Automatic Generation of Process Models for Fed-Batch Fermentations Based on the Detection of Biological Phenomena

Sebastian Herold

Automatic Generation of Process Models for Fed-Batch Fermentations Based on the Detection of Biological Phenomena

vorgelegt von
Dipl.-Ing.
Sebastian Herold
geb. in Wernigerode

von der Fakultät III – Prozesswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Matthias Kraume
Gutachter: Prof. Dr.-Ing. Rudibert King
Gutachter: Prof. Dr.-Ing. Achim Kienle

Tag der wissenschaftlichen Aussprache: 19. Januar 2015

Berlin 2015

Impressum

Verlag: epubli GmbH, Berlin, www.epubli.de

ISBN 978-3-7375-2957-0

Acknowledgment

The present thesis is the result of research conducted at the *Chair of Measurement and Control* at the *Technische Universität Berlin*.

First, I want to express my gratitude to the head of the chair, Prof. Dr.-Ing. Rudibert King. I certainly would not have finished my thesis successfully without his guidance, understanding, and patience. It really has been a pleasure to work with him.

Then, I thank Prof. Dr.-Ing. Achim Kienle for his willingness to review and evaluate this thesis and Prof. Dr.-Ing. Matthias Kraume for chairing the doctoral committee.

I am also grateful to my co-workers, both former and current ones, for creating a friendly working environment. My thanks go to all the members of the “Biogruppe” for their support, advice, and many fruitful discussions. Special thanks go to Norman Violet whose support helped me a lot during the last years.

I owe a very important dept to Angeline Aow who took on the very time-consuming task of proof-reading this thesis and correcting all the linguistic errors I made. Furthermore, I would like to thank Sebastian Curth for reading this thesis before I officially handed it in.

My deepest gratitude goes to my family for their support and trust along the way, come rain or come shine. Thank you for supporting me in the things I wanted to do, helping me to come through some crisis, and bearing my caprices.

And last, I thank Wesley Lim without whom my carbon footprint would have been less distinct in the last years, but my life would have been by far less adventurous. Thank you for reminding me that there is a life out there. Nothing compares to you.

Berlin, January 2015
Sebastian Herold

Abstract

Using dynamic models to describe biotechnological processes leads to a better understanding of the complex process dynamics and helps to find optimal conditions that improve the process significantly. However, the development of adequate mathematical models is generally difficult, tedious, time-consuming, and requires extensive prior experimentation.

This work presents an algorithm that automatically proposes process models from an automated processing and analysis of data from (fed-)batch experiments. For this purpose, the algorithm first uses different data smoothing and interpolation techniques to account for a typically noisy and poorly time-resolved data set. Then, the measurements are numerically compensated for the influence of feeding and sampling. To reveal the qualitative behavior of the measurements, a method is used that divides the compensated curves into several episodes in a probabilistic framework.

Based on these episodes and transitions between them, crucial information about the underlying reaction network can be obtained. For this, different biological phenomena describing the relation between several measured variables are defined. Rules to (dis-)prove these phenomena are applied to the data. The uncertainty of the phenomena detection towards influences like the number of taken samples and considered experiments and the measurement noise is analyzed by a bootstrap method.

The detected biological phenomena and the used measured variables then lead to an automated proposal of several model structures with different degrees of complexity. The best models are selected by Akaike's Information Criterion (AIC) and model-discriminating experiments are planned.

Furthermore, a procedure to detect model deficiencies is drafted. The phenomena detection is applied to simulations of a model and compared to the measurement-inherent phenomena. This approach is initially tested on simple case studies.

The presented algorithm is applied to fed-batch experiments of three different organisms (*Paenibacillus polymyxa*, *Streptomyces tendae*, and *Streptomyces griseus*). Small-size and medium-size structured models are generated, identified, and validated. The results show that the models still need to be improved, but, in many cases, are able to describe the dynamics satisfactorily. Thus, the presented approach helps to speed up the modeling process significantly.

Kurzfassung

Mit dynamischen Modellen für biotechnologische Prozesse können komplexe Prozessabläufe besser verstanden und optimale Versuchsbedingungen gefunden werden, welche den Prozess deutlich verbessern. Die Entwicklung geeigneter mathematischer Modelle ist allerdings schwierig und mühsam, erfordert viel Zeit und kann nur erfolgen, wenn ausreichend viele Experimente durchgeführt wurden.

Diese Dissertation stellt einen Algorithmus vor, der automatisch Modelle zur Prozessführung vorschlägt, nachdem Messdaten aus (Fed-)Batch-Experimenten automatisch bearbeitet und analysiert worden sind. Dazu werden zuerst unterschiedliche Methoden zur Datenglättung und -interpolation genutzt, um stark verrauschte und diskontinuierliche Daten später besser auswerten zu können. Danach wird der Einfluss der Zufütterung und der Probenahmen auf die Messdaten numerisch kompensiert. Das qualitative Verhalten der Messgrößen wird durch eine Methode untersucht, welche die ausgeglichenen Verläufe in mehrere Episoden probabilitisch unterteilt.

Ausgehend von diesen Episoden und den Übergängen zwischen ihnen können wichtige Informationen über das verborgene Reaktionsnetzwerk erlangt werden. Dazu werden verschiedene biologische Phänomene definiert, die das Verhalten der einzelnen Messgrößen zueinander beschreiben. Regeln, die die Phänomene be- oder widerlegen, werden auf die Messdaten angewendet. Die Unsicherheit der Phänomenerkennung bezüglich Einflussgrößen wie der Anzahl der Probenahmen und berücksichtigter Experimente sowie des Messrauschens wird durch eine Bootstrap-Methode untersucht.

Anhand der gefundenen biologischen Phänomene und der untersuchten Messgrößen werden mehrere Modellstrukturen unterschiedlicher Komplexität automatisch vorgeschlagen. Die besten Modelle werden durch Akaikes Informationskriterium ausgewählt und modell-diskriminierende Versuche werden geplant.

Darüber hinaus wird ein Verfahren skizziert, mit dem Modelldefizite erkannt werden sollen. Die Phänomenerkennung wird auf die Simulationen angewendet und mit den erkannten Phänomenen der Messdaten verglichen. Der Ansatz wird mit einfachen Fallstudien getestet.

Der vorgeschlagene Algorithmus wird genutzt, um Fed-Batch-Experimente dreier unterschiedlicher Organismen zu untersuchen (*Paenibacillus polymyxa*, *Streptomyces tendae* und *Streptomyces griseus*). Strukturierte Modelle kleiner und mittlerer Größe werden erzeugt, identifiziert und validiert. Die Ergebnisse zeigen, dass die Modelle noch verbessert werden müssen, aber in vielen Fällen die Dynamik ausreichend gut beschreiben. Der vorgeschlagene Ansatz ist demnach in der Lage, den Modellierungsprozess deutlich zu beschleunigen.

Contents

List of Figures

List of Tables

Notation

Abbreviations

AIC	Akaike's Information Criterion
C	Compartment
DNA	Deoxyribonucleic acid
inhib	Inhibiting dependency
limit	Limiting dependency
ME	Motivating example
P	Product
Pr	Proteins
RNA	Ribonucleic acid
S	Substrate
Sc	Score
St	Storage
STBV	Best validated model of *S. tendae*
UM3S	Unstructured model with three substrates
X	Biomass
Xa	Active biomass

Latin Letters

a	Parameter
B	Number of bootstrap samples
b	Paramater
$\mathbf{C}$	Covariance matrix
$\mathbf{C}_{\mathrm{in}}$	Matrix containing the concentrations in the feeding
C_0	Constant
c	Mass concentration
$\mathbf{D}$	Matrix for Whittaker smoothing
f	Differential equation
$\mathcal{G}$	Polynomial
g	Polynomial coefficient
g	Intracellular concentrations
h	Measurement equation
h_t	Scale parameter
K	Kernel
K	Kinetics constant
K	Number of parameters
$\mathbf{K}$	Pseudo-stoichiometric matrix
L	Whittaker-smoothing parameter

$L_{\mathrm{P}i}$	Length of confidence interval
M	Number of neighboring measurements on each side
$\mathcal{M}$	Model candidate
m	Mass
N	Number of measurement samples
N_{Exp}	Number of experiments used for the parameter identification
N_{Mod}	Number of model candidates used for the model discrimination
n	Sample size
P	Probability
$\mathcal{P}$	Polynomial
p	Polynomial coefficient
Q	Cost functional of Whittaker smoother
q	Number of measurement variables
R	Measure for smoothness
R	Denominator in weight calculation
$\mathbb{R}$	Set of real numbers
r	Reaction rate
r	Weight
S	Measure for deviation from original data
$\mathcal{S}$	Model structure
s	Slope in a point
t	Time
u	Flow rate
w	Weight
V	Volume
Y	Yield coefficient
$\mathbf{Y}$	Experimental data
x	State variable
y	Measurement variable
z	Slope of a straight line

Greek Letters

Δ	Difference, interval
Δ_i	AIC Difference
ϵ	Measurement noise
λ	Whittaker-smoothing parameter
μ	Specific reaction rate
ξ	Function variable
Φ	Cost functional

Subscripts

Aa	Amino acid

Am	Ammonium
C	Compartment
D	DNA
carb	Carbon source
Ep	Episode
Gc	Glucose
in	Feeding
Intp	Interpolation
i	Index
j	Index
k	Index
Lm	Landmark
M	Maintenance
Ml	Macrolactin
m	Maximum
Nm	Nikkomycin
Nu	Nucleotide
P	Product
$\mathrm{P}i$	Placeholder for a specific phenomenon
Ph	Phosphate
Pr	Proteins
Phen	Phenomenon
R	RNA
S	Substrate
Sm	Streptomycin
St	Storage
X	Biomass
μ	Index
ν	Index

Superscripts

(i)	Quantile
low	Lower bound
T	Transposed
up	Upper bound
$*$	Raw data
$\sim$	Compensated measurements
$\star$	Modification

Chapter 1

Introduction

Modern biotechnology, initiated by the birth of genetic engineering in the 1970s, has rapidly developed into a key technology of the 21st century (Demain, 2000a, Schügerl, 2001, biotechnologie.de, 2013). Biotechnology has a significant impact on human life as it applies to major industrial areas such as health care and pharmacy (‘red’ or medical biotechnology), crop production and agriculture (‘green’ biotechnology), and production of materials and chemicals (‘white’ or industrial biotechnology). For example, primary metabolites of microorganisms are used in the food industry: alcohol, amino acids (e.g., monosodium glutamate), organic acids, sugars, vitamins. Likewise, secondary metabolites (e.g., antibiotics, toxins, biopesticides, immunosuppressants, antitumor agents) are extremely important for health and nutrition (Demain, 2000a,b). Moreover, biotechnology offers environmental and economic benefits that provide new opportunities for sustainable production of existing and new products in the chemical industry (Gavrilescu and Chisti, 2005), that reduce pollution and its dependence on nonrenewable fuels and other resources. The increasing importance of biotechnology is also supported by its economic development. In 2012, the German biotechnology sector achieved a record turnover of EUR 2.9 billion, which increased by 32 % since 2008, the major part of which was generated by red biotechnology (EUR 2.02 billion, + 169 % since 2008). Industrial biotechnology contributed only EUR 193 million to turnover, but this sector has been the strongest growing sector since 2008 (+ 250 %) (biotechnologie.de, 2013).

A great number of the current biological production processes, mainly (fed-)batch fermentions, have the potential to be improved considerably (Schügerl, 2001, Roubos, 2002, Clementschitsch and Bayer, 2006, Kawohl et al., 2007). On the one hand, real-time process monitoring of different physical, chemical, and biological parameters is still limited, and developing better monitoring techniques may help to improve and optimize the production processes. On the other hand, in the industrial practice, trial-and-error approaches rather than more sophisticated methods are used to find optimal process conditions. Using dynamic models that describe past performance and predict the future performance of biotechnological processes can lead to better results (Schügerl, 2001). These models yield a better understanding of the complex process dynamics. Based on a mathematical description, optimization-based concepts can be applied to find optimal process conditions. Furthermore, state estimation techniques and advanced process control can be applied to the process, leading to a better performance (Schügerl, 2001). Especially the importance of state estimation cannot be underestimated after the U.S. Food and Drug Administration launched the Process Analytical Technology (PAT) initiative (U.S. Food and Drug Administration, 2004). Its major goal is to improve the understanding and control of the manufacturing process "through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality." In this context, dynamic models combined with state estimation can be seen as soft sensors for crucial yet immeasurable variables (Clementschitsch and Bayer, 2006, Rehbock et al., 2008, Herwig, 2010).

However, the development of adequate mathematical models for dynamic fed-batch fermentation processes is generally difficult, tedious, time-consuming, and requires extensive prior experimentation (Junker and Wang, 2006). At first, the data set used to develop a model originates from a difficult measuring situation. The measurements have to be obtained from drawn samples, with a sampling interval of several hours, resulting in poor, time-resolved and mostly noisy data sets. Then, depending on the goal of the model, an appropriate degree of complexity has to be defined. The simplest approach is given by unstructured models. They consider the biomass as a single compound. The conversion from initial substrates into final products is described by a 'black box' approach, using simple kinetic equations (Roubos, 2002, Bernard and Bastin, 2005). Due to their simplicity and few measurement variables required (substrates, biomass, products), mainly unstructured models have been proposed in the past. However, unstructured models are usually unable to explain or predict the behavior in a larger region of operation, and their ability to predict the cellular behavior under different cultivation conditions is quite limited (Gombert and Nielsen, 2000). For example, antibiotics production is not initiated until one of the substrates has depleted, leading to the reorganization of a cell's metabolism. Unstructured models cannot describe this behavior, and in the context of model-based control and optimization, they are often bound to fail (King, 1997). On the other hand, structured models consider the changing composition of the cells over time and describe the biological processes more accurately on the basis of physiological and biochemical principles and structures (Roubos, 2002). The most complex approach is represented by models derived from metabolic engineering which attempt to completely describe a subset of metabolic fluxes. Here, many substances need to be measured for modeling and a considerable amount of information with respect to the reaction network is needed. For models still being applicable to process control, less complex models are considered where the dynamic interactions of the most important substances are mimicked on a more simplistic scale (*small-size* and *medium-size* strucured models). These models contain several biotic state variables (e.g., DNA, RNA, protein content) which add up to the biomass. Few cell-intern compartments lump together several cellular functions, regulations, and dynamics of the cell. As a result, only few components need to be measured: the biomass, substrates and products, and some cell-intern components (like DNA, RNA, and proteins). For example, such structured models have been proposed by Nielsen et al. (1991a,b,c), Nikolajsen et al. (1991), King (1997), King and Büdenbender (1997), Paul et al. (1998), Bapat et al. (2006), Tang et al. (2007), Çelik et al. (2009).

To develop a mathematical model in such a way, a human expert will analyze experimental data and different nutrient situations in the fermentation data, e.g., limitations of important substrates need to be considered (King, 1997, King and Büdenbender, 1997, Kammerer and Gilles, 2000, Roubos, 2002). Based on the expert's knowledge about the cell's metabolism and a qualitative analysis of the measurements, he or she will detect correlations between different substances in the reaction network. However, before correct conclusions can be drawn, some preliminary steps are necessary. Measurement noise needs to be considered, i.e., the experimental data should be smoothed first. Then, for a fed-batch experiment, the feeding profile and the sampling have to be taken into account to get correct information about the measurement dynamics and the interaction of two or more reactants. Afterwards, the expert will describe the measurements qualitatively, biological phenomena can be found, and model components to describe specific phenomena will be proposed. After a parameter identification step, the model outcome will be compared to the measurements. The human modeler will then try to find model deficiencies and model improvements which will lead to another parameter identification step. It is obvious, in this context, that defining a model structure and then identifying the model parameters is an iterative procedure that will take up a lot of time.

In this work, a software-supported approach is presented that tries to imitate the human expert. An algorithm is presented that automatically discovers biological phenomena which describe correlations between changes in the qualitative behavior of different measurements. For this purpose, the measurements have to be compensated for the effects from feeding and sampling. Since the experiments are usually characterized by infrequent and noisy data, these measurements have to be interpolated. Contrary to Cheung and Stephanopoulos (1990), the qualitative behavior of the measurements is revealed by a heuristic but probabilistic approach that lowers the effect of inaccuracies in the data reconciliation and interpolation procedure for such approaches. After having defined and detected biological phenomena, model structures are proposed automatically. In the next step, they are transfered to a parameter identification block, finally leading to a list of identified models, which can be ordered, e.g., by their goodness-of-fit or more sophisticated methods considering the number of parameters. At last, model deficiencies are detected automatically—completing the aforementioned iterative model building procedure. By applying the presented methodology, it will be possible on the one hand to speed-up the modeling process significantly. On the other hand, many more model candidates will be tested compared to what a human modeler is able or willing to do. In Figure 1.1, an overview of the presented approach is outlined by a flow chart. As will be shown later, at some points, the user has to decide on several components of the algorithm that could not completely be automated. These 'manual' steps are essential for finding good model candidates, i.e., the software approach frees the human modeler from many tedious tasks but does not replace him or her. The figure shows these user inputs as well.

Outline

This thesis is organized as follows: In Chapter 2, it is shown how models used to describe cell growth can be derived. Chapter 3 provides an overview about necessary measurement reconciliation steps. Data smoothing and interpolation techniques are described and the probabilistic calculation of time periods with the same qualitative

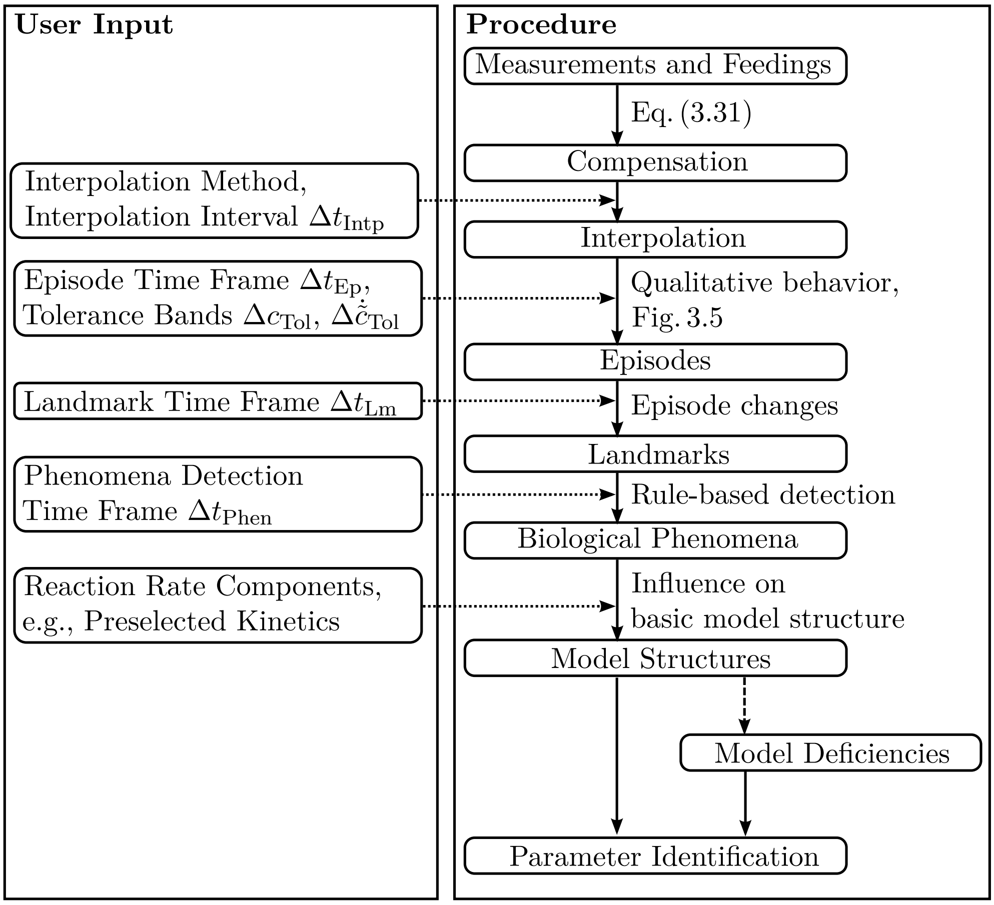

User Input
Interpolation Method, Interpolation Interval Δt_{Intp}
Episode Time Frame Δt_{Ep}, Tolerance Bands Δc_{Tol}, $\Delta \dot{\tilde{c}}_{\mathrm{Tol}}$
Landmark Time Frame Δt_{Lm}
Phenomena Detection Time Frame Δt_{Phen}
Reaction Rate Components, e.g., Preselected Kinetics
Procedure
Measurements and Feedings
Eq. (3.31)
Compensation
Interpolation
Qualitative behavior, Fig. 3.5
Episodes
Episode changes
Landmarks
Rule-based detection
Biological Phenomena
Influence on basic model structure
Model Structures
Model Deficiencies
Parameter Identification

