Development of a Dynamic Gray-Box Model of a Fermentation Process for Spore Production

Joschka Winz*, Supasuda Assawajaruwan, and Sebastian Engell

DOI: 10.1002/cite.202200237

Fermentation processes are difficult to describe using purely mechanistic relations as the underlying biochemical phenomena are complex and often not fully understood. In order to cope with this challenge, we developed an approach to augment standard dynamic model equations by data-based components that are fitted to data using machine learning techniques, which results in dynamic gray-box models. This methodology is applied here to the batch fermentation process of the sporulating bacterium *Bacillus subtilis*, using experimental data from a lab-scale fermenter. The key step in developing the model is the estimation of a training set for the machine learning submodels. The quality of the resulting model is analyzed, and the predictions are compared with real data.

Keywords: Dynamic modeling, Fermentation, Gray-Box Model, Machine learning, Sporulation

Received: December 16, 2022; *revised:* March 29, 2023; *accepted:* May 04, 2023

1 Introduction

Dynamic models of biochemical processes are needed to optimize the execution of the fermentation batches, to develop soft sensors, which estimate key variables that are not directly accessible by measurements, and to improve the control of the fermentations by advanced control. Different approaches to developing such models have long been discussed in the scientific literature. Early work was based on macroscopic balances of the main species (substrates, biomass, products) and on using simple formal kinetics as the Monod, Teissier, and Haldane kinetics as described in [1]. More detailed models consider the concentrations of a suitably chosen subset of the metabolites that are present within the cells, e.g. [2] for the process considered in this work. For the development of such models, knowledge of the metabolic reaction network within the cell is necessary, and the resulting models are of high dimension and may contain a large number of parameters and be therefore difficult to fit to data.

Since on the one hand the traditional models with standard pseudo-kinetics are not accurate enough and on the other hand to fit metabolic models is not feasible in a realistic setting because of lack of knowledge and data, interest emerged in combining the simple mechanistic approaches and data-based approaches to formulate gray-box or hybrid models of biological processes. Psichogios and Ungar presented an early application of combing mass balance equations with artificial neural networks (ANNs) for a fed-batch bioreactor [3]. Thompson and Kramer gave an overview of prior knowledge that can be combined with machine learning (ML) models and developed a gray-box model for the production of penicillin [4]. A new algorithm for training radial basis function (RBF) networks for describing the dynamics of a mammalian cell culture using a gray-box model was applied by Graefe et al. [5]. Wang et al. developed a gray-box model using a novel kind of support vector machine also for the penicillin production fermentation [6]. Other authors refer to models of this kind as hybrid or hybrid semi-parametric models [7].

More recently, the combination of information about the metabolic network with a kinetic gray-box model was proposed by Hebing et al. based upon the elementary modes (EM) of the metabolic network [8, 9]. The evolution of the EM was first parameterized over time along the batch and then kinetic expressions were fitted to this evolution. In a similar spirit, de Prada et al. [10] developed a gray-box model for the ABE process first estimating the temporal evolution of the growth rate and then fitting kinetic expressions from a broad set of functions using global nonlinear optimization.

In this work, a general recipe to develop a dynamic gray-box model is proposed, building upon the work in [11]. The approach is illustrated by the application to the fermenta-

[1]Joschka Winz https://orcid.org/0000-0002-1191-9675 (joschka.winz@tu-dortmund.de), [2]Supasuda Assawajaruwan, [1]Sebastian Engell

[1]TU Dortmund, Process Dynamics and Operations Group, Department of Biochemical and Chemical Engineering, Emil-Figge-Straße 70, 44227 Dortmund, Germany.

[2]Evonik Operations GmbH, Rodenbacher Chaussee 4, 63457 Hanau, Germany.

tion process to produce spores of *Bacillus subtilis*. The paper is structured as follows: Sect. 2 describes the process under consideration, the available measurements and their uncertainties. In Sect. 3, the choice of the structure of the dynamic model is discussed. The modeling methodology is presented in Sect. 4. Subsequently in Sect. 5 and 6, the model elements that describe the growth process and the sporulation are detailed and results on the prediction of the course of fermentation batches are shown. In Sect. 7, conclusions and an outlook on further research work are presented.

2 Fermentation of the Sporulating Bacterium *B. subtilis*

2.1 Sporulation

The term sporulation describes the process where a vegetative cell transitions into the form of an endospore, or spore for short. *B. subtilis* undergoes sporulation when faced with unfavorable conditions as, e.g., the depletion of nutrients like glucose. The spore represents a viable way for the long-term survival and storage of cells. Sporulation is a morphological change of the cells that takes place in several so-called sporulation stages; 6 to 8 stages have been identified, see [12].

2.2 Batch Process Procedure

In this section, the process of the fermentation of *B. subtilis*, strain 168, is described. The goal of the optimization of the operation of the process is to produce a maximum amount of mature spores. To this end, the number of vegetative cells first has to be maximized. Then, in order to make the cells commence sporulation, stress has to be invoked by no longer feeding additional substrate, causing substrate depletion.

The considered batch fermenter is a Sartorius system with a volume of 2 L and 1 L working volume. The degrees of freedom of this batch process are the initial concentrations in the fermentation medium, the batch temperature, the pH value and the oxygen concentration in the fermenter. If more substrate is present initially, more vegetative cells will be produced but the growth phase may take longer. The temperature, the pH value and the oxygen concentration can be controlled using a heating/cooling jacket, an acid/base injection and by adjusting the stirrer speed. In this case, the decision was taken to not control the pH value, instead it was set to 7.5 initially and then auto-regulated by the cells. The dissolved oxygen concentration is initially at the saturation value. During the later stages of the batch, it is reduced, controlling the partial pressure of oxygen to 30 % of the saturation pressure by adjusting the stirrer speed and the aeration rate in a cascade control.

Multiple batches were run with varying values of the degrees of freedom to provide a data base for model development. The initial concentration of substrate was set to 5, 12.5, 20 and 25 $g\,L^{-1}$. The temperature was varied from 25 to 45 °C. In some batches, the temperature was constant over time, in others, steps in the temperature were realized. In total 27 batches with different recipes were run that can be used for developing the dynamic model.

2.3 Offline Measurements

To better understand the dynamics of this fermentation process, several offline measurements were collected. The glucose concentration, denoted as S, is measured either by a HPLC or by a membrane measurement device. For quantifying the concentration of the cells, samples from the fermenter are diluted and cultivated in a Petri dish. An automatic counting device is used to count the number of colonies that have formed. This procedure, also called CFU determination, yields information on all cells that can form a colony. This can include cells in all stages, vegetative cells and mature spores as well as unstable spores. This measurement of the total cell concentration is denoted as X_t.

