Computational and Structural Biotechnology Journal 23 (2024) 3651–3661

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Optimizing bioprocessing efficiency with OptFed: Dynamic nonlinear modeling improves product-to-biomass yield

Guido Schlögel [a,b], Rüdiger Lück [c], Stefan Kittler [c], Oliver Spadiut [c], Julian Kopp [c], Jürgen Zanghellini [a], Mathias Gotsmy [a,d,*]

[a] Department of Analytical Chemistry, University Vienna, Währinger Straße, 1090 Vienna, Austria
[b] Doctorate School of Chemistry, University of Vienna, Währinger Straße, 1090 Vienna, Austria
[c] Integrated Bioprocess Development, Technical University Vienna, Getreidemarkt 9, 1060 Vienna, Austria
[d] Austrian Centre of Industrial Biotechnology, Krenngasse 37, 8010 Graz, Austria

ARTICLE INFO

Dataset link: https://github.com/gschloegel/OptFed/

Keywords:
Fed-batch
Bioproduction
Non-linear optimization
Process design

ABSTRACT

Biotechnological production of recombinant molecules relies heavily on fed-batch processes. However, as the cells' growth, substrate uptake, and production kinetics are often unclear, the fed-batches are frequently operated under sub-optimal conditions. Process design is based on simple feed profiles (e.g., constant or exponential), operator experience, and basic statistical tools (e.g., response surface methodology), which are unable to harvest the full potential of production.

To address this challenge, we propose a general modeling framework, OptFed, which utilizes experimental data from non-optimal fed-batch processes to predict an optimal one. In detail, we assume that cell-specific rates depend on several state variables and their derivatives.

Using measurements of bioreactor volume, biomass, and product, we fit the kinetic constants of ordinary differential equations. A regression model avoids overfitting by reducing the number of parameters. Thereafter, OptFed predicts optimal process conditions by solving an optimal control problem using orthogonal collocation and nonlinear programming.

In a case study, we apply OptFed to a recombinant protein L fed-batch production process. We determine optimal controls for feed rate and reactor temperature to maximize the product-to-biomass yield and successfully validate our predictions experimentally. Notably, our framework outperforms RSM in both simulation and experiments, capturing an optimum previously missed. We improve the experimental product-to-biomass ratio by 19% and showcase OptFed's potential for enhancing process optimization in biotechnology.

1. Introduction

Biotechnological production processes are the backbone of numerous industries, from pharmaceuticals to biofuels. Many of these processes are operated as a fed-batch, adding substrate and nutrients continuously when the initial batch medium is depleted [1]. This method allows for control of key parameters, such as nutrient concentration, and is fundamental to achieving high yields and product quality. Consequently, the optimization of such processes becomes a critical objective.

Optimizing fed-batch processes is challenging due to the complexity of cellular mechanisms, which are difficult to measure directly and can vary significantly based on factors such as product type, microorganism, induction mechanism, and product location. As a result, optimization often involves a trial-and-error approach [2]. In this context, theoretical and mathematical modeling offers a powerful and complementary alternative. By leveraging simulations and model-based design of experiments [3,4], we can reduce our dependence on costly and time-consuming trials, and enhance our ability to predict and optimize process performance in a more controlled and efficient manner.

In general, there are two distinct paths for optimizing biotechnological processes: statistical design of experiments, such as response surface methodology (RSM) [5], and the model-based approach [3,4,6,7]. Statistical methods, including RSM, offer straightforward and accessible means of optimization. However, they do not leverage biological knowledge, which could enhance the optimization process [8], and are limited to optimizing a predefined set of discrete variables [9]. In contrast, the

* Corresponding author.
E-mail address: mathias.gotsmy@univie.ac.at (M. Gotsmy).

https://doi.org/10.1016/j.csbj.2024.09.024
Received 30 July 2024; Received in revised form 27 September 2024; Accepted 29 September 2024

Available online 11 October 2024

model-based approach begins by understanding the underlying process, representing it accurately without relying on mathematical assumptions like the quadratic dependencies used in purely statistical methods [10].

For the model-based approach, empirical models are typically employed, often formulated as ordinary differential equations (ODEs). For instance, Monod's widely used model for population growth [11] is one such example. Product creation in these models is often linked to growth rate [4,12,13] or feed rate [14]. These models provide a simple yet effective means of understanding and predicting bioprocess behavior, particularly when dealing with the often limited and noisy data typically obtained from bioprocesses.

However, simplicity comes at a cost. The inherent simplifications in these empirical models constrain their applicability, limiting their ability to describe complex processes comprehensively. These straightforward models often fail to account for some phenomena observed under constant process conditions on a cellular level. This includes production deterioration over time due to metabolic adaptation, or product inhibition [15]. Despite their limitations, these models are frequently used because formulating more complex models is hindered by sparse data, biological variation, and the difficulties in selecting the best available model equations [16]. Different metrics can be used for model selection, e.g., AIC and AIC$_\text{C}$ [17], LASSO [18], and methods based on cross-validation [19,20]. Regardless of the used metric, the number of possible models can get very large (e.g., over $2^{18} \approx 2.6 \times 10^5$ possible models with 18 parameters) which requires defined search strategies (e.g., HIPPO) [21,22].

Moreover, decisions in the modeling process can introduce biases that influence the outcome [23,24], and using more complex models with small datasets may lead to overfitting [25].

The modeling process does not exist in isolation; its primary goal is to improve the efficiency of the process (e.g., by maximizing titer/biomass or minimizing operational costs). This objective often involves defining specific targets for each process, frequently utilizing the TRY metric (titer, rate, yield) [26]. The choice of optimization algorithms varies depending on the complexity of the model at hand, which is influenced by the available data set and its quality. While straightforward maximization algorithms suffice for discrete variables [27], optimizing continuous solutions, such as feed and temperature functions, poses mathematical challenges, notably due to the theoretically infinite number of control variables [28]. To address these challenges, various mathematical methods have been developed. For simpler models, Euler-Lagrange equation-based approaches can be employed [4]. In contrast, more intricate models, as encountered in our work, require the application of optimal control theory [29].

The realm of optimal control problems has been extensively explored, resulting in a multitude of solution methods [30]. Analytical methods, such as those grounded in Pontryagin's maximum principle [29], are well-suited for relatively straightforward problems but may not be practical for complex real-life scenarios. In most cases, numerical solvers become essential. The two primary categories of solvers are direct methods and those based on dynamic programming. Direct methods tackle the optimization and differential equations simultaneously, transforming the problem into a set of nonlinear differential equations [31]. In contrast, dynamic programming [32,33] relies on the principle of optimality within Hamilton-Jacobi-Bellman frameworks, iteratively solving the problem. While dynamic programming holds the promise of identifying global optima, it tends to be slower than direct methods and is often infeasible for high-dimensional problems. As a result, direct methods find more frequent applications, particularly in engineering contexts, where a wealth of software packages is available for implementation [34–37].

In this study, we present OptFed, a comprehensive framework using an ODE model to describe bioprocesses. The framework is divided into three stages: define, fit, and optimize. First, we define a general and flexible form of the ODE model. Next, its kinetic parameters are fitted to training data and the model size is reduced to avoid overfitting. To do this, we developed a heuristic algorithm that starts with the general model and removes terms and parameters that do not significantly improve the fit. In the third stage, based on the reduced model, we leverage optimal control theory to identify optimal values for control variables. In a case study, we apply OptFed to protein L production to maximize the product-to-biomass yield. Optimal values for the temperature and substrate feed control are predicted. A comparison to RSM highlights the improvements of OptFed to typically used statistical methods. Moreover, experimental validation results in a 19% improved product-to-biomass yield.

2. Methods

2.1. Modeling framework

Our modeling framework comprises three key components:

(I) define,
(II) fit, and
(III) optimize.

In the first stage, we establish a general process model capable of representing a wide range of biotechnological (fed-)batch production processes. In the second stage, the general model is fitted to specific process data, the kinetic parameters are estimated and insignificant terms are removed. In the third stage, the fitted process model is used to optimize control variables, ultimately maximizing a freely selectable objective function. A graphical overview of the modeling framework is given in Fig. 1.

A list of all parameters and their respective symbols and units is given in Supplementary Table S1.

2.1.1. Stage I: define

We consider the production of recombinant proteins by *E. coli* in a fed-batch process, described by the following standard system of differential equations [1]:

$$\dot{X} = \mu X - \frac{f}{V} X, \qquad X(0) = X_0, \tag{1a}$$

$$\dot{P} = \pi X - \frac{f}{V} P, \qquad P(0) = 0, \tag{1b}$$

$$\dot{G} = -\gamma X + \frac{f}{V}(G_\text{f} - G), \qquad G(0) = 0 \tag{1c}$$

$$\dot{V} = f, \qquad V(0) = V_0, \tag{1d}$$

where G, G_f, P, and X represent substrate concentrations in the reactor, substrate concentration in the feed, product, and total biomass, respectively, and V the current reactor volume. f, γ, μ, and π denote the feeding rate, substrate uptake rate per biomass, biomass growth rate, and specific product formation rate, respectively.

As the product is part of the biomass, we additionally define the metabolic active residual biomass (given as dry weight)

$$X_\text{r} = X - P. \tag{1e}$$

This defines the uptake per active residual biomass (γ°) as

$$\gamma^\circ = \gamma X / X_\text{r}. \tag{1f}$$

To connect the substrate uptake behavior with cellular growth and production, we assume that the total uptake can be divided into three additive components

$$\gamma^\circ = \gamma^\mu + \gamma^\pi + \gamma^\alpha, \tag{2}$$

where γ^μ, γ^π, and γ^α denote the specific substrate uptake supporting growth, product formation, and cellular maintenance, respectively. Here, maintenance summarizes all cellular processes not linked to growth or production.

