, i.e., to plan a design space, perform the experiments, record the on-line data from all the sensor systems and to gather the experimental off-line data by taking samples of the process endpoints and evaluating them. The common, fastest and easiest way to analyse the results obtained from the characterized design space is to visualize the endpoint measurements. Typically, this is done as a response surface to find the optimum in the investigated space, in which the production process, later on, will be carried out [51]. The drawbacks using this common approach are that no process knowledge or understanding is generated over how the process parameters and the process variables of interest are related. Thus, no process model is developed, preventing the possibility of understanding the process and CPP interactions.

A more advanced way to evaluate the endpoint is to develop a generic response surface model (RSM) of the analytical results [52]–[56]. For this most common modelling technique, linear, quadratic and interaction terms can be taken into account, fitting the experimental results to a surface as responses of the investigated parameters [57], [58]. The relative importance and the relationship between the individual CPPs and the process variables to be modelled can be accessed [20]. By utilizing this RSM, initial process understanding on a low level is generated. However, by only describing and modelling the process endpoints of the design space and neglecting the remaining process, almost all informational value is lost, i.e., the optimum of a process may not be at the end but some point before and therefore remains unrecognized. Moreover, process parameters, e.g., the cultivation temperature, are assumed to be constant, which is not the case typically. This assumption of a constant process, when in reality only the parameters are held constant, results in varying process trends and output qualities. Further, deviations and temporary errors remain hidden, e.g., a malfunction of a balance or a pump

defect. Neglecting these process dynamics is a major issue concerning accurate process characterization, evaluation and understanding [59].

The boundaries of the static process view, solely focussing on the endpoint, can be overcome by developing time-resolved process models. This supports on-line monitoring of process variables of interest enabling deeper process insights and observation of potential inconsistencies occurring in real-time, an advance in the right direction to enable APC. Authorities are also promoting regulatory requirements in this respect and proposing to abandon the pure endpoint evaluation. To make such an APC possible, the previously addressed drawbacks and weaknesses regarding process characterization and evaluation must be eliminated and the entire process duration needs to be considered. This time-resolved process modelling is of high value and a necessity for APC applications, providing that the adequate modelling approach is chosen. This implies periodic sampling for off-line analysis, which enables time-resolved trajectories of the variables of interest, investigation of the CPP impacts and detection of deviations occurring [60], [61]. To enable such time-resolved process modelling with the generated and recorded data, different modelling techniques are frequently applied [62] and wide-ranging MVDA is possible for further evaluation of the design space [63].

1.5 Process Modelling

The input data to such time-resolved process models need to be derived from the on-line monitoring tools and probes to obtain the model estimation in real-time. Although the above mentioned advanced on-line sensor technology is making progress, it is still not possible to directly measure important process variables on-line. Therefore, sampling cannot be abandoned, i.e., off-line analysis is necessary to determine the quantitative and qualitative product attributes, such as biomass, product titre, concentrations of media components, waste and by-products, but also as a reference to validate the on-line sensor measurements [64]. Efforts to solve this issue through mathematical modelling have already been attempted for more than 20 years [65].

As the first step to address this problem, unsupervised learning, a part of MVDA used for exploratory data analysis, can be applied to investigate the data and gain more knowledge by revealing hidden structures and identifying system components. For this purpose, the most used modelling techniques are principal component analysis (PCA) [66] and parallel factor analysis (PARAFAC) [67], [68]. Variables and structures identified using these analyses can be further used for supervised learning approaches, i.e., to model the relationship between a dependent target variable (e.g., the biomass) and independent predictor variables (e.g., process parameters). Furthermore, two modelling approaches are distinguishable,

non-parametric (black box) and parametric (white box) modelling [69]. Black box models are referred to as data-driven approaches, built only on experimental data; applied parameters do not have a physical meaning and no further process knowledge or understanding is needed, i.e., the model structure is inferred solely from the data. They can thus be easily developed by applying various regression techniques. White box models are based on first principles and empirical considerations, i.e., their structure is well-defined, knowledge-based and transparent since the parameters used possess a physical meaning. Since these models assume a pure mechanistic trend, they can extrapolate properly. However, biological processes especially have high variability, which can be a major issue; these white box predictions, therefore, frequently lead to inaccurate results [70]. Moreover, their development is laborious and the applied equations are too simple for the complexity of the processes [71]. In contrast, black box models are well-known for capturing the variability occurring in biological processes and therefore possess suitable interpolation capabilities. The black box model is limited to application within the range it has been trained on, i.e., it lacks accurate extrapolation capabilities [72]. Although the respective advantages and disadvantages for both approaches are well-known, both modelling techniques are becoming increasingly popular and are frequently applied in process modelling, to estimate target variables in the QbD concept [44], [73]–[76].

Two types of models exist: descriptive and predictive. Descriptive models such as soft sensors provide real-time information, i.e., changes in on-line signals can be used to correlate unspecified measurements from the on-line hardware sensors with the results derived from the off-line analysis. Specific process variables can be estimated in this way, but a value can only be estimated up to the current time point of the process [77]–[79]. For these applications, typically more advanced additional on-line tools such as sensors for spectral data are applied, e.g., Raman spectroscopy, near-infrared spectroscopy and 2D-fluorescence [80]. An additional advantage using 2D-fluorescence is that non-fluorescent compounds can also be estimated, if their stoichiometric relationship to fluorescent compounds in the process, e.g., $NADP^+$ [81], [82] or flavins [83], [84], is known [85]. Likewise, Raman spectroscopy is already used for real-time estimation of the antibody titre [86] and to measure and control glucose and lactate levels in CHO cultivations [87]. Since these still do not measure the variable itself but only the respective correlated signal and possess hardware-related constraints, this cannot be seen as a reliable solution. There is always the risk of correlation without causality, leading to misinterpretation. These opportunities, possibilities and limitations have already been widely discussed [88].

In contrast to descriptive models, predictive models are not based on process responses but on direct process inputs, e.g., the cultivation temperature or the amount of accumulated feed.

Since the input data can be simulated for the future, these models are also able to predict future values and provide an educated estimate about the trajectories [89].

In summary, the PAT initiative has encouraged the development of advanced sensor technology and supported and highlighted the value of process modelling. However, currently implemented state-of-the-art systems and modelling techniques still have considerable drawbacks and limitations, which remain to be solved.

1.6 Limitations

A major point to consider is the cost-intensive use and maintenance of the available advanced sensor technologies used for process modelling. With this system, the measured signals are only correlated to the process variables of interest, i.e., the system still lacks the ability to directly measure process variables, which is of high interest for QbD implementation. Another issue, which should not be neglected, is that the output of such sensor systems does not typically consist of only one variable but rather up to thousands in the case of, e.g., infrared spectroscopy. The knowledge and know-how required to interpret the high dimensional and heterogeneous process data generated are often not available for the whole process as has been reported earlier [90].

A further limitation in bioprocesses, complicating QbD implementation, is the biological component. This introduces an additional source of fuzziness to the system. The resulting process deviation due to all these factors leads to noisy data. An example is cell counting methods in mammalian processes, which possess high variability [91]; this complicates data interpretation and thus the calculation of specific rates, e.g., for the growth rate or the substrate consumption rate. Therefore, the estimation of these highly valuable specific rates is still a major issue. However, access to these process variables is of high interest and important to generate process knowledge, e.g., the examination of CPP impacts, more detailed process investigation and batch-to-batch comparison. However, using the state-of-the-art methods in bioprocess engineering, high analytical errors as yet prevent a precise determination of the specific rate values, rendering such investigations unacceptable.

Besides the knowledge gap mentioned above, the problem of solely focussing on the endpoint of a process, neglecting the remaining cultivation and inappropriate methods to calculate and compare process characteristics, by far the main deficiency is the lack of reliable and robust modelling of the entire fermentation process. With the presented state-of-the-art modelling methods, i.e., DoE studies in combination with RSM, most of the process information is missing. Although many methods and process models have already been published since the start of the PAT initiative, both black box and white box models have revealed boundaries,

e.g., error-prone estimations due to missing extrapolation properties or straight mechanistic assumptions [92]. These modelling approaches possess the above-mentioned unique advantages compared to the other, but both have severe disadvantages and limitations due to their respective model structures and fail to deliver an overall steady performance. Furthermore, interference in the process, e.g., to react to real-time process influences or to guide the process towards the desired direction, is not possible using such model structures. These issues can only be solved by combining both approaches in a single and more robust model with a superior generalization capability.

Implementation of a control system in the QbD concept, such as MPC for the upstream process, first requires the development of a robust predictive model, based only on controllable parameters [93]. In this way, the basis of the missing element to control the process (via the parameters), is introduced. Moreover, MPC will finally lead to a dynamic and flexible process. Once such a controllable model is developed and implemented, it will always be able to deliver a robust and uniform output, avoiding batch-rejection, as the PAT guidance demands. To develop such a robust predictive model, which can be used for MPC, the hybrid (semi-parametric) model structure can be used, i.e., a combinatorial structure of black box and white box models, incorporating both, process knowledge and process data [94], [95].

A second topical and prominent issue in the biopharmaceutical industry is the long development time of an upstream process, before the clinical studies. Budget restrictions and time pressure to speed up the return on investment are further arguments for process modelling to accelerate this upstream task [96]. Intensified design of experiments (iDoE) offer a novel approach to accelerate design space characterization and faster process development [97]. Both the major problems of current implementation of QbD concepts in upstream processes, i.e., accurate time-resolved predictive process modelling and the need to speed up process characterization, can be solved by the combinatorial use of hybrid models and iDoE.

1.7 Hybrid Modelling

The structure of a hybrid model benefits from the strengths of each single approach and eliminates the shortcomings of each modelling technique, the black box and white box. The hybrid model thus enables more advanced process modelling with higher complexity. These two parts can be in a serial or parallel sequence, i.e., the order in which the respective models are taken into account for the hybrid model [98]. The strengths of the two models cancel out each other’s weaknesses: the black box can quantify certain terms more precisely, which need to be assumed constant in the white box and the white box can extrapolate, whereas the black box alone would fail to extrapolate reasonably [99].

This incorporation of process data and process knowledge generally makes it an efficient and more cost-effective approach to deal with complex problems. Moreover, the development of a predictive model is possible, predicting the trend of a cultivation, i.e., for any given time, estimated values of the variables of interest are available [100]. Conclusions about the cultivation performance and the expected quality can be drawn in real-time, thereby modelling the process in the best way and meeting QbD requirements [101]. An exemplary presentation of a hybrid model using a serial structure is provided in **Fig. 2**. In this example, the black box is first used to estimate the, a priori to a process, highly valuable known unknowns, i.e., the specific rate expressions, using the chosen inputs to the black box. These are then transferred to the white box, using process knowledge to describe the process, at which point the estimated values for the rate expressions are inserted.

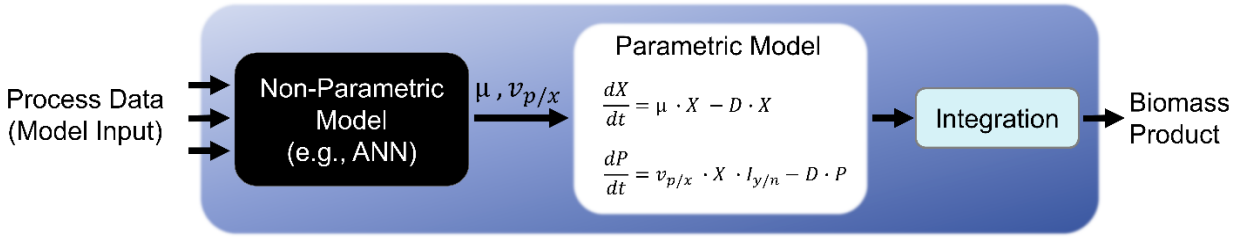

Figure 2. Setup of a hybrid model using a serial structure. The on-line process data are used as the input to the non-parametric (black box) part of the hybrid model to estimate the specific rates of the process variables to be modelled, e.g., the product titre and biomass. The estimated rates are used afterwards in the parametric (white box) by insertion into the respective terms. This results in a more accurate prediction of the specific rates and, by integrating, the respective concentrations are derived.

To develop a predictive model of the current processes and to establish new processes is of great value for all cultivation processes but especially in cultivations which are particularly complex, e.g., long-lasting mammalian processes associated with higher costs [102]. Further, the number of parameters needed can also be reduced by the development of a more robust and complex model that is capable of understanding and interpreting the data while maintaining or even enhancing the model's performance [98]. As a result, the behaviour and deviations of a process can be understood and it is therefore possible to understand the impact of changing CPPs during the process. Generally, with respect to the model performance, better and more accurate predictions are obtained using a hybrid model compared to purely mechanistic or data-driven models. The advantage of using time-resolved process models for DoE evaluation, regarding process understanding and the generation of knowledge is summarized in **Fig. 3**.

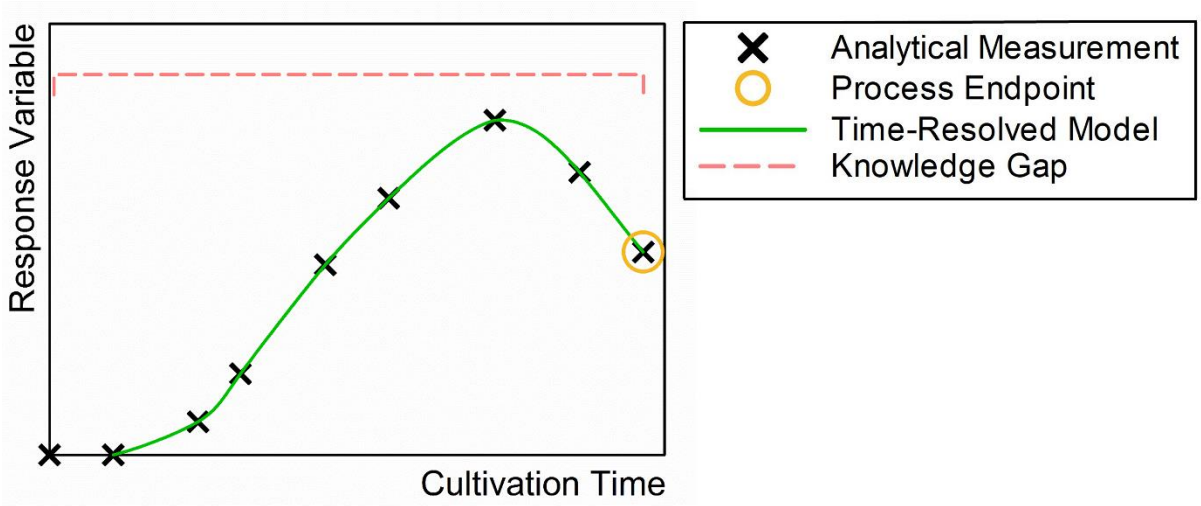

Figure 3. The capability of different model types to describe a bioprocess. By frequent sampling and measurement of a process response variable (black crosses), its trend can be investigated as a function of the cultivation time. Using response surface models, typically only the endpoints are described (yellow circle), potentially missing process optima and thereby leaving a so-called process knowledge gap behind (dashed red line). By using time-resolved process models, e.g., hybrid models, the complete trend of the response variable during the cultivation is described (solid green line).

Gathering this missing time-resolved process information is facilitated by hybrid models. However, limitations in this scope of the application still exist. Obstacles like the biological variance and analytical error cannot be eliminated. These also set the limits for the model performance itself, i.e., if the product titre can only be determined with an analytical error of about 10 %, the prediction of the model will not exceed this efficiency. With this setup, the full process can be monitored and described, with the result that the process variables of interest are always available, including the current value as well as the past and future predictions. Due to these advantages, hybrid modelling is gaining increasing popularity for bioprocess modelling. Nevertheless, this does not allow intervention in the process to change the outcome, if a parameter becomes unacceptable and the batch must be discarded. However, a predictive model provides the basis to approach the unsolved problem of a controllable hybrid model. For this purpose, the inputs to the model must be controllable and therefore real-time process adaptions, e.g., to bring the process back on track, avoid decreasing titres and keep the CQAs within the accepted range, are possible [103], [104]. In this way, the process is not only predictable but also controllable, as summarized in **Fig. 4**.

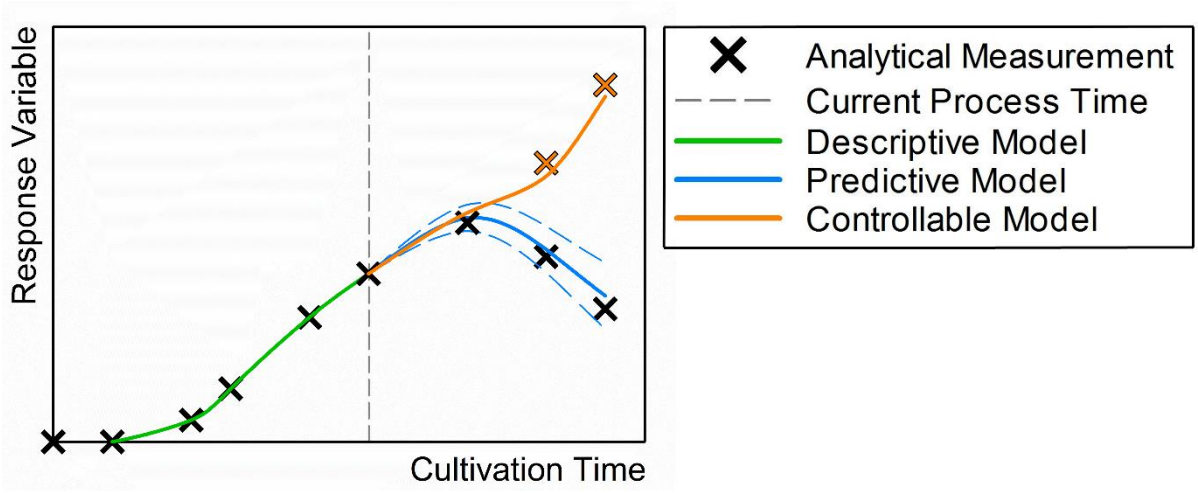

Figure 4. Characteristics of different time-resolved bioprocess models. The trend of a process response variable (black crosses), can be estimated from the start of the cultivation until the current process time, applying a descriptive model (solid green line). By using a predictive model, predictions for future values can be assumed with a certain uncertainty (solid blue line). By developing a predictive model, which can interfere with the process parameters, a controllable model is enabled (solid orange line). This leads to an alternative cultivation with different target responses (orange crosses).

However, although a hybrid model provides improved performance compared to other approaches, mispredictions are still possible. To access the chance of such a model uncertainty, bootstrapping can be applied, i.e., model averaging. Since bootstrapping allows full control over developing the final model, it has been proven a more flexible technique compared to cross-validation. For this approach, several models of each boot are selected and merged into one, receiving the information about the models' standard deviation (SD) and the prediction interval, which provides the probability of a misinterpretation and risk assessment [105]. The combination of both elements, hybrid modelling and bootstrapping, provides a robust and reliable hybrid model for bioprocess modelling.

Nevertheless, a negative aspect is the long development time before such a model is ready to be used, i.e., the experimental workload. The required process data has have to be generated, e.g., the design space must be characterized and the process optimized. This rapidly becomes a laborious and time-consuming task with the necessary related analytical effort. Moreover, the model must be trained, validated and tested. With an assumed overall duration of a week for microbial processes, including analytics and sterilization procedures and a typically chosen number of three CPPs each with three levels to be tested, the upstream process alone may take up to one year and much longer for mammalian cultures. However, the main advantage of using a hybrid model to predict the variables of interest during a process, is that the incorporation of process deviations and real-time dynamics is possible. Furthermore, this integration of process knowledge also allows CPPs to be changed during a process [106]. The concept of characterizing more than one CPP setpoint per fermentation by intentional

intra-experimental CPP setpoint shifts was derived from this possibility. As mentioned above, the so-called iDoE enables accelerated process characterization, optimization and significantly reduces the number of required experiments, therefore saving time in this operational unit.

1.8 Intensified Design of Experiments

To reduce the experimental workload by changing CPP setpoints during a cultivation, i.e., testing several settings within one intensified run, is a rather novel approach. It is alleged that by applying iDoE for process characterization, fewer experiments containing the same informational content may be sufficient for achieving the task, i.e., compared to a classical static design space, the total number of experiments can be significantly reduced. Additionally, it is claimed that due to the intra-experimental CPP setpoint changes, the reaction to these dynamic process changes can be captured [107]. By performing this iDoE, the history of the cell and a memory effect are often assumed to influence how the cells react to subsequent CPP setpoints [108]. Such an exemplary comparison of a two-dimensional design space characterized with either static or intensified cultivations is provided in **Fig. 5**, which highlights the temporal advantage of performing only three intensified – instead of nine static – experiments to cover the entire design space.

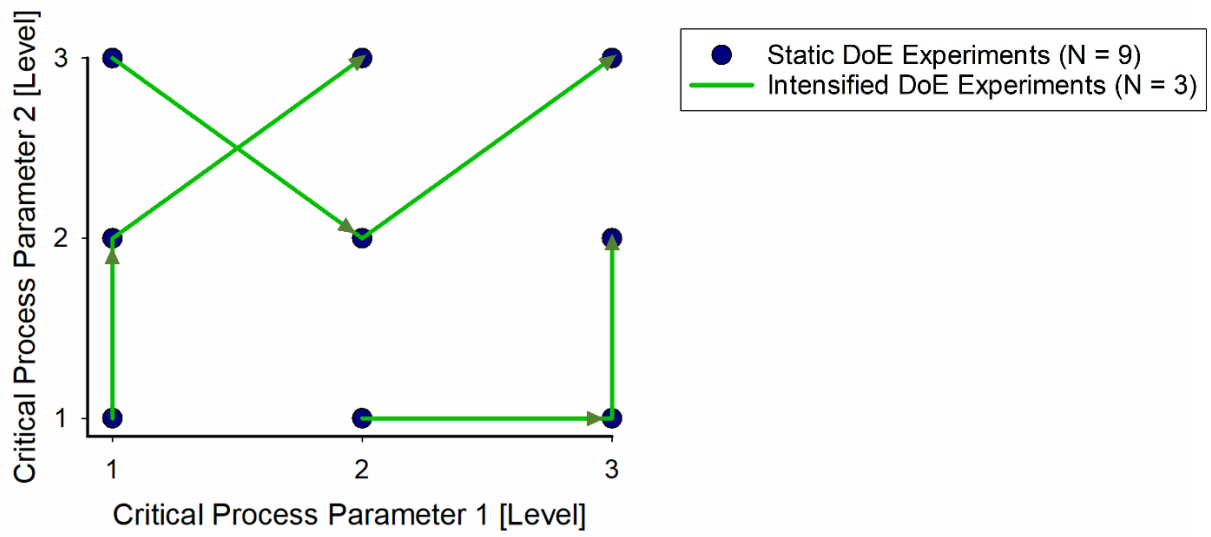

Figure 5. Design space characterized by static and intensified experiments. A two-dimensional design space, wherein each CPP is characterized at three levels and possesses nine CPP combination settings, which need to be investigated. By applying the static approach (dark blue dots) nine cultivations are necessary for a full factorial characterization. Using intensified experiments, i.e., intra-experimental CPP setpoint shifts, enables the characterization of three CPP combination settings per experiment, resulting in three intensified experiments to fully characterize the same design space (solid green lines).

The saved number of total experiments is not only reflected in the saved time but also in the reduced costs and resources due to fewer experiments. Moreover, it is assumed that the reduced amount of data for training the hybrid model may still maintain a high performance due to the knowledge generated about the process dynamics [109]. The increased process dynamics would remain unnoticed or would be modelled incorrectly if a state-of-the-art RSM or a single black or white box model were used. However, it is assumed that a hybrid model can deal with iDoE, describing the real-time process dynamics. This emphasizes the combinatorial use of intensified fermentations and hybrid modelling from economic and performance-related aspects, for instance: to meet the FDA's QbD and PAT requirements for quality assurance, to generate process knowledge, while simultaneously speeding up the time to market for newly developed drugs and biosimilars, by providing a faster supply of materials for clinical studies [110].

2 Objectives

This doctoral thesis aimed to develop and apply new methods for the implementation of QbD concepts in upstream processing, thereby eliminating present limitations. For this purpose, some of the most important issues addressed and investigated. This includes the

- precise calculation of the specific rates of a bioprocess
- development of soft sensors for process variables of interest
- importance of hybrid modelling for process characterization
- concept of iDoE for accelerating upstream process characterization

Required data were obtained by performing *E. coli* fed-batch fermentations (20 L) in a three-dimensional design space, investigating each parameter at three levels ($N = 3^3$). This design space was characterized completely twice, with static and intensified fed-batch fermentations. The developed methods were tested for their respective performance and compared to state-of-the-art techniques.

To emphasize QbD, the following hypotheses were proposed as the main objectives:

- A smoothing spline function enables access to the specific rates of a bioprocess despite high analytical errors (Publication I)
- Using 2D-fluorescence spectroscopy and standard process data, real-time on-line monitoring of the biomass concentration is possible (Publication II)
- Hybrid modelling is superior to state-of-the-art modelling techniques for upstream process characterization (Publication III)
- Process characterization can be significantly accelerated using intensified cultivations (Publication IV)

3 Results

3.1 Publication I – The Shortcomings of Accurate Rate Estimations in Cultivation Processes and a Solution for Precise and Robust Process Modeling

Hypothesis: *a smoothing spline function enables access to the specific rates of a bioprocess despite high analytical errors*

In this publication, a modelling technique, cubic smoothing splines, for calculating the specific growth rate and the specific substrate consumption rate is presented and compared to state-of-the-art methods. This is of high importance since a robust method to accurately estimate the cell growth or substrate consumption rate, unveiling the actual status of the cells, is essential to understand a bioprocess, e.g., batch-to-batch comparability and post-process investigation. In comparison, the presented technique has been proven to possess higher accuracy and higher precision than standard methods such as stepwise integration, which is not suitable for the calculation of non-linear trends. Additionally, high analytical errors further worsen a precise calculation. However, the smoothing spline is considered to be highly robust since the sampling frequency and high analytical errors do not significantly influence its preciseness. All these features make it a valuable tool to compare and investigate batches in the QbD concept.

3.2 Publication II – Soft Sensor Based on 2D-fluorescence and Process Data Enabling Real-time Estimation of Biomass in *Escherichia Coli* Cultivations

Hypothesis*: using 2D-fluorescence spectroscopy and standard process data, real-time on-line monitoring of the biomass concentration is possible*

To develop a soft sensor able to estimate the current biomass concentration in real-time, standard process data and a 2D-fluorescence probe were used. The established soft sensor can accurately estimate the biomass concentrations on-line for all design space settings. To test the generalization ability, fermentations with different settings, which mimicked process deviations, were performed and the performance of the soft sensor was assessed for all runs. It was demonstrated that the soft sensor was also able to estimate the biomass concentration of these fermentations. However, there were performance limitations with too many CPP deviations. Nevertheless, the developed soft sensor has been proven to be well-suited for on-line process monitoring of the biomass concentration in production processes.

3.3 Publication III – Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization

Hypothesis: *hybrid modelling is superior to state-of-the-art modelling techniques for upstream process characterization*

To prove the superiority of hybrid models over state-of-the-art modelling techniques, i.e., RSMs of the process endpoints and a time-resolved black box model, an extensive comparison was presented. A complete full factorial design space was characterized and the performance predicting the analytical values of the fermentations was assessed. For this evaluation, either the process endpoints or the complete trend of the biomass concentration and the soluble product titre were considered. This comparison demonstrated that the performance of a hybrid model is superior to that of state-of-the-art techniques to ensure constant product quality. Moreover, hybrid modelling was highlighted to be the method of choice for reliable time-resolved process modelling.

3.4 Publication IV – Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization

Hypothesis: *process characterization can be significantly accelerated using intensified cultivations*

To demonstrate potential time savings during process characterization by performing intensified cultivations, the same design space as in the previous publication (Publication III) was characterized using iDoE, i.e., cultivations with intra-experimental CPP setpoint shifts. An iDoE hybrid model based on only nine intensified cultivations was developed and its performance compared to the full factorial static hybrid model (27 cultivations) and a fractional factorial static hybrid model, trained only on the centre point and the corners (nine cultivations). It was shown that the full factorial static hybrid model had the most reliable performance. The iDoE hybrid model performed overall similar but with a slightly increased prediction error. However, only a third of the data was used to develop the hybrid model, compared to the full factorial static hybrid model, i.e., >66 % fewer practical experiments. Further, the presence of a possible memory effect caused by the CPP shifts, i.e., an altered behaviour due to previous CPP combination settings, which would make using the iDoE approach impracticable, was examined and rejected. The fractional factorial static hybrid model showed the highest error and SD compared to the other two hybrid models. It was thus shown that it is possible to significantly reduce bioprocess development times by combined application of iDoE and hybrid modelling.

4 Conclusions

All hypotheses were confirmed and the objectives accomplished as presented in the publications and this thesis. With the obtained results and findings, some of the current research gaps were closed, contributing to progress in the PAT initiative and to the implementation of QbD concepts in upstream processing.

The presented cubic smoothing spline solution to precisely estimate the specific rates of the cells in a bioprocess is of high value concerning batch-to-batch comparability and post-process investigation, i.e., process deviations are visible, can be examined and understood. Moreover, this method allowed the reasonable sampling frequency to be determined with respect to the accuracy of the off-line method. Furthermore, the estimative quality of state-of-the-art methods, e.g., stepwise linear integration, is highly susceptible to the frequency of the sampling interval, while the accuracy and preciseness of the cubic smoothing spline has been proven to be independent of the sampling interval to a great extent. Furthermore, it also demonstrates robustness towards high analytical measurement error, as is often the case in mammalian systems, e.g., the determination of the cell concentration.

Moreover, extracting real-time information from a bioprocess to estimate process variables, which are typically not accessible in real-time, e.g., the current biomass concentration, is of high interest. Using an advanced 2D-fluorescence sensor, in addition to the standard on-line process data, the development of a solely data-driven soft sensor was possible. Further, by altering process parameters, i.e., mimicking process deviations, the limitations of accurate estimations for this non-parametric model were tested. In this way it was shown that the soft sensor performs reliably, even when process parameters are changed. This makes it a valuable tool for production processes, providing information about the current biomass concentration at any time in the process. Moreover, with this established soft sensor, unit operations can always be carried out at the same setpoint, e.g., the induction can always take place at the same biomass concentration instead of a predetermined time point, to ensure process consistency.

To overcome the limitations of pure non-parametric and descriptive models, as presented by the soft-sensor approach, a predictive hybrid model was used to predict the process variables of interest, the biomass concentration and the soluble product titre. Therefore, the hybrid model was only developed on process parameters that can be controlled.

The performance of hybrid modelling was evaluated compared to state-of-the-art methods for process characterization, e.g., response surface modelling of the process endpoints and time-resolved non-parametric modelling. It has been demonstrated that a hybrid model outperforms a pure data-driven model and has almost equivalent performance to RSMs, predicting the process endpoints, with the additional advantage of providing the time-resolved

trajectory of the complete bioprocess, thus contributing to highly demanded process understanding. The established hybrid model was applied as a soft sensor and its performance was evaluated for predicting both process variables of interest in real-time, applied on new fermentations. Although the applied settings were completely new, the hybrid model was able to accurately predict the biomass and soluble product titre in real-time. Moreover, all the inputs used to train the hybrid model were easily and directly controllable, enabling MPC in future applications.

Finally, the issue of long development times and high experimental effort for process characterization was tackled using iDoE. Intra-experimental shifts of CPPs enabled the characterization of more than one CPP setpoint per fermentation, as is commonly performed. This means the exemplary three-dimensional design space was completely characterized by only nine iDoE cultivations, comprising two parameter shifts per fermentation, instead of 27 static cultivations. Regarding a possible memory effect, it was shown that the cells can handle this applied iDoE setup by rapidly adapting to new process conditions, already within one hour after the shift. This makes a comparison of both approaches valid and permits the usability of the intensified fed-batch fermentations to predict the outcome of the static fed-batch fermentations. Performance of the iDoE hybrid model was slightly inferior to the full factorial static hybrid model, but still highly accurate. A fractional factorial static hybrid model, developed on nine static fed-batch fermentations proved to be less accurate and less precise compared to the iDoE hybrid model, confirming the higher learning from intensified cultivations. Moreover, the iDoE hybrid model was able to predict the biomass concentration and soluble product titre of all static cultivations, leading to two thirds fewer practical experiments, i.e., a reduced time requirement by >66 % for process characterization. This further highlights the potential and advantages of applying iDoE.

5 List of Figures

6 List of Abbreviations

APC	advanced process control
CHO	Chinese hamster ovary
CPP	critical process parameters
CQA	critical quality attributes
DoE	design of experiments
E. coli	*Escherichia coli*
FDA	U.S. Food and Drug Administration
iDoE	intensified design of experiments
MPC	model predictive control
MVDA	multivariate data analysis
PARAFAC	parallel factor analysis
PAT	process analytical technology
PCA	principal component analysis
QbD	quality by design
QbT	quality by testing
RSM	response surface model
SD	standard deviation

7 References

[1] G. Walsh, “Biopharmaceutical benchmarks 2014,” *Nat. Biotechnol.*, **2014**, *32*, 992–1000.

[2] F. M. Wurm, “Production of recombinant protein therapeutics in cultivated mammalian cells,” *Nat. Biotechnol.*, **2004**, *22*, 1393–1398.

[3] G. L. Rosano and E. A. Ceccarelli, “Recombinant protein expression in *Escherichia coli*: advances and challenges,” *Front. Microbiol.*, **2014**, *5*, 1–17.

[4] S. Gnoth, M. Jenzsch, R. Simutis, and A. Lübbert, “Control of Cultivation Processes for Recombinant Protein Production: A Review,” *Bioprocess Biosyst. Eng.*, **2008**, *31*, 21–39.

[5] Y. Rouiller, A. Périlleux, N. Collet, M. Jordan, M. Stettler, and H. Broly, “A High-Throughput Media Design Approach for High Performance Mammalian Fed-Batch Cultures,” *MAbs*, **2013**, *5*, 501–511.

[6] Y.-M. Huang, W. Hu, E. Rustandi, K. Chang, H. Yusuf-Makagiansar, and T. Ryll, “Maximizing Productivity of CHO Cell-Based Fed-Batch Culture Using Chemically Defined Media Conditions and Typical Manufacturing Equipment,” *Biotechnol. Prog.*, **2010**, *26*, 1400–1410.

[7] J. Barrios-González, T. E. Castillo, and A. Mejía, “Development of high penicillin producing strains for solid state fermentation,” *Biotechnol. Adv.*, **1993**, *11*, 525–537.

[8] P. M. O’Callaghan *et al.*, “Cell Line-Specific Control of Recombinant Monoclonal Antibody Production by CHO Cells,” *Biotechnol. Bioeng.*, **2010**, *106*, 938–951.

[9] A. Mader, B. Prewein, K. Zboray, E. Casanova, and R. Kunert, “Exploration of BAC versus plasmid expression vectors in recombinant CHO cells,” *Appl. Microbiol. Biotechnol.*, **2013**, *97*, 4049–4054.

[10] T. W. Jeffries and Y. S. Jin, “Metabolic engineering for improved fermentation of pentoses by yeasts,” *Appl. Microbiol. Biotechnol.*, **2004**, *63*, 495–509.

[11] K. Byoungjin, P. Hyegwon, N. Dokyun, and L. Sang Yup, “Metabolic Engineering of *Escherichia Coli* for the Production of Phenol From Glucose,” *Biotechnol. J.*, **2014**, *9*, 621–629.

[12] L. Feeney *et al.*, “Eliminating Tyrosine Sequence Variants in CHO Cell Lines Producing Recombinant Monoclonal Antibodies,” *Biotechnol. Bioeng.*, **2013**, *110*, 1087–1097.

[13] V. M. DeZengotita, W. M. Miller, J. G. Aunins, and W. Zhou, “Phosphate Feeding Improves High-Cell-Concentration NS0 Myeloma Culture Performance for Monoclonal Antibody Production,” *Biotechnol. Bioeng.*, **2000**, *69*, 566–576.

[14] V. Singh, S. Haque, R. Niwas, A. Srivastava, M. Pasupuleti, and C. K. M. Tripathi, “Strategies for Fermentation Medium Optimization: An In-Depth Review,” *Front. Microbiol.*, **2017**, *7*, 2087.

[15] M. Luchner, G. Striedner, M. Cserjan-Puschmann, F. Strobl, and K. Bayer, “Online prediction of product titer and solubility of recombinant proteins in *Escherichia coli* fed-batch cultivations,” *J. Chem. Technol. Biotechnol.*, **2015**, *90*, 283–290.

[16] U.S. Food and Drug Administration, "FDA Report," *Final report of pharmaceutical cGMPs for the 21st Century–a risk-based approach*, U.S. Food and Drug Administration: Silver Spring, MD, **2004**.

[17] E. Ohage, R. Iverson, L. Krummen, R. Taticek, and M. Vega, “QbD implementation and Post Approval Lifecycle Management (PALM),” *Biologicals*, **2016**, *44*, 332–340.

[18] B. Kelley, M. Cromwell, and J. Jerkins, “Integration of QbD risk assessment tools and overall risk management,” *Biologicals*, **2016**, *44*, 341–351.

[19] S. M. Mercier, B. Diepenbroek, M. C. F. Dalm, R. H. Wijffels, and M. Streefland, “Multivariate Data Analysis as a PAT Tool for Early Bioprocess Development Data,” *J. Biotechnol.*, **2013**, *167*, 262–270.

[20] V. Kumar, A. Bhalla, and A. S. Rathore, "Design of experiments applications in bioprocessing: concepts and approach,” *Biotechnol. Prog.*, **2014**, *30*, 86–99.

[21] V. Mishra, S. Thakur, A. Patil, and A. Shukla, “Quality by Design (QbD) Approaches in Current Pharmaceutical Set-Up,” *Expert Opin. Drug Deliv.*, **2018**, *15*, 737–758.

[22] P. D. Bade, S. P. Kotu, and A. S. Rathore, “Optimization of a refolding step for a therapeutic fusion protein in the quality by design (QbD) paradigm,” *J. Sep. Sci.*, **2012**, *35*, 3160–3169.

[23] E. Trummer *et al.*, “Process Parameter Shifting: Part I. Effect of DOT, pH, and Temperature on the Performance of Epo-Fc Expressing CHO Cells Cultivated in Controlled Batch Bioreactors,” *Biotechnol. Bioeng.*, **2006**, *94*, 1033–1044.

[24] E. Trummer *et al.*, “Process Parameter Shifting: Part II. Biphasic Cultivation—A Tool for Enhancing the Volumetric Productivity of Batch Processes Using Epo-Fc Expressing CHO Cells,” *Biotechnol. Bioeng.*, **2006**, *94*, 1045–1052.

[25] A. S. Rathore and H. Winkle, “Quality by design for biopharmaceuticals,” *Nat. Biotechnol.*, **2009**, *27*, 26–34.

[26] A. Pekarsky, V. Konopek, and O. Spadiut, “The impact of technical failures during cultivation of an inclusion body process,” *Bioprocess Biosyst. Eng.*, **2019**, *42*, 1611–1624.

[27] M. Streefland, D. E. Martens, E. C. Beuvery, and R. H. Wijffels, “Process analytical

technology (PAT) tools for the cultivation step in biopharmaceutical production," *Eng. Life Sci.*, **2013**, *13*, 212–223.

[28] M. Melcher *et al.*, "The potential of random forest and neural networks for biomass and recombinant protein modeling in *Escherichia coli* fed-batch fermentations," *Biotechnol. J.*, **2015**, *10*, 1770–1782.

[29] J. Rantanen and J. Khinast, "The Future of Pharmaceutical Manufacturing Sciences," *J. Pharm. Sci.*, **2015**, *104*, 3612–3638.

[30] P. Biechele, C. Busse, D. Solle, T. Scheper, and K. Reardon, "Sensor systems for bioprocess monitoring," *Eng. Life Sci*., **2015**, *15*. 469–488.

[31] A. P. Teixeira, R. Oliveira, P. M. Alves, and M. J. T. Carrondo, "Advances in On-Line Monitoring and Control of Mammalian Cell Cultures: Supporting the PAT Initiative," *Biotechnol. Adv.*, **2009**, *27*, 726–732.

[32] M. Känsäkoski *et al.*, "Process analytical technology (PAT) needs and applications in the bioprocess industry: Review," *Espoo VTT Tech. Res. Cent. Finland. VTT Work. Pap.*, **2006**, *60*.

[33] M. Pohlscheidt, S. Charaniya, C. Bork, M. Jenzsch, T. L. Noetzel, and A. Luebbert, "Bioprocess and fermentation monitoring," *Encycl. Ind. Biotechnol.*, **2013**, 1469–1492.

[34] J. Glassey *et al.*, "Process analytical technology (PAT) for biopharmaceuticals," *Biotechnol. J.*, **2011**, *6*, 369–377.

[35] M. Luchner *et al.*, "Implementation of proton transfer reaction-mass spectrometry (PTR-MS) for advanced bioprocess monitoring," *Biotechnol. Bioeng.*, **2012**, *109*, 3059–3069.

[36] S. Ansorge, G. Esteban, and G. Schmid, "On-line monitoring of responses to nutrient feed additions by multi-frequency permittivity measurements in fed-batch cultivations of CHO cells," *Cytotechnology*, **2010**, *62*, 121–132.

[37] S. Ansorge, G. Esteban, and G. Schmid, "Multifrequency permittivity measurements enable on-line monitoring of changes in intracellular conductivity due to nutrient limitations during batch cultivations of CHO cells," *Biotechnol. Prog.*, **2010**, *26*, 272–283.

[38] R. Ulber, J.-G. Frerichs, and S. Beutel, "Optical sensor systems for bioprocess monitoring," *Anal. Bioanal. Chem.*, **2003**, *376*, 342–348.

[39] E. Hirsch *et al.*, "Inline noninvasive Raman monitoring and feedback control of glucose concentration during ethanol fermentation," *Biotechnol. Prog.*, **2019**, *35*, 2848–2855.

[40] H. L. T. Lee, P. Boccazzi, N. Gorret, R. J. Ram, and A. J. Sinskey, "In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy," *Vib. Spectrosc.*, **2004**, *35*, 131–137.

[41] M. Sandor, F. Rüdinger, R. Bienert, C. Grimm, D. Solle, and T. Scheper, “Comparative study of non-invasive monitoring via infrared spectroscopy for mammalian cell cultivations,” *J. Biotechnol.*, **2013**, *168*, 636–645.