The amount of mature spores X_s is quantified in a similar way to X_t, with the difference that an autoclave step is performed before the cultivation. The idea is that heating the sample to 80 °C for 15 min eliminates all vegetative cells and thus only the mature spores survive. However, there is no measurement of the amount of intermediate spores available. The quantification of X_t and X_s involves many manual steps and has a limited resolution as only a discrete number of colonies can be counted. Thus, these measurements are prone to statistical errors. To reduce this error, each analysis was conducted three times and the mean value was taken as the measurement. From comparing repeated measurements, the substrate measurement error σ_s is estimated to have a constant value of 0.0449 $g\,L^{-1}$. The error of the total concentration of cells is assumed to be proportional to the mean value and is estimated as 17.0 %. It is denoted as $\hat{\alpha}_t$. For the sporulated cells, the relative error $\hat{\alpha}_s$ is estimated as 18.25 %.

A typical set of experimental data for one batch is shown in Fig. 1. As can be seen, the batch was operated for 48 h, which is the standard operating procedure for all batches. Three phases can be identified. First, in the growth phase the cells consume all the available glucose and grow more or less exponentially. This phase ends when the substrate concentration S tends to zero. Since the substrate concentration is only measured a few times during the batch, it is difficult to determine exactly when this phase is over. Therefore, to define the end of the growth phase, the point in time when the pH value attains a minimum is utilized, in this case after 5.46 h. The pH value decreases during the growth phase due to the production of CO_2.

After the growth phase the sporulation phase follows. Here the bacteria react to the stress resulting from substrate depletion by beginning the sporulation process. This phase

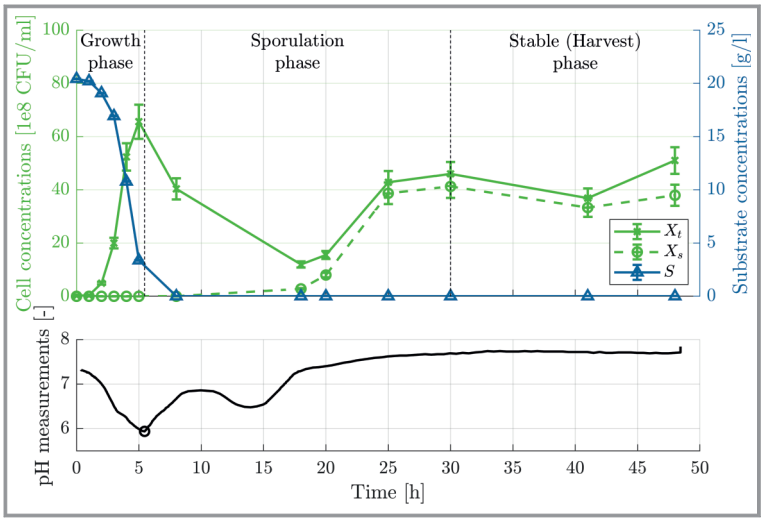

Figure 1. Experimental results of a fermentation batch. Top: evolution of the offline measurements, crosses, circles, triangles indicate the concentrations of total cells, sporulated cells and substrate. Bottom: evolution of the pH value, with a circle marking the lowest observed pH value.

is characterized by a decrease in measurement of the total concentration of cells X_t. After some time, mature spores X_s emerge. When the concentration of the mature spores stabilizes, this phase is over. For the batch shown in Fig. 1, this is the case at approximately 30 h. In the last phase, the process is in a steady state and will remain there as long as no further substrate is fed.

The following sections describe the procedure that we followed to develop a dynamic model of this process. One intended use of the dynamic model is the optimization of the operating conditions. Another possible application of model is process monitoring. A state estimation routine can be applied to combine noisy measurements with a dynamic model to give an operator better feedback about the state of the process at any point in time [11].

3 Structure of the Dynamic Model

The dynamic model is set up as a state-space model, i.e., it consists of coupled ordinary differential equations of first order. The state vector comprises the concentrations of vegetative cells, X_v, and of fully sporulated cells, X_s, and the substrate concentration S. To model the different sporulation stages that the cells undergo, additional states are introduced that indicate the concentrations of unstable or intermediate spores in stage i, $X_{u,i}$. In total n_{st} stages are considered with X_s being the stage of mature spores. For ease of notation, we introduce $n_{Xu} = n_{st}-1$ as the number of unstable spore species considered. The only dynamic input of the process is the temperature T. Thus, the dynamic state vector $\mathbf{x}$ and the input vector $\mathbf{u}$ read as:

$$\mathbf{x} = \left[X_v, X_s, S, X_{u,1}, \ldots, X_{u,n_{Xu}}\right]^T \in \mathbb{R}^{n_{Xu}+3} \tag{1}$$

$$\mathbf{u} = [T] \in \mathbb{R}^1 \tag{2}$$

The biochemical reactions that determine the dynamics of the fermentation system are the growth rate r_g, and the sporulation rate r_s along with the speeds of the transitions of the unstable spores from one unstable stage to the next, $r_{u,i}$. The resulting dynamic system then is of the form (3).

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{X}_v \\ \dot{X}_s \\ \dot{S} \\ \dot{X}_{u,1} \\ \cdots \\ \dot{X}_{u,n_{Xu}} \end{bmatrix} = \mathbf{f}_\phi(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} r_g - r_s \\ r_{u,n_{Xu}} \\ -r_g Y_g^{-1} \\ \eta r_s - r_{u,1} \\ \cdots \\ r_{u,n_{Xu}-1} - r_{u,n_{Xu}} \end{bmatrix} \tag{3}$$

Here $Y_g$is the substrate yield and η denotes the sporulation efficiency, which is only considered for the first spore transition with rate r_s. One cannot be sure that the other sporulation reactions $r_{u,i}$ have an efficiency of 100 %, but it is not possible to determine at which stage the cells do not survive the transition as no measurements of any of the variables $X_{u,i}$ are available. $\mathbf{f}$ denotes the vector of differential equations, and ϕ is used to refer to all quantities that determine the dynamics as, e.g., Y_g and η. The main issue in completing this model is to find suitable expressions for the reaction rates r_g, r_s, $r_{u,1}, \ldots, r_{u,n_{Xu}}$. This is discussed in the next sections.

4 Methodology of Dynamic Gray-Box Modeling

The identification of a dynamic gray-box model with embedded data-based submodels is a challenging task. After defining the model structure and selecting the variables that are described by data-based models, a model structure for these submodels has to be chosen. Not only is there a broad variety of data-based model structures, also the corresponding parameter estimation problem is difficult to solve if no good initial values are available. Besides that, it is not a priori clear which states or input variables should be used as inputs for each of the data-based submodels.