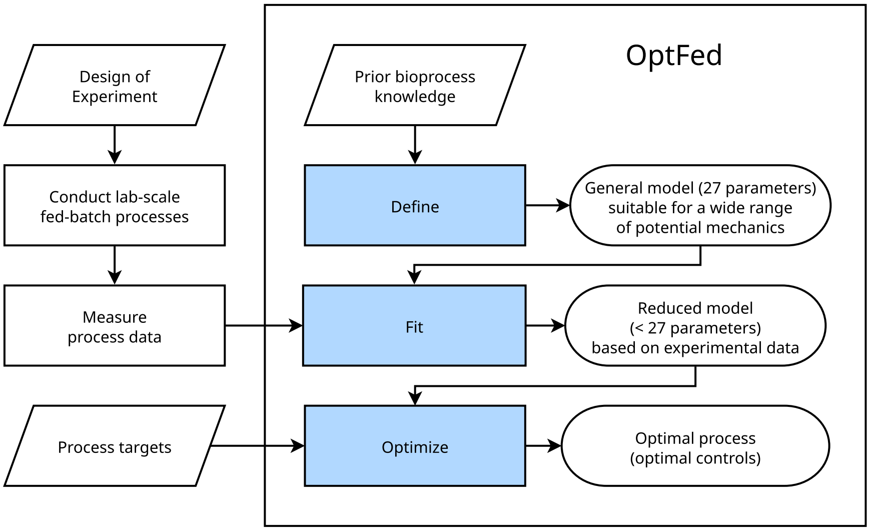


Fig. 1. Flow chart for OptFed. In this study, we focus on the stages highlighted in blue. Initially, we set up a general model including different inhibition effects (define). This model can describe different processes and uses many model parameters. In the second stage, the model is simplified to avoid overfitting and parameters are estimated (fit). In the third stage, optimal process control variables are predicted (optimize).

Using biomass-to-substrate yield $Y_{\frac{X_r}{G}}$ and product-to-substrate $Y_{\frac{P}{G}}$, we can connect glucose consumption to product formation and growth as follows:

$$\pi = \gamma^\pi Y_{\frac{P}{G}} X_r/X, \quad \text{and} \quad \mu = \pi + \gamma^\mu Y_{\frac{X_r}{G}} X_r/X. \tag{3}$$

The factor X_r/X accounts for the fact that only the metabolically active residual biomass X_r contributes to additional product formation and growth.

We assume that total substrate uptake as well as substrate uptake for product formation follow a non-competitively inhibited Michaelis–Menten process [14,38]

$$\gamma^\circ = \gamma^\circ_{\max}(T)\frac{G}{K^\circ_m + G}\prod_{i\in var_1}\frac{1}{1+i/K^\circ_i}, \quad var_1 = \{G, n, P/X, X\}, \tag{4a}$$

$$\gamma^\pi = \gamma^\pi_{\max}(T)\frac{\gamma^\circ - \gamma^\alpha}{K^\pi_m + \gamma^\circ - \gamma^\alpha}\prod_{i\in var_1}\frac{1}{1+i/K^\pi_i}, \tag{4b}$$

while substrate demand for maintenance is given by [39]

$$\gamma^\alpha = \gamma^\alpha_{\min}(T)\prod_{i\in var_2}(1+i/K^\alpha_i), \quad var_2 = \{\gamma^\circ, G, n, P/X, X\}. \tag{4c}$$

With these assumptions, γ^μ follows according to (2) to

$$\gamma^\mu = \gamma^\circ - \gamma^\pi - \gamma^\alpha. \tag{4d}$$

Here, K°_m, and K^π_m are Michaelis-Menten constants, while K°_i and K^π_i are inhibition constants, and K^α_i are activation constants, n is the number of generations ($n = \log_2[XV/(X_0V_0)]$). Finally, the minimum uptake rate $\gamma^\alpha_{\min}$, and the maximum uptake rates $\gamma^\circ_{\max}$ and $\gamma^\pi_{\max}$ are assumed to be temperature (T) dependent (excluding enzyme denaturation) [40]

$$\gamma^i_j(T) = E^i_0\frac{k_B T}{h}\frac{\exp\left(-\frac{\Delta G^i_{\text{cat}}}{RT}\right)}{1+\exp\left[\frac{\Delta H^i_{\text{eq}}}{R}\left(\frac{1}{T^i_{\text{eq}}}-\frac{1}{T}\right)\right]}, \quad \begin{array}{l} i\in\{\mu,\pi,\circ\}, \\ j\in\{\max,\min\}, \end{array} \tag{4e}$$

where E^i_0 is a (hypothetical) enzyme concentration, ΔG^i_{cat}, the activation energy, T^i_{eq} is the temperature where half of the enzymes are in an active state (the other half is inactivated by the high temperature), and ΔH^i_{eq} determines how abruptly the reaction rate declines with rising temperatures. The superscripts $\circ, \alpha$, and π differentiate variables for substrate uptake, maintenance, and production rate, respectively, differentiating the constants for γ°, γ^α and γ^π. Our model, defined in equations (1) and (4), contains 29 free parameters. The substrate yields $Y_{\frac{X_r}{G}}$, and $Y_{\frac{P}{G}}$ can be derived from genome-scale metabolic models [41], while the remaining 27 parameters need to be fitted from training data.

Bioreactor volume estimation Generally, change in the bioreactor volume is affected by five factors, the substrate feed, the base feed (for pH control), experimental sampling, the antifoam feed, and gaseous exchanges,

$$f = f_{\text{substrate}} + f_{\text{base}} + f_{\text{sampling}} + f_{\text{antifoam}} + f_{\text{gaseous}}. \tag{5}$$

In OptFed, we explicitly model the first three of them. While the substrate feed is kept variable (for optimization), the base feed is calculated as

$$f_{\text{base}} = XV(a\mu + b) \tag{6}$$

where the parameters a and b were fitted to the training data. Additionally, we accounted for volume change through experimental sampling in OptFed. For simplicity, and due to the fact that antifoam, feed, and gaseous exchanges (i.e., evaporation, O_2 uptake, and CO_2 excretion) are either minor or antagonistic contributors to volume change, we did not consider them in OptFed. More details on volume calculation are given in Supplementary Methods S1.1.

Substrate feed types Although OptFed is not restricted to a certain type of substrate feed, it can make sense to enforce them either for comparison to the training data or for simplification of experimental implementation. Here we present two feed types that we refer to throughout the manuscript.

An exponential feed is generally very popular as the cells' internal metabolic fluxes are (approximately) constant. It is calculated as

$$f_{\text{substrate}}(t) = f_0\exp(\mu_f\, t) \tag{7a}$$

where μ_f of the unit h^{-1} is the defining parameter. f_0 is usually derived from the properties of the process at hand.

Additionally, we also use a linear feed rate, calculated as

$$f_{\text{substrate}}(t) = \phi_1 + \phi_2 t \tag{7b}$$

where ϕ_1 and ϕ_2 may be varied.

2.1.2. Stage II: fit

In this step, we use training data to select a model of suitable size and fit its kinetic parameters.

Training data OptFed requires process data for different feed rates and temperatures. Data from a central composite design [10] commonly used for RSM proofed useful (Section 2.4).

Experimental data interpolation and rate calculation The differential method [42,43] is used to estimate uptake, growth, and production rates by fitting the concentration data and differentiating the fits. We fit splines, which are continuous in both their values and first derivatives, using SciPy's `UnivariateSpline` function [44] onto the experimental data points for control and state variables. By inserting these splines into equations (1) and (2), we calculate the experimental values for $\widehat{\gamma^\circ}$, $\widehat{\gamma^\alpha}$, $\widehat{\gamma^\pi}$, and $\widehat{\gamma^\mu}$ (Supplementary Methods S1.1). The hat notation indicates that these variables are derived from the experimental training data and are used to estimate the unknown parameters in (4).

Model parameter estimation Model selection is based on a heuristic algorithm. It is inspired by ANOVA [45] and uses cross-validation [19,20] for hyperparameter selection. By approximating uptake, growth, and production rates separately, we deal with three smaller models instead of one large, simplifying the estimation. Parameter identification is performed using differential evolution [46] with SciPy's `differential_evolution` function [44], utilizing the previously calculated $\widehat{\gamma^\circ}$, $\widehat{\gamma^\alpha}$, and $\widehat{\gamma^\pi}$. Each of the three rates is fitted separately.

We assume that the effects of each variable in (4) are independent, meaning each effect is a separate term in the equation. Each model term, containing one influencing variable and one or more parameters, can be removed (if they are deemed insignificant) and the model remains valid. In case there is no temperature effect, Eqn. (4e) simplifies to

$$\gamma^i_j(T) = c_{\gamma^i_j} \quad i \in \{\mu, \pi, \circ\},\ j \in \{\max, \min\}. \tag{8}$$

Each model term of Eqn. (4) is tested. If it does not significantly improve the model fit, it is removed according to the following algorithm:

1. **Initial Fit**: Fit the model with all currently considered parameters by minimizing the sum of quadratic errors over all processes and measurement points. Calculate the error residuals and total variance for the measurement points (Bounds used in the error minimization are shown in Supplementary Table S2).
2. **Leave-One-Out Fit**: Repeat the fitting process for models, each missing one parameter.
3. ***F*-Test**: Use an F-test to determine if the reduction in variance is significant (i.e., $p < \alpha$) and calculate the difference in variance with and without the parameter.
4. **Remove Insignificant Parameters**: Remove the parameter with the highest p-value in the F-test.
5. **Iterate**: Repeat 1 to 4 until only significant terms remain.

