Received: 26 July 2022 | Revised: 7 November 2022 | Accepted: 8 November 2022

DOI: 10.1002/biot.202200381

Biotechnology Journal

RESEARCH ARTICLE

Comparison of mechanistic and hybrid modeling approaches for characterization of a CHO cultivation process: Requirements, pitfalls and solution paths

Benjamin Bayer[1] | **Mark Duerkop**[1] | **Ralf Pörtner**[2] | **Johannes Möller**[2]

[1]Novasign GmbH, Vienna, Austria

[2]Institute of Bioprocess and Biosystems Engineering, Hamburg University of Technology, Hamburg, Germany

Correspondence
Johannes Möller, Institute of Bioprocess and Biosystems Engineering, Hamburg University of Technology, Hamburg,Germany
Email: johannes.moeller@tuhh.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 491268466; Federal Ministry of Education and Research, Grant/Award Number: 031B0577A-C

Abstract

Despite the advantages of mathematical bioprocess modeling, successful model implementation already starts with experimental planning and accordingly can fail at this early stage. For this study, two different modeling approaches (mechanistic and hybrid) based on a four-dimensional antibody-producing CHO fed-batch process are compared. Overall, 33 experiments are performed in the fractional factorial four-dimensional design space and separated into four different complex data partitions subsequently used for model comparison and evaluation. The mechanistic model demonstrates the advantage of prior knowledge (i.e., known equations) to get informative value relatively independently of the utilized data partition. The hybrid approach displayes a higher data dependency but simultaneously yielded a higher accuracy on all data partitions. Furthermore, our results demonstrate that independent of the chosen modeling framework, a smart selection of only four initial experiments can already yield a very good representation of a full design space independent of the chosen modeling structure. Academic and industry researchers are recommended to pay more attention to experimental planning to maximize the process understanding obtained from mathematical modeling.

KEYWORDS
bioprocess characterization, Chinese hamster ovary cells, design of experiments, machine learning, mechanistic modelling, parameter identification, quality by design, upstream bioprocessing

1 | INTRODUCTION

1.1 | Bioprocess characterization

Robust bioprocessing is emphasized by the quality by design (QbD) initiative, promoting process understanding to increase reproducibility and decrease the number of rejected batches.[1,2] One novel method for improving process understanding is mathematical modeling during bioprocess development, characterization, and optimization.[3] Mathematical bioprocess modeling can be applied to develop digital counterparts of the manufacturing process for a multitude of applications, such as bioprocess characterization,[4–6] supporting and further improving bioprocess reproducibility,[7–12] efficiency, productivity, saving experiments,[13–15] and lower the costs to meet rising demands.[16] However, mathematical models depend on and are often

Abbreviations: ANN, artificial neural network; CPP, critical process parameter; CQA, critical quality attribute; DoE, design of experiments; NRMSE, normalized root mean square deviation; PI, prediction interval; QbD, quality by design; SD, standard deviation.

limited by the underlying data, herein the investigated design space including the performed experiments. The design space is created by selecting critical process parameters (CPP) of the bioprocess, for example, cultivation temperature, dissolved oxygen setpoint, media or feed components, and feeding rates.[13,17,18] These CPPs supposedly have a crucial impact on critical quality attributes (CQA) of the biomanufacturing process and the final product,[6] for example, viable cells, product titer, or purity. Thoughtful planning and execution of these experiments, mostly using Design of Experiments (DoE), are of high importance to create and maximize the added value obtained from modeling.

1.2 | Overview of modeling approaches

All different kinds of bioprocess modeling approaches for the description of the timely progression of process state variables can be classified into three different categories.[19] The first category comprises regression models, such as Partial Least Squares, artificial neural networks (ANN) or support vector machines, or Gaussian processes.[20,21] These are only based on experimental data and are therefore often also called data-driven, empiric, black box, or non-parametric models.[22] The benefits of using such a model are that no previous process knowledge must be available, and the implementation is therefore rather fast. However, the requirement of many data, missing extrapolation capabilities, and lack of causality are the major drawbacks.[23]

The second category consists of mechanistic models. Contrary to the first category, these are based on mechanistic process knowledge and equations.[24] Therefore, they are also often called mechanistic, white box, or parametric models. In this contribution, the term "mechanistic model" refers to a system of ordinary differential equations. Since the definition of mathematical models is not standardized, such models are sometimes called "first-principle" or "phenomenological" models.[25] Such models are well suited for extrapolation but due to their purely mechanistic nature and high biological complexity, inaccurate predictions frequently occur in case an important parameter was not considered, or the model is too simple.[23]

The third category combines both modeling approaches in a joint structure, to merge their benefits and cancel out the individual limitations, that is, hybrid modeling, which is often also referred to as gray box or semi-parametric modeling.[26] Due to the combination of empirical process understanding and data, more rational use of the available knowledge and data is enabled to deal with complex issues. The order in which the individual model parts are arranged, how their weights are distributed, and the evaluation that takes place may vary in such a hybrid structure.[27,28] Further, the benefits of such models for accelerated upstream optimization and characterization,[14,15] chromatography,[29,30] or filtration processes[31,32] were already highlighted. For simplification of these technical terms, the individual modeling categories are hereafter called black box, mechanistic and hybrid models.

1.3 | Shedding light on the requirements and pitfalls of bioprocess modeling

In this research work, we aim to provide deeper insights into the requirements for effective bioprocess characterization during early-stage development by applying mathematical process modeling. Additionally, we highlight potential obstacles on the way along with recommendations on how to avoid them already at the beginning of the experimental planning. For this case study, a mechanistic and hybrid model were compared to predict a full design space for an antibody-producing CHO fed-batch process based on four different complex datasets. We focus on illustrating the impact of three common issues regarding bioprocess characterization by mathematical modeling: (i) choosing appropriate CPPs for setting up the design space, (ii) the importance of smartly selecting the best experiments for parameter estimation in mathematical modeling, and (iii) picking a suitable model approach regarding available bioprocess data and knowledge.

2 | EXPERIMENTAL SECTION

2.1 | Cell line and product

The dataset used in this work was generated in different studies with suspension-grown CHO DP-12 cells producing an interleukin-8 antibody.[5,33,34] Please see Möller et al. (2019)[33] for a detailed description of the experimental conditions, which are briefly summarized in the following. The cells were cultivated in chemically defined TC-42 medium (Xell AG, Germany) supplemented with 6 mmol l^{-1} glutamine, 0.1 mg l^{-1} LONG R3 IGF-1, and 200 nmol l^{-1} Methotrexate (all Sigma-Aldrich). As preculture, 125 ml single-use Erlenmeyer baffled flasks (40 ml working volume, Corning, USA) were inoculated with cryo-cultures ($1{\cdot}10^7$ cells ml^{-1}) and the cells were expanded.

Fed-batch experiments were performed in 125 ml single-use Erlenmeyer baffled flasks with a starting working volume of 30 or 40 ml. The feeding design was varied (feed: Chomacs basic feed, Xell AG) with regard to the start times of bolus feeding (48, 72, 96 h), the feeding rate (3–6 ml d^{-1}), and the concentrations of glucose (111–222 mmol l^{-1}) and glutamine (9–38 mmol l^{-1}) in the feed solution. The incubator (Kuhner, Switzerland) was controlled at 37°C, 5% CO_2 and 85% humidity, and a shaking speed of 220 rpm (12.5 mm shaking diameter). Sampling was performed daily from cultivation start to harvest.

2.2 | Analytical methods

Total cell concentration and cell-size distribution were measured with the Z2 particle counter (Z2, Beckmann Coulter, USA). The viability was determined with DNA staining using the DAPI method. Glucose, glutamine, and lactate concentrations were measured with the YSI 2900D (Yellow Springs Instruments, USA) biochemistry analyzer. Ammonium concentration was determined with an enzymatic test

kit (AK00091, Nzytech, Portugal). The antibody was quantified with a high-performance liquid chromatographic system (HPLC, Knauer Smartline, Germany) equipped with a Poros-A column (Thermo Fisher Scientific, USA; 0.1 ml) in accordance with the manufacturer's protocol.

