



Model-assisted DoE software: optimization of growth and biocatalysis in *Saccharomyces cerevisiae* bioprocesses

André Moser¹ · Kim B. Kuchemüller² · Sahar Deppe³ · Tanja Hernández Rodríguez³ · Björn Frahm³ · Ralf Pörtner² · Volker C. Hass¹ · Johannes Möller²

Received: 18 July 2020 / Accepted: 5 November 2020 / Published online: 20 January 2021
© The Author(s) 2021

Abstract

Bioprocess development and optimization are still cost- and time-intensive due to the enormous number of experiments involved. In this study, the recently introduced model-assisted Design of Experiments (mDoE) concept (Möller et al. in *Bioproc Biosyst Eng* 42(5):867, <https://doi.org/10.1007/s00449-019-02089-7>, 2019) was extended and implemented into a software (“mDoE-toolbox”) to significantly reduce the number of required cultivations. The application of the toolbox is exemplary shown in two case studies with *Saccharomyces cerevisiae*. In the first case study, a fed-batch process was optimized with respect to the pH value and linearly rising feeding rates of glucose and nitrogen source. Using the mDoE-toolbox, the biomass concentration was increased by 30% compared to previously performed experiments. The second case study was the whole-cell biocatalysis of ethyl acetoacetate (EAA) to (*S*)-ethyl-3-hydroxybutyrate (E3HB), for which the feeding rates of glucose, nitrogen source, and EAA were optimized. An increase of 80% compared to a previously performed experiment with similar initial conditions was achieved for the E3HB concentration.

Keywords Biocatalysis · Monte Carlo methods · Fed-batch strategy · Model-assisted design of experiments · Quality by design

Abbreviations

DCW	Dry cell weight
DO	Dissolved oxygen
DoE	Design of experiments
EAA	Ethyl acetoacetate
EtOH	Ethanol
E3HB	(<i>S</i>)-Ethyl-3-hydroxybutyrate
Glc	Glucose

mDoE

	Model-assisted design of experiments
Pep	Peptone
<i>RMSD</i>	Weighted mean squared deviation
RS	Response surface
RSM	Response surface method
RQ	Respiratory quotient
YE	Yeast extract

List of symbols

c_k	Concentration of component k in the bioreactor (g l^{-1})
$c_{k,\text{feed}}$	Concentration of component k in the feed (g l^{-1})
d	Desirability function (–)
D_i	Combined desirability function (–)
F	Total feeding rate (ml min^{-1})
F_{c_k}	Feeding rate of component k (ml min^{-1})
F_k	Feeding rate of component k (ml min^{-1})
$F_{V,\text{in}}$	Total volumetric feed rate ($\text{g l}^{-1}\text{h}^{-1}$)
i (index)	Recommended factor combination (–)

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00449-020-02478-3>.

✉ Johannes Möller
johannes.moeller@tuhh.de

¹ Faculty of Medical and Life Sciences, Furtwangen University of Applied Sciences, Villingen-Schwenningen, Germany

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

³ Biotechnology and Bioprocess Engineering, Ostwestfalen-Lippe University of Applied Sciences and Arts, Lemgo, Germany

k (index)	EAA, Glc, N, EtOH, E3HB (–)
$k_{\text{weighting}}$	Weighting factor in RMSD (–)
K_s	Saturation constant (g l^{-1})
K_{sl}	Gradient slope (–)
L	Lowest value/lower limit (–)
MW_k	Molecular weight of the state variable k (g mol^{-1})
MW_s	Molecular weight of the substrate S (g mol^{-1})
n	Number of data points (–)
r_{c_k}	Total uptake/consumption rate (h^{-1})
\bar{r}_i	Average value of response i (–)
r_s	Substrate uptake rate ($\text{g l}^{-1}\text{h}^{-1}$)
R^2	Coefficient of determination (–)
S_{BC}	Biocatalysis educt (g l^{-1})
S_C	Carbon substrate (g l^{-1})
S_N	Nitrogen substrate (g l^{-1})
t	Time (h)
U	Highest value/upper limit (–)
V	Working volume of the bioreactor (l)
w	Weighting factor in D_i (–)
X_d	Dead biomass (g l^{-1})
X_i	Inactive biomass (g l^{-1})
X_p	Product forming biomass (g l^{-1})
X_{pri}	Autocatalytically active biomass (g l^{-1})
X_s	Structurally active biomass (g l^{-1})
X_{si}	Structurally inactive biomass (g l^{-1})
X_v	Viable biomass (g l^{-1})
$X_{50,l}$	Low location parameter of x (–)
$X_{50,h}$	High location parameter of x (–)
y_i	Measured value (–)
\bar{y}_i	Mean of measured data (–)
y_s	Simulated value (–)
Y_h	y at high x (–)
$Y_{i/S}$	Yield coefficient (–)
Y_l	y at low x (–)
Y_{mid}	y at medium x (–)
v_i in model	Stoichiometric coefficient (–)
v_i in MC simulation	Predicted variability (–)

Introduction

Biotechnology is expected to make a significant contribution to the establishment of a bio-based economy, since it offers new product manufacturing approaches and resource-efficient technologies [2, 3]. However, the development of a bio-economy requires new sustainable and environmentally friendly industrial production processes [4, 5]. Experiments for their development and optimization are usually designed using one-factor-at-a-time approaches and statistical Design

of Experiments (DoE) methods. DoE methods inevitably require a large number of experiments to be performed and analytically evaluated [6, 7]. Although the use of high-throughput systems is well established, e.g., for the screening of new enzymes or drugs [8, 9], they can be used with simplifications only for the actual process development (e.g., dimensioning of bioreactors, design of process control strategies, and scale-up) [10–12]. Although DoE can be used to identify correlations between process parameters and their influence on the final productivity, the complex bioprocess is reduced to a few key numbers (e.g., final product concentration), and the dynamics of growth and metabolism are not sufficiently taken into account [1, 13, 14]. In addition, the heuristic conception of a DoE by choosing the limits of the parameter space poses a particular challenge [15–17]. Thus, there is a high risk that the experiments carried out were wrongly chosen and have only insufficient validity, which results in further costs and time delays [1, 18].

To overcome the previously mentioned limitations of DoE, a new model-assisted Design of Experiments (mDoE) concept was recently introduced for knowledge-driven bioprocess development and optimization [1, 14, 19]. In the mDoE approach, the recommended experiments in statistical DoE designs are simulated using mathematical process models instead of being performed experimentally. The DoE designs (i.e., experimental space) are then evaluated based on the simulations, which enables the definition of a well-defined experimental space with a significantly reduced number of experiments to be performed experimentally. Besides the significant reduction of the number of experiments, the use of mathematical process models is nowadays seen as a sustainable part of a knowledge-driven bioprocess development strategy, because they contribute to the scientific understanding of the process [12, 20–22]. So far, the successful application of the mDoE approach in the field of medium and feeding strategy optimization for an antibody-producing Chinese Hamster Ovary cell line was shown [1, 14, 19]. Among the here presented application in the field of bio-economy, mDoE is currently used in optimization studies with algae, stem cells, and different mammalian producer cell lines.

In this study, the mDoE concept is incorporated into a software toolbox (“mDoE-toolbox”) for efficient design and optimization of biotechnological processes with a reduced number of experiments. In general, the toolbox can be used for different applications such as cell culture, algae, and yeast. Here, the application of the mDoE-toolbox is shown in two optimization case studies with *Saccharomyces cerevisiae*. In the first case study (S1), the cultivation conditions of a fed-batch process were optimized to increase the biomass concentration. The chosen factors were the pH value of the medium and the feeding rates of linearly rising feeding rates for glucose (F_{Glc}) and nitrogen source (F_{N}). In the second

case study (S2), the concentration of (*S*)-ethyl-3-hydroxybutyrate (E3HB) was maximized in the biocatalytic conversion of ethyl acetoacetate (EAA) to E3HB based on constant feeding rates for EAA (F_{EAA}), glucose (F_{Glc}), and nitrogen source (F_N).

mDoE-toolbox

The mDoE concept (see [1, 14]) was incorporated into a software toolbox, implemented in MATLAB (V2018a) and R (V3.5.1). The main parts of the mDoE-toolbox are the combination of a mathematical process model, including model-parametric uncertainties with the computational planning and evaluation of DoE designs.

In the beginning, the objective of the study (i.e., maximization of product concentration and minimization of inhibitory component) is defined. Then, the biotechnological system is modeled first, as can be seen in the structural workflow in Fig. 1 box I. Thus, prior knowledge (e.g., pre-experiments and literature) about the strain is used to define mathematical expressions for cell growth, metabolism, and productivity [1]. It should be noticed that process modeling itself is a rather undefined work, and a variety of models and modeling approaches of different considered complexity exist in the literature [24–25]. The mDoE-toolbox is designed to be applied in the initial phase of process development for which very little data are available. Therefore, structurally simple [14, 26] or generalized models [27] are

applied, for which model parameters can be adapted based on few data points typically generated in medium tests or first cultivations. After defining a mathematical model, the model-parametric uncertainties are derived with Monte Carlo sampling based on the experimental uncertainty (i.e., measurement error, Fig. 1 box II). Therefore, the expected process variability based on the measurement errors is simulated and later used in the DoE evaluation [19, 28]. Next, the experimental factors and responses are defined (Fig. 1 box III) with individual boundary values, e.g., a tolerated concentration of an inhibitory component or a minimal required product concentration. A DoE design, such as an optimal design [29, 30] or Box–Behnken design [31, 32], is subsequently planned. Additionally, mDoE enables the *in silico* comparison of different DoE designs, which is not targeted in this study.

For each recommended factor combination i , the time courses of the modeled state variables (e.g., cell weight, substrate, and product concentration) are simulated multiple times (Monte Carlo simulations, Fig. 1 box IV) taking into account the previously determined parameter probability functions (box II). From these simulations, the average expected response \bar{r}_i (e.g., average maximal cell dry weight) and the variability v_i are calculated. Due to Monte Carlo simulations, v_i of the response is expressed as the difference between the 10% and 90% quantiles of simulations [19].

In the next step, \bar{r}_i and v_i are used for the computational evaluation (box V) of the former planned experimental

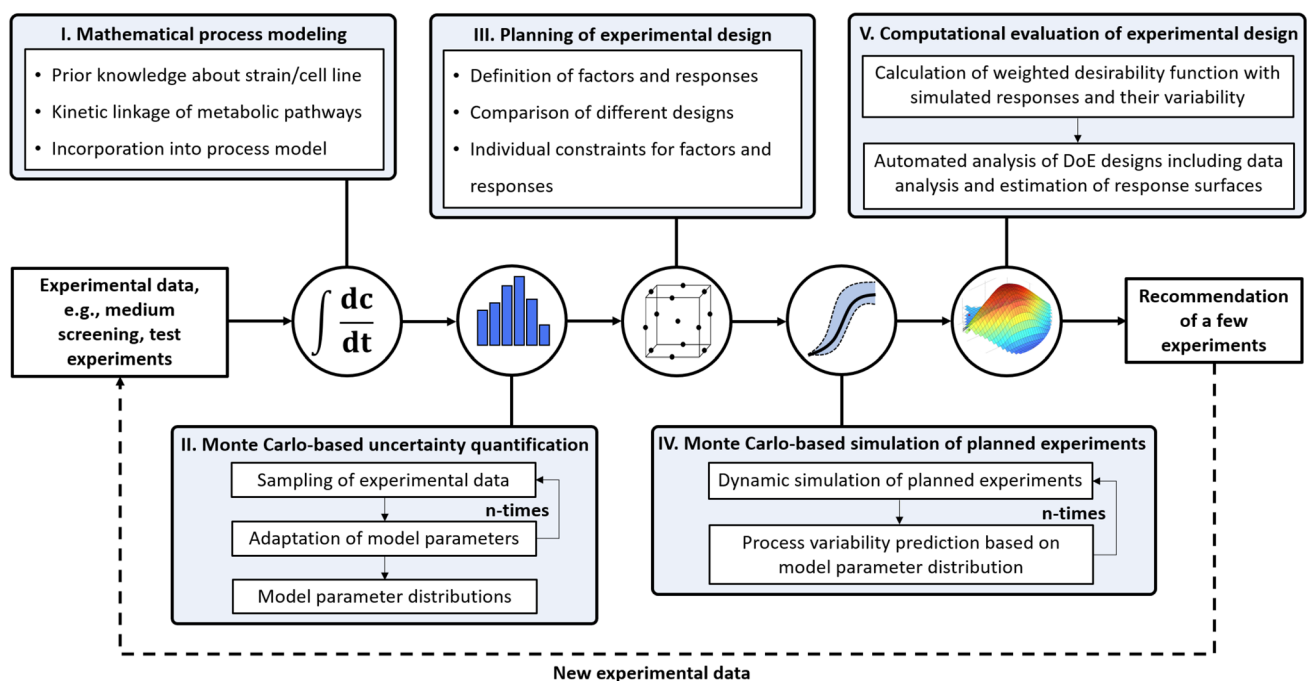


Fig. 1 Structural workflow of mDoE-toolbox consisting of the combination of mathematical process models and classical DoE under the consideration of model-parametric uncertainty based on experimental variability [1, 19]