Steps 1-5 are performed separately for each fitted rate ($\gamma^\circ, \gamma^\alpha$, and γ^π) and for 13 levels of α. The significance level depends on the training data (such as the number of processes and measurement points) and is computed through cross-validation. The significance level that results in the lowest error for the target variable ($\mathcal{Y}_{\frac{P}{X}}$) is selected (Supplemental Methods S1.2).

As for each iteration of our algorithm (Step 5), a term is removed from the model equation, the maximum amount of iterations is defined by the maximum amount of removable terms per rate equation plus one (i.e., seven).

To compare our heuristic approach to model selection with the corrected Akaike information criterion (AIC$_\mathrm{C}$) [17], we fit all possible parameter combinations for $\gamma^\circ, \gamma^\alpha$, and γ^π. Next, we calculate the AIC$_\mathrm{C}$ [Eqn. (14)] for each of the resulting models and selected the one with the lowest value.

2.1.3. Stage III: optimize

To optimize

$$\max_{f(t), G_\mathrm{f}, T(t), t_\mathrm{end}} \quad \mathcal{Y}_{\frac{P}{X}} = \frac{P(t_\mathrm{end})}{X(t_\mathrm{end})} \tag{9a}$$

$$\text{s. t.} \quad \text{Eqn. (1)} \tag{9b}$$

$$V \leq V_\mathrm{max} = 2.5 \text{ L} \tag{9c}$$

$$t_\mathrm{end} \leq t_\mathrm{max} = 12 \text{ h} \tag{9d}$$

we use `ipopt` 3.14.10 [47], which is a general nonlinear programming solver. For discretization and numerical differentiation [Eqn. (1)], we implemented an orthogonal collocation and optimization algorithm in `casadi` 3.3.5 [34] using Python. All required code to reproduce our analysis is available at https://github.com/gschloegel/OptFed.

To solve the differential equations in (1), we first scale the time coordinate by setting $t = t_\mathrm{end}\tau$, using the process' time t_end as a control variable. We then apply orthogonal collocation on 100 finite elements [48]. Specifically, we use Gauss–Legendre polynomials of degree one with collocation points at 0.5. For the substrate, due to system stiffness, we use Gauss–Radau collocation points at 1. Locally, we solve the differential equation using the backward Euler method. The controls (feed and temperature) are linear on each of the 100 intervals, allowing the optimization method to find optima with temperature gradients and unconventional feeding strategies.

To avoid rapid variations in the control variables $u = (f, G_\mathrm{f}, T, t_\mathrm{end})$, specifically in the feed and temperature profile $f(t_\mathrm{end}\tau)$, and $T(t_\mathrm{end}\tau)$, we add a regularization term to the objective function in (9a). This modified objective reads

$$\max_{f(t), G_\mathrm{f}, T(t), t_\mathrm{end}} \quad \mathcal{Y}_{\frac{P}{X}} = \frac{P(t_\mathrm{end})}{X(t_\mathrm{end})} - \sum_{i=1}^{\mathrm{length}(u)} \frac{c_i}{h} \sum_{j \in S} \left[\frac{u_i(j) - u_i(j+1)}{u_i(j) + u_i(j+1)} \right]^2, \tag{10}$$

where c_i is the penalty factor for each control variable u_i, h is the length of the finite elements, and S is the set of all sampling times.

In addition to the general optimization problem in (9), we defined a simplified version where all substrate is immediately used and no substrate accumulates, i.e., $\dot{G} = 0 = -\gamma^\circ X + \frac{f}{V}(G_\mathrm{f} - G)$. This mirrors the standard assumption in fed-batch processes and computationally avoids issues posed by stiff differential equations.

2.2. Response surface methodology

To benchmark OptFed, we evaluate its performance against response surface methodology (RSM) in process optimization. We calculated RSM using the `rsm` package [49] for R [50].

RSM requires that the control variables remain constant throughout the process and uses the process target metric (such as production concentration, yield, or productivity) to fit the model. Thus the RSM model is represented as:

$$\mathcal{Y}_{\frac{P}{X}} = c + c_f \mu_f + c_T T + c_{fT} \mu_f T + c_{f^2} \mu_f^2 + c_{T^2} T^2 \tag{11}$$

where c, c_f, c_T, c_{fT}, c_{f^2}, and c_{T^2} are fitted from the data minimizing the sum of quadratic errors. μ_f is the growth rate of the exponential feed (in $f = f_0 \exp(\mu_f t)$). The same model is calculated with the target variables P and X.

2.3. Model comparison and validation

To evaluate model fit [Eqn. (1)], individual (state) variables (e.g. P, P/X, X) or specific rates (e.g., γ°, γ^α, and γ^π), we use the coefficient of determination R^2 and its adjusted version R^2_adj, which are defined as:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/df_{\mathrm{res}}}{\mathrm{TSS}/df_{\mathrm{tot}}} \tag{12a}$$

with the residual and total sum of squares (RSS and TSS)

$$\mathrm{RSS} = \sum_{p\in\mathcal{P}} \sum_{m\in\mathcal{S}} \left(x_p(m) - \hat{x}_p(m)\right)^2 , \qquad \mathrm{TSS} = \sum_{p\in\mathcal{P}} \sum_{m\in\mathcal{S}} \left(x_p(m) - \langle \hat{x} \rangle\right)^2 , \tag{12b}$$

respectively. Here, $x_p(m)$ and $\hat{x}_p(m)$ represent the predicted and measured values respectively, $\langle \hat{x} \rangle$ represents the average of the observed values, $\mathcal{P}$ is the set of all processes, and $\mathcal{S}$ is the set of all sampling times. df_{res} and df_{tot} are the residual and total degrees of freedom, given by $df_{\mathrm{res}} = \#p - \#v - 1$ and $df_{\mathrm{tot}} = \#p - 1$, respectively, where $\#p = |\mathcal{P}| + |\mathcal{S}|$ represents the number of points and $\#v$ represents the number of variables.

Finally, we measure relative errors with respect to the mean of all data points:

$$x_{\mathrm{err}} = \frac{\hat{x} - x}{\langle \hat{x} \rangle} , \qquad \langle \hat{x} \rangle = \frac{1}{|\mathcal{P}| \cdot |\mathcal{S}|} \sum_{p\in\mathcal{P}} \sum_{m\in\mathcal{S}} \hat{x}. \tag{13}$$

Different models are compared using R^2 and R^2_{adj} on the state variables, focusing on P/X. In addition, we perform cross-validation using the leave-one-out strategy (predicting one process using all other processes) and compare R^2 for this as well. As an alternative metric for the goodness of the model, we calculate the corrected Akaike information criterion $\mathrm{AIC_C}$ [17]

$$\mathrm{AIC_C} = \#p \left(\log\left(\frac{\mathrm{RSS}}{\#v}\right) + 1\right) + 2\,\#v + \frac{\#v(1+\#v)}{\#p - \#v - 3}. \tag{14}$$

RSM only predicts end points of processes with constant exponential feed rates (μ_f) and constant temperatures (T). To ensure a fair comparison, here, we restricted OptFed to the same constraints.

In addition, we validate the OptFed framework by applying it to a case study and experimentally test the predicted optimal controls.

2.4. Case study

We illustrate our modeling framework by optimizing protein L production in a fed-batch fermentation of *E. coli*, and evaluate its effectiveness in comparison to RSM [49] in central composite design [10]. Data from twelve fermentations with varying specific substrate feed and temperature [51] are used as training input for both methods to predict an optimal bioprocess that maximizes the specific process yield $\mathcal{Y}_{\frac{P}{X}}$.

The dataset represents nine conditions (Fig. 2a):

- one center point (four runs),
- four star points (single runs) where either feeding rate or temperature varied from the center point,
- four factorial points (single runs) where both variables deviated from the center point.

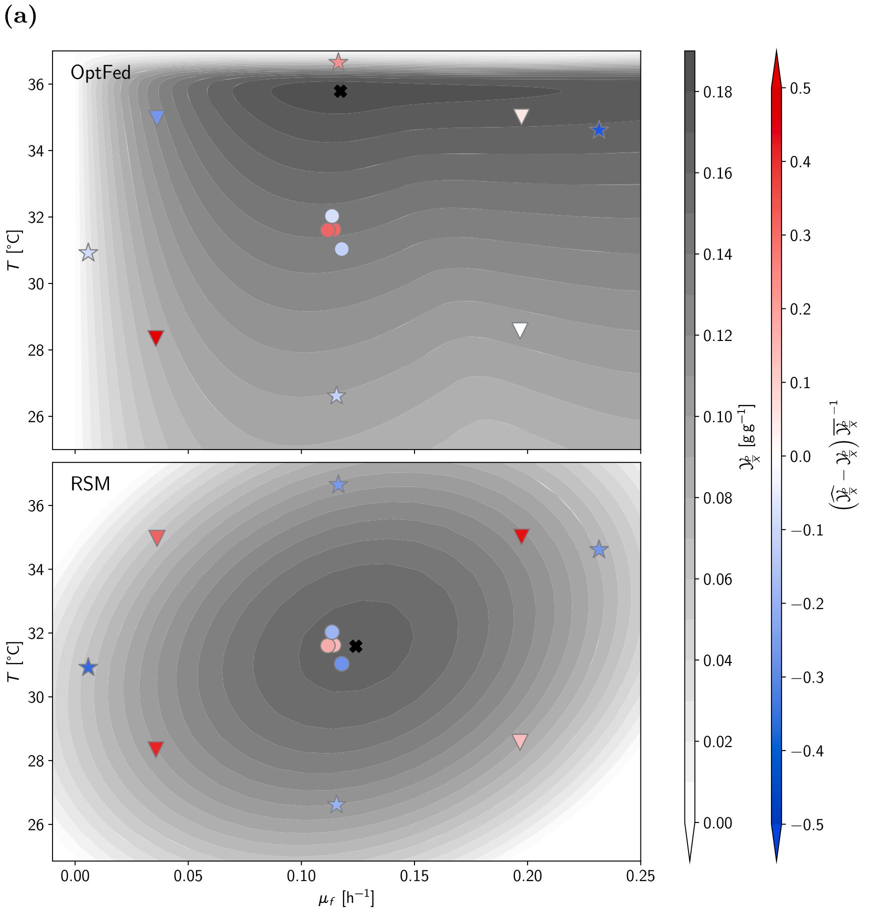
Training data Process data (biomass concentration, protein L concentration, and substrate concentration over time) from 12 fed-batch fermentations of *E. coli* strain BL21 DE3, with varying specific substrate feed and temperature [51], are used to fit the general process model.