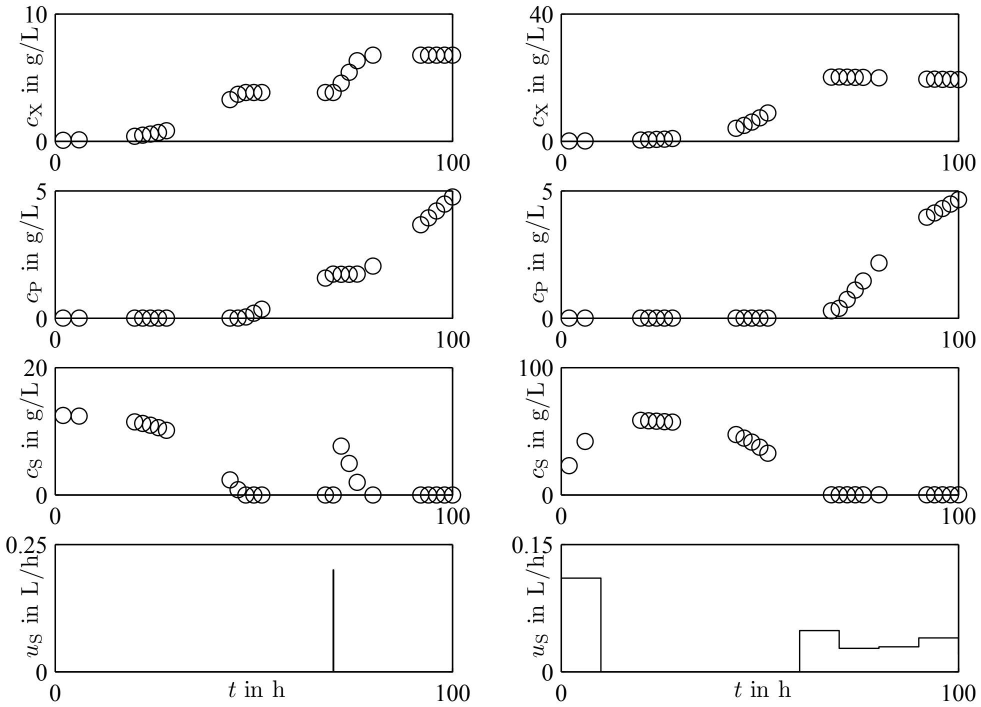
Figure 1.1: Flow chart of the presented algorithm and possible user inputs

behavior (*episodes*) and associated transition time points (*landmarks*) are introduced. In Chapter 4, it is shown how *biological phenomena* can be detected automatically. In Chapter 5, a possible approach for testing these phenomena for their uncertainty towards influences like measurement noise is given. Chapter 6 shows how biological phenomena that have been detected lead to proposed reaction patterns and model candidates. Furthermore, the parameter identification and model selection steps are described in this chapter. An approach to automatically detect model deficiencies is presented in Chapter 7. An experimental validation of the methodology is given in Chapter 8, using data from fed-batch experiments of three different strains. Conclusions are drawn in Chapter 9.

The presented methodology is initially tested with a simple example, which is introduced next.

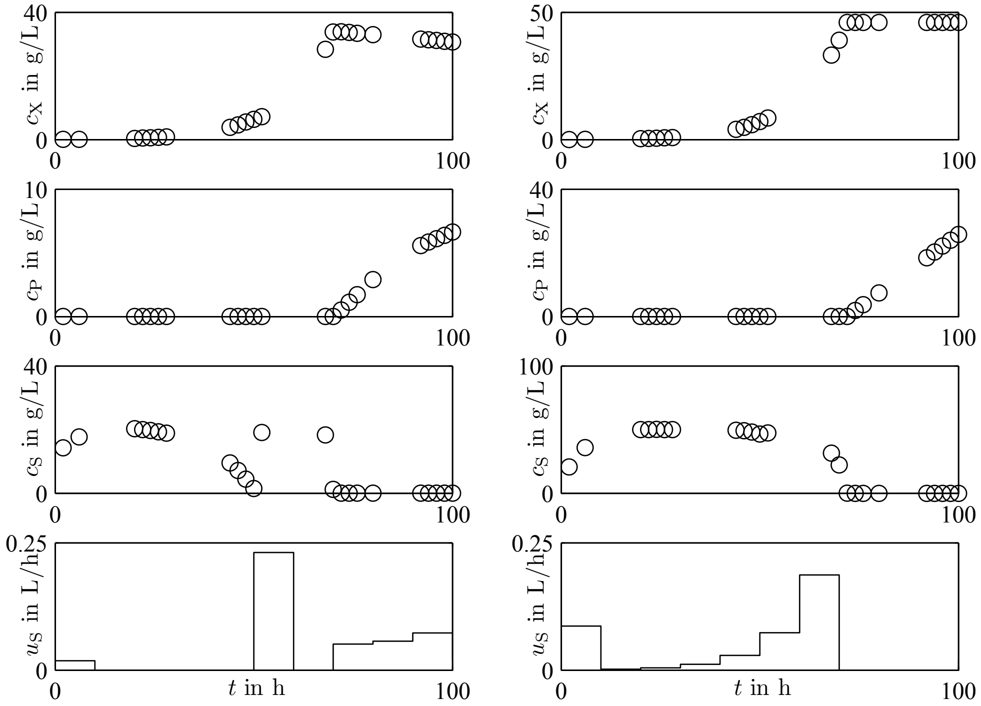
Motivating Example

The measurements of the in-silico experiments ME1–ME4 given in Figure 1.2 are



(a) ME1

(b) ME2



(c) ME3

(d) ME4

Figure 1.2: Measurements from four different in-silico experiments of the motivating example. The indices are X—biomass, P—product, and S—substrate. The feeding rates are given by u_S.

Table 1.1: Parameter values of the motivating example

	Parameter	Value	Unit
r_X	μ_Xm	0.10	1/h
	K_XS	1.00	g/l
r_P	μ_Pm	0.02	1/h
	K_PS	0.10	g/l

	Parameter	Value	Unit
r_M	μ_Mm	0.10	1/h
	K_M	0.01	g/l
	Y_SX	2.00	g/g

generated based on a simple unstructured model

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_\mathrm{X}(t) \\ m_\mathrm{S}(t) \\ m_\mathrm{P}(t) \end{pmatrix} = \begin{pmatrix} 0 \\ c_\mathrm{S,\,in}\cdot u_\mathrm{S}(t) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ r_\mathrm{M}(t)V(t) \\ 0 \end{pmatrix} + V(t)\begin{pmatrix} 1 & 0 \\ -Y_\mathrm{SX} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} r_\mathrm{X}(t) \\ r_\mathrm{P}(t) \end{pmatrix} \tag{1.1}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_\mathrm{S}(t) \quad , \tag{1.2}$$

describing the dynamic behavior of the biomass X, a substrate S, a product P, and the volume $V(t)$. The reaction rates are given by

$$r_\mathrm{X}(t) = \mu_\mathrm{Xm}\cdot\frac{c_\mathrm{S}(t)}{c_\mathrm{S}(t)+K_\mathrm{XS}}\cdot c_\mathrm{X}(t) \tag{1.3}$$

$$r_\mathrm{P}(t) = \mu_\mathrm{Pm}\cdot\frac{K_\mathrm{PS}}{c_\mathrm{S}(t)+K_\mathrm{PS}}\cdot c_\mathrm{X}(t) \tag{1.4}$$

$$r_\mathrm{M}(t) = \mu_\mathrm{Mm}\cdot\frac{c_\mathrm{S}(t)}{c_\mathrm{S}(t)+K_\mathrm{M}}\cdot c_\mathrm{X}(t) \quad . \tag{1.5}$$

The measured variables are the concentrations of the three substances,

$$\underline{y}(t) = \begin{pmatrix} c_\mathrm{X}(t) \\ c_\mathrm{S}(t) \\ c_\mathrm{P}(t) \end{pmatrix} \quad . \tag{1.6}$$

The parameter values can be seen in Table 1.1.

Chapter 2

Modeling Cell Growth

The formulation of mass balances of the major components of a process is an essential stage in the development of any model (Dunn et al., 2003) and, when describing biological processes, the most natural way to determine models that will enable the characterization of the process dynamics (Dochain, 2008). These component mass balances are generally formulated by

$$\begin{pmatrix}\text{Accumulation rate}\\ \text{of mass of component}\\ \text{in the system}\end{pmatrix} = \begin{pmatrix}\text{Mass flow}\\ \text{of component}\\ \text{into system}\end{pmatrix} - \begin{pmatrix}\text{Mass flow}\\ \text{of component}\\ \text{out of system}\end{pmatrix} \pm \begin{pmatrix}\text{Production/}\\ \text{consumption rate}\\ \text{of component}\end{pmatrix} . \tag{2.1}$$

Considering fermentations, the ‘system’ is the liquid phase in the reactor. The inflows and outflows in Eq. (2.1) are defined by the experiments. The modeling of the production/consumption rate of a component is the more complicated part (Schaber et al., 2009). Appropriate descriptions for cell growth, substrate consumption, product formation, etc. have to be found.

Depending on the purpose of the yet to be established model, a certain degree of complexity of the model structure has to be chosen (Schaber et al., 2009). Models that are meant to represent the metabolism realistically, and where detailed knowledge of the processes are necessary, need a higher complexity than models that are supposed to describe the processes in a more general manner and where empirical formulations are sufficient. The different approximations that are useful for describing the cell are divided into *unsegregated* or *segregated* models and *unstructured* and *structured* models (e.g., Bailey and Ollis, 1986, Dunn et al., 2003, Chmiel, 2006). In this work, models are built that are applicable to process control. They are supposed to mimic the dynamic interactions of the most important substances in fed-batch fermentations on a more simplistic scale and, at the same time, do not pose too many challenges in the model-building step. Only unsegregated representations will be considered, i.e., cells are not described individually and only average cellular properties are considered. Unstructured and structured models are explained in Sections 2.1 and 2.2, respectively.

2.1 Unstructured Models

Unstructured models do not describe the influence of a changing metabolism. They consider the total cell mass as a ‘black box’ that converts initial substrates into final products (Bernard and Bastin, 2005).

The growth of the biomass X is described by

$$\frac{\mathrm{d}m_{\mathrm{X}}(t)}{\mathrm{d}t} = r_{\mathrm{X}}(t) \cdot V(t) \quad , \tag{2.2}$$

where $r_{\mathrm{X}}(t)$ is the growth rate with respect to the volume $V(t)$. The growth rate is typically formulated as

$$r_{\mathrm{X}}(t) = \mu_{\mathrm{X}}(t) \cdot c_{\mathrm{X}}(t) \quad , \tag{2.3}$$

where $\mu_{\mathrm{X}}(t)$ is called the specific growth rate. In the case of unlimited growth, $\mu_{\mathrm{X}}(t)$ is constant,

$$\mu_{\mathrm{X}}(t) = \mu_{\mathrm{Xm}} \quad , \tag{2.4}$$

leading to an exponential evolution of $m_{\mathrm{X}}(t)$. However, growth is not unlimited but regulated by external factors (Shonkwiler and Herod, 2009). The nutrient supply has to be taken into account. If a sufficient amount of nutrition is present, the cell will be able to grow exponentially. In the case of limited resources, the growth will slow down or even stop. Initially, considering only one substrate S, $\mu_{\mathrm{X}}(t)$ can be formulated by

$$\mu_{\mathrm{X}}(t) = \mu_{\mathrm{X}}(c_{\mathrm{S}}(t)) = \mu_{\mathrm{Xm}} \cdot \mathrm{limit}(c_{\mathrm{S}}(t)) \quad , \tag{2.5}$$

where $\mathrm{limit}(c_{\mathrm{S}}(t))$ stands for any limiting dependency on the medium concentration $c_{\mathrm{S}}(t)$ that satisfies

$$\mathrm{limit}(c_{\mathrm{S}} = 0) = 0 \tag{2.6a}$$

and

$$\lim_{c_{\mathrm{S}} \to \infty} \mathrm{limit}(c_{\mathrm{S}}) = 1 \quad . \tag{2.6b}$$

Examples are given by Monod (1949) or Michaelis and Menten (1913),

$$\mu_{\mathrm{Monod}}(c_{\mathrm{S}}(t)) = \frac{c_{\mathrm{S}}(t)}{c_{\mathrm{S}}(t) + K_{\mathrm{XS}}} \tag{2.7}$$

or Moser (1958),

$$\mu_{\mathrm{Moser}}(c_{\mathrm{S}}(t)) = \frac{(c_{\mathrm{S}}(t))^{\lambda}}{(c_{\mathrm{S}}(t))^{\lambda} + K_{\mathrm{XS}}} \quad . \tag{2.8}$$

Possible limiting functions can be found in Figure 2.1.[1]

Usually, more than one substrate are essential for growth (e.g., a nitrogen, phosphorus, and carbon source are needed), and each substrate S_j is a limiting resource. Haefner

---

[1] Expressions like $\mu_{\mathrm{X}}(t) = k \cdot c_{\mathrm{S}}(t)^{\gamma}$, with k and γ being yet to be determined parameters, represent limiting dependencies, as well, but will not be used in this work.

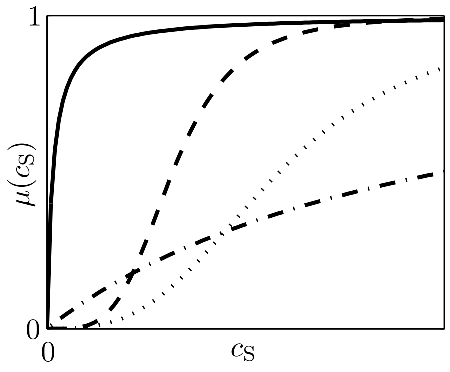


Figure 2.1: Possible limiting dependencies $\mu(c_\mathrm{S}) = \mathrm{limit}(c_\mathrm{S})$ on the substrate concentration c_S

(2005) shows five common methods for combining the regulating effects of these substrates. One possibility is the use of a multiplicative growth rate

$$\mu_\mathrm{X}(t) = \mu_\mathrm{Xm} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \quad . \tag{2.9}$$

If it is necessary, the cell death can be integrated by a separate death rate $r_\mathrm{dX}(t)$ and the mass balance will be described by

$$\frac{\mathrm{d}m_\mathrm{X}(t)}{\mathrm{d}t} = (r_\mathrm{X}(t) - r_\mathrm{dX}(t)) \cdot V(t) \quad . \tag{2.10}$$

Now, with the assumed reaction

$$\sum_j Y_{\mathrm{S}_j\mathrm{X}}\, \mathrm{S}_j \xrightarrow{r_\mathrm{X}} \mathrm{X} \quad , \tag{2.11}$$

where $Y_{\mathrm{S}_j\mathrm{X}}$ is the so-called yield coefficient, the mass balance for substrate S_j in a fed-batch reactor is

$$\frac{\mathrm{d}m_{\mathrm{S}_j}(t)}{\mathrm{d}t} = -Y_{\mathrm{S}_j\mathrm{X}} \cdot r_\mathrm{X}(t) \cdot V(t) + c_{\mathrm{S}_j,\mathrm{in}} \cdot u_{\mathrm{S}_j}(t) \quad . \tag{2.12}$$

For substrates S_carb serving as a carbon and energy source, a maintenance term

$$r_\mathrm{M}(t) \cdot V(t)$$

is included as well:

$$\frac{\mathrm{d}m_{\mathrm{S}_\mathrm{carb}}(t)}{\mathrm{d}t} = (-Y_{\mathrm{S}_\mathrm{carb}\mathrm{X}} \cdot r_\mathrm{X}(t) - r_\mathrm{M}(t)) \cdot V(t) + c_{\mathrm{S}_\mathrm{carb},\mathrm{in}} \cdot u_{\mathrm{S}_\mathrm{carb}}(t) \quad , \tag{2.13}$$

where the maintenance rate can be described by

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{S_{carb}}}(t)}{c_{\mathrm{S_{carb}}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \quad . \tag{2.14}$$

Product formation

In addition to cell growth, the product formation has to be described as well. However, there is no universal approach for including the product into the model. In fact, the product formation can be classified into several classes that determine how the stoichiometry of product formation and reaction rates will look like (Gaden, 1959, Bailey and Ollis, 1986):

1. The main product appears as a result of the primary energy metabolism, i.e., dissimilation of the primary carbohydrate.

   In this case, product formation and cell synthesis are coupled, and the product can simply be added to the right-hand side of Eq. (2.11),

   $$\sum_j Y_{\mathrm{S}_j\mathrm{X}}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{X}}} \mathrm{X} + Y_{\mathrm{P}}\,\mathrm{P} \quad . \tag{2.15}$$

   The mass balance for P is formulated by

   $$\frac{\mathrm{d}m_{\mathrm{P}}(t)}{\mathrm{d}t} = Y_{\mathrm{P}} \cdot r_{\mathrm{X}}(t) \cdot V(t) \quad . \tag{2.16}$$

   As an example, the ethanol production during the anaerobic growth of yeast can be considered.

2. The main product arises indirectly from the energy metabolism, i.e., it accumulates only under conditions of restricted or abnormal metabolism.

   Here, the product formation is not necessarily proportional to the cell growth. An extra reaction describing the product formation has to be introduced.

   $$\sum_j Y_{\mathrm{S}_j\mathrm{P}}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{P}}} \mathrm{P} \quad . \tag{2.17}$$

   Here, the dynamic behavior of the product P is described by

   $$\frac{\mathrm{d}m_{\mathrm{P}}(t)}{\mathrm{d}t} = r_{\mathrm{P}}(t) \cdot V(t) \quad , \tag{2.18}$$

   where the reaction rate $r_{\mathrm{P}}(t)$ has to be specified according to theoretical knowledge or experimental results. For instance, $r_{\mathrm{P}}(t)$ might be limited by the substrates S_j,

   $$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot c_{\mathrm{X}}(t). \tag{2.19}$$

Furthermore, Eq. (2.12) has to be adjusted to

$$\frac{\mathrm{d}m_{\mathrm{S}_j}(t)}{\mathrm{d}t} = (-Y_{\mathrm{S}_j\mathrm{X}} \cdot r_{\mathrm{X}}(t) - Y_{\mathrm{S}_j\mathrm{P}} \cdot r_{\mathrm{P}}(t)) \cdot V(t) + c_{\mathrm{S}_j,\mathrm{in}} \cdot u_{\mathrm{S}_j}(t) \quad . \tag{2.20}$$

As an example, citric acid is produced during the aerobic cultivation of molds, e.g., *Aspergillus niger* (Currie, 1917).

3. The product is a *secondary metabolite.*

   Secondary metabolites are produced by plants, bacteria, and fungi, but are typically uncoupled from basic metabolism, i.e., growth, development, and reproduction (Fraenkel, 1959). Secondary metabolites are synthesized when the cells and their environment are at appropriate conditions. Their accumulation is dictated by kinetic regulation and activity of the cell.