design (box III). Therefore, both \bar{r}_i and v_i are summarized into a combined objective/desirability function (desirability at experimental factor combinations i - D_i) for each planned experiment in the DoE designs [33, 34]. This enables an evaluation of each planned experiment with respect to its simulated average and its expected variability with the aim of simultaneously maximizing \bar{r}_i and reducing v_i . The evaluation of DoEs using D_i reflects a risk-based approach. An experiment with a high D_i is favorable and a low D_i indicates a high variability and/or a low average response, which is not desired. After calculating D_i for all planned and simulated experiments, the experimental design planned in box III is analyzed, and response surface (RS) plots are generated automatically for visualization. Only a few (e.g., 2–4) experiments with the highest D_i are recommended to be performed, experiments with low D_i are neglected, which enables a significantly reduced number of experiments compared to the initially planned DoE design (box II).

Using the mDoE-toolbox, the available knowledge can be captured in the mathematical model, which can serve as a basis for advanced process understanding and digital twins [36–37]. In this way, the new data obtained from the recommended experiments can be used to re-adapt the model parameters and their probability distribution or to modify the model structure if so far unknown effects were identified [19, 24, 28].

Materials and methods

S1: optimization of fed-batch strategy for maximization of dry cell weight

Genetically unmodified *Saccharomyces cerevisiae* (Agrano, Germany, commercial strain used for industrial food production) was cultivated using complex media consisting of water, glucose (Glc), yeast extract (YE), and soy peptone (Pep, all Roth, Germany). No preculture was carried out, and dried yeast was directly inoculated. An overview of the

performed experiments in S1 with the medium and feed compositions used is shown in Table 1.

Experiments for modeling

For the parameterization of the pH-related model part (see Section “Mathematical process model”), four cultivations with different pH were performed in 1 l baffled shake flasks (500 ml working volume, Schott, Germany), which were shaken at 170 rpm (1.9 cm shaking diameter, MaxQ4000, Thermo Fisher Scientific, USA) with initial $c_{DCW} = 10 \text{ g l}^{-1}$. The temperature was controlled at 30 °C. The pH was adjusted initially and maintained manually (pH = 3, 4, 5, 6, respectively) to the desired value. In all experiments, the pH was adjusted using 20 wt% potassium hydroxide solution or 20 wt% phosphoric acid (both VWR, Germany). One feed pulse (glucose and nitrogen source) of 50 ml was added to each flask after 24 h (feed concentration in Table 1).

After adjusting the pH part of the model, further model parameterization was done based on historical data of three bioreactor (2 l working volume, Biostat B, Sartorius, Germany) fed-batch cultivations with different initial concentrations and feed compositions (see Table 1). Gassing was manually set in relation to the state of the process between 1 and 2 vvm (max. 2 l min^{-1}), and stirring was held between 500 and 800 rpm to maintain a DO above 10%. The pH was automatically controlled at 5. Temperature was set to 30 °C.

Recommended experiments from mDoE-toolbox

For the experiments recommended from the mDoE-toolbox, the initial dry cell weight (DCW) was adjusted to $c_{DCW} = 2 \text{ g l}^{-1}$ and the initial conditions of the complex culture medium were prepared, as shown in Table 1. The recommended feeding strategy for the mDoE experiments was a linearly rising feeding rate starting at $t_{\text{start}} = 1 \text{ h}$. The final feed volume flow at $t_{\text{end}} = 48 \text{ h}$ was determined by the mDoE-toolbox for F_{Glc} and F_{N} separately. The feed rates were determined using the following function:

Table 1 Initial and feed volume, as well as initial and feed concentration of c_{DCW} , c_{Glc} , c_{YE} and c_{Pep} of every experiment in S1

	Medium					F_{Glc}	F_{N}	
	Volume (l)	Biomass (g l^{-1})	c_{Glc} (g l^{-1})	c_{YE} (g l^{-1})	c_{Pep} (g l^{-1})	c_{Glc} (g l^{-1})	c_{YE} (g l^{-1})	c_{Pep} (g l^{-1})
Modeling								
Cultivation I	1.00	3.0	6.0	1.4	2.3	290	50.0	76.0
Cultivation II	1.00	3.0	6.0	1.4	2.3	290	50.0	76.0
Cultivation III	0.75	20.0	7.9	1.4	2.3	290	85.0	135
pH-Exp	0.30	10.0	30.0	2.9	4.5	180	17.0	27.0
mDoE-toolbox								
S1	0.70	2.0	15.0	1.4	2.3	400	130	270

$$F_k(t) = \frac{F_{k, \text{end}}}{t_{\text{end}}} \times (t - t_{\text{start}}). \quad (1)$$

For $t < t_{\text{start}}$, $F_k(t)$ equals zero.

Linear feed strategies have been chosen instead of exponential strategies, as these are easier to handle and less risky. Slightly too large feed rates resulting from exponential strategies quickly result in over-feeding. The pH was held constant using 20 wt% potassium hydroxide solution or 20 wt% phosphoric acid (both VWR). The temperature was controlled at 30 °C.

S2: optimization of fed-batch strategy for biocatalysis

Genetically unmodified *Saccharomyces cerevisiae* (Agrano, Germany) served as the whole-cell biocatalyst. In the biocatalysis (S2), the media consist of Glc, YE, Pep, and EAA (all Roth, Germany). The temperature was set at 30 °C and pH 5. The pH was controlled with the addition of 20 wt% potassium hydroxide solution or 20 wt% phosphoric acid. The airflow rate was adjusted at 1–2 vvm and stirring rate at 800 rpm to maintain aerobic conditions (DO > 10%). Antifoam was fed when required. The experiments for the biocatalysis part (S2) are shown in Table 2.

Experiments for modeling

Two experiments for model parameterization were performed in 5 l (Biocatalysis I) (BioFlo, Eppendorf, Germany) and 20 l reactor (Biocatalysis II) (BioStat C, Sartorius). Initial and feed conditions are shown in Table 2. The first experiment was an initial test experiment and the second experiment was with a high initial c_{DCW} .

Recommended experiments from mDoE-toolbox

The recommended biocatalysis experiments were performed in 1 l stirred bioreactors (Medorex, Germany). Initial dry biomass density (cell dry weight c_{DCW}) of 40 g l⁻¹ was

chosen. Constant F_{Glc} , F_{N} , and F_{EAA} were defined as factors. Feeding started immediately after inoculation. Since the independently predicted feed flow rates for glucose and nitrogen were too low for the available pumps, they were fed together. No online off-gas measurement was performed.

Analytics

Concentrations of ethanol (c_{EtOH}), c_{Glc} , c_{EAA} , and c_{E3HB} were quantified with high pressure liquid chromatography using a Rezex ROA column (300 × 7.8 mm, Phenomenex, USA) and 0.005 N sulfuric acid as the aqueous mobile phase according to the manufacturer's protocol. c_{DCW} was determined by filtrating the medium through cellulose acetate filters (0.45 μm, VWR, US) and measuring the weight of the retentate after drying in a moisture analyzer (MA45, Sartorius, Germany). The percentages of oxygen and carbon dioxide in the off-gas were measured via an extractive gas analyzer (Sick, Germany). The respiratory quotient (RQ) is then calculated from the quotient of carbon dioxide produced divided by the oxygen consumed [38, 39]:

$$\text{RQ} = \frac{\text{CO}_{2, \text{produced}} [\text{mol}]}{\text{O}_{2, \text{consumed}} [\text{mol}]} \quad (2)$$

The pH value of the medium was measured in situ with an amperometric pH Probe (405-DPAS-SC-K8S, Mettler Toledo, US, and EasyFerm Plus PHI S8 225, Hamilton, US). The pH values in the shaking flask experiments were controlled offline with a benchtop pH meter (FiveEasy F20, Mettler Toledo, US). The DO in the biocatalysis experiments was measured with an optical dissolved oxygen (DO) probe (VisiFerm DO ECS 225 H0, Hamilton, US).

Mathematical process model

A novel structured compartment model, capable of being adapted to different biotechnological expression systems (e.g., bacteria, yeast, fungi, mammalian cell lines), was used to describe yeast growth, metabolism, and biocatalysis

Table 2 Initial and feed volume, as well as initial and feed concentration of c_{DCW} , c_{Glc} , c_{YE} , and c_{Pep} of the biocatalysis experiment of reference process S2; EAA is fed as a pure component

	Medium					F_{Glc}	F_{N}	
	Volume (l)	Biomass (g l ⁻¹)	c_{Glc} (g l ⁻¹)	c_{YE} (g l ⁻¹)	c_{Pep} (g l ⁻¹)	c_{Glc} (g l ⁻¹)	c_{YE} (g l ⁻¹)	c_{Pep} (g l ⁻¹)
Modeling								
Biocatalysis I	4.0	45.5	0.1	0	0	400	130	270
Biocatalysis II	10.0	78.2	0.1	0	0	400	130	270
mDoE-toolbox								
S2	0.6	40.0	2.0	1.6	2.4	Adjusted individually		

[27, 40]. The model was previously introduced by Brüning et al. [27] and is briefly explained in the following. The main part of the model is the segregation of the biomass into six distinct compartments, which are linked and individually described by mathematical equations representing different essential metabolic tasks. The detailed figure of the six model compartments can be found in the Electronic Supplementary Material (ESM) Fig. S1.