[42] A. E. Cervera, N. Petersen, A. E. Lantz, A. Larsen, and K. V. Gernaey, “Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation,” *Biotechnol. Prog.*, **2009**, *25*, 1561–1581.

[43] K. Hantelmann, M. Kollecker, D. Hüll, B. Hitzmann, and T. Scheper, “Two-dimensional fluorescence spectroscopy: a novel approach for controlling fed-batch cultivations,” *J. Biotechnol.*, **2006**, *121*, 410–417.

[44] S. M. Faassen and B. Hitzmann, “Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring,” *Sensors (Switzerland)*, **2015**, *15*, 10271–10291.

[45] S. Assawajaruwan, P. Eckard, and B. Hitzmann, “On-line monitoring of relevant fluorophores of yeast cultivations due to glucose addition during the diauxic growth,” *Process Biochem.*, **2017**, *58*, 51–59.

[46] A. Hagedorn, R. L. Legge, and H. Budman, “Evaluation of Spectrofluorometry as a Tool for Estimation in Fed-Batch Fermentations,” *Biotechnol. Bioeng.*, **2003**, *83*, 104–111.

[47] J. S. Guez, J. P. Cassar, F. Wartelle, P. Dhulster, and H. Suhr, “Real time in situ microscopy for animal cell-concentration monitoring during high density culture in bioreactor,” *J. Biotechnol.*, **2004**, *111*, 335–343.

[48] K. Joeris, J.-G. Frerichs, K. Konstantinov, and T. Scheper, “*In-situ* microscopy: Online process monitoring of mammalian cell cultures,” *Cytotechnology*, **2002**, *38*, 129–134.

[49] R. C. Rowland-Jones, F. van den Berg, A. J. Racher, E. B. Martin, and C. Jaques, “Comparison of spectroscopy technologies for improved monitoring of cell culture processes in miniature bioreactors,” *Biotechnol. Prog.*, **2017**, *33*, 337–346.

[50] L. Zhang and S. Mao, “Application of quality by design in the current drug development,” *Asian J. Pharm. Sci.*, **2017**, *12*, 1–8.

[51] A. S. Patil and A. M. Pethe, “Quality by design (QbD): A new concept for development of quality pharmaceuticals,” *Pharm. Res.*, **2008**, *4*, 781–791.

[52] F. Torkashvand *et al.*, “Designed amino acid feed in improvement of production and quality targets of a therapeutic monoclonal antibody,” *PLoS One*, **2015**, *10*, 1–21.

[53] J. Ramírez, H. Gutierrez, and A. Gschaedler, “Optimization of astaxanthin production by *Phaffia rhodozyma* through factorial design and response surface methodology,” *J. Biotechnol.*, **2001**, *88*, 259–268.

[54] J. Möller, K. B. Kuchemüller, T. Steinmetz, K. S. Koopmann, and R. Pörtner, “Model-assisted Design of Experiments as a concept for knowledge-based bioprocess

development," *Bioprocess Biosyst. Eng.*, **2019**, *42*, 867–882.

[55] S. J. Kalil, F. Maugeri, and M. I. Rodrigues, "Response surface analysis and simulation as a tool for bioprocess design and optimization," *Process Biochem.*, **2000**, *35*, 539–550.

[56] R. Balusu, R. R. Paduru, S. K. Kuravi, G. Seenayya, and G. Reddy, "Optimization of critical medium components using response surface methodology for ethanol production from cellulosic biomass by *Clostridium thermocellum* SS19," *Process Biochem.*, **2005**, *40*, 3025–3030.

[57] G. Q. Liu and X.-L. Wang, "Optimization of critical medium components using response surface methodology for biomass and extracellular polysaccharide production by *Agaricus blazei*," *Appl. Microbiol. Biotechnol.*, **2007**, *74*, 78–83.

[58] M. S. Tanyildizi, D. Özer, and M. Elibol, "Optimization of α-amylase production by *Bacillus* sp. using response surface methodology," *Process Biochem.*, **2005**, *40*, 2291–2296.

[59] T. Lundstedt *et al.*, "Experimental design and optimization," *Chemom. Intell. Lab. Syst.*, **1998**, *42*, 3–40.

[60] L. Zhao, H.-Y. Fu, Z. Weichang, and H. Wei-Shou, "Advances in process monitoring tools for cell culture bioprocesses," *Eng. Appl. Artif. Intell.*, **2015**, *15*, 459–468.

[61] J. Djuris and Z. Djuric, "Modeling in the quality by design environment: Regulatory requirements and recommendations for design space and control strategy appointment," *Int. J. Pharm.*, **2017**, *533*, 346–356.

[62] K.-M. Lee and D. F. Gilmore, "Statistical Experimental Design for Bioprocess Modeling and Optimization Analysis," *Appl. Biochem. Biotechnol.*, **2006**, *135*, 101–135.

[63] L. H. Chiang, R. Leardi, R. J. Pell, and M. B. Seasholtz, "Industrial experiences with multivariate statistical analysis of batch process data," *Chemom. Intell. Lab. Syst.*, **2006**, *81*, 109–119.

[64] D. Vier, S. Wambach, V. Schünemann, and K.-U. Gollmer, "Multivariate Curve Resolution and Carbon Balance Constraint to Unravel FTIR Spectra from Fed-Batch Fermentation Samples," *Bioengineering*, **2017**, *4*, 9–25.

[65] A. Chéruy, "Software sensors in bioprocess engineering," *J. Biotechnol.*, **1997**, *52*, 193–199.

[66] J. Shlens, "A tutorial on principal component analysis," *Int. J. Remote Sens.*, **2014**, *51*, 1–13.

[67] R. A. Harshman and M. E. Lundy, "PARAFAC: Parallel factor analysis," *Comput. Stat. Data Anal.*, **1994**, *18*, 39–72.

[68] R. Bro, “PARAFAC. Tutorial and applications,” *Chemom. Intell. Lab. Syst.*, **1997**, *38*, 149–171.

[69] D. M. Hallow *et al.*, “An example of utilizing mechanistic and empirical modeling in quality by design,” *J. Pharm. Innov.*, **2010**, *5*, 193–203.

[70] A. S. Rathore, “QbD/PAT for bioprocessing: Moving from theory to implementation,” *Curr. Opin. Chem. Eng.*, **2014**, *6*, 1–8.

[71] P. Wechselberger, P. Sagmeister, and C. Herwig, “Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes,” *Bioprocess Biosyst. Eng.*, **2013**, *36*, 1205–1218.

[72] H. Narayanan, M. Sokolov, M. Morbidelli, and A. Butté, “A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins,” *Biotechnol. Bioeng.*, **2019**, *116*, 2540–2549.

[73] K. Ohadi, H. Aghamohseni, R. L. Legge, and H. M. Budman, “Fluorescence-based soft sensor for at situ monitoring of Chinese hamster ovary cell cultures,” *Biotechnol. Bioeng.*, **2014**, *111*, 1577–1586.

[74] T. Schmidberger, C. Posch, A. Sasse, C. Guelch, and R. Huber, “Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling,” *Biotechnol. Prog.*, **2015**, *31*, 1119–1127.

[75] A. S. Rathore, S. Mittal, M. Pathak, and V. Mahalingam, “Chemometrics application in biotech processes: Assessing comparability across processes and scales,” *J. Chem. Technol. Biotechnol.*, **2014**, *89*, 1311–1316.

[76] K. Ohadi, R. L. Legge, and H. M. Budman, “Development of a Soft-Sensor Based on Multi-Wavelength Fluorescence Spectroscopy and a Dynamic Metabolic Model for Monitoring Mammalian Cell Cultures,” *Biotechnol. Bioeng.*, **2014**, *112*, 197–208.

[77] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, “A systematic approach for soft sensor development,” *Comput. Chem. Eng.*, **2007**, *31*, 419–425.

[78] C.-F. Mandenius and R. Gustavsson, “Mini-review: Soft sensors as means for PAT in the manufacture of bio-therapeutics,” *J. Chem. Technol. Biotechnol.*, **2015**, *90*, 215–227.

[79] P. Kadlec, B. Gabrys, and S. Strandt, “Data-driven Soft Sensors in the process industry,” *Comput. Chem. Eng.*, **2009**, *33*, 795–814.

[80] J. Claßen, F. Aupert, K. F. Reardon, D. Solle, and T. Scheper, “Spectroscopic sensors for in-line bioprocess monitoring in research and pharmaceutical industrial application,” *Anal. Bioanal. Chem.*, **2017**, *409*, 651–666.

[81] J. -K Li and A. E. Humphrey, “Use of fluorometry for monitoring and control of a

bioreactor," *Biotechnol. Bioeng.*, **1991**, *37*, 1043–1049.

[82] L. McIver *et al.*, "Characterisation of flavodoxin NADP+ oxidoreductase and flavodoxin; key components of electron transfer in *Escherichia coli*," *Eur. J. Biochem.*, **1998**, *257*, 577–585.

[83] A. Mukherjee, J. Walker, K. B. Weyant, and C. M. Schroeder, "Characterization of Flavin-Based Fluorescent Proteins: An Emerging Class of Fluorescent Reporters," *PLoS One*, **2013**, *8*, 1–15.

[84] M. J. McAnulty and T. K. Wood, "YeeO from *Escherichia coli* exports flavins," *Bioengineered*, **2014**, *5*, 386–392.

[85] E. Skibsted, C. Lindemann, C. Roca, and L. Olsson, "On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration," *J. Biotechnol.*, **2001**, *88*, 47–57.

[86] S. André, L. Saint Cristau, S. Gaillard, O. Devos, É. Calvosa, and L. Duponchel, "In-line and real-time prediction of recombinant antibody titer by *in situ* Raman spectroscopy," *Anal. Chim. Acta*, **2015**, *892*, 148–152.

[87] A. Zhang *et al.*, "Advanced process monitoring and feedback control to enhance cell culture process production and robustness," *Biotechnol. Bioeng.*, **2015**, *112*, 2495–2504.

[88] J. Randek and C.-F. Mandenius, "On-line soft sensing in upstream bioprocessing," *Crit. Rev. Biotechnol.*, **2017**, *38*, 106–121.

[89] S. Craven, J. Whelan, and B. Glennon, "Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller," *J. Process Control*, **2014**, *24*, 344–357.

[90] U.S. Food and Drug Administration, "Guidance for Industry: Pharmaceutical Quality System (Q10)," *Int. Conf. Harmon.*, **2008**.

[91] B. Sonnleitner, "Instrumentation of biotechnological processes.," *Adv. Biochem. Eng. Biotechnol.*, **2000**, *66*, 1–64.

[92] M. S. Hong, K. A. Severson, M. Jiang, A. E. Lu, J. C. Love, and R. D. Braatz, "Challenges and opportunities in biopharmaceutical manufacturing control," *Comput. Chem. Eng.*, **2018**, *110*, 106–114.

[93] J. F. Kepert *et al.*, "Establishing a control system using QbD principles," *Biologicals*, **2016**, *44*, 319–331.

[94] J. Schubert, R. Simutis, M. Dors, I. Havlik, and A. Lubbert, "Bioprocess optimization and control - application of hybrid modeling," *J. Biotechnol.*, **1994**, *35*, 51–68.

[95] D. Psichogios and L. Ungar, "A hybrid neural network-first principles approach to

process modeling," *AIChE J.*, **1992**, *38*, 1499–1511.

[96] C. Varsakelis, S. Dessoy, M. von Stosch, and A. Pysik, "Show Me the Money! Process Modeling in Pharma from the Investor's Point of View," *Processes*, **2019**, *7*, 596.

[97] M. von Stosch and M. J. Willis, "Intensified design of experiments for upstream bioreactors," *Eng. Life Sci.*, **2016**, *17*, 1173–1184.

[98] M. von Stosch, R. Oliveira, J. Peres, and S. Feyo de Azevedo, "Hybrid semi-parametric modeling in process systems engineering: Past, present and future," *Comput. Chem. Eng.*, **2014**, *60*, 86–101.

[99] M. von Stosch *et al.*, "Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry," *Biotechnol. J.*, **2014**, *9*, 719–726.

[100] R. Simutis and A. Lübbert, "Hybrid Approach to State Estimation for Bioprocess Control," *Bioengineering*, **2017**, *4*, 21.

[101] M. von Stosch, R. Oliveria, J. Peres, and S. F. De Azevedo, "Hybrid modeling framework for process analytical technology: Application to *Bordetella pertussis* cultures," *Biotechnol. Prog.*, **2012**, *28*, 284–291.

[102] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: New estimates of drug development costs," *J. Health Econ.*, **2003**, *22*, 151–185.

[103] C. Komives and R. S. Parker, "Bioreactor state estimation and control," *Curr. Opin. Biotechnol.*, **2003**, *14*, 468–474.

[104] J. H. Lee, "Model predictive control: Review of the three decades of development," *Int. J. Control. Autom. Syst.*, **2011**, *9*, 415–424.

[105] J. Pinto, C. R. de Azevedo, R. Oliveira, and M. von Stosch, "A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development," *Bioprocess Biosyst. Eng.*, **2019**, *42*, 1853–1865.

[106] M. von Stosch, J.-M. Hamelink, and R. Oliveira, "Hybrid modeling as a QbD/PAT tool in process development: an industrial *E. coli* case study," *Bioprocess Biosyst. Eng.*, **2016**, *39*, 773–784.

[107] M. von Stosch, J. M. Hamelink, and R. Oliveira, "Toward intensifying design of experiments in upstream bioprocess development: An industrial *Escherichia coli* feasibility study," *Biotechnol. Prog.*, **2016**, *32*, 1343–1352.

[108] T. Patarinska, D. Dochain, S. N. Agathos, and L. Ganovski, "Modelling of continuous microbial cultivation taking into account the memory effects," *Bioprocess Eng.*, **2000**, *22*, 517–527.

[109] O. Spadiut, S. Rittmann, C. Dietzsch, and C. Herwig, “Dynamic process conditions in bioprocess development,” *Eng. Life Sci.*, **2013**, *13*, 88–101.

[110] M. Duerkop, M. von Stosch, M. Mayer, and G. Striedner, “Beyond static process parameter generation and data analysis: An intensified Design of Experiment and hybrid modeling workflow that lends itself to model predictive control,” *Am. Pharm. Rev.*, **2018**, *21*, 1–9.

8 Publications

I B. Bayer, B. Sissolak, M. Duerkop, M. von Stosch, and G. Striedner, “The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling,” *Bioprocess Biosyst. Eng.,* **2020**, *43*, 169–178.

DOI: 10.1007/s00449-019-02214-6

II B. Bayer, M. von Stosch, M. Melcher, M. Duerkop, and G. Striedner, “Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in *Escherichia coli* cultivations,” *Eng. Life Sci*., **2020**, *20*, 26–35.

DOI: 10.1002/elsc.201900076

III B. Bayer, M. von Stosch, G. Striedner, and M. Duerkop, “Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization,” *Biotechnol. J.,* **2020**, *15*, 1900551.

DOI: 10.1002/biot.201900551

IV B. Bayer, G. Striedner, and M. Duerkop, “Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization,” *Biotechnol. J.,* **2020**, 2000121.

DOI: 10.1002/biot.202000121

Publication I

Bioprocess and Biosystems Engineering (2020) 43:169–178
https://doi.org/10.1007/s00449-019-02214-6

RESEARCH PAPER

The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling

B. Bayer[1] · B. Sissolak[2] · M. Duerkop[1] · M. von Stosch[3] · G. Striedner[1]

Received: 29 March 2019 / Revised: 21 June 2019 / Accepted: 10 September 2019 / Published online: 20 September 2019

Abstract

The accurate estimation of cell growth or the substrate consumption rate is crucial for the understanding of the current state of a bioprocess. Rates unveil the actual cell status, making them valuable for quality-by-design concepts. However, in bioprocesses, the real rates are commonly not accessible due to analytical errors. We simulated *Escherichia coli* fed-batch fermentations, sampled at four different intervals and added five levels of noise to mimic analytical inaccuracy. We computed stepwise integral estimations with and without using moving average estimations, and smoothing spline interpolations to compare the accuracy and precision of each method to calculate the rates. We demonstrate that stepwise integration results in low accuracy and precision, especially at higher sampling frequencies. Contrary, a simple smoothing spline function displayed both the highest accuracy and precision regardless of the chosen sampling interval. Based on this, we tested three different options for substrate uptake rate estimations.

Keywords Bioprocess development · Cubic smoothing spline · Fed-batch fermentation · Growth rate · Substrate uptake rate

Introduction

State variables, such as biomass, substrates, and product, are quantified via off-line measurements during cultivation processes of microbial, mammalian and yeast cells to understand how the process states evolve. To shed light into the biological subsystem, i.e., the cell state, as well as the metabolism [4, 6, 8, 12] or to compare different cultivations on the biological level, e.g., for media selection or cell line development [13, 16, 19], specific production/consumption rates are a necessity.

Principle approaches to rate estimation

There are several approaches for estimating rates of a bioprocess [7, 15, 21]. A very simple method is to calculate the first derivative of a cubic smoothing spline function [15, 21]. The result is a continuous rate over the whole course of a bioprocess such as a fed-batch process, where for every time point, a rate value can be derived.

Although the applicability of this non-parametric method on bioprocess data is known for a longer time [3, 15], it still does not seem to be the method of choice for researchers in upstream bioprocess engineering, or related fields of biology. In most cases, the integral approach, a simple stepwise integral estimation is used [5, 10, 11, 25]. Hereby two measurements, one derived from sampling time point t_i and the other from sampling time point t_{i+1}, are considered to estimate a rate for this interval (t_i, t_{i+1}). The same methodology is then applied to the next interval (t_{i+1}, t_{i+2}) and so on, estimating one rate value for each time interval, resulting in a trend over the course of the cultivation process. This, in turn, means that the rate is assumed to be constant for each

B. Bayer and B. Sissolak contributed equally.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00449-019-02214-6) contains supplementary material, which is available to authorized users.

✉ B. Bayer
benjamin.bayer@boku.ac.at

✉ B. Sissolak
bernhard.sissolak@boku.ac.at

[1] Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria

[2] Bilfinger Industrietechnik Salzburg GmbH, Salzburg, Austria

[3] School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, UK

sampling interval, for which it was calculated, independent on its length.

Parameters impacting rate estimation quality

Some parameters do have a high impact on the outcome of these rate estimations and if treated in the wrong way result in false estimations. For instance, dynamic process trends can remain unnoticed, e.g., if the sampling frequency is too low. In addition, if larger measurement errors are present, the rate is not feasible to describe the process anymore due to this inaccuracy. This can lead to a reduction of the accuracy of the rates and to a reasonably weakened hypothesis on the influences of certain variables or parameters. To make the calculations more applicable, different smoothing approaches for rates can be used. An often described and simple method is the moving average [9, 26]. Here, the rates from several sampling points are smoothed by taking the average value from a sampling window. In addition, more advanced moving average filters such as low-pass and Savitzky–Golay were already retrospectively used for rate modeling of bioprocesses [14, 17]. Such advanced filters require settings and appropriate knowledge for the ideal window size and smoothness, which are dependent on the process they are applied on. Using these methods, the true covariance matrix is often underestimated and the lack of automatic constraints for state variables may lead to suboptimal performances [23].

Accurate estimation of a rate

Key figures existing in every cultivation process are the growth rate μ, which is defined as the time derivative of the logarithm of the change in population size and specific substrate uptake rates, which are feed dependent. Although stepwise integral estimation gives a simple estimation of the growth rates, this calculation possesses several drawbacks. One discrete estimation from one sampling time point to the next one is suboptimal for non-linear trends. Due to inaccurate biomass measurements, which is, in particular, true for cell culture cultivations, cell growth rates vary strongly between the samplings, indicating a false process status. On the other hand, variations in the amount of fed substrate can have substantial impacts on the specific uptake rate estimation due to error propagation. A switch in the cell's behavior is more likely to happen continuously and not spontaneously. It can be expected that calculations and model building attempts with these obtained biased values can lead to unreliable results containing much noise. To yield better descriptions of cultivation processes continuous rates should be preferred over sudden changes to yield.

Since the "true" rate is not accessible in a real fermentation process, because of the existence of analytical measurement errors [20] and biological differences from cultivation to cultivation, we present a simulated case study, at which linear and inhibited cell growth were simulated in-silico. Noise was added to the dataset to mimic a range of typical analytical measurement errors. 100 single fed-batch processes were simulated to obtain a statistical meaningful dataset. We compared the performance of the stepwise integral estimation including post-smoothing with a simple moving average with the cubic smoothing spline function. Hereby, different sampling intervals and analytical measurement errors have been simulated and both approaches were elucidated with respect to their precision and accuracy to obtain the real rates. Additionally, we also highlight an optimal solution to describe the substrate uptake rates, since for estimating substrate uptake rates, the feeding rate and feeding substrate concentration need to be taken into account. Any analytical error in this part can have a huge impact on the level of noise in the data.

The unique combination of different rate calculations applied on data with varying sampling frequencies and analytical deviations is very valuable for process understanding and modeling.

Materials and methods

The detailed cultivation settings for the different simulated in-silico fed-batch fermentations (table 1) and all the necessary equations (Eqs. 1–4) are given in the *Bioprocess Simulation* section of the Online Resource 1.

Noise generation

To account for process and analytic related variance, randomly generated multivariate normal distributed numbers were added, accounting for different precision levels in each process variable. Such noise was added to volume (1%), substrate (1%), and biomass, for every sampling point. For the biomass, five different levels of coefficient of variation (CV) were utilized (2.5, 5, 7.5, 10 and 12.5%). The CV (Eq. 1) is the standardized standard deviation, independent of the extent of the value and, therefore, a good estimation for accuracy:

$$\mathrm{CV} = \frac{\sigma}{\bar{X}} \times 100. \tag{1}$$

The CV describes the magnitude of variation for 68.2% of the data with the standard deviation σ and the average value $\bar{X}$.

Stepwise integral estimation

The most commonly used method, the stepwise integral estimation, of calculating specific growth rates using the measured cell dry mass is described in the following equation:

$$\mu = \frac{\ln\left(\frac{X(t)}{X(t-1)}\right)}{\mathrm{d}t}. \tag{2}$$

As in Takuma et al. [22], μ is estimated for each time interval between two measurements by dividing the current total biomass $X(t)$ with the value of the previous measurement $X(t-1)$. This equation assumes that μ is constant for the described time interval.

Moving average

A moving average filter was applied to smooth the stepwise integral estimation by calculating the mean of the observations using a fixed window size as stated in the following equation:

$$\mu_{\mathrm{MA}} = \frac{\mu_{(t)} + \cdots + \mu_{(t+n-1)}}{n}, \tag{3}$$

with μ_{MA} as the smoothed value, μ the growth rate, and the chosen window size n.

Cubic smoothing spline

For the specific growth rate estimation via cubic smoothing spline, the MATLAB function *csaps(x,y,p)* was applied with x the total time of the process, the total cell mass y, and the chosen value for the fitting parameter p. This function is an implementation of the Fortran function SMOOTH [18]. The fitting parameter p determines the relative weight to either smooth or perfectly match the data. Here, the least-squares solution ($p=0$) is a straight line fit, while $p=1$ is the natural cubic spline interpolation matching each data point. To find the optimal fit, the p value was screened with a resolution of 0.1 and applied to the data. By choosing an appropriate value for p, the current growth rate can be determined by computing the functions respective time derivative (Eq. 4):

$$\frac{\mathrm{d}(xV)}{\mathrm{d}t} = \mu x V, \tag{4}$$

with x representing the biomass concentration and V the volume. The MATLAB script to apply the described cubic smoothing spline function to real data can be found in the Online Resource 2.

Specific substrate uptake rate

For the calculation of the specific substrate uptake rate in g/g/h (qS), different approaches were considered and compared with regard to the respective accuracy. For the following equations, uf represents the feed flowrate, Sf the substrate feed concentration, S the substrate concentration, V the volume, and x the biomass concentration. The change in substrate over time is determined by the amount of consumed and added substrate in the reactor (Eq. 5), accordingly:

$$\frac{\mathrm{d}(SV)}{\mathrm{d}t} = qSxV + ufSf. \tag{5}$$

Option 1

For the first approach, the total substrate consumption (i.e., accumulation minus input) was calculated and set into a relationship to the qS (Eq. 6). Accordingly, rearranging and integrating Eq. (5) resulted in:

$$\frac{\mathrm{d}(SV - S_0V_0 - \int ufSf\,\mathrm{d}t)}{\mathrm{d}t}\frac{1}{xV} = qS. \tag{6}$$

A cubic smoothing spline fit was performed on the total consumption ($SV - S_0V_0 - \int ufSf\,\mathrm{d}t$) and on the biomass term (xV).

Option 2

For the second approach, the total amount of substrate in the supernatant was taken into consideration for the spline function and set into relation with the qS (Eq. 7). The cubic smoothing spline fit was performed on the substrate term (SV) and on the biomass term (xV):

$$\left(\frac{\mathrm{d}(SV)}{\mathrm{d}t} - ufSf\right)\frac{1}{xV} = qS. \tag{7}$$

Option 3

The last approach is similar to the second one, but only takes the substrate concentration in the supernatant into account. Accordingly, it follows from Eq. (5):

$$\frac{\mathrm{d}(SV)}{\mathrm{d}t} = V\frac{\mathrm{d}S}{\mathrm{d}t} + S\frac{\mathrm{d}V}{\mathrm{d}t} = qSxV + ufS, \quad \text{with}\ \frac{\mathrm{d}V}{\mathrm{d}t} = uf, \tag{8}$$

$$V\frac{\mathrm{d}S}{\mathrm{d}t} - ufSf + ufS = qSxV, \quad \text{with}\ D = \frac{uf}{V}, \tag{9}$$

$$\left(\frac{\mathrm{d}S}{\mathrm{d}t} - D(Sf - S)\right)\frac{1}{xV} = qS. \tag{10}$$

For this, an additional variable must be introduced, the dilution rate D, which is defined as the ratio of uf to V (Eq. 10). The cubic smoothing spline fit was performed on the substrate concentration term (S) and on the biomass term (xV).

RMSE and MAPE calculation

The root-mean-square error (RMSE) was calculated according to Eq. (11) and the mean absolute percentage error (MAPE) according to Eq. (12), where $\hat{y}$ describes the actual value, y the desired target value and n the number of samples:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t0}^{i}(\hat{y}(t) - y(t))^2}{n}}, \tag{11}$$

$$\mathrm{MAPE} = \frac{\sum \frac{\left|y(t) - \widehat{y(t)}\right|}{y(t)}}{n} \times 100. \tag{12}$$

Results

Bioprocess simulation

The two different bioprocess setups are displayed in Fig. 1. Simulation 1 describes a bioprocess were the cells are not induced or do not exhibit any growth inhibition (Fig. 1a). The second simulation describes a typical biomass trend of an induced microbial process (Fig. 1b). Due to this setup, we obtained completely different trends for the biomass as well as for the substrate concentrations. This allows to test if the distinct curvature of those trends leads to any unwanted effects when the different methods calculating the growth rate are applied.

When a process is performed with exactly the same process parameters for an infinite number of runs and with the exact same time interval at which samples are drawn, still random errors are likely to occur. Due to the analytical method precision, which depends on the utilized device different amounts of CV can be expected. The CV of biomass determination, for instance, is obviously depending on the used method. Gravimetric dried biomass determination for *E. coli* is expected to be quite accurate, whereas the measurement of the viable cell count via a microscope using a

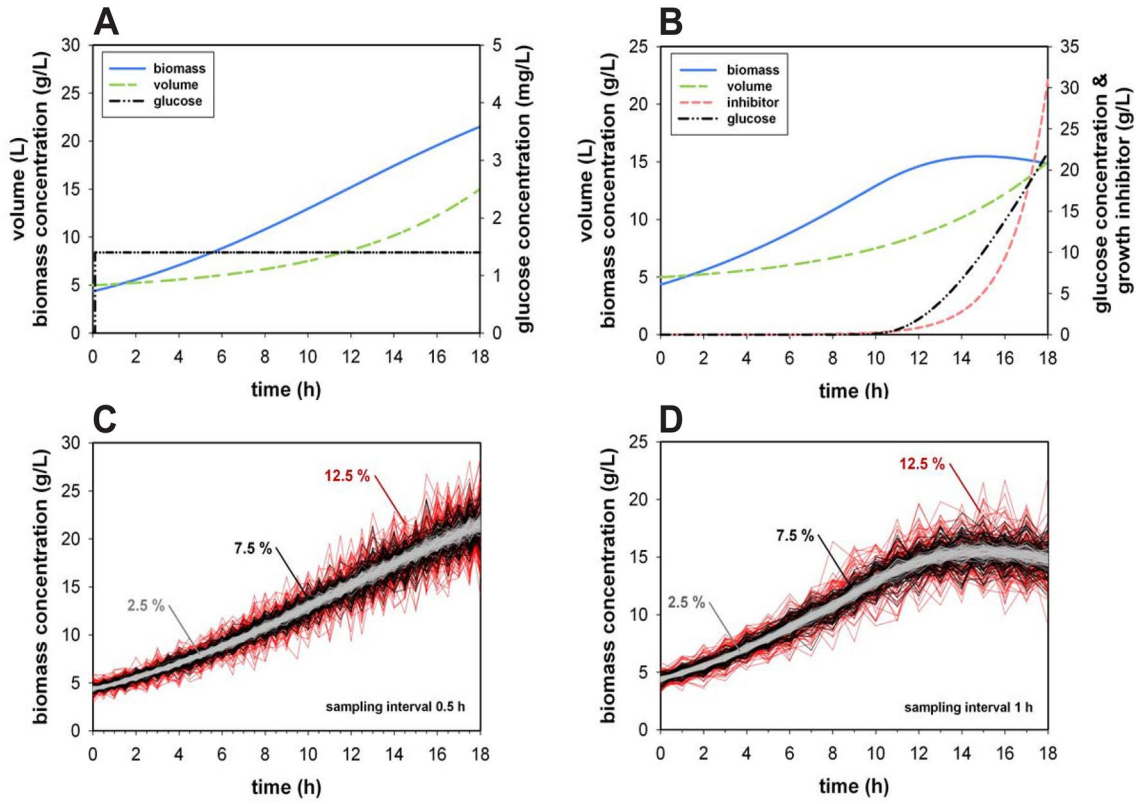

Fig. 1 Simulated **a** Monod and **b** non-competitive model process parameters and biomass concentration variation due to random sampling error at 12.5%, 7.5% and 2.5% CV for the Monod model (**c**) with a sampling interval of 0.5 h and the non-competitive model (**d**) with a sampling interval of 1 h are presented. For **c**, **d** the number of simulated fed-batch processes $n = 100$

hemocytometer can be rather imprecise [1, 2]. The generated variations between 2.5 and 12.5% already represent very precise cell measurements. For instance, at 7.5% CV, the biomass at 20 g/L varies with ±1.5 g/L, which is an absolutely realistic value (see Fig. 1c, d).

Rate estimations via stepwise integral estimation and elucidation of sampling interval impact

In the first step, the growth rates for the 100 simulated fed-batch experiments were calculated and the accuracy and precision of the growth rate estimations were determined. For each rate $\mu(i)$ at time point $t(i)$, the average and the standard deviation were calculated (n = 100). On average, the stepwise integral estimation is able to determine the rate quite precisely, independently if the growth rate is constant (Fig. 2a) or not (Fig. 2b). However, it is attended by low accuracy and further depends on the sampling interval and biomass accuracy. At an interval of 0.5 h, for instance, the minimal CV is already around 50% (Fig. 2c, d). Additionally, at a low biomass determination accuracy, the CV even increases fivefold. If the growth rate is following a dynamic trend, the maximum CV at the highest sampling frequency is almost 400%. For both bioprocesses, the CV for almost half of the dataset was higher than 50%.

This behavior of the stepwise integration has huge implications on the evaluation of the current growth rates. For instance, if the growth rate would be rapidly changed back and forth due to a modification in the experimental condition, the stepwise integration approach would not be able to recognize this and the information would remain hidden because of the weak performance.

Rate estimation via cubic smoothing spline

The cubic smoothing spline function was applied to the whole data for each run. The performance of the smoothing spline curve is displayed in Fig. 3. Additionally for the smoothing spline, also the perfect value for a general purpose of p was screened. A fitting parameter p of 1 led to a very low error but also to a generalization of the data and a p of 0 to an increasingly high error due to the simple straight line fit (Fig. 3a). Therefore, both were not displayed in Fig. 3b. To obtain the optimal p, the RMSE (Eq. 11) of

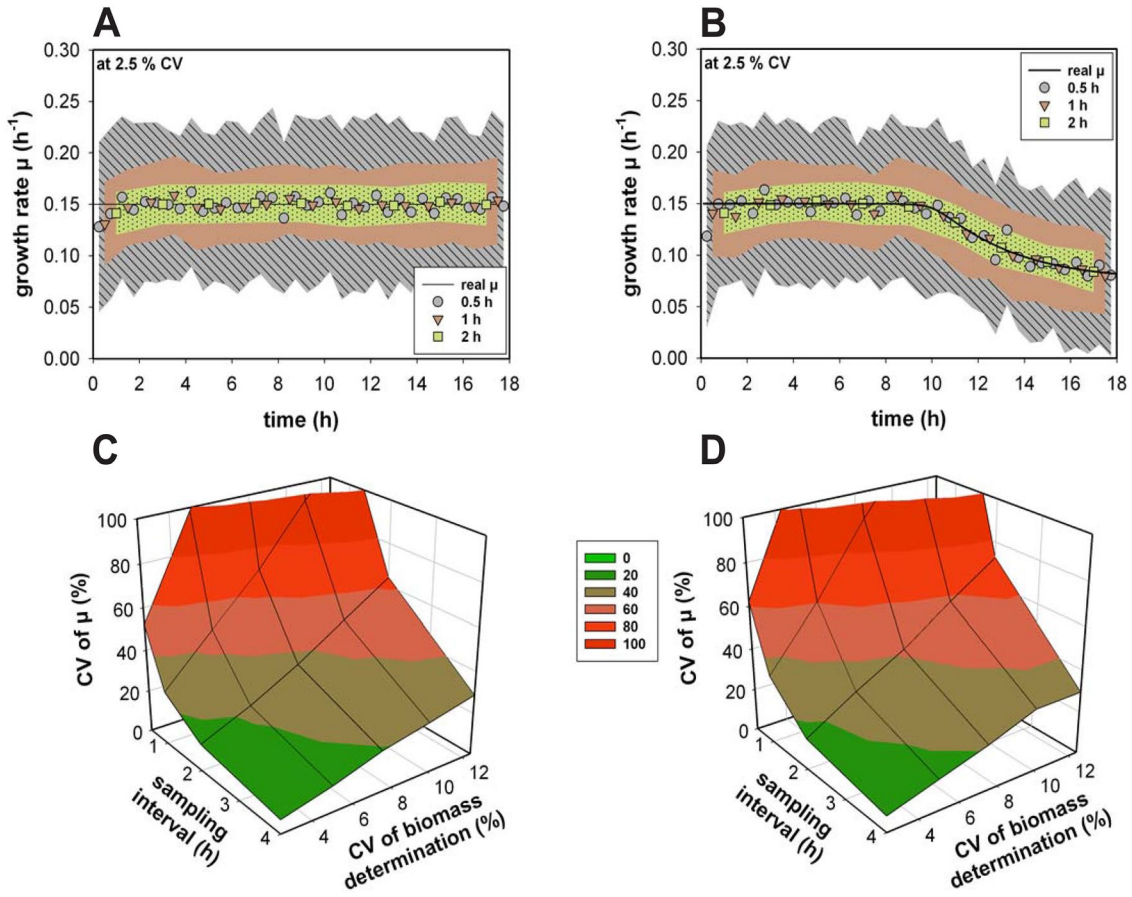

Fig. 2 **a**, **b** The estimated growth rates at different sampling intervals and their respective standard deviations (depicted by the area) at a biomass determination precision of 2.5% coefficient of variation (CV). **c**, **d** The resulting CV of the growth rate μ as a function of the sampling interval and at different biomass determination precisions for Monod model (**a**, **c**) and the non-competitive model (**b**, **d**) The number of simulated processes $n = 100$. Data above 100% are not depicted

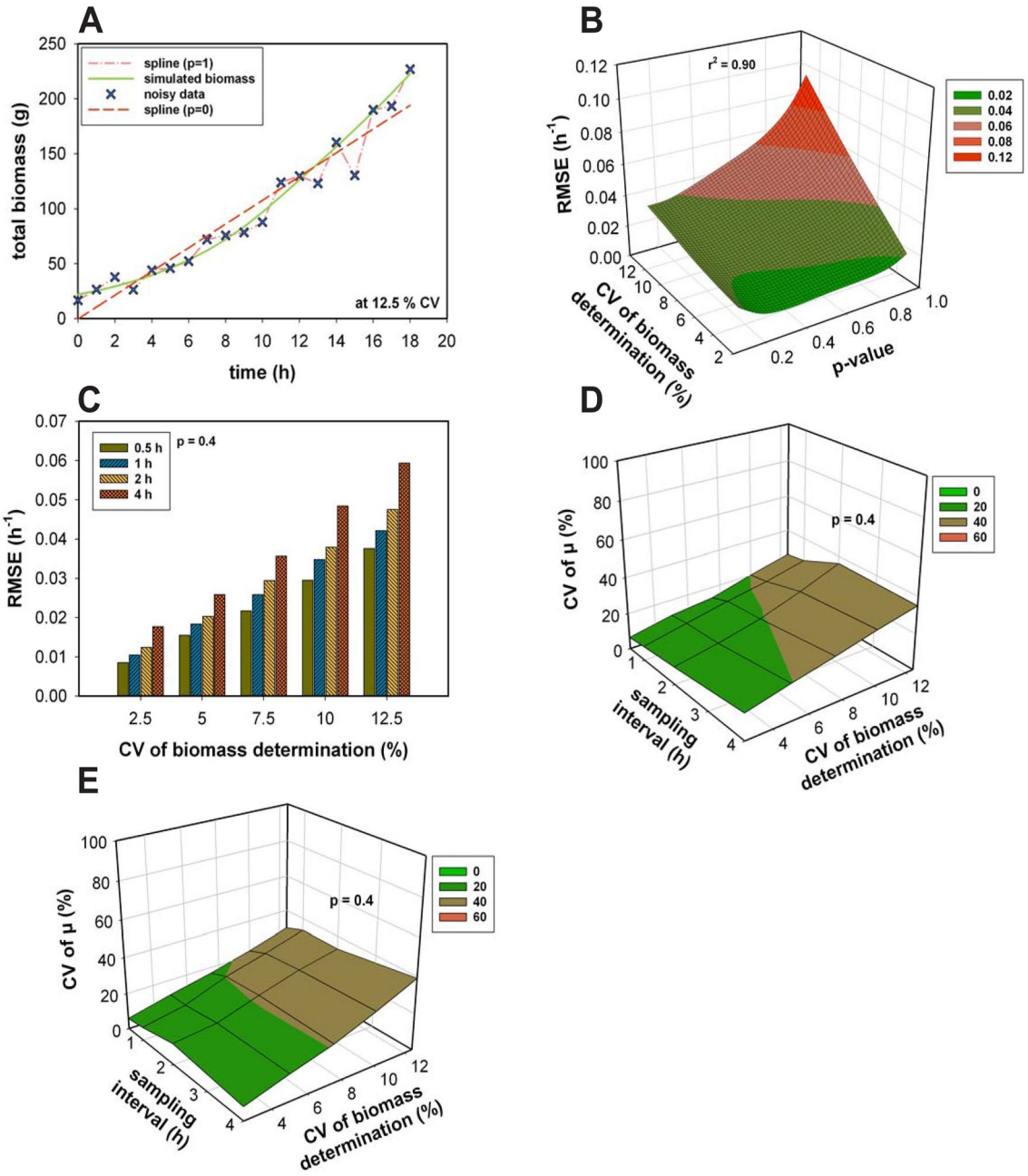

Fig. 3 **a** Spline fittings with p 0 and 1 of noisy biomass data (12.5% CV of biomass determination). **b** RMSE as a function of the sampling interval, the CV of biomass determination and the fitting parameter p of the spline function. **c** RMSE at a p of 0.4 at different sampling intervals. The coefficient of variation (CV) of the growth rate for the Monod model (**d**) and the non-competitive model (**e**) as a function of the sampling interval and CV of biomass determination for a fitting parameter p of 0.4. For **b**–**e** the number of simulated processes $n = 100$

the rates for 100 simulated fed-batch experiments at different sampling frequencies and CV for biomass determination was calculated (Fig. 3b) and described as a function of p, added noise, and sampling frequencies. The RMSEs of all the sampling intervals resulted in a similar shape. The surface exhibited a minimum at a p around 0.4 for all noise and sampling frequency combinations except for noise levels > 10% and the lowest sampling frequency of 4 h where a slightly lower p of 0.2 would be more preferable (see also Fig. 3c). Consequently, a fitting parameter of 0.4 was chosen for all further processes. At this magnitude, also the overall error at high sampling intervals and large measurement errors is reasonable low. Once the fit is applied sufficiently, the time derivative of this function represents the current growth rate. A very precise and accurate fit can be generated, which is sampling interval independent using the applied smoothing spline function. Even if the rate estimations became slightly inaccurate at the beginning and at the end of the processes,

still the precision for the rate estimations via spline is high. No differences between the estimation of a constant and a decreasing growth rate were evident. Also, if large noise was present, the spline was still able to estimate the rates correct and precise (Fig. 3d, e). With a biomass measurement error of 12.5%, the calculated CV ranged around 50% ($n = 100$).

Methodical comparison: stepwise integral estimation and cubic smoothing spline

The combination of stepwise integration and a moving average is a widely used approach for gathering smoothed rates. In the following, we elucidate the differences of using this combined method with the cubic smoothing spline.

The rate estimations described via the cubic smoothing spline outperformed the stepwise integral estimation. While the spline is considering the whole data, the stepwise integral estimation only takes two consecutive time points into account. Hence, smoothing splines can better deal with the error in the data compared to stepwise integral estimations. Regarding stepwise integral estimation, the error in the data is further propagated into the rate calculation. The spline fit already smooths the data before it gets even further processed. Considering this fact, it is obvious that spline functions are more accurate and precise.