   Important secondary metabolites are antibiotics, e.g., penicillin and streptomycin. Here, the antibiotics production is not initiated until one of the substrates can only be found in small amounts or is even depleted (Bajpai and Reuß, 1980, Mundry and Kuhn, 1991, Martín et al., 2011). This behavior can be described mathematically by

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \mathrm{inhib}(c_{\mathrm{S}_l}(t)) \cdots c_{\mathrm{X}}(t) \quad , \tag{2.21}$$

   where $\mathrm{inhib}(c_{\mathrm{S}_l}(t))$ stands for any inhibiting dependency that satisfies

$$\mathrm{inhib}(c_{\mathrm{S}_l} = 0) = 1 \tag{2.22a}$$

   and

$$\lim_{c_{\mathrm{S}_l} \to \infty} \mathrm{inhib}(c_{\mathrm{S}_l}) = 0 \quad . \tag{2.22b}$$

   Examples are given by Jerusalimski and Engamberdiev (1969),

$$\mu_{\mathrm{Jeru}}(c_{\mathrm{S}}(t)) = \frac{K_{\mathrm{PS}}}{c_{\mathrm{S}}(t) + K_{\mathrm{PS}}} \quad , \tag{2.23}$$

   or Aiba et al. (1968),

$$\mu_{\mathrm{Aiba}}(c_{\mathrm{S}}(t)) = \exp(-K_{\mathrm{PS}} \cdot c_{\mathrm{S}}(t)) \quad . \tag{2.24}$$

   Possible curves can be seen in Figure 2.2.

In matrix notation, the model can then be written as

$$\frac{\mathrm{d}\underline{m}(t)}{\mathrm{d}t} = \underbrace{\mathbf{K} \cdot \underline{r}(t) \cdot V(t)}_{\text{Metabolism}} + \underbrace{\mathbf{C}_{\mathrm{in}} \cdot \underline{u}^T(t)}_{\text{Inlet}} - \underbrace{\underline{\nu}(t)}_{\text{Maintenance}} \quad , \tag{2.25}$$

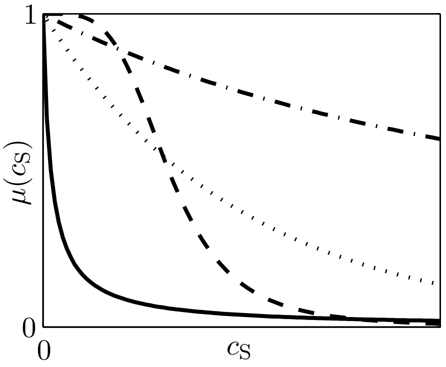


Figure 2.2: Possible inhibiting dependencies $\mu(c_S) = \text{inhib}(c_S)$ on the substrate concentration c_S

where $\mathbf{K}$ is the matrix of yield coefficients (pseudo-stoichiometric matrix), $\underline{r}$ is a vector containing the aforementioned reaction rates, $\mathbf{C}_{\text{in}}$ is a diagonal matrix containing the concentrations in the feeding, and the vector $\underline{u}^T$ comprises the feeding rates. The maintenance $\underline{\nu}$ only affects the substrate serving as a carbon and energy source (S_{carb}). Other components of this vector are set to zero.

Due to their simplicity, unstructured models have been proposed in many cases during the last decades. They only require measurements of substrates, the biomass, and the products to be built and the numerical values for many important process parameters can be determined easily (Dunn et al., 2003). However, unstructured models are usually unable to explain or predict the behavior in a larger region of operation, and their ability to predict the cellular behavior under different cultivation conditions is quite limited or insufficient (Gombert and Nielsen, 2000, Dunn et al., 2003). Such a situation has to be dealt with, e.g., in antibiotics production. Here, the cells are grown exponentially at the beginning of cultivation to produce a high cell mass concentration. Then, antibiotics production is initiated by a depletion of one of the substrates. This depletion will lead to a complete reorganization of the cell's metabolism for which an unstructured model cannot account for. Hence, when the unstructured models are used in the context of model-based control and optimization, experiments are often bound to fail (King, 1997).

2.2 Structured Models

Since events can occur that lead to the reorganization of a cell's metabolism, more detailed models are required that consider the changing composition of cells over time—structured models. One class of models is derived from metabolic engineering (e.g., Yarmush and Banta, 2003, Almaas et al., 2004, Vemuri and Aristidou, 2005, Nocon et al., 2014). These models attempt to comprehensively describe a subset of metabolic fluxes. They represent the most complex approach when describing biological systems. However, many substances have to be measured for modeling and a considerable amount of information regarding the reaction network is needed. Since the application of these models to process control is questionable and modeling might be too cumbersome, a compromise is sought. The structured models used here will contain only few biotic state variables (e.g., internal storage, active biomass, DNA, RNA, protein content) which add up to the biomass. These so-called compartments lump together several cellular functions, regulations and dynamics.

In the simplest case, the cell is divided into the active biomass Xa, responsible for replication, and storages $\mathrm{S}_j\mathrm{St}$, where substrates S_j are accumulated. Their growth is described by

$$\frac{\mathrm{d}m_{\mathrm{Xa}}(t)}{\mathrm{d}t} = r_{\mathrm{Xa}}(t) \cdot V_{\mathrm{X}}(t) \tag{2.26}$$

and

$$\frac{\mathrm{d}m_{\mathrm{S}_l\mathrm{St}}(t)}{\mathrm{d}t} = r_{\mathrm{S}_l\mathrm{St}}(t) \cdot V_{\mathrm{X}}(t) \quad , \tag{2.27}$$

respectively, where $V_{\mathrm{X}}(t)$ represents the cell volume. Within the modified reaction network, e.g.,

$$\sum_j Y_{\mathrm{S}_j\mathrm{Xa}}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{Xa}}} \mathrm{Xa} \tag{2.28}$$

and

$$\mathrm{S}_l \xrightarrow{r_{\mathrm{S}_l\mathrm{St}}} \mathrm{S}_l\mathrm{St} \quad , \tag{2.29}$$

the reaction rates have to be defined, e.g.,

$$r_{\mathrm{Xa}}(t) = \mu_{\mathrm{Xa}}(t) \cdot g_{\mathrm{Xa}}(t) \tag{2.30}$$

and

$$r_{\mathrm{S}_l\mathrm{St}}(t) = \mu_{\mathrm{S}_l\mathrm{St}}(t) \cdot g_{\mathrm{Xa}}(t) \quad , \tag{2.31}$$

with $g_{\mathrm{Xa}}(t)$ being the intracellular concentration regarding the total cell mass or cell volume $V_{\mathrm{X}}(t)$, i.e.,

$$g_{\mathrm{Xa}}(t) = \frac{m_{\mathrm{Xa}}(t)}{V_{\mathrm{X}}(t)} \quad . \tag{2.32}$$

Any other reaction scheme or reaction rate can be chosen, as well, according to the measurements of the conducted experiments. Compared to unstructured models, the measurement situation is the same. The biomass, substrates, and the product are measured. However, the biomass measurements now comprise several state variables,

$$c_{\mathrm{X}}(t) = \frac{m_{\mathrm{Xa}}(t) + \sum\limits_j m_{\mathrm{S}_j\mathrm{St}}(t)}{V(t)} \quad . \tag{2.33}$$

These models will be called small-size structured models in this work.

More complex structured models involve more measurable components than shown above: the biomass, substrates and products, and some cell-intern components (like DNA, RNA, and proteins). However, as detailed information about the metabolism is ignored in the lumping process, these medium-size structured models are still black-box in nature, though inspired by biological knowledge.

Some biological knowledge of DNA, RNA, and proteins

When cellular compartments like DNA, RNA, and proteins are integrated in a model, and no simple static relationships between the substrates and these compartments can be found, some biological knowledge about the synthesis, degradation, cellular functions, etc. of these compartments is useful for the modeling step. In what follows, it is shown how the biological processes involved can be integrated into a mathematical model. However, irrespective of this biological inspiration, it is not initially known which substances are to be considered and which reactions have to be described. Hence, the following mathematical description can only be seen as an example and not as a rule for how a structured model should look. Under certain circumstances, simpler model assumptions than the following might be sufficient, whereas in other cases, an even more complex model is necessary.

At first, the composition of the compartments can indicate how possible building-up reactions could look.

- The nucleic acids, i.e., DNA and RNA, are both polymers of *nucleotides* Nu which, in turn, are made up by phosphoric acid, five-carbon sugars (ribose or deoxyribose), and a nitrogenous base (e.g., Bailey and Ollis, 1986).

Considering defined media with one nitrogen (e.g., ammonium Am), one phosphorous (e.g., phosphate Ph), and one carbon source (e.g., glucose Gc), this relationship can be described by the reactions

$$Y_{\mathrm{AmNu}}\,\mathrm{Am} + Y_{\mathrm{PhNu}}\,\mathrm{Ph} + Y_{\mathrm{GcNu}}\,\mathrm{Gc} \xrightarrow{r_{\mathrm{Nu}}} \mathrm{Nu} \quad , \tag{2.34}$$

$$\mathrm{Nu} \xrightarrow{r_{\mathrm{D}}} \mathrm{DNA} \quad , \tag{2.35}$$

and

$$\mathrm{Nu} \xrightarrow{r_{\mathrm{R}}} \mathrm{RNA} \quad . \tag{2.36}$$

- Proteins (Pr) are polymers of *amino acids* Aa which are composed of amine and carboxylic acid (e.g., Bailey and Ollis, 1986).

The corresponding building-up reactions can then look like

$$Y_{\mathrm{AmAa}}\,\mathrm{Am} + Y_{\mathrm{GcAa}}\,\mathrm{Gc} \xrightarrow{r_{\mathrm{Aa}}} \mathrm{Aa} \tag{2.37}$$

and

$$\mathrm{Aa} \xrightarrow{r_{\mathrm{Pr}}} \mathrm{Pr} \quad . \tag{2.38}$$

This leads to the following mass balances:

$$\frac{\mathrm{d}m_{\mathrm{D}}(t)}{\mathrm{d}t} = r_{\mathrm{D}}(t) \cdot V_{\mathrm{X}}(t) \tag{2.39}$$

$$\frac{\mathrm{d}m_{\mathrm{R}}(t)}{\mathrm{d}t} = r_{\mathrm{R}}(t) \cdot V_{\mathrm{X}}(t) \tag{2.40}$$

$$\frac{\mathrm{d}m_{\mathrm{Pr}}(t)}{\mathrm{d}t} = r_{\mathrm{Pr}}(t) \cdot V_{\mathrm{X}}(t) \tag{2.41}$$

$$\frac{\mathrm{d}m_{\mathrm{Nu}}(t)}{\mathrm{d}t} = (r_{\mathrm{Nu}}(t) - r_{\mathrm{D}}(t) - r_{\mathrm{R}}(t)) \cdot V_{\mathrm{X}}(t) \tag{2.42}$$

$$\frac{\mathrm{d}m_{\mathrm{Aa}}(t)}{\mathrm{d}t} = (r_{\mathrm{Aa}}(t) - r_{\mathrm{Pr}}(t)) \cdot V_{\mathrm{X}}(t) \tag{2.43}$$

$$\frac{\mathrm{d}m_{\mathrm{Am}}(t)}{\mathrm{d}t} = (-Y_{\mathrm{AmNu}} \cdot r_{\mathrm{Nu}}(t) - Y_{\mathrm{AmAa}} \cdot r_{\mathrm{Aa}}(t)) \cdot V_{\mathrm{X}}(t) + c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \tag{2.44}$$

$$\frac{\mathrm{d}m_{\mathrm{Ph}}(t)}{\mathrm{d}t} = -Y_{\mathrm{PhNu}} \cdot r_{\mathrm{Nu}}(t) \cdot V_{\mathrm{X}}(t) + c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \tag{2.45}$$

$$\frac{\mathrm{d}m_{\mathrm{Gc}}(t)}{\mathrm{d}t} = (-Y_{\mathrm{GcNu}} \cdot r_{\mathrm{Nu}}(t) - Y_{\mathrm{GcAa}} \cdot r_{\mathrm{Aa}}(t) - r_{\mathrm{M}}(t)) \cdot V_{\mathrm{X}}(t) + c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \quad . \tag{2.46}$$

A mass balance for the residual biomass, i.e., the part of the biomass not described by DNA, RNA, proteins, nucleotides, or amino acids, has to be established as well. Additionally, compartment degradation rates can be included as well. The mass balances can be written in matrix notation,

$$\frac{\mathrm{d}\underline{m}(t)}{\mathrm{d}t} = \underbrace{\mathbf{K} \cdot \underline{r}(t) \cdot V_{\mathrm{X}}(t)}_{\text{Metabolism}} + \underbrace{\mathbf{C}_{\mathrm{in}} \cdot \underline{u}^T(t)}_{\text{Inlet}} - \underbrace{\underline{\nu}(t)}_{\text{Maintenance}} \quad . \tag{2.47}$$

Second, the flow of information within the cell can provide factors to which the growth rates r_i are proportional.

- When a DNA molecule is replicated, the two strands of DNA are separated and each strand then acts as a template for a new, complementary strand (e.g., Shonkwiler and Herod, 2009). Hence, DNA codes its own accurate *replication*. This whole process is autocatalytic.

The rate $r_{\mathrm{D}}(t)$ can be written as

$$r_{\mathrm{D}}(t) = \mu_{\mathrm{D}}(t) \cdot g_{\mathrm{D}}(t) \quad . \tag{2.48}$$

- The *central dogma of molecular genetics* states that DNA passes its information on to RNA (*transcription*), and RNA passes it to the proteins (*translation*) (e.g., Shonkwiler and Herod, 2009).

Hence, the RNA and proteins formation rates can be expressed by

$$r_{\mathrm{R}}(t) = \mu_{\mathrm{R}}(t) \cdot g_{\mathrm{D}}(t) \tag{2.49}$$

and

$$r_{\mathrm{Pr}}(t) = \mu_{\mathrm{Pr}}(t) \cdot g_{\mathrm{R}}(t) \quad . \tag{2.50}$$

At last, regulation mechanisms in the cell can be used to define the specific rates μ_i. One fundamental distinction between the multiple mechanisms examines if the activity of a protein already present in the cell is changed or if the net rate of synthesis and, therefore, the cellular concentration of that protein is altered (Neidhardt et al., 1990). Both regulation types are managed by many different mechanisms. The activity can be changed by covalent modification (e.g., phosphorylation) or by reversible association with another molecule. The cellular concentration can be altered by changing its synthesis or, more rarely, its degradation rate. To do this, the cell provides multiple regulation mechanisms. However, a more detailed description of the regulation mechanisms is beyond this work's scope. The modeler, however, has to keep the complexity of the regulating processes in mind, and should consult related literature when necessary, e.g., Bailey and Ollis (1986), Neidhardt et al. (1990), Sanchez and Demain (2002), Cornish-Bowden (2004), Chmiel (2006), Bisswanger (2008).

An example for a structured model, that is inspired by the mentioned biological knowledge, can be found for the strain *Streptomyces tendae* established by King (1997) in Appendix C.1. The same basic model structure could also be applied to other strains (King and Büdenbender, 1997, Büdenbender, 2004) and has been used for process control and state estimation (Heine, 2004, Kawohl et al., 2007).

2.3 Tools and Techniques to Support Modeling

Since building dynamic models of biochemical networks has become an essential step in biosciences (e.g., systems biology and metabolic engineering) (Schaber et al., 2011), computational tools that support and facilitate the modeling process have become increasingly important (e.g., Ross, 2012). Extensive reviews on current tools and techniques have been published by Wiechert (2002), Crampin et al. (2004), Alves et al. (2006), Gostner et al. (2014). In the last years, many software packages have been developed that can simulate and visualize cellular models, e.g., Virtual Cell (Loew and Schaff, 2001) and COPASI (Hoops et al., 2006). Likewise, general purpose packages are available that are capable of analyzing complex dynamic systems, e.g., calculating parameter sensitivities and performing bifurcation analyses (Mangold et al., 2005, Schmidt and Jirstrand, 2006, Mirschel et al., 2009, Rodriguez-Fernandez and Banga, 2010, Droste et al., 2011). They are especially useful in the field of systems biology. Furthermore, the creation of the Systems Biology Markup Language (SBML) (Hucka et al., 2003) has set a standard for representing computational models in systems biology and allows for sharing and exchanging cellular models in a comprehensive way.

However, these software packages usually require already formulated models. Only little attention has been paid to tools that support the user in developing models from measurements (Clewley, 2012). Nevertheless, some methods and concepts have been introduced and described to reveal information about the reaction network. Cheung and Stephanopoulos (1990) provide a format that describes the measurements qualitatively. For simple process models, several approaches to identify the reaction network and the reaction rates are given (Bernard and Bastin, 2005, Hulhoven et al., 2005, Marquardt, 2005). Unfortunately, these approaches usually assume that every state variable is measured. Hence, they cannot be used when unmeasured components need to be integrated into the model. Several methods regarding structure identification have been established. They either try to find an adequate structure by testing multiple different hypotheses and handling multiple different model candidates at the same time (Haunschild et al., 2005, Wahl et al., 2006, Flöttmann et al., 2008, Violet et al., 2009, Schaber et al., 2011) or by generating alternative model candidates from already existing ones. In the latter case, these alternative models are often generated via the means of Genetic Programming (GP)[2]. It has been applied to biological systems by, e.g., Marenbach et al. (1997), Freyer et al. (1998), Sugimoto et al. (2005), Cho et al. (2006).

In this work, an approach is shown that identifies possible model structures based on the automatic detection of biological phenomena. This approach is initially proposed for simple process models by King et al. (2002) and will be extended here.

---

[2] GP is a learning algorithm based on evolutionary algorithms that modifies and extends structures by imitating principles of natural selection and reproduction. For each model candidate, a fitness value is evaluated that considers both accuracy and complexity (Marenbach et al., 1997).

Chapter 3

Measurement Reconciliation

Before information about the reaction network can be revealed automatically, the measurements

$$\mathbf{Y}^* = \begin{pmatrix} y_1^*(t_1) & y_1^*(t_2) & \dots & y_1^*(t_N) \\ y_2^*(t_1) & y_2^*(t_2) & \dots & y_2^*(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ y_q^*(t_1) & y_q^*(t_2) & \dots & y_q^*(t_N) \end{pmatrix} = \begin{pmatrix} \underline{y}^*(t_1) & \underline{y}^*(t_2) & \dots & \underline{y}^*(t_N) \end{pmatrix} \quad , \tag{3.1}$$

with q being the number of measured variables and N being the number of measurement samples, are subjected to a reconciliation step. Measurement noise, discontinuous data, and the influence of feeding and sampling need to be considered before the data can be described qualitatively and tested for correlations. For ease of presentation, only one measured variable $\underline{y}^{*T} = [y^*(t_1),\, y^*(t_2),\, \dots,\, y^*(t_N)]$ instead of the whole data set $\mathbf{Y}^*$ is used to explain the following methods.

3.1 Data Smoothing

Analyzing measurements from experiments usually means handling noisy data $\underline{y}^{*T} = [y^*(t_1),\, y^*(t_2),\, \dots,\, y^*(t_N)]$. The noise, however, should be reduced to a certain extent before any analysis can be conducted. By smoothing $\underline{y}^{*T}$, a modified data set $\underline{y}^T$ is created that tries to keep important patterns in the data and, at the same time, to reduce the noise. Simonoff (1996) gives an overview of common smoothing techniques. Here, two different smoothing techniques are presented and considered for later use in the measurement reconciliation process.

3.1.1 Kernel Smoother

A kernel smoother is an estimation technique in non-parametric statistics that defines a set of weights w_j for each time of measurement t_j, $j = 1, \dots, N$. The shape of the weighting function $w(t)$ is described via the so-called kernel $K(\xi)$—a density function

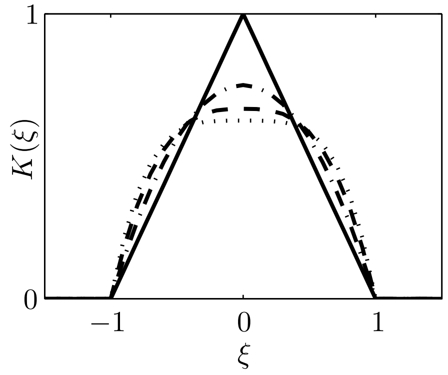


Figure 3.1: Graph of kernel function (3.5) for different p. Solid line: $p = 1$ (triangular kernel), dash-dot line: $p = 2$ (Epanechnikov kernel), dashed line: $p = 3$, dotted line: $p = 4$.

that satisfies the conditions

$$\int_{-\infty}^{+\infty} K(\xi)\ \mathrm{d}\xi = 1 \quad , \quad \int_{-\infty}^{+\infty} \xi K(\xi)\ \mathrm{d}\xi = 0 \quad , \quad \int_{-\infty}^{+\infty} \xi^2 K(\xi)\ \mathrm{d}\xi > 0 \quad . \tag{3.2}$$

The modified and smooth measurements are defined by

$$y(t_i) = \sum_{j=1}^{N} w_j y^*(t_j) \quad , \tag{3.3}$$

where the weights w_j are calculated by

$$w_j = w(t_j) = \frac{K\left(\dfrac{t_i - t_j}{h_t}\right)}{\sum\limits_{k=1}^{N} K\left(\dfrac{t_i - t_k}{h_t}\right)} \tag{3.4}$$

and h_t is a scale parameter that will be discussed below.