2.3 | Datasets for modeling

The dataset consisted of 33 fed-batch shake flask cultivations with different experimental conditions.[33,40] These experiments were based on a previous study for the knowledge-driven design of feeding strategies using a novel DoE method (model-assisted DoE).[33] Four numeric factors were investigated in the DoE: the glucose concentration in the feed (c_{GlcF}), the glutamine concentration in the feed (c_{GlnF}), the bolus feed rate (F), and the feed start. Overall, four initial experiments and a full DoE with 29 experiments (performed in two blocks, Block 1 and Block 2, respectively) were included in the dataset. In this manuscript, a datapoint is defined as a timely observation (one sampling), including all the analytical results for the seven response variables at this time as well as the CPP setpoints. The considered nomenclature is as follows:

i. "Initial 4": Four cultivations initially used for model parameter determination (model training), see Möller et. al (2019)[33] ($N = 4$, 68 datapoints)
ii. "Optimized 4": Four selected experiments from the full design space intended to cover more process variation compared to the initial experiments ($N = 4$, 79 datapoints)
iii. "Initial 4 + Block 1": Four initial cultivations and Block I of the DoE plan (15 cultivations) ($N = 19$, 384 datapoints)
iv. "Initial 4 + Block 1 + Block 2": Total dataset considering all cultivations ($N = 33$, 671 datapoints)

Please see Table S1–S3, Supporting Information, for a detailed list of the experimental settings and the data repository for the whole data.[40]

3 | MECHANISTIC MODEL

The mechanistic model used in this study was already applied in different studies.[5,33–35] In brief, the unstructured model is based on the description of the main substrates glucose (c_{Glc}) and glutamine (c_{Gln}) as well as the main metabolic waste products lactate (c_{Lac}) and ammonium (c_{Amm}). By this, cell growth behavior is modeled ($X_{t\_}$ total cell density, and $X_{v\_}$ viable cell density) using kinetic parameters $K_{S,i}$ (i = Glc, Gln), a maximal growth rate (μ_{max}), a cell lysis constant (K_{Lys}) of dead cells, and a minimal ($\mu_{d,min}$) and a maximal death rate ($\mu_{d,max}$). The mechanistic is included as Monod-like mathematical terms for the specific growth rate μ (Equation 3), the specific death rate μ_d (Equation 4), and the uptake rates q_{Glc} and q_{Gln} of the substrates glucose and glutamine. The changes in the lactate and ammonium concentrations are proportional to the uptake rates of glucose (lactate) or glutamine (ammonium) (Equations 7–11). The antibody production rate (Equations 5 and 6) is expressed constantly and modeled as proportional to X_v. More details of the model parameter adaptation, identifiability, and sensitivity analysis of this mathematical model can be found in Möller et al. (2020)[5] The mathematical model is presented in the data repository.[40]

$$\frac{dX_t}{dt} = \mu X_v - K_{Lys}(X_t - X_v) - DXv \tag{1}$$

$$\frac{dX_v}{dt} = (\mu - \mu_d)\, X_v - DXv \tag{2}$$

with

$$\mu = \mu_{\max} \frac{c_{Glc}}{c_{\mathrm{Glc}} + K_{S,Glc}} \frac{c_{\mathrm{Gln}}}{c_{\mathrm{Gln}} + K_{S,Gln}} \frac{K_{I,\mathrm{Amm}}}{c_{\mathrm{Amm}} + K_{I,\mathrm{Amm}}} \tag{3}$$

with

$$\mu_d = \mu_{d,\min} + \mu_{d,\max} \frac{K_{S,Glc}}{K_{S,Glc} + c_{Glc}} \frac{K_{S,G\,\mathrm{ln}}}{K_{S,G\,\mathrm{ln}} + c_{G\,\mathrm{ln}}} \frac{c_{Amm}}{K_{S,Amm} + c_{Amm}} \tag{4}$$

$$\frac{dc_{Ab}}{dt} = \gamma X_v - Dc_{Ab} \tag{5}$$

$$\text{constraints: } \frac{dc_{Ab}}{dt} = 0 \text{ if } c_{Amm} \geq K_{Amm} \text{ or } c_{G\,\mathrm{ln}} < 0.05\ \mathrm{mmol\,l^{-1}} \tag{6}$$

$$\frac{dc_{Glc}}{dt} = \left(-q_{Glc,\max} \frac{c_{Glc}}{c_{Glc} + k_{Glc}} \left(\frac{\mu}{\mu + \mu_{\max}} + 0.5\right)\right) X_v + D\left(c_{Glc_F} - c_{Glc}\right) \tag{7}$$

$$\frac{dc_{Lac}}{dt} = \left(Y_{Lac,Glc} \frac{c_{Glc}}{c_{Lac}} q_{Glc} - q_{Lac,uptake}\right) X_v - Dc_{Lac} \tag{8}$$

with

$$c_{Glc} < 0.5\mathrm{mmol\,l^{-1}} : q_{Lac,uptake} = q_{Lac,uptake,\max} \tag{9}$$

$$\frac{dc_{G\,\mathrm{ln}}}{dt} = \left(-q_{G\,\mathrm{ln},\max} \frac{c_{G\,\mathrm{ln}}}{c_{G\,\mathrm{ln}} + k_{G\,\mathrm{ln}}}\right) X_v + D\left(c_{\mathrm{Gln}_F} - c_{G\,\mathrm{ln}}\right) \tag{10}$$

$$\frac{dc_{Amm}}{dt} = \left(Y_{Amm,G\,\mathrm{ln}}\, q_{G\,\mathrm{ln}}\right) X_v - Dc_{Amm} \tag{11}$$

$$D = \frac{F}{V} \tag{12}$$

D = dilution rate (h^{-1}) (Equation 12), describing the ratio between the flow of all volume additions into the shake flasks (i.e., feed), and the overall volume V comprising the initial volume plus the added volume. Therefore, the addition of the main substrates (glucose and glutamine) due to feeding was included to distinguish between the current concentration in the reactor (e.g., for Glutamine as c_{Gln}) and the concentration of the feed (e.g., for Glutamine as c_{GlnF}).

3.1 | Hybrid modeling

3.1.1 | Hybrid model development

The four numeric factors (CPPs) were selected as model inputs to estimate the seven response variables: X_t (10^6 cells ml^{-1}), X_v (10^6 cells ml^{-1}), c_{Ab} (P, g l^{-1}), c_{Glc} (g l^{-1}), c_{Lac} (g l^{-1}), c_{Gln} (g l^{-1}) and, c_{Amm} (g l^{-1}), as well as the predictions of the previous time step of each response variable (last calculated value), taking the process history into account. Before model building, the input variables were standardized using the z-score. A serial hybrid model structure was implemented to predict the response variables. First, an artificial neural network (ANN) consisting of an input layer, one hidden layer, and one output layer was set up. For the hybrid model, an ANN was chosen due to the reasons that non-linear trends can be depicted as well as multiple response variables with one single algorithm. The ANN parameters were selected and optimized by utilizing the respective training dataset only (listed in Section 2.3). The neurons of the hidden layer used hyperbolic tangent transfer functions, while the output layer used linear transfer functions. The Levenberg–Marquardt regularization algorithm was applied during hybrid model identification to estimate the respective specific rates as propagated estimations for the mechanistic part, as a function of the model inputs (Equation 13), that is, the four CPPs and the predictions of the previous time step of each response variable (t-1). Namely, the growth rate of all cells μ_t, the viable cells μ, the product formation rate $v_{p/x}$, the glucose consumption rate q_{Glc}, lactate formation rate q_{Lac}, glutamine consumption rate $q_{Gln,}$ and ammonia formation rate $q_{Amm.}$

$$\mu_t,\ \mu,\ v_{p/x},\ q_{Glc},\ q_{Lac},\ q_{G\ln},\ q_{Amm} = f\left(c_{Glc_F},\ c_{Gln_F},\ F,\ feed\ start,\ response\ variables_{t-1}\right) \quad (13)$$

These rates were embedded in the subsequent mechanistic part to propagated predictions as a function of the model inputs (Equations 14–20).