A very common approach to further process the rates derived from stepwise integral estimations is to apply a moving average filter to smooth the data. For this study, we have chosen an averaging window size of 3 and 4. As expected the larger is the window size, the smaller the variations. Even with a window size of 3, the RMSE was reduced to an acceptable level. At a window size of 4, the error in the rate estimations in some cases was even better than the ones calculated with the cubic smoothing spline (Fig. 4).

However, due to the moving average, the rate change will seem to occur at different time points than it is the case. This is, in particular, a problem for non-constant rates (Fig. 4b). This effect will get even stronger at lower sampling frequencies. Further, averaging rates over several time points reduces the ability to describe the dynamics in the system, whereas exactly this should be described by the rates. The more likely process changes occur and the larger the averaging window is, the more likely they are overseen. Hence, the increased precision is traded for a reduced rates description.

The user also has to face the so-called endpoint problem. Due to the application of the moving average, the end of the process is not determined. Depending on the window size, the timeline of the rates will be inevitable shorter. Consequently, the utilization of moving average will reduce variation in the prediction, but will also lead to a reduced descriptiveness of the process and to misleading assumptions.

Specific substrate uptake rate estimations via the cubic smoothing spline

Other important process characteristics are substrate uptake rates. In this specific case, the amount of fed substrate must be incorporated into the calculation and with it any possible variations and errors, which might come along. Since we already verified the superiority of a cubic smoothing spline we only focused on the performance of this approach. A simulation of 100 fed-batch processes using the non-competitive model was performed in which a feed variation of 1% occurs. The sampling interval was chosen to be 1 h and the worst case of 12.5% CV for the biomass determination was used

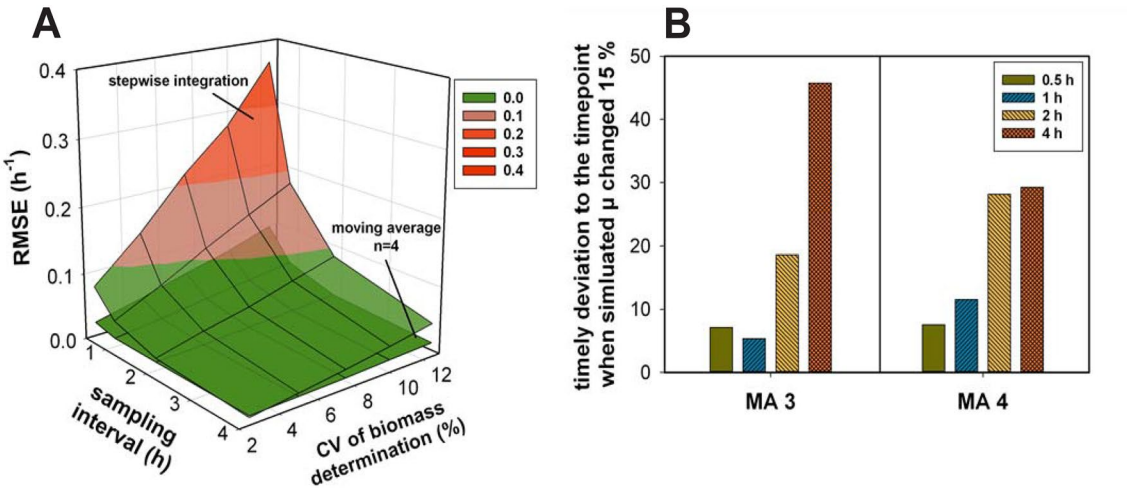

Fig. 4 Comparing the RMSE values of the stepwise integral estimations (**a**) and stepwise integral estimations using a moving average ($n = 4$) as a function of the sampling interval and CV of biomass determination. **b** The timely deviation (%) from the time point when the simulated μ changed 15% (non-competitive model) derived from utilizing moving average with a window size of 3 and 4. The number of simulated processes $n = 100$

and the fitting parameter p was set to 0.4. There are three possible options for the estimation of a feed-dependent rate. Either the total amount of consumed substrate (Option 1), the total amount of substrate in the supernatant (Option 2) or the substrate concentration in the supernatant (Option 3) can be taken into consideration for the cubic smoothing spline fitting (Fig. 5a–c).

All three options can in average accurately describe the specific substrate uptake rate (Fig. 5d). However, the incorporation of the feed into the calculation beforehand increased the precision to a great extent (Option 1) and also the feeding noise can be almost completely erased. Interestingly, between option 2 and 3, respectively, using the total amount of substrate or the substrate concentration, no significant difference was observed (see Fig. 5e). Only at the end of the fed-batch process, option 2 underestimates the specific substrate uptake rate. However, already 1% variation in the feeding system can have a substantial impact. As a consequence of using the wrong approach, the error will increase almost fourfold (Fig. 5f) from around 5% up to 20% MAPE (Eq. 12). If the feed is not incorporated into the calculation beforehand, such as it

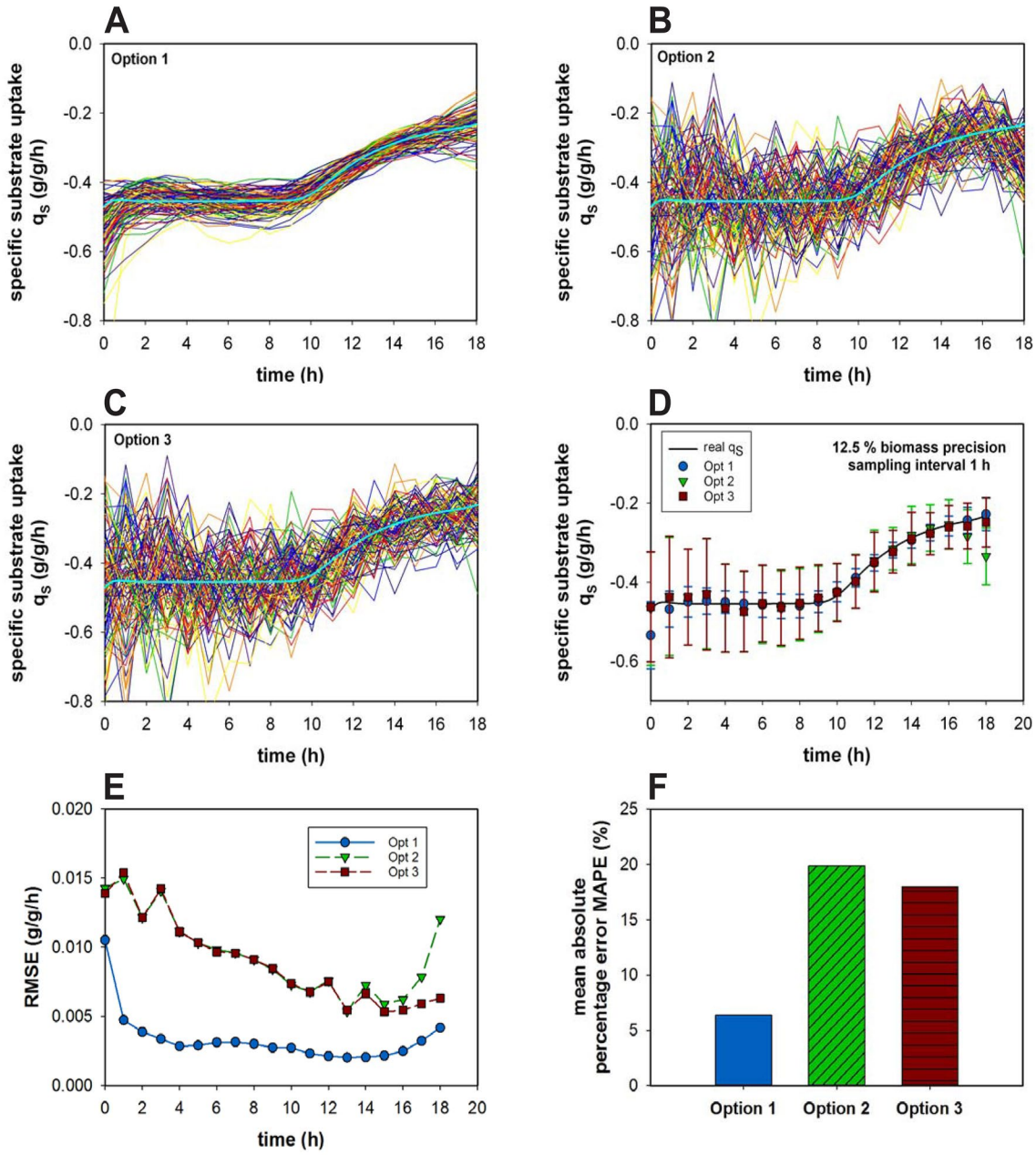

Fig. 5 Specific substrate uptake rate estimation via option 1 (**a**) 2 (**b**) and 3 (**c**) over the time course of a fed-batch ($n = 100$) for a sampling interval of 1 h and precision of 12.5% CV for the biomass determination are presented. The averaged values and their respective standard deviations of the three different options over the time course of the process (**d**), the resulting RMSE values for each option and sampling point (**e**), and MAPE for all three options (**f**) are displayed. The number of simulated processes $n = 100$

is the case in Option 2 and 3, the feeding error propagates further into the rate estimation.

Discussion

Stepwise integral estimation issues

The key to process development and process modeling is to estimate rates accurately and precisely. In average ($n = 100$), the stepwise integral approach calculated an accurate rate value. This was expected considering that a large number of repetitive experiments should always meet in average the desired target value. But, we demonstrated that the stepwise integral estimation will end up in large variations. It is not surprising that the inaccuracy rises with an increased sampling frequency [24], but such an increasing variation at higher sampling frequencies was on first sight rather unexpected. Due to the magnitude of the sampling errors, the slope of the linear function will either be more positive or negative, in comparison to the real value. Every new sampling point will add its failure to it and, consequently, the deviation will increase over the time course of the cultivation. Therefore, with an increased sampling frequency, the rate estimation error increases although the measurement error remains constant. Since this behavior is counterintuitive, it is most likely overseen. This is a major disadvantage since for accurate process characterization and to gather process know-how a large dataset, thus a high sampling frequency, is a necessity. The application of the moving average would be a simple tool to reduce such variances but the user will eventually end up in less accurate values. Therefore, rates calculated by stepwise integral estimation should be handled carefully for modeling purposes.

Application of cubic spline and specific substrate rate estimation

In this study, we focused on the cubic smoothing spline function as an alternative to rate estimations via stepwise integral estimation. With a reduced precision of the analytical determination, also the variation in the estimation increased but not to the same extent as when the stepwise integral estimation was applied. In the best case, at a high sampling frequency and biomass determination inaccuracy, the CV was around a factor of 4 lower. Moreover, the cubic smoothing spline was not affected by the sampling frequency. In real bioprocesses, a good trade-off between sampling frequency, process dynamics and the analytical error should be considered. For high analytical errors and slow process dynamic changes, a high sampling interval does not increase precision and accuracy.

Additionally, we elucidated three different approaches for estimating substrate uptake rates via the established spline fit. If the substrate feed is not incorporated beforehand a cubic spline is performed, feed variations can have a substantial impact on the propagated error. Hence, it is important to first calculate the total amount of consumed substrate before the rates are estimated.

The only "drawback" using the cubic smoothing spline function is that one degree of freedom is present, the fitting parameter p. Therefore, before processing the optimal p must be reconsidered with respect to the given magnitude of the x ordinate. Another powerful alternative to spline functions can be found in Gaussian distributions. It was shown that for processes with high sampling numbers (100–1000), the Gaussian distribution outperforms the spline function while for samplings below 100, it is vice-versa [21]. Typically, mammalian cell culture processes lead to only 10–20 observations. Likewise, also microbial fermentations do not comprise such a high sampling frequency, also resulting in only 15–25 observations per process. These considerations and the remarkably easy use of this method due to no data pre- or post-processing are clearly stating the advantage of the smoothing spline compared with other methods.

Conclusion

In this study, the specific growth rate and the specific substrate uptake rate were chosen as representative examples. It was shown that cubic spline estimations are a simple but powerful tool to determine rates, compared to the most commonly used standard procedure the stepwise integral estimation. The presented method:

- is easy to apply and to implement for off-line analytical purposes,
- is to a major extent sample interval independent,
- can cope with large analytical variances,
- allows the user to assess a rate value at every time point.

In addition, we showed that a small error in the feeding system can lead to huge impacts in the estimation of specific substrate uptake rates. Hereby, it is important to take the feeding into account before the actual spline fit takes part.

For this level of complexity, the spline is sufficiently enough and more complex algorithms such as the Gaussian distribution or functions with more degrees of freedom (e.g., Kalman filters) are not necessary. It is easy to implement into existing codes and can add a reasonable value to process development and process comparability.

Acknowledgements Open access funding provided by University of Natural Resources and Life Sciences Vienna (BOKU). We would like

to thank Bilfinger Industrietechnik Salzburg and the Austrian Research Promotion Agency (FFG) for their support. (Research Studio Austria, 859219 and Competence Headquarters, 849725).

Compliance with ethical standards

Conflict of interest The authors have declared no conflicts of interest.

References

1. Bratbak G, Dundas IAN (1984) Bacterial dry matter content and biomass estimations. Appl Environ Microbiol 48(4):755–757
2. Cadena-Herrera D, Lara JEE, Ramírez-Ibañez ND, López-Morales CA, Pérez NO, Flores-Ortiz LF, Medina-Rivero E (2015) Validation of three viable-cell counting methods: manual, semi-automated, and automated. Biotechnol Rep 7:9–16. https://doi.org/10.1016/j.btre.2015.04.004
3. Craven P, Wahba G (1978) Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31(4):377–403. https://doi.org/10.1007/BF01404567
4. Ferreira AR, Dias JML, Teixeira AP, Carinhas N, Portela RMC, Isidro IA (2011) Projection to latent pathways (PLP): a constrained projection to latent variables (PLS) method for elementary flux modes discrimination. BMC Syst Biol 5(1):181. https://doi.org/10.1186/1752-0509-5-181
5. Franz C, Kern J, Karl B (2005) Sensor combination and chemometric modelling for improved process monitoring in recombinant *E. coli* fed-batch cultivations. J Biotechnol 120:183–196. https://doi.org/10.1016/j.jbiotec.2005.05.030
6. Galleguillos SN, Ruckerbauer D, Gerstl MP, Borth N, Hanscho M, Zanghellini J (2017) What can mathematical modelling say about CHO metabolism and protein glycosylation? Comput Struct Biotechnol J 15:212–221. https://doi.org/10.1016/j.csbj.2017.01.005
7. Glassey J, Gernaey KV, Clemens C, Schulz TW, Oliveira R, Striedner G, Mandenius C-F (2011) Process analytical technology (PAT) for biopharmaceuticals. Biotechnol J 6:369–377. https://doi.org/10.1002/biot.201000356
8. Hefzi H, Ang KS, Hanscho M, Borth N, Lee D, Lewis NE (2016) Consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism. Cell Syst 3:434–443. https://doi.org/10.1016/j.cels.2016.10.020
9. Herwig C, Marison I, Stockar U Von (2001) On-line stoichiometry and identification of metabolic state under dynamic process conditions. Biotechnol Bioeng 75(3):345–354
10. Li J, Jaitzig J, Lu P, Süssmuth RD, Neubauer P (2015) Scale-up bioprocess development for production of the antibiotic valinomycin in *Escherichia coli* based on consistent fed—batch cultivations. Microb Cell Fact. https://doi.org/10.1186/s12934-015-0272-y
11. Mairhofer J, Scharl T, Marisch K, Cserjan-Puschmann M, Striedner G (2013) Comparative transcription profiling and in-depth characterization of plasmid-based and plasmid-free *Escherichia coli* expression systems under production conditions. Appl Environ Microbiol 79(12):3802–3812. https://doi.org/10.1128/AEM.00365-13
12. Niklas J, Schräder E, Sandig V, Noll T, Heinzele E (2011) Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1. HN using time resolved metabolic flux analysis. Bioproc Biosyt Eng 34:533–545. https://doi.org/10.1007/s00449-010-0502-y
13. Noh SM, Shin S, Lee GM (2018) Comprehensive characterization of glutamine synthetase-mediated selection for the establishment of recombinant CHO cells producing monoclonal antibodies. Sci Rep 1–11. https://doi.org/10.1038/s41598-018-23720-9
14. Ohadi K, Legge RL, Budman HM (2014) Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures. Biotechnol Bioeng 112(1):197–208. https://doi.org/10.1002/bit.25339
15. Oner MD, Erickson LE, Yang SS (1986) Utilization of spline functions for smoothing fermentation data and for estimation of specific rates. Biotechnol Bioeng 28(6):902–918. https://doi.org/10.1002/bit.260280618
16. Pan X, Streefland M, Dalm C (2017) Selection of chemically defined media for CHO cell fed-batch culture processes. Cytotechnology 69:39–56. https://doi.org/10.1007/s10616-016-0036-5
17. Paulsson D, Gustavsson R, Mandenius C (2014) Filtering of metabolic heat signals. Sensors 14:17864–17882. https://doi.org/10.3390/s141017864
18. R. J, de Boor C (2006) A practical guide to splines. Math Comput 34(149):325. https://doi.org/10.2307/2006241
19. Sieck JB, Cordes T, Budach WE, Rhiel MH, Suemeghy Z, Leist C, Soos M (2013) Development of a scale-down model of hydrodynamic stress to study the performance of an industrial CHO cell line under simulated production scale bioreactor conditions. J Biotechnol 164(1):41–49. https://doi.org/10.1016/j.jbiotec.2012.11.012
20. Sonnleitner, B. (2007). Bioanalysis and biosensors for bioprocess monitoring. Springer, Berlin, pp 1–64. https://doi.org/10.1007/3-540-48773-5_1
21. Swain PS, Stevenson K, Leary A, Montano-Gutierrez LF, Clark IBN, Vogel J, Pilizota T (2016) Inferring time derivatives including cell growth rates using Gaussian processes. Nat Commun 7(May):1–8. https://doi.org/10.1038/ncomms13766
22. Takuma S, Hirashima C, Piret JM (2007) Dependence on glucose limitation of the pCO_2 Influences on CHO cell growth. Metab IgG Prod 97(6):1479–1488. https://doi.org/10.1002/bit
23. Ungarala S, Dolence E, Li K (2007) Constrained extended Kalman filter. IFAC Proc 2:63–68
24. Wechselberger P, Herwig C (2012) Model-based analysis on the relationship of signal quality to real-time extraction of information in bioprocesses. AIChE J 28(1):265–275. https://doi.org/10.1002/btpr.700
25. Wechselberger P, Sagmeister P (2013) Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes. Bioproc Biosyst Eng 36:1205–1218. https://doi.org/10.1007/s00449-012-0848-4
26. Zahel T, Sagmeister P, Suchocki S, Herwig C (2016) Accurate information from fermentation processes-optimal rate calculation by dynamic window adaptation. Chem-Ing-Tech 88(6):798–808. https://doi.org/10.1002/cite.201500085

Publication I

Supporting Information

Online Resources

Bioprocess Simulation

We simulated the fed-batch phase of an *E. coli* fermentation for two different process setups using MATLAB (2016b, MathWorks, Massachusetts, USA). Each simulation started with the indicated parameter values as stated in Table 1. 100 individual fed-batches for each setting were simulated. These processes were simulated with varying sampling intervals of 0.5, 1, 2 and 4 h, respectively and a negligibly small sampling volume. Also, the biomass X was calculated as stated in Eq. 1 via the growth rate μ and time t.

$$X(t) = X(t_0)\, e^{\mu\,(t-t_0)} \tag{1}$$

Further, the exponential feeding strategy was established as indicated in Eq. 2.

$$uf = \frac{1}{Sf}\, X(t_0)\, Y_{xs}\, \mu\, e^{\mu\,(t-t_0)} \tag{2}$$

With the feed flow rate uf, the feed glucose concentration Sf, the total biomass at the feed start $X(t_o)$, the biomass per glucose yield Y_{xs} and the set growth rate μ.

Table 1. Cultivation settings for the simulated fed-batch processes.

Parameter	**Value**
starting biomass concentration	4.4 g/L
vessel volume	5 L
feeding strategy	exponential
growth rate	0.15 h^{-1}
feed duration	18 h
feed glucose concentration	100 g/L
inhibitor concentration	80 g/L
yield biomass/glucose	0.33 g/g
sampling interval	0.5, 1, 2, 4 h

Monod model

The first *in-silico* process setup was based on a glucose limitation and therefore the apparent growth rate was adjusted by the feeding rate only, following the order of Monod (Eq. 3) as shown in Figure 1A.

$$\mu = \frac{\mu_{\max}\ S}{K_S + S} \tag{3}$$

With the specific growth rate μ, the assumed maximum growth rate μ_{max} of 0.9 h^{-1}, the limiting glucose/substrate concentration S, and the substrates affinity constant K_S with an assumed value of 0.007 (Senn, Lendenmann, Snozzi, Hamer, & Egli, 1994).

Non-competitive model

For the second process setup the exponential addition of an inhibitor, mimicking the product formation, was taken into account (Eq. 4). The inhibitor concentration was selected in a way that adequately mimics a decreased growth rate as presented in Figure 1B. Hence, the simulation was performed in such a way that the inhibitor was included in the feed.

$$\mu = \frac{\frac{\mu_{\max} S}{K_S + S}}{1 + \frac{I}{K_i}} \quad (4)$$

Using the non-competitive model the Monod equation with the same assumed values is extended with an additional term to describe the growth inhibition, containing the inhibitor concentration *I* and its affinity constant K_i with an assumed value of 1.7.

MATLAB script for growth rate estimation via cubic smoothing spline

Depending on the dataset the optimal *p* can differ. We observed a time dependency with respect to *p* using minutes instead of hours for microbial or hours instead of days for mammalian processes, which can lead to different cubic spline results. Thus, to expect optimal fitting results for a fixed *p*, the time axis should be in a similar range. In this case, the optimal *p* of 0.4 is valid for processes in the double-digit range (e.g.: 20 hours or days). Moreover, we performed *csaps* for real microbial as well as cell culture processes and always established a good fitting performance with a *p* of 0.4.

For this section user input is needed - set the value for the fitting parameter p and import your Excel file including your process data for the growth rate calculation

```
clear ; clc ; close all

% create an Excel file, it only has to include the columns for the
 following process parameters in the indicated order to work (exclude
 any headers)
% column A = absolute time of the process
% column B = viable cell concentration/biomass
% column C = vessel volume (in the same unit as the concentration in
 Column B)

% if your file is complete, choose your wanted value for the fitting
 parameter p
fitP = 1 ;

% import your Excel file by assigning it to the variable 'importData'
importData = importdata('testfile.xlsx') ;

% the output of this script will be a newly generated Excel file
% It contains the sampling points with the respective viable cell
 concentration/biomass, total viable cells/biomass and the growth
 rate, all calculated via the cubic smoothing spline function
% the used value for the fitting parameter will also be saved in this
 Excel file
% the file will be created into the same folder as this Matlab script
 is located

% run the code (F5)
```

no further user input is needed for this section - rate calculation and export to Excel file

```
% assigns the variables to the column number in your Excel file
Time_Total = importData(:,1) ;
VCC = importData(:,2) ;
Volume = importData(:,3) ;

% function generation and value calculation
fnX = csaps(Time_Total,VCC.*Volume,fitP);
GrowthRate_Spline_Samplings = fnval(fnder(fnX,1),Time_Total)./
fnval(fnX,Time_Total);

% plots the measured and calculated cell concentration
```

```
figure(1)
plot(Time_Total, VCC.*Volume, 'kd', 'MarkerSize',
 10,'markerfacecolor','k')
title('Measured versus calculated Values')
xlabel('Total Time')
ylabel('Total Viable Cells')
hold on
fnplt(fnX)
legend('Measurement', 'Calculated Trend', 'Location', 'northwest')
hold off

% plots the calculated growth rate via smoothing Spline with the
 chosen fitting parameter p
figure(2)
hold on
plot(Time_Total, GrowthRate_Spline_Samplings)
title('Calculated Growth Rate via Smoothing Spline')
xlabel('Total Time')
ylabel('Growth Rate')
legend('Spline Estimation') ;
hold off

% creates variables for the Excel export
ExcelSheet = 1 ;
Export_Names = {'Total Time - Sampling Points', 'Calculated Biomass/
Viable Cell Concentration via Smoothing Spline', 'Calculated Total
 Biomass/Viable Cells via Smoothing Spline', 'Calculated Growth Rate
 via Smoothing Spline',  'used Value of the Fitting Parameter'} ;

TVC_Calculated = fnplt(fnX) ; TVC_Calculated = TVC_Calculated' ;
alignment = knnsearch(TVC_Calculated(:,1),Time_Total) ;
VCC_Calculated = TVC_Calculated(alignment,2)./Volume ; TVC_Calculated
 = TVC_Calculated(alignment,2) ;

% creates Excel export file
xlswrite('SmoothingSpline_Calculations.xlsx', Export_Names,
 ExcelSheet, 'A1:E1') ;
xlswrite('SmoothingSpline_Calculations.xlsx', Time_Total,
 ExcelSheet, 'A2') ;
xlswrite('SmoothingSpline_Calculations.xlsx', VCC_Calculated,
 ExcelSheet, 'B2') ;
xlswrite('SmoothingSpline_Calculations.xlsx', TVC_Calculated,
 ExcelSheet, 'C2') ;
xlswrite('SmoothingSpline_Calculations.xlsx',
 GrowthRate_Spline_Samplings, ExcelSheet, 'D2')
xlswrite('SmoothingSpline_Calculations.xlsx', fitP,
 ExcelSheet, 'E2') ; clearvars
```

Publication II

Received: 29 April 2019 | Revised: 3 September 2019 | Accepted: 18 October 2019

DOI: 10.1002/elsc.201900076

Engineering in Life Sciences

RESEARCH ARTICLE

Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in *Escherichia coli* cultivations

Benjamin Bayer[1] | **Moritz von Stosch**[2] | **Michael Melcher**[3,4] | **Mark Duerkop**[1,5] | **Gerald Striedner**[1]

[1]Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria

[2]School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, United Kingdom

[3]Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, Vienna, Austria

[4]Austrian Centre of Industrial Biotechnology, Graz, Austria

[5]Novasign GmbH, Vienna, Austria

Correspondence
Benjamin Bayer, Department of Biotechnology, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria.
Email: benjamin.bayer@boku.ac.at

Funding information
Österreichische Forschungsförderungsgesellschaft, Grant/Award Number: 859219

Abstract

In bioprocesses, specific process responses such as the biomass cannot typically be measured directly on-line, since analytical sampling is associated with unavoidable time delays. Accessing those responses in real-time is essential for Quality by Design and process analytical technology concepts. Soft sensors overcome these limitations by indirectly measuring the variables of interest using a previously derived model and actual process data in real time. In this study, a biomass soft sensor based on 2D-fluorescence data and process data, was developed for a comprehensive study with a 20-L experimental design, for *Escherichia coli* fed-batch cultivations. A multivariate adaptive regression splines algorithm was applied to 2D-fluorescence spectra and process data, to estimate the biomass concentration at any time during the process. Prediction errors of 4.9% (0.99 g/L) for validation and 3.8% (0.69 g/L) for new data (external validation), were obtained. Using principal component and parallel factor analyses on the 2D-fluorescence data, two potential chemical compounds were identified and directly linked to cell metabolism. The same wavelength pairs were also important predictors for the regression-model performance. Overall, the proposed soft sensor is a valuable tool for monitoring the process performance on-line, enabling Quality by Design.

KEYWORDS

bioprocess engineering, chemometric modeling, multivariate adaptive regression spline, process monitoring, Quality by Design

1 | INTRODUCTION

1.1 | Recombinant protein production

At present, operators try to ensure process performance consistency by operating the process according to a fixed protocol, with deviations leading to post-process investigations. However, process inputs succumb inevitably to variability, and the quality is examined only at the end of the process. At this point, it is determined whether the outputs meet the required standards or whether the batch must be

Abbreviations: 5x-CV, five-fold cross-validation; CPP, critical process parameter; DoE, design of experiments; ex/em, excitation/emission; LoBo-CV, leave-one-batch-out cross-validation; MARS, multivariate adaptive regression splines; PARAFAC, parallel factor analysis; PAT, process analytical technology; PC, principal component; PCA, principal component analysis; QbD, Quality by Design; RMSE, root mean squared error; SGR, specific growth rate; VIP, importance of the input variables.

withdrawn [1]. Narrowing the output specifications to guarantee higher quality standards results in increasing numbers of rejected batches. This consequently leads to an enormous loss of energy, time, money, and goods [2]. Another point to consider is that the application of fixed process settings gives rise to variable outputs. This can be troublesome if a certain biomass is needed for a specific process operation, for example, induction in *Escherichia coli*. Thus, it is of great interest to know the current biomass concentration at any time during the process.

PRACTICAL APPLICATION

We propose a workflow to establish a soft sensor with an exceptional generalization capacity and wide applicability. The presented soft sensor is able to accurately estimate biomass concentrations on-line. Therefore, no analytical time delay occurs. This is of great interest to manufacturers, for monitoring and controlling their processes. For example, using this soft sensor, the induction could be always initiated at a defined biomass concentration. Moreover, the described modeling algorithm lists the predictive importance of all possible model parameters, enabling process understanding under the QbD concept. Furthermore, the soft sensor performance was tested by applying it to fermentations with different parameter settings as used for the design space characterization (up to three altered parameters). Despite the new fermentation settings, accurate estimations were obtained, which demonstrates the ability of the soft sensor to monitor the biomass concentration of different processes in real time.

1.2 | Quality by design and process analytical technology

Pharmaceutical manufacturing is tightly controlled by the authorities. The current procedure, namely Quality by Testing, is disadvantageous from an economic aspect and is associated with long development times. To tackle these batch-to-batch variations and inconsistencies and to increase process understanding, the FDA published the process analytical technology (PAT) guidance for the biopharmaceutical industry in 2004. This guidance proposes the use of risk assessments for the identification of critical process parameters (CPPs), whose impact on the product's critical quality attributes (CQAs) should be studied during process development [3]. Typically, this is accomplished utilizing design of experiments (DoE) approaches to analyze the CPPs' multidimensional impacts on the CQAs. Subsequently, a monitoring strategy must be defined to ensure that the process performs as expected and to provide an opportunity to counteract any input variations that may occur. This allows for more robust and uniform outputs with respect to quality assurance and proper risk management [4].

The gathered process knowledge should be used to switch from a Quality by Testing to a Quality by Design (QbD) approach [5]. This will lead to a well-understood process to the extent that the monitored variables and the quality are guaranteed by the process itself. Although plenty of information about QbD and its application is already available, QbD is still far from being implemented as the new state of the art, in particular for upstream bioprocess operations [6], due to the lack of appropriate monitoring tools.

1.3 | Advanced on-line sensor systems and soft sensors

Progress has been seen regarding on-line monitoring tools for the PAT concept, not only for microbial but also for mammalian cell cultures. Many optical sensor systems using different spectroscopic techniques are currently used in the industry [7]. For instance, simple in situ microscopic techniques are already in use [8], as well as more advanced Raman spectroscopy [9] or infrared spectroscopy [10] techniques. Fluorescence spectroscopy techniques are also associated with the group of advanced sensors. Fluorescence spectroscopy is based on determining the specific excitation and emission wavelengths of a compound in order to identify it qualitatively and quantitatively, in the range of the measured 2D-fluorescence spectrum [11,12]. This sensor type, together with other spectroscopic methods, is suitable for on-line applications, since continuous, non-invasive, and nondestructive measurements are possible and no sample needs to be drawn, thereby eliminating the risk of contamination. In addition, the determination of various compounds within a single measurement renders these techniques fast and robust, as well as cost-efficient. 2D-fluorescence spectroscopy is very sensitive and allows fluorescing molecules to be monitored inside and outside the cell. This technique has already been used to monitor microbial cultures and has been shown to reveal information about the physiological status of the cells [13].

Changes in the on-line signals, (e.g., fluorescence) can be used for chemometric modeling, to build so-called soft sensors for estimating various bioprocess quality attributes or variables of interest in real time. In particular, multivariate data analysis (MVDA) is used to investigate the correlations between on-line and off-line measurements. With the help of machine learning methods, these on-line signals can be translated into the corresponding off-line variables [14]. Hence, it is possible to estimate and monitor specific complex variables via unspecific on-line signals in real time and

moreover, to estimate non-fluorescent substances via their stoichiometric relationship to fluorescent compounds within the process [15].

1.4 | Multivariate data analysis and regression models

Unsupervised methods for exploratory data analysis, for example, parallel factor analysis (PARAFAC) [16] or principal component analysis (PCA) [17], are applied to gain deeper knowledge and to reveal information hidden within the data. In this way, important chemical compounds can be identified and further insights into the physiology of the cell can be obtained. An effective way of extracting information from process data and building soft sensors exploiting the hardware sensors used is MVDA [18]. These types of soft sensors are based on data and do not necessarily need further knowledge or mechanistic understanding. Some frequently applied machine learning methods make use of partial least squares regression, but non-linear methods such as random forest, artificial neural networks and support vector machines are also in use [19]. A powerful approach that takes strong interactions between variables into account and is also able to model non-linearities, is the multivariate adaptive regression spline (MARS). Due to its dynamic adaptability in selecting subsets of local variables, this algorithm can be seen as an ideal candidate for process modeling. The MARS algorithm has not been used for soft sensor-building in upstream processes to date. MARS is considered as an extension of linear models and is well suited to dealing with high dimensional input data [20].

Data preprocessing should also be taken into consideration before model-building, in order to develop a more robust model, for example, by using the z-score, that is, autoscaling [21]. This enables more accurate comparability between different processes by contemplating only the change over time instead of the quantity of the measured units [22]. The common way to validate the developed model is to apply the model to an independent test set, also referred to as an external validation set, which has not been used for model training.

This work presents a new soft sensor based on 2D-fluorescence data and other on-line process data. MARS was used for model-building, due to its simplicity compared to other algorithms that can deal with a large number of input variables, multi-collinearity and non-linearity. The soft sensor performance for on-line monitoring of the biomass is assessed for the 27 distinct experiments of a complete DoE study, as well as for two DoE-independent test runs. Exploratory data analysis was performed to gain insight into the data and to investigate the fluorescence spectra. The important wavelengths for biomass sensing and the potential underlying chemical compounds accounting for cell metabolism, were identified and described in detail using PCA and PARAFAC.

2 | MATERIALS AND METHODS

2.1 | Process conditions

E. coli (HMS174 (DE3)) was cultivated in fed-batch fermentations at a 20-L scale, expressing recombinant human Cu/Zn superoxide dismutase. All details of the bacterial strain, plasmid, cultivation, induction conditions, and on-line and off-line monitoring, have already been described elsewhere [23, 24]. The impact of three CPPs on the process performance using DoE, was studied. These were temperature (30, 34, and 37°C), the induction ratio (0.2, 0.5, and 0.9 μmol IPTG/g cell dry mass) and the specific growth rate (SGR) (0.10, 0.15, and 0.20 hours $(h)^{-1}$). The SGR was held constant by an exponential substrate feed. All corresponding reactor volumes of the fed-batch fermentations are provided in Supporting Information Figure S1. This resulted in a 3-D design space with 27 CPP combinations. This design extends the space investigated in the earlier study.

2.2 | Data set

The data set consisted of 33 fermentations, with 27 experiments from the DoE study, together with two duplicates and one quintuplicate. Furthermore, two differing CPP settings, still located in the investigated space, were used as a test set. All CPP settings within the design space are listed in Supporting Information Table S1. The biomass (target variable) was measured once prior to the induction and thereafter at hourly intervals, via thermogravimetric analysis. The five variables available on-line (accumulated feed in grams, base in grams, inductor in μmol, temperature in°C and inlet air in standard liters per minute), as well as the 120 excitation/emission (ex/em) wavelength pairs measured by a 2D-fluorescence probe (BioView®, Delta Light & Optics, Denmark), were utilized as input data for model-building. The inlet air and the stirrer speed (not used for model-building), were used to keep the dissolved oxygen set point at 30% during the fermentations. The 2D-fluorescence probe measured the cultivation broth ranging from ex270/em310 up to ex550/em590, in 20-nm steps.

Exploratory data analysis and soft sensor development were performed using MATLAB (2016b, MathWorks, USA), together with the three freely available packages ARES-Lab [25], N-way [26], and drEEM [27]. A graphical overview of the complete development process for the soft sensor, from the data gathering stage to until the final model, is provided in Supporting Information Figure S2.

2.3 | Data preprocessing

2.3.1 | Standardization of the fluorescence data

To take the change in the measured spectra into account, rather than the absolute quantity, the 120 ex/em pairs were standardized along the time domain. This was done for each observation prior to modeling using the MATLAB function *zscore*.

2.3.2 | Time alignment

The on-line data set used for training (values available every 3 min), consisted of 125 variables and 11126 observations and was time-aligned to the respective sampling points of the single target variable (values available every hour), consisting of 690 observations (12 to 25 per fermentation).

2.4 | Exploratory data analysis of the fluorescence data using PCA and PARAFAC

PCA and PARAFAC, as described by Bro [28], were used on the complete fluorescence data set to gain more specific insights into the data and the underlying structures. First, a PCA was performed on the fluorescence data, to unveil the latent structures that explain most of the variance in the data. To determine the location of the underlying fluorescent compounds in the spectrum, PARAFAC was also applied. PARAFAC, unlike PCA, decomposes the fluorescence matrices not only into scores and loadings but also into a third dimension, resulting in three different modes. In the case of the fluorescence data, the first mode represents the sample and is directly proportional to its concentration. The second mode represents the excitation and the third mode represents the emission wavelength of the respective analyte. By joining the second and third modes, the location of the respective factor in the 2D-fluorescence spectrum is displayed. Thus, PARAFAC overcomes the rotational freedom of PCA, making it a better choice for the analysis of fluorescence spectra.

2.5 | Model development

For model training, all fermentations were used. In total, three different models were developed: one using the five available on-line process variables mentioned above, one with only the fluorescence data, and one with both types of data merged. The best input to the model with respect to accurate biomass estimation for internal validation, was used as the final model. The established single-response models were based on the MARS algorithm. This algorithm is well suited to regression modeling of high-dimensional data. It is flexible and based on the expansion of spline basis functions as described by [29]. The model-building comprises two phases, the forward selection followed by the backward deletion of input variables. Detailed information about the MARS algorithm, the workflow for building the MARS model, the basic functions included in the final model and exemplary trajectories of the used inputs, are provided in the Supporting Information Figure S3.

2.5.1 | Relative input variable importance

The importance of the input variables (VIP) was assessed for subsequent use. The VIP is defined by the square root of the generalized cross-validation of the MARS model excluding that variable (still including all basis functions), minus the square root of the corresponding full MARS model's generalized cross-validation. For ease of interpretation, all relative VIPs were scaled in such a way that the most important variable possessed a value of 100.

2.5.2 | Model performance criteria

To guarantee that the developed models possess optimal generalization capabilities, various performance criteria were considered. To validate the models, the root mean squared error (RMSE) (Eq. (1)) and the percentage model error (Eq. (2)) were computed, together with the number of observations (N), the respective value of the biomass concentration (y),the index (i = 1:N) and its estimated counterpart ($\hat{y}$).

$$RMSE = \sqrt{\frac{1}{N} * \sum \left(y_{(i)} - \hat{y}_{(i)}\right)^2} \tag{1}$$

$$Error\ [\%] = \frac{100}{N} * \sum \frac{\left|y_{(i)} - \hat{y}_{(i)}\right|}{y_{(i)}} \tag{2}$$

The SD in Eq. (3) was calculated using the measured value (y), the mean value (y_{mean}) and the number of observations (N) for each time point (t).

$$SD_{(t)} = \sqrt{\frac{1}{N} * \sum \left(y_{(t)} - y_{\text{mean}(t)}\right)^2} \tag{3}$$

The confidence band is provided by calculating the upper and lower 95% confidence interval (CI) in Eq. (4) for each value (y) and the respective SD for each time point (t).

$$95\ \%\ CI_{(t)} = y_{(t)} \pm 1.96\ \times SD_{(t)} \tag{4}$$

2.5.3 | Model validation

Two internal validations were performed. The five-fold cross-validation (5x-CV) in which a random 20% of the data were not considered for the model-building, was used to test the performance. This procedure was repeated four more times until every observation was used for model validation. The second validation used the leave-one-batch-out cross-validation (LoBo-CV) method. For the LoBo-CV, there was always a complete fermentation that was not considered for

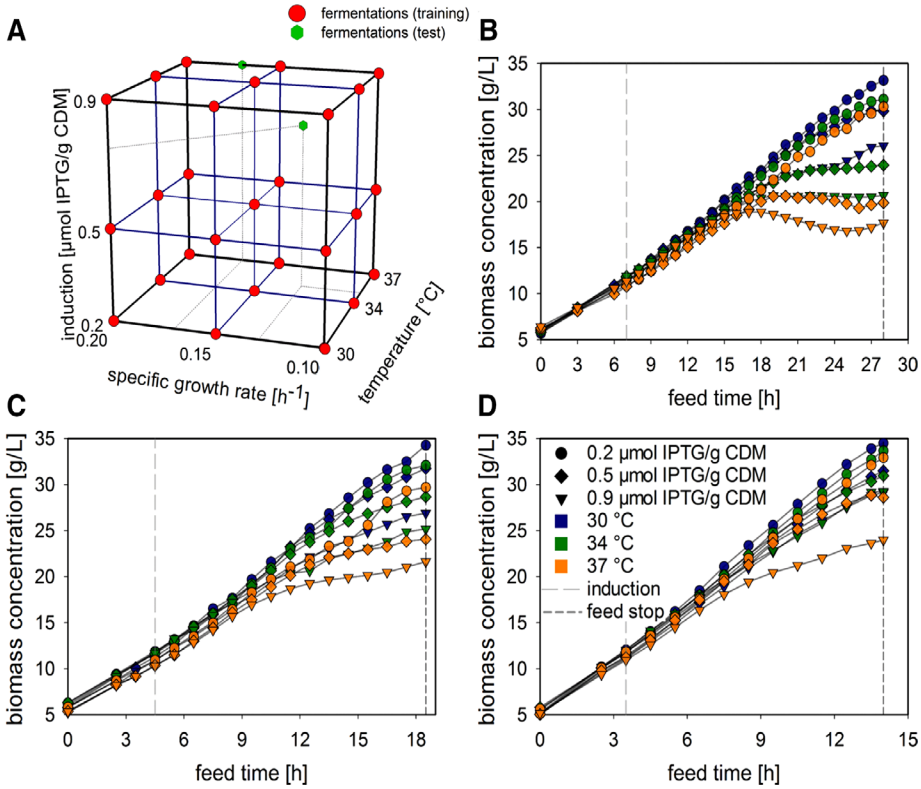

FIGURE 1 Experimental biomass results of the investigated design space. (A) DoE for the three factors (red), and test fermentations (green). (B–D): Biomass trends as a function of the SGR for slow (B), medium (C) and fast (D) growth. The induction ratios are presented with different symbols, that is, 0.2 (dot), 0.5 (square), and 0.9 (triangle), while the varying temperatures are displayed in different colors (30°C in blue, 34°C in green, and 37°C in orange)

training, and the model, which was built on all the other fermentations was validated on this particular fermentation. Again, this procedure was repeated until each fermentation had been used once as a validation set. The performance of the three established models regarding the internal validation was used as the quality criterion for choosing the best input for the final model. The final model was applied to a test set (external validation) to investigate how it performed on new data. The external validation consisted of the two different fermentation settings, as described previously, which had not been used for validation.