There are several different types of kernel functions that can be used. Some are given in Simonoff (1996). Here, the proposal of Hilberg (1989) is modified, i.e., the kernel

$$K(\xi) = \begin{cases} \dfrac{p+1}{2p}(1 - |\xi|^p) & \text{if } |\xi| \le 1 \quad , \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

is implemented where the user can choose p. When $p = 1$, a triangular density function is used as the kernel, and for $p = 2$, $K(\xi)$ is the Epanechnikov kernel (Epanechnikov, 1969). For different p, the graph of the kernel (3.5) can be found in Figure 3.1.

In this work, the scale parameter h_t is specified in a way that two demands are taken into account. First, only $2M+1$ measurements $y^*(t_j)$, $j \in [i-M,\, i+M]$, shall be considered when calculating $y(t_i)$. The value for M is chosen by the user. Second, in the case of nonuniform sampling, more importance has to be attached to measurements that are closer to $y^*(t_i)$ than to those that are farther away. Therefore,

$$h_t = M\frac{t_N - t_1}{N-1} \quad . \tag{3.6}$$

3.1.2 Whittaker Smoother

Another smoothing technique is by Whittaker (1922) and was rediscovered by Eilers (2003). Here, an algorithm is developed that balances the two conflicting goals of the smoothing procedure: On the one hand, the modified data $\underline{y}^T$ should be smooth. But on the other hand, it should not deviate too much from the original data $\underline{y}^{*T}$. However, the smoother $\underline{y}^T$ will be, the farther away it will be from $\underline{y}^{*T}$.

Now, two measures S and R are introduced to describe the discussed goals. The deviation of the smooth data $\underline{y}^T$ from the original data $\underline{y}^{*T}$ can be described by the sum of squared differences between original and modified data,

$$S = \sum_{j=1}^{N} (y^*(t_j) - y(t_j))^2 \quad . \tag{3.7}$$

The smoothness is measured by the sum of squared second-order differences of the smooth data set,

$$R = \sum_{j=3}^{N} (y(t_j) - 2y(t_{j-1}) + y(t_{j-2}))^2 \quad . \tag{3.8}$$

Then, S and R are combined in

$$Q = S + \lambda R \quad , \tag{3.9}$$

where a value for λ is yet to be allocated. The idea of the Whittaker smoother is to find data $\underline{y}^T$ that minimizes Q. The choice of λ determines the influence of the smoothness on this cost functional. The larger λ is, the smoother $\underline{y}^T$ will be.

Substituting Eqs. (3.7) and (3.8) into (3.9), Q can be calculated by

$$Q = \left(\underline{y}^{*T} - \underline{y}^T\right)\left(\underline{y}^{*T} - \underline{y}^T\right)^T + \lambda \underline{y}^T \mathbf{D}^T \mathbf{D} \left(\underline{y}^T\right)^T \quad , \tag{3.10}$$

where

$$\mathbf{D} = (d_{ij}) = \begin{cases} 1 & \text{if } j = i \quad , \\ -2 & \text{if } j = i+1 \quad , \\ 1 & \text{if } j = i+2 \quad , \\ 0 & \text{otherwise} \quad , \end{cases} \tag{3.11}$$

with $\mathbf{D} \in \mathbb{R}^{(N-2)\times N}$. Q then becomes minimal, when

$$\underline{y}^T = \underline{y}^{*T} \left(\mathbf{I} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1} \quad . \tag{3.12}$$

As a modification, the user does not choose λ but L, with $\lambda = 10^L$.

3.2 Piecewise Cubic Interpolation

As mentioned in Chapter 1, the qualitative description of the measurements is a necessary step to detect biological phenomena. To automate this step, the first derivatives of the measurements need to be calculated. This is done best with continuous data at hand. For this reason, the measurements are interpolated before they are described qualitatively. The interpolation task is not trivial, due to the slow dynamics, as fermentations may last several days and sampling intervals could be as long as several hours. This infrequent sampling makes interpolation even worse when no samples are taken overnight. Furthermore, measurements in biological systems are typically noisy, so the rare data set is uncertain to some extent as well. Hence, the choice of the interpolation method is crucial for a successful analysis. Common interpolation methods are shown by de Boor (2001). Here, only the piecewise cubic interpolation method is discussed and considered as it yields best results.

3.2.1 Conventional Cubic Spline Interpolation

Given the smoothed measurements $[y(t_1),\, y(t_2),\, \ldots,\, y(t_N)]$, an interpolation $\mathcal{P}$ can be constructed that consists of piecewise cubic polynomials $\mathcal{P}_i$, $i = 1, \ldots, N-1$,

$$\mathcal{P}(t) = \mathcal{P}_i(t) \quad , \quad t \in [t_i,\, t_{i+1}] \quad , \tag{3.13}$$

with

$$\mathcal{P}_i(t) = p_{0,\,i} + p_{1,\,i}(t - t_i) + p_{2,\,i}(t - t_i)^2 + p_{3,\,i}(t - t_i)^3 \quad . \tag{3.14}$$

Each polynomial $\mathcal{P}_i$ has to satisfy the conditions

$$\mathcal{P}_i(t_i) = y(t_i) \quad , \quad \mathcal{P}_i(t_{i+1}) = y(t_{i+1}) \quad , \quad \left.\frac{\mathrm{d}\mathcal{P}_i}{\mathrm{d}t}\right|_{t_i} = s_i \quad , \quad \left.\frac{\mathrm{d}\mathcal{P}_i}{\mathrm{d}t}\right|_{t_{i+1}} = s_{i+1} \quad . \tag{3.15}$$

The slopes s_j, $j = 1, \ldots, N$, are free parameters. The polynomial $\mathcal{P}(t)$ is continuous and has a continuous first derivative.

The coefficients are computed by

$$p_{0,\,i} = y(t_i) \quad , \tag{3.16a}$$

$$p_{1,\,i} = s_i \quad , \tag{3.16b}$$

$$p_{2,\,i} = \frac{3z_i - 2s_i - s_{i+1}}{t_{i+1} - t_i} \quad , \tag{3.16c}$$

$$p_{3,\,i} = \frac{s_i + s_{i+1} - 2z_i}{(t_{i+1} - t_i)^2} \quad , \tag{3.16d}$$

where

$$z_i = \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \quad . \tag{3.16e}$$

Now, the slopes s_j have to be determined. When cubic spline interpolations are calculated, the polynomial $\mathcal{P}(t)$ needs to be twice continuously differentiable, i.e., for $i = 2, \ldots, N - 1$,

$$\left.\frac{\mathrm{d}^2\mathcal{P}_{i-1}}{\mathrm{d}t^2}\right|_{t_i} = \left.\frac{\mathrm{d}^2\mathcal{P}_i}{\mathrm{d}t^2}\right|_{t_i} \quad . \tag{3.17}$$

Applying Eqs. (3.14)–(3.17) leads to

$$2p_{2,\,i-1} + 6p_{3,\,i-1}(t_i - t_{i-1}) = 2p_{2,\,i} \quad . \tag{3.18}$$

Finally, with Eqs. (3.16c) and (3.16d), the linear system

$$s_{i-1}(t_{i+1}-t_i)+2s_i(t_{i+1}-t_{i-1})+s_{i+1}(t_i-t_{i-1}) = 3\left((t_{i+1} - t_i)z_{i-1} + (t_i - t_{i-1})z_i\right) \tag{3.19}$$

is established. With Eq. (3.19), $N-2$ unknown slopes $s_2, \ldots, s_{N-1}$ can be calculated. The remaining slopes s_k, $k = 1, N$, have to be chosen somehow. Typical approaches are:

- Complete cubic spline: $s_k = \frac{\mathrm{d}y_k}{\mathrm{d}t}$, if $\frac{\mathrm{d}y_k}{\mathrm{d}t}$ is known,
- Natural cubic spline: $\left.\frac{\mathrm{d}^2\mathcal{P}}{\mathrm{d}t^2}\right|_{t_k} = 0$,
- Not-a-knot condition: Choose s_k so that $\mathcal{P}_1 = \mathcal{P}_2$ and $\mathcal{P}_{N-2} = \mathcal{P}_{N-1}$.

3.2.2 Akima Interpolation

In many cases, conventional cubic splines tend to over- and undershoot or produce wiggles (King et al., 2002). Akima (1970) introduces a method that calculates curves which are similar to those manually drawn by “a well-trained scientist or engineer” and which show a more natural behavior, i.e., without oscillations.

Proposition

The slopes s_j, $j = 1, \ldots, N$ are approximated locally, i.e., $\mathcal{P}_j$ depends only on information from, or near $[t_j,\, t_{j+1}]$, instead of the whole set of points. They are determined under a geometrical condition. Generally,

$$s_j = w_{j-1} z_{j-1} + w_j z_j \quad . \tag{3.20}$$

The weights w_{j-1} and w_j are defined by

$$w_{j-1} = \begin{cases} 0.5 & \text{if } z_{j-2} = z_{j-1} \neq z_j = z_{j+1} \quad , \\ \dfrac{|z_{j+1} - z_j|}{|z_{j+1} - z_j| + |z_{j-1} - z_{j-2}|} & \text{otherwise} \quad , \end{cases} \tag{3.21a}$$

or

$$w_j = \begin{cases} 0.5 & \text{if } z_{j-2} = z_{j-1} \neq z_j = z_{j+1} \quad , \\ \dfrac{|z_{j-1} - z_{j-2}|}{|z_{j+1} - z_j| + |z_{j-1} - z_{j-2}|} & \text{otherwise} \quad , \end{cases} \tag{3.21b}$$

respectively. With Eq. (3.16e) it can be seen that five measurements $y(t_{j-2}), \ldots, y(t_{j+2})$ are needed to calculate s_j, i.e., additional end points $y(t_{-1})$, $y(t_0)$, $y(t_{N+1})$, $y(t_{N+2})$ have to be estimated to determine s_1, s_2, s_{N-1}, and s_N. Two quadratic polynomials $\mathcal{G}_k(t)$, $k = 1, N$, are assumed to describe these additional points,

$$\mathcal{G}_k(t) = g_{0,k} + g_{1,k}(t - t_k) + g_{2,k}(t - t_k)^2 \quad , \tag{3.22}$$

with

$$t_{k+2} - t_k = t_{k+1} - t_{k-1} = t_k - t_{k-2} \quad . \tag{3.23}$$

The coefficients of the piecewise cubic polynomials $\mathcal{P}_i$ are then calculated by Eqs. (3.16a)–(3.16d).

Modifications

Fried and Zietz (1973) investigate the utility of the Akima method for the analysis of biological data. They note that overshoots occur when one of the slopes z_i or z_{i+1} is small, and the other one is large. However, this effect can be reduced by some modifications.

At first, two conditions

$$|z_{i-1}| > C_0 \wedge |z_i| < C_0^{-1} \quad , \tag{3.24a}$$

and

$$|z_{i-1}| < C_0^{-1} \wedge |z_i| > C_0 \quad , \tag{3.24b}$$

are evaluated, where C_0 is a large number, here $C_0 = 1000$. Then, new weights

$$w_{j-1}^* = \sqrt{|(z_j - z_{j-2})(z_{j+1} - z_j)|}/R \tag{3.25a}$$

and

$$w_j^* = \sqrt{|(z_{j-1} - z_{j-2})(z_{j+1} - z_{j-1})|}/R \tag{3.25b}$$

are calculated, where

$$R = \sqrt{|(z_j - z_{j-2})(z_{j+1} - z_j)|} + \sqrt{|(z_{j-1} - z_{j-2})(z_{j+1} - z_{j-1})|} \quad . \tag{3.25c}$$

If condition (3.24a) applies and $w_j^*/w_{j-1}^* > w_j/w_{j-1}$, then w_j and w_{j-1} (Eqs. (3.21a) and (3.21b)) are substituted by w_j^* and w_{j-1}^*. In the case of (3.24b), the substitution takes place if $w_{j-1}^*/w_j^* > w_{j-1}/w_j$.

Another modification affects the determination of $y(t_{-1})$, $y(t_0)$, $y(t_{N+1})$, $y(t_{N+2})$. Instead of the parabolic extrapolation (3.22), points are added that have the same values as $y(t_1)$ or $y(t_N)$, respectively, i.e., $y(t_{-1}) = y(t_0) = y(t_1)$ and $y(t_{N+2}) = y(t_{N+1}) = y(t_N)$.

3.2.3 Yeh Interpolation

An alternative to calculating the slopes s_j is shown by Yeh and Small (1989), motivated by poor results of the numerical integration of the interpolation functions received by older methods.

Here, three different cases are analyzed and the corresponding slopes s_j are calculated.

Case 1: $z_{i-1}z_i > 0$, $i = 2, \ldots, N-1$. This case represents monotone data over $[t_{i-1}, t_{i+1}]$. The slopes are defined by

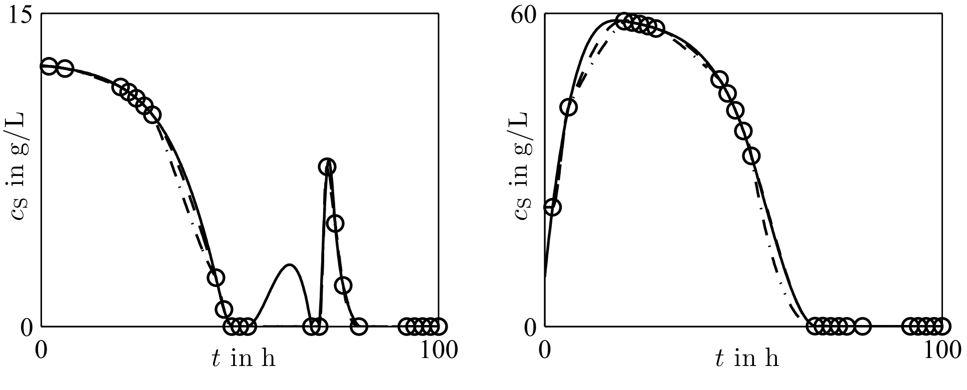
$$s_i = \begin{cases} \dfrac{z_{i-1}z_i}{w_i z_i + (1-w_i)z_{i-1}} & \text{if } y(t_{i-1}) \cdot y(t_i) \cdot y(t_{i+1}) = 0 \quad , \\[2ex] \dfrac{r_{i-1}r_i y(t_i)}{w_i r_i + (1-w_i)r_{i-1}} & \text{otherwise} \quad , \end{cases} \tag{3.26}$$

where

$$w_i = \frac{1}{3}\left(1 + \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}\right) \quad , \tag{3.27a}$$

$$r_{i-1} = \frac{1}{t_i - t_{i-1}} \ln\left(\frac{y(t_i)}{y(t_{i-1})}\right) \quad , \tag{3.27b}$$

$$r_i = \frac{1}{t_{i+1} - t_i} \ln\left(\frac{y(t_{i+1})}{y(t_i)}\right) \quad . \tag{3.27c}$$



(a) ME1 (b) ME2

Figure 3.2: Comparison of different interpolation methods. Measurements of substrate concentration (○) of two different experiments, cubic spline interpolations with not-a-knot condition (solid line), Akima interpolations (dashed line), and Yeh interpolations (dash-dot line) are shown.

Case 2: $z_{i-1}z_i \leq 0$, $i = 2, \ldots, N-1$. Here, all intervals are described where the data is not monotonic. The slopes are calculated according to

$$s_i = \begin{cases} w_i z_{i-1} + (1 - w_i) z_i & \text{if } y(t_i) = \max\limits_j y(t_j) \quad , \\ 0 & \text{otherwise} \quad . \end{cases} \tag{3.28}$$

Case 3: End points. The first two cases applied to interior data points whose slopes are calculated by the data point itself and the left and right neighboring point. Since one of these neighbors is missing in the case of the end points, their slopes are calculated differently:

$$s_1 = \frac{3z_1}{2} - \frac{s_2}{2} \quad , \tag{3.29}$$

$$s_N = \frac{3z_{N-1}}{2} - \frac{s_{N-1}}{2} \quad . \tag{3.30}$$

Figure 3.2 compares the result of the mentioned methods. As can be seen in Figure 3.2(a) at around 60 h, conventional cubic splines tend to overshoot whereas the Akima and Yeh interpolations are able to maintain a constant value.

A lot of effort has been spent to automate the steps of interpolation. However, after many tests with real data from very different cultivations, a satisfying solution could not be found. Instead, both the Akima method (together with the mentioned modifications) and the Yeh method are implemented to determine the interpolation of the measurements. The results are presented to the user who, then, individually selects one of the methods as well as the interval when the interpolating Polynomial $\mathcal{P}(t)$ should be evaluated. This interval will be called Δt_{Intp} in the following.

Doing so, the interpolations highlight what the experienced modeler deems to be adequate. This ‘manual’ step, though supported by automatic interpolations, is essential for finding good model candidates. The software approach presented in this work frees the human modeler from many tedious tasks but does not replace him or her.

3.3 Compensation for Feeding and Sampling

The approach presented in this work is not restricted to batch cultivations without sampling but considers fed-batch cultivations of microorganisms as well. Feeding and selective sampling, e.g., via a membrane, change the dynamic evolution of the observed quantities. Hence, variations in the measurements cannot only be attributed to biological phenomena but likewise to these external manipulations. Furthermore, feeding and sampling lead to a dilution. If a substrate is fed, the concentrations of the other reactants decrease because of the change in volume. A decreasing concentration does not necessarily mean that a (bio-)reaction takes place. Feeding and sampling may indicate biological phenomena not present or may hide true biological phenomena. Thus, the feeding flow rates have to be considered as an extra source of information to gain insight into the reactions taking place. Instead of looking at two sources of information at the same time and to facilitate the following steps, these sources of information are combined by compensating for feeding and sampling. It has to be pointed out that the compensation is supposed to highlight the qualitative nutrient situation to detect phenomena but not quantitative values as shown below. The compensation is best explained by starting at the end of the experiment.

A continuous measurement is assumed first to simplify the description of the compensation for feeding and sampling. The interpolation of non-continuous data is discussed below. Figure 3.3 shows the result of a qualitative compensation of a measurement. A substrate concentration $c_{\mathrm{S}}(t)$ and the corresponding feeding rate $u_{\mathrm{S}}(t)$ are depicted, as well as the compensated concentration $\tilde{c}_{\mathrm{S}}(t)$, whose dynamic is independent of the feeding.