$$\frac{dX_t}{dt} = \mu_t X_t - DX_t \quad (14)$$

$$\frac{dX_v}{dt} = \mu X_v - DX_v \quad (15)$$

$$\frac{dc_{Ab}}{dt} = v_{p/x} X_v - Dc_{Ab} \quad (16)$$

$$\frac{dc_{Glc}}{dt} = q_{Glc} X_v + D\left(c_{Glc_F} - c_{Glc}\right) \quad (17)$$

$$\frac{dc_{Lac}}{dt} = q_{Lac} X_v - Dc_{Lac} \quad (18)$$

$$\frac{dc_{G\ln}}{dt} = q_{G\ln} X_v + D\left(c_{Gln_F} - c_{G\ln}\right) \quad (19)$$

$$\frac{dc_{Amm}}{dt} = q_{Amm} X_v - Dc_{Amm} \quad (20)$$

The hybrid model development was accomplished in the Novasign GmbH (Vienna, Austria) Hybrid Modeling Toolbox, which can be downloaded and assessed for free upon registration (https://portal.novasign.at/). Even though the source code is not open to the public, the related documentation is online available and provides detailed insights into the preprocessing steps, model-building workflow, evaluation criteria, and utilized functions. The hybrid models used in this study can be found in a data repository.[40]

3.1.2 | Hybrid model validation

Cross-validation was performed to evaluate model performance and to avoid overfitting (poor performance on new data the model was not trained on). The experiments used for training were split into a training and a validation partition using random data partitioning. The initial model was built on the training partition, and the model parameters were optimized until a minimal error in the validation partition was found. This split ratio varied with respect to the available number of experiments in the used dataset while ensuring sufficient variance in the validation partition, that is, for small datasets ("Initial 4" and "Optimized 4") one experiment was used for validation, while in datasets with more experiments (additional utilization of "Block 1" and "Block 2") four or six experiments of the training data (≈20%) were utilized for validation, respectively. Accordingly, based on the number of utilized experiments for training, this random data partitioning step was repeated multiple times to generate a sound number of individual models, that is, for the small datasets the step was repeated four times, while for larger datasets this step was repeated 20 times. Additionally, the number of neurons in the hidden layer also varied, depending on the utilized dataset for model training. Thereto related the number of ANN parameters varied with respect to the different training datasets, i.e., the final optimized ANNs had 54 (Initial 4 & Optimized 4), 144 (Initial 4 + Block 1), and 198 (Initial 4 + Block 1 + Block 2) connections. To ensure the correctness of the ANN, all necessary information was disclosed as proposed by Walsh et al.[36]

3.1.3 | Hybrid model averaging

To assess the risk of model misprediction (overfitting), the predictions from individual hybrid models of the cross-validation (displaying the smallest errors on the validation partition) were averaged. This model averaging of multiple models represents a robust way to deal with model uncertainties. For this final averaged hybrid model, the standard deviation (SD) (Equation 21) and the prediction interval (PI) (Equation 22) were computed, where $\hat{y}_{average}$ is the estimation of the averaged model, $\hat{y}_{model}$ is the estimation of the respective model, i the index of these models, and n is the number of observations for each time point.

$$SD_{(t)} = \sqrt{\frac{1}{n-1} \cdot \sum \left(\hat{y}_{average_{(t)}} - \hat{y}_{\bmod el(i)_{(t)}}\right)^2} \quad (21)$$

$$PI_{(t)} = \hat{y}_{average} \pm SD_{(t)} \quad (22)$$

To assess the performance on new data as well as the risk of estimation uncertainties, the averaged hybrid model was applied to the independent test dataset, containing all experiments (external validation). With an increasing number of experiments utilized for model training and validation (see data partitions in 2.3, I–IV), the test data was accordingly reduced. For data visualization, the units of the measured and predicted outputs were converted to guarantee a uniform comparison.

3.2 | Statistical comparison of model performance

For both modeling approaches (mechanistic and hybrid model), the model parameters were adapted based on the four datasets (see Section 2.3). The adapted models were statistically compared to reflect the whole design space (i.e., dataset: "Initial 4 + block 1 + block 2"). The adapted models were evaluated in comparison to the experimental data using the coefficient of determination R^2 (Equation 23) and the normalized root mean square deviation (NRMSE, Equation 24),[37] based on the simulation $\hat{y}_{(t)}$, the measured data $y_{(t)}$, the mean of the measured data $\overline{y_{(t)}}$, the total number of observations N and the difference between the maximal and minimal measured data ($y_{i,max} - y_{i,min}$):

$$R^2 = \frac{\sum_{i=1}^{n} \left(y_{(t)} - \hat{y}_{(t)}\right)^2}{\sum_{i=1}^{n} \left(y_{(t)} - \overline{y_{(t)}}\right)^2} \quad (23)$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{n} \left(y(t) - \hat{y}(t)\right)^2}}{\left(y_{i,max} - y_{i,min}\right)} \quad (24)$$

4 | RESULTS

4.1 | Experimental data partitions for bioprocess modeling

The parameters of the mechanistic, as well as the hybrid model, were identified using four different datasets with a varying number of experiments (see Section 2.3). The respective setting of each cultivation and their distribution in the design space are illustrated in Figure 1.

As presented in Figure 1, to display the four-dimensional design space, each start day of feeding is depicted as an individual three-dimensional cube consisting of the remaining three CPPs. The initial four experiments (green diamonds) were distributed in the design space spanned by block 1 (black diamonds) and block 2 (blue diamonds). The initial experiments were designed in the respective study[33] for modeling purposes and as proof-of-concept cultivations for the experimental conditions before the DoE was performed. When these initial experiments were performed, the design space was not known, and the experiments were based on the available process understanding. Therefore, the experimental settings were not specifically designed to cover the design space. The chosen settings for these experiments covered the entire value range of the glutamine concentration in the feed and the start day of the feeding. For this dataset ("Initial 4"), higher feed rate values were not covered and instead lower settings than in the remaining design space were executed. Additionally, the glucose concentration in the feed always remained at the same level. The reason for keeping the glucose concentration constant was the availability of a glucose-free feeding medium in the experimental study.[33]

To evaluate if a different design of the initial four experiments would lead to an increased modeling precision of both mathematical models, we chose four more widely distributed experiments from the full dataset ("Initial 4 + Block 1 +Block 2"). These cultivations cover the entire design space (orange diamonds). Herein, it was ensured that every CPP's minimum and maximum value as well as values in between were covered, providing high data variety. Due to this availability of the minimal and maximal CPP level for each investigated variable, it was presumed that both models can learn from the entire range, which provides a more robust basis for parameter identification even though the same number of experiments was utilized.

Using the four different datasets, the model parameters of the mechanistic and the hybrid model were estimated, and the modeling precision of the design space was compared (see also Section 2.3). The results of this comparison are displayed in the following for X_v, c_{Ab}, and c_{Glc}. The remaining results are given in Figure S1–S4 in the Supporting Information.

4.2 | Model performance predicting the viable cell density

The individual performances of both modeling approaches using the four data partitions to predict X_v are presented as scatter plots in Figure 2 along with statistical information.

For the adaptation of the model parameters of the mechanistic model (Figure 2A,D), the individual scatter plots are visually comparable regarding their shape and distribution around the identity line (dot-dashed). The highest R^2 ($R^2 = 0.87$) was achieved with the dataset "Optimized 4" (Figure 2B) for which also the NRMSE is the lowest (NRMSE = 0.12). The lowest R^2 was determined using dataset "Initial 4 + Block 1 + Block 2" (Figure 2D, $R^2 = 0.73$) and the highest NRMSE was calculated for the "Initial 4" (Figure 2A, NRMSE = 0.23).

Utilizing the initial four experiments to identify the parameters in the hybrid model for predicting X_v (Figure 2E) resulted in decent predictions (NRMSE = 0.12 & R^2 = 0.88) but mostly underestimations of the analytical values. The predictive accuracy improved by changing the training data partition to the optimized four experiments (Figure 2F, NRMSE = 0.09 & R^2 = 0.93). However, the additional usage of block 1 (Figure 2G) and block 2 (Figure 2H) only further increased the performance to a small extent, that is, a NRMSE = 0.09 and $R^2 = 0.92$ was achieved by using all available experimental data.