In dynamic gray-box modeling, the estimation of suitable training sets containing input and output values of the data-based submodels in a first step has been acknowledged as an effective and efficient way to tackle the issues of structure selection of the submodels and computational complexity of the parameter estimation. This is because once there

is input-output-data for the submodels available, conventional input selection and model structure selection methods and parameter estimation approaches can be applied without having to consider the dynamics of the process. Using information about the metabolites of a cell, Hebing et al. [9], estimate the values of the rates of cell internal reactions to analyze the suitability of a chosen set of reaction pathways for subsequent kinetic model development. This is based on the methodology of Leighty and Antoniewicz [13, 14] called dynamic metabolic flux analysis (DMFA). Scheffold et al. [15] estimate the training sets using a state observer and de Prada et al. [10] use an optimization-based approach.

A related approach has been developed by Brendel et al. [16], who propose to approach the modeling of reaction rates in a chemical system with known stoichiometry and a large number of noisy measurements in consecutive steps. The first step is the estimation of the reaction fluxes. The estimation of the reaction fluxes is conducted using smoothing splines, such that the estimated fluxes can be used to identify kinetic models.

In our previous work, we extended the optimization-based approaches to nonlinear systems and analyzed the effect of the type and of the extent of regularization in detail and proposed methods for model structure selection [17, 18]. An adapted variant of this methodology is applied in this work and is visualized in Fig. 2.

The first step is the derivation of the first principles-based model equations. The next step is to analyze these model equations and to select the expressions that will be modeled using data-based approaches, here also denoted as embedded variables. Once these variables have been specified, a training set is estimated. For variables that are changing dynamically and where one expects to have rich enough measurements to describe their values at each point in time, the training set estimation is done dynamically. This means that we try to determine a trajectory over time of the embedded variable that leads to a good model fit to the measured data. For variables for which sufficient data is not available to estimate time-varying values, only one constant value over the whole batch is identified.

In the fourth step, the estimated training data is used to find a set of input variables that correlate with each of the embedded variables. With this set of inputs and the estimated values of the embedded variables as outputs, standard regression problems can be solved relatively easily for multiple data-based model structures in a trial-and-error fashion. Thus, the result of this step is a suitable data-based model structure and a set of estimated parameters of the data-based model. These parameters are then used as initial values to solve a full dynamic parameter estimation problem where all degrees of freedom are varied to find the best values of the parameters of the data-based submodels as well as of the biological and physico-chemical parameters as, e.g., the yield coefficient. The approach used here exploits the presence of subsequent phases in the batch that are handled independently. Additionally, it deals with the situation that for some embedded variables there are not sufficient measurements to explain the evolution of the variable over time.

In the following sections, this methodology is applied to the process at hand.

5 Identification of a Gray-Box Model for the Growth Phase

5.1 Estimation of the Growth Rate

In the growth phase, only the growth reaction rate r_g is of interest, as sporulation does not occur. The only dynamic states that depend on r_g are X_v and S. Since the concentration of vegetative cells evolves over multiple orders of magnitude, it is crucial to include a first-order dependency

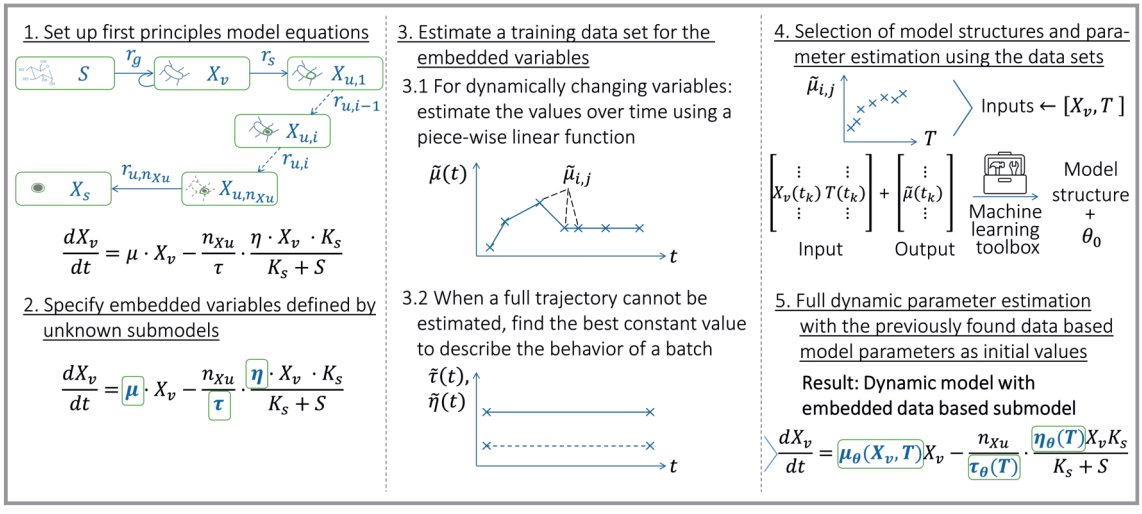

Figure 2. Overview over the steps to identify a dynamic gray-box model.

of r_g on X_V. Thus, the estimation will be conducted for μ, with $r_g = \mu \cdot X_v$. Assuming a continuous piece-wise linear trajectory of μ, denoted $\tilde{\mu}(t)$, as defined in Eq. (4), the dynamic model can be integrated in closed form as shown in Eqs. (5) and (6).

$$\tilde{\mu}(t) = \left(t - t_{i,j}\right)\frac{\tilde{\mu}_{i+1,j} - \tilde{\mu}_{i,j}}{t_{i+1,j} - t_{i,j}} + \tilde{\mu}_{i,j}, \quad t \in \left[t_{i+1,j}, t_{i,j}\right] \tag{4}$$

$$X_v\left(t_{i+1,j}\right) = X_v\left(t_{i,j}\right)\exp\left(\left(t_{i+1,j} - t_{i,j}\right)\frac{\tilde{\mu}_{i+1,j} + \tilde{\mu}_{i,j}}{2}\right) \tag{5}$$

$$S\left(t_{i+1,j}\right) = S\left(t_{i,j}\right) - Y_g^{-1}\left(X_v\left(t_{i+1,j}\right) - X_v\left(t_{i,j}\right)\right) \tag{6}$$