In interval IV, the measured substrate concentration is zero. However, the substrate is fed in the interval as seen from $u_{\mathrm{S}}(t)$. Hence, the substrate is consumed in some reaction. To highlight this, $c_{\mathrm{S}}(t)$ has to be compensated. Every measurement has to be increased by the amount of substrate which has been fed till the corresponding time t. By doing so, a decreasing evolution of the compensated concentration $\tilde{c}_{\mathrm{S}}(t)$ is obtained, indicating a consumption of the substrate inside the reactor. The sampling is compensated for in a similar manner. The amount of the substrate S which has been sampled up to a time t has to be subtracted. Considering that the measurements are not continuous but are taken at a certain sampling time t_j, the compensated total mass $\tilde{m}_{\mathrm{S}}(t_j)$ of substance S at time $t = t_j$ is calculated by

$$\tilde{m}_{\mathrm{S}}(t_j) = c_{\mathrm{S}}(t_j) \cdot V(t_j) + \sum_{i=j}^{N} u_{\mathrm{S}}(t_i) \cdot c_{\mathrm{S,\,Feed}} \cdot \Delta t_i - \sum_{i=j}^{N} \Delta m_{\mathrm{S}}(t_i) \quad , \tag{3.31}$$

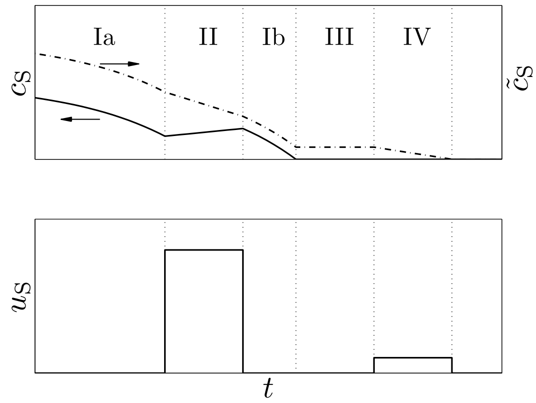


Figure 3.3: Compensation of feeding in measurements. $c_S(t)$ (solid line)—measured concentration, u_S—corresponding feeding, $\tilde{c}_S(t)$ (dash-dot line)—compensated concentration according to Eq. (3.31).

where $c_S(t_j)$ is the measured substrate concentration, $V(t_j)$ the volume, N the last time point considered, $u_S(t_i)$ the constant flow rate in the interval $[t_i,\, t_i + \Delta t_i]$, $c_{S,\,\mathrm{Feed}}$ the substrate concentration in the feed, and $\Delta m_S(t_i)$ the mass of reactant S in a sample taken at time t_i. It has to be mentioned again that the compensation starts at the end of the experiment and only feeding and sampling in the interval $[t_j,\, t_N]$ are considered. Similarly, a general reactant R which is not fed to the reactor is compensated with $u_R = 0$. Compensated concentrations $\tilde{c}(t_j)$ can then be calculated by

$$\tilde{c}_S(t_j) = \frac{\tilde{m}_S(t_j)}{V(t_N)} \quad . \tag{3.32}$$

Considering the situation in interval IV, the feeding rate and the consumption rate are in equilibrium, the amount of substrate fed is consumed immediately. Effectively, the cumulative, i.e., compensated substrate $\tilde{c}_S(t)$ increases when looking backward in time.

In interval III, the measured substrate vanishes as well. Since there is no feeding, the compensated concentration $\tilde{c}_S$ is constant and unequal to zero. No substrate consumption takes place.

Interval II shows an increasing substrate concentration. Possible causes can be either lysing cells, which provide additional substrates, or a substrate feeding. As a matter of fact, at this time, substrate is fed and therefore it has to be compensated for. The compensated concentration decreases when looking forward in time, i.e., the substrate is actually consumed. However, it is fed more than the microorganisms are able to consume which leads to an increase in the measured concentration c_S. It is obvious that, for a later analysis, a look at the measured concentration without taking feeding into account would lead to wrong conclusions.

In the intervals Ia and Ib, the measured substrate concentrations decrease. Since there is no feeding, the compensated and the measured concentrations show the same parallel trend. The substrate is actually consumed.

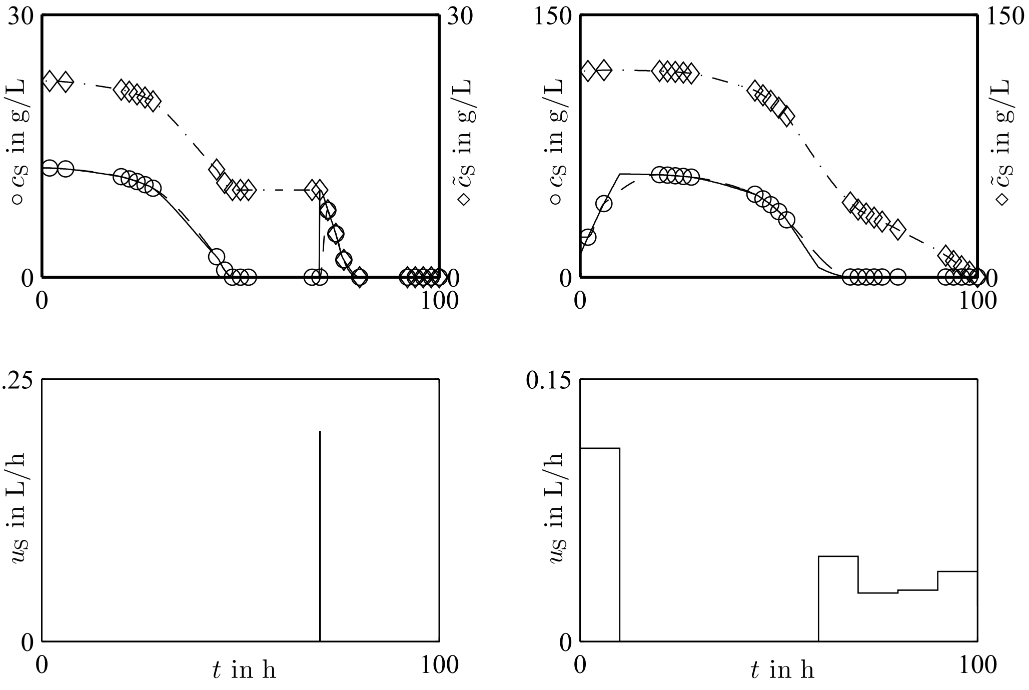
Now, the assumption of continuous measurements—that was only chosen for ease of presentation—is relaxed. In the next step, the measured and compensated data sets are interpolated by cubic splines as described in Section 3.2 at times $t_i = k \cdot \Delta t_{\mathrm{Intp}}$, with $k = 0, 1, 2, \ldots$. In addition to the interpolations, the first derivatives $\dot{c}(t)$ are also calculated as they will be needed to characterize an episode.

Moreover, instead of interpolating compensated and non-compensated concentrations independently of each other, these interpolations can be linked together, i.e., the interpolations of the compensated measurements are used to calculate the interpolations of the non-compensated measurements. For this purpose, the compensated concentrations $\underline{\tilde{c}}_{\mathrm{S}}^{T} = [\tilde{c}_{\mathrm{S}}(t_1), \tilde{c}_{\mathrm{S}}(t_2), \ldots, \tilde{c}_{\mathrm{S}}(t_N)]$ are calculated according to Eqs. (3.31) and (3.32) and interpolated as described above. Then, the interpolation of the real concentrations, $c_{\mathrm{S}}(t_k) = c_{\mathrm{S}}(k \cdot \Delta t_{\mathrm{Intp}})$, $k = 0, 1, 2 \ldots$, are calculated by

$$c_{\mathrm{S}}(t_k) = \frac{1}{V(t_k)} \cdot \left(\tilde{c}_{\mathrm{S}}(t_k) \cdot V(t_N) - \sum_{t_i=t_k}^{t_N} u_{\mathrm{S}}(t_i) \cdot c_{\mathrm{S,\,Feed}} \cdot \Delta t_i + \sum_{t_i=t_k}^{t_N} \Delta m_{\mathrm{S}}(t_i) \right) \quad . \tag{3.33}$$

Figure 3.4 compares the linked and non-linked interpolations of the substrate S in the experiments ME1 and ME2 ($\Delta t_{\mathrm{Intp}} = 0.2\,\mathrm{h}$). As can be seen, the interpolations that are calculated from the compensated measurements are able to show dynamics that the conventional interpolations do not. In Figure 3.4(a), the dynamics resulting from the pulse conducted at $t = 70\,\mathrm{h}$ are described well by the linked interpolations. The concentration of the fed substrate also increases with a pulse whereas the conventional interpolation only connects the measurements before and after the pulse. The same effect can be seen in Figure 3.4(b) as well. When the flow rate u_{S} is changed between two measurements (e.g., $t = 10\,\mathrm{h}$), the influence on the dynamics of c_{S} is described by the linked interpolation. Hence, better results are usually obtained with linked interpolations since it reduces the risk of an unwanted removal of the feeding and sampling dynamics.

The approaches described above apply to substrates and product concentrations in the fermenter broth. For compartment, i.e., structured models, the intracellular compounds have to be discussed separately. If intracellular concentrations are given with respect to the total cell mass or volume, i.e., $g_i(t) = m_i(t)/V_{\mathrm{X}}(t)$, these are not affected by feeding or sampling. The cell volume V_{X} can be expressed in terms of biomass if a constant density of the cell is assumed. However, if concentrations of intracellular components are given based on the fermenter volume, the compensation method applies as well.



(a) ME1

(b) ME2

Figure 3.4: Comparison of linked and non-linked interpolations. Shown are the compensated concentrations $\tilde{c}_S$ (◇) with their interpolations (dash-dot line) and the measured concentrations c_S (○) with the interpolations calculated from the 'compensated interpolations' (solid line) of two different experiments. The non-linked interpolations (dashed line) are shown as well.

3.4 Episodes

As mentioned in Chapter 1, the definition of a biological phenomenon is based on changes in the qualitative behavior of several substances in the process network. Therefore, a qualitative representation has to be found. Cheung and Stephanopoulos (1990) introduce such a representation based on the gradient and the curvature of a signal at a certain time. This leads to seven possible qualitative states they call *episodes*. Here, the episodes are detected on the basis of the compensated curves $\tilde{c}(t)$. However, the curvatures of these graphs are not considered in this work. Only the gradients are used for the episode detection—leading to three possible episodes: increasing, decreasing, and constant.

Although the compensated concentrations $\tilde{c}(t)$ are used to determine the episodes, the influence of the non-compensated measurements $c(t)$, however, cannot be completely neglected. Taking a look back at Figure 1.2 (page 5), a correlation between the substrate S and the product P can be assumed. In the pulse experiment (ME1, Figure 1.2(a)) at around $t = 50\,\text{h}$, the substrate is depleting while the product starts to grow. When S is fed at $t = 70\,\text{h}$, the growth of P stops and does not continue until

S depletes again at around $t = 80\,\mathrm{h}$. The same relationship between S and P can be detected in the fed-batch experiment ME2 (Figure 1.2(b)), starting at $t = 70\,\mathrm{h}$. From Eq. (1.4),

$$r_{\mathrm{P}}(t) = r_{\mathrm{Pm}} \cdot \frac{K_{\mathrm{PS}}}{c_{\mathrm{S}}(t) + K_{\mathrm{PS}}} \cdot c_{\mathrm{X}}(t) \quad ,$$

describing the growth rate of P, it becomes clear that the smaller $c_{\mathrm{S}}(t)$ is, the larger $r_{\mathrm{P}}(t)$ will be. When $c_{\mathrm{S}}(t)$ falls below a certain threshold, changes in $r_{\mathrm{P}}(t)$ become perceivable and, thus, visible in $c_{\mathrm{P}}(t)$.

The result of the compensation can be seen in Figure 3.4. In experiment ME1 (Figure 3.4(a)), using conventional episodes, the trend of the compensated substrate concentration $\tilde{c}_{\mathrm{S}}(t)$ can be divided into four episodes (decreasing–constant–decreasing–constant). From this, it does not become clear that the aforementioned threshold is hit. An additional episode 'zero' is able to show this important information. In experiment ME2 (Figure 3.4(b)), a conventional episode detection will only find that $\tilde{c}_{\mathrm{S}}(t)$ is decreasing the whole time, since substrate is fed when low substrate concentrations occur. However, as discussed above, low substrate concentrations have an impact on other measured variables—independent of the feeding. To be able to detect such a relation, a further episode 'decreasing with measurements being zero' is introduced.

Therefore, when the compensated curves $\tilde{c}(t)$ show a constant or decreasing behavior, the measurements $c(t)$ are tested for being zero. Hence, five possible episodes can be detected: increasing (+), decreasing with measurements not being zero (−), decreasing with measurements being zero ($\ominus$), constant and not zero (c), and zero (0).

From Figure 3.3, interval II, it becomes clear why compensated values are chosen. The gradient of $\tilde{c}_{\mathrm{S}}$ clearly indicates that the substrate is being consumed, whereas the evaluation of c_{S} might lead to a wrong, opposite conclusion.

The use of the derivative of a filtered interpolation to calculate episodes of the original measurements may be error-prone because the result will strongly depend on the quality of the data reconciliation and interpolation. This problem would be even more severe, when the curvature would be included as well. To deal with this problem, a probabilistic approach is used here to lower the effect of the imperfect interpolation onto the detection of episodes.

For this purpose, and in contrast to Cheung and Stephanopoulos (1990), a procedure is applied where the episodes are not evaluated at a certain time t_i, but in a moving time frame Δt_{Ep} around t_i. At first, a tolerance band $\Delta\dot{\tilde{c}}_{\mathrm{Tol}}$ is calculated (see below) in which the values $\dot{\tilde{c}}(t)$ are assumed to be zero. Then, it is determined which part of $\dot{\tilde{c}}(t)$ in the aforementioned time frame is enclosed by values greater than zero ($A_+(t_i)$), which part is enclosed by values lower than zero ($A_-(t_i)$), and which part is within the tolerance band $\Delta\dot{\tilde{c}}_{\mathrm{Tol}}$ ($A_c(t_i)$) (see Figure 3.5). The areas A_j, $j \in \{+, -, c\}$, in a time interval $[t_a, t_b]$ are numerically approximated by the trapezoidal rule,

$$A_j\Big|_{t_a}^{t_b} = A_j(c)\Big|_{t_a}^{t_b} = \int_{t_a}^{t_b} \dot{\tilde{c}}(\tau)\ \mathrm{d}\tau \approx \frac{1}{2}(t_b - t_a)\left(\dot{\tilde{c}}(t_a) + \dot{\tilde{c}}(t_b)\right) \quad . \tag{3.34}$$

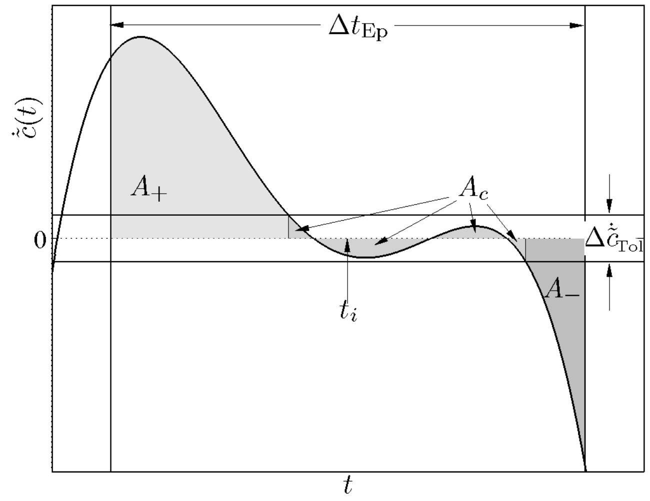


Figure 3.5: Calculation of episodes. In a moving time frame Δt_{Ep} it is determined which part of the concentration gradient $\dot{\breve{c}}(t)$ is enclosed by values within a tolerance band $\Delta\dot{\breve{c}}_{\mathrm{Tol}}$ where the values are assumed to be zero (A_c), by values greater than zero (A_+), and by values lower than zero (A_-). Episode probabilities are calculated by $P_j = A_j / \sum A_j$. For more details see text.

All areas are described by their absolute values. By relating each individual area $A_+(t_i)$, $A_-(t_i)$, or $A_c(t_i)$ to the aggregated area

$$A_\Sigma(t_i) = A_+(t_i) + A_-(t_i) + A_c(t_i) \quad , \tag{3.35}$$

the result is a measure of probability that within Δt_{Ep} the compensated measurement increases, decreases, or is constant. Two episodes are described by decreasing ($-$ or $\ominus$) and constant graphs (c or 0), respectively. To distinguish between cases with variables being equal or unequal to zero by introducing an additional tolerance Δc_{Tol}, the calculation procedure has to be applied to the interpolation of the measured (not compensated) data set as well to determine values for A_+^* (measurements unequal to zero) and A_0^* (measurements equal to zero). The aggregated area $A_\Sigma^*(t_i)$ is defined by

$$A_\Sigma^*(t_i) = A_+^*(t_i) + A_0^*(t_i) \quad , \tag{3.36}$$

negative concentration values (and thus $A_-^*(t_i)$) do not exist.

The probability of an increasing episode is then calculated by

$$P_+(t_i) = P_+(t_i, c) = \frac{A_+(t_i)}{A_\Sigma(t_i)} \quad , \tag{3.37a}$$

and the other probabilities by

$$P_-(t_i) = P_-(t_i,\, c) = \frac{A_-(t_i)}{A_\Sigma(t_i)} \frac{A^*_+(t_i)}{A^*_\Sigma(t_i)} \quad , \tag{3.37b}$$

$$P_\ominus(t_i) = P_\ominus(t_i,\, c) = \frac{A_-(t_i)}{A_\Sigma(t_i)} \frac{A^*_0(t_i)}{A^*_\Sigma(t_i)} \quad , \tag{3.37c}$$

$$P_c(t_i) = P_c(t_i,\, c) = \frac{A_c(t_i)}{A_\Sigma(t_i)} \frac{A^*_+(t_i)}{A^*_\Sigma(t_i)} \quad , \tag{3.37d}$$

$$P_0(t_i) = P_0(t_i,\, c) = \frac{A_c(t_i)}{A_\Sigma(t_i)} \frac{A^*_0(t_i)}{A^*_\Sigma(t_i)} \quad . \tag{3.37e}$$

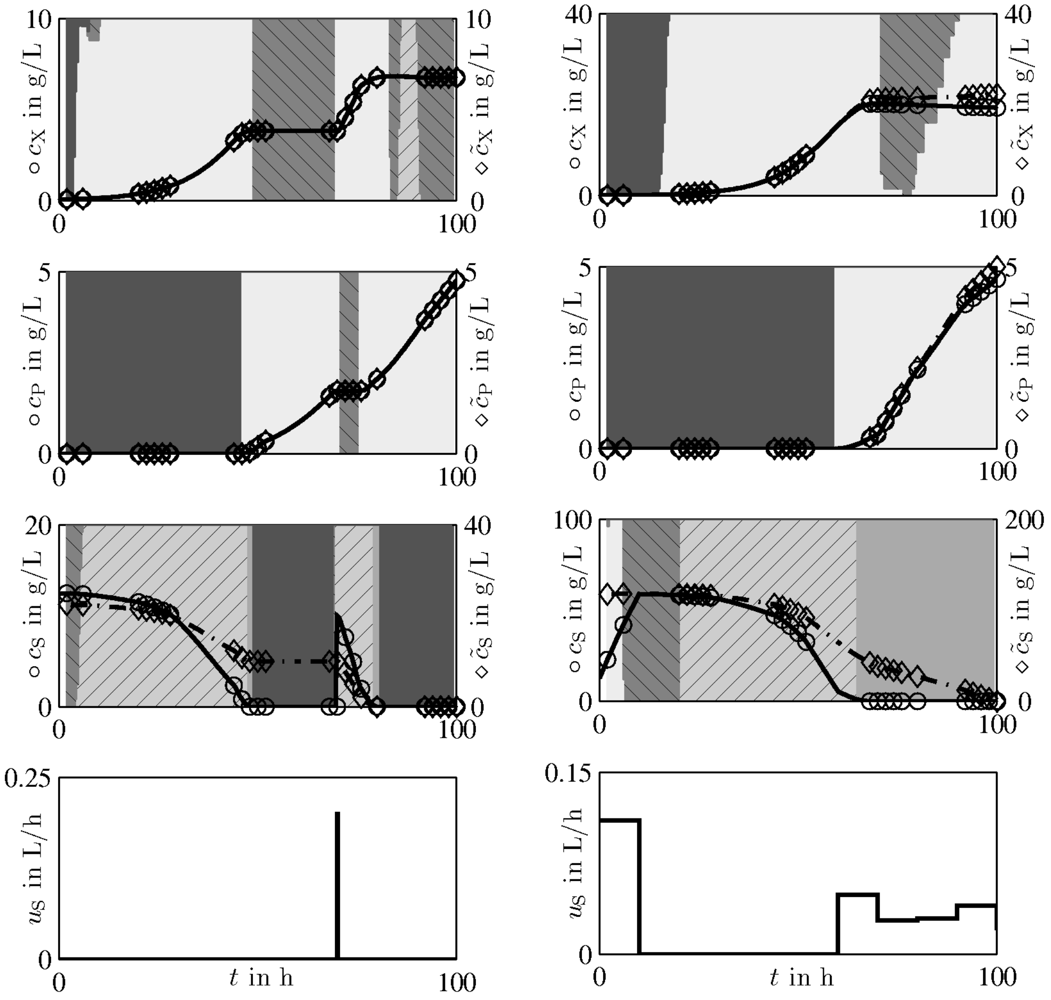
The episode probabilities are calculated at all interpolation times t_i with $\Delta t_{\text{Ep}}/2 \leq t_i \leq t_{\text{end}} - \Delta t_{\text{Ep}}/2$ where t_{end} is the end of the experiment.