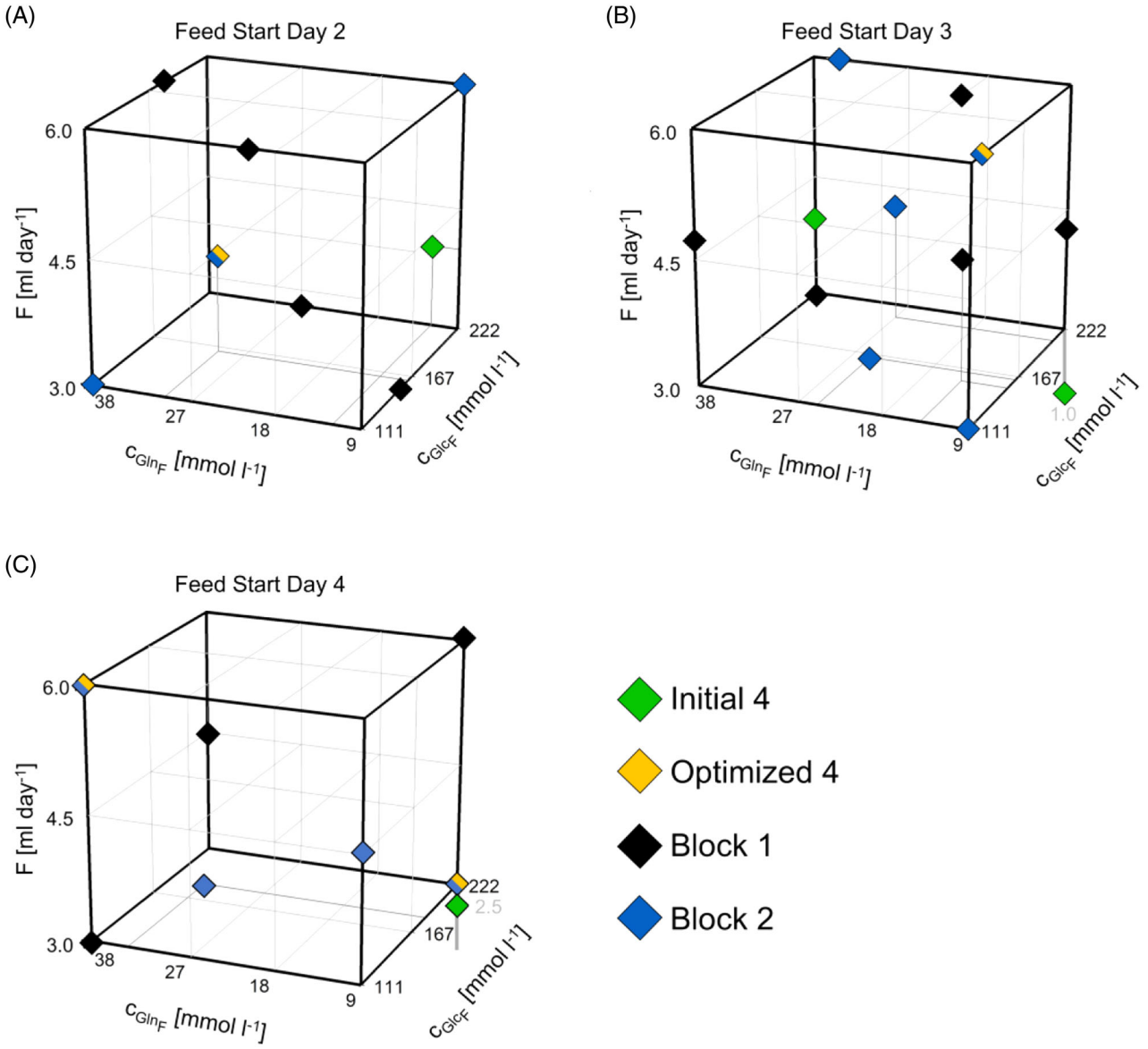

FIGURE 1 Experimental design space and utilized experiments for model training. To enable a three-dimensional representation of the design space, each feed start (A: day 2; B: day 3; C: day 4) is separately presented using the other CPPs: the feed rate, feed glucose, and feed glutamine concentration. The performed experiments (diamond symbols) are indicated according to their respective data partition in the model training: the initial four experiments (green), the optimized four starting experiments (orange), block 1 (black), and block 2 (blue)

4.3 | Model performance predicting the product titer

The individual performances of both modeling approaches using the four data partitions to predict c_{Ab} are presented as scatter plots in Figure 3 along with statistical information.

The R^2 of the adapted mechanistic model (Figure 3A,D) was low (R^2 = –0.54) for the initial data (Figure 3A) and increased for all other datasets up to a maximum of R^2 = 0.71 (Figure 3D) and the NRMSE decreased in the same way from 0.36 (Figure 3A) to 0.16 (Figure 3D). The data are widely distributed in the two-dimensional representation for the dataset 'Initial 4' (Figure 3A) and the data points moved narrower to the identity line when the optimized data was used (Figure 3B).

The hybrid model was not able to accurately predict the product titer by only using the initial four experiments (NRMSE = 0.19 & R^2 = 0.50) as can be seen by the data distribution in the scatter plot (Figure 3E). However, the utilization of the optimized four experiments highly improved the model performance (Figure 3F), already leading to accurate predictions (NRMSE = 0.09 & R^2 = 0.97). While the addition of block 1 to the training data only marginally improved the model performance (Figure 3G, NRMSE = 0.08 & R^2 = 0.95), the utilization of block 2 further decreased the modeling error (Figure 3H, NRMSE = 0.06 & R^2 = 0.97).

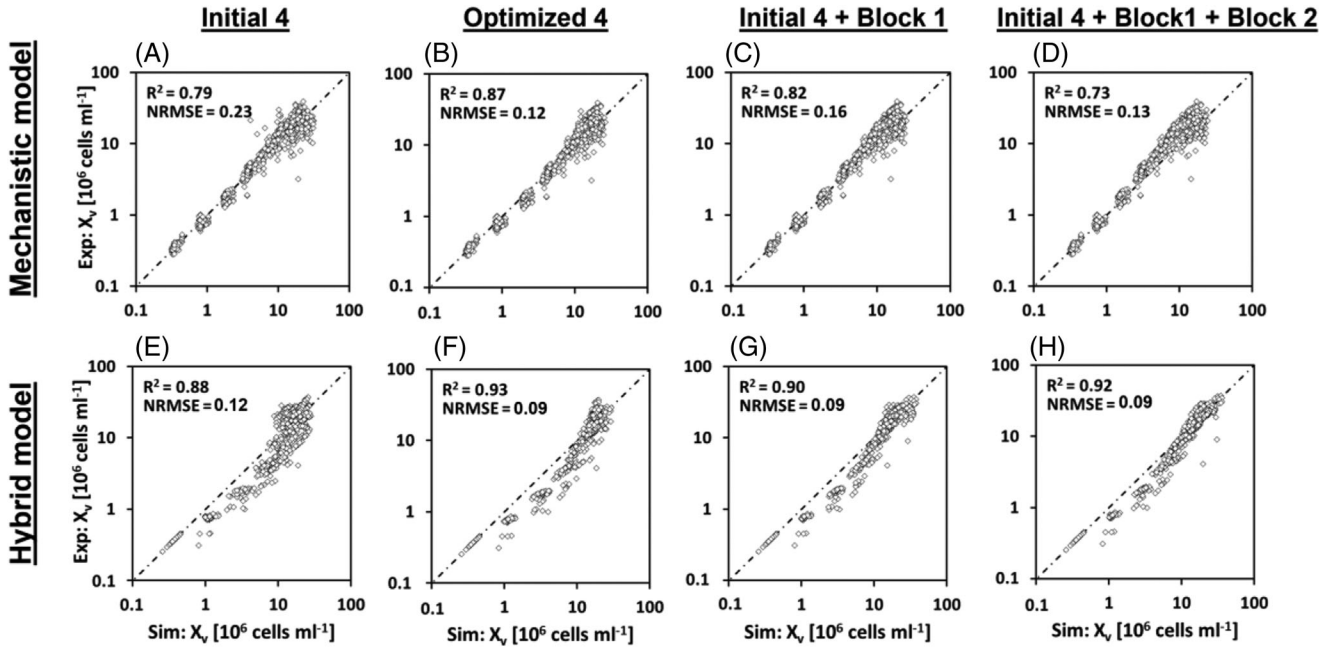

FIGURE 2 Performances of the mechanistic and hybrid model estimating the X_v trained on different numbers of experiments. A–D: scatter plots of the experimental versus estimated values of the mechanistic model using either the initial four experiments (A), the optimized four (B), the initial four experiments and block 1 (C), or the initial four experiments, block 1 and block 2 for model parameter identification. E–H: scatter plots of the experimental versus estimated values of the hybrid model. The utilized experiments for model parameter identification are in the same sequence as for the mechanistic model. The identity line (dot-dashed) is given as a reference. Statistical information (R^2 and NRMSE) is presented for all models