With these values of the concentrations as functions of the trajectory of μ, an estimation problem for the yield coefficient, the initial values and the parameters of the piece-wise linear evolution of μ, denoted $\tilde{\mu}_{i,j}$, can be solved. In this optimization problem the measurements of S and X_t are considered because the total cell concentration X_t in this phase is equal to the concentration of vegetative cells X_v. Besides the parameters of the model of the growth rate, the yield coefficient Y_g is also optimized. This is contrast to some works based on DMFA, where these coefficients are computed from a metabolic network [9, 13]. The parameter estimation problem is formulated as

$$\begin{matrix}\min \\ \tilde{\mu}_{i,j},\quad j = 1 \ldots n_b,\quad i = 1 \ldots n_{s,j} \\ Y_g \\ X_v\left(t_{0,j}\right),\quad S\left(t_{0,j}\right),\quad j = 1 \ldots n_b\end{matrix} \quad MSE_{exp} + \lambda_\mu REG_\mu \tag{7}$$

with

$$MSE_{exp} = \frac{1}{\sum_{j=1}^{n_b} n_{s,j}}\sum_{j=1}^{n_b}\sum_{i=0}^{n_{s,j}}\left(\frac{X_t^{exp}\left(t_{i,j}\right) - X_v\left(t_{i,j}\right)}{\hat{\alpha}_t X_t^{exp}\left(t_{i,j}\right)}\right)^2 + \left(\frac{S^{exp}\left(t_{i,j}\right) - S\left(t_{i,j}\right)}{\sigma_S}\right)^2 \tag{8}$$

and

$$REG_\mu = \frac{1}{\sum_{j=1}^{n_b}\left(n_{s,j} - 1\right)}\sum_{j=1}^{n_b}\sum_{i=0}^{n_{s,j}-1}\left(\frac{\mu_{i+1,j} - \mu_{i,j}}{t_{i+1,j} - t_{i,j}}\right)^2 \tag{9}$$

In Eq. (8) $X_v(t_{i,j})$ and $S(t_{i,j})$ are computed as shown in Eqs. (5) and (6). $\tilde{\mu}_{i,j}$ denotes the value of the relative growth rate at time $t_{i,j}$. λ_μ is a regularization parameter which was determined using the L-curve criterion [19] to be 7. The number of batches is referred to as n_b, the number of samples in batch j as $n_{s,j}$.

5.2 ML Model Fitting for Growth Phase

After the rate estimation step, a kinetic model μ_θ that is parameterized in a vector θ can be fitted to the estimated values of the growth rate. In this case it was found that an ANN model with just three nodes in a single hidden layer using a tanh activation function performs well. The training was done using the Levenberg-Marquardt algorithm with Bayesian regularization. Different sets of inputs were tested. With the inputs $[S,T]$ an MSE of $0.070\,h^{-2}$ was observed. The error decreased significantly when adding the concentration of the vegetative cells to the set of inputs; for the inputs $[S,X_v,T]$ the resulting MSE was $0.041\,h^{-2}$. When removing S from this set, the MSE only increased to $0.044\ h^{-2}$. Because this increase is minor, the finally chosen set of inputs is $[X_v,T]$. Since the growth stops when the substrate has been depleted, the ANN model $\mu_{ANN,\theta}(X_v,T)$ is multiplied by a standard Monod term, giving rise to the final submodel shown in Eq. (10).

$$\mu_\theta(X_v, S, T) = \mu_{ANN,\theta}(X_v, T)\frac{S}{K_s + S} \tag{10}$$

Here K_s denotes the substrate inhibition constant, which was fixed to $0.1\,g\,L^{-1}$.

The vector of parameters of the ANN after this step is denoted as θ_0. After this model was fitted to the estimated values of the rates, a full dynamic parameter estimation problem was solved for the growth phase, formulated as shown in Eqs. (13)–(16). The MSE_{exp} is defined as in Eq. (8), but the values of $X_v(t_{i,j})$ and $S(t_{i,j})$ now cannot be computed explicitly anymore. Instead, they were found by integrating the system using the IDAS solver from the SUNDIALS suite [20]. Constraints based on biological insight were enforced as soft constraints with the penalty term shown in Eqs. (11) and (12), which is enforced on $n_{unlab} = 100$ different combinations of X_v,T distributed in a 10×10 grid. Eq. (12) expresses that the growth rate is not expected to increase with the cell concentration X_v. To ensure that model behaves as can be expected from biological insight, λ_{PC} was chosen to have a high value of $1 \cdot 10^6$.

$$MSE_{PC} = \frac{1}{n_{unlab}}\sum_{j=1}^{n_b} J_{PC}\left(X_{v,unlab,j}, T_{unlab,j}\right) \tag{11}$$

$$J_{PC}\left(X_{v,unlab,j}, T_{unlab,j}\right) = \begin{cases} 0 & \text{if } \dfrac{\partial \mu_{ANN,\theta}}{\partial X_v}\left(X_{v,unlab,j}, T_{unlab,j}\right) < 0 \\ \left(\dfrac{\partial \mu_{ANN,\theta}}{\partial X_v}\left(X_{v,unlab,j}, T_{unlab,j}\right)\right)^2 & \text{else} \end{cases} \tag{12}$$

The optimization was also solved using the Levenberg-Marquardt algorithm, computing the Jacobians using adjoint sensitivities from IDAS and automatic differentiation from CasADi [21]. As the initial values of the ANN param-

eters, the vector $\theta_0$from the previous step is used. As a simplification, the Monod term is assumed to be 1 because in the growth phase the substrate is not depleted leading to $\frac{S}{K_s + S} \approx 1$ for the observed values of S. The parameter estimation problem was formulated as

$$\min_{\substack{\theta, Y_g, X_v(t_{0,j}) \\ S(t_{0,j}), j = 1 \ldots n_b}} MSE_{exp} + \lambda_\theta REG_\theta + \lambda_{PC} MSE_{PC} \quad (13)$$

s.t.

$$\dot{X}_v = \mu_\theta(X_v, T) X_v \quad (14)$$

$$\dot{S} = -\mu_\theta(X_v, T) X_v Y_g^{-1} \quad (15)$$

Here, REG_θ is used to penalize drastic changes in the ANN parameters.

$$REG_\theta = \frac{1}{n_\theta} \sum_{k=1}^{n_\theta} \left(\theta_{0,k} - \theta_k\right)^2 \quad (16)$$

The results of the ANN predictions after the training on the estimated values of the growth rate and after the full dynamic parameter estimation are shown in Fig. 3 for $\lambda_\theta = 25$. The optimized value of the yield coefficient Y_g is $3.15 \cdot 10^{11}$ CFU $g^{-1}_{Glucose}$ while it was $3.796 \cdot 10^{11}$ CFU $g^{-1}_{Glucose}$ after the rate estimation step.

As can be seen from the Fig. 3, the estimated values of μ show a dependency on the temperature and on the concentration of vegetative cells. While a temperature dependency could be expected, it is surprising that the number of cells has such a strong effect. Similar effects were observed by other researchers [22], but it is not entirely clear whether the inhibition is caused by the cell density or just statistically correlated to it. One possible explanation is that the decrease of the pH value (see Fig. 1), which is clearly correlated to the cell concentration, is the true cause of the inhibition. This parameter estimation was also conducted without the biologically motivated constraints, which lead to a minor increase in growth rate at high cell concentrations and low temperatures, where there is no data available. This unwanted phenomenon is prevented by the constraints, however the effect on the model predictions in the relevant part of the input space is negligible.