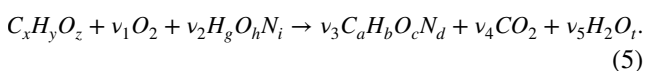
The following compartments are considered: an autocatalytically active biomass (Xpri), a product forming (Xp), a biocatalytically inactive (Xi), a structurally active (Xs) and inactive (Xsi), and a dead biomass (Xd) compartment. Biomass synthesis is based on a carbon (SC) and a nitrogen substrate (SN) and biocatalysis is modeled based on an educt (SBC). Furthermore, physicochemical state variables, such as DO, pH, and temperature, have a direct influence on cell metabolism, biomass growth and/or biocatalytic activity [27]. The uptake rates (rS) of the substrates (S) are rate-limiting steps, which are modeled by Monod kinetics typically used in bioprocess modeling [14, 26, 28, 40]:

$$rS = rS_{\max} \left(\frac{S}{K_s + S} \right) \times \prod_{i=1}^n f_{\text{Dsig}}(x_i). \quad (3)$$

K_s is the half-saturation constant. The Monod-like term for the uptake rates is multiplied with the product of multiple double sigmoidal functions (f_{Dsig}) of the state variables (x_i), which describe changes of the cell metabolism [27, 41]:

$$f_{\text{Dsig}}(x) = \left(Y_l + \frac{Y_{\text{mid}} - Y_l}{1 + e^{-K_{sl}(x - X_{50,l})}} \right) \times \left(1 + \frac{(Y_h/Y_{\text{mid}} - 1)}{1 + e^{-K_{sh}(x - X_{50,h})}} \right). \quad (4)$$

The value of a state variable is described by x . Y_l is the value of f_{Dsig} at low x , and Y_h is the value at high x . Y_{mid} is the value between $X_{50,l}$ and $X_{50,h}$, which are location parameters of the low/high side of the function. K_{sl} determines the gradient of the slope [27]. The sigmoidal functions are also used to describe the influence of operating parameters on the activation and inactivation rates as well as yield coefficients. This structure enables the description of complex changes in multiple metabolic pathways and their intensity, e.g., for biomass formation, overflow metabolisms, biocatalysis, and complete oxidation under aerobic and anaerobic conditions. Moreover, the product of $f_{\text{Dsig}}(x)$ is used to account for combined influences such as substrate/product inhibition and/or pH, DO, or temperature on the uptake rates. A parameterization strategy for the double sigmoidal functions is described in [41]. Each pathway is represented by the same, generalized stoichiometric function:



The stoichiometric coefficients v_i were determined previously [42] and used according to:

$$Y_{i/S} = v_i \left(\frac{MW_i}{MW_S} \right), \quad (6)$$

where MW_i is the molecular weight for the state variable (e.g., biomass, O_2 , by-product) and MW_S for the substrate [27, 43]. The yield coefficients, describing the formation of a substance i based on the substrate S ($Y_{i/S}$) are used in the calculation of production and uptake rates, whereas rates of each pathway are summed up to total rates. The total rates rc_i are then used in general mass balances for each component with the concentrations of components in the feed $c_{i,\text{feed}}$ and their concentration in the bioreactor c_i :

$$\frac{dc_i}{dt} = \underbrace{rc_i^+ \times Xv}_{\text{production}} - \underbrace{rc_i^- \times Xv}_{\text{uptake}} + \underbrace{c_{i,\text{feed}} \times \frac{F_{c_i}}{V}}_{\text{input}} - \underbrace{c_i \times \frac{F_{V,\text{in}}}{V}}_{\text{dilution}}. \quad (7)$$

Monte Carlo-based uncertainty quantification

To quantify the variability of the model simulations in the mDoE-toolbox, the model-parametric uncertainties are determined using Monte Carlo sampling with repeated parameter adaptations [19]. In brief, the determined standard deviation of each experimental data point was considered to be independent and normally distributed. For the initial values, the standard deviation was assumed to be 5%. The individual biomass compartments considered in the Six-compartment model (Section “**Mathematical process model**”) could not be experimentally determined and were presumed with a standard deviation of 10%. The standard deviations of the set pH value, temperature, DO, and feeding rates, as well as their concentrations, were defined to be 5% based on the typical standard deviations in bioprocesses (i.e., expert knowledge) [19, 44, 45]. The model-parametric uncertainty was determined based on the experimental uncertainty using multiple parameterization runs (Monte Carlo samples). Due to limited computational power, 116 adaptations were performed in case study S1 and 240 adaptations in case study S2. Model parameters were adapted by minimizing the weighted root-mean-square deviation $RMSD$ [14, 19, 27, 46]. The $RMSD$ is calculated from the squared difference between the measured value y_m and the simulated value y_s , multiplied by a factor for weighting individual data points $k_{\text{weighting}}$, and divided by the number of data points n in the data set:

$$RMSD = \sqrt{\sum_{i=1}^n \frac{(y_{s,i} - y_{m,i})^2}{n} \cdot k_{\text{weighting}}}. \quad (8)$$

Only high $c_{DCW} > 100 \text{ g l}^{-1}$ were weighted by 0.5 and no weighting was used for other state variables. The individually adapted model parameters, their medians, 10% and 90% quantiles (ESM: Tables S1 and S2), as well as their distributions are shown in the ESM.

The simulations using the determined parameters were additionally evaluated using the coefficient of determination (R^2), which includes the differences between simulated $y_{s,i}$ and experimental data y_i as well as the differences between experimental data and their mean \bar{y} [1, 19, 47]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_{s,i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (9)$$

R^2 lies between minus infinity and one. If R^2 is one, the data points correspond precisely to the solution of the model. If R^2 is less than zero, the mean of the measured data points is closer to the mean result than the solution of the model [19].

Planning of experimental design

The factor settings of the DoE designs were determined as described in the following: First, a large number of points ($> 10^6$) were randomly distributed in the three-dimensional design space (i.e., three investigated factors). Then, clusters in these points were determined using the k -means algorithm [48, 49]. These clusters are partitioned into the k sets corresponding to the number of experiments in the DoE design. The resulting k cluster centers are the factor combinations (i.e., planned experiments) of the DoE design. Based on this algorithm and previous studies, a total of 29 experiments were planned, which were later individually simulated and evaluated using the mDoE-toolbox. The main advantage of this method, among other approaches, is the universal application to any number of investigated factors and experimental spaces of any shape [50].

Monte Carlo-based simulation of planned experiments

Instead of performing each planned cultivation from the DoE design, they were first simulated (see Fig. 1) with the developed process model. Due to computational power, 30 Monte Carlo simulations (Section “Monte Carlo-based uncertainty quantification”) (C-eStIM, Germany) were performed for each planned experiment, so that the propagated uncertainty of the simulations was quantified. The model parameter values were drawn using Latin Hypercube Sampling using the R-Package “lhs” (V1.0.2) [51, 52]. The interval boundaries are determined by calculating the 10% and 90% quantiles using the Type R-7 method provided in R.

Computational evaluation of experimental design

First, for each planned factor combination i , the average expected response (\bar{r}_i) is calculated based on the Monte Carlo simulations. Furthermore, v_i is calculated as the difference of the 10% and the 90% quantile, and is used as a measure of the expected process variability. In the mDoE-toolbox, the maximization of \bar{r}_i is targeted with a simultaneously minimization of v_i . Therefore, for both \bar{r}_i and v_i , individual desirability functions $d(\bar{r}_i$ or v_i) are calculated by rescaling between 0 and 1. $d(\bar{r}_i$ or v_i) are based on the minimal response $L(\bar{r}$ or $v)$ and the maximal response $U(\bar{r}$ or $v)$ of all \bar{r}_i and v_i (vector including all i experiments donated as \bar{r} or v). Therefore, the desirability function $d(\bar{r}_i)$ is in the optimization range ($U(\bar{r}) - L(\bar{r})$). For the maximization of \bar{r}_i , $d(\bar{r}_i)$ is calculated as follows:

$$d(\bar{r}_i) = \left(\frac{\bar{r}_i - L(\bar{r})}{U(\bar{r}) - L(\bar{r})} \right). \quad (10)$$

For the minimization of v_i , $d(v_i)$ is inversely calculated, i.e., a high v_i has a low $d(v_i)$ and *vice versa*:

$$d(v_i) = \left(\frac{v_i - U(v)}{L(v) - U(v)} \right). \quad (11)$$

In the mDoE-toolbox, $d(\bar{r}_i)$ and $d(v_i)$ are combined into one numerical value D_i to quantify the average value and its variability of each planned experiment (see Section 1.1) into D_i including weighting factors w :

$$D_i = w_1 \times d(\bar{r}_i) + w_2 \times d(v_i) \quad (12)$$

for which

$$w_1 + w_2 = 1. \quad (13)$$

By this approach, a risk-based evaluation of the planned designs is enabled and $w(v_i)$ reflects the percentage at which v_i is considered. In this study, $w_1 = 0.8$ and $w_2 = 0.2$. Contour and 3D plots were generated with Gnuplot 5.2.8.

Results and discussion

The mDoE-toolbox software was tested on two optimization studies with *Saccharomyces cerevisiae* (S1 and S2, respectively). First, the aim was to maximize c_{DCW} after 48 h (S1) based on the experimental factors pH, as well as F_{Glc} and F_{N-} . Second, the biocatalytic conversion from EAA to E3HB was optimized (S2). E3HB should be maximized based on F_{EAA} , F_{Glc} and F_{N-} . EAA shows inhibitory effects above a concentration of $c_{EAA} = 0.5 \text{ g l}^{-1}$ [53, 54]. In both processes, ethanol formation is crucial due to its inhibitory effect on cell growth and biocatalysis [55, 56]. In addition, *Saccharomyces*

cerevisiae produces ethanol even under aerobic conditions if glucose concentration is above a certain limit. This phenomenon is known as the Crabtree effect and should be minimized to optimize growth and biocatalysis [57, 58].

Monte Carlo-based uncertainty quantification

Model parameters and their distributions were determined using Monte Carlo sampling (see Fig. 1 box II), as explained in Section “Monte Carlo-based uncertainty quantification”. Therefore, three sets of experiments (see Table 1 and 2) were used. The first set consists of shaking flask experiments to adapt the parameters for the pH model. The second set consists of three historical fed-batch cultivations to model the

growth of yeast, uptake rates, and production rates in relation to critical process parameters, e.g., glucose and ethanol concentration (both in Table 1). The last set was designed based on literature and was used to identify the parameters for the biocatalysis (Table 2).

Growth and metabolic model parameters (S1)

The growth and metabolic model parameters targeted in case study S1 were adjusted using data of three fed-batch cultivations (see Table 1). These model parameters describe the uptake of glucose, ethanol, and nitrogen, the activation, inactivation, and mortality rates, as well as the general yield coefficients for glucose and ethanol. Furthermore, the

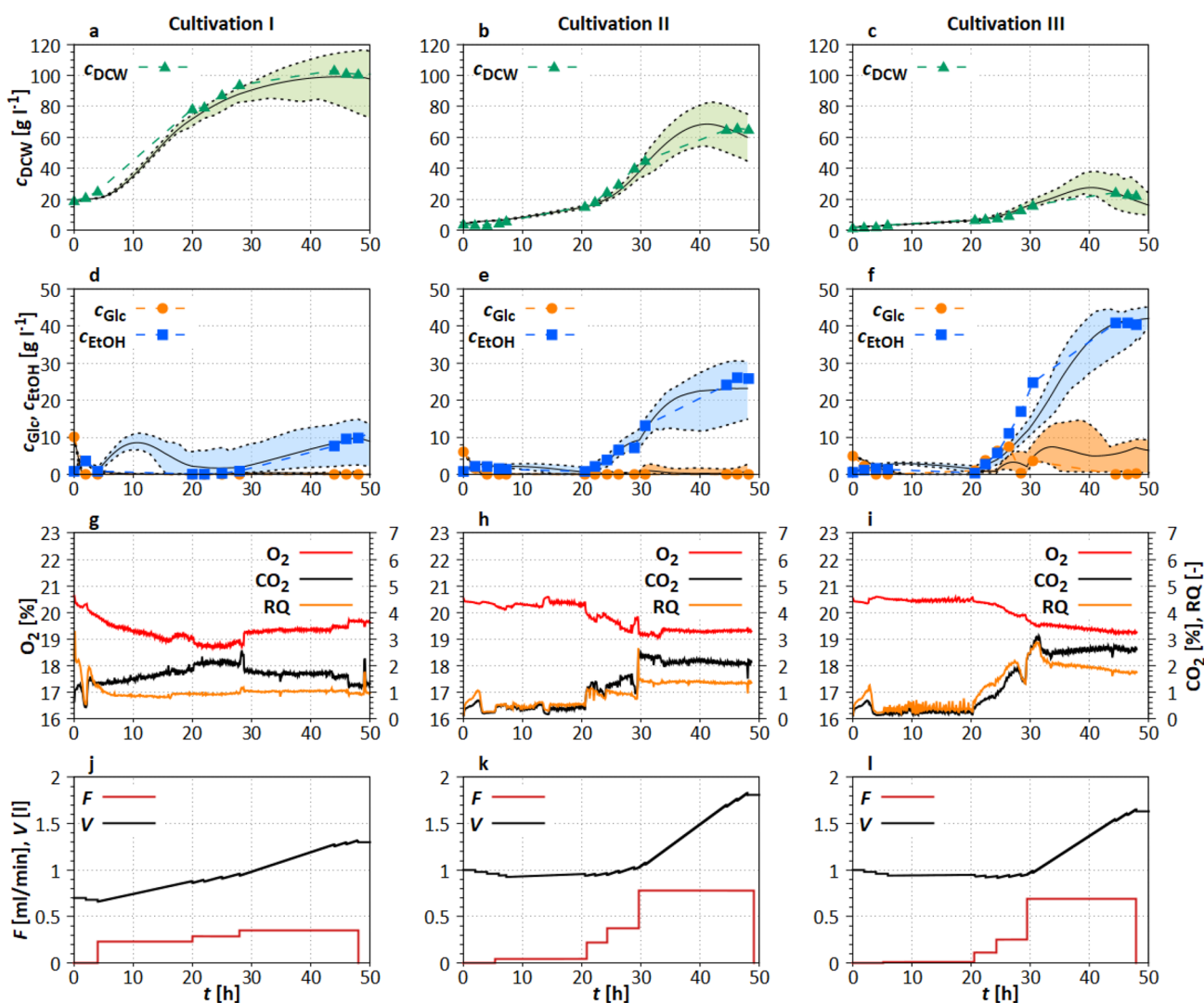


Fig. 2 Comparison of experimental data to adapted model of three initial cultivations (see Table 1), **a–f** solid lines represent the mean of 116 Monte Carlo simulations (Section “Monte Carlo-based simulation of planned experiments”); dashed lines represent the 10% and