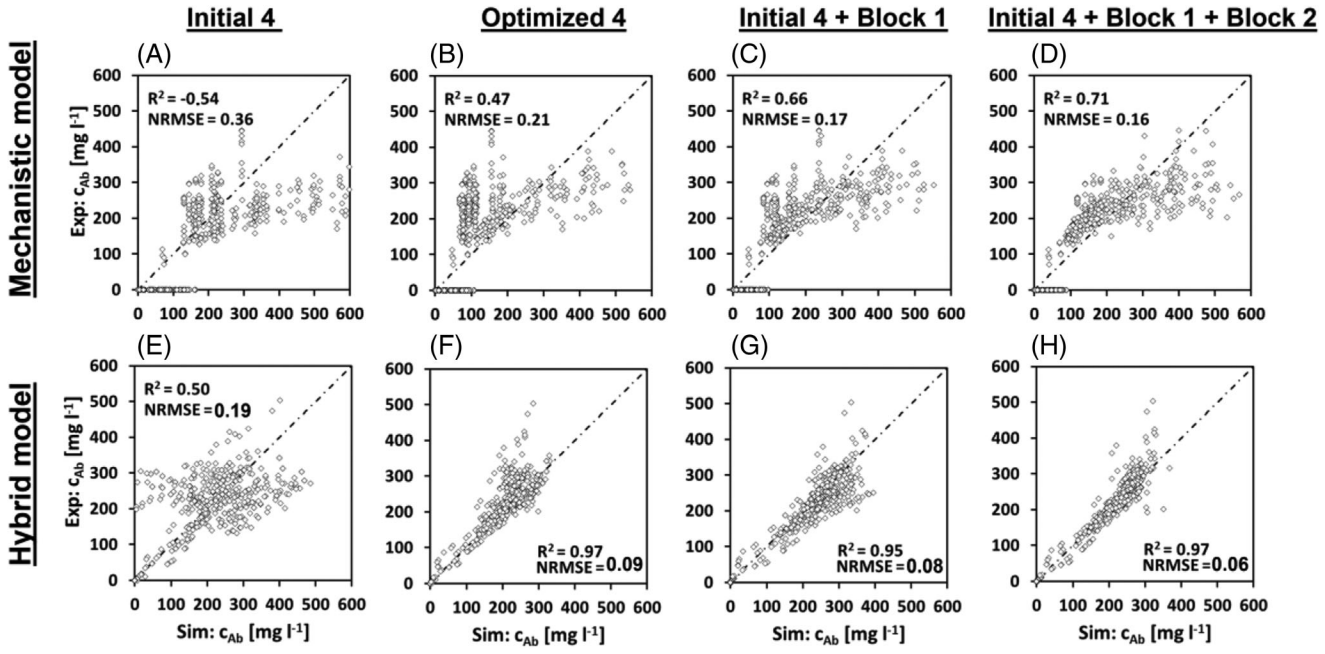

FIGURE 3 Performances of the mechanistic and hybrid model estimating c_{Ab} trained on different numbers of experiments. A–D: scatter plots of the experimental versus estimated values of the mechanistic model using either the initial four experiments (A), the optimized four (B), the initial four experiments and block 1 (C), or the initial four experiments, block 1 and block 2 for model parameter identification. E-H: scatter plots of the experimental versus estimated values of the hybrid model. The utilized experiments for model parameter identification are in the same sequence as for the mechanistic model. The identity line (dot-dashed) is given as a reference. Statistical information (R^2 and NRMSE) is presented for all models

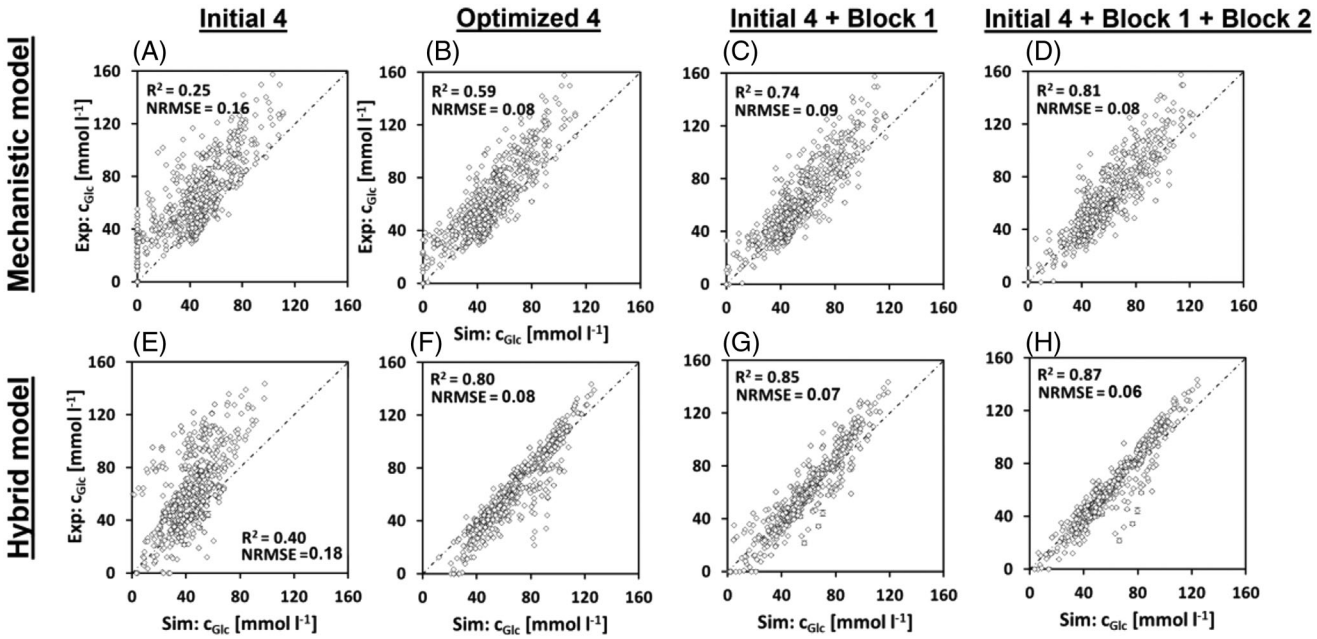

FIGURE 4 Performances of the mechanistic and hybrid model estimating c_{Glc} trained on different numbers of experiments. A–D: scatter plots of the experimental versus estimated values of the mechanistic model using either the initial four experiments (A), the optimized four (B), the initial four experiments and block 1 (C), or the initial four experiments, block 1 and block 2 for model parameter identification. E–H: scatter plots of the experimental versus estimated values of the hybrid model. The utilized experiments for model parameter identification are in the same sequence as for the mechanistic model. The identity line (dot-dashed) is given as a reference. Statistical information (R^2 and NRMSE) is presented for all models

4.4 | Model performance predicting the glucose concentration

The individual performances of both modeling approaches using the four data partitions to predict the main growth substrate c_{Glc} are presented as scatter plots in Figure 4 along with statistical information.

Using just the dataset "Initial 4," c_{Glc} was underestimated with the mechanistic model for the majority of data points (Figure 4A). However, this simulation quality was sufficient for the estimation of the experimental space, as was shown by Möller et. al (2019).[33] R^2 increased to 0.59 and the NRMSE decreased to 0.08 when the optimized dataset was used for the estimation of the model parameters (Figure 4B). A further increase in the number of data did not improve the model predictions further (Figure 4C,D).