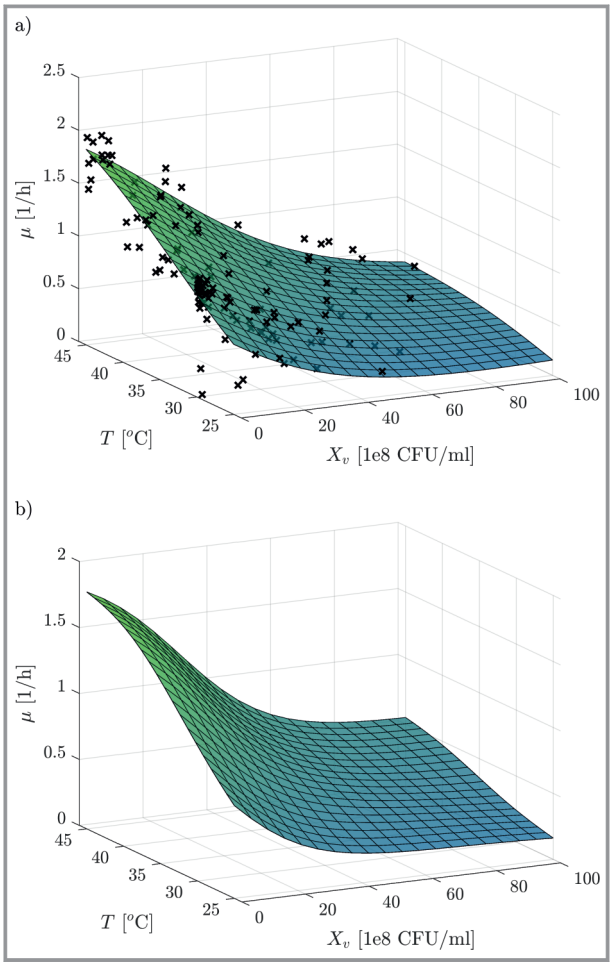

Figure 3. Results of fitting the specific growth rate μ as a function of the temperature and the concentration of vegetative cells X_v. a) Estimated values of the growth rates shown as crosses, ANN approximation after fitting only the estimated rates shown as a surface, b) ANN prediction of the growth rate after the full dynamic parameter estimation with biologically motivated constraints.

6 Identification of the Model for the Sporulation Phase

6.1 Dynamic Model for the Sporulation Phase

There has been some previous work on modeling the dynamics of the sporulation of bacteria. Atehortúa et al. [23] developed a dynamic model for the sporulation that describes the sporulation rate using an empirical exponential kinetics approach that depends on the current and the initial substrate concentrations. They did not consider the delay in the sporulation. This is also neglected in the work of Park et al. [24], who developed a more detailed dynamic model that considers different species in the sporulation cycle explicitly. Das and Sen [25] modeled the sporulation kinetics by approximating the pure delay by a first order Taylor expansion. A different approach to model the delay of the sporulation

was proposed by Gauvry et al. [26] who define a probability of sporulation over time that is used to determine the fraction of cells that have committed to sporulation.

Here, we investigate whether it is possible to model the sporulation process without choosing a kinetic model structure a priori and using a data-based approach to model the effects that are visible in the data. This model takes the form of a finite set of ordinary differential equations. It is similar to an *n*th-order linear dynamic system but additionally there are inhibition terms added to make the response of the have steeper slopes. The reaction rates $r_{u,i}$ are modeled in the following way:

$$r_s = \frac{n_{Xu}}{\tau}\frac{\eta X_v K_s}{K_s + S} \tag{17}$$

$$r_{u,1} = \frac{n_{Xu}}{\tau}\frac{X_{u,1}}{\left(\frac{\eta X_v}{K_y}\right)^{\gamma} + 1} \tag{18}$$

$$r_{u,i} = \frac{n_{Xu}}{\tau}\frac{X_{u,i}}{\left(\frac{X_{u,i-1}}{K_y}\right)^{\gamma} + 1}, \quad i = 2 \ldots n_{Xu} \tag{19}$$

Here, τ is the time constant of each stage, K_y and γ are parameters that determine the shape of the delayed response. A drawback of this formulation is the fact that the dependency of the reaction rate $r_{u,i}$ on the intermediate species $X_{u,i-1}$ can lead to unwanted side effects, when modeling fed-batch systems. This can happen because a high value of one intermediate cell concentration $X_{u,i-1}$ decreases the reaction rate $r_{u,i}$. In a batch system this makes sense because the reaction rate $r_{u,i}$ will be low as long as the reaction $r_{u,i-1}$ is not completed assuming that both occur subsequently. But in a fed-batch system, the reactions do not occur strictly subsequently. In this case a population model could be used to generate the desired delay effect. In a batch experiment this cannot be the case as the intermediate species react successively since the trajectory of the vegetative cells $X_v$is monotonously decreasing because no additional substrate is fed after the end of the growth phase.

Responses of the concentrations of (intermediate) sporulated cells for a sudden increase in the number of vegetative cells at complete substrate depletion are shown in Fig. 4, for $n_{Xu} = 6$. As can be seen, the vegetative cells react quickly to the depletion of the substrate giving rise to a cascade of intermediate spores $X_{u,i}$. Each species of $X_{u,i}$ follows a similar pattern of rise and decline as one would expect in such a system. Towards the end, the mature spores X_s emerge.

6.2 Estimation of the Time Constant of the Sporulation Process

The dynamic model put forward in Sect. 6.1 is characterized by the parameters γ, K_y τ and η. It is not clear a priori on

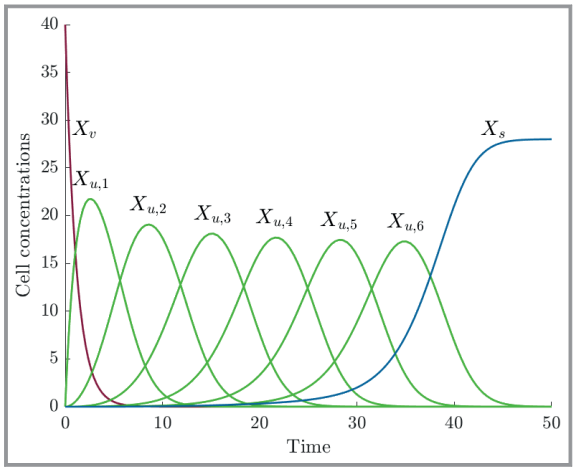

Figure 4. Exemplary step response of the dynamic sporulation model.