In short, protein L accumulated intracellularly, and glycerol was the sole carbon source. IPTG (isopropyl β-D-1-thiogalactopyranoside) was used to induce the product promoter during the feed phase. The biomass concentration at induction varied between 20 g L^{-1} to 45 g L^{-1}. The control variables were the exponential feed rate coefficient (μ_f) and the temperature (T). Each process had a production phase of 12 h. Biomass and product concentrations were measured every 2 h, while temperature and feed rate were measured online. Due to the small reactor size, each sampling reduced the reactor volume non-negligibly, which was considered in the analysis. Details about the experimental setup and analytical methods are described in Section 2.4.1.

2.4.1. Experimental procedures

Experimental data for model fitting and validation in OptFed was obtained from 15 bioreactor cultivations. The cultivations were carried out in two bioreactor systems having a similar working volume. The 12 initial cultivations were executed in a DASGIP© Parallel Bioreactor System (max. working volume 2 L; Eppendorf, Hamburg, Germany) as described in [51]. In contrast to the original paper, where 11 out of 12 performed processes were used, according to the DoE, we use all 12 processes as training data. In addition, three validation runs were performed in a Minifors 2 bioreactor system (max. working volume 2.5 L; Infors HT, Bottmingen, Switzerland). For all cultivations a defined minimal medium according to DeLisa [52] was used, supplemented with an initial concentration of 20 g L^{-1} glycerol as the main carbon source and 0.02 g L^{-1} kanamycin as a selection marker. The temperature was set to 37 °C during batch phase, 35 °C during fed-batch phase and controlled at defined levels during induced fed-batch phase in accordance with the experimental plan. The pH was monitored with an EasyFerm pH electrode (Hamilton, Reno, NV, USA) and kept constant at 6.7 via addition of 12.5% NH_4OH. A probe for monitoring dissolved oxygen (dO_2) was installed (Visiferm DO425, Hamilton, Reno, NV, USA). The dissolved oxygen in the cell broth was kept over 40% through continuous stirring (1400 rpm) and aeration of 2 vvm. If needed, pure oxygen was added to the air flow. Furthermore, the off-gas was analyzed with respect to O_2 and CO_2 concentrations via a Bluevary sensor (BlueSens Gas analytics, Herten, Germany) for real-time monitoring of the metabolic activity of the cells. The process parameters were logged and controlled using the bioprocess management system eve© (Infors HT, Bottmingen, Switzerland). The expression of recombinant protein L was induced by a one-point addition of sterile Isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. After addition of the inducer, samples were taken every two hours for further process and product analytics.

An *E. coli* BL21 (DE3) strain transformed with a pET-24a(+) plasmid was used for the cultivations (GenBank accession no. AAA67503). The plasmid carries the codon-optimized genes coding the 5B (binding) protein L with a C-terminal His$_6$-tag. The recombinant protein L is expressed intracellularly. The cells were harvested and subsequent analytics were done. All subsequent analytical steps were realised with samples of 35 mL cell broth each. The cell broth was centrifuged (10 min, 21 000 rpm, 4 °C) and the supernatant was separated from the cell pellet and aliquoted (1 mL) for anion exchange chromatography. Biomass concentration was quantified by dry cell weight (DCW) in triplicates. Therefore, the cell pellet was washed with saline (0.9 wt.% NaCl), centrifuged with the same settings and dried at 105 °C for 48 h. In addition, the biomass concentration was determined via optical density measurements at 600 nm wavelength (OD600) in triplicates. Residual glycerol and metabolites in the cell-free supernatant were analyzed by a high performance liquid chromatography (HPLC) system (UltiMate 3000; Thermo Fisher, Waltham, MA) equipped with an Aminex HPX-87 H column (Bio-Rad Laboratories, Hercules, CA, USA). HPLC standards with various concentrations of protein L (0.063-1.0 g L^{-1}), glycerol (0.781-50 g L^{-1}) and acetate (1-10 g L^{-1}) were prepared separately. A sample volume of 10 mL cell broth was centrifuged (15 min, 14 000 rpm, 4 °C) and the separated cell pellet was re-suspended in 40 mL lysis buffer (10 mM EDTA, 100 mM Tris, pH 7.4) and homogenized subsequently (7 passages, 1200 bar; PandaPLUS, Gea AG, Germany). After centrifugation of the crude cell lysate (20 min, 14 000 rpm, 4 °C), the supernatant was analyzed using a reversed-phase HPLC method for protein L quantification based on a PpL standard calibration curve. The UltiMate 3000 HPLC system was equipped with a BioResolve reversed-phase Polyphenyl column (Waters Corporation, MA, USA). Further information about the analytical procedures can be found in [51].



(b)

Variable	R^2(fit)		adjusted R^2(fit)		CV-R^2(cross validation)	
P/X	0.73	(0.56)	0.69	(0.19)	0.64	(-1.50)
P	0.82	(0.90)	0.79	(0.81)	0.75	(0.45)
X	0.83	(0.55)	0.81	(0.17)	0.66	(-1.14)

Fig. 2. Estimation of final product titer and model errors (a) predicted specific yield in OptFed (top) versus RSM (bottom). Both models are constructed using data from twelve fermentations, including a center point (circle, four replicates), four factorial points (triangles), and four star points (stars; each single runs). The predicted specific product yields $\mathcal{Y}_{\frac{P}{X}}$ at the end of the process are shown in shades of gray, with the black cross indicating the optima. Both optima are calculated for exponential feed rates at constant temperature throughout the production phase. Model errors, indicating differences between predicted and measured $\mathcal{Y}_{\frac{P}{X}}$, are indicated by the color of the markers: red for overestimation, white for accurate prediction, and blue for underestimation. (b) Goodness of fit [R^2(fit), and adjusted R^2(fit)] as well as goodness of fit for leave-one-out cross-validation (CV) [as measured by CV-R^2(cross validation)]. Values in brackets refer to the RSM model, values in standard print to OptFed.

3. Results

We illustrate our modeling framework by optimizing protein L production in a fed-batch fermentation of *E. coli*, and evaluate its effectiveness in comparison to RSM [49] in central composite design [10]. Data from twelve fermentations with varying specific substrate feed and temperature [51] are used as input for both methods to predict an optimal bioprocess that maximizes the specific process yield $\mathcal{Y}_{\frac{P}{X}}$ (product per biomass ratio at the end of the process).

3.1. *Response surface methodology predicts limited optimization potential*

RSM utilizes a second-degree polynomial model to forecast the specific yield $\mathcal{Y}_{\frac{P}{X}}$ as a function of the specific feeding rate and temperature (Equation (11)). It predicts an optimal specific yield of 0.16 g g^{-1} near the center point at $\mu_f = 0.12$ h^{-1} and $T = 31$ °C (Fig. 2a). However, the predicted yield improvement is small (+1%) yet uncertain (adjusted $R^2 = 0.19$; Fig. 2b), and none of the model's parameters are statistically significant (at a 0.95 confidence level, Supplementary Table S3).

3.2. *OptFed identifies significant optimization potential at high temperature*

3.2.1. *Model simplification and parameter estimation*

Initially, we mitigate measurement errors of the state variables by fitting them with cubic splines (Supplementary Figure S1). Based on these splines, we parameterize our general model using the algorithm described in the Methods Section 2.1.2. We found that only 12 out of 27 parameters of the general process model are statistically significant and required ($\alpha = 0.2$, Supplementary Figure S2), with only a minor drop in the explained variance (see Tables 1 and 2). The difference of R^2 and adjusted R^2 is higher for the full model as the degrees of freedom

Table 1

List of parameters remaining in the selected model and their fitted values. The increased RSS (residual sum of squares) column next to a parameter shows the increase in fitting error (calculated with Equation (12)) if this parameter would be removed from the reduced model. K_m° is not removed, as removal leads to physically impossible negative substrate concentrations. A comprehensive list of all parameters of the initially designed model is found in Supplementary Table S1).