Since there was only one glucose feed concentration present in the original four experiments, the hybrid model was not able to accurately predict the experiments distributed in the design space (Figure 4E, NRMSE = 0.18 & R^2 = 0.40) due to missing variation and the possibility to distinguish them. The exchange of the training data partition to the optimized four experiments (Figure 4F) led to a highly improved model performance (NRMSE = 0.08 & R^2 = 0.80). This performance was further increased by the supplementary use of block 1 (Figure 4G, NRMSE = 0.07 & R^2 = 0.85) and block 2 (Figure 4H, NRMSE = 0.06 & R^2 = 0.87), narrowing the data distribution around the identity line.

4.5 | Model performance & improvement by adding experiments

A holistic comparison of both modeling approaches is provided in Figure 5 to visualize the model improvement. Herein, the average NRMSE and R^2 of each utilized data partition are given.

The examination of the different model performances utilizing different training data partitions demonstrates the importance of the proper selection of experiments for both types of models (mechanistic and hybrid model, respectively). Using the 'Initial 4′ experiments for the model parameter identification resulted in inaccurate model predictions, which is apparent when considering the NRMSE (Figure 5A) as well as the R^2 (Figure 5B). By selecting the optimized four experiments for model training, the model performance highly increased with respect to the NRMSE (mechanistic: 0.14 ± 0.04, hybrid: 0.09 ± 0.02) and R^2 (mechanistic: 0.70 ± 0.16, hybrid: 0.90 ± 0.06). This demonstrates the importance of the quality of the data and the subsequent information content, over the pure quantity of data (also see Figure 2–4 and Figure S1-4). The additional use of the experiments from block 1 for model training ("Initial 4 + Block 1") further decreased the NRMSE (mechanistic: 0.13 ± 0.03 hybrid: 0.08 ± 0.02) and increased the R^2 (mechanistic: 0.73 ± 0.08 hybrid: 0.90 ± 0.05) of the model predictions. While the same trend was present for the experiments from block 1 and block 2 ('Initial 4 + Block 1 + Block 2′) with

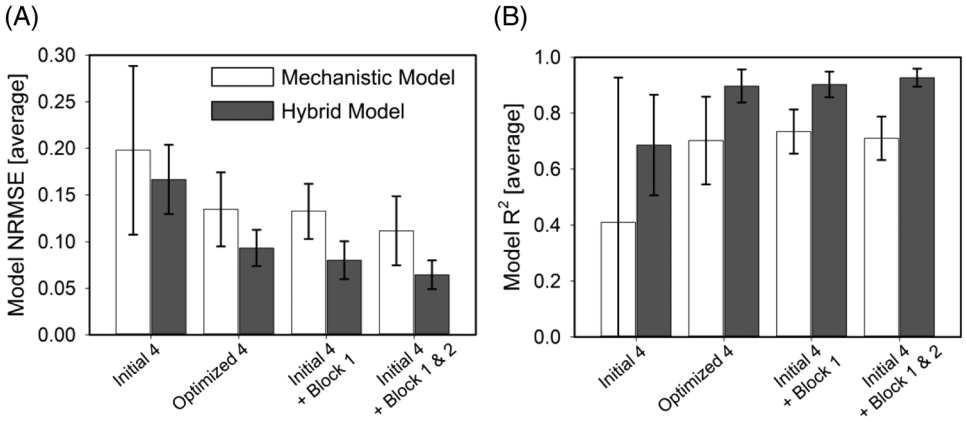

FIGURE 5 Performance comparison of the mechanistic and hybrid model with respect to the average NRMSE and R^2. The average values ± standard deviation of the NRMSE (A) and R^2 (B) of each utilized data partition for model parameter identification are displayed. The statistical values of the glutamine and lactate prediction were excluded from this calculation for the mechanistic model (white bars) since no reasonable predictions were achieved (see Supporting Information Figures), while for the hybrid model (gray bars) all response variables were considered

a declining learning rate which was observed by continuously adding more experiments.

5 | DISCUSSION

In this study, we were focusing on illustrating the impact of three common issues, which hinder process characterization by modeling: (i) choosing appropriate CPPs for setting up the design space, (ii) the importance of smartly selecting the best experiments for parameter estimation in mathematical modeling, and (iii) picking a suitable model approach regarding available bioprocess data and knowledge.

5.1 | Choosing appropriate CPPs for setting up the design space

Setting up a DoE with several CPPs for a mammalian cell culture bioprocess can result in a large experimental space, rapidly exceeding the number of feasible laboratory experiments and posing a particular challenge in data management and handling. However, the investigation of impactful CPPs and their experimental investigation is a necessity for bioprocess characterization and mathematical process modeling. Therefore, choosing appropriate identified CPPs concerning quantity and quality to determine their impact on key process variables is a major challenge. For the purely mechanistic model, a mathematical dependency on a particular CPP must be established upfront and implemented in the model structure. For the hybrid model, a less strict understanding is required but, if the chosen CPPs do not display an impact on process variation, the knowledge gain is limited, the subsequently developed process model will not help to foster process understanding and a lot of experimental resources are wasted. Thus, it is of high importance to thoroughly consider the key process variables by means of risk analysis of potential CPPs to set up the experimentsand the design space, capturing the process behavior and enabling successful process characterization. The more important factors are considered the more experimental effort will be required but the subsequent models can easier be applied during scale-up [18] or as model predictive control applications in the future. Thereby an expected variation (e.g., pump failures) in one variable can potentially be corrected by adaptation of the other CPPs.[10]

5.2 | Importance of smartly selecting the best experiments for parameter estimation in mathematical modeling

In this study, four CPPs on several levels were chosen. If a full factorial DoE design would have been chosen instead of a fractional factorial design, 108 experiments would span the whole design space resulting in an immense laboratory workload. Typically, only a fraction of all possible experiments herein can be performed and utilized for process characterization (Figure 1). Therefore, it is important to wisely choose a limited number of experiments to maximize knowledge gain. The impact of selecting the right experiments to be conducted and subsequently used for mathematical process modeling is explicitly visible when the modeling results are compared. Especially in the hybrid modeling approach, this difference between selecting the initial four experiments or the four optimized experiments for model training can lead to high differences regarding process understanding and predictive performance (e.g., Figures 3 and 4E,F). By adding more and more experiments to the training data, the knowledge gap in the design space and thereby the model error and its uncertainty are reduced (Figures 2–4. C,D, G,H), which diminishes the impact of the small initial dataset. However, this effect is less obvious for the mechanistic modeling approach in which the better incorporation of prior knowledge in the model structure itself yields more similar results independent of the chosen experiments. Especially if only a limited number of experi-

ments can be performed and used for modeling (initially or at all), the experimental settings should be selected carefully, maximizing input differences in each run, to obatain meaningful results and the highest possible added value. Nevertheless, both modeling approaches already demonstrate good performance with the optimized four experiments highlighting the advantage of integrating process knowledge over pure black box modeling again proved to be a valuable choice.[6]

5.3 | Picking a suitable model approach regarding available bioprocess data and knowledge

The findings from (i) and (ii) underline that modeling is not a detached all-in-one solution to deal with bioprocess issues but rather depends on smart experimental planning and thoughtful execution. The amount and quality of the data as well as prior process understanding have a high impact on the modeling results. Based on these pillars, the selection of a well-suited methodology for mathematical process modeling should follow. Based on the outcome of our study, no explicit answer is inherently correct in this respect. In consideration of the complexity of mathematical modeling, available process knowledge and the quality and quantity of experimental data, hybrid or mechanistic models may be sufficient. An advantage of the mechanistic model is that with the initial set of experiments, the mechanistic model can give a fair estimation of the expected modeling outputs (X_v, c_{Ab}, c_{Glc}), but the quality of the prediction does not significantly increase with more data. Since the mechanistic model has a clear rigid structure, only the described mechanisms can be included and the flexibility of the model to represent unknown effects is limited.[38,39] This effect can be seen in the reduced R^2 for the description of X_v with all available data (see. Figure 2D). However, potentially a few representative cultivations could be sufficient to already yield fair predictions of the design space if the bioprocess is understood well. Thereby, the model less depends on the amount and quality of data. However, with an increase in the quantity of available data, or by choosing optimized initial experiments it is obvious that the hybrid model outperforms the mechanistic model in modeling precision. This can be explained by the fact that the rates in the mechanistic equations within the hybrid model are functions of all the CPPs rather than dedicated numerical and defined numbers following predefined mathematical structures. Due to the high complexity of bioprocesses, hybrid models have been proven to be beneficial to deal with known unknowns for the estimation of specific rates. Utilizing a hybrid model structure, these parameters are estimated less strictly instead of taking fixed assumptions. The availability of more experimental data, tuning the data-driven part, further improves the hybrid model performance with respect to model accuracy and certainty (Figure 5). Moreover, besides the in this study utilized ANN as black box part of the hybrid model, other suitable algorithms can be applied if they match the specific modeling requirements for the task (e.g., herein the possibility to model non-linear trends and multiple response variables). However, such a comparison of different black box algorithms and their impact on the model performance was not in the scope of this study.