It is understood that the sizes of the tolerance band $\Delta\dot{\tilde{c}}_{\text{Tol}}$ and the time frame Δt_{Ep} have considerable influence on the detection of biological phenomena. The choice of Δt_{Ep} depends on the dynamics of the microorganism under consideration. If it is too big, important changes in the attributes can be overlooked, if it is too small, little changes can be overrated. For the organism analyzed in this work and with experiments lasting 100 h–150 h, good results are achieved with $\Delta t_{\text{Ep}} = 2\,\text{h}$. However, the user can adjust the value for Δt_{Ep} to other microorganisms. Finding an appropriate size for Δc_{Tol} and especially $\Delta\dot{\tilde{c}}_{\text{Tol}}$ is a more tedious task. The different measured variables need different specific values which can even differ from experiment to experiment. As a result, it is expected that this cannot be fully automated. Instead, the human expert will still be needed in the future for this critical step in model building. The method and tools developed here, however, will significantly help him or her in making a decision on a rational basis and speed up the overall process of model building. In this work, $\Delta c_{\text{Tol}} = 1/20 \cdot \sigma(c(t))$ and $\Delta\dot{\tilde{c}}_{\text{Tol}} = \min(1/12 \cdot \sigma(\dot{\tilde{c}}(t)),\, 5 \times 10^{-3})$ are chosen by default, with $\sigma(\cdot)$ being the standard deviation of $(\cdot)$ in an experiment calculated with respect to the median instead of the average value.

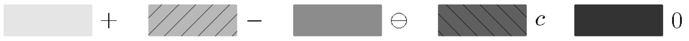
Figure 3.6 shows the detected episodes of the biomass, the substrate, and the product concentrations of the experiments ME1 and ME2 (see Chapter 1). Additionally, the results of the compensation and the interpolation can be seen.

3.5 Landmarks

Having calculated the five episode probability curves for each measured variable, the next step is to detect transitions of episodes, which are indicated by the so-called *landmarks* (Cheung and Stephanopoulos, 1990). For instance, the landmark $(+c)$ shows that the episode $+$ changes into c. As there are five possible episodes before and after the transition, $5^2 = 25$ possible landmarks can be determined, five of which obviously do not indicate any transitions. Others do not make sense such as $(+0)$ or $(c0)$. Some important landmarks, which will be used later for the detection of



(a) ME1 (b) ME2



(c) Episodes

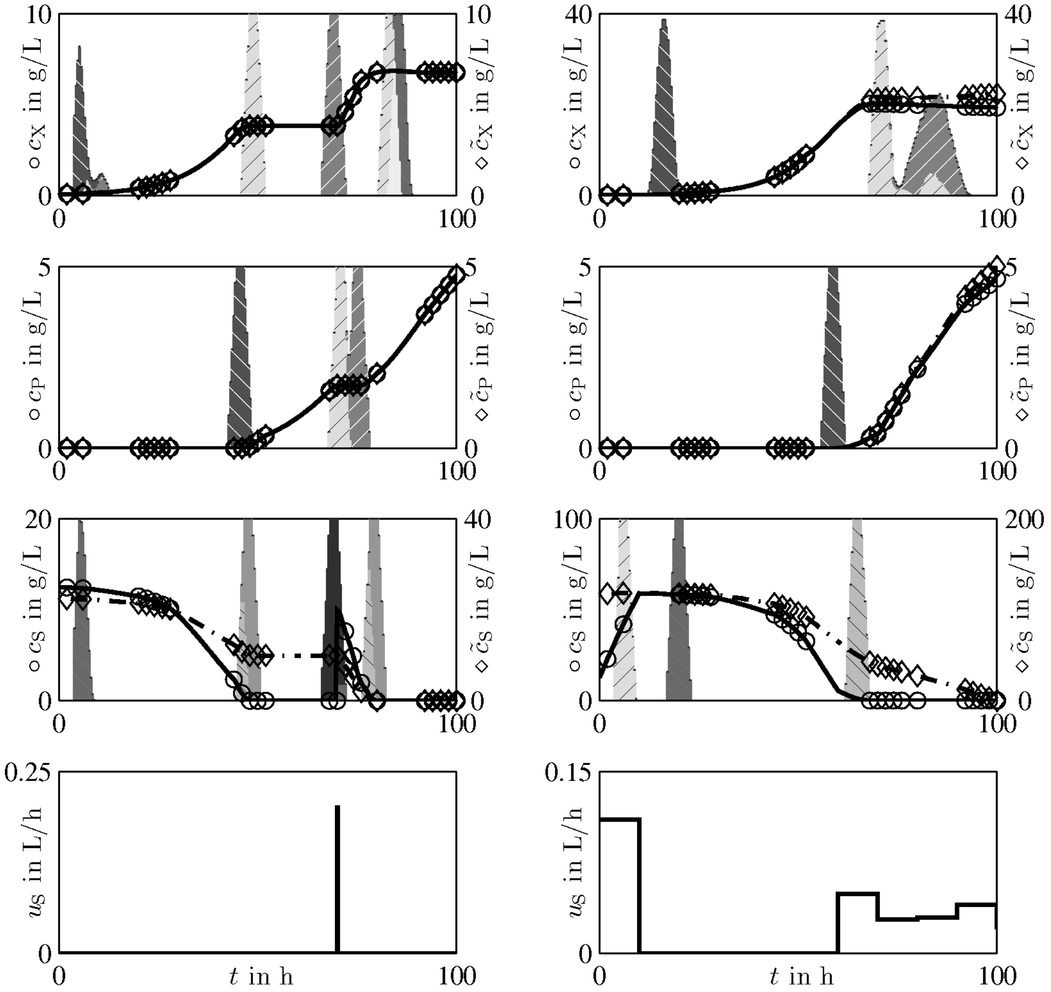
Figure 3.6: Episode detection of concentrations from the in-silico experiments ME1 and ME2. Shown are the measured concentrations c_i (○) and their interpolations (solid line), the compensated concentrations $\tilde{c}_i$ (◇) and their interpolations (dash-dot line). The episode probabilities are displayed as stacked bars, the sum of all bars is 1 at each time instant. For better readability the bars are scaled to the length of the chosen concentration axis. The different shades of gray are assigned to different episodes. Possible episodes are: increasing (+), decreasing with measurements not being zero (−), decreasing with measurements being zero (⊖), constant and not zero (c), and zero (0). The indices are X—biomass, P—product, and S—substrate. The feeding rates are given by u_S.

biological phenomena, are $(+-)$, $(+c)$, $(-+)$, $(-\ominus)$, (-0), $(c+)$, $(c-)$, $(0+)$, and $(0-)$. The latter one seems to be unusual or impossible in contrast to $(c-)$, considering that a concentration cannot decrease after it has already become zero. But, again, not the measured but the compensated concentrations are used to determine the episodes. The landmark $(0-)$ means that a substance is fed after it has been depleted for a certain period of time (e.g., see $c_\mathrm{S}(t)$ in Figure 3.7(a), at around $t = 70\,\mathrm{h}$). The 0 in $(0-)$ is only used to indicate that the real substrate was zero before the landmark.

Similar to the episodes, the landmark probabilities are calculated at all interpolation times $\Delta t_\mathrm{Lm}/2 \le t_i \le t_\mathrm{end} - \Delta t_\mathrm{Lm}/2$, where Δt_Lm is a moving time frame. The probability $P_{(kl)}(t_i)$ of landmark (kl), $k,\ l \in \{+,\ -,\ \ominus,\ c,\ 0\}$, is determined by

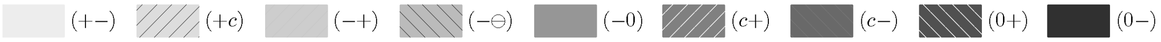
$$P_{(kl)}(t_i) = P_{(kl)}(t_i,\ c) = \bar{P}_k(t_i-) \cdot \bar{P}_l(t_i+) \quad , \tag{3.38}$$

where $\bar{P}_k(t_i-)$ is the average probability of episode k in the time interval $[t_i-(\Delta t_\mathrm{Lm}/2),\ t_i)$ and $\bar{P}_l(t_i+)$ is the average probability of episode l in $(t_i,\ t_i + (\Delta t_\mathrm{Lm}/2)]$. In this work, for all the organisms analyzed, $\Delta t_\mathrm{Lm} = 8\,\mathrm{h}$ is chosen. However, this value can be adapted to other microorganisms by the user as well. The detected landmarks of the experiments ME1 and ME2 can be found in Figure 3.7.



(a) ME1

(b) ME2



(c) Landmarks

Figure 3.7: Landmark detection of concentrations from the in-silico experiments ME1 and ME2. The landmark probabilities are displayed as stacked bars, the sum of all bars is 1 at each time instant. The different shades of gray are assigned to different chosen landmarks. For more information, see Figure 3.6.

Chapter 4

Biological Phenomena

The measurement reconciliation steps described in Chapter 3 finally lead to the detection of episodes and landmarks which describe the qualitative behavior of a measurement variable at a certain time and the change of that behavior. Information that is obvious for human experts when looking at the measurements is now provided in a format that a computer can use to imitate the next step in the model-building process: finding correlations between different measurement variables to formulate interactions and dependencies between the substances in the process under consideration.

Possible interactions between several substances can be found best when one of them changes its qualitative behavior. If other substances change their behavior as well, it seems possible that the relevant substances are connected with each other in the underlying reaction network. An example can be seen in Figure 3.7(a), where all three measurement variables show a change in their qualitative behavior—depicted by the different landmarks—at around $t = 50\,\mathrm{h}$. They are connected with each other in a way that is yet to be determined. Therefore, to automatically find correlations between two or more reactants, the measurements are analyzed for several landmark combinations that occur at around the same time.

That way, finding that two substances are independent of each other can be done easily. Every time, a change in a substance A occurs without an effect on the qualitative behavior of a substance B, and vice versa, this can be understood as a sign that these substances are not correlated. Examples can be found in Figure 4.1, where for two substances A and B combinations of several episode transitions, i.e., landmarks, are shown that do not indicate any interaction between these two substances.

However, to find dependencies between two or more substances, analyzing every possible combination of measured variables and landmarks is not effective as the possible number of these combinations is too large: Considering combinations of only two substances where each has to be checked for 25 different landmarks, $25^2 = 625$ landmark combinations are possible here. With m different measured variables, there exist $\binom{m}{2}$ possible combinations of two substances which have to be checked for all possible landmark combinations each. So, with three measurement variables at hand (e.g., biomass, one substrate, and one product), $\binom{3}{2} \cdot 25^2 = 1875$ combinations are possible, five measurement variables (e.g., biomass, three substrates, and one product) lead to 6250 combinations, and having eight substances measured (e.g., biomass,

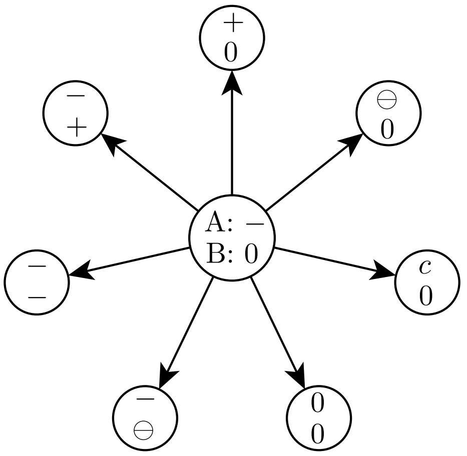


Figure 4.1: Landmark combinations that do not indicate any interaction between two substances A and B. The circle in the middle indicates the episodes before the transition: − (A) and 0 (B). The other circles indicate the episodes after the transition. In each case, an episode change can be observed in only one substance, the other one keeps its initial episode.

three substrates, three cell-intern measurements, and one product), the number of to-be-checked combinations increases to 17 500.

Instead of checking the measurements for every possible combination, the presented approach concentrates on meaningful combinations, see below. Therefore, at first, the substances are distinguished among the types substrates, products, biomass, or compartments, and only reasonable combinations thereof are considered. Then, scenarios—so-called *biological phenomena*—are developed that can actually happen to the pair of substances considered and impossible or implausible landmark combinations are excluded. If, for example, the biomass and a substrate are analyzed for their interactions, it will not make any sense to check any correlation to a landmark (0+) in the biomass measurements. It should not be possible that biomass will suddenly start growing when no biomass is present in the beginning.

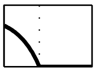
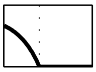
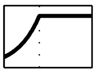
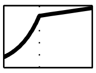
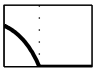
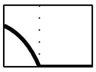
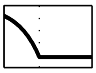
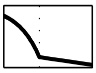
The biological phenomena describe the relation and interaction between two or more substances as a cause-and-effect chain: a change in one measured variable occurs and causes a change in another one. At first, the roles of cause and effect are assigned to a considered set of substances. Then, analyzing the substance acting as the cause, only those landmarks are considered that can actually occur and have an effect on the other substances. Finally, examining the substances acting as the effect, only those transitions are taken into account that can really happen when the aforementioned change in the cause takes place. That is, by creating a biological phenomenon and to assign the roles and landmarks correctly, biological knowledge is needed to a certain extent. For each phenomenon, rules to prove or disprove it are established, and the measurements are tested for these rules automatically. To consider a certain response time that might pass between cause and effect, the landmark combinations are analyzed within a certain time window Δt_{Phen}. As can be seen in Figure 3.7(b), the right choice for Δt_{Phen} is important for the success of the automated phenomena detection. Here, landmarks are found in all three concentrations between $t = 60\,\mathrm{h}$ and $t = 70\,\mathrm{h}$. If the value for Δt_{Phen} were too small, an automated detection would miss what a human expert would have found.

First, the complete list of the biological phenomena used is provided in Section 4.1. Then, the following sections will show how, depending to the measurement variables at hand, the biological phenomena are derived, and which rules to (dis-)prove the biological phenomena have to be examined. Last, an approach is presented that determines how strongly each phenomenon can be accepted or rejected.

4.1 The Phenomena at a Glance

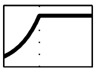
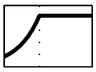
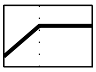
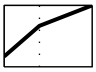
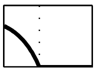
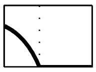
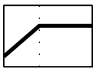
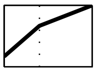
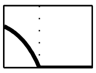
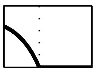
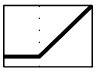
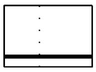
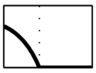
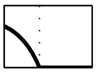
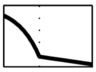
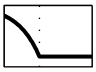
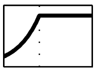
Before the phenomena are discussed in detail, an overview over the phenomena is given in Table 4.1.

Table 4.1: List of biological phenomena and the corresponding necessary landmark combinations (kl), k, $l \in \{+, -, \ominus, c, 0\}$ to (dis-)prove them based on compensated mass concentrations or cell-related concentrations

			Landmark combination	
	Biological Phenomenon	Reactants	Pro	Contra
P1	The growth is limited by a substrate.	Substrate	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Biomass	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P2	Two substrates are consumed simultaneously.	Substrate 1	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Substrate 2	$(-c)$ 	$(--)/(c-)/(0-)$ 

To be continued on next page

Table 4.1: List of biological phenomena – continued from previous page

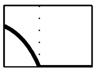
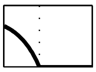
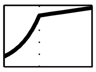
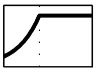
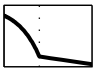
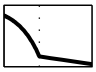
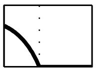
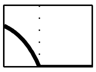
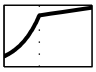
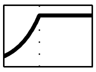
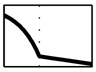
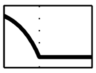
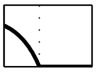
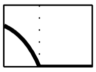
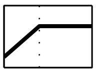
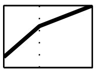
			Landmark combination	
	Biological Phenomenon	Reactants	Pro	Contra
P3	The biomass and the product grow simultaneously.	Biomass	$(+c)/(+-)$ 	$(+c)/(+-)$ 
		Product	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P4	The product formation is limited by a substrate.	Substrate	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Product	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P5	Secondary metabolism: The product formation is inhibited by a substrate.	Substrate	$(-0)/(-\ominus)/$ $(-c)$[a]$/(--)$[a] 	$(-0)/(-\ominus)/$ $(-c)$[a]$/(--)$[a] 
		Product	$(c+)/(0+)$ 	$(cc)/(00)$ 
P6	Maintenance	Substrate 1 (not C-source)	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Substrate 2 (C-source)	$(--)$ 	$(-c)/(-0)$ 
		Biomass	$(+c)$ 	

To be continued on next page

---

[a] When this landmark is analyzed, it has to be checked if the concentrations of the corresponding substances are low. For further explanation, see text.

Table 4.1: List of biological phenomena – continued from previous page

	Biological Phenomenon	Reactants	Landmark combination	
			Pro	Contra
P7	Storage A:[b] While a substrate is depleted, the cell uses a previously filled storage for this substrate to continue growing.	Substrate 1[c]	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Biomass	$(++)$ 	$(+c)$ 
		Substrate 2[c]	$(--)$ 	
		Substrate 3[c]	$(--)$ 	
P8	Storage B: While a substrate is depleted, another substrate is stored in the cell.	Substrate 1[c]	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Biomass	$(++)$ 	$(+c)$ 
		Substrate 2[c]	$(--)$ 	
		Substrate 3[c]	$(-c)$ 	
P9	The formation of a compartment is limited by a substrate.	Substrate	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Compartment	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 

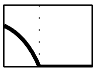
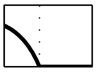
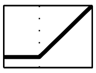
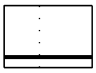
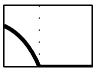
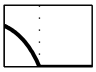
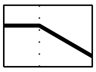
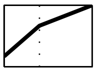
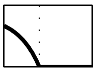
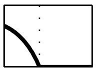
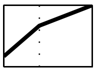
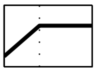
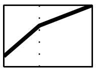
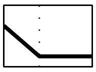
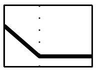
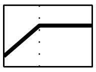
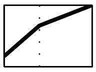
To be continued on next page

---

[b] As an alternative to this storage, it is also possible that Substrates 2 and 3 are stored according to *Storage B*.

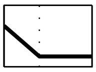
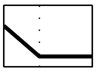
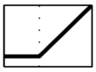
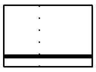
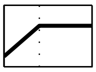
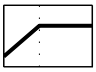
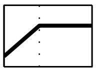
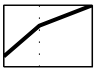
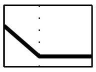
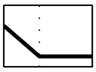
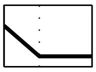
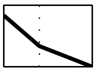
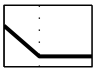
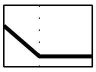
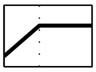
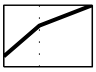
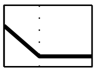
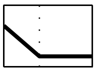
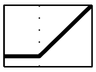
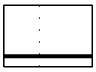
[c] The rules for this phenomenon assume the cultivation in a chemically defined medium with three essential substrates (one nitrogen, one phosphate, and one carbon source).