what quantities these parameters depend, thus before finding functional relations with e.g. the temperature, first a training set $\tilde{\gamma}_j, \tilde{\tau}_j, \tilde{\eta}_j$ is estimated similar to the approach in Sect. 5, with the difference that here only a single value for each batch is estimated to avoid overfitting. As both γ and K_y describe the sharpness of the response, K_y was fixed to $1 \cdot 10^8\,\mathrm{CFU\,mL^{-1}}$. The values of the other parameters were optimized for each batch separately by solving following optimization problems:

$$\min_{\tilde{\gamma}_j, \tilde{\tau}_j, \tilde{\eta}_j} \sum_{i=i_{s0}}^{n_{s,j}} \left(X_s^{exp}(t_{i,j}) - X_s(t_{i,j})\right)^2 \tag{20}$$

s.t.

$$\dot{\mathbf{x}} = \mathbf{f}_{\phi_j}(\mathbf{x}, \mathbf{u}) \tag{21}$$

$$\phi_j = \left[\mu_\theta(X_v, T), Y_g, \tilde{\gamma}_j, \tilde{\tau}_j, \tilde{\eta}_j\right]^T \tag{22}$$

$$X_v(t_{i_{s0},j}) = X_v(t_{0,j}) + \int_0^{t_{i_{s0},j}} \mu_\theta(X_v, T) X_v \, \mathrm{d}t \tag{23}$$

$$S(t_{i_{s0},j}) = X_{u,1}(t_{i_{s0},j}) = \ldots = X_{u,n_{Xu}}(t_{i_{s0},j}) = X_s(t_{i_{s0},j}) = 0 \tag{24}$$

with $j = 1 \ldots n_b$.

In the objective functions, the deviations between the simulated and the experimental spore concentrations are minimized in the sporulation and in the stable phase. This optimization is subject to the dynamic model, parameterized by the growth model that was estimated before and the decision variables. The initial values for the states are 0 except for the vegetative cell concentration X_v. Here, the initial value is computed by integrating the dynamic system up until the end of the growth phase, i.e., to the point where the substrate is depleted, denoted as $t_{i_{s0},j}$. The yield coeffici-

ent Y_g and the parameters of the growth rate μ_θ are taken from the solution of the optimization problem Eqs. (13)–(16). The results of this optimization are shown in Fig. 5. It should be noted that batches in which the sporulation was not finished after 48 h were not considered for these plots as the resulting parameters are only very rough guesses. The prediction intervals depict the range of values in which a future observation will fall with a probability of 95 %.

In Fig. 5a and 5b it can be seen that both the sporulation delay time τ and the sporulation efficiency η show a temperature dependency that can be described sufficiently accurately using a quadratic model. The minimum delay is predicted to be at around 40 °C, and the maximum efficiency at around 35 °C. Note that the value of the temperature is the average temperature over the sporulation phase. For the exponent γ no clear dependency could be found, even when considering the initial amount of substrate as a degree of freedom, thus it has to be considered as a constant here, to avoid assuming a wrong dependency, which could lead to an unreliable generalization and unreliable optimization results.

Overall, there is a large variation from batch to batch, even under identical conditions. Thus, there seems to be a significant statistical fluctuation that cannot be captured by a deterministic model. This is visualized by the prediction intervals PI. These are computed at a sample input x_0 as shown in Eqs. (25)–(27), which are taken from [27].

$$\mathrm{PI} = \left[x_0^T\hat{\beta} - \mathrm{PIHW}, x_0^T\hat{\beta} + \mathrm{PIHW}\right] \tag{25}$$

$$\mathrm{PIHW} = t_{n_{data}-n_{para}}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}\left(1 + x_0^T\left(\underline{\mathbf{X}}^T\underline{\mathbf{X}}\right)^{-1}x_0\right)^{0.5} \tag{26}$$

The vector of parameters of the assumed linear model is represented by $\hat{\beta}$. It contains the n_{para} parameters that were fit to the n_{data} data points. The data matrix is referred to as $\underline{\mathbf{X}}$. $t_{n_{data}-n_{para}}\left(1-\frac{\alpha}{2}\right)$ is the critical value of the Student's t distribution with (n_{data} – n_{para}) degrees of freedom and a significance level of 1–α/2.

$\hat{\sigma}$ denotes the variance in the data, which is estimated from the deviation to the model prediction as shown in Eq. (27).

$$\hat{\sigma} = \sqrt{\frac{\mathbf{r}^T\mathbf{r}}{n_{data} - n_{para}}} \tag{27}$$

Here, the vector of residuals is denoted as **r**.

Finally, the submodels for $\tau(T)$ and $\eta(T)$ were fine-tuned by solving the optimization problem in Eqs. (20)–(24) again but accumulating the cost function for all batches and varying the parameters of the submodels $\tau(T)$ and $\eta(T)$ together with γ and K_γ.

6.3 Contribution of the Intermediate Cells to the Concentration of Total Cells

The developed dynamic model describes the growth and the sporulation dynamics sufficiently accurately for applications like batch trajectory optimization and soft sensing. But one phenomenon is not taken into account in the model: the fact that the measurements of the total cell concentration X_t decrease before the rise of the spore concentration. This can be seen in Fig. 1 for the measurements at 17 and 20 h. Biologically it is not clear why this occurs, especially given that other researchers did not observe this effect in similar studies [23–26, 28, 29]. One hypothesis is that a part of the cells die after the depletion of the substrate and a part of the cells can establish another phase of growth with stored glucose or cell cannibalism. This explanation is not in line with the fact that the total cell concentration X_t increases simultaneously with the mature spore concentration X_s. One would expect another sporulation delay and thus a two-step increase of the spore concentration. Therefore, another hypothesis is formulated: the intermediate cells contribute only in a reduced manner to the total number of cells, for example due to agglutination, which leads to

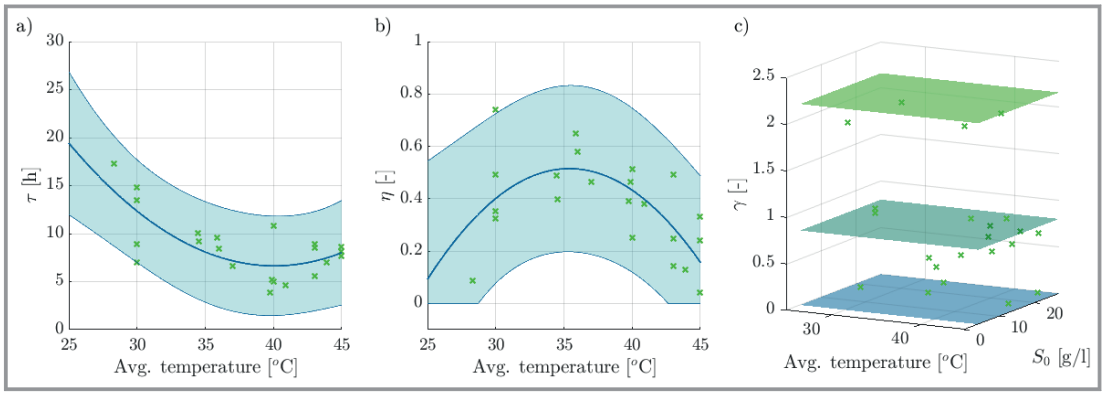

Figure 5. Results of the estimation of the parameters of the sporulation model for each batch separately. Estimated values are shown as crosses and the resulting model as a line. The 95 % prediction intervals are shown as shaded areas for a) delay time, b) sporulation efficiency. In c), the sporulation exponent is shown, the surface in the middle shows the mean value, the upper and lower surfaces indicate the 95 % prediction interval.

less colonies forming in the analytics. This can be modeled as

$$X_t = X_v + X_s + \zeta \sum_{i=1}^{n_{Xu}} X_{u,i} \tag{28}$$

Here the factor $\zeta \in [0, 1]$ represents the fraction with which the intermediate cells contribute to X_t. In the case considered here it is regressed to a value of 0.41.