6 | CONCLUSION AND OUTLOOK

Even if computational tools for mathematical process modeling of upstream bioprocesses are available, the requirements during their implementation in early-stage bioprocess development possess a particular challenge, and obstacles might already be found during the experimental planning. This study provides insights into three common issues that prevent optimal process modeling and how to avoid these problems with the objective of optimal bioprocess characterization.

i. The aspect of choosing appropriate CPPs with sufficient levels, setting up the design space, to investigate the key process variables of interest is the first crucial step.
ii. The selection of meaningful experiments, allowing the model to develop a cause-and-effect linkage even though only a fraction of all possible experiments was performed.
iii. In case both these experimental prerequisites are fulfilled, selecting a suitable model structure and algorithm for bioprocess modeling is required. These should be based on the experimental data and the availability of process knowledge. If mechanistic understanding is given, the usage of a mechanistic model yields good results with an extremely limited set of data. However, since the hybrid model learns more from the data than the mechanistic counterpart, steps (i and ii) are utterly crucial to set up powerful hybrid models with minimum effort.

If the described requirements are taken into mind and are complied, the opportunity to develop meaningful models and successfully apply them for improving bioprocessing and proceeding toward digitalization is within one's reach.

Nomenclature

Symbol	Unit	Definition
μ	[h^{-1}]	specific growth rate of the viable cells
μ_d	[h^{-1}]	death rate
$\mu_{d,max}$	[h^{-1}]	maximal death rate
$\mu_{d,min}$	[h^{-1}]	minimal death rate
μ_t	[h^{-1}]	specific growth rate of all cells
Y	[mg $cell^{-1}$ h^{-1}]	antibody formation rate
c_{Ab}	[mg l^{-1}]	product concentration
c_{Amm}	[mmol l^{-1}]	ammonia concentration
c_{GlcF}	[mmol l^{-1}]	glucose concentration in the feed
c_{Glc}	[mmol l^{-1}]	glucose concentration in the reactor
c_{GlnF}	[mmol l^{-1}]	glutamine concentration in the feed
c_{Gln}	[mmol l^{-1}]	glutamine concentration in the reactor
c_{Lac}	[mmol l^{-1}]	lactate concentration
D	[h^{-1}]	dilution rate
F	[l h^{-1}]	feeding rate

Symbol	Unit	Definition
I	[]	index (Glc, Gln, Amm, Lac, Ab)
k_i	[mmol l^{-1}]	kinetic constant for component i
$K_{i,Amm}$	[mmol l^{-1}]	inhibitory constant of ammonia
K_{Lys}	[-]	cell lysis parameter
$K_{S,i}$	[mmol l^{-1}]	Monod kinetic constant for component i
N	[]	Number of observations
q_{Ab}	[mg $cell^{-1}$ h^{-1}]	cell-specific antibody formation rate
q_{Amm}	[mmol $cell^{-1}$ h^{-1}]	cell-specific ammonia formation rate
q_{Glc}	[mmol $cell^{-1}$ h^{-1}]	cell-specific glucose uptake rate
$q_{Glc,max}$	[mmol $cell^{-1}$ h^{-1}]	maximum uptake rate of glucose
q_{Gln}	[mmol $cell^{-1}$ h^{-1}]	cell-specific glutamine uptake rate
$q_{Gln,max}$	[mmol $cell^{-1}$ h^{-1}]	maximum uptake rate of glutamine
q_{Lac}	[mmol $cell^{-1}$ h^{-1}]	cell-specific lactate uptake rate
$q_{Lac,uptake}$	[mmol $cell^{-1}$ h^{-1}]	uptake rate of lactate
$q_{Lac,uptake,\ max}$	[mmol $cell^{-1}$ h^{-1}]	maximum uptake rate of lactate
V	[l]	reactor volume
$v_{p/x}$	[mg $cell^{-1}$ h^{-1}]	antibody formation rate
X_t	[10^6 cells ml^{-1}]	total cell density
X_v	[10^6 cells ml^{-1}]	viable cell density
$Y_{Lac,Glc}$	[-]	yield coefficient of lactate formation to glucose consumption
$Y_{Amm,Gln}$	[-]	yield coefficient of ammonia formation to glutamine consumption
$\hat{y}_{average_{(t)}}$	[]	estimation of the averaged model
$\hat{y}_{model}$	[]	estimation of the respective model
$\hat{y}_{(t)}$	[]	simulation of a datapoint
$y_{(t)}$	[]	measured data
$\overline{y_{(t)}}$	[]	mean of measured data
$y_{i,max}$	[]	maximum of measured data
$y_{i,min}$	[]	minimum of measured data

AUTHOR CONTRIBUTIONS

Ralf Pörtner: Conceptualization; Formal analysis; Software; Supervision; Writing – review & editing. Johannes Möller: Conceptualization; Formal analysis; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing

ACKNOWLEDGEMENTS

Publishing fees funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 491268466 and the Hamburg University of Technology (TUHH) in the funding programme *Open Access Publishing*. Johannes Möller and Ralf Pörtner acknowledge partial funding by the German Federal Ministry of Education and Research (BMBF, Grant 031B0577A-C).

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

Benjamin Bayer and Mark Duerkop hold shares of Novasign GmbH. Johannes Möller and Ralf Pörtner declare no conflict of interest.

DATA AVAILABILITY STATEMENT

In addition to the supplementary materials, the experimental data, hybrid model structure and mechanistic model is available in a data repository.[40]

ORCID

Benjamin Bayer https://orcid.org/0000-0001-5241-4924
Mark Duerkop https://orcid.org/0000-0003-4750-6474
Ralf Pörtner https://orcid.org/0000-0003-1163-9718
Johannes Möller https://orcid.org/0000-0001-9546-055X

REFERENCES

1. Rathore, A. S. (2014). QbD/PAT for bioprocessing: moving from theory to implementation. *Current Opinion in Chemical Engineering*, *6*, 1.
2. Zhang, L., & Mao, S. (2017). Application of quality by design in the current drug development. *Asian Journal of Pharmaceutical Sciences*, *12*, 1.
3. Mandenius, C.-F., & Brundin, A. (2008). Bioprocess optimization using design-of-experiments methodology. *Biotechnology Progress*, *24*, 1191.
4. Möller, J., Bhat, K., Riecken, K., Pörtner, R., Zeng, A - P., & Jandt, U. (2019). Process-induced cell cycle oscillations in CHO cultures: Online monitoring and model-based investigation. *Biotechnology and Bioengineering*, *116*, 2931.
5. Möller, J., Hernández Rodríguez, T., Müller, J., Arndt, L., Kuchemüller, K. B., Frahm, B., Eibl, R., Eibl, D., & Pörtner, R. (2020). Model uncertainty-based evaluation of process strategies during scale-up of biopharmaceutical processes. *Computers & Chemical Engineering*, *134*, 106693.
6. Bayer, B., von Stosch, M., Striedner, G., & Duerkop, M. (2020). Comparison of modeling methods for DoE-based holistic upstream process characterization. *Biotechnology Journal*, *15*, 1900551.
7. Bayer, B., Maccani, A., Jahn, J., Duerkop, M., Kapeller, E., Pletzenauer, R., Kraus, B., Striedner, G., & Hernandez Bort, J. A. (2022). Proton-transfer-reaction mass spectrometry (PTR-MS) for online monitoring of glucose depletion and cell concentrations in HEK 293 gene therapy processes. *Biotechnology Letter*, *44*, 77.
8. Bayer, B., von Stosch, M., Melcher, M., Duerkop, M., & Striedner, G. (2020). Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in *Escherichia coli* cultivations. *Engineering in Life Sciences*, *20*, 26.
9. Abt, V., Barz, T., Cruz Bournazou, M. N., Herwig, C., Kroll, P., Möller, J., Pörtner, R., & Schenkendorf, R. (2018). Model-based tools for optimal experiments in bioprocess engineering. *Current Opinion in Chemical Engineering*, *22*, 244.
10. Sommeregger, W., Sissolak, B., Kandra, K., von Stosch, M., Mayer, M., & Striedner, G. (2017). Quality by control: Towards model predictive control of mammalian cell culture bioprocesses. *Biotechnology Journal*, *12*, 1600546.
11. Bayer, B., Sissolak, B., Duerkop, M., von Stosch, M., & Striedner, G. (2020). The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling. *Bioprocess Biosystem Engineering*, *43*, 169.
12. Pappenreiter, M., Sissolak, B., Sommeregger, W., & Striedner, G. (2019). Oxygen uptake rate soft-sensing via dynamic kLa computation: Cell Volume and metabolic transition prediction in mammalian bioprocesses. *Frontier Bioengineering Biotechnology*, *7*, 195.