90% quantiles of the simulations; **g–i** online data of the off-gas measurement as well as the calculated respiratory quotient. **j–l** Calculated V and feeding rates (F). Experimental settings and the used reactor are shown in Section “Experiments for modeling”

parameters of the sigmoidal functions for the influence of glucose limitation and ethanol inhibition on the glucose and ethanol uptake as well as the biomass inactivation rate were identified (see ESM Section 3.1). The comparison of the experimental data with the Monte Carlo-based simulations including 10% and 90% quantiles are shown in Fig. 2a–f. Furthermore, gassing rates and experimental online data (Fig. 2g, h), and calculated total volume V and total feeding F (Fig. 2j–l) are shown.

Cultivations I and II were initial test cultivations aiming to achieve a high biomass density. For this purpose, different initial biomass concentrations were chosen. Cultivation III was performed with a feeding strategy which should lead to ethanol inhibition (Table 1).

Cultivation I In cultivation I, the biomass (Fig. 2a) increases from initially $c_{DCW} = 18 \text{ g l}^{-1}$ to $\approx 100 \text{ g l}^{-1}$ ($t = 42 \text{ h}$). An initial c_{DCW} that high would not be used in “real” bioprocesses and was just utilized for the purpose of model parameter adaptation. The ethanol concentration was relatively low below $c_{EtOH} = 10 \text{ g l}^{-1}$ throughout the experiment and glucose concentration was not measurable after $t = 2 \text{ h}$. These results are reflected in the course of the respiratory quotient (Fig. 2b, RQ), which was constantly around

one, indicating a low ethanol formation. Feeding rate was increased stepwise to $F = 0.35 \text{ ml min}^{-1}$ (Fig. 2j).

Cultivation II In cultivation II, a biomass density of $c_{DCW} = 64 \text{ g l}^{-1}$ was achieved at the end of the process and an ethanol concentration of $c_{EtOH} = 28 \text{ g l}^{-1}$ was determined (Fig. 2d), leading to low growth inhibition. Glucose was directly consumed when fed, and therefore, the measured glucose concentrations were about 0 g l^{-1} after $t = 10 \text{ h}$. The rather strong ethanol production after $t = 30 \text{ h}$ is reflected in the RQ (Fig. 2e). Feeding (Fig. 2h) was higher than in cultivation I despite a lower initial biomass density.

Cultivation III As can be seen in Fig. 2c, only $c_{DCW} = 24 \text{ g l}^{-1}$ was formed at $t = 44 \text{ h}$, but over $c_{EtOH} = 40 \text{ g l}^{-1}$ was produced during the same time period. This trend was also confirmed by the RQ (Fig. 2f), which was clearly above one from $t = 20 \text{ h}$ onwards, indicating an increased CO_2 formation during ethanol production. Feeding (Fig. 2i) was designed to induce ethanol inhibition (i.e., over-feeding) and was increased stepwise up to $F = 0.7 \text{ ml min}^{-1}$ [59].

Overall, the model parameters could be adapted well to the process data with an R^2 above 0.85 (total for c_{DCW} , c_{Glc} , c_{EtOH}) comparing the experimental data to the mean of the simulations for every experiment (Table 3).

In addition, the modeling of high biomass densities and ethanol inhibition was adapted sufficiently. The width of the uncertainty band (10% and 90% quantiles) of the simulations (Fig. 2a–f) was narrow, indicating a reliable estimation of model parameters. All model parameters and their individual distribution are shown in ESM Figs. S3–S5.

Table 3 Total R^2 for average model parameters in Monte Carlo-based uncertainty quantification (S1—cultivation and S2—biocatalysis)

Experiment	R^2
S1 Cultivation I	0.97
Cultivation II	0.98
Cultivation III	0.85
S2 Biocatalysis I	0.92
Biocatalysis II	0.96

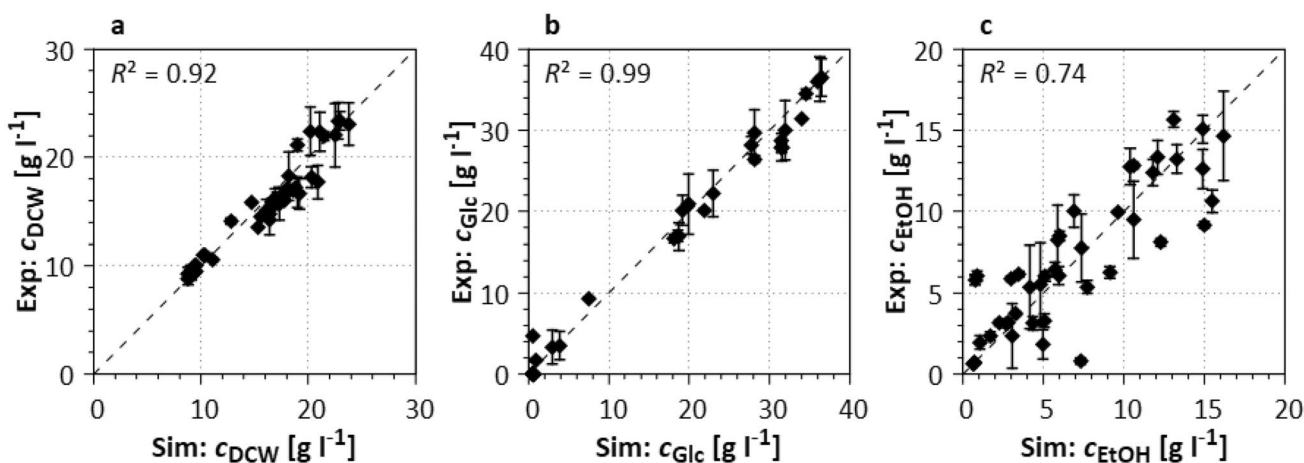


Fig. 3 Comparison of experimental (Exp) data and simulated (Sim) data for the adaptation of pH-related model parameters (see Section “Experiments for modeling”). Error bars show standard deviation of two parallel shaking flask experiments for each investigated pH

(4, 5, 6, and 7, respectively). The quality of fit is represented by R^2 (optimal simulation $x = y$). For each pH, the individual growth curves are shown in the ESM Fig. S2. Experimental settings are specified in Section “Experiments for modeling”

pH-dependent model parameters (S1)

As can be seen in Fig. 3, the influence of a wide pH range (4–7) on the growth and metabolism of *Saccharomyces cerevisiae* can be sufficiently reflected by the model with the adapted model parameters in case study S1. The concentrations of biomass and glucose can be well reflected with an R^2 above 0.92. The simulation of the ethanol concentration is sufficient ($R^2 = 0.74$).

For all cultivations (see ESM Fig. S2 for individual plots), cells grew up to approx. $c_{DCW} = 20 - 25 \text{ g l}^{-1}$ in the first hours including strong ethanol production. After glucose depletion, cell growth stagnated, and ethanol was taken up until the feed pulse at $t = 24 \text{ h}$ was fed. After the feed pulse, the biomass density increased again, while glucose was consumed and ethanol produced. No strong growth inhibitions were seen for the different pH values investigated. The biomass densities at the end of the process were found to be slightly higher at pH 5 (compared to pH 4 and 6, respectively), and 10% higher than at pH 3. Furthermore, the glucose and ethanol consumption rates increased with increasing pH. The growth rates at pH 4 and pH 5 were equally the highest with $0.105 \pm 0.005 \text{ h}^{-1}$.

Using the experimental data, the model parameters related to the pH-dependent glucose metabolism and uptake, and the segregation of glucose into biomass and ethanol production were determined. These parameters were not used in the Monte Carlo parameter adaptation (Section “[Monte Carlo-based uncertainty quantification](#)”) to speed up the calculations of the parameterization algorithm and were kept constant thereafter.

Biocatalytic model part (S2)

The biocatalysis model used in S2 was adapted on data of two experiments, based on the literature [53, 54]. The focus was on those model parameters, characterizing the biocatalysis, and therefore partly differs from the previously chosen parameters in case study S1. The new parameters are listed in ESM Section 3.2 and describe the uptake of EAA, glucose, ethanol and nitrogen, the activation, inactivation, and mortality rate, as well as the general yield coefficients for EAA, glucose, and ethanol. Furthermore, the parameters of the sigmoidal functions quantifying the influence of ethanol inhibition, glucose limitation on glucose uptake, and EAA inhibition on the biomass inactivation rate were adapted to the experimental data. In Fig. 4, the comparison between experimental and simulation data of the biocatalysis parameterization experiments is shown.

Biocatalysis I The first experiment aimed at the production of E3HB with $c_{DCW} = 45 \text{ g l}^{-1}$. The feed and initial concentrations of the parameterization experiments are listed in Table 2. In the first experiment, biomass (Fig. 4a)

decreased to $c_{DCW} = 25 \text{ g l}^{-1}$ ($t = 26 \text{ h}$), first due to dilution by feeding and towards the end due to a higher mortality rate induced by toxically high concentrations of EAA. Ethanol concentration (Fig. 4c) rose from 30 g l^{-1} to over 40 g l^{-1} at $t = 26 \text{ h}$. Feeding was designed to increase c_{EAA} (Fig. 4e). Therefore, a constant $F_{Glc} = 0.65 \text{ ml min}^{-1}$ and a constant $F_{EAA} = 0.067 \text{ ml min}^{-1}$ were set during the first 28 h. Then, F_{Glc} was reduced and F_{EAA} was increased to identify potential EAA inhibition. Thus, c_{EAA} rose to over 4 g l^{-1} , and the resulting inhibition is reflected in rising c_{Glc} , despite the reduced glucose feed. E3HB constantly increased up to $c_{E3HB} = 24 \text{ g l}^{-1}$ ($t = 32 \text{ h}$) and then stopped increasing due to the EAA and potential ethanol inhibition.

Biocatalysis II In the second experiment, a higher initial biomass concentration of $c_{DCW} = 80 \text{ g l}^{-1}$ was used, and ethanol and glucose concentrations were kept below 0.1 g l^{-1} during the whole experiment (Fig. 4b, d). As in Cultivation I (S1), this high c_{DCW} was used for model parameter adaptation. After 48 h, $c_{E3HB} = 44 \text{ g l}^{-1}$ was reached, whereas c_{EAA} was constantly low and the biomass density decreased due to dilution. Constant glucose ($F_{Glc} = 3 \text{ ml min}^{-1}$) and EAA feeding rates ($F_{EAA} = 0.05 \text{ ml min}^{-1}$) were set. F_{EAA} was increased to $0.085 \text{ ml min}^{-1}$ in two steps.