Table 4.1: List of biological phenomena – continued from previous page

	Biological Phenomenon	Reactants	Landmark combination	
			Pro	Contra
P10	The formation of a compartment is inhibited by a substrate.	Substrate	$(-0)/(-\ominus)/(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-0)/(-\ominus)/(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Compartment	$(c+)$ 	$(cc)/(00)$ 
P11	The degradation of a compartment is inhibited by a substrate.	Substrate	$(-0)/(-\ominus)/(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-0)/(-\ominus)/(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Compartment	$(c-)/(+-)$ 	$(++)/(-+)/(c+)/(+c)/(-c)/(cc)$ 
P12	Intermediate compartment: The depletion of one substrate has no influence on any compartment.	Substrate	$(-0)/(\ominus 0)$ 	$(-0)/(\ominus 0)$ 
		Compartment	$(++)/(-+)/(c+)$ 	$(+c)$ 
		All other Compartments	$(++)/(cc)$ 	
P13	The formation of a compartment is limited by another one.	Compartment 1	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Compartment 2	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 

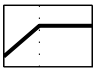
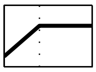
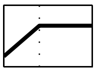
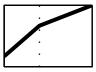
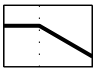
To be continued on next page

Table 4.1: List of biological phenomena – continued from previous page

			Landmark combination	
	Biological Phenomenon	Reactants	Pro	Contra
P14	The formation of a compartment is inhibited by another one.	Compartment 1	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Compartment 2	$(c+)$ 	(cc) 
P15	Two compartments grow simultaneously.	Compartment 1	$(+c)/(+-)$ 	$(+c)/(+-)$ 
		Compartment 2	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P16	Two compartments are degraded simultaneously.	Compartment 1	$(-c)$ 	$(-c)$ 
		Compartment 2	$(-c)$ 	$(--)/(c-)/(0-)$ 
P17	The product formation is limited by a compartment.	Compartment	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Product	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P18	The product formation is inhibited by a compartment.	Compartment	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 	$(-c)^{\mathrm{a}}/(--)^{\mathrm{a}}$ 
		Product	$(c+)/(0+)$ 	$(cc)/(00)$ 

To be continued on next page

Table 4.1: List of biological phenomena – continued from previous page

			Landmark combination	
	Biological Phenomenon	Reactants	Pro	Contra
P19	The product and a compartment grow simultaneously.	Compartment	$(+c)/(+-)$ 	$(+c)/(+-)$ 
		Product	$(+c)/(+-)$ 	$(++)/(-+)/(c+)$ 
P20	Degradation.	Biomass, DNA, or Product	$(c-)/(+-)$ 	

4.2 Simple Phenomena for Unstructured Models

4.2.1 Limitation

As already mentioned in Section 2.1, the growth of an organism is not unlimited and the nutrient situation, i.e., the amount of substrate(s) present is considered when modeling growth. If the growth is limited by a specific substrate, the organism can only grow as long as there is a sufficient amount of this substrate present. How can such a limiting dependency—described by phenomenon P1: *growth limited by a substrate*—be found automatically in the measurements?

Considering only one substrate S and the biomass X and assuming the simple reaction scheme

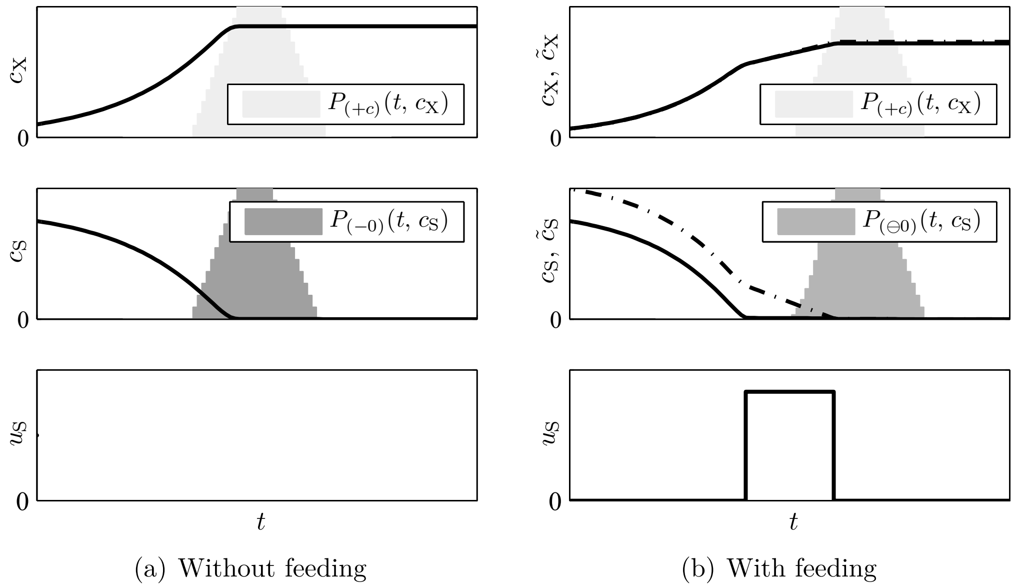
$$Y_{\mathrm{SX}}\,\mathrm{S} \xrightarrow{r_{\mathrm{X}}} \mathrm{X} \quad , \tag{4.1}$$

that leads to a set of differential equations (see Section 2.1)

$$\frac{\mathrm{d}m_{\mathrm{X}}(t)}{\mathrm{d}t} = \mu_{\mathrm{Xm}} \cdot \mu_{\mathrm{X}}(t) \cdot m_{\mathrm{X}}(t) \tag{4.2}$$

$$\frac{\mathrm{d}m_{\mathrm{S}}(t)}{\mathrm{d}t} = -Y_{\mathrm{SX}} \cdot \mu_{\mathrm{Xm}} \cdot \mu_{\mathrm{X}}(t) \cdot m_{\mathrm{X}}(t) + c_{\mathrm{S,\,in}} \cdot u_{\mathrm{S}}(t) \quad , \tag{4.3}$$

where $\mu_{\mathrm{X}}(t) = \mu_{\mathrm{X}}(c_{\mathrm{S}}(t)) = \mathrm{limit}(c_{\mathrm{S}}(t))$. The qualitative behavior of $c_{\mathrm{X}}(t)$ and $c_{\mathrm{S}}(t)$, $u_{\mathrm{S}}(t) = 0$, are shown in Figure 4.2(a). As can be seen, the biomass grows (episode +) as long as there is substrate, which is decreasing (episode −), and stops to grow (c)



(a) Without feeding

(b) With feeding

Figure 4.2: Detection of limited growth. Shown are a possible solution to Eqs. (4.2) and (4.3) as well as the probabilities of necessary landmarks to detect the phenomenon *growth limited by a substrate.* (a) No substrate is fed, the landmarks (-0) (substrate) and $(+c)$ (biomass) are correlated. (b) Substrate is fed while it has been depleted in the fermenter, $(\ominus 0)$ (substrate) and $(+c)$ (biomass) prove the phenomenon. Solid lines indicate the measured concentrations c_i, dash-dot lines characterize the compensated concentrations $\tilde{c}_i$.

when the substrate depletes (0). The landmarks (-0) in the substrate and $(+c)$ in the biomass are correlated. Hence, to prove phenomenon P1, the measurements have to be checked for this landmark combination.

If substrate is fed, another landmark combination has to be checked as well. In Figure 4.2(b), an example is shown where the substrate is fed immediately after it has been depleted in the fermenter. As can be seen, the compensated substrate concentration $\tilde{c}_S$ (dash-dot line) continues to decrease while the measured concentration c_S is zero, leading to the episode $\ominus$. The limiting effect becomes obvious after the feeding stops. Therefore, the landmark combination $(\ominus 0)$ (substrate) and $(+c)$ (biomass) represents another possibility to prove phenomenon P1.

Furthermore, considering the possibility of cell death, described by $r_{dX}(t)$ in Eq. (2.10),

$$\frac{\mathrm{d}m_X(t)}{\mathrm{d}t} = (r_X(t) - r_{dX}(t)) \cdot V(t) \quad ,$$

the biomass should be tested for the landmark $(+-)$ as well: When the substrate depletes, the growth rate $r_X(t)$ becomes zero and the influence of $r_{dX}(t)$ cannot be compensated anymore. The biomass degrades.

To disprove the phenomenon P1, the biomass should still grow when the substrate is vanished. The landmark combinations $(-0)/(\ominus 0)$ (substrate) and $(++)/(-+)/(c+)$ (biomass) will detect that no limiting dependency on the substrate exists.

Similarly, the phenomenon P4: *product formation limited by a substrate* can be tested. Therefore, the landmark combinations $(-0)/(\ominus 0)$ (substrate) and $(+c)/(+-)$ (product) will prove this phenomenon, and $(-0)/(\ominus 0)$ (substrate) and $(++)/(-+)/(c+)$ (product) will disprove it.

4.2.2 Simultaneous Consumption/Formation

Usually, more than one substrate is essential for growth, and therefore, several substrates have to be integrated into the growth model. Considering two substrates, the reaction changes to

$$Y_{\mathrm{S_1X}}\,\mathrm{S_1} + Y_{\mathrm{S_2X}}\,\mathrm{S_2} \xrightarrow{r_\mathrm{X}} \mathrm{X} \quad , \tag{4.4}$$

and the mass balances are

$$\frac{\mathrm{d}m_\mathrm{X}(t)}{\mathrm{d}t} = \mu_\mathrm{Xm} \cdot \mu_\mathrm{X}(t) \cdot m_\mathrm{X}(t) \tag{4.5}$$

$$\frac{\mathrm{d}m_\mathrm{S_1}(t)}{\mathrm{d}t} = -Y_\mathrm{S_1X} \cdot \mu_\mathrm{Xm} \cdot \mu_\mathrm{X}(t) \cdot m_\mathrm{X}(t) \tag{4.6}$$

$$\frac{\mathrm{d}m_\mathrm{S_2}(t)}{\mathrm{d}t} = -Y_\mathrm{S_2X} \cdot \mu_\mathrm{Xm} \cdot \mu_\mathrm{X}(t) \cdot m_\mathrm{X}(t) \quad . \tag{4.7}$$

The specific growth rate considers a limiting influence of both substrates, e.g., according to Eq. (2.9). Figure 4.3 shows a possible numerical solution of Eqs. (4.5)–(4.7). As can be seen, growth stops when $\mathrm{S_1}$ vanishes, thus $\mathrm{S_2}$ is not being consumed either. When the landmark combination $(-0)/(\ominus 0)$ ($\mathrm{S_1}$) and $(-c)$ ($\mathrm{S_2}$) can be found, this can be seen as a sign that both substrates are consumed simultaneously (P2), i.e., they have to appear in the same reaction as educts. However, when substrate 2 is a carbon source $\mathrm{S_{carb}}$, the simultaneous consumption cannot be proven by the proposed landmarks. Instead of keeping a constant value, the (compensated) concentration $\tilde{c}_\mathrm{S_{carb}}$ will still decrease as the cell will need it for maintenance.

Phenomenon P2 can be disproven if, after the depletion of substrate 1, substrate 2 is still decreasing. Therefore, the landmark combination $(-0)/(\ominus 0)$ (substrate 1) and $(+-)/(--)/(\ominus -)/(c-)/(0-)$ (substrate 2) have to be tested.

In addition to the simultaneous comsumption, two substances can be analyzed for a simultaneous formation as well. The phenomenon P3 analyzes, if the biomass and the product are produced simultaneously, i.e., are products of the same reaction—which is the case if the product is a result of the primary energy metabolism, see Section 2.1. Hence, when the biomass stops growing, the product has to stop as well. On the other hand, if the product continues growing after the biomass changes its behavior, it is likely that the biomass and product formations are not induced by the same factors. The landmark combinations $(+c)/(+-)$ (biomass) and $(+c)/(+-)$ (product) are used

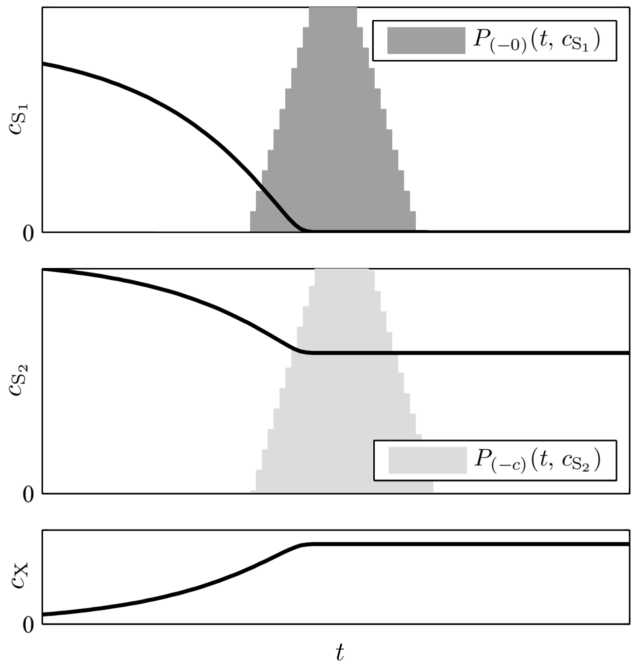


Figure 4.3: Detection of simultaneous consumption. Shown are a possible solution to Eqs. (4.5)–(4.7) as well as the probabilities of the landmarks (-0) (substrate 1) and $(-c)$ (substrate 2)—a possible landmark combination to detect the phenomenon *simultaneous consumption of two substrates.*

to prove phenomenon P3. It is disproven by the combinations $(+c)/(+-)$ (biomass) and $(++)/(-+)/(c+)$ (product).

4.2.3 Inhibition

As already discussed in Section 2.1, in addition to the growth reaction (4.1), products can be formed that are uncoupled from the basic metabolism. These secondary metabolites, e.g., antibiotics, are produced at appropriate conditions, e.g., after a depletion of one of the substrates. A rule to find this phenomenon will be derived on the basis of Eqs. (4.2) and (4.3),[1]

$$\frac{\mathrm{d}m_X(t)}{\mathrm{d}t} = \mu_{Xm} \cdot \mu_X(t) \cdot m_X(t)$$
$$\frac{\mathrm{d}m_S(t)}{\mathrm{d}t} = -Y_{SX} \cdot \mu_{Xm} \cdot \mu_X(t) \cdot m_X(t) \quad .$$

---

[1] The lack of one of the substrate will actually lead to a reorganization of the cell's metabolism which an unstructured model cannot account for. However, for convenience, the rule to detect an inhibitory effect will be derived on the basis of this unstructured model.

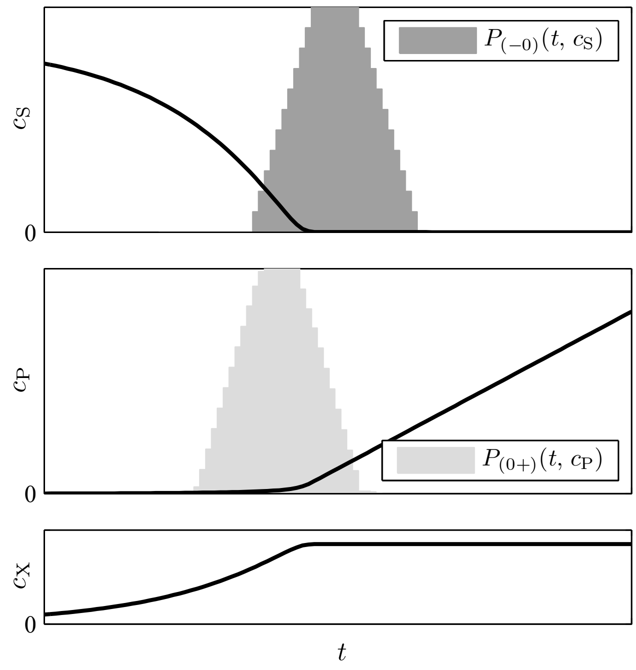


Figure 4.4: Detection of inhibited product formation. Shown are a possible solution to Eqs. (4.2), (4.3), and (4.8) as well as the probabilities of the landmarks (−0) (substrate) and (0+) (product)—a possible landmark combination to detect the phenomenon *product formation inhibited by a substrate.*

Additionally, the dynamic behavior of product P is supposed to be described by

$$\frac{\mathrm{d}m_{\mathrm{P}}(t)}{\mathrm{d}t} = \mu_{\mathrm{Pm}} \cdot \mu_{\mathrm{P}}(t) \cdot m_{\mathrm{X}}(t) \quad , \tag{4.8}$$

where $\mu_{\mathrm{P}}(t) = \mu_{\mathrm{P}}(c_{\mathrm{S}}(t)) = \mathrm{inhib}(c_{\mathrm{S}}(t))$ describes an inhibiting effect of the substrate on the product formation. The qualitative behavior shown in Figure 4.4 indicates which landmark combinations can be tested to check phenomenon P5: *product formation inhibited by a substrate*. Here, as long as there is substrate in the reactor, the product formation does not start. The landmark combination (−0) (substrate) and (0+) (product) signalizes the relationship between the two substances.

However, other landmark combinations apply as well. Looking once again at Figure 3.7, it can be seen in experiment ME2 (Figure 3.7(b)) that the product formation starts when the substrate depletes. Since, at the same time, substrate is fed, the landmark (−0) cannot be detected, the compensated substrate concentration $\tilde{c}_{\mathrm{S}}$ keeps decreasing. But the inhibiting relationship still exists. Now, it should become clear why the additional episode $\ominus$ is introduced. By only considering the compensated concentrations $\tilde{c}$, wrong conclusions could be drawn, e.g., an inhibiting dependency on a substrate could never be detected although the amount of this substrate in the reactor—which is the amount the cell is opposed to—is zero or lower than the detection limit of the measurement system that is used. Therefore, the landmark combination $(-\ominus)$ (substrate) and $(0+)$ (product) represents another possibility to check for phenomenon P5. Then, the inhibitory effect might not only be loosened when the substrate vanishes, but also be perceivable at low substrate concentrations, i.e., when the substrate is still decreasing or is at a low constant level. The landmarks $(--)$ and $(-c)$ have to be included as well, but the probabilities for these landmarks have to be adapted to the concentrations $c_{\mathrm{S}}(t)$:

$$P^*_{(-l)}(t,\, c_{\mathrm{S}}) = P_{(-l)}(t,\, c_{\mathrm{S}}) \cdot \left(1 - \frac{c_{\mathrm{S}}(t)}{c_{\mathrm{Sm}}}\right), \quad l \in \{-,\, c\} \quad , \tag{4.9}$$

where c_{Sm} is the maximum value of $c_{\mathrm{S}}(t)$. That way, transitions at low substrate concentrations are ranked higher than those at high concentrations. Furthermore, the product does not necessarily need to be zero before the transition, the landmark $(c+)$ in the product concentration applies here as well.

If the product is not synthesized once the substrate vanishes, the inhibitory effect can be disproven, i.e., the landmark combinations $(-0)/(--)/(-c)$ (substrate) and $(00)/(cc)$ have to be checked.

4.3 Storage Detection

When the biomass still grows after a substrate has vanished, a limiting dependency on this substrate has to be withdrawn (see P1/Contra in Table 4.1). The substrate would not appear in the growth reaction. However, assuming that microorganisms are cultivated in a chemically defined medium, i.e., all the chemical components are known and there are only one nitrogen, one phosphate, and one carbon source, this essential component for growth cannot be excluded. For example, in experiments run with *Paenibacillus polymyxa* in defined media, it can be observed that the amount of biomass still increases after the phosphate has disappeared (e.g., Figure 4.5). To account for this behavior, a storage will be included into the model and according to the behavior of the other substrates, two phenomena are established to check for different types of storages.

Phenomenon P7: *Storage A*, assumes that the depleted substrate has been stored in the cell. After the depletion, the cell then uses the stored substrate to continue growing. As long as the cell-intern storage is filled, the other substrates are still consumed to provide for growth. Thus, to detect this phenomenon, the landmark combination $(-0)/(\ominus 0)$ (substrate 1) and $(++)$ (biomass) is extended by the landmark $(--)$ (all other essential substrates).