13. Nold, V., Junghans, L., Bisgen, L., Drerup, R., Presser, B., Gorr, I., Schwab, T., Knapp, B., & Wieschalka, S. (2021). *Engineering in Life Sciences*, https://doi.org/10.1002/elsc.202100123
14. Bayer, B., Striedner, G., & Duerkop, M. (2020). Hybrid modeling and intensified DoE: An approach to accelerate upstream process characterization. *Biotechnology Journal*, *15*, 2000121.
15. Bayer, B., Dalmau Diaz, R., Melcher, M., Striedner, G., & Duerkop, M. (2021). Digital twin application for model-based DoE to rapidly identify ideal process conditions for space-time yield optimization. *Processes*, *9*, 1109.
16. Rathore, A. S., & Winkle, H. (2009). Quality by design for biopharmaceuticals. *Nature Biotechnology*, *27*, 26.
17. Moser, A., Kuchemüller, K. B., Deppe, S., Hernández Rodriguez, T., Frahm, B., Pörtner, R., Hass, V. C., & Möller, J. (2021). Model-assisted DoE software: optimization of growth and biocatalysis in *Saccharomyces cerevisiae* bioprocesses. *Bioprocess Biosystem Engineering*, *44*, 683.
18. Bayer, B., Duerkop, M., Striedner, G., & Sissolak, B. (2021). Model transferability and reduced experimental burden in cell culture process development facilitated by hybrid modeling and intensified design of experiments. *Frontier Bioengineering Biotechnology*, *9*, 740215.
19. Pörtner, R. (Ed.) (2021). *Cell Culture Engineering and Technology*. Springer Nature, Switzerland AG.
20. di Sciascio, F., & Amicarelli, A. N. (2008). Biomass estimation in batch biotechnological processes by Bayesian Gaussian process regression. *Computers & Chemical Engineering*, *32*, 3264.
21. Glassey, J., Gernaey, K. V., Clemens, C., Schulz, T. W., Oliveira, R., Striedner, G., & Mandenius, C.-F. (2011). Process analytical technology (PAT) for biopharmaceuticals. *Biotechnology Journal*, *6*, 369.
22. Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven Soft Sensors in the process industry. *Computers & Chemical Engineering*, *33*, 795.
23. Solle, D., Hitzmann, B., Herwig, C., Pereira Remelhe, M., Ulonska, S., Wuerth, L., Prata, A., & Steckenreiter, T. (2017). Between the poles of data-driven and mechanistic modeling for process operation. *Chemie Ingenieur Technik*, *89*, 542.
24. Mears, L., Stocks, S. M., Albaek, M. O., Sin, G., & Gernaey, K. V. (2017). Mechanistic fermentation models for process design, monitoring, and control. *Trends in Biotechnology*, *35*, 914.
25. Kroll, P., Hofer, A., Stelzer, I. V., & Herwig, C. (2017). Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. *Process Biochemistry*, *62*, 24.
26. von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J.- M., Luebbert, A., Mayer, M., Oliveira, R., O'Kennedy, R., Rice, P., & Glassey, J. (2014). Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnology Journal*, *9*, 719.
27. von Stosch, M., Oliveira, R., Peres, J., & Feyo de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, *60*, 86.
28. von Stosch, M., Oliveria, R., Peres, J., & De Azevedo, S. F. (2012). Hybrid modeling framework for process analytical technology: Application to Bordetella pertussis cultures. *Biotechnology progress*, *28*, 284.
29. Narayanan, H., Luna, M. F., Sokolov, M., Arosio, P., Butté, A., & Morbidelli, M. (2021). Hybrid models based on machine learning and an increasing degree of process knowledge: Application to capture chromatographic step. *Industrial & Engineering Chemistry Research*, *60*, 10466.
30. Narayanan, H. H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., & Butté, A. (2021). Hybrid models for the simulation and prediction of chromatographic processes for protein capture. *J. Chromatogr. A.*, *1650*, 462248.
31. Krippl, M., Dürauer, A., & Duerkop, M. (2020). Hybrid modeling of cross-flow filtration: Predicting the flux evolution and duration of ultrafiltration processes. *Separation and Purification Technology*, *248*, e117064.
32. Krippl, M., Kargl, T., Duerkop, M., & Dürauer, A. (2021). Hybrid modeling reduces experimental effort to predict performance of serial and parallel single-pass tangential flow filtration. *Separation and Purification Technology*, *276*, 119277.
33. Möller, J., Kuchemüller, K. B., Steinmetz, T., Koopmann, K. S., & Pörtner, R. (2019). Model-assisted design of experiments as a concept for knowledge-based bioprocess development. *Bioprocess Biosyst. Eng.*, *42*, 867.
34. Arndt, L., Wiegmann, V., Kuchemüller, K. B., Baganz, F., Pörtner, R., & Möller, J. (2021). Model-based workflow for scale-up of process strategies developed in miniaturized bioreactor systems. *Biotechnology Progress*, *37*, e3122.
35. Kern, S., Platas-Barradas, O., Pörtner, R., & Frahm, B. (2016). Model-based strategy for cell culture seed train layout verified at lab scale. *Cytotechnology*, *68*, 1019.
36. Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. E. (2021). ELIXIR machine learning focus group. *Nature Methods*, *18*, 1122.
37. Ulonska, S., Kroll, P., Fricke, J., Clemens, C., Voges, R., Müller, M. M., & Herwig, C. (2018). Workflow for target-oriented parametrization of an enhanced mechanistic cell culture model. *Biotechnology Journal*, *13*, 1700395.
38. Anane, E., López C, D. C., Barz, T., Sin, G., Gernaey, K. V., Neubauer, P., & Cruz Bournazou, M. N. (2019). Output uncertainty of dynamic growth models: Effect of uncertain parameter estimates on model reliability. *Biochemical Engineering Journal*, *150*, 107247.
39. Hernández Rodríguez, T., Posch, C., Schmutzhard, J., Stettner, J., Weihs, C., Pörtner, R., & Frahm, B. (2019). Predicting industrial-scale cell culture seed trains "A Bayesian framework for model fitting and parameter estimation, dealing with uncertainty in measurements and model parameters, applied to a nonlinear kinetic cell culture model, using an MCMC method. *Biotechnology and Bioengineering*, *116*, 2944.
40. Bayer, B., Duerkop, M., Pörtner, R., & Möller, J. (2022). Figshare. https://doi.org/10.6084/m9.figshare.c.6250950

SUPPORTING INFORMATION

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

How to cite this article: Bayer, B., Duerkop, M., Pörtner, R., & Möller, J. (2023). Comparison of Mechanistic and Hybrid Modeling Approaches for Characterization of a CHO Cultivation Process: Requirements, Pitfalls and Solution Paths. *Biotechnology Journal*, *18*, e2200381. https://doi.org/10.1002/biot.202200381