	Name	Unit	Value	increased RSS
Parameters (fitted using training data)				
$c_{\gamma^\circ_{\max}}$	maximal uptake rate	$\mathrm{g\,g^{-1}\,h^{-1}}$	0.49	
$c_{\gamma^\alpha_{\min}}$	maintenance requirement without growth	$\mathrm{g\,g^{-1}\,h^{-1}}$	2.4×10^{-5}	
E_0^π	hypothetical enzyme concentration		8.8×10^{-9}	
K_m°	dependence on substrate concentration	g	1.0×10^{-3}	
K_g^α	growth dependent maintenance	$\mathrm{g\,g^{-1}\,h^{-1}}$	1.0×10^{-4}	142%
K_m^π	production dependence on available substrate	$\mathrm{g\,g^{-1}\,h^{-1}}$	0.62	188%
K_G°	substrate inhibition	$\mathrm{g\,L^{-1}}$	89	119%
K_P^α	increase caused by product	$\mathrm{g\,L^{-1}}$	0.11	37%
K_n^π	inhibition for higher no. of generations		1.5	50%
$\Delta G^\pi_{\mathrm{cat}}$	catalytic activation energy	$\mathrm{J\,mol^{-1}}$	5.2×10^{4}	38%
$\Delta H^\pi_{\mathrm{eq}}$	enthalpic (conversion of active to inactive enzyme)	$\mathrm{J\,mol^{-1}}$	4.8×10^{6}	
T^π_{eq}	temperature where half the enzyme is active	K	310	
E_0^π	hypothetical enzyme concentration	K	8.8×10^{-9}	
Constants (from literature)				
$Y_{\frac{X_r}{G}}$	biomass yield per substrate	$\mathrm{g\,g^{-1}}$	0.627	
$Y_{\frac{P}{G}}$	product yield per substrate	$\mathrm{g\,g^{-1}}$	0.652	
k_B	Boltzmann constant	$\mathrm{J\,K^{-1}}$	1.38×10^{-23}	
R	Gas constant	$\mathrm{J\,mol^{-1}\,K^{-1}}$	8.314	
h	Plank constant	J h	2.39×10^{-30}	

Table 2

Overview of the goodness of fit of the kinetic functions. The values of the parameters of the reduced OptFed are given in Table 1.

	reduced OptFed			full OptFed		
	γ°	γ^α	γ^π	γ°	γ^α	γ^π
R^2	0.54	0.68	0.70	0.58	0.74	0.73
adjusted R^2	0.52	0.67	0.68	0.24	0.71	0.70
# of parameters	3	3	6	9	9	9

are higher. The effect is especially pronounced for γ° as only 3 out of 12 processes are used for estimation (for other processes the substrate concentration is below the limit of quantification). Further reducing the model would increase the fitting error by at least one-third (Supplementary Figure S3). Thus, with the parameter values listed in Table 1, the reduced model reads:

$$\gamma^\circ = c_{\gamma^\circ_{\max}} \frac{G}{K_m^\circ + G} \frac{K_G^\circ}{K_G^\circ + G} \tag{15a}$$

$$\gamma^\alpha = c_{\gamma^\alpha_{\min}} \left(1 + \gamma^\circ\ K_g^\alpha\right) \left(1 + P/X\ K_P^\alpha\right) \tag{15b}$$

$$\gamma^\pi = \gamma^\pi_{\max}(T) \frac{\gamma^\circ - \gamma^\alpha}{K_m^\pi + \gamma^\circ - \gamma^\alpha} \frac{1}{1 + n\ K_n^\pi} \tag{15c}$$

$$\gamma^\mu = \gamma^\circ - \gamma^\pi - \gamma^\alpha \tag{15d}$$

with

$$\gamma^\pi_{\max}(T) = E_0^\pi \frac{k_B T}{h} \frac{\exp\left(-\frac{\Delta G^\pi_{\mathrm{cat}}}{RT}\right)}{1 + \exp\left(\frac{\Delta H^\pi_{\mathrm{eq}}}{R}\left(T^{\pi\,-1}_{\mathrm{eq}} - T^{-1}\right)\right)}. \tag{15e}$$

In the reduced OptFed process model (15), the substrate uptake rate γ° follows Michaelis-Menten kinetics with self-inhibition by the substrate. The substrate-to-maintenance flux γ^α increases multilinearly with the substrate uptake rate γ°, and the product-to-biomass yield P/X. While both fluxes are temperature-independent, the product formation rate π is modeled as a temperature-dependent Michaelis-Menten-like kinetic with non-competitive inhibition by the number of generations n after induction. Supplementary Figure S3 illustrates the quality of our model's fit on rates. Despite some noise and occasional large errors in individual data points, the overall trend is well-predicted.

We compare the measurement data for product and biomass with the model predictions and validate these predictions using cross-validation. The adjusted R^2 values remain above 0.52 for P, P/X, and X, both with and without cross-validation (Fig. 2b and Supplementary Figures S4 and S5). Thus, we conclude that the model is a reliable choice, especially compared to RSM.

Additionally, we compare our heuristic model selection algorithm (Section 2.1.2) with model selection based on the $\mathrm{AIC_C}$. The selected kinetic parameters are almost identical for both methods, only $\mathrm{AIC_C}$ includes terms for G in γ^α and γ^π. With our heuristic algorithm, these variables were removed in the last step of our elimination as they do not improve the model fit significantly (p-values of 0.25 and 0.30). While optimal feed rates are different (about ±22%), temperature optimum is similar (±0.03 °C). Using the model optimum for one model and testing it with the other misses the optimal $\mathcal{Y}_{\frac{P}{X}}$ by less than 0.6%.

3.2.2. Process model optimization

Using the model develop above, we apply the optimization algorithm described in Methods Section 2.1.3 to determine the optimal feeding strategy and temperature settings. Additionally, we consider the following constraints:

- The initial biomass matches the mean value of the center point runs of the experimental data, $X_0 = 30\ \mathrm{g\,L^{-1}}$.
- The initial bioreactor volume is set at 1.3 L, with no maximum volume constraint.
- The feed glycerol concentration is fixed at 390 $\mathrm{g\,L^{-1}}$.
- Sampling of 35 mL reduces the current reactor volume at $t = 2, 4, 6, 8$, and 10 h.
- A linear regression model (Supplementary Figure S6) approximates the base addition as

$$f_{\mathrm{base}} = XV \times 10^{-7} \times (13\,900\mu - 1.1\ \mathrm{h^{-1}}).$$

Fig. 3 illustrates the predicted optimal fed-batch process for protein L production. At 35.8 °C and after 12 h, we predict a biomass concentration of 52 $\mathrm{g\,L^{-1}}$ and a product concentration of 9.6 $\mathrm{g\,L^{-1}}$, resulting in

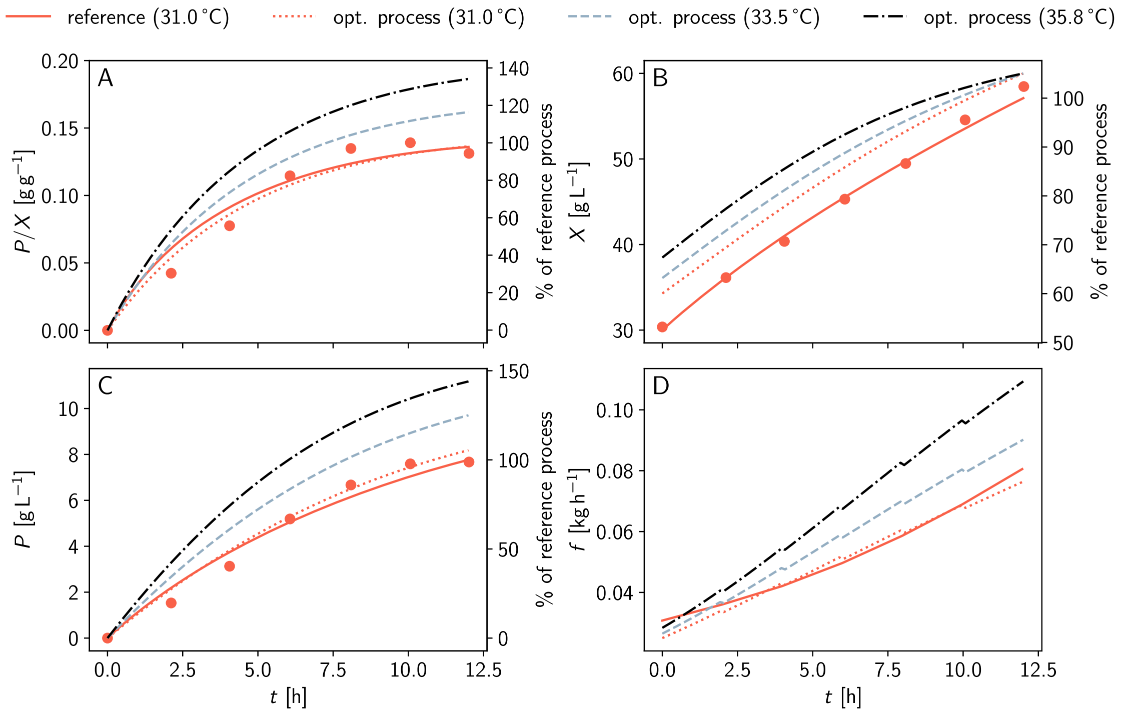


Fig. 3. Predicted optimal process with OptFed. The red dashed-dotted lines represent OptFed's prediction for the reference process, while grey lines show the predicted optimal behavior at $T = 35.8$C. Black dashed lines indicate the optimal process with approximated linear feed rate. The simplification of the feed profile has a negligible impact on the process (grey full and black dashed lines). After 12 h, a 37.4% improvement in specific product yield is predicted. For reference, measurements from the training data are shown (red circles, center point, same initial biomass).

an optimal product yield of 0.19 $\mathrm{g\,g^{-1}}$. This represents a 37.1% increase compared to the reference process.

Increasing the temperature is key for optimized performance (Fig. 3). According to the model, maximum production rate $\gamma^{\pi}_{\max}$ increases with temperature up to the optimal temperature of 35.8 °C and decreases sharply for higher temperatures (Supplementary Figure S3). Optimizing the feed but keeping the temperature at 31 °C increases the specific product yield by just 0.3%. Conversely, keeping the feed constant and raising only the temperature boosts the maximally obtainable specific product yield by 37.0%. A summary of optimization results can be found in Supplementary Table S4.

Fig. 2a compares the predictions of RSM and OptFed. Under identical process constraints (constant exponential feed rate and constant temperature), OptFed identifies an optimum at elevated temperatures that RSM misses.