3 | RESULTS

3.1 | Experimental biomass results

Biomass trends of the 27 characterized DoE points (Figure 1A) are presented, separated into the three SGRs ($\mu = 0.10$, $\mu = 0.15$, and $\mu = 0.20$) (Figure 1B–D). The biomass concentration trajectory shows the variation as a function of each CPP combination, providing an insight into the challenge presented to the soft sensor. The respective time points of induction (after one doubling time) and the feed stop (after four doubling times in total), are given.

A distinctive tendency towards higher biomass concentrations is visibly associated with lower induction and lower temperature settings, which were uniform for all SGR settings. For $\mu = 0.10$, the maximum difference between the settings is 15.5 g/L (Figure 1B, ranging from 17.7 to 33.2 g/L). An increase in temperature or induction subsequently causes lower biomass concentrations over the whole fermentation. This effect is diminished by increasing the SGR. For $\mu = 0.15$, the maximum difference is 12.7 g/L (Figure 1C, 21.6 to 34.3 g/L) and for $\mu = 0.20$ it is only 10.5 g/L (Figure 1D, 24 to 34.5 g/L). The CPP combinations for the test set were also located in regions where high impacts on the biomass are reported. Therefore, it can be assumed that they will be quite challenging for the soft sensor to estimate, producing a suitable quality criterion for external validation.

3.2 | Exploratory data analysis of the 2D-fluorescence spectra

To gain deeper process understanding, unsupervised learning was performed, and the measured data derived from the advanced on-line probe were inspected. A PCA of the 2D-fluorescence data set revealed that three to four principal components (PC) describe almost all of the variance in the data. These are PC 1 (58.1%), PC 2 (30.8%), PC 3 (5.8%), and PC 4 (3.4%), as shown in Figure 2A. To determine the ex/em pairs that are accountable for the changes in the spectrum, PARAFAC was performed on the fluorescence data. Two

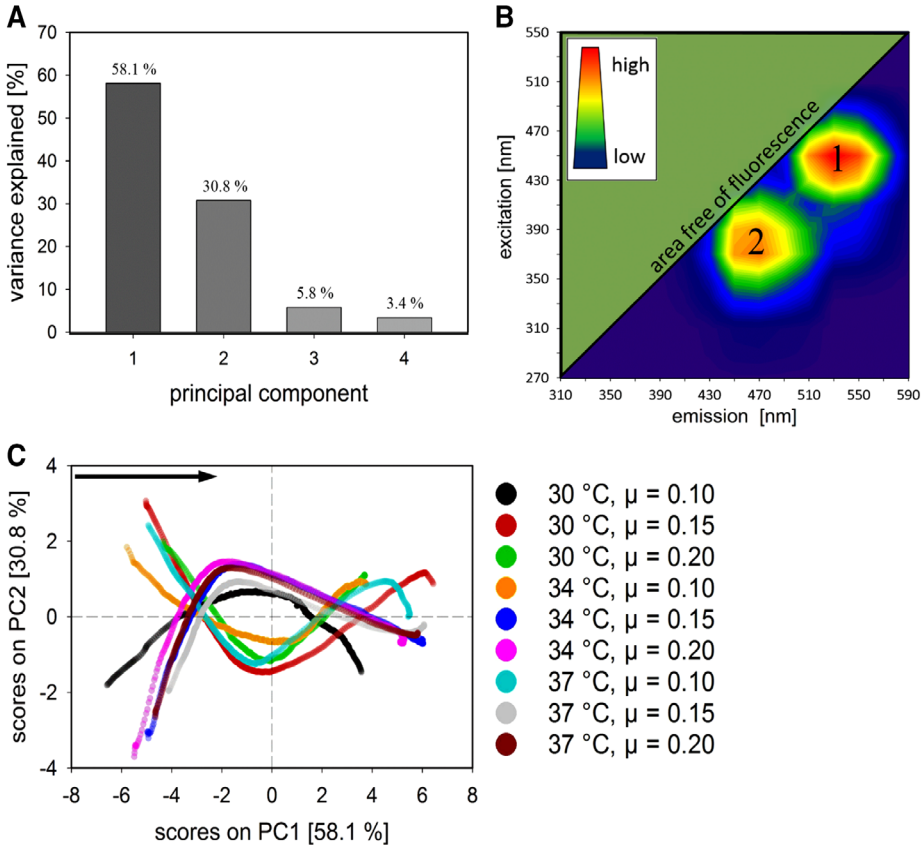

FIGURE 2 Results of the exploratory data analysis of the 2D-fluorescence spectra. (A) Variance explained by principal components for the fluorescence spectra. (B) The location of the compounds ex450/em530 (factor 1) and ex370/em470 (factor 2) determined using PARAFAC in the 2D-fluorescence matrix. The fluorescence-free area and the color scale representing the intensity from dark blue (lowest value) to red (highest value), are shown. (C) Scatter plot of the scores for PC 1 versus PC 2 for all CPP combinations carried out at an induction ratio of 0.9. The direction (black arrow) and the different CPP settings (color scale) are indicated

main factors were identified in the 2D-fluorescence spectrum, as shown in Figure 2B, namely, ex450/em530 (factor 1) and ex370/em470 (factor 2). These factors correspond to underlying fluorescent chemical compounds inside the cell and the broth, which provide additional insight for soft sensor building. In the previous findings, it was shown that processes possess the highest variance with respect to biomass trends and endpoint values at an induction ratio of 0.9 (Figure 1). Thus, the PCA scores (PC 1 versus PC 2) for the nine CPP combinations carried out at this ratio are presented in Figure 2C. The different shapes represent different progressions in the fluorescence spectra, caused by the respective CPP combinations. For PC 1, no major difference was found between the fermentations. All scores followed the same course, while the second PC displayed two score groups with contrary trends. All settings at $\mu = 0.10$ displayed unique trajectories (black, orange and turquoise). The courses of the red (30°C and $\mu = 0.15$) and green (30°C and $\mu = 0.20$) trajectories follow the shape of the turquoise one to some extent, but not markedly. It is not surprising that the scores of the blue (34°C and $\mu = 0.15$) and brown (37°C and $\mu = 0.20$) CPP settings are almost identical, since their biomass trends and endpoints (endpoints at 25.2 and 24.0 g/L) are also very similar. Shapes also matching these two are observed for the pink (34°C and $\mu = 0.20$) and grey (37°C and $\mu = 0.15$) scores. Due to comparable process behaviors with respect to the biomass trajectory, similar PCA trends are indicated and their locations in the score plot confirmed these findings. The pink trend (endpoint at 29.3 g/L) is located above the identical blue and brown trends, while the grey trend (endpoint at 21.6 g/L) is below them. This also reflects the biomass concentrations. It can be concluded that different CPP settings lead to varying 2D-fluorescence spectra. By decomposing and investigating these spectra, conclusions about their progress can be made. All these findings strongly suggest that valuable process information is present in the 2D-fluorescence data.

3.3 | Comparison of the input variables for soft sensor development

Subsequently, after the exploratory data analysis of the online data, the optimal data set for model-building was tested using three different types of input. The performances of soft sensors using only process data, using only 2D-fluorescence data and using merged input data (both types) were considered. For the decision-making, the best performance with LoBo-CV (internal validation) was investigated and presented (Figure 3). Four fermentations from the investigated space

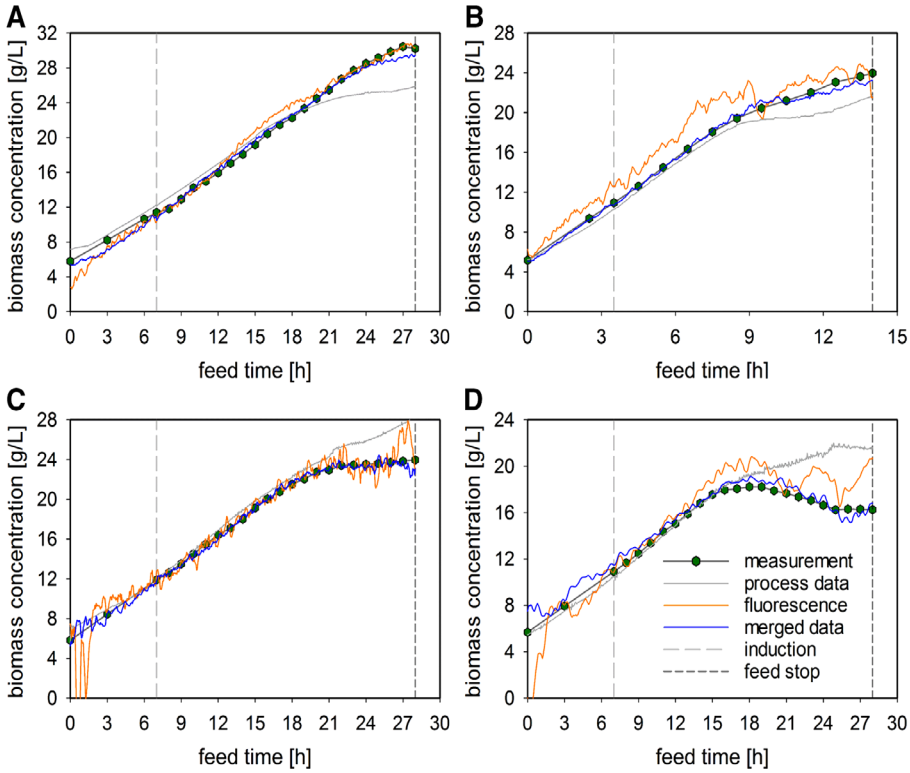

FIGURE 3 Performance comparison for the different inputs to the model. The biomass trends of four fermentations, the respective time point of induction and the feed stop are shown. (A) 30°C, μ = 0.10 and induction ratio = 0.5; (B) 37°C, μ = 0.20 and induction ratio = 0.9; (C) 34°C, μ = 0.10 and induction ratio = 0.9; (D) 37°C, μ = 0.10 and induction ratio = 0.9. The estimations of the established models using the three different inputs (process data (grey), 2D-fluorescence data (orange) and merged data (blue)) applied to the LoBo-CV, are shown

TABLE 1 Performance results of the developed models with respect to R^2 and RMSE (both rounded to two decimal places), indicated as concentrations, and the percentage error (rounded to one decimal place), are presented. The different inputs for building the model are indicated. The results for the external validation (test sets #1 and #2) are only given for the final model (using merged data)

	R^2	RMSE (g/L)	Error (%)
	Process data/ fluorescence/merged	Process data/ fluorescence/merged	Process data/ fluorescence/merged
Training	0.97 / 0.99 / 0.99	1.15 / 0.74 / 0.45	5.7 / 3.7 / 2.2
5x-CV	0.97 / 0.97 / 0.99	1.28 / 1.20 / 0.58	6.3 / 5.9 / 2.9
LoBo-CV	0.96 / 0.92 / 0.98	1.42 / 2.04 / 0.99	7.0 / 10.0 / 4.9
Test #1	- / - / 0.66	- / - / 2.62	- / - / 13.8
Test #2	- / - / 0.98	- / - / 0.69	- / - / 3.8

and the respective model performances are shown. To allow for meaningful comparison and statements, different biomass trends are presented. Two fermentations with consistently increasing concentrations (Figure 3A and B), one reaching a plateau (Figure 3C) and one with a decreasing concentration (Figure 3D), were chosen. The performance criteria (R^2, RMSE, and the percentage error) for the three established models, are provided in Table 1. The estimation of the soft sensor using solely process data (grey) always, without exception, underestimates or overestimates the measured values, with an RMSE of 1.42 g/L (7%). Using 2D-fluorescence data (orange) as input to the model led to visually more reliable models, even though the RMSE was higher, reaching 2.04 g/L (10.0%). This occurs due to peaks and fluctuations in the estimations, which are not observed for the model developed using the process data. Only the established soft sensor using both kinds of data (blue) can accurately estimate every trend, resulting in an RMSE of 0.99 g/L (4.9%). This demonstrates that both data sets possess relevant and complementary information for building a robust model and further highlights the importance and advantage of this advanced sensor type.

3.4 | Soft sensor performance of the final model

Based on the results shown in Figure 3, the soft sensor developed using the merged data set was chosen as the final model. To evaluate the quality of the established soft sensor, its

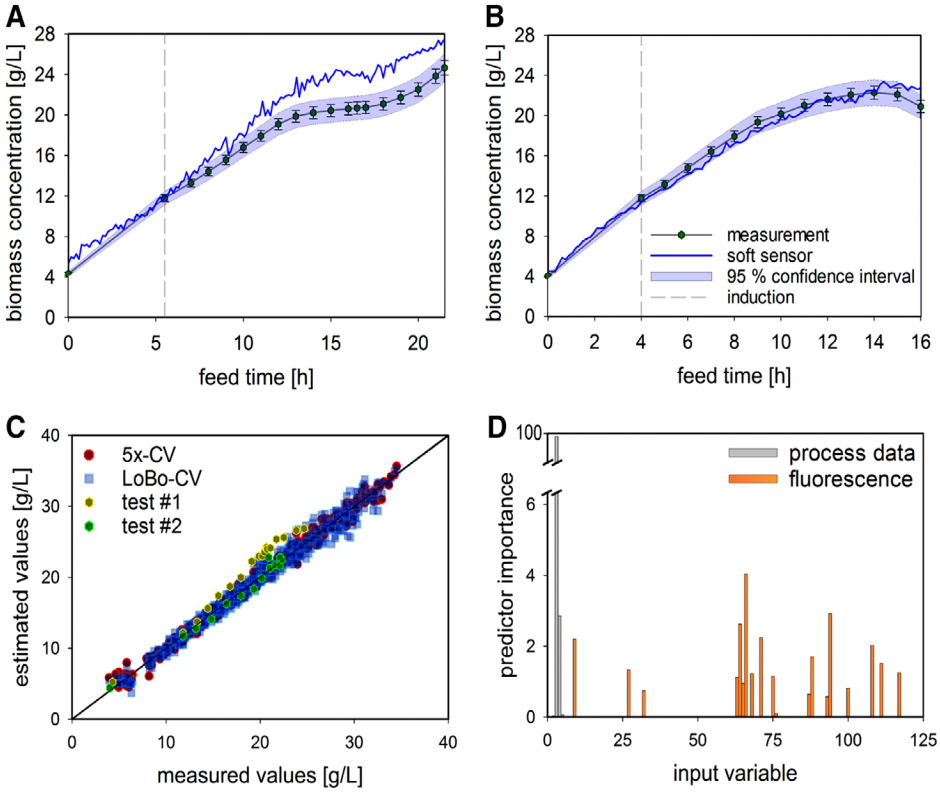

FIGURE 4 Performance of the final model on the external validation data set. (A–B) The biomass trends for the first (A) and the second (B) test set ± SD including the 95% CI and the estimation from the developed soft sensor, are presented. (C) Scatter plot of the model performances on the 5x-CV (red dots), the LoBo-CV (blue squares) and the two test fermentations (yellow and green dots), are shown. (D) The VIP of the process (grey) and 2D-fluorescence data (orange) derived from the MARS algorithm

performance on the test set, consisting of two fermentations with (partly) different CPP settings, was considered, that is, external validation was performed. The model's estimation of biomass concentration for the test fermentation, which was executed using three different CPP settings, shows an overestimation after the first half of the process (Figure 4A) with respect to the off-line measured concentrations. Although the general shape of the trajectory is reproduced, the endpoint is still overestimated at 27 g/L, rather than the analytically measured 24.7 g/L. The estimation of the process with only one different CPP is able to follow the off-line trend and results in a more satisfying endpoint value of 22.7 g/L, rather than 20.9 g/L (Figure 4B). The higher deviation from the measured values for the first test process is also visible in the scatter plot (Figure 4C), while the results for the second process are located within the error magnitude of the two internal cross-validations. Since the training error is negligibly small, at 0.45 g/L (2.2%), these results are not displayed in the scatter plot, for greater clarity. The remaining variance in the test sets might result from a factor that is not considered in the model input or from a function of included factors interpreted in an insufficient way by the model. To evaluate which input variables are important for the estimation, the VIP was determined for the five process parameters as well as for the 120 ex/em pairs. As shown in Figure 4D, only a few inputs are important for building the model. These were therefore retained in the backward deletion phase of model-building and included in the algorithm. The list of all variables with VIP scores above zero, in descending order, is presented in Supporting Information Table S2. The highest importance for the process parameters was given to the accumulated feed (scoring the maximum value of 100) and the accumulated inductor (scoring 2.8). For the 2D-fluorescence data, only 19 of the 120 available variables were taken into account in model-building. The chosen variables are mostly collinear, due to the fact that they are neighbors (ex/em ± 20 nm). These collinear ex/em pairs do not carry extra information but are still included in the model for noise reduction and enhanced robustness. This finally results in only two important ex/em pairs, namely, ex450/em530 (scoring 2.9) and ex370/em470 (scoring 4). These are identical to the two ex/em pairs determined via PARAFAC. The final model performance with respect to the RMSE of the 5x-CV is fairly small, at 0.58 g/L (2.9%), and the LoBo-CV also displays good accuracy with an RMSE of 0.99 g/L (4.9%). The results for the external validation show an RMSE score of 2.62 g/L (13.8%) with three altered CPPs and 0.69 g/L (3.8%) with one altered CPP, highlighting the estimation qualities of the established soft sensor. This is in good accordance with the off-line measurement used as the reference, where an SD of 3.41% was observed. The performance of the final model on the test set is presented in Table 1.

4 | DISCUSSION

The impact of the CPP settings on the biomass is shown in Figure 1. The different concentrations and endpoints result from diverse metabolic burdens, for example, recombinant protein production, stressing the cells. With slow SGR settings, more resources are available for protein synthesis. When the SGR increases, cells focus more on their own growth and neglect protein production. As a result, fewer product is present and stress levels are lower. This results in higher biomass concentrations, even though the other CPP settings, except the SGR, stay the same. The product formation, metabolism and cellular stress levels also increase with higher temperature settings, again leading to decreased biomass values. The maximum impact is seen when considering the opposing corners of the investigated design space, resulting in more than 50% difference, at 16.3 and 34.5 g/L.

The insights into the fluorescence data via PCA and PARAFAC (Figure 2) strongly support the assumption that, in fact, it is possible to monitor intracellular fluorescent substrates, especially the ex/em wavelengths of the two chemical compounds identified via PARAFAC are similar to those from flavins (riboflavin, flavin mononucleotide and flavin adenine dinucleotide) for factor 1 [30] and nicotinamide adenine dinucleotide phosphate for factor 2 [31]. These molecules are directly linked to cell physiology and important metabolic pathways. Flavins are overproduced during exponential growth and act as electron carriers [32]. Similarly, nicotinamide adenine dinucleotide phosphate is a major component of the electron transfer chain [33]. It is conceivable that different DoE settings result in diverse concentrations and consumption rates. Since metabolic activities are temperature-dependent, it is indicated that these changes and deviations in cell physiology caused by different CPP combinations are measured by the 2D-fluorescence probe. This is also hinted at by the two observable score groups of PC 2. They are caused by two different observed ex/em wavelength clusters in the loadings of PC 2. One group contained a cluster with wavelengths ex450/em510-550 and the second group contained a cluster with wavelengths ex510-530/em550-590. These show different trends over the fermentation and cause the opposing trend in the PCA score plot. However, more investigation into the cell is required in the future, for example, taking cell lysis into account or deliberately provoking metabolic shifts and measuring the response in the fluorescence spectra.

In addition, the added value and advantage of using a 2D-fluorescence probe for on-line biomass estimation is demonstrated across CPP settings, and also for fermentations with altered CPP settings. With the process data alone, only discrete and accumulating/rising values are introduced into the model. Thus, fermentations with steady or decreasing concentrations are especially difficult to estimate, as previously shown (Figure 3). Using 2D-fluorescence in addition allowed the cell physiology and metabolism to be examined and the potential underlying chemical compounds to be identified.

The test fermentation with three altered process settings (Figure 4A) led to completely new metabolic patterns for which the MARS model was not trained, and therefore resulted in a high residual value. To overcome these boundaries, other CPP levels could be considered and additional sensors could be utilized. However, these approaches would need to be accompanied by several additional experiments. To avoid this time-consuming step most simply, a mechanistic part (white box) can be taken into account to describe this missing term. This exploitation of both model advantages is called hybrid modeling, and has already been reported elsewhere [34]. Potentially, with this added value, more challenging processes can also be monitored on-line in the future, such as the so-called intensified DoE. Hence, through intra-experimental set-point changes, the dynamics of the specified design space can be captured [35].

MARS proved to be a suitable algorithm for soft sensor development (Figure 4). Its characteristics, for example, its ability to handle nonlinearity and multicollinearity, make it an ideal candidate for working with this complex input data and creating meaningful models. Its VIP also determined the importance of two particular ex/em pairs for accurate biomass estimation. Moreover, these were identical to the factors identified by PARAFAC. As discussed above, these wavelengths probably represent chemical compounds that are representative of the current biomass state. It is comprehensible that the highest VIP for biomass estimation is possessed by the amount of added feed medium (controlling the SGR). However, the two ex/em variables seem to be responsible for fine-tuning the precise biomass estimation by the soft sensor, taking the various metabolic burdens into account. This enables precise on-line monitoring of the biomass with real-time availability of the current value, which can be exploited in the QbD concept. All these findings demonstrate that the established soft sensor is a valuable PAT tool.

The study did not contain experiments using animals or human subjects.

ACKNOWLEDGMENT

We would like to thank the Austrian Research Promotion Agency (FFG) for their support (Research Studio Austria, 859219).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Benjamin Bayer https://orcid.org/0000-0001-5241-4924

REFERENCES

1. Rosano, G. L. and Ceccarelli, E. A., Recombinant protein expression in Escherichia coli: Advances and challenges. *Front. Microbiol.* 2014, *5*, 1–17.
2. Gnoth, S., Jenzsch, M., Simutis, R. and Lübbert, A., Control of cultivation processes for recombinant protein production: A review. *Bioprocess Biosyst. Eng.* 2008, *31*, 21–39.
3. FDA, Guidance for industry. PAT — A framework for innovative pharmaceutical development, manufacturing, and quality assurance. 2004.
4. Rantanen, J. and Khinast, J., The future of pharmaceutical manufacturing sciences. *J. Pharm. Sci.* 2015, *104*, 3612–3638.
5. Patil, A. S. and Pethe, A. M., Quality by design (QbD): A new concept for development of quality pharmaceuticals. *Int. J. Pharm. Qual. Assur.* 2013, *4*, 13–19.
6. Zhang, L. and Mao, S., Application of quality by design in the current drug development. *Asian J. Pharm. Sci.* 2017, *12*, 1–8.
7. Glassey, J., Gernaey, K. V., Clemens, C., Schulz, T. W., et al., Process analytical technology (PAT) for biopharmaceuticals. *Biotechnol. J.* 2011, *6*, 369–377.
8. Joeris, K., Frerichs, J.-G., Konstantinov, K. and Scheper, T., *In-situ* microscopy: Online process monitoring of mammalian cell cultures. *Cytotechnology* 2002, *38*, 129–134.
9. Lee, H. L. T., Boccazzi, P., Gorret, N., Ram, R. J. and Sinskey, A. J., In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy. *Vib. Spectrosc.* 2004, *35*, 131–137.
10. Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A. and Gernaey, K. V., Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnol. Prog.* 2009, *25*, 1561–1581.
11. Faassen, S. M. and Hitzmann, B., Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. *Sensors (Switzerland)* 2015, *15*, 10271–10291.
12. Assawajaruwan, S., Eckard, P. and Hitzmann, B., On-line monitoring of relevant fl uorophores of yeast cultivations due to glucose addition during the diauxic growth. *Process Biochem.* 2017, *58*, 51–59.
13. Teixeira, A. P., Oliveira, R., Alves, P. M. and Carrondo, M. J. T., Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnol. Adv.* 2009, *27*, 726–732.
14. Mandenius, C. F. and Gustavsson, R., Mini-review: Soft sensors as means for PAT in the manufacture of bio-therapeutics. *J. Chem. Technol. Biotechnol.* 2015, *90*, 215–227.
15. Skibsted, E., Lindemann, C., Roca, C. and Olsson, L., On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration. *J. Biotechnol.* 2001, *88*, 47–57.
16. Harshman R. A. and Lundy, M. E., PARAFAC: Parallel factor analysis. *Comput. Stat. Data Anal.* 1994, *18*, 39–72.
17. Shlens, J., A tutorial on principal component analysis. *Internet Artic.* 2005, 1–13.
18. Mercier, S. M., Diepenbroek, B., Dalm, M. C. F., Wijffels, R. H. and Streefland, M., Multivariate data analysis as a PAT tool for early bioprocess development data. *J. Biotechnol.* 2013, *167*, 262–270.
19. Kadlec, P., Gabrys, B. and Strandt, S., Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* 2009, *33*, 795–814.
20. Friedman, J., Multivariate adaptive regression splines. *Ann. Stat.* 1991, *2*, 1152–1174.
21. Curtis, A., Smith, T., Ziganshin, B. and Elefteriades, J., The mystery of the Z-score. *Aorta* 2016, *4*, 124–130.
22. Rathore, A. S., Mittal, S., Pathak, M. and Mahalingam, V., Chemometrics application in biotech processes: Assessing comparability across processes and scales. *J. Chem. Technol. Biotechnol.* 2014, *89*, 1311–1316.
23. Melcher, M., Scharl, T., Spangl, B., Luchner, M. et al., The potential of random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed-batch fermentations. *Biotechnol. J.* 2015, *10*, 1770–1782.
24. Luchner, M., Striedner, G., Cserjan-Puschmann, M., Strobl, F. and Bayer, K., Online prediction of product titer and solubility of recombinant proteins in Escherichia coli fed-batch cultivations. *J. Chem. Technol. Biotechnol.* 2015, *90*, 283–290.
25. Jekabsons, G., Adaptive regression splines toolbox for MATLAB/octave. 2016, 1–33.
26. Andersson, C. A. and Bro, R., The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.* 2000, *52*, 1–4.
27. Murphy, K. R., Stedmon, C. A., Graeber, D. and Bro, R., Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal. Meth.* 2013, *5*, 6557.
28. Bro, R., PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* 1997, *38*, 149–171.
29. Friedman, J. H. and Roosen, C. B., An introduction to multivariate adaptive regression splines. *Stat. Meth. Med. Res.* 1995, *4*, 197–217.
30. Mukherjee, A., Walker, J., Weyant, K. B. and Schroeder, C. M., Characterization of flavin-based fluorescent proteins: An emerging class of fluorescent reporters. *PLoS One* 2013, *8*, 1–15.
31. Li, J.-K. and Humphrey, A. E., Use of fluorometry for monitoring and control of a bioreactor. *Biotechnol. Bioeng.* 1991, *37*, 1043–1049.
32. McAnulty, M. J. and Wood, T. K., Yeeo from Escherichia coli exports flavins. *Bioeng. Bugs* 2014, *5*, 386–392.
33. McIver, L., Leadbeater, C., Campopiano, D. J., Baxter, R. L. et al., Characterisation of flavodoxin NADP+oxidoreductase and flavodoxin; key components of electron transfer in Escherichia coli. *Eur. J. Biochem.* 1998, *257*, 577–585.
34. Von Stosch, M., Oliveira, R., Peres, J. and Feyo de Azevedo, S., Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Comput. Chem. Eng.* 2014, *60*, 86–101.
35. Von Stosch, M. and Willis, M. J., Intensified design of experiments for upstream bioreactors. *Eng. Life Sci.* 2016, 1–9.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Bayer B, von Stosch M, Melcher M, Duerkop M, Striedner G. Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in *Escherichia coli* cultivations. *Eng Life Sci*. 2020;20:26–35. https://doi.org/10.1002/elsc.201900076

Publication II

Supporting Information

Supporting Information

All the critical process parameter settings, presented in the Materials & Methods section of the main manuscript (2.2), investigated in the design space for model building and internal validation, as well as the settings for the external validation (test sets), are presented in Table S1.

Table S1. All investigated CPP combinations of the design space, namely, specific growth rate, induction strength and cultivation temperature (each with three levels) used as training are listed as well as the settings for the test set. If more than one fermentation was performed, the number of repetitions is indicated.

CPP setting	temperature [°C]	specific growth rate [h^{-1}]	induction ratio [µmol IPTG/g cell dry mass]
1	30	0.1	0.2
2	30	0.15	0.2
3	30	0.2	0.2
4	34	0.1	0.2
5	34	0.15	0.2
6	34	0.2	0.2
7	37	0.1	0.2
8	37	0.15	0.2
9	37	0.2	0.2
10 (N = 2)	30	0.1	0.5
11	30	0.15	0.5
12 (N = 2)	30	0.2	0.5
13	34	0.1	0.5
14	34	0.15	0.5
15	34	0.2	0.5
16	37	0.1	0.5
17	37	0.15	0.5
18	37	0.2	0.5
19	30	0.1	0.9
20	30	0.15	0.9
21	30	0.2	0.9
22	34	0.1	0.9
23	34	0.15	0.9
24	34	0.2	0.9
25 (N = 5)	37	0.1	0.9
26	37	0.15	0.9
27	37	0.2	0.9
28 (test set #1)	35	0.13	0.75
29 (test set #2)	37	0.17	0.9

The relative importance of each input variable for the multivariate adaptive regression spline model-building, presented in the Results section of the main manuscript (3.4) in Fig. 4D, is listed in descending order in Table S2. Only variables that were used for building the final model are listed, i.e., scoring a VIP above zero.

Table S2. List of all VIP scores (above zero) of the final model in descending order.

rank	input variable	VIP score
1	accumulated feed	100.0
2	ex370/em470	4.0
3	ex450/em530	2.9
4	accumulated inductor	2.8
5	ex350/em470	2.6
6	ex370/em410	2.2
7	ex270/em370	2.2
8	ex450/em550	2.0
9	ex390/em570	1.7
10	ex470/em510	1.5
11	ex290/em450	1.3
12	ex490/em550	1.3
13	ex350/em550	1.2
14	ex370/em490	1.2
15	ex350/em450	1.1
16	ex350/em490	1.0
17	ex430/em510	0.8
18	accumulated base	0.8
19	ex290/em550	0.8
20	ex390/em550	0.6
21	ex410/em510	0.6
22	temperature	0.4
23	ex370/em510	0.1

The reactor volumes of the presented *E. coli* fed-batch cultivations in the Materials & Methods section of the main manuscript (2.1) utilizing an exponential feeding rate profile are presented (Fig. S1). The batch medium was calculated to produce 22.5 g biomass. This value was used as specific growth rate setpoint to calculate the exponential feeding strategy to constantly provide the respective set specific growth rate during the whole feeding phase. Volume differences are observable due to batch-to-batch variations and different base consumption patterns due to the varying critical process parameter combinations.

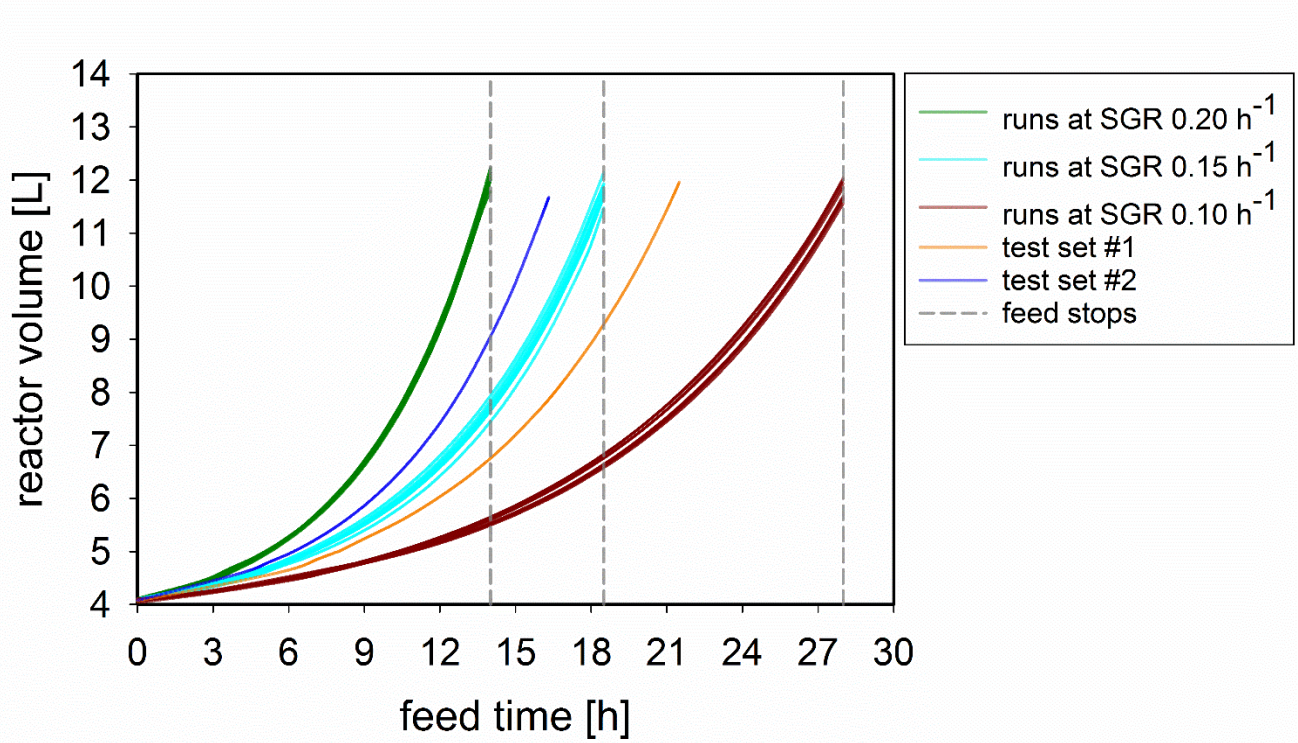

Figure S1. Reactor volumes of all DoE fed-batch fermentations applying an exponential feeding strategy. The reactor volumes as a function of the SGR, slow (dark red), medium (cyan) and fast (dark green), are shown for every CPP setting of the DoE study. The time of the respective feed stop (14, 18.5 and 28 h) is indicated (dashed grey lines). Test set #1 (orange) and test set#2 (blue) are displayed, without the respective feed stops (21.5 and 16 h).

The applied workflow of the soft sensor development, using MATLAB 2016b and additionally the freely available toolbox packages, described in the Materials & Methods section of the main manuscript (2.2), is presented in the simplified form of a graphical overview (Fig. S2).

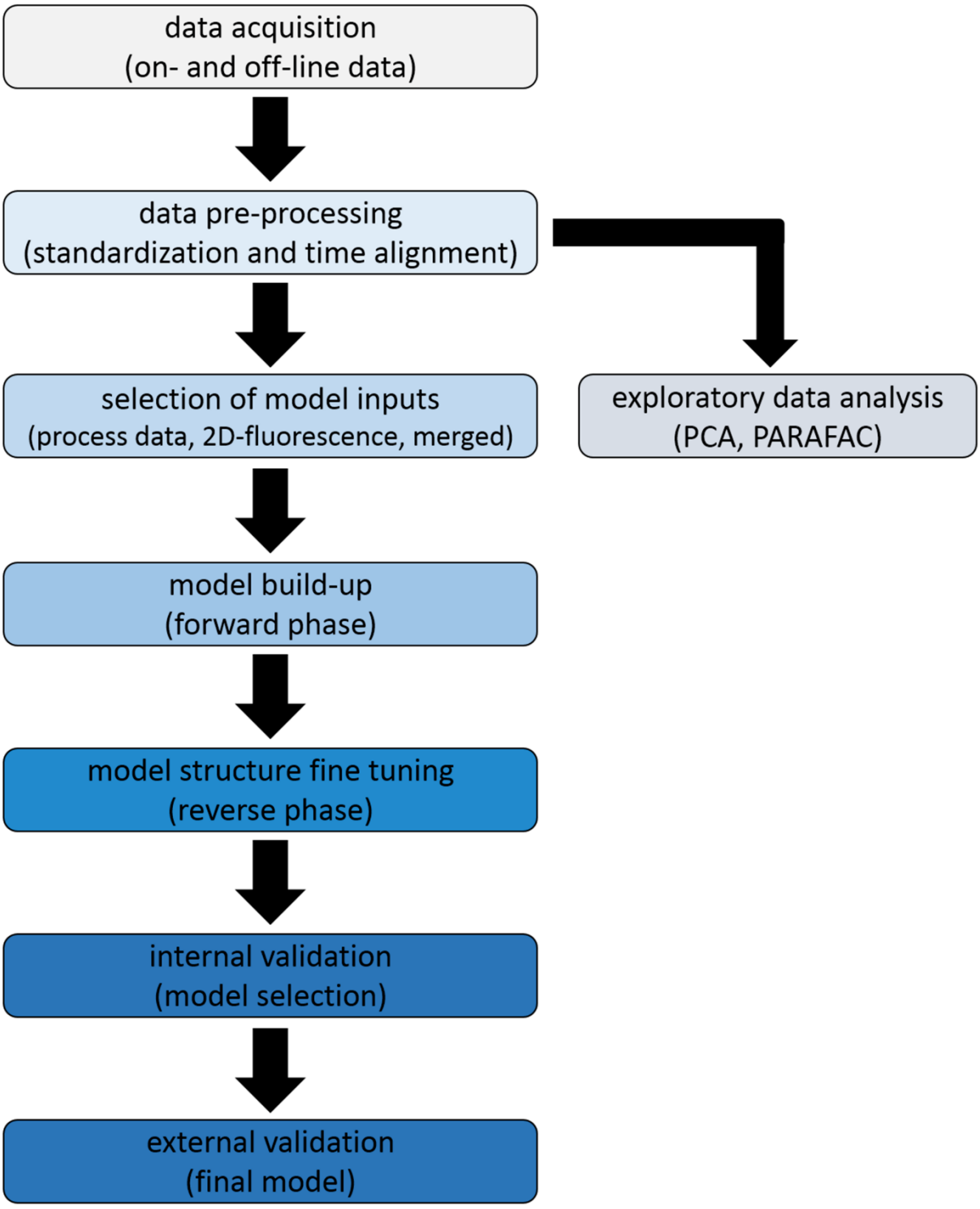

Figure S2. The schematic stepwise workflow of the model development.

The MARS algorithm used for building the soft sensor is considered as a flexible tool for regression modeling. Its strengths, compared to other algorithms used for regression modeling, are higher flexibility to capture relationships and interactions between input variables. This makes it an ideal candidate for handling high dimensional data with additive and (multi)collinearity characteristics, as it is the case for the 2D-fluorescence data. This dynamic adaptability is enabled by the selection of a subset of local variables. These are depending on distinctive conditions to fit the target value.

In general, the MARS algorithm can be seen as a specialization of a general multivariate regression and an expansion to spline basis functions, in which the number of basis functions is automatically limited by the number of inputs. The MARS algorithm builds hierarchical models in a two-step procedure; forward and backward. Starting with a set of basis functions, by stepwise selection a subset of these is chosen, which are suitable for modeling the target variable.

In the first phase (forward selection procedure), the model comprises only the intercept term and more basis functions are added iteratively to consecutively reduce the training error. This phase is executed until any of the conditions to stop are met, e.g., the number of coefficients equals the number of observations, new basis functions do not change the R^2 above the set threshold or the R^2 reached 1. The result of the forward phase is a large model overfitting the data. This general structure of the main model is given in Eq. 1. The estimated value ($\hat{y}$) of the model, the intercept *(b),* a basis function *(BF)* and its respective covariate vector (*v)* for the number of used functions (*i)* is given.

$$\hat{y} = b + v_{(1)} * BF_{(1)} + \ldots + v_{(i)} * BF_{(i)} \quad (1)$$

Each truncated cubic basis function comprises one of the input variables and three respective knot locations (a central knot, a lower and an upper side knot) to handle the local conditions. After completing the first phase, the, in this study, developed overfitting model comprised 65 of these basis functions.

To optimize the established overfitting model, the second phase (backward deletion procedure) is executed to simplify and generalize the model. This is done by stepwise deleting the least important basis function (smallest reduction of the training error) until the model again only consists of the

intercept term. For every reduced model, the generalized cross-validation (GCV) is determined (Eq. 2). The reduced model for which the lowest GCV, with optimal performance on validation data, is obtained, is selected as the final model. The GCV is calculated using the models mean squared error (MSE_{train}), the number of observations (N) and the effective number of parameters (n_p) in the model.

$$GCV = MSE_{train} / \left(1 - \frac{n\_p}{N}\right)^2 \quad (2)$$

The number of initially used input variables was reduced from 125 to 23 (Table S2), which were applied in the basis functions. The number of basis functions remaining in the final model is pruned to 42, including the intercept term. The number of final basis functions exceeds the number of finally used input variables, which means that some input variables were assigned multiple times to the remaining basis functions. The obtained final model was applied to the test set fermentations (external validation) to demonstrate its performance on new data, which had not been used for validation.

The input variables used in the final MARS model (see Table S2) are shown in Fig. S3. The fed-batch fermentation performed at CPP setting 25 (see Table S1) is shown as an example to demonstrate their respective trajectory during the fed-batch fermentation. The on-line available process variables in Fig. S3A present an exponential (feed, base and inductor, according to the feeding profile) or constant (temperature, setpoint at 37°C) trend. Otherwise the ex/em wavelength pairs in Fig. S3B consist of diverse shapes, e.g., (multi)collinear, nonlinear and constant trends. For an easier comparison and a simplified visual inspection of the 2D-fluorescence data, each ex/em wavelength pair was scaled from zero to one. The two ex/em wavelength pairs identified by PARAFAC and MARS are highlighted.