6.4 Accuracy of the Full Model

The full model consists of the differential equations shown in Eq. (3) with the reaction rates as depicted in Eqs. (17)–(19) and $r_g = \mu_\theta \cdot X_v$. The outputs are X_v, S and X_t as described in Eq. (28). To analyze the prediction accuracy of the model, the model predictions for the batch used in Fig. 1 and an additional batch are shown in Fig. 6.

As can be seen, the model represents the measurements well for the considered batches. The temperature in the batch shown at the top of Fig. 6 is initially 35 °C and is changed to 40 °C after 6 h. For the batch in the bottom the temperature is 30 °C throughout the batch. Additionally, the initial substrate concentrations are different as can be seen in the figure. As the interest is mostly on predicting the concentration of the sporulated cells from the perspective of the intended production of spores, only the prediction interval for the sporulated cells is visualized. This is done by evaluating the extreme cases of the prediction interval shown in Fig. 5. The broad prediction interval is a result of the fact that the process varies statistically, indeed the same operating conditions of a batch did not result in the same measurements. This is also visible in results shown in related work [28].

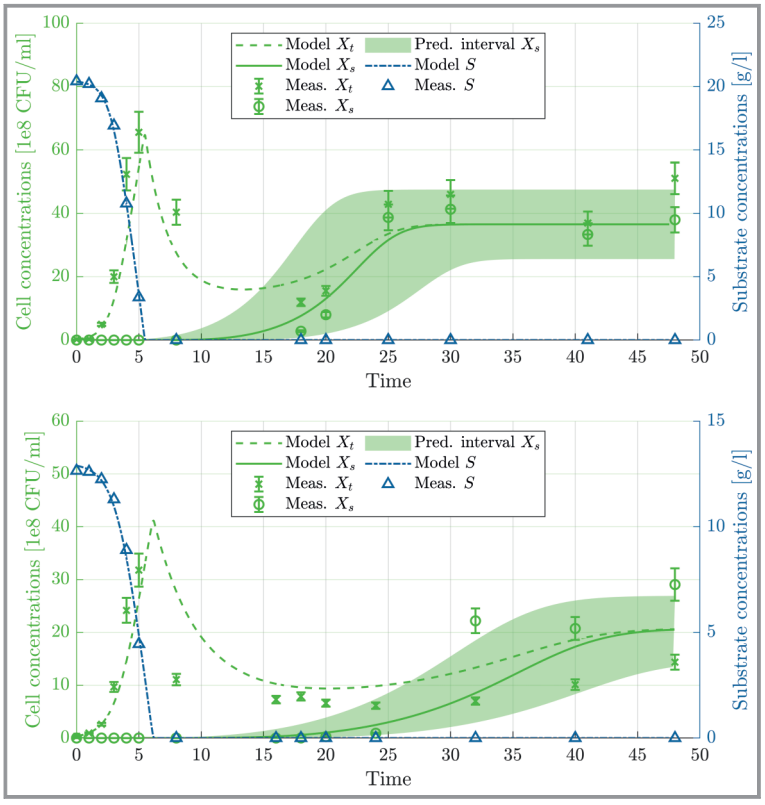

Figure 6. Comparison of model predictions with actual measurements for two batches. Crosses, circles and triangles show measurements of the concentration of all cells, sporulated cells and substrate. Dashed, full, and dash-dotted lines represent the model predictions of these concentrations. The shaded area is an estimate of the 95 % prediction interval of the model of the sporulation process.

Despite the broad prediction interval, which is mainly a consequence of the variability of the process itself, the model describes the dynamics sufficiently well and can be used to investigate the expected influence of process parameters on the batch operation. The tradeoff in determining an optimal temperature is visible for example in the submodels of τ and η, which show different optima with respect to temperature. The dynamic model provides a basis for optimizing the batch recipe while considering the broad range in which the measurements can lie. Due to the amount of noise present in this process it is difficult to reliably determine the process conditions that have the highest expected value of some optimization objective. Therefore, one can use a deterministic dynamic model to represent this expected value and optimize over it. Additionally, it can be used during the operation of the plant to monitor the process.

7 Conclusion

In this work, a hybrid dynamic model for the growth and sporulation of *B. subtilis* cells was developed on the basis of experiments over a range of values of the temperature over the batch and of the initial substrate concentration. Since this process is difficult to describe based on first principles only, a gray-box model structure was identified. This was done by first estimating values of the specific growth rate over time, which then provide the data for fitting a ML model of the growth rate as a function of the process states. A similar approach was used for the sporulation delay time and for the sporulation efficiency, which were both modeled as functions of temperature. What could be observed is that the sporulation is a statistical process that varies strongly even for identical process conditions. Nonetheless the model provides a reliable estimate of the expected trajectory given a recipe, which provides the basis for an optimization of the batch recipe. Future work will deal with the question of how an online state and parameter estimator can be used to address the problem of the statistical variation of the

behavior of the cells in the growth and sporulation phases so that the prediction of the end-time of the batch can be improved and the operation can be stopped at an earlier point in time to improve the productivity of the process.

Acknowledgment

This research has been supported by the project "KI-Inkubator-Labore in der Prozessindustrie – KEEN", funded by the Bundesministerium für Wirtschaft und Klimaschutz (BMWK) under grant number 01MK20014T. This support is gratefully acknowledged. Open access funding enabled and organized by Projekt DEAL.