Unlike the training data's fermentations, which used an exponential feed, we find that an almost linearly increasing feed rate is best to maximize the product-to-biomass yield (Fig. 3D). Therefore, we decided to approximate the predicted optimal feed function with a simple linear equation (Supplementary Figure S6)

$$f_{\mathrm{opt}} = 0.022\ \mathrm{g\,h^{-1}} + 0.0053\ \mathrm{g\,h^{-2}}\ t. \tag{16}$$

This adjustment changes the final product and biomass concentrations and the product-to-biomass yield $\mathcal{Y}_{\frac{P}{X}}$ by less than 0.1%, but significantly eases practical implementation (Fig. 4). Generally, variations of the feed function have little influence as long as the initial and final biomass concentrations remain constant.

Our model's predictions are validated by experimentally running the optimal fermentation process with the simplified linear feed (Fig. 4). Additionally, an intermediate process at 33.5 °C (halfway between the temperature of the center point and the optimum) is carried out.

For the first 6.5 h, the processes closely matched the predictions (Fig. 4). The optimal process at 36 °C achieved a 19% increase in specific product yield (compared to the predicted 21%), while the process at 33.5 °C achieved a 5% increase (compared to the predicted 18%). Data points beyond 6.5 h are affected by unstable temperatures and therefore not considered.

4. Discussion

Recent years have seen a surge in research on biotechnological process parameterization and optimization strategies [21,53–55]. However, studies usually focus on one of the two aspects. Here, we developed a comprehensive modeling framework, OptFed, to strategically combine them.

OptFed employs a general phenomenological process model fitted with experimental time series data from multiple fermentations. The algorithm discards parameters with insufficient statistical power to minimize overfitting, simplify the model, and increase the model's reliability. Using this parameterized model, we applied nonlinear optimization to predict a fermentation profile for optimal specific protein L yield in *E. coli*, resulting in a near-linear feed at an elevated temperature of 35.8 °C. This approach predicted a 37% increase in specific product yield compared to the training data, significantly outperforming the standard RSM, which only predicted a 1% improvement. However, during the experimental validation of the optimized process, we encountered issues with temperature stability shortly after 6.5 h and decided not to use data collected thereafter. This instability is due to an undersized cooling capacity of the bioreactor.

Despite this, at 6.5 h the specific protein yield was up by 19%, close to the predicted 21% at that time. The primary increase in specific product yield is attributed to the elevated temperature, accounting for over 99% of the improvement. According to OptFed, product concentration increases approximately linearly with temperature up to the optimum. This dependence is completely missed by RSM, highlighting the advantage of our approach.

Compared to RSM, our process equations constrain the possible solution space to more realistic outcomes. In fact, combining mechanistic modeling with purely statistical approaches has already previously been

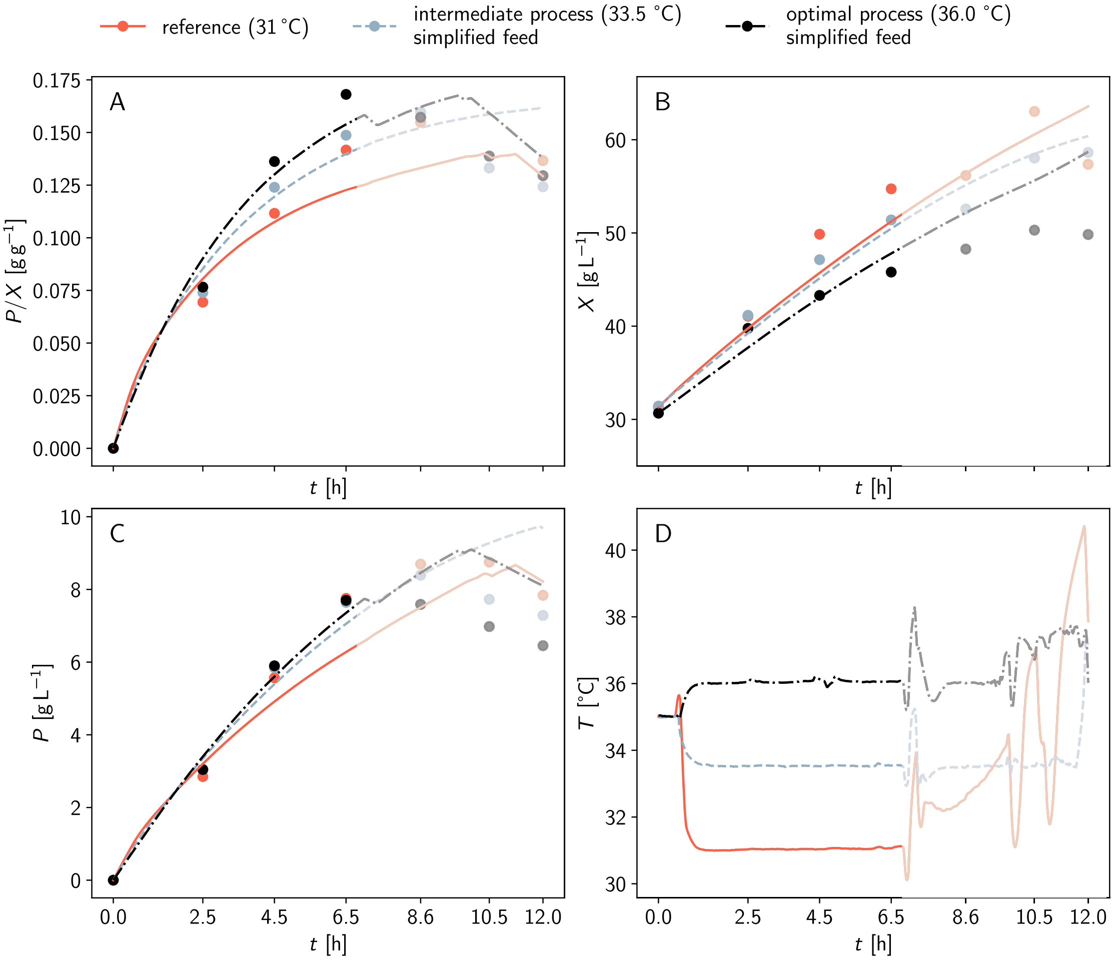


Fig. 4. Validation of the predicted optimal fermentation. Panels A to C compare experimental data (circles) with model predictions (lines). Corresponding feed rates, substrate uptake rates, growth rates, and product formation rates are illustrated in Supplementary Figure S9. Panel D shows the temperature profile of the fermentations. For $t > 6.5$ h temperatures are unstable and the data is not considered. (opaque region in panels A to D). A comparison of optimal predicted and experimentally achieved controls is given in Supplementary Figure S7.

shown to perform better than pure RSM [54]. For example, RSM may predict negative values of the target metric (Fig. 2), an effect that cannot be observed with OptFed. Moreover, our process equations implicitly ensure mass balance due to the calculation of the substrate-to-growth rate (γ^{μ}) from the difference of substrate uptake and the other substrate draining fluxes (Equation (2)).

Maintenance flux, as we use it throughout the manuscript, is defined as a catchall-term for several metabolic effects. It comprises (1) the (non-)growth associated maintenance [56], (2) all non-optimal growth and production due to byproduct formation [57], and (3) any overflow metabolism during high substrate uptake rates [58]. Consequently, our yields are derived from a metabolomic model, excluding maintenance requirements, which differs from experimental yields where maintenance is included. Maintenance accounts for more than half of the uptake at the end of production (Supplementary Figure S8), leading to seemingly higher-than-usual yields ($Y_{\frac{X_r}{G}}$, $Y_{\frac{P}{G}}$) in Table 1.

Furthermore, we estimate uptake rates during production based on experimental data. As uptake can be significantly reduced during production [59] this avoids possible overfeeding in the predicted optimum. In the case study, we observe a reduced uptake rate, but uptake is not a limiting factor for optimization.

The mathematical problems in our procedure, such as model identification and calibration, are difficult to solve because models often have too many interrelated parameters and lack sufficient high-quality training data. This can cause optimization algorithms to be unstable and not converge, presenting significant challenges [60]. Our model, for example, requires high-quality time course data. Random (relative) errors for biomass and product concentrations should be in the range of 3% and 15%, respectively, to reliably identify the correct optimum (Supplementary Figure S10). However, these experimental uncertainties are manageable with current process monitoring technology [61].

Compared to the center point of the training data, almost all of the improvement in our case study originates from the increase in temperature. However, due to different product formation kinetics, this may be different for other products. Therefore, we cannot derive a general rule which feed profiles and temperatures are more advantageous in other setups. For example, [62] find a high influence of the feeding strategy on the production of inclusion bodies. This is most likely to a difference in product and process setup.

In Section 2.1.3, we optimize for the product-to-biomass yield ($\mathcal{Y}_{\frac{P}{X}}$), a rather unconventional metric, compared to the titer, productivity, and (product-to-substrate) yield commonly used [63]. However, in this case study, the maximization of $\mathcal{Y}_{\frac{P}{X}}$ is of critical importance for the ease of downstream processing. This strategy bears the risk of converging to a process that yields excellent $\mathcal{Y}_{\frac{P}{X}}$ but very low amounts of product overall, which is also unfavorable. Here, we mitigated the risk by comparing our optimal $\mathcal{Y}_{\frac{P}{X}}$ process to an optimal PV process (Supplementary Figure S11). Volume can be easily scaled by scaling batch volume and feed rate. The reachable biomass concentration depends on the reactor design (cooling capabilities, oxygen supply). A biomass limit of 60 g L^{-1} (reached in training processes) could increase the product by 16%.