Overall, the parameterized model parameters reflect the kinetics of biocatalysis satisfactorily. The calculated R^2 are shown Table 3 and are higher than 0.92 for both simulations for the experimental data compared to the mean simulation (summarized for c_{E3HB} , c_{EAA} , c_{DCW} , c_{EtOH}).

The biocatalytic metabolite concentrations are well reproduced by the model for both experiments. The 10% and 90% quantiles of the simulations (Fig. 4a–d) are small. The model parameters and their individual distributions are shown in ESM Figs. S6–S9.

Optimization of fed-batch process with mDoE-toolbox (S1)

The factors (pH, F_{Glc} , F_N , respectively) have been selected based on literature and experience in the cultivation of *Saccharomyces cerevisiae*. pH is described to have a strong influence on viability and growth rate [60]. Glucose is essential for cell growth, but high glucose feeding rates lead to ethanol formation due to the Crabtree effect and possibly ethanol inhibition [58–59]. A nitrogen source is essential for cell growth, but feeding needs a tight control to avoid dilution and to enable steady growth [61].

mDoE-toolbox (S1)

Planning of experimental design and Monte Carlo simulations Using the mDoE-toolbox explained in Sect. 1.1, rather widely distributed initial boundary

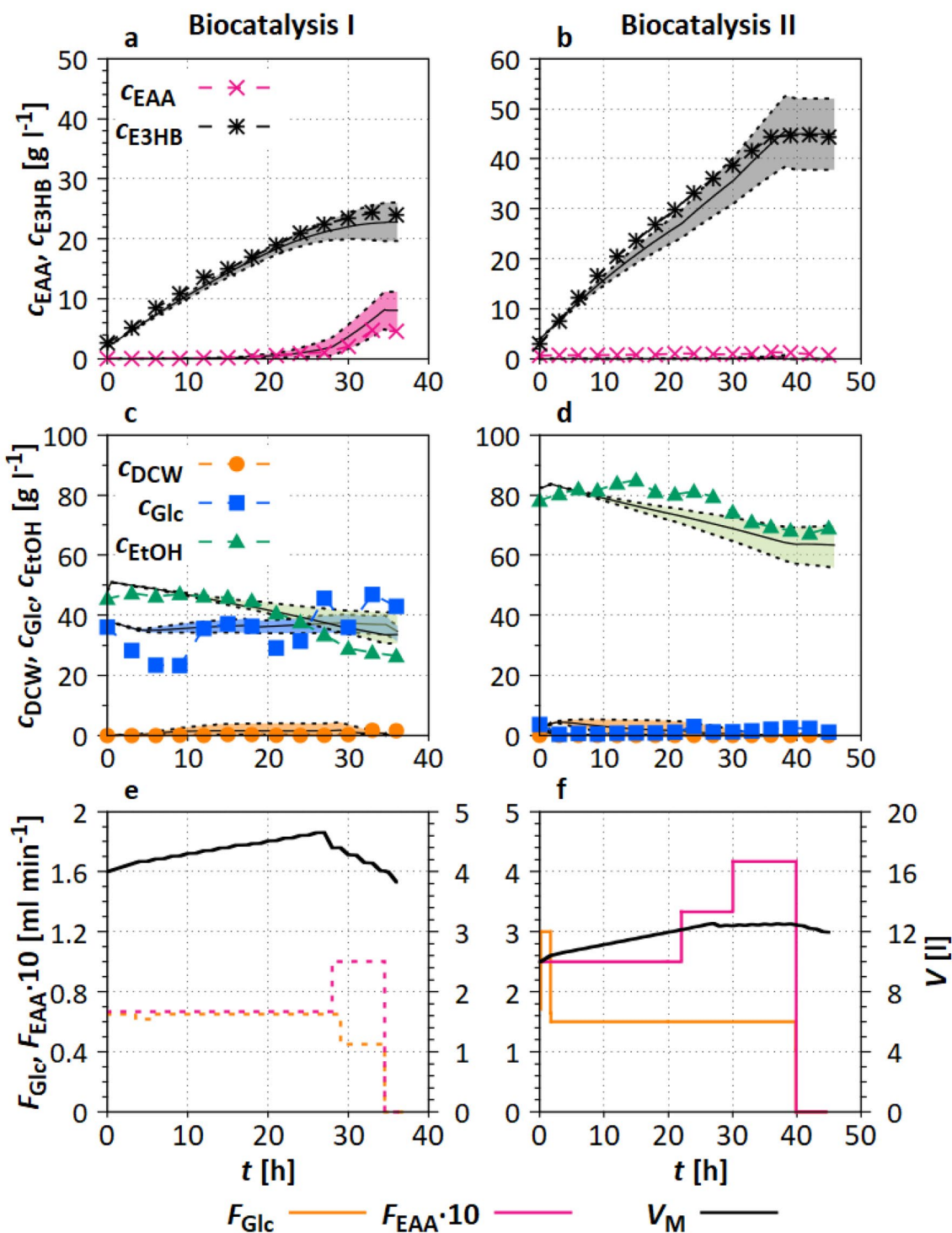


Fig. 4 Model parameter adaptation for biocatalysis (see Table 2). **a–d** solid line represents the mean of 116 Monte Carlo simulations (Section “Monte Carlo-based simulation of planned experiments”), dashed lines represent the 10% and 90% quantiles of the simulations;

e, f calculated V and F . Experimental settings and used reactors are specified in Section “S2: optimization of fed-batch strategy for biocatalysis”

values are defined first for the planning of the experimental design (Fig. 1, box III) [1]. Therefore, start of feeding was at $t = 1$ h. The initial boundaries for F_{Glc} were set between $0.1 \leq F_{Glc} \leq 1 \text{ ml min}^{-1}$, and F_N between

$0.05 \leq F_N \leq 0.6 \text{ ml min}^{-1}$. pH was varied between 3 and 7. Using the mDoE-toolbox, a total of 29 design points (i.e., planned experiments) were distributed in the

three-dimensional design space, determined by the previously defined boundaries (see Section “[Planning of experimental design](#)”). For each of the planned experiments, Monte Carlo simulations (Fig. 1, box IV) were performed, as explained in Section “[Monte Carlo-based simulation of planned experiments](#)”.

Computational evaluation of experimental design

From the Monte Carlo simulations, \bar{r}_i and v_i were calculated for the maximum c_{DCW} for each experimental setting i , which were further used to derive the desirability D_i (Fig. 1, box V). In Fig. 5, the desirability functions are plotted for $\text{pH} = 5.0$, $F_N = 0.20 \text{ ml min}^{-1}$, and $F_{Glc} = 0.42 \text{ ml min}^{-1}$, for which the highest D_i was calculated. Figure 5 shows the contour and 3D plots at these process conditions.

D_i is high for flow rates of $F_{Glc} < 0.5 \text{ ml min}^{-1}$ and $F_N < 0.3 \text{ ml min}^{-1}$ at $\text{pH} = 5.0$ (Fig. 5a). The influence of the pH value (Fig. 5b) on the desirability function is low, and a pH between 4 and 6 provides the best results. These results were additionally confirmed with the RSM in Fig. 5c and were low $F_N < 0.3 \text{ ml min}^{-1}$, and a pH between 4 and 6 shows again the best results. Using the mDoE-toolbox,

the experimental space was computationally simulated and evaluated. For the factors investigated, the resulting RS plots could hardly be predicted on experience solely. The in silico calculation and computational evaluation of the planned experimental design offer a knowledge-driven approach, which is the major advantage of the mDoE-toolbox.

mDoE-suggested experimental settings Based on the evaluation of the experimental space, further experiments

Table 4 Factor combination of the four yeast cultivation experiments determined with the usage of the mDoE-toolbox. F_{Glc} and F_N are the final maximum feed rates of the linearly rising feed at the end of the experiments

Exp.	F_{Glc} (ml min^{-1})	F_N (ml min^{-1})	pH (-)
#1	0.28	0.12	4
#2	0.28	0.12	6
#3	0.41	0.18	5
#4	0.54	0.25	5

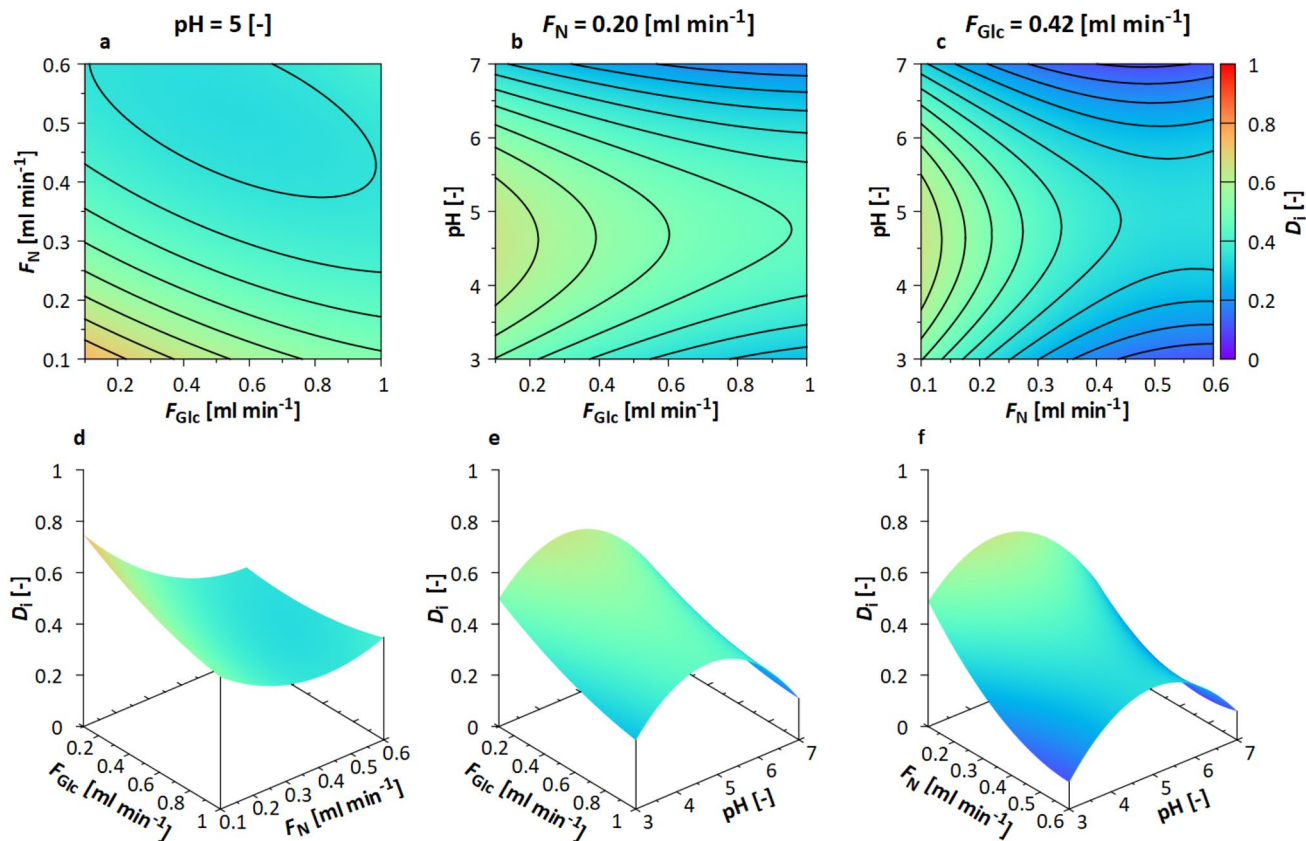


Fig. 5 Contour and 3D plot for the response surfaces based on D_i for the optimization of pH, F_{Glc} , and F_N in S1. The responses were calculated with Monte Carlo simulations in the mDoE-toolbox, as explained in Section 1.1. **a–c** Graphs are adjusted to $\text{pH} = 5$,

$F_N = 0.20 \text{ ml min}^{-1}$ and $F_{Glc} = 0.42 \text{ ml min}^{-1}$, respectively, lines show differences of 0.05; **d–f** 3D representation of RS curves. The individual quadratic RS equations are shown in ESM Table S1

were recommended (see Sect. 2.8) from the toolbox. To statistically validate the recommended factor settings, four experiments located in the high D_i region were chosen, which are listed in Table 4.