Symbols used

$\mathbf{f}$	[various]	Differential equations
K	[CFU mL^{-1}, g L^{-1}]	Inhibition constant
n_b	[–]	Number of batches
$n_{s,j}$	[–]	Number of samples in batch j
r	[CFU mL^{-1}h^{-1}]	Reaction rate
$\mathbf{r}$	[various]	Residual vector
S	[g L^{-1}]	Substrate concentration
t	[h]	Time
$\mathbf{u}$	[°C]	Input variables
X	[CFU mL^{-1}]	Cell concentration
$\underline{\mathbf{X}}$	[various]	Data matrix
$\mathbf{x}$	[various]	State variables
x_0	[various]	Sample input
Y	[CFU g^{-1}]	Yield coefficient

Greek letters

$\hat{\alpha}$	[–]	Relative measurement error
$\hat{\beta}$	[various]	Parameter vector of submodel
γ	[–]	Exponent parameter
ζ	[–]	Factor for reduction in measurement
η	[–]	Sporulation efficiency
θ	[various]	Parameters
λ	[–]	Regularization constant
μ	[h^{-1}]	Specific growth rate
$\tilde{\mu}$	[h^{-1}]	Specific growth rate as function of time
σ	[g L^{-1}]	Absolute measurement error
τ	[h]	Sporulation time constant
ϕ	[various]	Parameters

Sub- and Superscripts

exp	Experimental
g	Growth
S	Sporulated
t	Total
u, I	Unstable spore of stage *i*
v	Vegetative

Abbreviations

MSE	Mean squared error
PI	Prediction interval
PIHW	Prediction interval half width
REG	Regularization

References

[1] G. Bastin, D. Dochain, *On-Line Estimation and Adaptive Control of Bioreactors*, Elsevier, Amsterdam **1990**.
[2] J. W. Jeong, J. Snay, M. M. Ataai, *Biotechnol. Bioeng.* **1990**, *35 (2)*, 160–184. DOI: https://doi.org/10.1002/bit.260350208
[3] D. C. Psichogios, L. H. Ungar, *AIChE J.* **1992**, *38 (10)*, 1499–1511. DOI: https://doi.org/10.1002/aic.690381003
[4] M. L. Thompson, M. A. Kramer, *AIChE J.* **1994**, *40 (8)*, 1328–1340. DOI: https://doi.org/10.1002/aic.690400806
[5] J. Graefe, P. Bogaerts, J. Castillo, M. Cherlet, J. Wérenne, P. Marenbach, R. Hanus, *Bioprocess Eng.* **1999**, *21 (5)*, 423. DOI: https://doi.org/10.1007/s004490050697
[6] X. Wang, J. Chen, C. Liu, F. Pan, *Chem. Eng. Res. Des.* **2010**, *88 (4)*, 415–420. DOI: https://doi.org/10.1016/j.cherd.2009.08.010
[7] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, *Comput. Chem. Eng.* **2014**, *60*, 86–101. DOI: https://doi.org/10.1016/j.compchemeng.2013.08.008
[8] L. Hebing, T. Neymann, T. Thüte, A. Jockwer, S. Engell, *IFAC-PapersOnLine* **2016**, *49 (7)*, 621–626. DOI: https://doi.org/10.1016/j.ifacol.2016.07.237
[9] L. Hebing, T. Neymann, S. Engell, *Biotechnol. Bioeng.* **2020**, *117 (7)*, 2058–2073. DOI: https://doi.org/10.1002/BIT.27340
[10] C. de Prada, D. Hose, G. Gutierrez, J. L. Pitarch, *IFAC-PapersOnLine* **2018**, *51 (2)*, 523–528. DOI: https://doi.org/10.1016/J.IFACOL.2018.03.088
[11] L. Hebing, F. Tran, H. Brandt, S. Engell, *Ind. Eng. Chem. Res.* **2020**, *59 (6)*, 2566–2580. DOI: https://doi.org/10.1021/acs.iecr.9b05504
[12] P. J. Piggot, J. G. Coote, *Bacteriol. Rev.* **1976**, *40*, 55.
[13] M. R. Antoniewicz, *Curr. Opin. Biotechnol.* **2013**, *24 (6)*, 973–978. DOI: https://doi.org/10.1016/j.copbio.2013.03.018
[14] R. W. Leighty, M. R. Antoniewicz, *Metab. Eng.* **2011**, *13 (6)*, 745–755. DOI: https://doi.org/10.1016/j.ymben.2011.09.010
[15] L. Scheffold, T. Finkler, U. Piechottka, *Comput. Chem. Eng.* **2021**, *146*, 107204. DOI: https://doi.org/10.1016/J.COMPCHEMENG.2020.107204
[16] M. Brendel, D. Bonvin, W. Marquardt, *Chem. Eng. Sci.* **2006**, *61 (16)*, 5404–5420. DOI: https://doi.org/10.1016/j.ces.2006.04.028
[17] J. Winz, S. Engell, *Comput.-Aided Chem. Eng.* **2022**, *51*, 1483–1488. DOI: https://doi.org/10.1016/B978-0-323-95879-0.50248-4
[18] J. Winz, S. Engell, *IFAC-PapersOnLine* **2022**, *55 (7)*, 86–93. DOI: https://doi.org/10.1016/j.ifacol.2022.07.426
[19] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia **1998**.
[20] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, C. S. Woodward, *ACM Trans. Math. Software* **2005**, *31 (3)*, 363396.

[21] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, M. Diehl, *Math. Prog. Comput.* **2019**, *11 (1)*, 1–36. DOI: https://doi.org/10.1007/s12532-018-0139-4
[22] D. E. Y. 1959 Contois, *J. Gen. Microbiol.* **1959**, *21 (1)*, 40–50. DOI: https://doi.org/10.1099/00221287-21-1-40
[23] P. Atehortúa, H. Álvarez, S. Orduz, *Bioprocess Biosyst. Eng.* **2007**, *30 (6)*, 447–456. DOI: https://doi.org/10.1007/s00449-007-0141-0
[24] S. Park, B. E. Rittmann, W. Bae, *Biotechnol. Bioeng.* **2009**, *104 (5)*, 1012–1024. DOI: https://doi.org/10.1002/bit.22456
[25] S. Das, R. Sen, *Bioresour. Technol.* **2011**, *102 (20)*, 9659–9667.
[26] E. Gauvry, A.-G. Mathot, O. Couvert, I. Leguérinel, M. Jules, L. Coroller, *Appl. Environ. Microbiol.* **2019**, *85 (10)*, e00322–19. DOI: https://doi.org/10.1128/AEM.00322-19
[27] L. Fahrmeir, T. Kneib, S. Lang, B. Marx, *Regression: Modelle, Methoden und Anwendungen*, Springer, Heidelberg **2013**.
[28] E. Gauvry, A.-G. Mathot, O. Couvert, I. Leguérinel, L. Coroller, *Int. J. Food Microbiol.* **2021**, *337*, 108915. DOI: https://doi.org/10.1016/j.ijfoodmicro.2020.108915
[29] E. Baril, L. Coroller, O. Couvert, M. El Jabri, I. Leguerinel, F. Postollec, C. Boulais, F. Carlin, P. Mafart, *Food Microbiol.* **2012**, *32 (1)*, 79–86. DOI: https://doi.org/10.1016/j.fm.2012.04.011