While our optimum is stable for changes in feeding strategy, we see a sharp drop in productivity when the temperature is raised above the optimum. This means, the optimal temperature is a good first guess, but more data is required when we move from screening to the design of the production setup. Based on the existing data, we applied a Monte Carlo estimation [64–66] (Supplementary Figure S12). We observe that we can guarantee (at a significance level of 0.05) a production rate within 10% of the optimum by reducing the temperature from 35.8 °C to 34.9 °C.

OptFed focuses on the model selection. Based on this further improvements are possible. Sensitivity analyses [67] could help to make the optimum more stable considering the uncertainties in parameter fitting. We also limit ourselves by using existing data. Model-based design of experiment [68] could provide better training data or can be used to plan additional experiments to improve the model.

5. Conclusion

In this study, we presented OptFed, a phenomenological model-based bioprocess optimization framework that (also) allows us to seamlessly integrate preexisting biological and process knowledge. Unlike other tools, we emphasized the parameterization of process equations using experimental bioprocess data. To prevent overfitting, OptFed employs a multi-step fitting strategy that retains only the terms that significantly reduce model error, discarding others.

This approach addresses key challenges in industrial process design by eliminating the reliance on trial-and-error methods and standard, predefined feeding strategies. We demonstrated that OptFed not only accurately describes the training data but also predicts optimized process controls. Experimental validation shows a 19% increase in specific protein L yield compared to the control.

While effective, OptFed's performance depends on the quality of the training data. Future work will explore expanding the model to address more complex biological phenomena, incorporate multi-objective optimization, and validate its application across a broader range of bioprocesses and products.

We are confident that OptFed is a valuable tool for bioprocess optimization and will benefit the industry in the future.

CRediT authorship contribution statement

Guido Schlögel: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Rüdiger Lück:** Writing – review & editing, Investigation. **Stefan Kittler:** Writing – review & editing, Investigation. **Oliver Spadiut:** Writing – review & editing, Funding acquisition. **Julian Kopp:** Writing – review & editing, Investigation. **Jürgen Zanghellini:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Mathias Gotsmy:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

The authors thank the European Union's Horizon Europe Research and Innovation Programme for their funding facilitating this study through the Enviromed project (Grant agreement No. 101057844).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.09.024.

Data availability

All code required to reproduce our analysis, as well as all training and validation data, is available at https://github.com/gschloegel/OptFed/.

References

[1] Lim Henry C, Shin Hwa Sung. Fed-batch cultures: principles and applications of semi-batch bioreactors. Cambridge University Press; 2013.

[2] Rodrigues Maria Isabel, Iemma Antonio Francisco. Experimental design and process optimization. Boca Raton: CRC Press. ISBN 978-0-429-16186-5, December 2014.

[3] Modak JM, Lim HC, Tayeb YJ. General characteristics of optimal feed rate profiles for various fed-batch fermentation processes. Biotechnol Bioeng 1986;28(9):1396–407.

[4] Maurer Michael, Kühleitner Manfred, Gasser Brigitte, Mattanovich Diethard. Versatile modeling and optimization of fed batch processes for the production of secreted heterologous proteins with pichia pastoris. Microb Cell Fact 2006;5(1):1–10.

[5] Steinberg David M, Bursztyn Dizza. Response surface methodology in biotechnology. Qual Eng March 2010;22(2):78–87. https://doi.org/10.1080/08982110903510388. Publisher: Taylor & Francis.

[6] de Oliveira Rafael D, Le Roux Galo AC, Mahadevan Radhakrishnan. Nonlinear programming reformulation of dynamic flux balance analysis models. Comput Chem Eng 2023;170:108101.

[7] Klamt Steffen, Mahadevan Radhakrishnan, Hädicke Oliver. When do two-stage processes outperform one-stage processes? Biotechnol J 2018;13(2):1700539.

[8] Mermoud Grégory. Model-based optimization. In: Mermoud Gregory, editor. Stochastic reactive distributed robotic systems: design, modeling and optimization. Springer tracts in advanced robotics. Cham: Springer International Publishing. ISBN 978-3-319-02609-1, 2014. p. 175–9.

[9] Carvalho João CM, Vitolo Michele, Sato Sunao, Aquarone Eugênio. Ethanol production by Saccharomyces cerevisiae grown in sugarcane blackstrap molasses through a fed-batch process. Appl Biochem Biotechnol September 2003;110(3):151–64. https://doi.org/10.1385/ABAB:110:3:151.

[10] Khuri André I, Mukhopadhyay Siuli. Response surface methodology. WIREs: Comput Stat 2010;2(2):128–49. https://doi.org/10.1002/wics.73.

[11] Monod Jacques. The growth of bacterial cultures. Annu Rev Microbiol October 1949;3(1):371–94. Publisher: Annual Reviews.

[12] Lopes Marta B, Martins Gabriel, Calado Cecília RC. Kinetic modeling of plasmid bioproduction in Escherichia coli DH5α cultures over different carbon-source compositions. J Biotechnol September 2014;186:38–48. https://doi.org/10.1016/j.jbiotec.2014.06.022. https://www.sciencedirect.com/science/article/pii/S0168165614003137.

[13] Klumpp Stefan, Zhang Zhongge, Hwa Terence. Growth rate-dependent global effects on gene expression in bacteria. Cell December 2009;139(7):1366–75. https://doi.org/10.1016/j.cell.2009.12.001. ISSN 0092-8674, 1097-4172. https://www.cell.com/cell/abstract/S0092-8674(09)01505-0.

[14] Kager Julian, Bartlechner Johanna, Herwig Christoph, Jakubek Stefan. Direct control of recombinant protein production rates in E. coli fed-batch processes by nonlinear feedback linearization. Chem Eng Res Des June 2022;182:290–304. https://doi.org/10.1016/j.cherd.2022.03.043. https://www.sciencedirect.com/science/article/pii/S0263876222001460.

[15] Weber Jan, Hoffmann Frank, Rinas Ursula. Metabolic adaptation of Escherichia coli during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. Biotechnol Bioeng 2002;80(3):320–30. https://doi.org/10.1002/bit.10380.

[16] Jannasch Holger W, Egli Thomas. Microbial growth kinetics: a historical perspective. Antonie Van Leeuwenhoek September 1993;63(3):213–24. https://doi.org/10.1007/BF00871219.

[17] Hurvich Clifford M, Tsai Chih-Ling. Regression and time series model selection in small samples. Biometrika June 1989;76(2):297–307. https://doi.org/10.1093/biomet/76.2.297.

[18] Lee Jason D, Sun Dennis L, Sun Yuekai, Taylor Jonathan E. Exact post-selection inference, with application to the lasso. Ann Stat June 2016;44(3):907–27. https://doi.org/10.1214/15-AOS1371. ISSN 0090-5364, 2168-8966, Publisher: Institute of Mathematical Statistics. https://projecteuclid.org/journals/annals-of-statistics/volume-44/issue-3/Exact-post-selection-inference-with-application-to-the-lasso/10.1214/15-AOS1371.full.

[19] Zhang Yongli, Yang Yuhong. Cross-validation for selecting a model selection procedure. J Econom July 2015;187(1):95–112. https://doi.org/10.1016/j.jeconom.2015.02.006. https://www.sciencedirect.com/science/article/pii/S0304407615000305.

[20] Yates Luke A, Aandahl Zach, Richards Shane A, Brook Barry W. Cross validation for model selection: a review with examples from ecology. Ecol Monogr 2023;93(1):e1557. https://doi.org/10.1002/ecm.1557.

[21] Sánchez Benjamín J, Soto Daniela C, Jorquera Héctor, Gelmi Claudio A, Pérez-Correa José R. HIPPO: an iterative reparametrization method for identification and calibration of dynamic bioreactor models of complex processes. Ind Eng Chem Res December 2014;53(48):18514–25. https://doi.org/10.1021/ie501298b. Publisher: American Chemical Society.