Performed experiments

The four recommended experiments were performed (see Section “Recommended experiments from mDoE-toolbox”). The comparison of the experimental data to the model simulations, including the parametric 10% and 90% uncertainty-based prediction bands, is depicted in Figure 6.

Biomass In all four cultivations, cells grew until maximum biomass densities between $c_{DCW} = 45 \text{ g l}^{-1}$ (Fig. 6d) and $c_{DCW} = 80 \text{ g l}^{-1}$ (Fig. 6a). Growth was predicted sufficiently for all cultivations and only partly underestimated between $t = 18$ and 24 h in cultivation #4 (Fig. 6c). The width of the uncertainty band (10% and 90% quantiles) was relatively narrow for cultivation #1–#3 (Fig. 6a–c), which reflects that the variability of these experimental settings are predicted to be low. Broad uncertainty bands indicating a high variability of the experimental settings were predicted for cultivation #4 (Fig. 6d), due to high c_{EtOH} near inhibitory conditions.

Glucose and ethanol Glucose was consumed during the cultivations and was constantly very low (i.e., fully consumed) after $t \approx 20$ h. In cultivations #1 and #2, c_{EtOH}

increased only at the end of the cultivation in relatively low amounts below $c_{EtOH} = 12 \text{ g l}^{-1}$, for which no growth inhibition was seen. In cultures #3 and #4 (Fig. 6g, h), ethanol was not consumed and produced up to a maximum inhibitory concentration of $c_{EtOH} = 45.5 \pm 1.8 \text{ g l}^{-1}$ in cultivation #4 (Fig. 6h) [59, 61].

Overall, $c_{DCW} = 80 \text{ g l}^{-1}$ was achieved (cultivation #1) after the application of the mDoE-toolbox in a study with three influencing factors (F_{Glc} , F_N , pH). This reflects an improvement of $\approx 30\%$ in relation to the cultivation II (Fig. 2 b) with similar initial conditions. Simultaneously, c_{EtOH} at $t = 48$ h could be reduced by 50% compared to cultivation II (Fig. 2 e), resulting in higher substrate usage and a safer process operation point due to less possibility of inhibition.

The investigated factors are difficult to assess in traditional DoE studies due to their dynamic nature, i.e., the feeding rate itself changes during the process. A model-based approach strongly supports the evaluation of such dynamically changing factors.

Optimization of biocatalytic conversion of EAA to E3HB (S2)

In case study S2, the feeding rates F_{EAA} , F_{Glc} , and F_N were manipulated to optimize the biocatalytic conversion of EAA to E3HB. The optimization objective is defined as

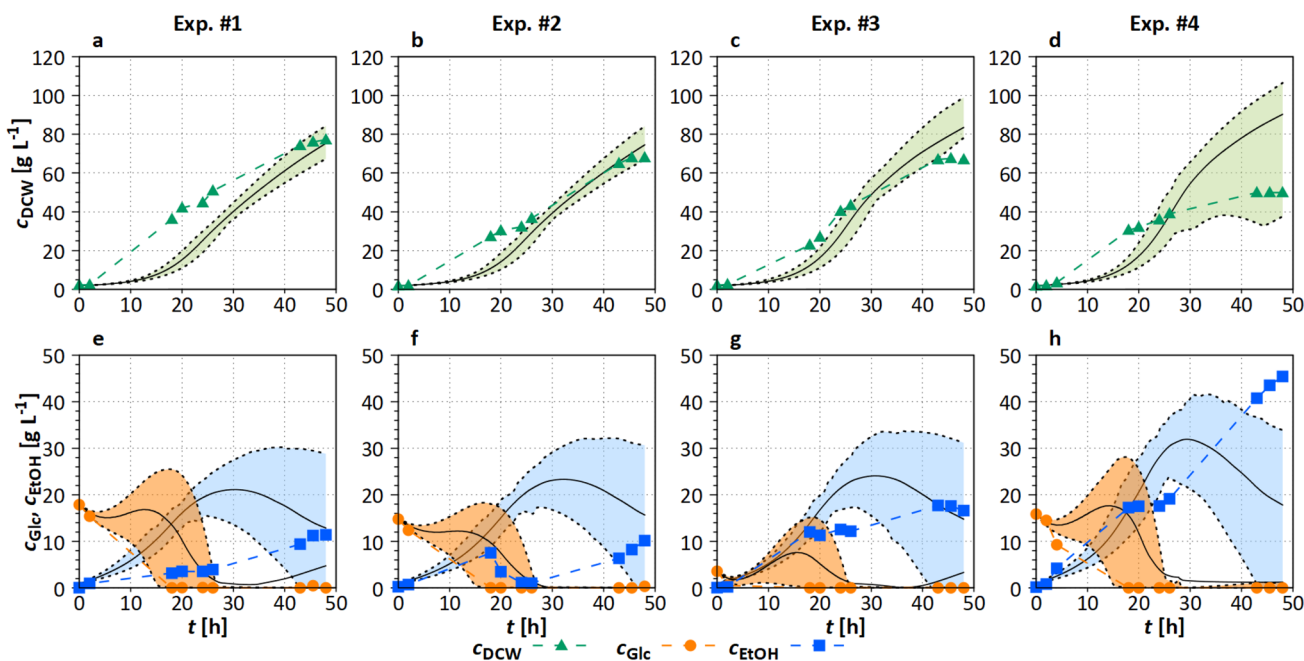


Fig. 6 Experimental data of the four performed cultivations compared to the simulated data from the mDoE-toolbox for c_{DCW} , c_{Glc} , and c_{EtOH} . The solid line is the mean of 30 simulations (Section “Monte Carlo-based simulation of planned experiments”);

dashed line represents the 10% and 90% quantiles of the simulations, and online off-gas data and individual feeding profiles can be found in ESM Fig. S10. Experimental settings and the used reactor are specified in Section “Recommended experiments from mDoE-toolbox”

the maximization of the biocatalytic product concentration (c_{E3HB}). Among the other factors, F_{EAA} is very critical, because c_{EAA} has to be kept below 0.5 g l^{-1} to avoid inhibition [53].

mDoE-toolbox (S2)

Planning of experimental design and Monte Carlo simulations The boundaries for F_{EAA} were $0 \leq F_{EAA} \leq 0.04 \text{ ml min}^{-1}$, based on literature to avoid inhibition through high EAA concentrations [53, 54]. F_{Glc} and F_N were defined to meet the demands the maintenance metabolism and the amount required for biocatalysis to be $0 \leq F_{Glc} \leq 0.2 \text{ ml min}^{-1}$ and $0 \leq F_N \leq 0.03 \text{ ml min}^{-1}$. The constitution of the feeds can be found in Table 2. For each simulated and performed biocatalysis, $c_{DCW} = 30 \text{ g l}^{-1}$ was directly inoculated with no prior cultivation. The same design of the experiments as in S1 was applied (see Fig. 1 box III) and 29 experiments were planned initially using the mDoE-toolbox. For each planned experiment, Monte Carlo simulations were performed (see Fig. 1, box IV).

Computational evaluation of experimental design

D_i was calculated and the response surface was predicted (Fig. 1 box V). As can be seen in Fig. 7, a maximum was determined at $F_{EAA} = 0.02 \text{ ml min}^{-1}$, $F_{Glc} = 0.11 \text{ ml min}^{-1}$, and $F_N = 0.01 \text{ ml min}^{-1}$.

Excessive EAA feeding is predicted to inhibit and deactivate biocatalysis. A high D_i is achieved only with a low F_{EAA} , as can be seen from the shape of the 3D plots (Fig. 7d–f). Too small values of F_{Glc} lead to low c_{Glc} limiting the biocatalysis in the simulations. Too high F_{Glc} resulting in an increased c_{Glc} , which leads to ethanol formation due to the Crabtree effect. In addition, high feeding rates (F_{Glc} and F_N) always result in dilution. The impact of F_N on achieving high c_{E3HB} is rather low. Therefore, a small F_N is desirable.

mDoE-suggested experimental settings Based on the Monte Carlo simulations and the computational evaluation of the experimental space, experiments located in the high D_i regions were identified and four of them were chosen. The experimental settings are listed in Table 5.

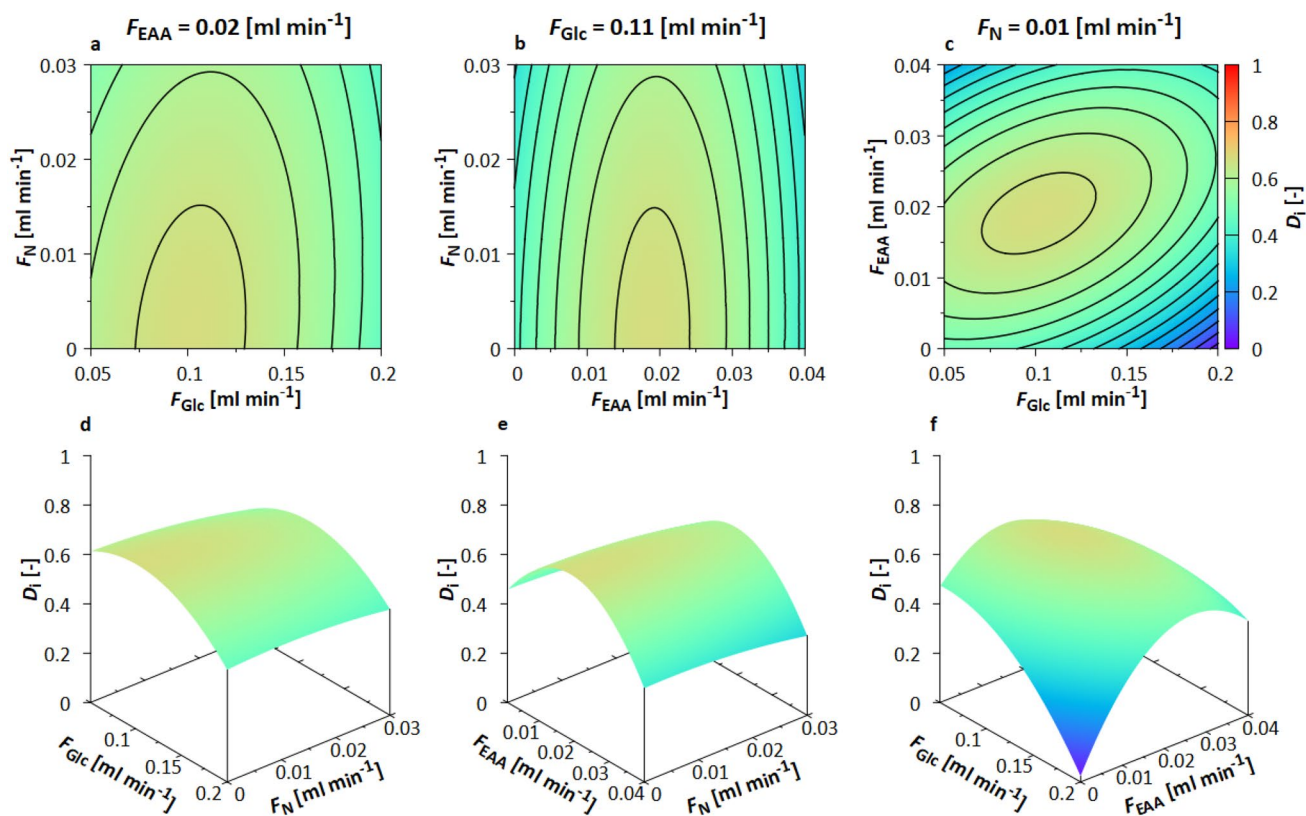


Fig. 7 Contour and 3D plot for the response surfaces based on D_i for F_{EAA} , F_{Glc} , and F_N in case study S2. The responses were calculated with Monte Carlo simulations in the mDoE-toolbox, as explained in Section 1.1. **a–c** Graphs are adjusted to $F_{EAA} = 0.02 \text{ ml min}^{-1}$,

$F_{Glc} = 0.11 \text{ ml min}^{-1}$ and $F_N = 0.01 \text{ ml min}^{-1}$ respectively, lines show differences of 0.05; **d–f** 3D representation of RS curves. The individual quadratic RS equations are shown in ESM Table S1

Table 5 Factor combination of the four recommended biocatalytic experiments determined with the mDoE-toolbox

Exp.	F_{EAA} (ml min ⁻¹)	F_{Glc} (ml min ⁻¹)	F_N (ml min ⁻¹)
#1	0.024	0.09	0.010
#2	0.018	0.12	0.010
#3	0.014	0.06	0.020
#4	0.010	0.09	0.002

Performed experiments

As can be seen in Fig. 8, final E3HB concentrations between $c_{E3HB} = 10 - 50 \text{ g l}^{-1}$ were reached in the four recommended experiments.