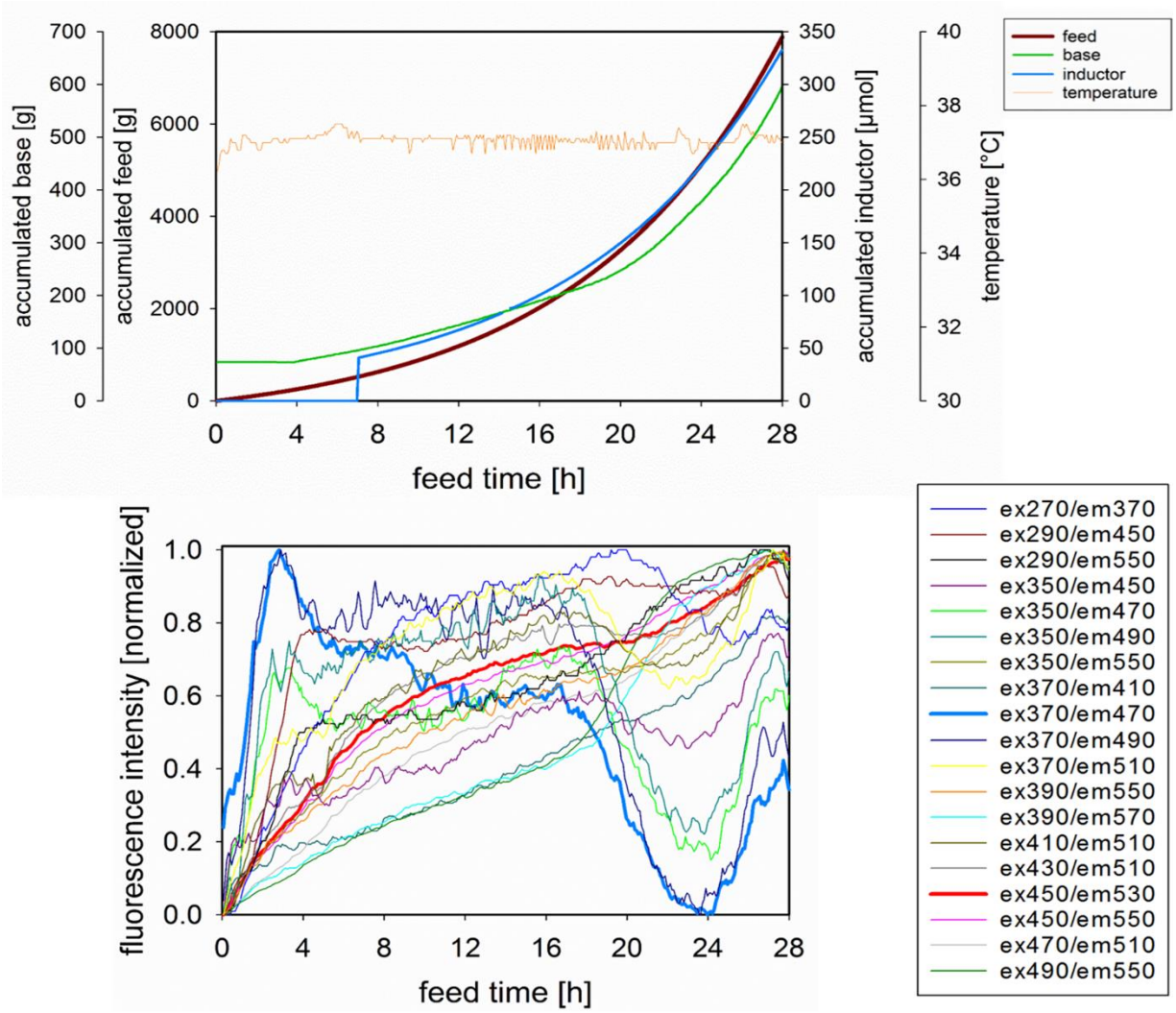

Figure S3. Trajectories of the input variables over the whole fed-batch fermentation at CPP setting 25. A: The on-line available process variables used as input to the MARS model, namely, the feed (dark red), base (green), inductor (blue) and temperature (orange), are displayed. B: The 19 ex/em wavelength pairs used as input to the MARS model are displayed. Each ex/em wavelength pair was scaled from zero to one. The two ex/em wavelength pairs, identified both by PARAFAC and MARS, are highlighted (ex370/em470 in blue and ex450/em530 in red).

Publication III

Biotechnology Journal
www.biotechnology-journal.com

Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization

*Benjamin Bayer, Moritz von Stosch, Gerald Striedner, and Mark Duerkop**

Upstream bioprocess characterization and optimization are time and resource-intensive tasks. Regularly in the biopharmaceutical industry, statistical design of experiments (DoE) in combination with response surface models (RSMs) are used, neglecting the process trajectories and dynamics. Generating process understanding with time-resolved, dynamic process models allows to understand the impact of temporal deviations, production dynamics, and provides a better understanding of the process variations that stem from the biological subsystem. The authors propose to use DoE studies in combination with hybrid modeling for process characterization. This approach is showcased on *Escherichia coli* fed-batch cultivations at the 20L scale, evaluating the impact of three critical process parameters. The performance of a hybrid model is compared to a pure data-driven model and the widely adopted RSM of the process endpoints. Further, the performance of the time-resolved models to simultaneously predict biomass and titer is evaluated. The superior behavior of the hybrid model compared to the pure black-box approaches for process characterization is presented. The evaluation considers important criteria, such as the prediction accuracy of the biomass and titer endpoints as well as the time-resolved trajectories. This showcases the high potential of hybrid models for soft-sensing and model predictive control.

1. Introduction

1.1. Quality by Design and Process Characterization

The U.S. Food and Drug Administration (FDA) proposed the process analytical technology (PAT) guidelines in 2004[1] to foster the application of science and risk-based technologies in the biopharmaceutical industry. Answering to the guidelines, industry adopts more and more a Quality by Design (QbD) paradigm to gain deeper process insights and to counteract batch-to-batch variability (e.g., fluctuations in the cultivation temperature and temporary pump malfunctions).[2] At the beginning of QbD implementation, a technical risk assessment is used to identify critical quality attributes (CQA) of the respective product and thereupon the critical process parameters (CPP) that might affect the CQAs.[3] Once the CPPs are known, process characterization studies are performed to assess and understand their impact on the CQAs. These characterization studies nowadays mostly adopt a combination of statistical design of experiments (DoE) with response surface modeling (RSM), in particular for upstream bioprocess characterization.[4–6]

1.2. Design of Experiments

The basic purpose of DoEs is to systematically investigate the relationship between defined CPPs (input factors) and the process response of interest, for example, the final product concentration. Different levels of each CPP (e.g., different pH set points or cultivation temperatures) are typically examined to have a better understanding of the respective impacts.[7] The multidimensional combination of the identified and chosen CPPs, each with a selected number of investigated levels, set up the space to be investigated. Herein, the space in which the desired product quality is achieved is referred to as the design space.[8] Depending on the specific aim of the study different designs can be applied, by performing cultivations in this space, characterizing it, and gaining process knowledge, for example, an initial screening design, investigating only the corners of the design space, or a full factorial design for process characterization and optimization. For the first use, many reduced designs can be applied, for example,

B. Bayer, Prof. G. Striedner, Dr. M. Duerkop
Department of Biotechnology
University of Natural Resources and Life Sciences
Vienna 1190, Austria
E-mail: mark.duerkop@novasign.solutions
Dr. M. von Stosch
School of Chemical Engineering and Advanced Materials
Newcastle University
Newcastle upon Tyne NE1 7RU, UK
Dr. M. Duerkop
Novasign GmbH
Vienna 1190, Austria

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/biot.201900551

DOI: 10.1002/biot.201900551

definite screening, fractional factorial or Box-Behnken designs to name a few. All of these designs require a different number of experiments and generate different amounts of information.[9] As a result, the area of interest can be identified in which the original DoE can be extended to derive robust process settings and to describe their aftermath.[10] However, to completely characterize this space, the number of experiments suggested by DoEs can easily exceed a feasible number of experiments, if too many factors or levels are chosen or the experiment's inherent complexity is very high. This is typically accompanied by long durations of the studies and tremendous expenses.[11]

1.3. Response Surface Modeling

In process development and optimization settings, typically RSM are adopted to analyze the generated analytical results,[12] for example, using one out of the many available statistical software tools. Most often solely the impact of the factors on the endpoints of the processes is investigated.[13,14] This only leads to a snapshot of the end of the process, since the rest of the process is neglected, which can lead to significantly different conclusions than if the complete process profile would be considered. For instance, RSM are utilized to investigate the relationship between the CPPs and one or more process variables of interest, for example, the biomass and product titer.[15,16] However, when using such models for endpoint descriptions, many temporary process influences are not taken into account. Also, process dynamics are neglected, wherefore the aftermath of process deviations is not understood and the impact of changes in the CPPs during the process cannot be described.[17]

1.4. Process Modeling

More generally, there are two ways of modeling, called parametric (white box) and non-parametric (black box) modeling.[18] Non-parametric models refer to purely data-driven approaches, for which no further process knowledge is needed, the model structure is inferred from data, for example, artificial neural networks (ANN). However, once the model is used for predictions out of the characterized space, it lacks the ability to extrapolate. The structure[19] of a parametric model is based on knowledge and empirical considerations and therefore is a more suitable match for extrapolation. Since this type of model solely assumes a mechanistic trend, and does not account for empiric observations, the predictability is also often inaccurate. These parametric models are rarely used in upstream process development since their development is time consuming and laborious.[20,21] Despite their shortcomings, both approaches are already in use to model and predict processes.[22–26]

1.5. Hybrid Modeling

One way to exploit the advantages of both modeling approaches is called hybrid (semi-parametric) modeling. Hereby, it is possible to benefit from the positive aspects and make up for the respective drawbacks.[27] Such a model can either combine the black- and white-box parts in a parallel or serial structure. The difference is the order in which the different parts and their respective weights are taken into account in the overall model. These features generally make it a more cost-effective modeling approach to deal with complex problems[28]; for example, the black box can be used to estimate the rate expressions used in the white box.[29] In contrast to simple black-box endpoint models, certain hybrid model structures are able to describe the entire process and not only the endpoints. As a result, the behavior and deviations of a process can be understood and therefore it is possible to understand the impact of changing the CPPs during the process.[30]

1.6. Bridging the Research Gap

Many methods, approaches, and tools were developed since the start of the PAT initiative, to implement QbD in upstream process development, characterization, and optimization.[31] Current methods to evaluate the results obtained from DoE, for example, RSM, solely focus on the endpoint of a process, neglecting the remaining cultivation, missing the majority of the process information. The main objective of this study was to show how this gap can be closed in the biopharmaceutical industry by utilizing hybrid modeling, with respect to process understanding and compared to the performance of the most commonly adopted techniques.

We utilized experimental data from *Escherichia coli* fed-batch cultivations at the 20L scale expressing the recombinant protein human Cu/Zn superoxide dismutase (hSOD). A space with three factors, each with three levels, was characterized by performing fed-batch fermentations at each CPP combination setting, leading to 27 distinctive DoE conditions. With this generated fed-batch fermentation data, the best structure and input to the three modeling approaches were searched. Consecutively, the performance of the hybrid model was compared to the widely adopted RSM methodology and a solely data-driven ANN model.

2. Experimental Section

2.1. Process Conditions

All *E. coli* (HMS174 (DE3)) fed-batch cultivations were performed at the 20L scale expressing hSOD. The feed phase was carried out for four doubling times and induction always took place after the first doubling time, that is, the product formation persisted for the remaining three doubling times. Detailed information about the media, feed, strain, plasmid, cultivation, induction conditions, on- and off-line monitoring were already published elsewhere.[32,33] The investigated design of these studies was later extended, resulting in a design space with three CPPs, each at three levels, that is, 27 CPP combinations in total.[34] The investigated CPPs comprise the cultivation temperature (30, 34, and 37 °C), the induction strength (0.2, 0.5, and 0.9 µmol IPTG/g cell dry mass [CDM]) and the intended specific growth rate (0.10, 0.15, and 0.20 h^{-1}), which was used to compute the feeding rate, as reported in detail elsewhere.[35]

2.2. Data Sets

The data set containing the fed-batch fermentations originated from the previous DoE study and consisted of 31 fermentations, covered the 27 CPP conditions, together with two duplicates and one triplicate. The two off-line process variables, namely the biomass concentration and the soluble product titer, were measured once prior to the induction and hourly afterward. The biomass concentration was determined using thermogravimetric analysis,[36] while for the soluble product titer, an ELISA assay was performed.[37] The on-line available process variables (every 3 minutes) comprised the standard measurements, for example, temperature, feed balance, inductor balance, base balance, stirrer speed, and inlet air. The detailed analytical results for the biomass and the soluble product titer of the fed-batch fermentations are given in Figure S1, Supporting Information. To estimate the uncertainty of the biomass and titer production process, we repeated experiments of one particular fermentation condition seven times. We calculated the errors over all samples and obtained 3.6% and 7.6% for biomass and titer accuracy, respectively. These values represent the threshold of accuracy a certain model can obtain.

2.3. Data Preprocessing

2.3.1. Standardization of the On-Line Data

Prior to the time-resolved process modeling, the on-line available process variables used for the model building were standardized along with the time domain, using the z score.

2.3.2. Interpolation of the Off-Line Data

The off-line measurements for the biomass concentration were available every hour, while the measurements for the soluble product titer were only available every 2 h. For a valid weighted evaluation, a value for the soluble product titer to each biomass value was provided by interpolation to the sampling frequency of the biomass using Hermite polynomials. For this type of interpolation, an initial value and at least four additional values of each variable (i.e., five in total) to be interpolated have to be provided.

2.4. Process Modeling

The respective fed-batch fermentations used for model building (train and validation data set) and model testing (test data set) for the different modeling approaches are described below.

1. Response surface model: all fed-batch fermentations (DoE #1-27) were used ($N = 31$).
2. ANN black-box model: 25 fed-batch fermentations (DoE #1, #3-16, #18-20, and #22 27) for model training and 6 fed-batch fermentations (DoE #2, #4, #9, #17, #21, and #22) for model testing were used ($N = 25+6$).
3. Hybrid model: the same training and test fed-batch fermentations as for the ANN black-box model were used to allow for a fair comparison ($N = 25 + 6$).

The six fed-batch fermentations in the test set were chosen in a way that one fermentation of each replicate run and three individual fed-batch fermentations, which were not present in the training and validation data set were used and that each induction strength was represented by two fed-batch fermentations. The complete list of all CPP combination settings and the respective naming is given in Table S1, Supporting Information.

The black-box model and the hybrid model were developed in the C# hybrid modeling toolbox (Novasign GmbH, Vienna, Austria).

2.4.1. Response Surface Model

The RSM of the full factorial DoE, modeling the endpoints, was generated with MATLAB (2016b, MathWorks, USA) using the function *rstool*, taking independent, constant, linear, interaction, and squared terms into account, to fit the surface and the global confidence intervals set to 95 %.

2.4.2. Artificial Neural Network Black-Box Model

An ANN, that is, the use of propagating predictions, applying a Levenberg–Marquardt algorithm was used to describe the concentration profiles of biomass and soluble product titer. Herein, the predicted values were calculated by numerical stepwise integrating the values from the previously estimated time point, as elaborated elsewhere.[38] Three input variables were used, the cultivation temperature (in °C), the cumulative inductor mass (in kg) and the cumulative feed (in L). The output variables were the biomass concentration and the soluble product titer. These model input variables were chosen due to their assumed significant impact on the soluble product titer.[33] Additionally, these parameters were good candidates to also allow the view on model predictive control (MPC) as the future perspective, since these were direct process inputs and their simple controllability. To identify the most suitable model structure, the number of hidden layers and neurons was systematically varied and the model was chosen for which the best performance in terms of the Akaike information criterion (minimal value) was obtained.

A single hidden layer of four neurons proved to be the best performing structure. Linear transfer functions for the input and output layers and hyperbolic transfer functions for the hidden layer were used.

2.4.3. Hybrid Model

A serial hybrid model was developed by complementing the knowledge-based white-box model with an ANN. Material balances were adopted to describe the evolution of biomass (Equation (1)) and product (Equation (2)). Here, the black-box part was adopted to likewise model the unknown rate expressions, μ and $v_{p/x}$:

$$\frac{dX}{dt} = \mu * X - D * X \tag{1}$$

$$\frac{dP}{dt} = \mathrm{v}_{p/x} * \mathrm{X} * I_{\gamma/n} - \mathrm{D} * \mathrm{P} \tag{2}$$

With the biomass concentration (X) in g L^{-1}, the soluble product titer (P) in g L^{-1}, the inductor switch ($I_{\gamma/n}$), that is, possessing either the value 0 (no induction) or 1 (induction), and the dilution factor (D), which contains the feed addition and sampling, both in liters.

2.4.4. Model Validation

To validate the quality of the established model, an internal validation was performed. Therefore, the training data set was randomly split, fermentation wise, into a training and validation set (ratio 0.8). The training partition was used to train the black-box model and applied on the remaining validation partition. The training stopped once no further improvement in the model performance for the validation partition was achieved, that is, further model training would only lead to overfitting. To provide a large variety of different models, this partitioning of the fed-batch fermentations into boots was repeated 40 times, that is, the fed-batch fermentations used for model building were shuffled and randomly assigned to either the training or validation partition 40 times. For all established models the RMSE (Equation (3)) and the percentage model error (Equation (4)) were calculated with the measured value (y), its estimated counterpart ($\hat{y}$) for each time point (t) and the total number of observations (N).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} * \sum \left(y_{(t)} - \hat{y}_{(t)}\right)^2} \tag{3}$$

$$\mathrm{Error}\,[\%] = \frac{100}{N} * \sum \frac{\left|y_{(t)} - \hat{y}_{(t)}\right|}{y_{(t)}} \tag{4}$$

2.4.5. Bootstrapping

For the ANN black-box and the hybrid model, bootstrap aggregation was applied to enhance the robustness and evaluate its predictive uncertainty.[39] Each bootstrapped model consisted of six individual models (each derived from a different boot) to ensure that every fed-batch fermentation was present in the bootstrap-aggregated model and to guarantee a meaningful evaluation. For each bootstrapped model, the standard deviation (SD) (Equation (5)) was calculated with the bootstrapped value of the prediction ($\hat{y}_{\mathrm{btstrp}}$) (i.e., the mean value of all used models), the respective predicted counterpart from the models ($\hat{y}_{\mathrm{model}}$), the index ($i = 1{:}6$) and the number of observations for each time point (n). The SD was used to compute the prediction interval (PI) (Equation (6)). The respective PIs were derived from the SD of the incorporated models. These bootstrap-aggregated models were applied on the test set (external validation), to assess the predictability of the final models on new data.

$$\mathrm{SD}_{(t)} = \sqrt{\frac{1}{n-1} * \sum \left(\hat{y}_{\mathrm{btstrp}(t)} - \hat{y}_{\mathrm{model}(i)_{(t)}}\right)^2} \tag{5}$$

$$\mathrm{PI}_{(t)} = \hat{y}_{\mathrm{btstrp}(t)} \pm \mathrm{SD}_{(t)} \tag{6}$$

3. Results

3.1. Model Performance Comparison for Process Endpoints

We compared the model performance of the RSM, the black-box model, and the hybrid model at the process endpoints. The endpoint values of the generated black-box model and the hybrid model were taken and plotted as surfaces. This allows for a comparison to the widely adopted RSM applications, with respect to the accuracy. The response surfaces were separated by the induction strengths, referred to as induction planes. The three approaches, a commonly developed RSM using a full quadratic fitting function and the ANN black box as well as the hybrid model, were compared to the measured endpoint values. One exemplary induction plane (0.5 µmol IPTG/g CDM) of the biomass concentration (**Figure 1**) and the soluble product titer (**Figure 2**) is presented.

The summarized modeling results are listed in **Table 1**. Herein, all induction planes are incorporated for the calculations of the R^2, the RMSE, and the percentage error. Even though the same value is calculated for the RMSE, for example, for the RSM and the hybrid model, these resulted in different values in the percentage error. This is due to the relativity in the calculation method of the percentage error.

An explicit trend toward higher biomass concentrations with CPP combinations using faster specific growth rates, lower induction strengths, and lower temperatures was observed in all induction planes. Compared to the experimental response for the biomass concentration (Figure 1A), the RSM exhibits high similarity (Figure 1B). The response surface of the black-box model (Figure 1C) is of limited quality and does not allow for meaningful comparison, for example, displaying a declining biomass concentration in the region, where the highest biomass concentration was observed, accompanied by high SDs. Two of the replicates in the test set were performed in this induction plane, respectively, fed-batch fermentation #4 (30 °C and μ 0.10) and #22 (30 °C and μ 0.20). Even though these data sets were present in the training and the test set, the estimation from the black-box model was error prone indicating poor fitting performance. In contrast, the hybrid model (Figure 1D) achieves high comparability to the experimental response and the RSM, while displaying small SDs. Although RSM generates insights into the process at the endpoint, by solely considering this data point, it is not possible to find process optima, which are not located at the end of the process due to the limited evaluation.

The modeling of the soluble product titer proved to be more challenging. The RSM (Figure 2B) captures the overall shape but is not able to accurately represent the high values of the experimental response (Figure 2A). The black-box model (Figure 2C) performed similarly to the RSM. The prediction for replicate #4 in the test set almost matched the true response, while the test set replicate #22 was still displayed inaccurately. The hybrid model (Figure 2D) was able to predict these replicates more accurately and, overall, displayed the highest similarity to the experimental response while maintaining small SDs.

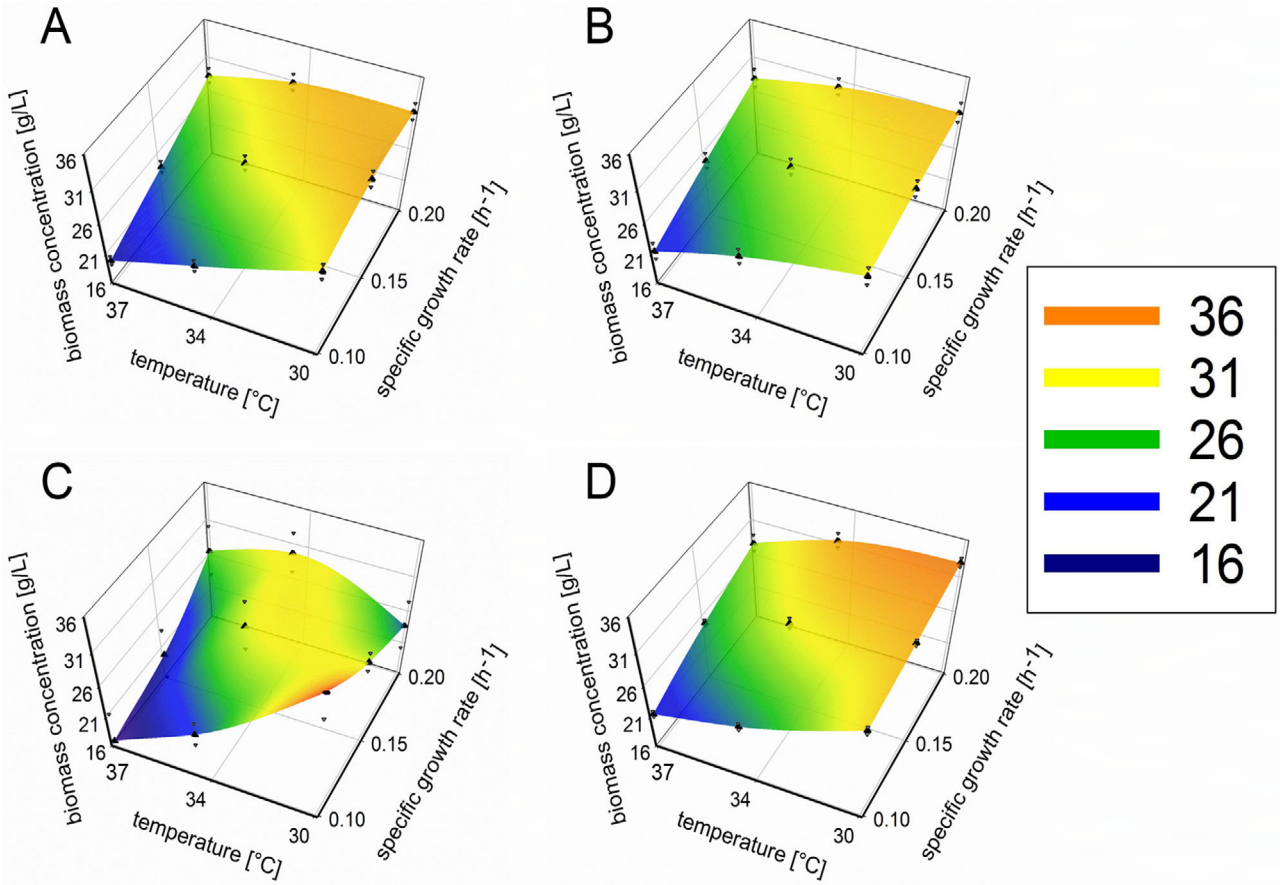

Figure 1. Response surface predictions of the endpoint values of the biomass concentration. A) The experimental response, B) the full quadratic RSM, C) the bootstrap-aggregated black-box model, and D) the bootstrap-aggregated hybrid model are displayed as a function of the temperature and the intended specific growth rate for the induction plane 0.5 µmol IPTG/g CDM. The color indicates the values of the biomass concentration from dark blue (lowest value) to orange (highest value). For the analytical measurements and the values derived from the respective model (triangles) the SD is indicated.

3.2. Time-Resolved Model Performance Comparison

By the application of time-resolved models, the limitation of solely endpoint modeling and neglecting process dynamics, as is the case using RSM, can be overcome. To determine which time-resolved model is most suitable to accurately model the entire process, the predictions of the black-box model and the hybrid model were compared. The robustness and susceptibility to errors on the fed-batch fermentations from the test set were considered. The superiority of the hybrid model compared to a pure black-box model, with respect to predictability, is shown for the biomass concentration (**Figure 3**) and the soluble product titer (**Figure 4**). The complete comparison of the two model performances applied to the test set is given in Table S2, Supporting Information. The complete overview of the scatter and the time-resolved plots for the training and the test set, including the error bars of the prediction (deliberately omitted here), as was generated by the Novasign hybrid modeling toolbox, is shown in Figures S2 (ANN black box) and S3 (hybrid model), Supporting Information.

Considering the biomass concentration, a high spreading of the values in both, the training and test partition, for the bootstrap-aggregated ANN black-box model (Figure 3A) was observed, while the bootstrap-aggregated hybrid model displayed a tight distribution around the least-squares line (Figure 3B). The predictive error for the biomass concentration decreased threefold due to the incorporated knowledge in the structure of the hybrid model. Similar results were obtained when considering the time-resolved model performance on the test set. The predicted values from the bootstrap-aggregated ANN black-box model clearly differed from the analytical measurements, resulting in broad PIs. This was even observed for the prediction of the replicates (Figure 3C,D), indicating an incapacity to fit the data.

The bootstrap-aggregated hybrid model did not display such issues and predicted the trend correct with small PIs. Likewise, this was the case for the fed-batch fermentations, which were not present in the training set. The biomass predictions of the bootstrap-aggregated ANN black-box model (Figure 3E) did not match the analytical results, being even outside the PIs. The bootstrap-aggregated hybrid model was able to predict these trends accurately. Only for one fed-batch fermentation, also the bootstrap-aggregated hybrid model failed to match the analytical values, still predicting the correct trend but in an insufficient manner, while the bootstrap-aggregated ANN black-box model again displayed fluctuating values (Figure 3F).

For the soluble product titer, the results of the scatter plots for the bootstrap-aggregated ANN black-box model (Figure 4A) and

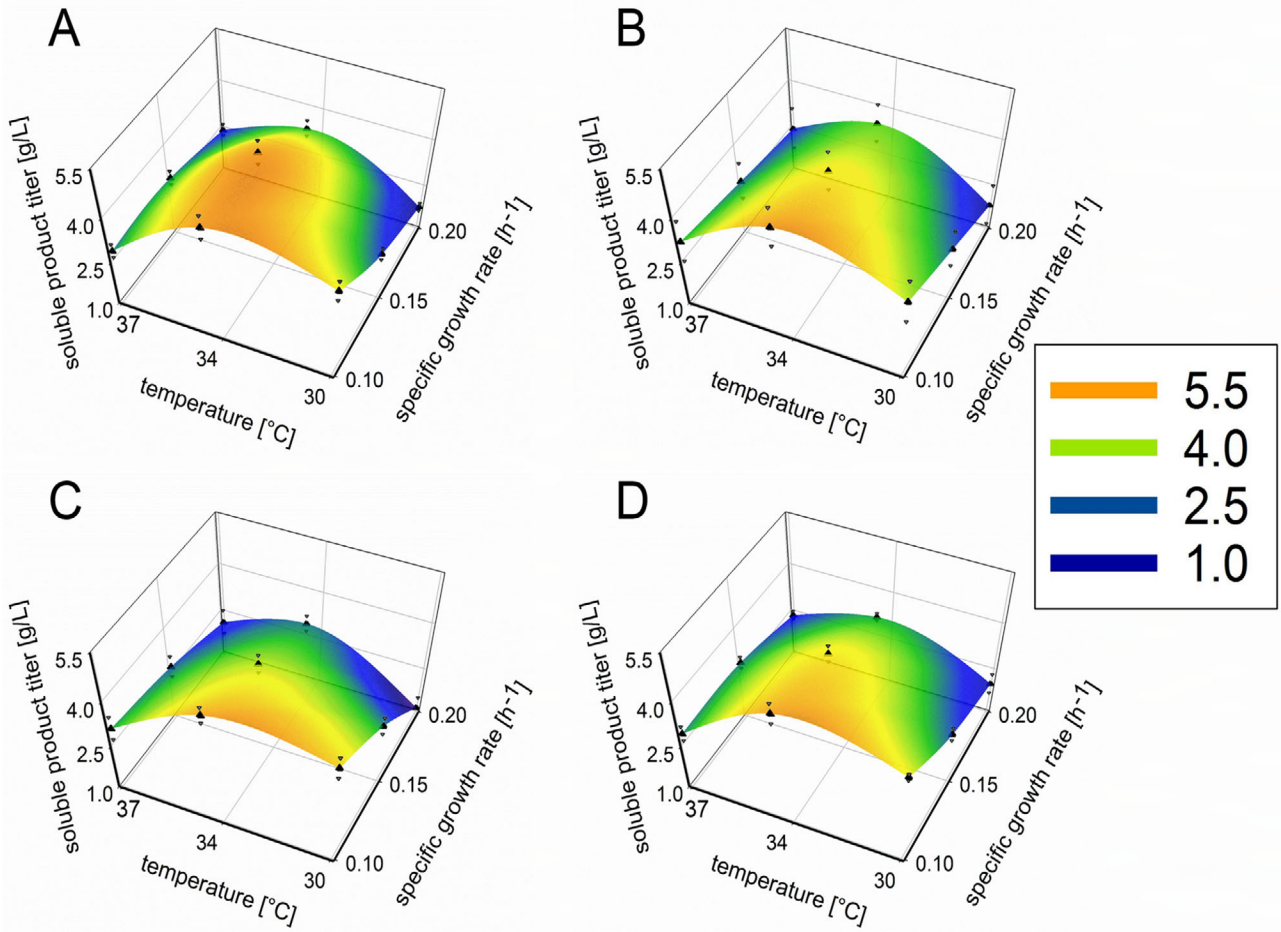

Figure 2. Response surface predictions of the endpoint values of the soluble product titer. A) The experimental response, B) the full quadratic RSM, C) the bootstrap-aggregated black-box model, and D) the bootstrap-aggregated hybrid model are displayed as a function of the temperature and the intended specific growth rate for the induction plane 0.5 µmol IPTG/g CDM. The color indicates the values of the soluble product titer from dark blue (lowest value) to orange (highest value). For the analytical measurements and the values derived from the respective model (triangles) the SD is indicated.

Table 1. Performance results of the RSM, the bootstrap-aggregated black-box model, and the bootstrap-aggregated hybrid model predicting the endpoint values of the biomass concentration and the soluble product titer. For each process variable, the model, the RMSE, and the percentage error are given (rounded to two decimal places). Additional model characteristics are indicated. The number of fed-batch fermentations used for model training and model testing is indicated in brackets.

Endpoint values	Biomass			Product		
Performance criteria	state of the art RSM ($n = 31$)	ANN black-box model ($n = 25+6$)	Hybrid model ($n = 25 + 6$)	state of the art RSM ($n = 31$)	ANN black-box model ($n = 25 + 6$)	Hybrid model ($n = 25 + 6$)
RMSE [g L^{-1}]	0.95	4.32	0.99	0.47	0.47	0.33
Error [%]	2.84	11.62	3.06	21.45	19.11	9.54
Entire process modeling	×	✓	✓	×	✓	✓
Captures process dynamics	×	✓	✓	×	✓	✓
Incorporation of process knowledge	×	×	✓	×	×	✓

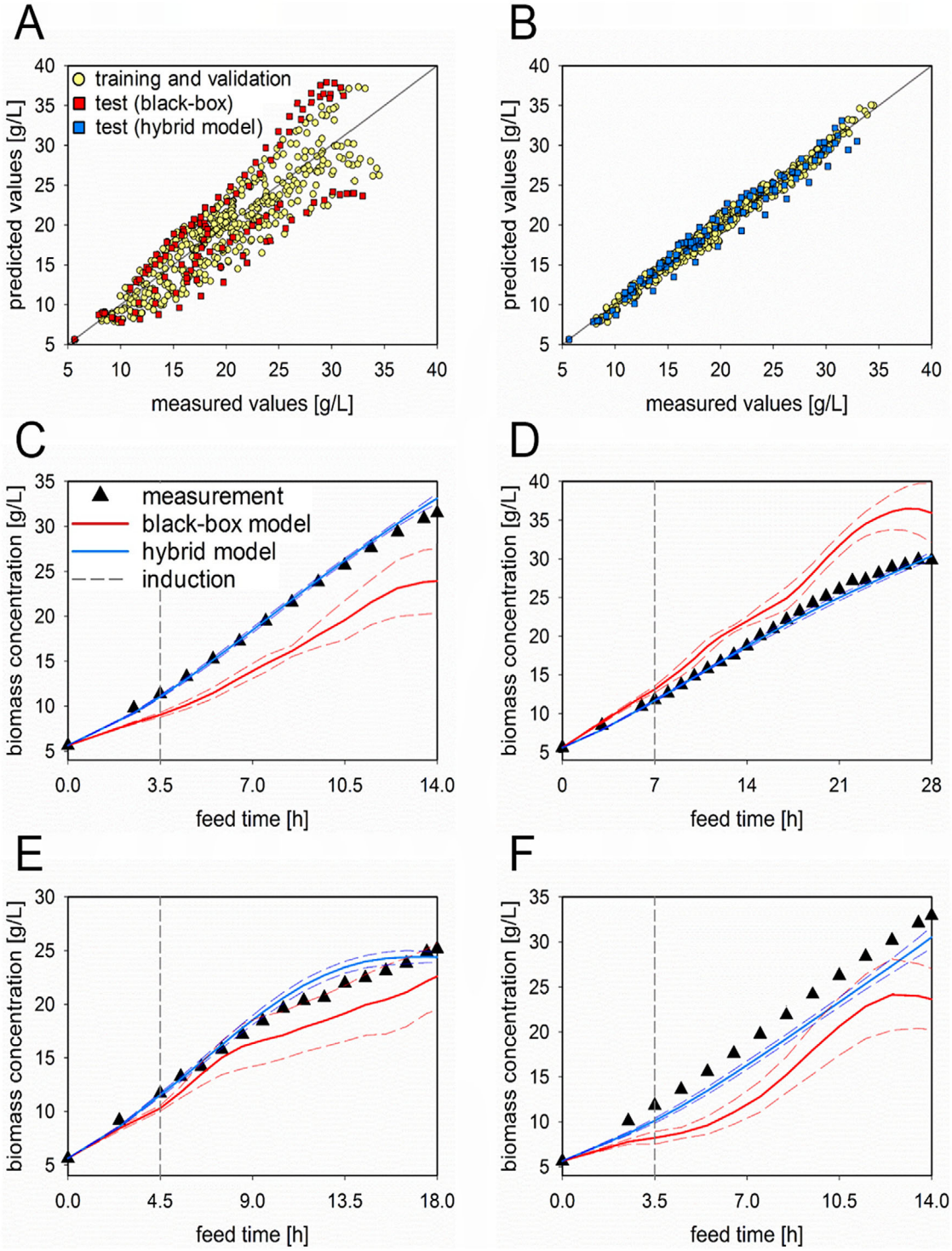

Figure 3. Performance comparison of the bootstrap-aggregated ANN black-box and the bootstrap-aggregated hybrid modeling approaches applied to the test set. The analytical results for the biomass concentration and the predictions from the two models are presented. A) The scatter plot of the ANN black-box model, B) the hybrid model and the time-resolved comparison of four fed-batch fermentations of the test set are displayed, including the respective PI of the model. C) DoE #22, D) DoE #4, E) DoE #17, and F) DoE #21.

the bootstrap-aggregated hybrid model (Figure 4B) did not differ as much as for the biomass. Still, the predictions from the test set of the bootstrap-aggregated hybrid model presented narrower distributions compared to the bootstrap-aggregated ANN black box. Likewise as for the biomass concentrations, the bootstrap-aggregated hybrid model was able to predict the soluble product titer trends of the fed-batch fermentations accurately. The predictions obtained from the bootstrap-aggregated ANN black-box model did not match the analytical measurements of the replicate in Figure 4C, but performed sufficiently on the second replicate (Figure 4D). The predicted trends of the fed-batch fermentation in Figure 4E was slightly underestimated by the

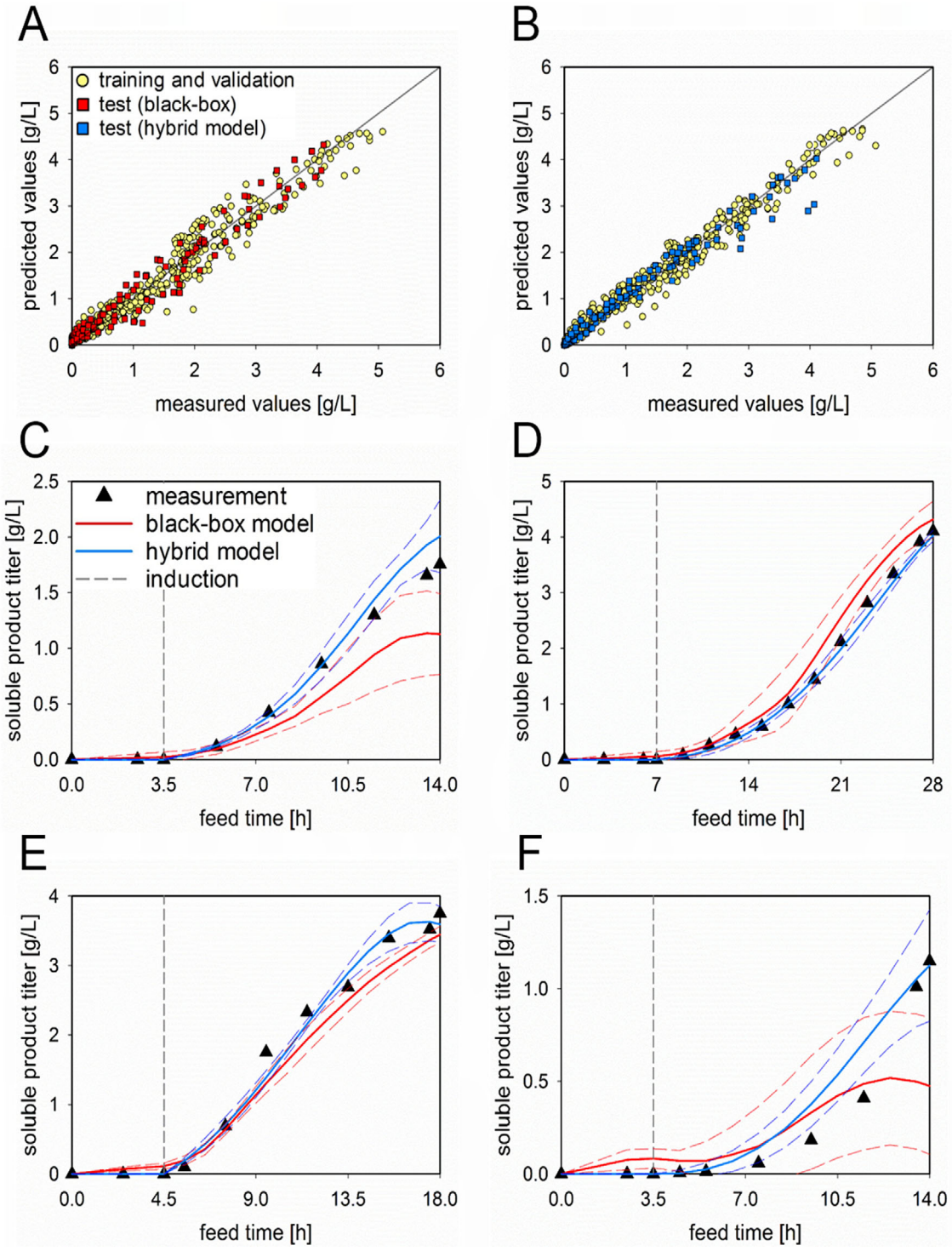

Figure 4. Performance comparison of the bootstrap-aggregated ANN black-box and the bootstrap-aggregated hybrid modeling approaches applied to the test set. The analytical results for the soluble product titer and the predictions from the two models are presented. A) The scatter plot of the ANN black-box model, B) the hybrid model and the time-resolved comparison of four fed-batch fermentations of the test set are displayed, including the respective PI of the model. C) DoE #22, D) DoE #4, E) DoE #17, and F) DoE #21.

bootstrap-aggregated ANN black-box model, but well described by the bootstrap-aggregated hybrid model. The only fed-batch fermentation in which the bootstrap-aggregated hybrid model did not perform well, was the same as for the biomass concentration before and also the bootstrap-aggregated ANN black-box model displayed the highest deviation from the analytical measurements in this cultivation (Figure 4F). Both bootstrap-aggregated models displayed the broadest PIs. To find possible explanations for the prediction difficulty of the fed-batch fermentation in Figures 3F and 4F, the respective rates, calculated from the off-line measurements, which are estimated in the black box of the hybrid model structure, of all fed-batch fermentations from the test set were investigated and are shown in Figure S4, Supporting Information.