[22] Jaqaman Khuloud, Danuser Gaudenz. Linking data to models: data regression. Nat Rev Mol Cell Biol November 2006;7(11):813–9. https://doi.org/10.1038/nrm2030. https://www.nature.com/articles/nrm2030. Publisher: Nature Publishing Group.
[23] Varma Sudhir, Simon Richard. Bias in error estimation when using cross-validation for model selection. BMC Bioinform February 2006;7(1):91. https://doi.org/10.1186/1471-2105-7-91.
[24] Gigerenzer Gerd, Homo Henry Brighton. Heuristicus: why biased minds make better inferences. Top Cogn Sci 2009;1(1):107–43. https://doi.org/10.1111/j.1756-8765.2008.01006.x.
[25] Hawkins Douglas M. The problem of overfitting. J Chem Inf Comput Sci January 2004;44(1):1–12. https://doi.org/10.1021/ci0342472. Publisher: American Chemical Society.
[26] Nielsen Jens, Keasling Jay D. Engineering cellular metabolism. Cell 2016;164(6):1185–97.
[27] Nocedal Jorge, Wright Stephen J. Numerical optimization. 2nd ed edition. Springer series in operations research. New York: Springer. ISBN 978-0-387-30303-1, 2006. OCLC: ocm68629100.
[28] Kalise Dante, Kunisch Karl, Rao Zhiping. Hamilton-Jacobi-Bellman equations: numerical methods and applications in optimal control. De Gruyter. ISBN 978-3-11-054359-9, August 2018. Publication Title: Hamilton-Jacobi-Bellman Equations.
[29] Liberzon Daniel. Calculus of variations and optimal control theory. Princeton University Press. ISBN 978-0-691-15187-8, 2012.
[30] Srinivasan B, Palanki S, Bonvin D. Dynamic optimization of batch processes: I. Characterization of the nominal solution. Comput Chem Eng January 2003;27(1):1–26. https://doi.org/10.1016/S0098-1354(02)00116-3. https://www.sciencedirect.com/science/article/pii/S0098135402001163.
[31] Vassiliadis VS, Sargent RWH, Pantelides CC. Solution of a class of multistage dynamic optimization problems. 1. Problems without path constraints. Ind Eng Chem Res September 1994;33(9):2111–22. https://doi.org/10.1021/ie00033a014. Publisher: American Chemical Society.
[32] Bellman Richard E. Dynamic programming. Princeton University Press. ISBN 978-1-4008-3538-6, August 2021. Publication Title: Dynamic Programming.
[33] Bojkov Bojan, Luus Rein. Time-optimal control by iterative dynamic programming. Ind Eng Chem Res June 1994;33(6):1486–92. https://doi.org/10.1021/ie00030a008. Publisher: American Chemical Society.
[34] Andersson Joel AE, Gillis Joris, Horn Greg, Rawlings James B, Diehl Moritz. CasADi: a software framework for nonlinear optimization and optimal control. Math Program Comput March 2019;11(1):1–36. https://doi.org/10.1007/s12532-018-0139-4.
[35] Beal Logan DR, Hill Daniel C, Martin R Abraham, Hedengren John D. GEKKO optimization suite. Processes August 2018;6(8):106. https://doi.org/10.3390/pr6080106. https://www.mdpi.com/2227-9717/6/8/106. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
[36] Hart William E, Watson Jean-Paul, Woodruff David L. Pyomo: modeling and solving mathematical programs in Python. Math Program Comput August 2011;3(3):219. https://doi.org/10.1007/s12532-011-0026-8.
[37] Yang Jingyi, Yang Yuebao, Li Mingtao. OptControl.jl: an interpreter for optimal control problem. http://arxiv.org/abs/2207.13229. arXiv:2207.13229 [math], July 2022.
[38] Michaelis Leonor, Menten Maud Leonora. Die Kinetik der Invertinwirkung. Biochem Z 1913;49:333–69. http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/17273.
[39] van Bodegom Peter. Microbial maintenance: a critical review on its quantification. Microb Ecol May 2007;53(4):513–23. https://doi.org/10.1007/s00248-006-9049-5.
[40] Daniel Roy M, Danson Michael J, Eisenthal Robert, Lee Charles K, Peterson Michelle E. The effect of temperature on enzyme activity: new insights and their implications. Extremophiles January 2008;12(1):51–9. https://doi.org/10.1007/s00792-007-0089-7.
[41] Monk Jonathan M, Koza Anna, Campodonico Miguel A, Machado Daniel, Seoane Jose Miguel, Palsson Bernhard O, et al. Multi-omics quantification of species variation of escherichia coli links molecular features with strain phenotypes. Cell Syst 2016;3(3):238–51.
[42] Froment Gilbert F, Bischoff Kenneth B, De Wilde Juray. Chemical reactor analysis and design, vol. 2. New York: Wiley; 1990.
[43] Bardow André, Marquardt Wolfgang. Incremental and simultaneous identification of reaction kinetics: methods and comparison. Chem Eng Sci July 2004;59(13):2673–84. https://doi.org/10.1016/j.ces.2004.03.023. https://www.sciencedirect.com/science/article/pii/S0009250904002015.
[44] Virtanen Pauli, Gommers Ralf, Oliphant Travis E, Haberland Matt, Reddy Tyler, Cournapeau David, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
[45] Rutherford Andrew. Introducing Anova and Ancova: a GLM approach. Introducing statistical methods. London: SAGE Publications Ltd. ISBN 978-0-7619-5160-5, 2001. https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=251737&site=ehost-live.
[46] Storn Rainer, Price Kenneth. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim December 1997;11(4):341–59. https://doi.org/10.1023/A:1008202821328.
[47] Wächter Andreas, Biegler Lorenz T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math Program March 2006;106(1):25–57. https://doi.org/10.1007/s10107-004-0559-y.
[48] Biegler Lorenz T. Nonlinear programming: concepts, algorithms, and applications to chemical processes. Society for Industrial and Applied Mathematics; January 2010. ISBN 978-0-89871-702-0, 978-0-89871-938-3. http://epubs.siam.org/doi/book/10.1137/1.9780898719383.
[49] Lenth Russell V. Response-surface methods in R, using rsm. J Stat Softw 2010;32:1–17. https://doi.org/10.18637/jss.v032.i07.
[50] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. https://www.R-project.org/.
[51] Kittler Stefan, Ebner Julian, Besleaga Mihail, Larsbrink Johan, Darnhofer Barbara, Birner-Gruenberger Ruth, et al. Recombinant protein L: production, purification and characterization of a universal binding ligand. J Biotechnol November 2022;359:108–15. https://doi.org/10.1016/j.jbiotec.2022.10.002. https://www.sciencedirect.com/science/article/pii/S0168165622002371.
[52] DeLisa Matthew P, Li Jincai, Rao Govind, Weigand William A, Bentley William E. Monitoring GFP-operon fusion protein expression during high cell density cultivation of Escherichia coli using an on-line optical sensor. Biotechnol Bioeng 1999;65(1):54–64. https://doi.org/10.1002/(SICI)1097-0290(19991005)65:1<54::AID-BIT7>3.0.CO;2-R.
[53] Bauer Jasmin, Klamt Steffen. Optmsp: a toolbox for designing optimal multi-stage (bio) processes. J Biotechnol 2024;383:94–102.
[54] Pinto José, de Azevedo Cristiana Rodrigues, Oliveira Rui, von Stosch Moritz. A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development. Bioprocess Biosyst Eng 2019;42:1853–65.
[55] Raj Kaushik, Venayak Naveen, Mahadevan Radhakrishnan. Novel two-stage processes for optimal chemical production in microbes. Metab Eng November 2020;62:186–97. https://doi.org/10.1016/j.ymben.2020.08.006. https://www.sciencedirect.com/science/article/pii/S1096717620301269.
[56] Coltman Benjamin Luke, Rebnegger Corinna, Gasser Brigitte, Zanghellini Jürgen. Characterising the metabolic rewiring of extremely slow growing komagataella phaffii. Microb Biotechnol 2024;17(1):e14386.
[57] Aristidou Aristos A, San Ka-Yiu, Bennett George N. Improvement of biomass yield and recombinant gene expression in escherichia coli by using fructose as the primary carbon source. Biotechnol Prog 1999;15(1):140–5.
[58] Xu Bo, Jahic Mehmedalija, Enfors Sven-Olof. Modeling of overflow metabolism in batch and fed-batch cultures of escherichiacoli. Biotechnol Prog 1999;15(1):81–90.
[59] Müller Don Fabian, Wibbing Daniel, Herwig Christoph, Kager Julian. Simultaneous real-time estimation of maximum substrate uptake capacity and yield coefficient in induced microbial cultures. Comput Chem Eng May 2023;173:108203. https://doi.org/10.1016/j.compchemeng.2023.108203. Publisher: Pergamon. https://www.sciencedirect.com/science/article/pii/S0098135423000728.
[60] Abt Vinzenz, Barz Tilman, Cruz-Bournazou Mariano Nicolas, Herwig Christoph, Kroll Paul, Möller Johannes, et al. Model-based tools for optimal experiments in bioprocess engineering. Curr Opin Chem Eng 2018;22:244–52.
[61] Kager Julian, Herwig Christoph. Monte Carlo-based error propagation for a more reliable regression analysis across specific rates in bioprocesses. Bioengineering November 2021;8(11):160. https://doi.org/10.3390/bioengineering8110160. https://www.mdpi.com/2306-5354/8/11/160. Publisher: Multidisciplinary Digital Publishing Institute.
[62] Slouka Christoph, Kopp Julian, Strohmer Daniel, Kager Julian, Spadiut Oliver, Herwig Christoph. Monitoring and control strategies for inclusion body production in E. coli based on glycerol consumption. J Biotechnol April 2019;296:75–82. https://doi.org/10.1016/j.jbiotec.2019.03.014. https://www.sciencedirect.com/science/article/pii/S0168165619300951.
[63] Zhuang Kai, Yang Laurence, Cluett William R, Mahadevan Radhakrishnan. Dynamic strain scanning optimization: an efficient strain design strategy for balanced yield, titer, and productivity. dyssco strategy for strain design. BMC Biotechnol 2013;13:1–15.
[64] Buckland ST. Monte Carlo confidence intervals. Biometrics 1984;40(3):811–7. https://doi.org/10.2307/2530926. https://www.jstor.org/stable/2530926. Publisher: International Biometric Society.
[65] Krausch Niels, Barz Tilman, Sawatzki Annina, Gruber Mathis, Kamel Sarah, Neubauer Peter, et al. Simulations for the analysis of non-linear parameter confidence intervals in optimal experimental design. Front Bioeng Biotechnol May 2019;7. https://doi.org/10.3389/fbioe.2019.00122. https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2019.00122/full. Publisher: Frontiers.
[66] Preacher Kristopher J, Selig James P. Advantages of Monte Carlo confidence intervals for indirect effects. Commun Methods Meas April 2012;6(2):77–98. https://doi.org/10.1080/19312458.2012.679848. Publisher: Routledge.
[67] Schenkendorf René, Xie Xiangzhong, Rehbein Moritz, Scholl Stephan, Krewer Ulrike. The impact of global sensitivities and design measures in model-based optimal experimental design. Processes April 2018;6(4):27. https://doi.org/10.3390/pr6040027. https://www.mdpi.com/2227-9717/6/4/27. Number: 4. Publisher: Multidisciplinary Digital Publishing Institute.
[68] Franceschini Gaia, Macchietto Sandro. Model-based design of experiments for parameter precision: state of the art. Chem Eng Sci October 2008;63(19):4846–72. https://doi.org/10.1016/j.ces.2007.11.034. https://www.sciencedirect.com/science/article/pii/S0009250907008871.