EAA and E3HB Even if the variation in F_{EAA} in the four experiments was relatively low ($0.10 \leq F_{EAA} \leq 0.24 \text{ ml min}^{-1}$), the maximum c_{E3HB} decreased strongly with decreasing F_{EAA} from cultivations #1-#4. EAA was constantly consumed during the bioprocesses and reached a maximum of $c_{EAA} = 2 \text{ g l}^{-1}$ in experiment #1 (Fig. 8a). The width of the 10% and 90% quantiles of the c_{EAA} simulations increases for higher F_{EAA} , due to an increasing probability of EAA inhibition.

In cultivations #3 and #4 (Fig. 8c, d), c_{E3HB} is lower than the prediction. Since no EAA was detectable, a higher by-product formation rate might have occurred in the experiments with lower product concentrations (Fig. 8c, d).

Biomass, glucose, and ethanol Despite a low F_N , biomass density increased by at least 15 g l^{-1} in every experiment (Fig. 8e-h). This was consistent with c_{Glc} and c_{EtOH} , which were below 5 g l^{-1} for cultivations #2-#4. In cultivation #1, increasing c_{EAA} leads to inhibition and reduction of the metabolic activity, resulting in a lower glucose consumption and an increasing glucose concentration up to $c_{Glc} = 1 \text{ g l}^{-1}$. This induces ethanol formation reaching $c_{EtOH} = 10 \text{ g l}^{-1}$ ($t = 48 \text{ h}$, Fig. 8e).

It could be shown that even the feeding strategy for a biocatalytic process with complex reaction mechanisms could be designed and optimized with the application of the mDoE-toolbox. By calculating an optimal feeding profile for EAA, glucose, and nitrogen source, $c_{E3HB} = 44 \text{ g l}^{-1}$ (Exp. #1) was achieved. This is an improvement of 80% in comparison to the experiment Biocatalysis I (Fig. 4a), with similar initial conditions. The same E3HB concentration was reached with a 60% lower initial c_{DCW} , when compared to the high cell density experiment (Biocatalysis II, Fig. 4b) and 10% compared to the literature [53, 54]. Furthermore, the application of the mDoE-toolbox reduced the initial experimental space by over 90% (compare Table 4 with Fig. 5). Only four recommended experiments had to be performed to find improved operating conditions, resulting in 80% higher product concentrations. The reactor volume was 80% smaller compared to Biocatalysis I, which may have had an additional influence on the improved product concentration. However, the transferability of process understanding obtained using

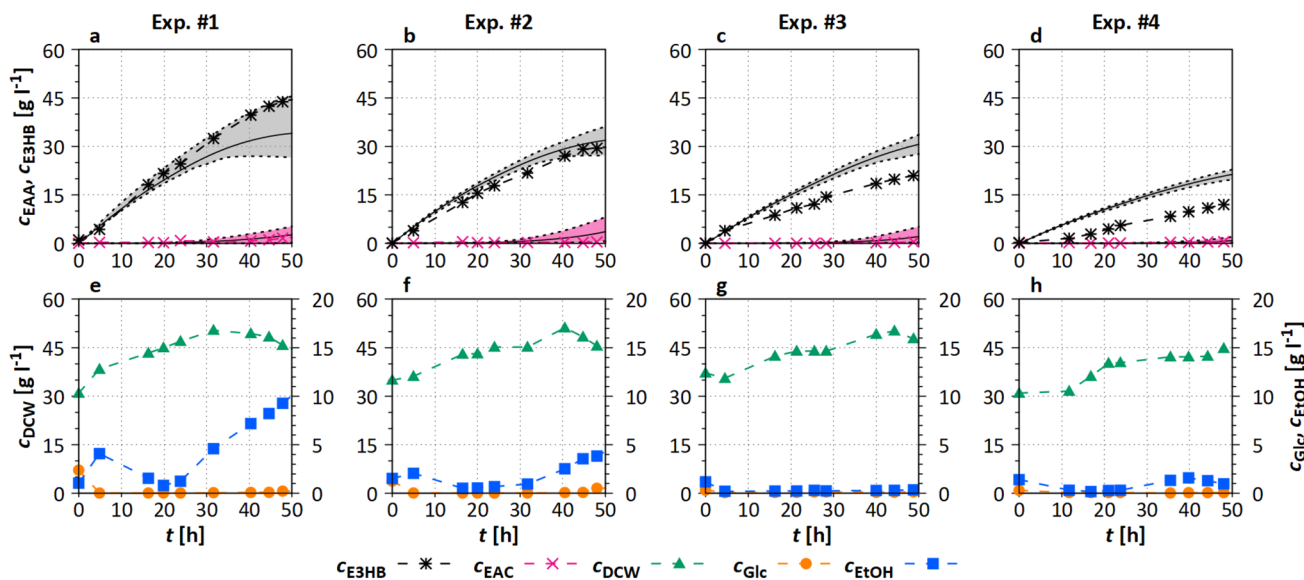


Fig. 8 Experimental data of the four performed cultivations compared to the simulated data from the mDoE-toolbox for c_{EAA} , c_{E3HB} , and c_{Glc} . The solid line is the mean of 30 simulations (Section “Monte Carlo-based simulation of planned experiments”);

dashed line represents the 10% and 90% quantiles of the simulations. Experimental settings and the used reactor are specified in Section “Recommended experiments from mDoE-toolbox”

mathematical process models was recently shown to be transferable between different scales [19, 37].

Conclusion

In this study, the mDoE-toolbox was introduced to enable a more knowledge-driven experimental design, and to strongly reduce the number of experiments during bioprocess development and optimization. The application of the toolbox was shown for two different case studies with *Saccharomyces cerevisiae*. In case study S1, a fed-batch process was optimized to maximize the final biomass density depending on the factors pH, and linearly rising substrate feeding rates. Just four experiments were needed to achieve a 30% increase of the final biomass density compared to Cultivation II, instead of 29 initially planned experiments, which would have been performed in the fully experimental evaluation of the DoE. In case study S2, the biocatalytic production of E3HB was optimized based on constant substrate feeding rates. Just four experiments were needed instead of 29 initially planned experiments. An improvement of 80% in the final E3HB concentration was experimentally achieved compared to the experiment Biocatalysis I with similar initial conditions. Although this reaction is well known, an improvement of about 10% of c_{E3HB} was achieved compared to the literature [53, 54]. In addition, this result was obtained in less than half the cultivation time [54]. In both processes, the optimization of rather difficult to assess factors, such as timely changing feeding rates (S1) or feeding of inhibitory components (S2), was possible through modeling.

In summary, the usage of the mDoE-toolbox enables optimization studies with dynamic factors in statistics-based biotechnology research employing a reduced number of experiments. Further research will focus on online model parameter adaptation and the consequent online re-design of experiments to increase the obtained process understanding further.

Acknowledgements The authors would like to thank V. Gockel, M. Meissner, and M. Menden for their laboratory work at Furtwangen University. Krathika Bhat is kindly acknowledged for English proof reading. The mDoE-toolbox can be made available for scientific collaborations.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors acknowledge partial funding by German Federal Ministry of Education and Research (BMBF, Grant 031B0577A-C).

Compliance with ethical standards

Conflict of interest The authors declare that there are no conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Möller J, Kuchemüller KB, Steinmetz T, Koopmann KS, Pörtner R (2019) Model-assisted design of experiments as a concept for knowledge-based bioprocess development. *Bioproc Biosyst Eng* 42(5):867. <https://doi.org/10.1007/s00449-019-02089-7>
- Dubois JL (2011) Requirements for the development of a bioeconomy for chemicals. *Curr Opin Environ Sustain* 3(1):11. <https://doi.org/10.1016/j.cosust.2011.02.001>
- Lokko Y, Heijde M, Schebesta K, Scholtès P, Van Montagu M, Giacca M (2018) Biotechnology and the bioeconomy towards inclusive and sustainable industrial development. *New Biotechnol*. <https://doi.org/10.1016/j.nbt.2017.06.005>
- Scarlat N, Dallemand JF, Monforti-Ferrario F, Nita V (2015) The role of biomass and bioenergy in a future bioeconomy: policies and facts. *Environ Dev* 15:3. <https://doi.org/10.1016/j.envdev.2015.03.006>
- Guo M, Song W (2019) The growing U.S. bioeconomy: drivers, development and constraints. *New Biotechnol* 49:48. <https://doi.org/10.1016/j.nbt.2018.08.005>
- Mandenius CF, Brundin A (2008) Bioprocess optimization using design-of-experiments methodology. *Biotechnol Progr* 24(6):1191. <https://doi.org/10.1002/btpr.67>
- Glauche F, Pilarek M, Bournazou MNC, Grunzel P, Neubauer P (2017) Design of experiments-based high-throughput strategy for development and optimization of efficient cell disruption protocols. *Eng Life Sci* 17(11):1166. <https://doi.org/10.1002/elsc.20160030>
- Vervoort Y, Linares AG, Roncoroni M, Liu C, Steensels J, Verstrepen KJ (2017) High-throughput system-wide engineering and screening for microbial biotechnology. *Curr Opin Biotech* 46:120. <https://doi.org/10.1016/j.copbio.2017.02.011>
- Dörr M, Fibinger MP, Last D, Schmidt S, Santos-Aberturas J, Böttcher D, Hummel A, Vickers C, Voss M, Bornscheuer UT (2016) Fully automatized high-throughput enzyme library screening using a robotic platform. *Biotechnol Bioeng* 113(7):1421. <https://doi.org/10.1002/bit.25925>
- Bareither R, Bargh N, Oakeshott R, Watts K, Pollard D (2013) Automated disposable small scale reactor for high throughput bioprocess development: a proof of concept study. *Biotechnol Bioeng* 110(12):3126. <https://doi.org/10.1002/bit.24978>
- Savizi ISP, Soudi T, Shojaosadati SA (2019) Systems biology approach in the formulation of chemically defined media for recombinant protein overproduction. *Appl Microbiol Biotechnol* 103(20):8315. <https://doi.org/10.1007/s00253-019-10048-1>
- Abt V, Barz T, Cruz-Bournazou MN, Herwig C, Kroll P, Möller J, Pörtner R, Schenkendorf R (2018) Model-based tools for optimal experiments in bioprocess engineering. *Curr Opin Chem Eng* 22:244. <https://doi.org/10.1016/j.coche.2018.11.007>