Moreover, all time-resolved depictions of the bootstrap-aggregated ANN black-box model already predict product before induction. Due to the inductor switch in the white box of the hybrid model, such product overestimations before induction were eliminated. Moreover, the chosen CPPs for the ANN model input may be suboptimal for predicting the variables of interest. Therefore, these might not be the most suitable factors to completely describe the behavior in the characterized space. The predictability of the soluble product titer increased, reducing the range of the PIs, applying the bootstrap-aggregated hybrid model since herein the biomass which can be predicted with high accuracy is linked to the product.

4. Discussion

It has been shown that the widely adopted RSM is able to fit the endpoint values of the characterized space quite accurately (Figures 1B and 2B). This fitting function contained constant, linear, and squared terms and interactions between the CPPs. However, typically only one point in time, mostly the endpoints, are used for model development, disregarding completely the evolution of the process over time. Herein, several drawbacks are recognizable, when process parameters are assumed to be constant during the process. For instance, the impact of temporal deviations in the CPPs on the endpoint response cannot be assessed. Also the desired process optimum, for example, the highest specific concentration may be overlooked. In contrast, the developed ANN black-box model and the hybrid model enable modeling of the biomass concentration and the soluble product titer of the process in a time-resolved manner. Moreover, by the utilization of bootstrapping the uncertainty of the developed model is visible.

When taking only the endpoint values into account, the superiority of the hybrid model over the black-box model could be determined (Figures 1C,D and 2C,D). Also, the hybrid model performed comparatively well to the RSM and displayed the smallest SD, while providing the additional advantage of modeling the entire process. Therefore, being able to identify process optima during the whole process. The same tendency was observed for the soluble product titer, for which in general less accurate prediction results were obtained.

Most probably, this behavior can be explained by the combination of various factors. The most prominent issue could have been a non-producing subpopulation without plasmids which increases during the cultivations but was not measured analytically. Especially for the cultivations at low feeding rates with high induction and high temperatures (DoE #9) this behavior can be seen in the increased growth rates at the end of the cultivation (Figure S4C, Supporting Information) Therefore, the modeling could not account for it. Another point to consider was the formation of inclusion bodies at certain CPP combinations, which decreased the soluble product. The initial variability of the plasmid copy number and the occurring cell lysis were also considered as influencing factors, which could not be taken into account. Addressing total SOD production did also not yield more accurate results due to the higher uncertainty of inclusion body analytics via SDS gel electrophoresis.

With respect to modeling the entire process, it also has been demonstrated that the bootstrap-aggregated hybrid model outperforms the bootstrap-aggregated black-box model (Figures 3 and 4). Due to the hybrid model structure, the model gains the advantage of understanding whether occurring variations are due to changes in the metabolism (e.g., the exemplary rate expressions in Figure S4, Supporting Information) or process operations, for example, the feeding, resulting in small SDs and tight PIs. The bootstrap-aggregated black-box model does not possess this ability; resulting in imprecise predictions for the process variable trajectories. This limitation was not observed using the bootstrap-aggregated hybrid model, due to the corrective action of the white box.

These results are showing that hybrid modeling has become a reliable and highly beneficial concept for upstream process characterization, including better process characterization and building a dynamic model. The developed bootstrap-aggregated hybrid model keeps up with commonly used techniques predicting the endpoint values of the process and even outperforms these techniques with respect to time-resolved process modeling. Moreover, the three CPPs used as inputs to the hybrid model are controllable, which in principle enables MPC in future applications.

Even though we could not account for every limitation in the presented data set, as mentioned above, the bootstrap-aggregated hybrid model performed reliably and with high predictability. To gain superior results in possible future DoE studies, more analytical techniques should be applied to also consider the above-mentioned factors. This includes on-line measurements such as off-gas analysis (CO_2 and O_2) and off-line analysis such as total organic carbon and measuring the DNA content in the suspension and supernatant to also access the carbon balance and to account for cell lysis.

In summary, to tackle these previously mentioned research gaps, regarding process understanding and QbD implementation, for example, in process characterization and optimization, the established bootstrap-aggregated hybrid model enables time-resolved modeling of the entire process for two target variables with a single model structure. This also renders possible the identification of process optima differing from the endpoint, potentially increasing the space time yield. We are aware of the fact that, while modeling the biomass concentration works well, our three chosen CPPs are probably insufficient inputs to ideally model the product formation for the ANN black box or the hybrid model. However, these CPPs were selected since they are easily controllable and predictable multisteps ahead, which is the prerequisite for subsequently building a model predictive or model-based control strategy. Therefore, these are promising control parameters to implement hybrid models for predictive soft sensors and also to implement MPC strategies, advancing from predictive to controllable models. While the first tool for advanced process control is already in use (a showcase of an exemplary fermentation, recorded in our OPC environment, is presented in Video S1, Supporting Information), the latter is currently in proof of concept and results will be presented in future publications.

An unsolved yet critical and major concern still is the long period of time spent on experiments until such a hybrid model is applicable. A novel approach to accelerate process characterization and optimization is presented by the concept of intensified DoE. Herein, by the utilization of intra-experimental shifts more than

one CPP combination setting is addressed per cultivation run, reducing the total number of experiments.[40,41] This promising approach in combination with the herein developed bootstrap-aggregated hybrid modeling strategy will be investigated in future publications.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors would like to thank the Austrian Research Promotion Agency (FFG) for their support (Research Studio Austria, 859219). The authors would also like to thank the laboratory of Igor Škrjanc and especially Dejan Dovzan from the faculty of electrical engineering at the University of Ljubljana and Roger Dalmau Diaz (University of Natural Resources and Life Sciences, Vienna) for developing the prototype of the Novasign Hybrid Modeling Toolbox.

Conflict of Interest

Gerald Striedner and Mark Dürkop hold shares of Novasign GmbH.

Keywords

hybrid modeling, process control, Quality by Design

Received: December 10, 2019
Revised: January 28, 2020
Published online: February 17, 2020

[1] U.S. Food and Drug Administration, *Pharmaceutical CGMPs for the 21st Century—A Risk Based Approach*; Final Report; U.S. Food and Drug Administration: Silver Spring, MD, **2004**.
[2] A. Pekarsky, V. Konopek, O. Spadiut, *Bioprocess Biosyst. Eng.* **2019**, *42*, 1611.
[3] A. S. Rathore, H. Winkle, *Nat. Biotechnol.* **2009**, *27*, 26.
[4] F. Torkashvand, B. Vaziri, S. Maleknia, A. Heydari, M. Vossoughi, F. Davami, F. Mahboudi, *PLoS One* **2015**, *10*, 1.
[5] J. Ramírez, H. Gutierrez, A. Gschaedler, *J. Biotechnol.* **2001**, *88*, 259.
[6] J. Möller, K. B. Kuchemüller, T. Steinmetz, K. S. Koopmann, R. Pörtner, *Bioprocess Biosyst. Eng.* **2019**, *42*, 867.
[7] C. Mandenius, A. Brundin, *Biotechnol. Prog.* **2008**, *24*, 1191.
[8] U.S. Food and Drug Administration, *Guidance for Industry: Q8(R2) pharmaceutical development*; Final Draft; International Conference on Harmonisation, **2009**.
[9] V. Mishra, S. Thakur, A. Patil, A. Shukla, *Expert Opin. Drug Delivery* **2018**, *15*, 737.
[10] A. S. Patil, A. M. Pethe, *Int. J. Pharm. Qual. Assur.* **2013**, *4*, 13.
[11] B. K. Ahuja, S. Suresh, R. Sabyasachi, *Indian Drugs* **2015**, *52*, 5.
[12] V. Kumar, A. Bhalla, A. S. Rathore, *Biotechnol. Prog.* **2014**, *30*, 86.
[13] S. J. Kalil, F. Maugeri, M. I. Rodrigues, *Process Biochem.* **2000**, *35*, 539.
[14] R. Balusu, R. R. Paduru, S. K. Kuravi, G. Seenayya, G. Reddy, *Process Biochem.* **2005**, *40*, 3025.
[15] G. Q. Liu, X. L. Wang, *Appl. Microbiol. Biotechnol.* **2007**, *74*, 78.
[16] M. S. Tanyildizi, D. Özer, M. Elibol, *Process Biochem.* **2005**, *40*, 2291.
[17] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3.
[18] D. M. Hallow, B. M. Mudryk, A. D. Braem, J. E. Tabora, O. K. Lyngberg, J. S. Bergum, L. T. Rossano, S. Tummala, *J. Pharm. Innovation* **2010**, *5*, 193.
[19] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, University of Oxford, UK **1996**.
[20] A. S. Rathore, *Curr. Opin. Chem. Eng.* **2014**, *6*, 1.
[21] P. Wechselberger, P. Sagmeister, C. Herwig, *Bioprocess Biosyst. Eng.* **2013**, *36*, 1205.
[22] S. M. Faassen, B. Hitzmann, *Sensors* **2015**, *15*, 10271.
[23] K. Ohadi, H. Aghamohseni, R. L. Legge, H. M. Budman, *Biotechnol. Bioeng.* **2014**, *111*, 1577.
[24] K. Ohadi, R. L. Legge, H. M. Budman, *Biotechnol. Bioeng.* **2015**, *112*, 197.
[25] T. Schmidberger, C. Posch, A. Sasse, C. Guelch, R. Huber, *Biotechnol. Prog.* **2015**, *31*, 1119.
[26] A. S. Rathore, S. Mittal, M. Pathak, V. Mahalingam, *J. Chem. Technol. Biotechnol.* **2014**, *89*, 1311.
[27] M. von Stosch, S. Davy, K. Francois, V. Galvanauskas, J.-M. Hamelink, A. Luebbert, M. Mayer, R. Oliveira, R. O'Kennedy, P. Rice, J. Glassey, *Biotechnol. J.* **2014**, *9*, 719.
[28] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, *Comput. Chem. Eng.* **2014**, *60*, 86.
[29] B. Bayer, B. Sissolak, M. Duerkop, M. von Stosch, G. Striedner, *Bioprocess Biosyst. Eng.* **2020**, *43*, 169.
[30] M. von Stosch, J.-M. Hamelink, R. Oliveira, *Bioprocess Biosyst. Eng.* **2016**, *39*, 773.
[31] B. S. Riley, X. Li, *AAPS PharmSciTech* **2011**, *12*, 114.
[32] M. Melcher, T. Scharl, B. Spangl, M. Luchner, M. Cserjan, K. Bayer, F. Leisch, G. Striedner, *Biotechnol. J.* **2015**, *10*, 1770.
[33] K. Marisch, K. Bayer, M. Cserjan-Puschmann, M. Luchner, G. Striedner, *Microb. Cell Fact.* **2013**, *12*, 58.
[34] B. Bayer, M. von Stosch, M. Melcher, M. Duerkop, G. Striedner, *Eng. Life Sci.* **2020**, *20*, 26.
[35] M. Luchner, G. Striedner, M. Cserjan-Puschmann, F. Strobl, K. Bayer, *J. Chem. Technol. Biotechnol.* **2015**, *90*, 283.
[36] M. Cserjan-Puschmann, W. Kramer, E. Duerrschmid, G. Striedner, K. Bayer, *Appl. Microbiol. Biotechnol.* **1999**, *53*, 43.
[37] T. Porstmann, R. Wietschke, H. Schmechta, R. Grunow, B. Porstmann, R. Bleiber, M. Pergande, S. Stachat, R. von Baehr, *Clin. Chim. Acta* **1988**, *171*, 1.
[38] C. R. De Azevedo, J. Peres, M. von Stosch, *Eng. Appl. Artif. Intell.* **2015**, *38*, 24.
[39] D. A. Freedman, *Annals of Statistics* **1981**, *9*, 1218.
[40] M. von Stosch, J. M. Hamelink, R. Oliveira, *Biotechnol. Prog.* **2016**, *32*, 1343.
[41] M. von Stosch, M. J. Willis, *Eng. Life Sci.* **2016**, *17*, 1173.

Publication III

Supporting Information

Supporting Information

Analytical Results of the Design of Experiments

The analytical results of the fed-batch fermentations performed in the three-dimensional design space, introduced in the Materials & Methods section of the main manuscript (2.1 & 2.2), are displayed as a function of the feed time in Fig. S1. The biomass (displayed as concentration and total to demonstrate the applied exponential feeding strategy) and the soluble product titer are shown for all 27 CPP settings, as stated in Table S1. Only one fed-batch fermentation is shown per duplicate and triplicate run for those CPP settings for which fermentations were repeated.

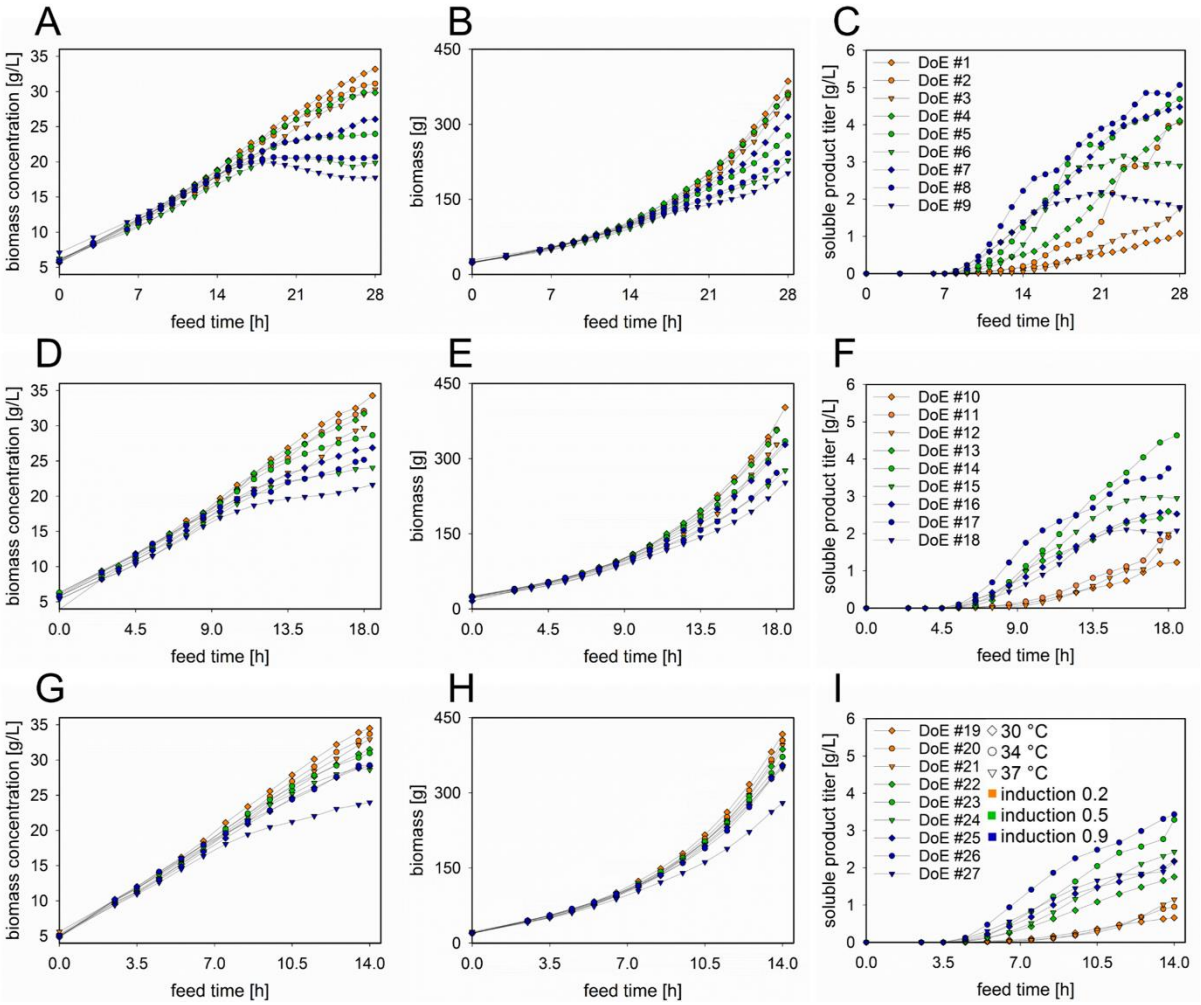

Figure S1. Analytical results for the biomass and soluble product titer of the fed-batch fermentations. The variables are displayed as a function of the feed time. For visual clarity, the fed-batch fermentations are separated into the three specific growth rates, i.e., $\mu = 0.10\ h^{-1}$ (A-C), $\mu = 0.15\ h^{-1}$ (D-F) and $\mu = 0.20\ h^{-1}$ (G-I). The biomass concentration (A, D, G), the total biomass (B, E, H), and the soluble product titer (C, F, I) are displayed. To indicate the induction strength of these fermentations, a color code was used, i.e., orange 0.2, green 0.5, and blue 0.9.

Critical Process Parameter Combination Settings for the Design of Experiments

The investigated CPPs, introduced in the Materials & Methods section of the main manuscript (2.1 & 2.2), are presented in a tabular form in Table S1.

Table S1. CPP settings of the characterized design space. Each parameter, namely, the specific growth rate in h^{-1}, the induction strength in µmol IPTG/g cell dry mass, and cultivation temperature in °C, was investigated at three levels. The settings for the fed-batch fermentations (DoE) are listed. The number of repetitions is indicated

CPP setting	specific growth rate	temperature	induction strength
#1	0.10	30	0.2
#2	0.10	34	0.2
#3	0.10	37	0.2
#4 (n = 2)	0.10	30	0.5
#5	0.10	34	0.5
#6	0.10	37	0.5
#7	0.10	30	0.9
#8	0.10	34	0.9
#9 (n = 3)	0.10	37	0.9
#10	0.15	30	0.2
#11	0.15	34	0.2
#12	0.15	37	0.2
#13	0.15	30	0.5
#14	0.15	34	0.5
#15	0.15	37	0.5
#16	0.15	30	0.9
#17	0.15	34	0.9
#18	0.15	37	0.9
#19	0.20	30	0.2
#20	0.20	34	0.2
#21	0.20	37	0.2
#22 (n = 2)	0.20	30	0.5
#23	0.20	34	0.5
#24	0.20	37	0.5
#25	0.20	30	0.9
#26	0.20	34	0.9
#27	0.20	37	0.9

Modeling Results of the Bootstrap Aggregating Models

The complete modeling results of the fed-batch fermentations from the training and the test set introduced in the sections Materials & Methods (2.4) and Results (3.1 & 3.2) of the main manuscript are displayed. The overview of the bootstrap-aggregated ANN black-box model (Fig. S2) and the bootstrap-aggregated hybrid model (Fig. S3) are displayed as it is generated by the Novasign hybrid modeling toolbox. Herein, for each observation the standard deviation (scatter plots) and the prediction interval (time-resolved plots) are given. The standard deviation in the scatter plots was intentionally left out in the main manuscript, for the sake of clarity.

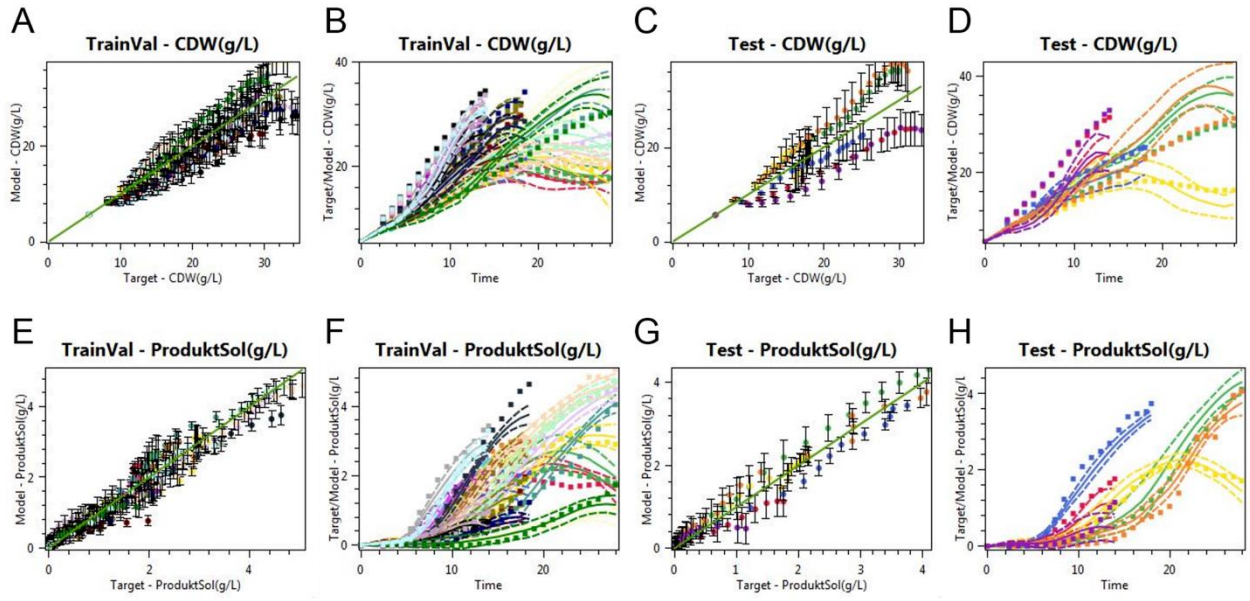

Figure S2. Overview of the modeling results of the bootstrap-aggregated ANN black-box model. The plots from the Novasign hybrid modeling toolbox show the modeling results for the biomass (A-D) and the product (E-H). The scatter plots for the training (A & E), the test data (C & G) and the time-resolved plots for the training (B & D) and the test data (F & H) display each fed-batch fermentation in a different color. The respective standard deviation and the prediction interval of the bootstrap-aggregated model for each fed-batch fermentation are indicated.

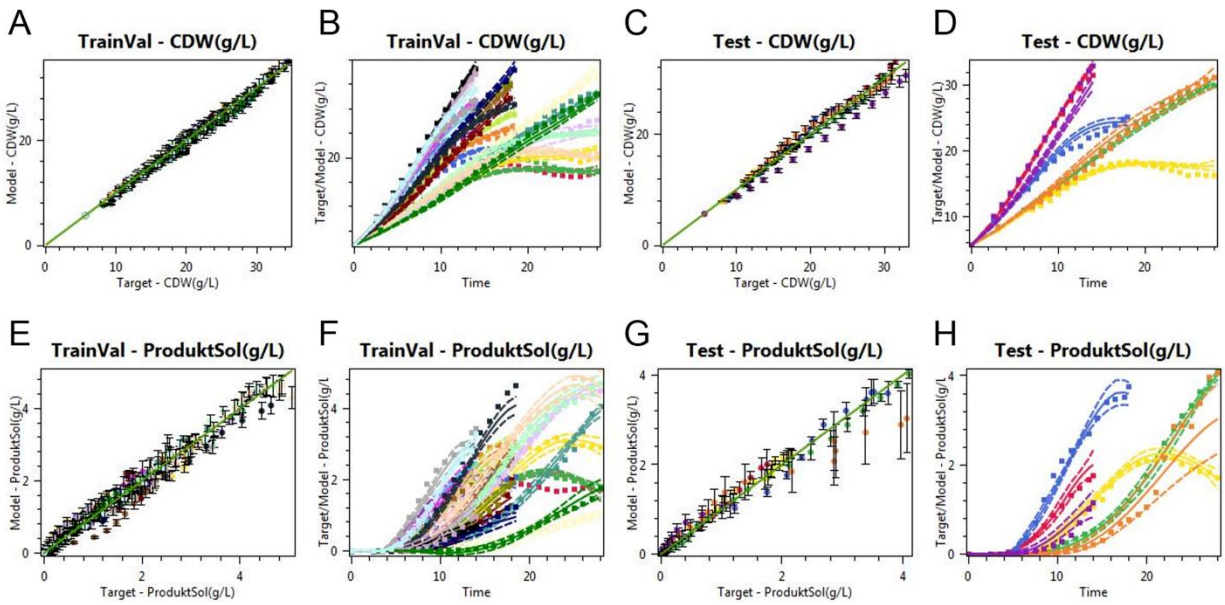

Figure S3. Overview of the modeling results of the bootstrap-aggregated hybrid model. The plots from the Novasign hybrid modeling toolbox show the modeling results for the biomass (A-D) and the product (E-H). The scatter plots for the training (A & E), the test data (C & G) and the time-resolved plots for the training (B & D) and the test data (F & H) display each fed-batch fermentation in a different color. The respective standard deviation and the prediction interval of the bootstrap-aggregated model for each fed-batch fermentation are indicated.

The direct and complete comparison of the two modeling attempts shows the advantage of using a hybrid model, as also shown in detail in Fig. 3 and Fig. 4 in the main manuscript.

Regarding the biomass, the bootstrap-aggregated ANN black-box showed significant drawbacks to predict the analytical values, for the training (Fig. 2SA & B) and the test set (Fig. 2SC & D). In addition, the rather high standard deviation and the wide prediction intervals indicate poor model performances from the individual models. The prediction of the soluble product titer performed better. The bootstrap-aggregated model was able to fit the training and the test set likewise (Fig. 2SE & G). The standard deviation and prediction intervals were narrowed (Fig. 2SF & H) compared to its biomass prediction.

Otherwise, the bootstrap-aggregated hybrid model showed exceptional performance on predicting the biomass, displaying small standard deviations and tight prediction intervals in both, the training (Fig. 3SA & B) and the test data (Fig. 3SC & D). In addition, the soluble product

titer was predicted with a good performance and small standard deviations as seen in Fig. 3SE & G. Overall, the bootstrap-aggregated hybrid model displayed tight prediction intervals for the product titer (Fig. 3SF & H). However, it struggled with one fed-batch fermentation in the test set, namely #21, as indicated by the high standard deviation and broad prediction interval.

Summarized, the ANN black-box model was not able to predict the biomass concentration, only performed decently on the soluble product titer and the additional utilization of bootstrapping also did not lead to an adequate model performance. The bootstrap-aggregated hybrid model on the other hand, excellently predicted the biomass concentration, achieved good results for the soluble product titer and displayed high robustness due to the bootstrapping, as indicated by the small standard deviations and prediction intervals. Possible difficulties of predicting the values of fed-batch fermentation #21 in the test set are discussed in the subsequent section “Specific Rates of the Fed Batch Fermentations in the Test Set”.

Specific Rates of the Fed-Batch Fermentations in the Test Set

The specific growth rate and the specific rate of the soluble product formation from all fed-batch fermentations of the test set, mentioned in the Results section of the main manuscript (3.2), are displayed as a function of the feed time in Fig. S4. The displayed specific rates are derived from the off-line measurements. Moreover, the specific growth rate is plotted only from the first sampling until the end. In the hybrid model structure, these were estimated in the black-box to provide the values for the white-box. For each replicate, the mean value and the respective standard deviation are plotted.

As suggested and presented in an earlier publication, for robust and accurate calculation of the specific growth rate (μ), a cubic smoothing spline was used, applying the MATLAB function *csaps(x,y,p)* [1]. Herein, *x* represents the total feed time of the process, *y* the total biomass and a value of 0.4 for the fitting parameter *p*.

The specific soluble product formation rate ($\mathbf{q}_p$ soluble) was derived from the off-line analytical measurements and calculated according to Eq. 1 with *P* the total soluble product yield (in mg), *X* the total biomass (in g) and *T* the total time of the process (in h).

$$\boldsymbol{q}p\ (soluble) = \frac{P(t+1)-P(t)}{\frac{(X(t+1)+X(t))/2}{T(t+1)-T(t)}} \quad (1)$$

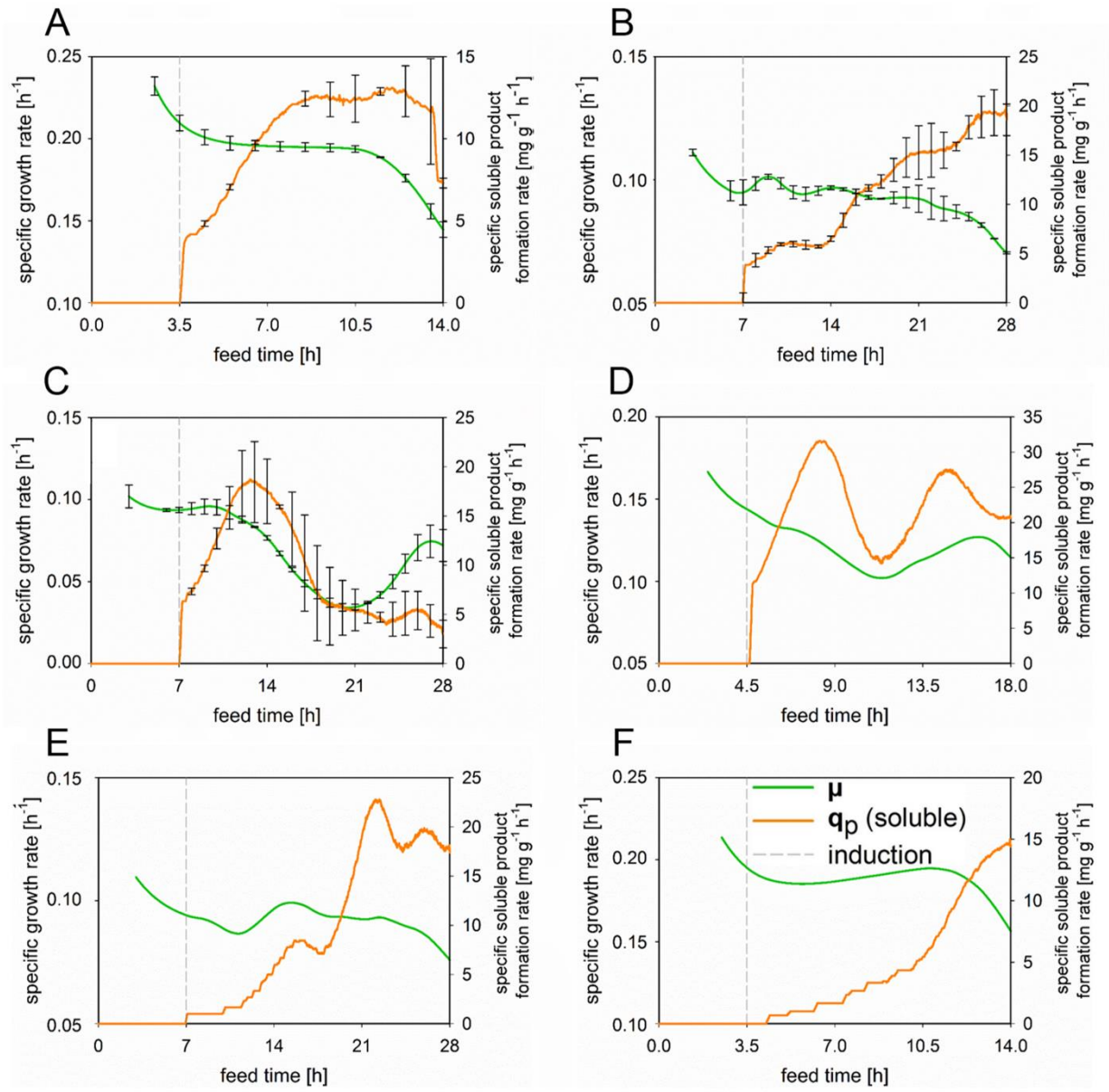

Figure S4. Calculated values for the specific rates, derived from the off-line measurements, of all fed-batch fermentations in the test set. The specific growth rate (green line) and the specific soluble product formation rate (orange line) are displayed as a function of the feed time. For the replicates, the mean values and the standard deviations are plotted. DoE #22 (A), DoE #4 (B), DoE #9 (C), DoE #17 (D), DoE #2 (E) and DoE #21 (F).

The different CPP settings, at which the fed-batch fermentations were performed, had significant impacts on the biomass concentration and the soluble product titer as previously shown. These also resulted in different patterns for the specific rates. The intended specific growth rate was influenced by the induction strength of the respective fed-batch fermentation, i.e., if the induction strength increased, the specific growth rate declined. This impact was diminished by higher

specific growth rate settings, since by providing a higher feeding rate the cell neglects product formation and focusses on its growth instead. This was observed in all fed-batch fermentations and was already discussed elsewhere [2]. Regarding the specific soluble product formation rate, even though explicit trends are observable, some inconsistencies/fluctuations in the plotted values are derived due to the measurement error in the off-line analytics. Here, it is also seen that higher values are obtained at slower specific growth rate settings, even though the induction strength and temperature were the same.

Exemplary cases for this are DoE #22 (Fig. 2SA) and DoE #4 (Fig. 2SB). Both were performed at a cultivation temperature of 30°C and an induction strength of 0.5, but DoE #22 with μ 0.20 and DoE #4 with μ 0.10, leading to an increased specific soluble product formation rate. The highest values were observed for the fed-batch fermentations performed at induction strength 0.9, respectively DoE #9 (Fig. 2SC) and DoE #17 (Fig. 2SD). However, these rates started to decrease halfway through the fed-batch due to the high metabolic burden and in the case of DoE #9 almost dropped to zero. This happened due to the formation of inclusion bodies at this particular CPP setting, therefore no soluble product formation took place anymore.

At the induction strength 0.2 the lowest impact on the specific growth rate as well as on the specific soluble product formation rate was observed, i.e., DoE #2 (Fig. 2SE) and DoE #21 (Fig. 2SF). The specific rates of these two fed-batch fermentations also followed the same pattern as described above, i.e., a slow specific growth resulted in a higher specific soluble product formation rate. However, at this induction strength, the slope of the specific soluble product formation rate only increased very slowly, compared to the other fed-batch fermentations. The fed-batch fermentations performed at the higher induction strengths displayed the same values shortly after the induction, which the two fed-batch fermentations performed at induction strength 0.2 reached only after three-fourths of the whole process. The different expression patterns of these fed-batch fermentations, especially DoE #21, could also be the reason why the hybrid model struggled to predict these values accurately. Moreover, as it is seen in the replicates, the specific growth rate has a rather small standard deviation. This enables a more accurate

prediction for the hybrid model compared to the product formation rate, for which higher standard deviations were computed for all CPP combination settings.

Performance Comparison of the Time-Resolved Models

The two developed time-resolved models (the ANN black-box model and the hybrid model) are presented in the sections Material & Methods (2.4) and Results (3.1 & 3.2) of the main manuscript. Herein, it has been shown that the bootstrap-aggregated hybrid model, in general, outperforms the bootstrap-aggregated black-box model. This complete comparison of the time-resolved model performances applied to the test set is given in Table S2. Due to the fact that there is no detectable product formation until induction, the percentage and relative percentage cannot be calculated for this stage.

Table S2. Performance results of the bootstrap-aggregated black-box model and the bootstrap-aggregated hybrid model for predicting the biomass concentration and the soluble product titer over the entire process. For each process variable the R^2, RMSE and (if accessible) the percentage error is given (rounded to two decimals) for each model. The presented values refer to the model performances obtained for the respective test set. The number of fed-batch fermentations used for model-training and model-testing is indicated in brackets

entire process	**Biomass**		**Product**	
Performance Criteria	ANN Black-Box Model (n = 25+6)	Hybrid Model (n = 25+6)	ANN Black-Box Model (n = 25+6)	Hybrid Model (n = 25+6)
R^2	0.75	0.98	0.96	0.97
RMSE [g/L]	4.36	1.10	0.25	0.22
error [%]	17.50	4.24	-	-

References

[1] B. Bayer, B. Sissolak, M. Duerkop, M. Von Stosch, G. Striedner, *Bioprocess Biosyst. Eng.*, **2020**, *43,* 169.

[2] B. Bayer, M. Von Stosch, M. Melcher, M. Duerkop, G. Striedner, *Eng. Life Sci.*, **2020**, *20,* 26.

Publication IV

Biotechnology Journal
www.biotechnology-journal.com

Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization

*Benjamin Bayer, Gerald Striedner, and Mark Duerkop**

Process characterization is necessary in the biopharmaceutical industry, leading to concepts such as design of experiments (DoE) in combination with process modeling. However, these methods still have shortcomings, including large numbers of required experiments. The concept of intensified design of experiments (iDoE) is proposed, that is, intra-experimental shifts of critical process parameters (CPP) that combine with hybrid modeling to more rapidly screen a particular design space. To demonstrate these advantages, a comprehensive experimental design of *Escherichia coli (E. coli)* fed-batch cultivations (20 L) producing recombinant human superoxide dismutase is presented. The accuracy of hybrid models trained on iDoE and on a fractional-factorial design is evaluated, without intra-experimental shifts, to simultaneously predict the biomass concentration and product titer of the full-factorial design. The hybrid model trained on data from the iDoE describes the biomass and product at each time point for the full-factorial design with high and adequate accuracy. The fractional-factorial hybrid model demonstrates inferior accuracy and precision compared to the intensified approach. Moreover, the intensified hybrid model only required one-third of the data for model training compared to the full-factorial description, resulting in a reduced experimental effort of >66%. Thus, this combinatorial approach has the potential to accelerate bioprocess characterization.

1. Introduction

1.1. Process Characterization

The need for a paradigm shift in the biopharmaceutical industry for quality assurance has long been recognized. This was put into focus by the introduction of the process analytical technology (PAT) guide by the U.S. Food and Drug Administration (FDA). This guide calls for enhanced process understanding that emphasizes a new quality by design (QbD) approach, in which the requested product quality is assured by the process itself and does not have to be tested afterward.[1] Among the most prominent approaches to enable QbD is design of experiments (DoE), which is used in studies to gather process knowledge.[2] To set up a DoE for a product, critical process parameters (CPP) that impact the products' critical quality attributes (CQA) must be defined.[3] To systematically investigate these CPP combinations and their multifactorial influence and to keep the number of experiments manageable, different designs can be applied, for example, Doehlert, Box–Behnken and central composite designs.[4] Process modeling is frequently applied to evaluate such designs.[5]

1.2. Process Modeling

The most common modeling technique, used in combination with DoE studies, is the response surface methodology. In this technique, the experimental results of a design space are represented on a surface as responses of the CPPs, and this is used to find the optima for the investigated conditions.[6] This technique is widely used in media development and optimization for production processes, for example, utilizing an additional amino acid feed for Chinese hamster ovary cells or reformulating macronutrients in *Bacillus sp.* for product titer enhancement.[7–9] Time-resolved process models have also increased in popularity and applicability. There is a distinction between descriptive and predictive models. Descriptive models, such as soft sensors derived from spectral data (e.g., Raman spectroscopy, near-infrared spectroscopy, or 2D fluorescence), provide real-time information, that is, certain spectral information for only up to the current time point of the process.[10,11] Predictive models, on the other hand, are able to predict future values of state variables and, since the input data can be simulated for the future, provide an educated guess about the trajectory.[12] In addition, exploratory data analysis, such as principal component analysis (PCA) and parallel factor analysis (PARAFAC), are commonly applied to expand the existing process knowledge. With this means, hidden structures and latent

B. Bayer, Prof. G. Striedner, Dr. M. Duerkop
Department of Biotechnology
University of Natural Resources and Life Sciences
Vienna 1190, Austria
E-mail: mark.duerkop@novasign.at
Dr. M. Duerkop
Novasign GmbH
Vienna 1190, Austria

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/biot.202000121

DOI: 10.1002/biot.202000121

variables, the so-called principal components (PC), in the investigated spectra can be determined.[13,14] Generally, there are two different modeling approaches: nonparametric (black-box) and parametric (white-box) models. Nonparametric models are built on experimental data only and do not need any further process knowledge. Various regression techniques are available and commonly applied to develop nonparametric models.[15] In contrast, parametric models use empirical knowledge and first principles, that is, their structure is well defined and transparent.[16] Both modeling approaches possess separate unique advantages as well as disadvantages and limitations due to their respective model structures.

1.3. Hybrid Modeling

The concept of combining a nonparametric and a parametric model into a single semi-parametric model structure is called hybrid modeling. This allows the incorporation of both process knowledge and data-driven information. The hybrid model structure overcomes the shortcomings of each separate modeling technique, for example, the black box can be used to calculate parameters in the white box, which therefore do not have to be solely assumed, reducing errors[17] at the cost of increased complexity. For instance, the values of specific rate expressions are known unknowns a priori to a bioprocess and must first be determined, for example, by process modeling. However, by solely utilizing a white-box model, these rate values must be assumed from data using a defined causal method, but in a hybrid model, they first can be estimated in a defined black box and then transferred to the white box.

For example, by utilizing process variables that have an influence on these rates as input to the black box, the incorporation of this impact can also be taken into account in the white box, generating hybrid model predictions closer to the analytical values. Artificial neural networks (ANN) are frequently utilized for this process.[18] Accurate rate estimations are of great importance for robust bioprocess modeling, and achieving these as precisely as possible is of high interest.[19] Due to these advantages, hybrid modeling is gaining in popularity for bioprocess modeling. Even though a hybrid model provides improved performance compared to other approaches,[20] the possibility of misprediction still exists. To ascertain the chance of such model uncertainty, cross validation is commonly performed in machine learning to calculate the average misprediction possibility.[21] However, bootstrapping can also be applied for this task and has been proven to be a more flexible technique, allowing full control over developing the final model. In this method, a number of models are merged into one, which leads to a probability of misinterpretation and risk assessment occurring from different data permutations.[22] However, these techniques are linked to an increased computational workload, since several hybrid models must be developed. This workload linearly increases, either with the number of chosen folds for the cross validation or the number of applied bootstraps.

The combination of both elements, hybrid modeling and bootstrapping, provides a robust and reliable hybrid model for bioprocess modeling. Nevertheless, the experimental workload, that is, the generation of the required process data and the related analytical effort, rapidly ends up being laborious and time consuming.

1.4. Intensified Design of Experiments

A promising approach to reducing the experimental workload is to change the CPPs during the cultivation. With these intra-experimental CPP set-point changes, the reaction to dynamic changes in the process can be captured.[23] By performing this intensified DoE (iDoE), one should note that the history of the cell contributing to a memory effect is often assumed to influence how the cells react to subsequent CPP set points.[24] One of the main challenges is to describe the process dynamics in response to intra-experimental changes and to estimate the behavior of the cells under constant conditions. Therefore, a time-resolved hybrid model can be built on iDoE data to describe the occurring process dynamics,[25] because it captures the whole process. This emphasizes a combinatorial approach, using hybrid modeling and iDoE, to generate process knowledge and simultaneously accelerate process characterization.

1.5. A Combinatorial Approach Leading to Accelerated Process Development

To significantly reduce the number of required experiments for developing a process model, we present the concept of iDoE. As basis for comparison, we used a completely characterized three-dimensional design space of a previously derived *Escherichia coli* (*E. coli*) fed-batch study at the 20L scale with 27 distinct CPP combination settings.[20] The intensified experiments were performed in the same design space but contained two CPP set-point shifts during each cultivation, so that three CPP combination settings were tested overall within one fed-batch fermentation. This led to nine iDoE cultivations to completely characterize the same design space. Consecutively, to examine a possible memory effect of the cells, the online process data and the 2D-fluorescence spectra of the static and intensified fed-batch fermentations were investigated and compared using exploratory data analysis (PCA and PARAFAC) to test for any differences.

A hybrid model was built on iDoE data, and its performance was compared to a previously developed full-factorial static hybrid model, built on the complete design space. To further challenge the iDoE approach, a fractional-factorial static hybrid model, built only on the center point and corners of the design space, was assessed to challenge the potential time reduction and advantages regarding process characterization using iDoE.

2. Experimental Section

2.1. Experimental Design

For all fed-batch cultivations, *E. coli* (HMS174 (DE3)) was utilized for expressing recombinant human superoxide dismutase at the 20 L scale. The experimental design consisted of a full-factorial design space with three CPPs: the specific growth rate (μ) controlled by the substrate feeding rate, the cultivation temperature

(T), and the induction strength (I). The values for the three levels, respectively, are $\mu = 0.10$, 0.15, and 0.20 h^{-1}; $T = 30$, 34, and 37 °C; and $I = 0.2$, 0.5, and 0.9 μmol IPTG g^{-1} cell dry mass. This results in 27 CPP combination settings, as presented elsewhere.[26] The complete list of all performed fed-batch cultivations of the DoE is given in Table S1, Supporting Information. For all these cultivations in the design space, the analytical measurements for the biomass concentration (in g L^{-1}) and the soluble product titer (in g L^{-1}) were assessed by thermogravimetric analysis[27] and ELISA,[28] respectively. The analytical error of the biomass and product titer determination was assessed from seven replicate runs in a previous study[20] with 3.6% and 7.6%, respectively. The fed-batch phase was always carried out for four doubling times. The induction of the cells took place after the first doubling time, enabling recombinant protein production for three doubling times. All information about the applied exponential feeding strategy for the fed-batch phase, the utilized *E. coli* strain, the expression vector system, the online monitoring, and the offline measurements were presented elsewhere.[29–31] In addition to the standard online available process variables, such as pH, temperature, inlet air, stirrer speed, base consumption, accumulated feed, inducer, and head pressure, a 2D fluorescence probe (BioView, Delta Light and Optics, Denmark) was utilized to measure the cultivation broth in 20 nm steps (from ex270/em310 up to ex550/em590), resulting in 120 excitation/emission wavelength variables. These measurements were used to examine if differences on the cellular and process level are visible between the DoE and iDoE.

To investigate and quantify the metabolic burden, possible toxicity, and induced stress due to recombinant protein production, the production load (PL), a summary of all these factors, was utilized.[32] Therefore, fed-batch cultivations outside of the presented design space were performed at an induction strength of 0, that is, the same CPP combination settings as above for the cultivation temperature and the specific growth rate were used but without induction. These fed-batch cultivations are listed in Table S3, Supporting Information, and the results of the investigation of the PL are presented in Figure S6, Supporting Information.

2.2. Intensified Design of Experiments

Two intra-experimental shifts from one CPP set point to another were performed in each fed-batch fermentation to cover three different parameter combinations of the design space within one fermentation. The complete list of all iDoE CPP combination settings and the performed shifts per fed-batch fermentation are provided in Table S2, Supporting Information. These shifts were done in compliance with already published constraints.[25] Since the inducer was not consumed by the cells, a shift toward lower inducer concentrations was not feasible without heavy and impractical dilution of the fermentation broth. Therefore, the 3D design space was subdivided into three 2D induction planes for the iDoE approach, and shifts were only performed for the temperature and the specific growth rate in the respective induction plane. These shifts were carried out after each theoretically calculated cell doubling, post induction, leading to three phases per fed-batch fermentation. The intensified fed-batch fermentations of the three induction planes were coordinated to guarantee that each CPP combination setting was passed in every phase if overlaid. **Figure 1** provides a detailed graphical overview of the operating procedure of the intra-experimental shifts and the performed intensified fed-batch fermentations, shown in the design space and separated in the induction planes.

2.3. Data Sets

The static data set derived from an earlier study consisted of 31 fed-batch fermentations (27 CPP combination settings and four replicates) covering the complete design space.[20] Values for the standard online process parameters were available every minute, while the measurement frequency of the 2D fluorescence probe leads to a value every three minutes. The biomass was measured a single time before induction and then hourly, and the soluble product titer was measured every 2 h from the time point of induction to the last sampling at the end of the process. In total, 589 samples to determine the biomass concentration and 306 samples to analyze the product titer were acquired.

The data set containing the intensified fed-batch fermentations consisted of nine cultivations that were designed to cover the complete design space. The sampling interval and analytical methods for the biomass and soluble product titer analysis were performed as with the static data set, with the sampling interval increased to 30 min after each shift for 2–3 h. In total, 213 samples to determine the biomass concentration and 153 samples to analyze the product titer were acquired. The online available process variables were recorded at the same frequency as that for the static cultivations. The detailed analytical results for the biomass and the soluble product titer of the intensified fed-batch fermentations are provided in Figure S1, Supporting Information. Also, a detailed example of how the CPP changes affect the variables to be modeled and how rapidly these adapt to the new CPP set points is provided in Figure S2, Supporting Information. Further, to exclude a potential memory effect due to the direction of the CPP shifts, the comparison of one experiment, performed reversely to iDoE #3, is presented in Figure S3, Supporting Information.

2.4. Data Preprocessing

The data used were stored as Excel spreadsheets with columns representing variables and rows representing observations. Prior to exploratory data analysis and process modeling, every measurement of the available online variables was standardized, along with the time domain, using the z score. This procedure was done to exclude quantitative effects and to specifically account for the change over time. If there was a missing analytical value at a sampling time point for one of the two target variables, the missing value was interpolated using Hermite polynomials, which guaranteed an equally weighted and valid evaluation.

2.5. PCA and PARAFAC

PCA and PARAFAC, as described by Bro,[33] were applied for exploratory data analysis of the online available process data. Both

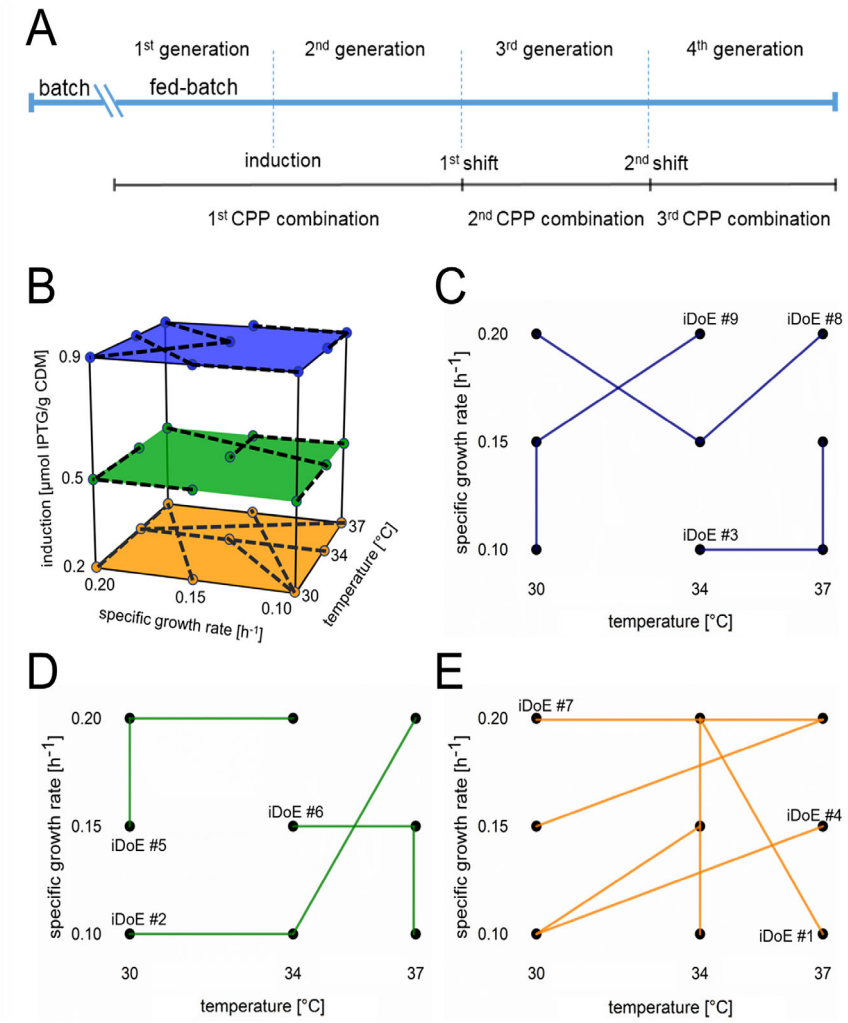

Figure 1. Operating procedure of the intensified fed-batch fermentations. A) The fed-batch phase subdivided into the single doubling times. The time point of induction and the time point of the CPP shifts, switching to a different point in the design space each after one theoretical doubling time, are indicated. B) The intensified fed-batch fermentations are presented as the entire design space and separately for each induction plane, that is, C) blue 0.9, D) green 0.5, and E) orange 0.2. For (C–E), the starting points of the iDoE fermentations are depicted by the respective ID.

techniques were performed with MATLAB (2016b, MathWorks, USA) and two freely available toolboxes, N-way[34] and drEEM.[35] PCA was performed on the complete online process data, that is, the standard process variables and the 2D fluorescence data from each approach (DoE and iDoE). PARAFAC was applied to the 2D fluorescence data only. The more detailed and complete comparison of the static and intensified fed-batch cultivations is presented in Figure S5, Supporting Information.

2.6. Hybrid Modeling

2.6.1. Data Sets

Different data sets were used to train the hybrid models:

1) Full-factorial static hybrid model: The first static hybrid model, used as the qualitative reference and derived from an earlier publication,[20] consisted of 25 static fed-batch fermentations (DoE #1, #3–16, #18–20, and #22–27) for model training and 6 static fed-batch fermentations (DoE #2, #4, #9, #17, #21, and #22), including one fermentation each from the duplicate and triplicate runs and three runs chosen by randomization, for model testing ($N = 25 + 6$).
2) Fractional-factorial static hybrid model: The second static hybrid model was likewise developed as a full-factorial static hybrid model counterpart, but only nine static fed-batch fermentations, the center point, and the corners of the design space, that is, a fractional-factorial design, were used for model training (DoE #1, #3, #7, #9, #14, #19, #21, #25, and #27). This model was developed to allow a comparison between the fractional-factorial static and the iDoE approach with respect to the same amount of input data for model training. The test set contained all static fermentations (DoE #1–27) ($N = 9 + 31$).
3) iDoE hybrid model: To build the third hybrid model, based on iDoE, all intensified fed-batch fermentations (iDoE #1–9) were considered. To allow a comparison between the full-static and the iDoE hybrid model, the same six static fed-batch fermentations as for the static hybrid model (DoE #2, #4, #9,

#17, #21, and #22) were initially used as the test set ($N = 9 + 6$).
4) In addition, for a full comparison of how a model based on iDoE can describe the static design space, a second test set containing all static fermentations was introduced (DoE #1–27) ($N = 9 + 31$).

A graphical overview of the respective utilized experiments used for training the three hybrid models is presented in Figure S4, Supporting Information. Hybrid model development and evaluation were accomplished in the stand-alone C# hybrid modeling toolbox (Novasign GmbH, Vienna, Austria), which can be downloaded. Furthermore, the static and the intensified data sets used for modeling are provided as Supporting Information. These are preprocessed and can be used for individual modeling purposes.

2.6.2. Nonparametric Black Box

To predict the values of the response variables, a serial hybrid model structure was chosen. The nonparametric model, an ANN that applies a Levenberg-Marquardt algorithm and is embedded in the hybrid model, was applied to model the known unknowns for the parametric part, that is, the specific growth rate (μ (h^{-1})) and the soluble product formation rate ($\nu_{p/x}$ (g g^{-1} h^{-1})) as propagated predictions.

The ANN had three layers. The nodes of the hidden layer used tangential hyperbolic transfer functions, while the input and output layers used linear transfer functions. There were three inputs: the cultivation temperature (°C), the cumulative inductor mass (mg), and the cumulative feed (L).

2.6.3. Parametric White Box

The hybrid model was developed based on material balances, which were derived for biomass and the soluble product titer assuming an ideally mixed fed-batch reactor. Further, it was assumed that the biomass catalyzes all reactions, therefore, specific rates were used. That is, the estimated rate expressions derived from the nonparametric part are used in the parametric part, as shown in Equation 1 and Equation 2. These equations assume an ideal population, that is, 100 % producing cells and do not consider any emerging subpopulation due to the PL.

$$\frac{dX}{dt} = \mu \cdot X - D \cdot X \tag{1}$$

$$\frac{dP}{dt} = \nu_{p/x} \cdot X \cdot I_{y/n} - D \cdot P \tag{2}$$

where X is the biomass concentration (g L^{-1}), P is the soluble product titer (g L^{-1}), $I_{y/n}$ is the inductor switch (set to either zero for no induction or one for induction), and D is the dilution rate (h^{-1}) to describe the relationship between feed addition (L h^{-1}) and the reactor volume (L).

2.6.4. Model Validation

The performance of the model with respect to the fit of the experimental data was evaluated using the root mean square error (RMSE) (Equation 3) and the normalized RMSE (NRMSE) (Equation 4). This calculation used the measured value (y), its estimated counterpart ($\hat{y}$) for each sampling point (t), the mean of the measured values ($\bar{y}$), and the total number of observations (N).

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum \left(y_{(t)} - \hat{y}_{(t)}\right)^2} \tag{3}$$

$$NRMSE\ [\%] = \frac{RMSE}{\bar{y}} \cdot 100 \tag{4}$$

The model was validated using internal cross validation, that is, in the beginning, the hybrid model was derived using the training data. The data were split into training and validation partitions. The training partition was used to build the model, which was then applied to the remaining validation partition. Once no further model improvement was achieved, the model training stopped.

This data partitioning and model development was repeated nine times to account for all possible permutations of eight training and one validation data set. By studying different numbers of nodes, two to eight in steps of one, in the hidden layer of the embedded ANN, the ANN parameters were identified, and four nodes in a single hidden layer were chosen that give the best performance for the fractional-factorial and iDoE hybrid models.

2.6.5. Bootstrap Aggregation

The assessment of the risk of model misprediction based on the random data partitioning during model building used bootstrap aggregation of the individual hybrid models,[36] which can be imagined as model averaging. This averaging of the predictions of multiple models into one gives the operator more control in model selection and represents a robust way to deal with model uncertainties. This approach is similar to a leave-one-batch-out cross validation approach but allows for better control to select individual models of each boot. The bootstrap-aggregated fractional-factorial and iDoE hybrid models each consisted of five individual models, each derived from a different boot, for which the standard deviation (SD) (Equation 5) and the prediction interval (PI) (Equation 6) were calculated to access the model performance:

$$SD_{(t)} = \sqrt{\frac{1}{n-1} \cdot \sum \left(\hat{y}_{btstrp(t)} - \hat{y}_{model(i)_{(t)}}\right)^2} \tag{5}$$

$$PI_{(t)} = \hat{y}_{btstrp(t)} \pm SD_{(t)} \tag{6}$$

Therefore, the value of the bootstrap-aggregated prediction ($\hat{y}_{btstrp}$), the predicted counterpart from the respective model ($\hat{y}_{model}$), the index ($i = 1{:}5$), and the number of observations for each time point (n) were used. Each generated hybrid model was

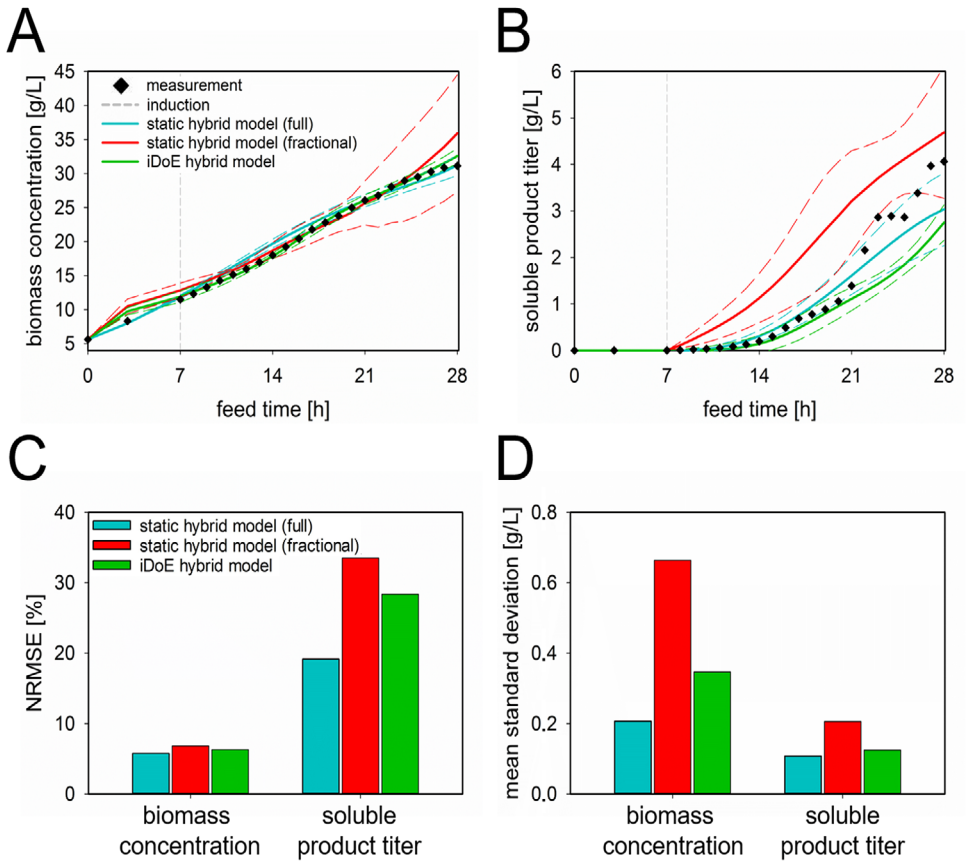

Figure 2. Comparative evaluation of the predictive quality of the three developed hybrid models. The model predictions for one exemplary fed-batch fermentation from the test set (DoE #2) are displayed for A) the biomass concentration and B) the soluble product titer. The analytical results (squares), the time point of induction (dashed gray line), and the predictions of the hybrid models (solid lines), including the respective PIs (dashed lines), are indicated: the full-factorial static hybrid model (turquoise), the fractional-factorial static hybrid model (red), and the iDoE hybrid model (green). C) The NRMSE and D) the mean SD of the model predictions for the complete respective test set, using the same color code as above.

derived from a different boot to ensure high generalization ability. This bootstrap-aggregated hybrid model was used for model testing to assess the predictability of the model on new data (external validation) and to investigate the risk of predictive uncertainty.

3. Results

3.1. Setup of the Intensified Design of Experiments

The general operating procedure for the intensified fed-batch fermentations, including the CPP shifts, is presented in Figure 1. Figure 1A indicates the biomass doubling times (generations), the time point of induction, and the performed CPP shifts, that is, the switch to different parameter combinations in the design space, always after one calculated doubling time. Also, the complete design space (Figure 1B) and the more detailed operating scheme of each induction plane are shown, including the location of the starting CPP combination setting for each intensified fed-batch fermentation and the setting after each shift (Figure 1C-E). The intensified experiments were performed so that, if the induction planes are overlaid, each CPP combination setting is characterized in every cell generation.

The comparability of the DoE and iDoE approaches on a cellular level was assessed. The impact of the intra-experimental CPP shifts on the biomass concentration and the soluble product titer is presented in Figure S1, Supporting Information. It was demonstrated that the cells rapidly adapt to new CPP combination settings after a CPP shift (Figure S2, Supporting Information) on the basis of an exemplary iDoE cultivation. To exclude any possible memory effect caused by the CPP shifts, that is, altered behavior with respect to the investigated process variables due to previous CPP combination settings, an additional experiment and exploratory data analysis of the online process data were conducted (Figures S3 and S5, Supporting Information). This exploratory data analysis examined the 2D fluorescence spectra of the static and iDoE approaches as well as the standard online available process variables linked to biomass formation, that is, the cultivation temperature, the accumulated feed, the accumulated inductor, and the base consumption.

3.2. Performance of the Developed Hybrid Models

To investigate and compare the predictive performance of the three developed hybrid models for the biomass and soluble product titer, one exemplary fed-batch fermentation from the test set is presented (**Figure 2**A,B). This fed-batch fermentation (DoE #2) was chosen since it was present in every test set and was not a replicated cultivation of the training set. Further, the respective

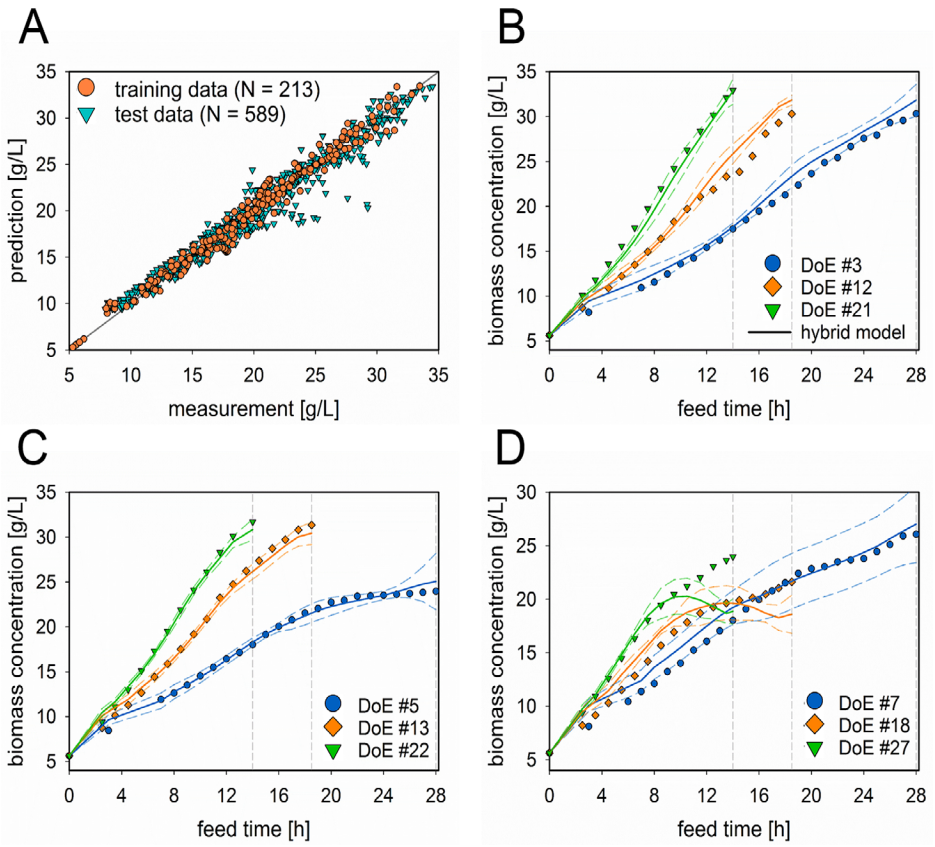

Figure 3. Performance of the bootstrap-aggregated iDoE hybrid model predicting the biomass concentration of the 31 test cultivations. A) The scatter plot of the hybrid model on the training data (orange dots) and the test data (cyan triangles). The model predictions for the individual fed-batch fermentations are displayed for each induction strength. For B) I 0.2, C) I 0.5, and D) I 0.9, the analytical results (symbols), the respective prediction (colored lines), and the PI (dashed lines) of the iDoE hybrid model are indicated. The IDs of the presented fed-batch fermentations are listed.

risk of misprediction, that is, the PI, as well as the NRMSE and the mean SD of the entire test set of each hybrid model, were also incorporated in this comparative evaluation.

This evaluation revealed that all hybrid models perform on a comparable level with respect to predicting the biomass. Regarding the PIs, the full-factorial static hybrid model and the iDoE hybrid model perform on a comparable level. In contrast, the predictions from the fractional-factorial static hybrid model are less precise, displaying broad PIs. A broad PI indicates a high risk of misprediction, because the different models selected for bootstrapping are very different in their prediction. This was observed for both process variables: the biomass concentration (Figure 2A) and the soluble product titer (Figure 2B). Regarding the NRMSE (Figure 2C) and mean SD (Figure 2D), a direct comparison shows that the full-factorial static hybrid model displays the lowest values for both process variables. The fractional-factorial static hybrid model displayed the highest values for both the NRMSE and the mean SD. The iDoE hybrid model has inferior performance compared to the full-factorial static hybrid model, but its performance is superior to the fractional-factorial static hybrid model.

3.3. iDoE Hybrid Model Performance on Predicting the Biomass Concentration

A comprehensive demonstration of the performance of the iDoE hybrid model using the test set is presented in **Figure 3**. The iDoE hybrid model predicted the biomass concentration for all 31 static cultivations in the scatter plot (Figure 3A) and on nine particular static fed-batch fermentations, that is, three fed-batch fermentations per induction strength, each performed with one of the three intended specific growth rates (Figure 3B–D).

The iDoE hybrid model displayed an exceptional ability to predict the analytical biomass results of the static runs with high accuracy over the entire cultivation time. With induction strengths of 0.2 (Figure 3B) and 0.5 (Figure 3C), the model predictions for all fed-batch fermentations were highly accurate, including the tightly distributed PIs. At the induction strength of 0.9 (Figure 3D), higher CPP impacts on the state variables were observed, impeding the predictions. This was also visible by the broader PIs. The presented static fed-batch fermentations DoE #18 and DoE #27 are not perfectly covered by the iDoE hybrid model, which displayed a decrease in the biomass concentration. This misprediction was not observed for the third representative cultivation (DoE #7), for which the biomass trend was predicted well.

3.4. iDoE Hybrid Model Performance on Predicting the Soluble Product Titer

As with the biomass concentration, the performance of the iDoE hybrid model for predicting the soluble product titer of all 31 static cultivations is showcased in **Figure 4**. To obtain a consistent

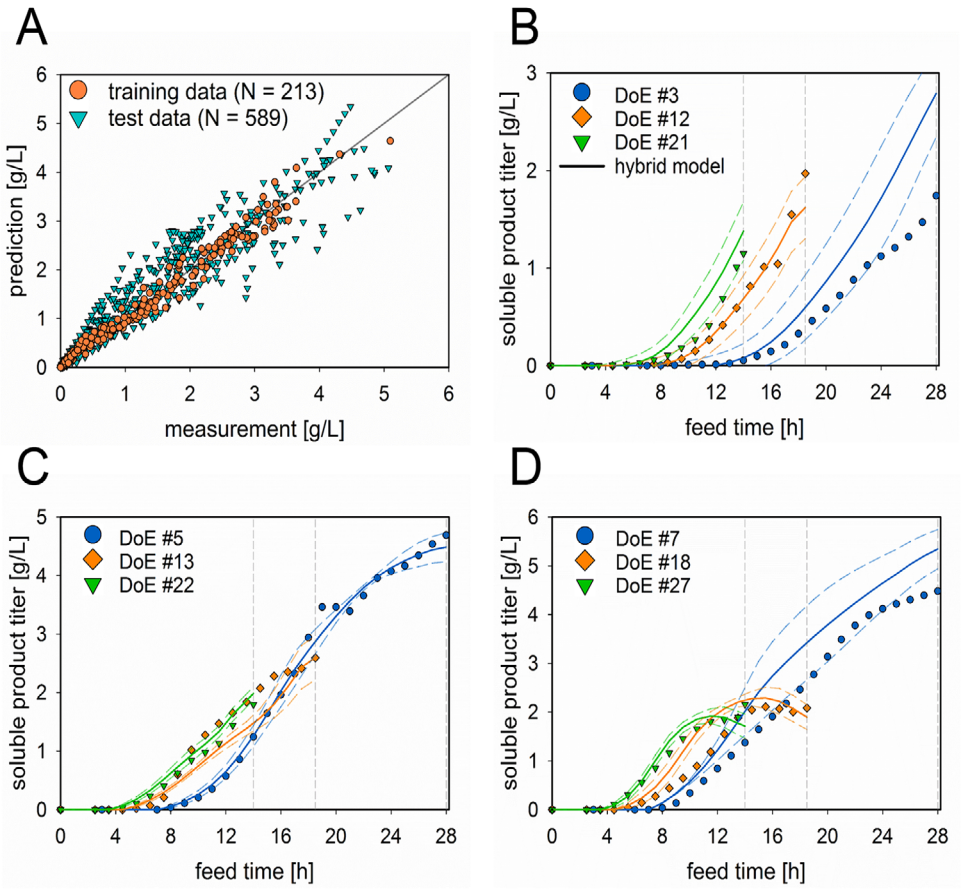

Figure 4. Performance of the bootstrap-aggregated iDoE hybrid model in predicting the soluble product titer of the 31 test cultivations. A) The scatter plot of the hybrid model on the training data (orange dots) and the test data (cyan triangles). The model predictions for the individual fed-batch fermentations are displayed for each induction strength. For B) I 0.2, C) I 0.5, and D) I 0.9, the analytical results (symbols), the respective prediction (solid lines), and the PI (dashed lines) of the iDoE hybrid model are indicated. The IDs of the presented fed-batch fermentations are listed.

impression of the iDoE model quality, the cultivations shown are the same as those in Figure 3.

The iDoE hybrid model was able to predict the soluble product titer of the full-factorial design space with adequate accuracy, as presented for the individual induction strengths. For the induction strength of 0.2 (Figure 4B), accurate predictions were observed for two out of three cultivations. Only cultivation DoE #3 displayed an overestimation of the analytical values. Similar behavior was also observed for one cultivation that was performed with an induction strength of 0.5 (Figure 4C), namely DoE #13, while predictions of the other two cultivations matched the analytical values and displayed small PIs. Similar results were obtained for a strength of 0.9 (Figure 4D). The trend for two cultivations was not predicted completely well (DoE #7 and DoE #18), while the remaining cultivation was predicted accurately.

A complete comparison of the predictive quality of the static and intensified bootstrap-aggregated hybrid models, applied on the respective test sets, is presented in Table 1 using the R^2, RMSE, and NRMSE as criteria for model comparison.

Table 1. Model comparison of the developed bootstrap-aggregated static hybrid models and the bootstrap-aggregated iDoE hybrid model. The results of the three models applied to the respective test sets are presented, including the respective R^2, RMSE, and NRMSE, all rounded to two decimal places, and the required number of experiments. The number of fed-batch fermentations used for model training and model testing is indicated in brackets.

	Full-factorial static hybrid model[20] ($N = 25 + 6$)		iDoE hybrid model ($N = 9 + 6$)		Fractional-factorial static hybrid model ($N = 9 + 31$)		iDoE hybrid model ($N = 9 + 31$)	
Target variable	Biomass	Product	Biomass	Product	Biomass	Product	Biomass	Product
R^2	0.98	0.97	0.98	0.88	0.97	0.91	0.97	0.91
RMSE [g L^{-1}]	1.10	0.22	1.12	0.33	1.29	0.39	1.19	0.33
NRMSE [%]	5.77	19.14	5.86	28.65	6.83	33.53	6.30	28.37
No. of experiments	31		9		9		9	

4. Discussion

The evaluation of the iDoE approach (Figure 1) on a full -factorial design space and the quality of the generated data sets was of primary interest in this study. The investigation into the comparability of the DoE and iDoE setups on a cellular level was crucial and of high interest for the consecutive modeling steps. Applying the described intra-experimental shifts, it was demonstrated that the cells can cope with the iDoE setup and the easy to express recombinant human superoxide dismutase. We showed that cells are able to rapidly adapt to new process conditions within 1 h after the change of conditions (Figure S2, Supporting Information) and, regarding the outcome, that the shift direction does not matter (Figure S3, Supporting Information). Hence, the defined intervals between the changes of process conditions were not too short, providing cells adequate residence times for each CPP setting. Moreover, the exploratory data analysis of the DoE and iDoE data in Figure S5, Supporting Information, displayed no significant differences on the cellular and process level between both approaches. These results also strongly support the assumption that appropriate CPP shifts do not provoke persistent cellular memory effects. In addition to information on dynamics in response to changes, iDoE data displays similar information content as data from conventional static experiments, which can be seen as sound basis for more detailed data interpretation via modeling approaches. However, the general usability of the iDoE approach must be investigated, for example, for more input factors, a variety of target proteins with different characteristics, for example, cytotoxicity, and especially the applicability on other organisms.

From a modeling and prediction perspective, hybrid models utilizing either iDoE or DoE data were evaluated with respect to their prediction performance. The hybrid model established with iDoE data was able to predict the biomass for the test data set (Figure 3), containing all static DoE cultivations ($N = 31$), with an accuracy similar to the static full-factorial hybrid model. There was also good accordance with the analytical error of the biomass determination (3.6%). The prediction performance for soluble product titer was on an acceptable level and again comparable to predictions with the full-factorial hybrid model (Figure 4). In general, it also has to be kept in mind that the test set for the full-factorial static hybrid model consisted of only six cultivations which were used to calculate the RMSE, NRMSE, and mean SD. Therefore, its performance in comparison to the iDoE hybrid model, using 31 cultivations in the test set, should not be overrated. We also verified the potential of the idea to save time and costs simply by reducing the number of static experiments. Therefore, a fractional-factorial data set, comprising only the center point and the corners of the static DoE, was used for model building (Figure S4, Supporting Information). This approach resulted in a model with significantly reduced prediction quality and a strongly increased model uncertainty (Figure 2). The results, summarized in Table 1, clearly demonstrate the superiority of iDoE data which is most probably based on a significantly increased information entropy, as every single iDoE experiment contains data from three different CPP settings. With the prospect for process control, accurately modeling the response to a variable that changes over time is a great advantage. Even though the inputs to the ANN, that is, the three chosen CPPs, do not fully describe all possible process responses, these are easily controllable, thus enabling model predictive control applications in the future.

Further, the somehow limited prediction performance for the soluble product titer observed for all hybrid models built in this study is most likely caused by the formation of nonproducing subpopulations during the induction phase of the process. In *E. coli* cell banks, the presence of a small population of plasmid-free cells is a known phenomenon and even application of selection pressure along the production process rendered to be of limited efficiency in suppressing this subpopulation during the production phase.[37] As we assume a homogeneous population of producer cells, in Equation 2, this decoupled the biomass from the product formation. The key problem in this context is that there is no information on the distribution of producer and nonproducer in our datasets. There is no analytical method available that facilitates differentiation between producing and nonproducing subpopulations with the required accuracy or, for example, without introducing an additional fluorescent protein, which is not applicable for industrial production processes. As the load level, triggered by product formation, directly impacts the difference in growth rates of these subpopulations, we introduced the PL concept. The obtained PL values (up to 30%), presented in Figure S6, Supporting Information, were in good accordance with the reference literature[30] and, further, are reasonable from a metabolic point of view, that is, an increase in the induction strength, as well as the cultivation temperature, raised the PL. We are aware that the pure description of the PL is not the solution to the occurring limitations of accurately predicting the soluble product titer but rather a starting point to this multidimensional restriction of the model performance, which we are not able to fully explain. However, this limitation will remain at the moment, since a precise analytical method to approach this nonproducing population is not available. As the limiting problem for more accurate predictions is known, we anticipate a predictive improvement by incorporating a suitable term in the white box, that is, solely taking the producer population for the product formation into account and not the entire predicted biomass. This again highlights the advantages of knowledge incorporation using hybrid models. However, for systems with constitutive product formation, and therefore without selection pressure introduced by induction, the predictability of the product is assumed to be much higher.

In conclusion, the concept of performing iDoE is rather new in upstream processing, and it has never before been tested in a comprehensive comparison between static and dynamic cultivations of a full-factorial design space. Besides the more common considerations, for example, which process variables will be analyzed and which data will be measured, a major and more conceptional matter must be considered when an iDoE is designed: the conduction of the intra-experimental CPP shifts. Adequate time for the adaption of the cells in each phase, the number of shifts per cultivation, and a reasonable direction of all shifts and their magnitudes are highly important in generating meaningful process data.

Further, we do not propose iDoE for the discovery of an unknown design space where particular CPP combinations can lead to irreversible cell damage. If cells do not recover after a shift, the following learning rate of the hybrid model might be very low and inaccurate. Additionally, the number of required cultivations was

significantly reduced compared to the full-factorial static hybrid model, that is, the application of iDoE in this study led to an acceleration of the process characterization time by more than 66%, while the analytical effort increased by 20% due to the temporal higher sampling frequency. This is also highly valuable from the economical aspect to keep up in terms of budget restrictions, that is, saving working hours and raw materials for the preparation and execution of the fermentation of two thirds of the complete experimental setup.

Furthermore, the iDoE provided a lot of information about the behavior of cells to changes in the process which enabled the development of a hybrid model with a good generalization ability. Since a lot of information was gathered on how cells react on process changes, utilizing the iDoE hybrid model structure for advanced process control is the logical next step. The great potential for using iDoE in combination with hybrid models to speed up process characterization, as was demonstrated in this work, is of high interest for the biopharmaceutical industry, for example, to keep up with timelines, budget restrictions, and return on investments.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors would like to thank the Austrian Research Promotion Agency (FFG) for their support (Research Studio Austria, 859219). The authors would like to thank Roger Dalmau Diaz (University of Natural Resources and Life Sciences, Vienna) for developing the prototype of the Novasign Hybrid Modeling Toolbox and Lina Vranitzky for her support during the fed-batch cultivations and for conducting the ELISA measurements. The authors would also like to thank Moritz von Stosch and Michael Melcher for critical review and input during the preparation of this manuscript.

Conflict of Interest

Gerald Striedner and Mark Dürkop hold shares of Novasign GmbH.

Keywords

machine learning, process control, quality by design

Received: March 16, 2020
Revised: May 11, 2020
Published online:

[1] A. S. Rathore, H. Winkle, *Nat. Biotechnol.* **2009**, *27*, 26.
[2] V. Kumar, A. Bhalla, A. S. Rathore, *Biotechnol. Prog.* **2014**, *30*, 86.
[3] L. Zhang, S. Mao, *Asian J Pharm. Sci.* **2017**, *12*, 1.
[4] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3.
[5] K.-M. Lee, D. F. Gilmore, *Appl. Biochem. Biotechnol.* **2006**, *135*, 101.
[6] S. J. Kalil, F. Maugeri, M. I. Rodrigues, *Process Biochem.* **2000**, *35*, 539.
[7] M. S. Tanyildizi, D. Özer, M. Elibol, *Process Biochem.* **2005**, *40*, 2291.
[8] F. Torkashvand B. Vaziri, S. Maleknia, A. Heydari, M. Vossoughi, F. Davami, F. Mahboudi, *PLoS One* **2015**, *10*, e0140597.
[9] G. Q. Liu, X.-L. Wang, *Appl. Microbiol. Biotechnol.* **2007**, *74*, 78.
[10] C. F. Mandenius, R. Gustavsson, *J. Chem. Technol. Biotechnol.* **2015**, *90*, 215.
[11] J. Claßen, F. Aupert, K. F. Reardon, D. Solle, T. Scheper, *Anal. Bioanal. Chem.* **2017**, *409*, 651.
[12] S. Craven, J. Whelan, B. Glennon, *J. Process Control* **2014**, *24*, 344.
[13] R. A. Harshman, M. E. Lundy, *Comput. Stat. Data Anal.* **1994**, *18*, 39.
[14] J. Shlens, *Int. J. Remote Sens.* **2014**, *51*, 1.
[15] P. Kadlec, B. Gabrys, S. Strandt, *Comput. Chem. Eng.* **2009**, *33*, 795.
[16] L. Mears, S. M. Stocks, M. O. Albaek, G. Sin, K. V. Gernaey, *Trends Biotechnol.* **2017**, *35*, 914.
[17] M. von Stosch, S. Davy, K. Francois, V. Galvanauskas, J.-M. Hamelink, A. Luebbert, M. Mayer, R. Oliveira, R. O'Kennedy, P. Rice, J. Glassey, *Biotechnol. J.* **2014**, *9*, 719.
[18] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Oxford, UK **1996**.
[19] B. Bayer, B. Sissolak, M. Duerkop, M. von Stosch, G. Striedner, *Bioprocess Biosyst. Eng.* **2020**, *43*, 169.
[20] B. Bayer, M. Von Stosch, G. Striedner, M. Duerkop, *Biotechnol. J.* **2020**, *15*, 1900551.
[21] S. Varma, R. Simon, *BMC Bioinformatics* **2006**, *7*, 91.
[22] J. Pinto, C. R. de Azevedo, R. Oliveira, M. von Stosch, *Bioprocess Biosyst. Eng.* **2019**, *42*, 1853.
[23] M. von Stosch, J. M. Hamelink, R. Oliveira, *Biotechnol. Prog.* **2016**, *32*, 1343.
[24] T. Patarinska, D. Dochain, S. N. Agathos, L. Ganovski, *Bioprocess Eng.* **2000**, *22*, 517.
[25] M. von Stosch, M. J. Willis, *Eng. Life Sci.* **2017**, *17*, 1173.
[26] B. Bayer, M. von Stosch, M. Melcher, M. Duerkop, G. Striedner, *Eng. Life Sci.* **2020**, *20*, 26.
[27] M. Cserjan-Puschmann, W. Kramer, E. Duerrschmid, G. Striedner, K. Bayer, *Appl. Microbiol. Biotechnol.* **1999**, *53*, 43.
[28] T. Porstmann, R. Wietschke, H. Schmechta, R. Grunow, B. Porstmann, R. Bleiber, M. Pergande, S. Stachat, R. von Baehr, *Clin. Chim. Acta* **1988**, *171*, 1.
[29] K. Marisch, K. Bayer, M. Cserjan-Puschmann, M. Luchner, G. Striedner, *Microb. Cell Fact.* **2013**, *12*, 58.
[30] M. Luchner, G. Striedner, M. Cserjan-Puschmann, F. Strobl, K. Bayer, *J. Chem. Technol. Biotechnol.* **2015**, *90*, 283.
[31] M. Melcher, T. Scharl, B. Spangl, M. Luchner, M. Cserjan, K. Bayer, F. Leisch, G. Striedner, *Biotechnol. J.* **2015**, *10*, 1770.
[32] P. Rugbjerg, N. Myling-Petersen, A. Porse, K. Sarup-Lytzen, M. O. A. Sommer, *Nat. Commun.* **2018**, *9*, 787.
[33] R. Bro, *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149.
[34] C. A. Andersson, R. Bro, *Chemom. Intell. Lab. Syst.* **2000**, *52*, 1.
[35] K. R. Murphy, C. A. Stedmon, D. Graeber, R. Bro, *Anal. Methods* **2013**, *5*, 6557.
[36] D. A. Freedman, *Ann. Stat.* **1981**, *9*, 1218.
[37] A. Schuller, M. Cserjan-Puschmann, C. Tauer, J. Jarmer, M. Wagenknecht, D. Reinisch, R. Grabherr, G. Striedner, *Microb. Cell Fact.* **2020**, *19*, 58.