13. Mandenius CF, Graumann K, Schultz TW, Premstaller A, Olsson IM, Petiot E, Clemens C, Welin M (2009) Quality-by-design for biotechnology-related pharmaceuticals. *Biotechnol J* 4(5):600. <https://doi.org/10.1002/biot.200800333>
14. Kuchemüller KB, Pörtner R, Möller J (2020a) Efficient optimization of process strategies with model-assisted design of experiments. Springer US, New York, pp 235–249. https://doi.org/10.1007/978-1-0716-0191-4_13
15. Brunner M, Fricke J, Kroll P, Herwig C (2017) Investigation of the interactions of critical scale-up parameters (ph, po₂ and pco₂) on cho batch performance and critical quality attributes. *Bioproc Biosyst Eng* 40(2):251. <https://doi.org/10.1007/s00449-016-1693-7>
16. von Stosch M, Willis MJ (2017) Intensified design of experiments for upstream bioreactors. *Eng Life Sci* 17(11):1173. <https://doi.org/10.1002/elsc.201600037>
17. Manzon D, Claeys-Bruno M, Declomesnil S, Carité C, Sergent M (2020) Quality by design: comparison of design space construction methods in the case of design of experiments. *Chemometr Intell Lab* 200:104002. <https://doi.org/10.1016/j.chemo lab.2020.104002>
18. von Stosch M, Hamelink JM, Oliveira R (2016) Hybrid modeling as a qbd/pat tool in process development: an industrial *E. coli* case study. *Bioproc Biosyst Eng* 39(5):773. <https://doi.org/10.1007/s00449-016-1557-1>
19. Möller J, Hernández Rodríguez T, Müller J, Arndt L, Kuchemüller KB, Frahm B, Eibl R, Eibl D, Pörtner R (2020) Model uncertainty-based evaluation of process strategies during scale-up of biopharmaceutical processes. *Comput Chem Eng* 134:106693. <https://doi.org/10.1016/j.compchemeng.2019.106693>
20. Carrondo MJT, Alves PM, Carinhas N, Glassey J, Hesse F, Merten OW, Micheletti M, Noll T, Oliveira R, Reichl U, Staby A, Teixeira AP, Weichert H, Mandenius CF (2012) How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology? *Biotechnol J* 7(12):1522. <https://doi.org/10.1002/biot.201200226>
21. Kontoravdi C, Samsatli NJ, Shah N (2013) Development and design of bio-pharmaceutical processes. *Curr Opin Chem Eng* 2(4):435. <https://doi.org/10.1016/j.coche.2013.09.007>
22. Jarka GVGK, Christoph CWST, Rui O, Gerald S, Carl-Fredrik M (2011) Process analytical technology (pat) for biopharmaceuticals. *Biotechnol J* 6(4):369. <https://doi.org/10.1002/biot.201000356>
23. Bailey JE (1998) Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnol Progr* 14(1):8. <https://doi.org/10.1021/bp9701269>
24. Möller J, Korte K, Pörtner R, Zeng AP, Jandt U (2018) Model-based identification of cell-cycle-dependent metabolism and putative autocrine effects in antibody producing cho cell culture. *Biotechnol Bioeng* 115(12):2996. <https://doi.org/10.1002/bit.26828>
25. Kroll P, Hofer A, Ulonska S, Kager J, Herwig C (2017) Model-based methods in the biopharmaceutical process lifecycle. *Pharm Res* 34(12):2596. <https://doi.org/10.1007/s11095-017-2308-y>
26. 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(4):1019. <https://doi.org/10.1007/s10616-015-9858-9>
27. Brüning S, Gerlach I, Pörtner R, Mandenius CF, Hass VC (2017) Modeling suspension cultures of microbial and mammalian cells with an adaptable six-compartment model. *Chem Eng Technol* 40(5):956. <https://doi.org/10.1002/ceat.201600639>
28. Möller J, Bhat K, Riecken K, Pörtner R, Zeng AP, Jandt U (2019) Process-induced cell cycle oscillations in cho cultures: online monitoring and model-based investigation. *Biotechnol Bioeng* 116(11):2931. <https://doi.org/10.1002/bit.27124>
29. Jin Z, Han SY, Zhang L, Zheng SP, Wang Y, Lin Y (2013) Combined utilization of lipase-displaying *Pichia pastoris* whole-cell biocatalysts to improve biodiesel production in co-solvent media. *Bioproc Biosyst Eng* 36(10):1102. <https://doi.org/10.1016/j.biortech.2012.12.020>
30. Isidro IA, Portela RM, Clemente JJ, Cunha AE, Oliveira R (2016) Hybrid metabolic flux analysis and recombinant protein prediction in *Pichia pastoris* x-33 cultures expressing a single-chain antibody fragment. *Bioproc Biosyst Eng* 39(9):1351. <https://doi.org/10.1007/s00449-016-1611-z>
31. Mondal NK, Samanta A, Dutta S, Chatteraj S (2017) Optimization of Cr(VI) biosorption onto *Aspergillus niger* using 3-level box-behnken design: Equilibrium, kinetic, thermodynamic and regeneration studies. *J Genet Eng Biotechnol* 15(1):151. <https://doi.org/10.1016/j.jgeb.2017.01.006>
32. Abdel-Fattah YR (2002) Optimization of thermostable lipase production from a thermophilic *Geobacillus* sp. using box-behnken experimental design. *Biotechnol Lett* 24(14):1217. <https://doi.org/10.1023/A:1016167416712>
33. Candiotti LV, Zan MMD, Camara MS, Goicoechea HC (2014) Experimental design and multiple response optimization. using the desirability function in analytical methods development. *Talanta* 124:123. <https://doi.org/10.1016/j.talanta.2014.01.034>
34. Derringer G, Suich R (1980) Simultaneous optimization of several response variables. *J Qual Technol* 12(4):214. <https://doi.org/10.1080/00224065.1980.11980968>
35. Narayanan H, Luna MF, von Stosch M, Cruz Bournazou MN, Polotti G, Morbidelli M, Butte A, Sokolov M (2020) Bioprocessing in the digital age: the role of process models. *Biotechnol J* 15(1):1900172. <https://doi.org/10.1002/biot.20190172>
36. Herwig C (2020) Applied basic science in process analytics and control technology. *Anal Bioanal Chem* 412(9):2025. <https://doi.org/10.1007/s00216-020-02465-3>
37. Kuchemüller KB, Pörtner R, Möller J (2020) Digital twins and their role in model-assisted design of experiments. Springer US, New York
38. Zeng AP, Byun TG, Posten C, Deckwer WD (1994) Use of respiratory quotient as a control parameter for optimum oxygen supply and scale-up of 2,3-butanediol production under microaerobic conditions. *Biotechnol Bioeng* 44(9):1107. <https://doi.org/10.1002/bit.260440912>
39. Heyman B, Tulke H, Putri SP, Fukusaki E, Büchs J (2020) Online monitoring of the respiratory quotient reveals metabolic phases during microaerobic 2,3-butanediol production with *Bacillus licheniformis*. *Eng Life Sci* 20(3–4):133. <https://doi.org/10.1002/elsc.201900121>
40. Moser A, Appl C, Brüning S, Hass VC (2020) Mechanistic mathematical models as a basis for digital twins. Springer US, New York. https://doi.org/10.1007/10_2020_152
41. Gerlach I, Brüning S, Gustavsson R, Mandenius CF, Hass VC (2014) Operator training in recombinant protein production using a structured simulator model. *J Biotechnol* 177:53. <https://doi.org/10.1016/j.jbiotec.2014.02.022>
42. Hass VC (2005) Verbesserung der bioverfahrenstechnischen ausbildung durch einen virtuellen bioreaktor. *Chem-Ing-Tech* 77(1–2):161. <https://doi.org/10.1002/cite.200407053>
43. Gerlach I, Hass VC, Brüning S, Mandenius CF (2013) Virtual bioreactor cultivation for operator training and simulation: application to ethanol and protein production. *J Chem Technol Biotechnol* 88(12):2159. <https://doi.org/10.1002/jctb.4079>
44. Wechselberger P, Sagmeister P, Herwig C (2013) Model-based analysis on the extractability of information from data in dynamic fed-batch experiments. *Biotechnol Progr* 29(1):285. <https://doi.org/10.1002/btpr.1649>
45. 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 a mcmc method. *Biotechnol Bioeng* 116(11):2944. <https://doi.org/10.1002/bit.27125>
46. Nelder JA, Mead R (1965) A simplex method for function minimization. *Comput J* 7(4):308. <https://doi.org/10.1093/comjnl/7.4.308>
 47. Colin Cameron A, Windmeijer FA (1997) An r-squared measure of goodness of fit for some common nonlinear regression models. *J Econometr* 77(2):329. [https://doi.org/10.1016/S0304-4076\(96\)01818-0](https://doi.org/10.1016/S0304-4076(96)01818-0)
 48. D'haeseleer P (2005) How does gene expression clustering work? *Nat Biotechnol* 23(12):1499. <https://doi.org/10.1038/nbt1205-1499>
 49. Möller J, Rosenberg M, Riecken K, Pörtner R, Zeng AP, Jandt U (2020) Quantification of the dynamics of population heterogeneities in cho cultures with stably integrated fluorescent markers. *Anal Bioanal Chem* 412(9):2065. <https://doi.org/10.1007/s00216-020-02401-5>
 50. Lloyd S (1982) Least squares quantization in pcm. *IEEE T Inf Theory* 28(2):129. <https://doi.org/10.1109/TIT.1982.1056489>
 51. McKay MD, Beckman RJ, Conover WJ (1979) Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2):239. <https://doi.org/10.1080/00401706.1979.10489755>
 52. Gargalo CL, Cheali P, Posada JA, Carvalho A, Gernaey KV, Sin G (2016) Assessing the environmental sustainability of early stage design for bioprocesses under uncertainties: an analysis of glycerol bioconversion. *J Clean Prod* 139:1245. <https://doi.org/10.1016/j.jclepro.2016.08.156>
 53. Hirschmann R, Borodkin N, Baganz F, Hass V (2018) Towards the integration of the anaerobic ethyl (s)-3-hydroxybutyrate production process into a biorefinery concept. *Chem Eng Trans* 70:559. <https://doi.org/10.3303/CET1870094>
 54. Kometani T, Yoshii H, Kitatsuji E, Nishimura H, Matsuno R (1993) Large-scale preparation of (s)-ethyl 3-hydroxybutanoate with a high enantiomeric excess through baker's yeast-mediated bioreduction. *J Ferment Bioeng* 76(1):33. [https://doi.org/10.1016/0922-338X\(93\)90049-E](https://doi.org/10.1016/0922-338X(93)90049-E)
 55. Lin Y, Tanaka S (2006) Ethanol fermentation from biomass resources: current state and prospects. *Appl Microbiol Biotechnol* 69(6):627. <https://doi.org/10.1007/s00253-005-0229-x>
 56. Bai F, Anderson W, Moo-Young M (2008) Ethanol fermentation technologies from sugar and starch feedstocks. *Biotechnol Adv* 26(1):89. <https://doi.org/10.1016/j.biotechadv.2007.09.002>
 57. Sonnleitner B, Käppeli O (1986) Growth of *Saccharomyces cerevisiae* is controlled by its limited respiratory capacity: formulation and verification of a hypothesis. *Biotechnol Bioeng* 28(6):927. <https://doi.org/10.1002/bit.260280620>
 58. Fiechter A, Fuhrmann G, Käppeli O (1981) Regulation of glucose metabolism in growing yeast cells. *Adv Microbiol Physiol* 22:123. [https://doi.org/10.1016/S0065-2911\(08\)60327-6](https://doi.org/10.1016/S0065-2911(08)60327-6)
 59. Casey GP, Ingledew WM (1986) Ethanol tolerance in yeasts. *CRC Crit Rev Microbiol* 13(3):219. <https://doi.org/10.3109/10408418609108739>
 60. Arroyo-López FN, Orlic S, Querol A, Barrio E (2009) Effects of temperature, ph and sugar concentration on the growth parameters of *Saccharomyces cerevisiae* *S. kudriavzevii* and their interspecific hybrid. *Int J Food Microbiol* 131(2):120. <https://doi.org/10.1016/j.ijfoodmicro.2009.01.035>
 61. Larsson C, von Stockar U, Marison I, Gustafsson L (1993) Growth and metabolism of *Saccharomyces cerevisiae* in chemostat cultures under carbon-, nitrogen-, or carbon- and nitrogen-limiting conditions. *J Bacteriol* 175(15):4809. <https://doi.org/10.1128/jb.175.15.4809-4816.1993>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.