Publication IV

Supporting Information

Supporting Information

CPP Combination Settings for the Static Cultivations

The CPPs, investigated in this DoE study, are introduced in the Materials & Methods section of the main manuscript (2.1). The static CPP combination settings for the static fed-batch fermentations (DoE) are presented in Table S1.

Table S1. Static CPP combination settings of the characterized design space. Each parameter of the design space, namely, the specific growth rate in h^{-1}, the induction strength in µmol IPTG/g cell dry mass (CDM), and the cultivation temperature in °C, was investigated at three levels. The settings for the static fed-batch fermentations (DoE) and the number of repetitions are listed

CPP setting	specific growth rate [h^{-1}]	temperature [°C]	induction strength [µmol IPTG/g CDM]
DoE #1	0.10	30	0.2
DoE #2	0.10	34	0.2
DoE #3	0.10	37	0.2
DoE #4 (n = 2)	0.10	30	0.5
DoE #5	0.10	34	0.5
DoE #6	0.10	37	0.5
DoE #7	0.10	30	0.9
DoE #8	0.10	34	0.9
DoE #9 (n = 3)	0.10	37	0.9
DoE #10	0.15	30	0.2
DoE #11	0.15	34	0.2
DoE #12	0.15	37	0.2
DoE #13	0.15	30	0.5
DoE #14	0.15	34	0.5
DoE #15	0.15	37	0.5
DoE #16	0.15	30	0.9
DoE #17	0.15	34	0.9
DoE #18	0.15	37	0.9
DoE #19	0.20	30	0.2
DoE #20	0.20	34	0.2
DoE #21	0.20	37	0.2
DoE #22 (n = 2)	0.20	30	0.5
DoE #23	0.20	34	0.5
DoE #24	0.20	37	0.5
DoE #25	0.20	30	0.9
DoE #26	0.20	34	0.9
DoE #27	0.20	37	0.9

CPP Combination Settings for the Intensified Cultivations

The intra-experimental CPP shifts are introduced in the Materials & Methods section of the main manuscript (2.2). The iDoE CPP combination settings, as well as the CPP shifts, for the intensified fed-batch fermentations are presented in Table S2.

Table S2. iDoE CPP combination settings of the characterized design space. Each parameter of the design space, namely, the specific growth rate in h^{-1}, the induction strength in µmol IPTG/g CDM, and the cultivation temperature in °C, was investigated at three levels. The settings for the intensified fed-batch fermentations are listed and the CPP shifts are listed

CPP setting	specific growth rate [h^{-1}]	temperature [°C]	induction strength [µmol IPTG/g CDM]	shift #1	shift #2
iDoE #1	0.10	37	0.2	37°C → 34 °C 0.10 h^{-1} → 0.20 h^{-1}	0.20 h^{-1} → 0.10 h^{-1}
iDoE #2	0.10	30	0.5	30°C → 34 °C	34°C → 37 °C 0.10 h^{-1} → 0.20 h^{-1}
iDoE #3	0.10	34	0.9	34°C → 37 °C	0.10 h^{-1} → 0.15 h^{-1}
iDoE #4	0.15	37	0.2	37°C → 30 °C 0.15 h^{-1} → 0.10 h^{-1}	30°C → 34 °C 0.10 h^{-1} → 0.15 h^{-1}
iDoE #5	0.15	30	0.5	0.15 h^{-1} → 0.20 h^{-1}	30°C → 34 °C
iDoE #6	0.15	34	0.5	34°C → 37 °C	0.15 h^{-1} → 0.10 h^{-1}
iDoE #7	0.20	30	0.2	30°C → 37 °C	37°C → 30 °C 0.20 h^{-1} → 0.15 h^{-1}
iDoE #8	0.20	37	0.9	37°C → 34 °C 0.20 h^{-1} → 0.15 h^{-1}	34°C → 30 °C 0.15 h^{-1} → 0.20 h^{-1}
iDoE #9	0.20	34	0.9	34°C → 30 °C 0.20 h^{-1} → 0.15 h^{-1}	0.15 h^{-1} → 0.10 h^{-1}

Analytical Results of the Intensified Design of Experiments

The analytical results of the intensified fed-batch fermentations performed in the three two-dimensional induction planes are displayed as a function of the feed time in Fig. S1. The biomass (displayed as concentration and total) and the soluble product titer are shown for all nine iDoE settings, which are listed in Table S2. The utilized induction plane is indicated with the same color code as in Fig. 1 in the main manuscript, i.e., orange (0.2), green (0.5) and blue (0.9). The varying feed durations occur due to the shifts of the specific growth rate.

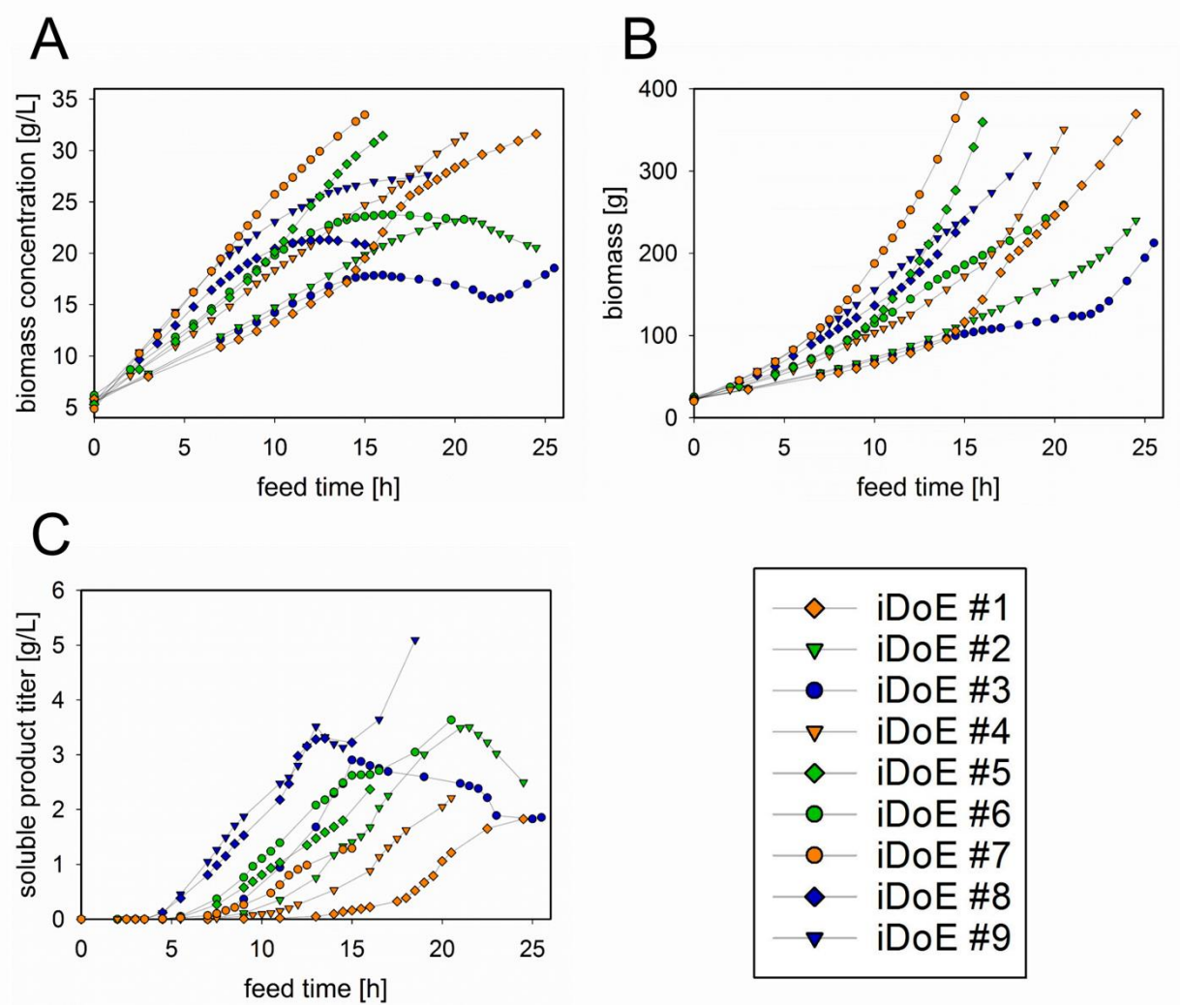

Figure S1. Analytical results for the biomass and soluble product titer of the intensified fed-batch fermentations. The variables are displayed as a function of the feed time for the biomass concentration (A), the total biomass (B) and the soluble product titer (C). To indicate the operated induction plane of these fermentations, the same color code is used for the data symbols as for the respective induction strength in Fig. 1, i.e., orange 0.2, green 0.5, and blue 0.9.

The intensified fed-batch fermentations are discussed in the Materials & Methods section (2.2 & 2.3) of the main manuscript and a general overview of the analytical results is provided in Fig. S1. A more detailed example of how the intra-experimental CPP changes affect the trends of the biomass concentration and the soluble product titer, as well as how fast both process variables adapt to the new CPP settings, is presented in Fig. S2. The exemplary intensified fed-batch fermentation was performed using the CPP settings of iDoE #1, as listed in Table S2. With the initial settings of $\mu = 0.10\ h^{-1}$, 37 °C and the set induction of 0.2 µmol IPTG/g CDM. The first CPP shift was performed after 14 h of feed addition, to 34 °C and $\mu = 0.20\ h^{-1}$. The second CPP shift was performed after 17.5 h of feed addition, to $\mu = 0.10\ h^{-1}$.

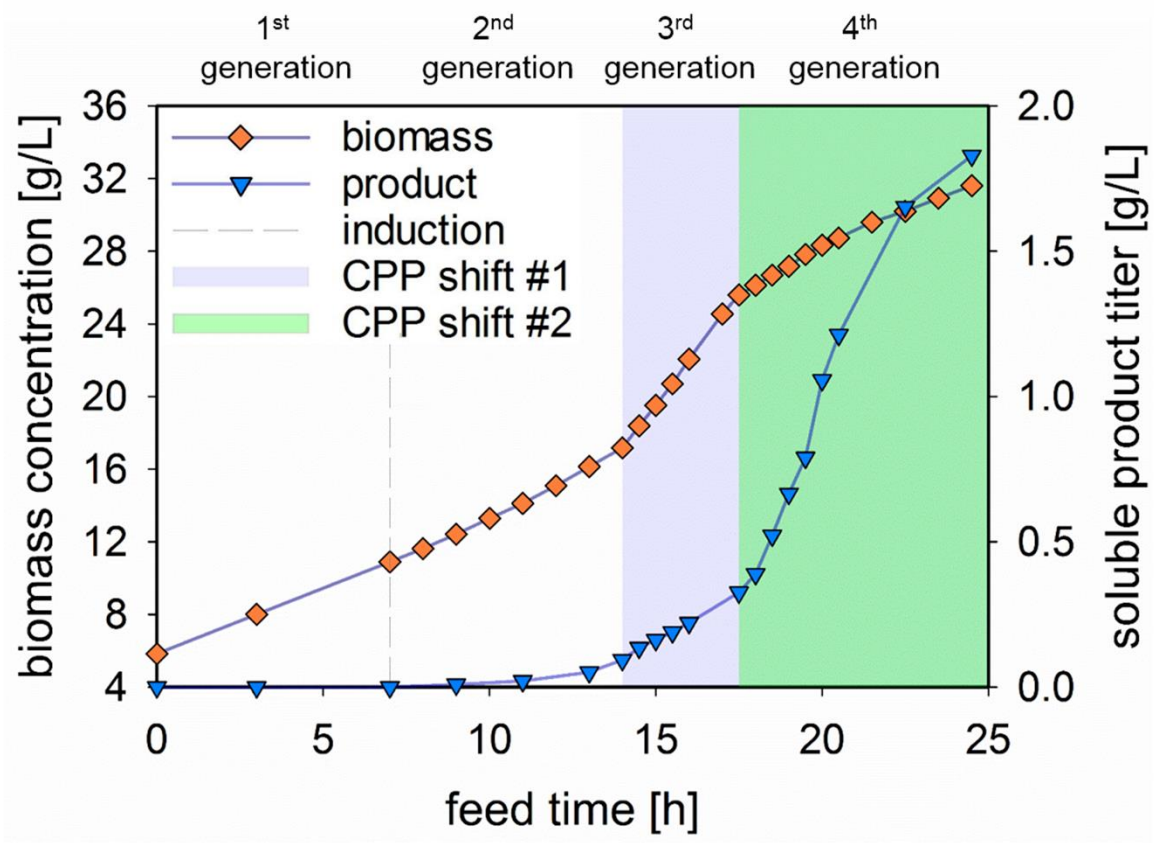

Figure S2. Illustrative example of an intensified fed-batch fermentation (iDoE #1). The biomass (orange squares) and soluble product titer (blue triangle) are presented as a function of the feeding time. The dashed grey line indicates the time point of induction after one doubling time. The remaining doubling times, at which CPP shift #1 (3rd generation) and CPP shift #2 (4th generation) took place, are highlighted in light blue and light green, respectively.

Investigation of a memory effect due to the direction of CPP shifts

To verify that no memory effect from one CPP shift to another takes place and to exclude a potential dependency on the shift direction, an additional intensified fed-batch fermentation was executed, i.e., iDoE #3 in the opposite direction. This exemplary fed-batch fermentation was chosen since the herein executed CPP combination settings (highest induction strength and high cultivation temperatures) displayed the highest impact on the biomass concentration and product titer in the static fed-batch fermentations. Thus, if the CPP shifts or the starting point cause a memory effect, it will most likely be observed in this comparison.

The starting point and CPP shift direction of iDoE #3 are listed in Table S2, with the initial settings of μ = 0.10 h^{-1}, 34 °C and the set induction of 0.9 µmol IPTG/g CDM. The first CPP shift was performed to 37 °C and the second CPP shift to μ = 0.15 h^{-1}. Therefore, the reversed iDoE #3 started with the CPP combination settings from the process end of iDoE #3, i.e., 37 °C and μ = 0.15 h^{-1}. Accordingly, the first CPP shift was performed to μ = 0.10 h^{-1} and the second CPP shift to 34 °C,

The biomass and the product titer of both intensified fed-batch fermentations are displayed as concentration and in total for this comparison in Fig. S3. Herein, to enable this comparison, the entire expressed product (soluble and insoluble fraction) is displayed and the feed time is normalized. Both these steps were done to avoid any distortion due to the different cultivation durations by the applied intended specific growth rate.

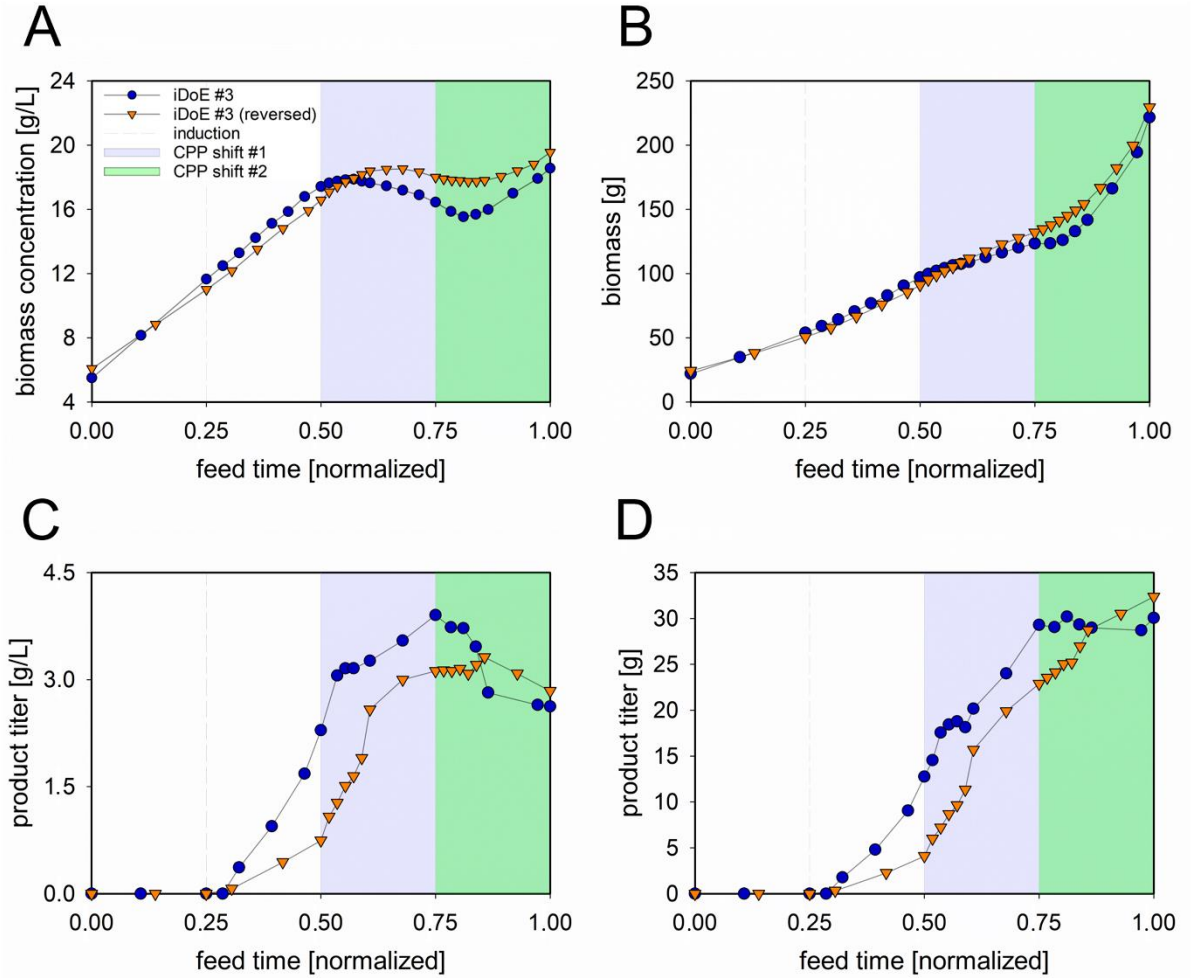

Figure S3. Comparison of two opposingly performed intensified fed-batch fermentations to investigate a potential memory effect. The analytical results for the biomass (concentration (A) and in total (B)) and the product (concentration (C) and in total (D)) for iDoE #3 (blue squares) and the reversely performed iDoE #3 (reversed) (orange triangles) are displayed. The dashed grey line indicates the time point of induction after one doubling time. The remaining doubling times, at which CPP shift #1 (3rd generation) and CPP shift #2 (4th generation) took place, are highlighted in light blue and light green, respectively.

The investigation of a potential impact of the starting point and direction of the CPP shifts in Fig. S3 revealed no visible difference between both opposingly performed intensified fed-batch fermentations. Even though the counterpart of iDoE #3 was performed in the opposite direction, and therefore has undergone the same CPP combination settings in a reversed order, both fed-batch fermentations displayed a highly similar outcome for the biomass and the product. With respect to the analytical results, iDoE #3 displayed a biomass concentration of 18.56 g/L at the end of the process, while its reversed counterpart reached 19.56 g/L (Fig. S3A). This equals a total

biomass of 222g (iDoE #3) and 230g (iDoE #3 reversed) (Fig. S3B). The same is observed for the product titer with 2.62 g/L (iDoE #3) and 2.84 g/L (iDoE #3 reversed) in Fig. S3C and the total product with 30.07 g (iDoE #3) and 32.39 g (iDoE #3 reversed) in Fig. S3D. These slight differences between the end values of both processes are in the range of the analytical uncertainty and process deviation.

Moreover, it is shown that both fed-batch fermentations adapt to the new CPP combination settings after a shift, regardless of the previous settings. These findings further support the assumption that despite these harsh process conditions, no memory effect took place and that in our chosen setup, the direction of the shifts does not have an impact on the outcome of the process.

Training Data utilized for the different Hybrid Models

The three developed hybrid models, i.e., the full-factorial static hybrid model, the fractional-factorial static hybrid model and the iDoE hybrid model, were introduced in the Materials & Methods section (2.6.1) of the main manuscript. The varying numbers of experiments, used as inputs for model building, all derived from the investigated three-dimensional design space, are presented in Fig. S4. In the intensified approach, the induction strengths were separated into three induction planes, wherein the intra-experimental shifts were performed.

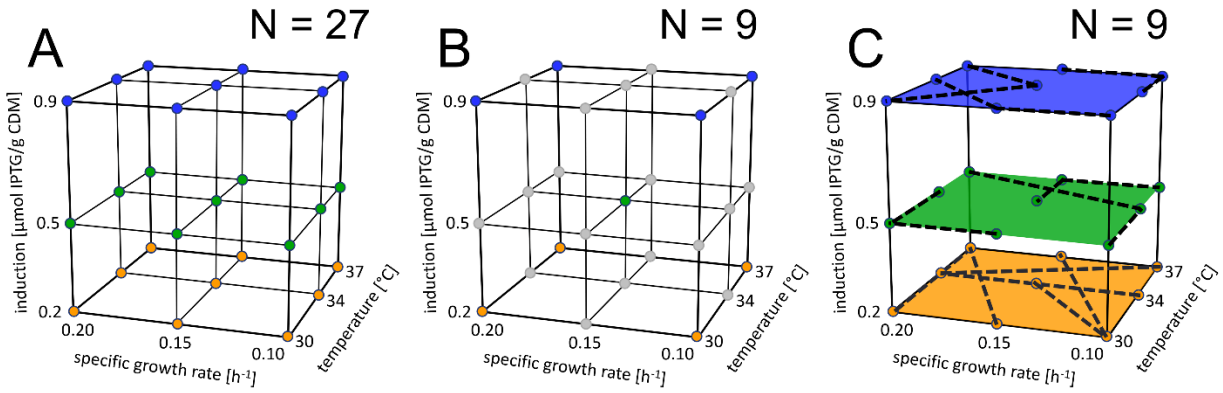

Figure S4. Graphical representation of the utilized data for training the hybrid models. The full-factorial static hybrid model was developed using the complete three-dimensional design space (A), while for the fractional-factorial static hybrid model only the center point and the corners were utilized (B). The iDoE hybrid model was developed using all intensified cultivations, performed in the three two-dimensional induction planes (C). The levels of the three CPPs, namely the via the feed adjusted specific growth rate, the induction strength and the temperature are indicated. The induction strengths and the induction planes are highlighted in different colors (orange 0.2, green 0.5 and blue 0.9). In the case of the fractional-factorial static hybrid model, CPP combination settings not used for model training are grayed out.

For the development of the previously derived full-factorial static hybrid model, the maximum number of cultivations, i.e., all 27 CPP combination settings of the design space, were assessed (Fig. S4A). To bridge between the static and intensified approach and to elaborate the comparison of these, also the fractional-factorial static hybrid model was developed (Fig. S4B). Herein, the number of cultivations for model training was reduced to nine, i.e., only the center point and the corners were used and the test set was expanded to the complete static data set, i.e., the same setup as for the iDoE approach was assessed. The iDoE hybrid model was developed using the

nine intensified cultivations, i.e., three per induction plane, allowing to characterize the entire design space (Fig. S4C). With these three developed models, fair comparison between the approaches was possible by evaluating the performance of a full-factorial static hybrid model, a fractional-factorial static hybrid model and the iDoE hybrid model.

Exploratory Data Analysis of the Static and Intensified Cultivations

The online available process data and the 2D fluorescence data were presented in the Materials & Methods section (2.1) of the main manuscript. The exploratory data analysis of the DoE and iDoE cultivations was described in the Materials & Methods section (2.5) of the main manuscript. PCA and PARAFAC were performed to detect latent structures, accountable for the variance in the data and to investigate a possible memory effect by applying iDoE.

To investigate how these structures contributed to explaining the variance in the online process data, namely, the cultivation temperature, the accumulated feed, the accumulated inductor, and the base consumption, PCA was performed. Underlying differences between both cultivation approaches, indicating metabolic differences or a potential memory effect, can be uncovered if they exhibit different patterns. By applying PARAFAC on the 2D fluorescence data, the spectra were decomposed into three modes, each containing specific information about underlying factors. The first mode represents the compounds' time course, the second mode represents the exact excitation wavelength, and the third mode represents the exact emission wavelength of the respective compound. By merging the information of the second and third mode, the location of each factor is identified in the 2D spectrum, which makes PARAFAC a suitable candidate for analyzing 2D fluorescence spectra. The PCA and PARAFAC results of the two fermentation approaches are compared in Fig. S5.

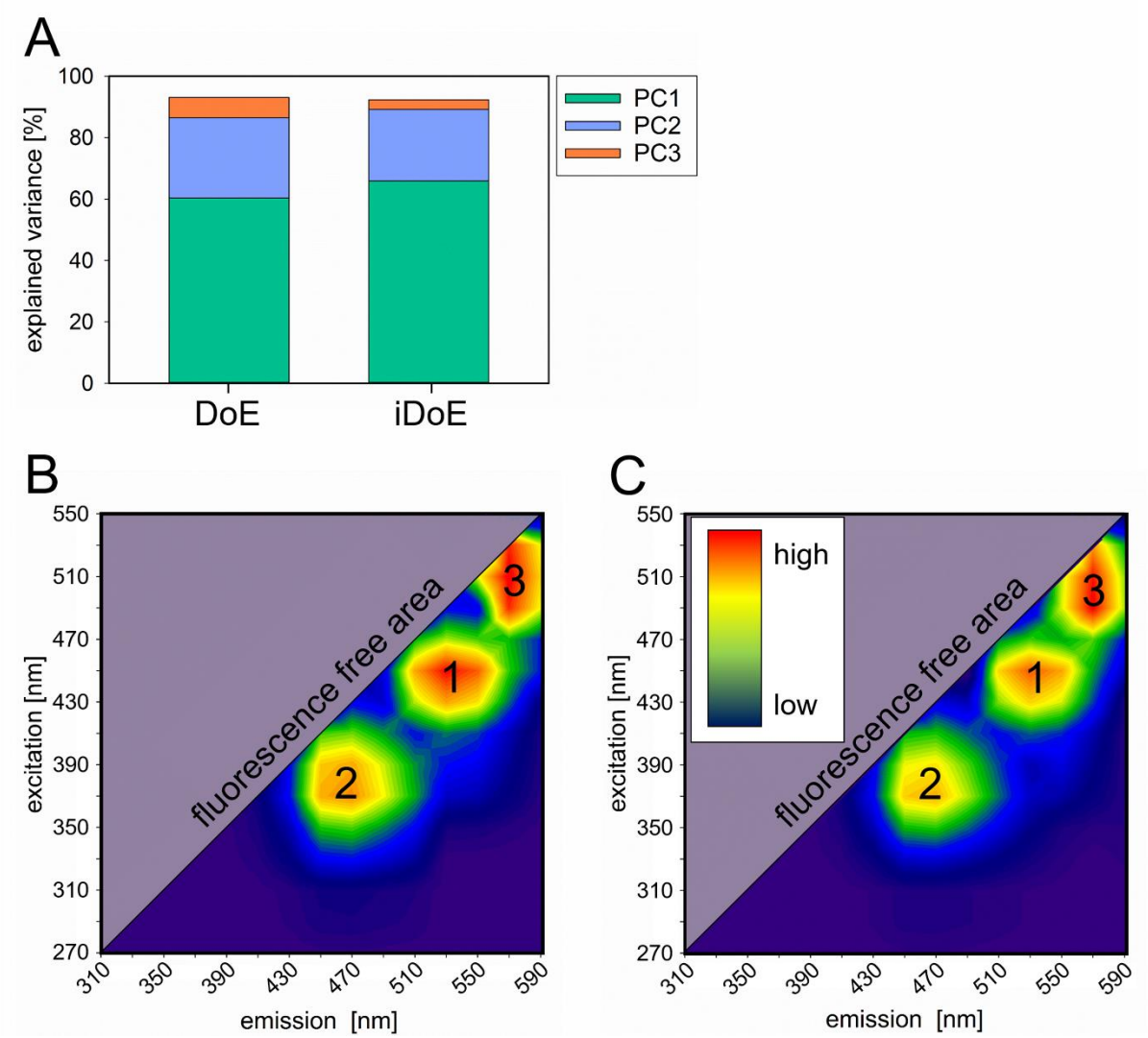

Figure S5. Comparison of the online process data of the static and intensified fed-batch fermentations, using PCA and PARAFAC. The explained variance by the principal components for the complete online data (standard process variables and the fluorescence spectra) of the static DoE (respectively, 60.3 % (PC1), 26.1 % (PC2) and 6.7 % (PC3)) and iDoE (respectively, 65.9 % (PC1), 23.3 % (PC2) and 3.1 % (PC3)) is shown (A). PARAFAC of the 2D fluorescence data located the three most relevant underlying fluorescent compounds, in the 2D fluorescence spectra. For the 2D fluorescence spectrum of the static fed-batch fermentations (B) the wavelength combinations ex450/em530 (1), ex370/em470 (2) and ex510/em570 (3) were identified. For the intensified fed-batch fermentations (C) the same excitation/emission combinations were identified, except for the third compound, represented by ex490/em570 instead. The fluorescence free area and the fluorescence intensity (color bar ranging from blue (low) to red (high)) are indicated.

The PCA results of the complete online process data demonstrate that for both approaches, two PCs explain already approximately 90 % of the variance in the online data (Fig. S5A), displaying a similar distribution. Besides the three CPPs, also the base consumption was considered for the PCA and the comparison of the DoE and iDoE data because it is highly correlated to the biomass formation. This makes it an ideal and well-suited indicator for deviations in the metabolism due to the CPP shifts, e.g., if the cells would stop consuming the base after a shift, which is not the case in the static cultivations. This different consumption trend would be visible in the PCA. However, the PCA did not reveal any significant difference between both approaches, indicating a similar process behavior and no metabolic aberrations.

The measurements obtained from the 2D fluorescence probe are not specific for any analyte, i.e., ex/em wavelength pairs and the respective fluorescence intensity. To be able to draw conclusions from the obtained 2D fluorescence spectra, exploratory data analysis was performed to set these into causal relationships. With respect to the added value of utilizing such a probe during cultivations, its measurements are non-invasive, non-destructive, available without any time delay and take the entire cell broth (cells and supernatant) into account, also delivering information about the cell status, compared to, e.g., HPLC measurements. The comparison of the 2D fluorescence spectra of the static and intensified fed-batch fermentation using PARAFAC was considered and granted a deeper insight, regarding process understanding on a basic metabolic level. 2D fluorescence enables the sensing of metabolic compounds associated with the cell concentration, as it has been discussed elsewhere [1], (Fig. S5B & C). Herein, it has been shown that the first two fluorescent compounds in both approaches are identical. While for a previous paper only the first two compounds were presented, this time the first three were investigated to look deeper into the 2D fluorescence data. The excitation wavelength of the third compound was shifted from 510 nm (in the static) to 490 nm (in the intensified approach). However, it is probably the same compound and the shift is caused by the calibration only. Even though it cannot be excluded that the patterns of nonfluorescent compounds change, the three identified peaks in the 2D fluorescence spectra were identical in both the static and iDoE.

These findings of the exploratory data analysis indicate that during intra-experimental CPP shifts no metabolic switch took place, as also demonstrated in Fig. S3. Further, the cells can rapidly adapt to the new CPP settings (also seen in Fig. S2) and are not permanently altered due to previous CPP settings. Since no memory effect of the cells was observed, the applicability of the iDoE hybrid model approach to predict the outcome of the static fed-batch fermentations is given with high likelihood.

Metabolic Burden of Recombinant Protein Production

The CPP combination settings of the noninduced fed-batch fermentations, introduced in the Materials & Methods section of the main manuscript (2.1), utilized to calculate the production load (PL), are presented in Table S3.

Table S3. CPP combination settings of the noninduced fed-batch fermentations. The parameter settings for the noninduced fed-batch fermentations (Non-Ind), namely, the specific growth rate in h^{-1}, the induction strength in µmol IPTG/g CDM, and the cultivation temperature in °C, are listed

CPP setting	specific growth rate [h^{-1}]	temperature [°C]	induction strength [µmol IPTG/g CDM]
Non-Ind #1	0.10	30	0
Non-Ind #2	0.10	34	0
Non-Ind #3	0.10	37	0
Non-Ind #4	0.15	30	0
Non-Ind #5	0.15	34	0
Non-Ind #6	0.15	37	0
Non-Ind #7	0.20	30	0
Non-Ind #8	0.20	34	0
Non-Ind #9	0.20	37	0

It is known that the induction of the cells to start product formation, decreases the specific growth rate in relation to the respective noninduced CPP combination setting. This metabolic burden, introduced by recombinant protein production during the cultivation, is assumed to enhance the growth of a nonproducing subpopulation in the previous study and therefore was investigated more closely [2], as presented in the Materials & Methods section (2.1) of the main manuscript. If additionally, a small nonproducing subpopulation is already present at the time point of induction, this population without plasmid can grow faster than its counterpart, which is forced to produce the target protein. To make an educated guess about metabolic burden for each CPP combination setting, the additional term of the PL, i.e., the decrease in the specific growth rate of the producing population compared to the respective noninduced cultivation, as presented

elsewhere [3], was introduced (Eq. S1). To calculate the PL, the average specific growth rate of the induced fed-batch cultivation ($\mu_{induced}$) after induction and the respective noninduced cultivation ($\mu_{noninduced}$) was considered. Regarding the calculation of the PL, we are aware that herein $\mu_{induced}$ consists of a mixed population of both, producer and nonproducer cells, making this calculation less precise.

$$PL = 1 - \frac{\mu_{induced}}{\mu_{noninduced}} \quad \text{(S1)}$$

For the static DoE cultivations, this value is specified for each DoE set-point and was assessed for all CPP combination settings. These calculated PL values were fitted to a surface, using a full quadratic function, to individually display them for each induction plane and further provided as the mean of each induction plane in Fig. S6.

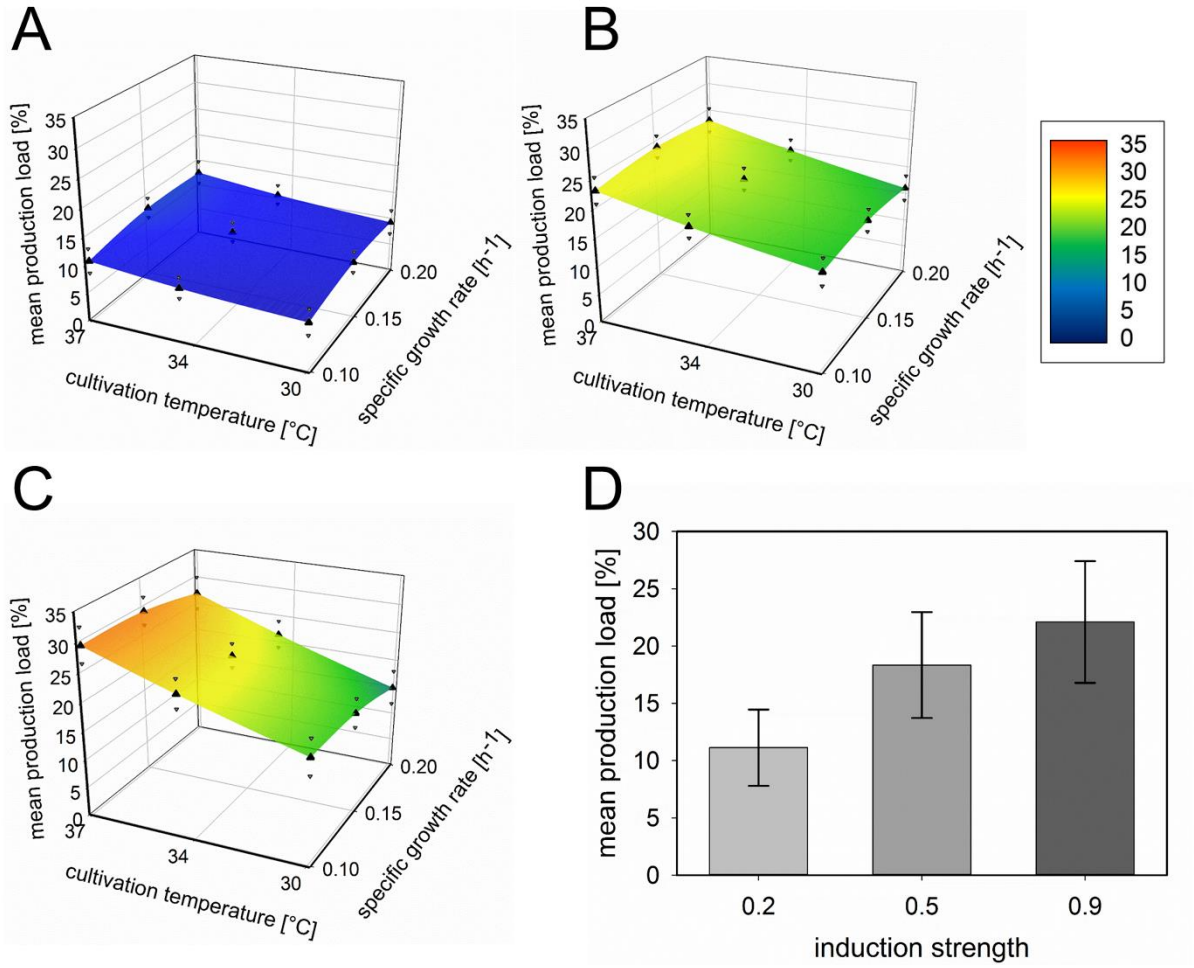

Figure S6. Values of the production load for the entire design space. The full quadratic response surface model of the production load and the SD (triangles) is displayed as a function of the temperature and the specific growth rate. This is done for each induction plane separately, i.e., *I* 0.2 (A), *I* 0.5 (B) and *I* 0.9 (C). The color scale indicates the values from dark blue (lowest value) to red (highest value). The mean values of each induction plane and the respective SD are indicated as bars (D).

The calculated PLs are highly dependent on the induction strength and also follow a consistent trend towards higher cultivation temperatures. This results in a PL range of 8.5-12.7 % for the induction strength 0.2 (Fig. S6A), 16.1-24.2 % for the induction strength 0.5 (Fig. S6B), and 15.0-29.5 % for the induction strength 0.9 (Fig. S6C). The mean PLs after the induction of 11.1 % (±3.3 %, I = 0.2), 18.3 % (±4.6 %, I = 0.5), and 22.1 % (±5.3 %, I = 0.9) were calculated and shown in Fig. S6D.

Although there is currently no analytical method available to precisely distinguish between producers and nonproducers, these values give an indication about the gradient with which the nonproducing cells have an advantage over the producing cells per induction plane, and more specific per CPP combination setting. This formed subpopulation does not have an influence on predicting the biomass, because, in fact, it is also biomass. However, it decreases the performance on predicting the soluble product titer, since Eq. 2 (main manuscript) takes all cells into account for the calculation of the soluble product titer, i.e., it is assumed that more cells are producing the recombinant protein, as it is the case.

Also, the formation of inclusion bodies due to CPP combination settings is an interesting factor that may be present in the PL and potentially interferes with the prediction of the soluble product titer. In the static fed-batch fermentations, we measured insoluble target protein in 6/29 CPP combination settings (up to 50% of the total protein amount), all linked to a cultivation temperature of 37°C at the induction strengths 0.5 and 0.9, i.e., DoE #6, #9, #15, #18, #24 and #27. Likewise, such behavior was observed for the iDoE fed-batch fermentations at similar settings, resulting in smaller values since herein the process conditions were shifted to more favorable CPP combination settings. Here, insoluble protein was measurable in 4/9 iDoE CPP combination settings (up to 29% of the total protein amount), i.e., iDoE #2, #3, #6 and #8. The amount of formed inclusion bodies displays a similar trend towards higher cultivation temperatures and higher induction strengths likewise as the PL. This suggests that an increasing PL promotes this inclusion body formation and additionally complicates the prediction of the soluble product titer.

Therefore, these findings are a first indication of how to approach the issues with respect to predicting the soluble product titer more accurately. Nevertheless, a major issue remains to be solved before the incorporation of this term in the white box is possible, i.e., we are only able to measure the entire mixed population and cannot discriminate between producers and nonproducers and an analytical method to precisely quantify the size of the nonproducing

subpopulation, to confirm this hypothesis, is still unknown. We also tried streaking on plates with and without antibiotics, but the error introduced by heavy dilution to reach countable numbers of clones per plate, is too high to obtain reasonable numbers for the subpopulation.

References

[1] B. Bayer, M. von Stosch, M. Melcher, M. Duerkop, G. Striedner, *Eng. Life Sci.*, **2020**, *20*, 26.

[2] B. Bayer, M. von Stosch, G. Striedner, M. Duerkop, *Biotechnol. J.*, **2020,** *15,* 1900551.

[3] P. Rugbjerg, N. Myling-Petersen, A. Porse, K. Sarup-Lytzen, M. O. A. Sommer, *Nat. Commun.*, **2018**, *9*, 787.