Another mechanism will be analyzed by phenomenon P8: *Storage B*. Here, it is assumed that, instead of the depleted substrate (S_1), another substrate (S_2) is stored in the cell and that the increasing biomass can be explained by the increasing amount of stored substrate. Due to the lack of S_1, the biomass cannot replicate anymore

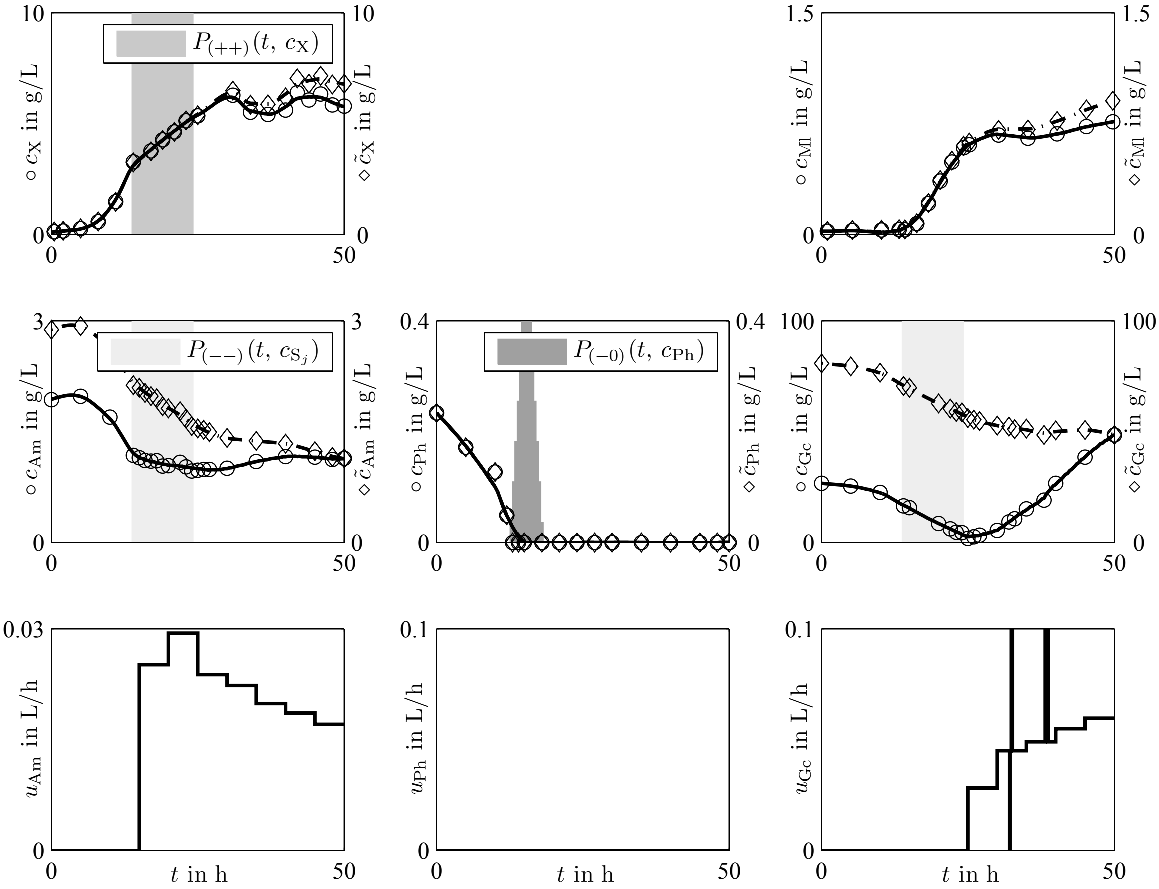


Figure 4.5: Detection of storages. Measurements from a fermentation with *P. polymyxa*. At around $t = 20\,\mathrm{h}$, phosphate depletes but the amount of biomass still increases and ammonium and glucose are still being consumed. The same information as in Figure 3.6 is shown. The landmark probabilities are displayed in the important time window only. The subscript Ml stands for macrolactin (product).

and the third substrate (S_3, assuming defined media) is not being consumed after the depletion of S_1. Consequently, the phenomenon P8 is detected by the landmark combination $(-0)/(\ominus 0)$ (substrate 1), $(++)$ (biomass), $(--)$ (substrate 2), and $(-c)$ (substrate 3).

Furthermore, the landmark combination used to find *Storage A* can be used to explain another mechanism. Instead of substrate 1 being stored according to *Storage A*, it is possible, as well, that the substrates 2 and 3 are stored according to the mechanisms of *Storage B*.

The combination $(-0)/(\ominus 0)$ (substrate) and $(+c)$ (biomass) will reject both storages since phenomenon P1 can be proven then.

4.4 Biological Phenomena for Structured Models

Once measurements of cell-intern components, i.e., compartments C_i (like DNA, RNA, or proteins) are considered to be integrated into the model, the dynamical behavior of the biomass X will not be described explicitly by an extra state variable, but, in fact, as the sum of several biotic state variables. The interactions between these compartments have to be analyzed and their relations to the substrates and products have to be found. The biological phenomena describing correlations between biomass measurements and other measurements can, in the context of structured models, only be interpreted as indirect cause-and-effect chains that will not help to find structured models automatically. Hence, phenomena have to be derived that consider the cell-intern measurements rather than the biomass. Assuming that the cell-intern measurements at hand do not add up to the biomass, virtual measurements for the residual biomass $\mathrm{Xr} = \mathrm{X} - \sum \mathrm{C}_i$ can be created and integrated into the phenomena detection process. When doing so, however, it has to be kept in mind that phenomena considering Xr will become invalid if other phenomena suggest that Xr should be further divided into several (non-measured) compartments to describe the cellular behavior better.

As can be seen in Section 4.2, any substance A that causes a phenomenon will later appear in a regulatory kinetic expression. Usually, the concentration $c_{\mathrm{A}}(t)$ *in the reactor* is used in this expression, e.g., $\mathrm{limit}(c_{\mathrm{A}}(t))$ or $\mathrm{inhib}(c_{\mathrm{A}}(t))$. When cell-intern regulations are described that consider cell-intern compartments C_i, the amount of C_i *in the cell* becomes relevant as this is the amount the environment, where the regulation takes place, is exposed to. Therefore, with respect to compartmental measurements, only landmarks derived from cell-related concentrations $g_{\mathrm{C}_i}(t)$ are used for the phenomena detection.

4.4.1 Adaptation of Already Mentioned Phenomena

Biological phenomena that have already been discussed in Section 4.2 can be adapted to cell-intern measurements C_i, e.g., DNA, RNA, and proteins, as well. Reasonable combinations between compartments on one side and substrates, products, or different compartments on the other side are regarded as new phenomena.

Limitation

The phenomenon P9: *formation of a compartment limited by a substrate* can be similarly (dis-)proven to the substrate-limited growth (P1). The landmark combinations $(-0)/(\ominus 0)$ (substrate) and $(+c)/(+-)$ (compartment) indicate that there is a limiting dependency, the combinations $(-0)/(\ominus 0)$ (substrate) and $(++)/(-+)/(c+)$ (compartment) contradict this assumption.

Furthermore, two phenomena can be established that describe the limiting influence of a compartment on the formation of a different compartment (P13) or on the product formation (P17), respectively. However, considering typical cell-intern measurements like DNA, RNA, and proteins, these compartments cannot disappear to keep the cell alive, and an alternative to the landmark (-0) in the limiting compartment has to be used. Here, the landmarks $(--)$ and $(-c)$ are considered, but the landmark probabilities are adjusted in a way that only low values for $g_{\mathrm{C}_i}(t)$ will lead to the detection of the phenomenon,

$$P^*_{(-l)}(t, g_{\mathrm{C}_i}) = P_{(-l)}(t, g_{\mathrm{C}_i}) \cdot (1 - g_{\mathrm{C}_i}(t)), \quad l \in \{-, c\} \quad . \tag{4.10}$$

Inhibition

Similar to the substrate-inhibited product formation (P5), the phenomena *formation of a compartment inhibited by a substrate* (P10), *formation of a compartment inhibited by another compartment* (P14), and *product formation inhibited by a compartment* (P18) are formulated. The landmark combinations are the same, but the landmark (-0) is omitted in the compartment measurements (see Table 4.1).

Furthermore, not only can the compartment formation be inhibited, but also the compartment degradation. For instance, for the strain *Streptomyces tendae*, it could be observed that both the RNA and protein degradation reactions were inhibited by the amount of ammonium (King, 1997), see (C.5) and (C.7). Hence, phenomenon P11: *compartment degradation inhibited by a substrate* is established. Here, the landmark combinations $(-0)/(-\ominus)/(-c)/(--)$ (substrate) and $(c-)/(+-)$ (compartment) are used to confirm this assumption. The degradation of the compartment cannot start until the amount of substrate is sufficiently small. The landmark probabilities are adapted to the substrate concentration values according to Eq. (4.9). To neglect phenomenon P11, the landmark combinations $(-0)/(-\ominus)/(-c)/(--)$ (substrate) and $(++)/(-+)/(c+)/(+c)/(-c)/(cc)$ (compartment) have to be checked. If the compartment grows or remains constant after a substrate depletion, an inhibiting influence of the substrate on the compartment degradation cannot be detected.

Simultaneous degradation/formation

The phenomenon P16: *simultaneous degradation of two compartments* is established similar to P2. Assuming that two compartments are needed to form another substance in the reaction network, the same reasoning as in Section 4.2.2 can be used. When one compartment stops degrading, the other one should stop as well. If the second compartment continues degrading, the assumed relationship is disproven. The landmark combinations to check for P16 should be the same as for P2. However, since the landmark (-0) will not be found in the measurements of typical cell-intern compartments, the landmark combination $(-c)$ (compartment 1) and $(-c)$ (compartment 2) is considered to prove this phenomenon. It is rejected by the combinations $(-c)$ (compartment 1) and $(--)/(c-)/(0-)$ (compartment 2).

Similar to P3, the phenomena P15 and P19 test if two compartments or a compartment and the product are produced simultaneously. The landmark combinations $(+c)/(+-)$ (compartment (1)) and $(+c)/(+-)$ (compartment 2 or product) are used to prove the phenomena P15 or P19, respectively. They are disproven by the combinations $(+c)/(+-)$ (compartment (1)) and $(++)/(-+)/(c+)$ (compartment 2 or product).

4.4.2 Detection of Intermediate Compartments or Precursors

Considering the phenomenon P9 in Table 4.1, a limiting dependency of a compartment on a substrate is rejected when the substrate vanishes and the compartment still grows. Now, assume that this is detected for all compartments which make up the total biomass. For instance, in cultivations with *Streptomyces* strains in defined media, it is often observed that the total amount of DNA, RNA, and proteins increases for some time when phosphate is depleted in the culture broth (King and Büdenbender, 1997). An example can be found in Figure 4.6, where measurements of an experiment with the strain *Streptomyces tendae* are shown. This is consistent with other publications which show that *Streptomyces* strains accumulate phosphate, e.g., Mundry and Kuhn (1991), Sin et al. (2008), Martín et al. (2011), Allenby et al. (2012).

As a consequence, every compartment is independent of phosphate. Using phenomenon P9 alone would discard phosphate from the model completely. However, given that the cultivation took place in a chemically defined medium, phosphate cannot be discarded completely from the model structure. Such a situation, instead, hints to an intermediate compartment or precursor which has not been considered so far as it has not been measured. For explaining growth, though, this intermediate compartment will be crucial. In fact, as already mentioned in Section 2.2, phosphate is needed to build nucleotides which, in turn, constitutes DNA and RNA. As a simple limiting dependency on phosphate is neglected in this example, an intermediate acting as the nucleotides has to be introduced.

As a result, the phenomenon P12: *Intermediate compartment* is introduced. Here, in addition to the landmark combinations $(-0)/(\ominus 0)$ (substrate) and $(++)/(-+)/(c+)$ (compartment 1), all other compartments have to be taken into account as well. Their measurements are analyzed for the landmarks $(++)/(cc)$. P12 is disproven when the landmark combination $(-0)/(\ominus 0)$ (substrate) and $(+c)$ (compartment 1) can be found, i.e., a limiting dependency on the substrate exists.

4.5 The Phenomenon Score

Since there is only a certain probability that a landmark appears, the biological phenomena, too, can only be detected with a given probability. In a window Δt_{Phen} around time t_i, every rule to prove or disprove a phenomenon is applied to the compensated measurements. For each rule, the maxima of the corresponding landmark probabilities $P_{(kl)_{\text{pro}}}(t_i,\, c_j)$, $P_{(kl)_{\text{contra}}}(t_i,\, c_j)$ in Δt_{Phen} are multiplied by each other to

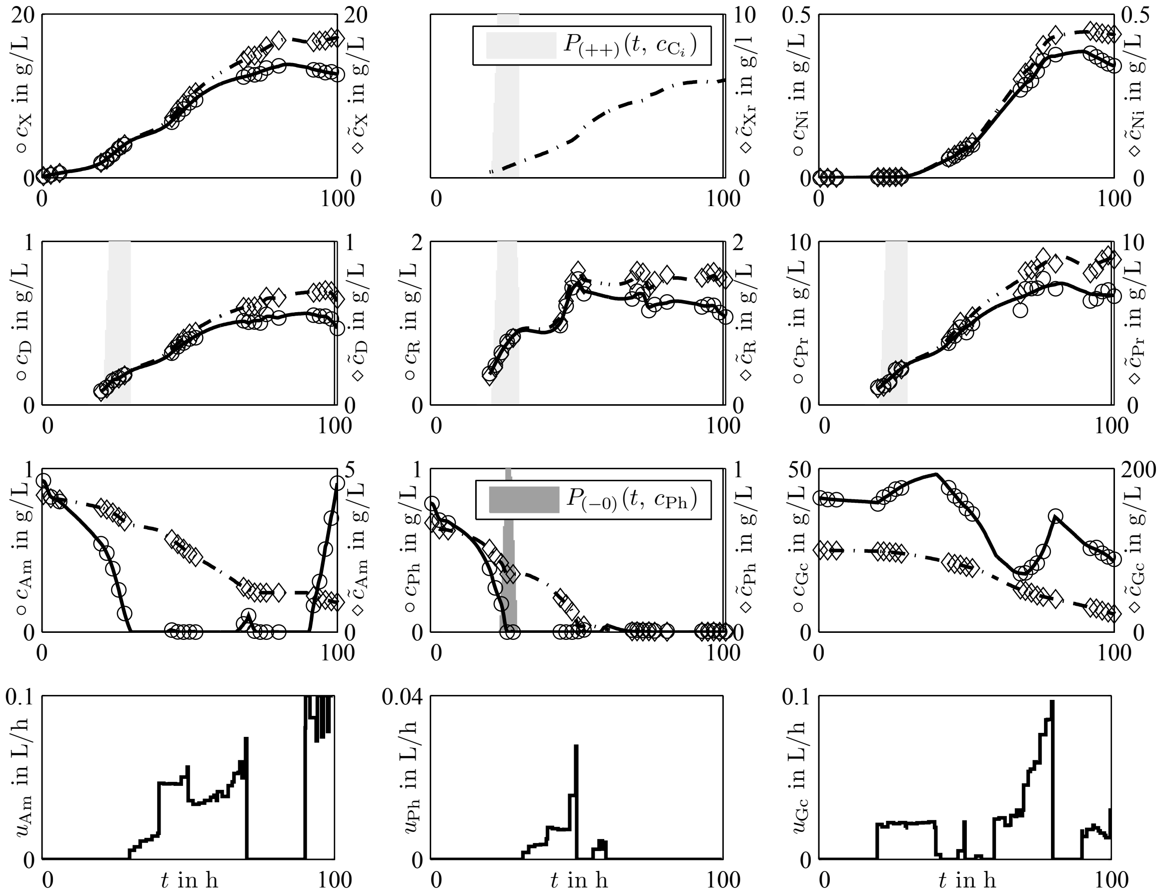


Figure 4.6: Detection of precursors. Measurements from a fermentation with *S. tendae*. At around $t = 20\,\mathrm{h}$ phosphate depletes but the cell-intern measurements still grow. The same information as in Figure 3.6 is shown. The landmark probabilities are displayed in the important time window only. The subscript Ni stands for nikkomycin (product). For the 'remaining' biomass Xr only the compensated curve is shown, as this represents a calculated quantity and not a measurement.

get the probabilities

$$P_{\mathrm{pro}}(t_i) = \prod_j \max_{t\in[t_-,\,t_+]} P_{(kl)_{\mathrm{pro}}}(t,\, c_j) \tag{4.11a}$$

or

$$P_{\mathrm{contra}}(t_i) = \prod_j \max_{t\in[t_-,\,t_+]} P_{(kl)_{\mathrm{contra}}}(t,\, c_j) \tag{4.11b}$$

for and against this rule at time t_i, with $t_- = t_i - \Delta t_{\mathrm{Phen}}/2$ and $t_+ = t_i + \Delta t_{\mathrm{Phen}}/2$. This procedure leads to probability information of the rules which (dis-)prove the phenomena. However, phenomena can only be considered if the resulting episodes in the associated landmark combinations are supported by at least two measurements for each substance. For all microorganisms considered in this work, $\Delta t_{\text{Phen}} = 4\,\text{h}$ is chosen. This value can be adjusted by the user.

In order to gain detailed insight into the underlying reaction network, several experiments running under different conditions, e.g., feeding strategies, should be considered. Thus, it is possible that specific requirements to check a phenomenon are not fulfilled in every experiment and consequently not every phenomenon can be checked in every experiment. For example, to check if the growth of the biomass is limited by a substrate, the substrate has to vanish according to the established rule. This does not occur in every experiment. For this reason, a way has to be found for how to handle the results of each rule in each experiment which, if existing at all, could even be contradictory. The following solution is applied. Each time a phenomenon can be checked, probabilities are calculated that this biological phenomenon can be accepted or has to be rejected (P_{pro} and P_{contra}, respectively, see (4.11a) and (4.11b)). Then, a score Sc is determined that indicates how strongly a phenomenon can be accepted or rejected,

$$\text{Sc} = \frac{\sum P_{\text{pro}} - \sum P_{\text{contra}}}{n_{\text{Phen}} + 1} \quad , \tag{4.12}$$

with n_{Phen} indicating how often the phenomenon can be analyzed, i.e., proven or disproven. This number is increased by one to assign a higher importance to phenomena that can be checked more often. For example, if a phenomenon is detected only once with $P_{\text{pro}} = 1$ and $P_{\text{contra}} = 0$, then $\text{Sc} = 1/2$. If it is detected ten times, the score will be 10/11. In general, it is $-1 < \text{Sc} < 1$. The closer Sc is to 1 or -1, the more reliable it is that the phenomenon is proven or disproven, respectively.

Chapter 5

Uncertainty Analysis of the Detection of Biological Phenomena

The score Sc introduced in Section 4.5 represents a quantitative measure of the reliability of every detected phenomenon. A high absolute score means that the phenomenon's impact on the model to be developed should be considered with a higher degree of confidence than those of phenomena with a lower absolute score. Since the presented approach is measurement-driven, the score Sc depends on the measurement situation, i.e., in which cultivation phase the measurements are taken, how many are used for the detection, how large the measurement noise is, and how many experiments are considered. Thus, the question arises regarding how sensitive the identification of Sc is against these influences.

5.1 Bootstrap Approach

To answer this question, the so-called bootstrap method, as introduced by Efron (1979), is proposed, similar to the determination of parameter uncertainties in parameter identification (e.g., Joshi et al., 2006). Here, the bootstrap method is adapted to the proposed algorithm to detect biological phenomena.

To perform the analysis, the experiments used to determine the scores of the different phenomena have to repeated B times, where B is a sufficiently large number. Due to the measurement noise, this leads to a set of slightly different experimental data $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_B$. These experimental data are then used to calculate corresponding phenomena scores $\underline{\mathrm{Sc}}_1, \underline{\mathrm{Sc}}_2, \ldots, \underline{\mathrm{Sc}}_B$. Afterwards, the statistical properties of the resulting distribution of the set of scores can be determined.

However, the set of scores need to be checked for outliers that could have a strong impact on the statistics. Therefore, an approach described by Montgomery et al. (2001) is used to identify outliers: the quantiles $\mathrm{Sc}_{\mathrm{P}i}^{(0.25)}$, $\mathrm{Sc}_{\mathrm{P}i}^{(0.5)}$, and $\mathrm{Sc}_{\mathrm{P}i}^{(0.75)}$ for each phenomenon Pi are determined that divide the sorted set of scores $\{\mathrm{Sc}_{\mathrm{P}i}^{j}\}_{j=1}^{B}$ into four equal parts, i.e., 25 % of the data can be found between $\mathrm{Sc}_{\mathrm{P}i}^{(0.25)}$ and $\mathrm{Sc}_{\mathrm{P}i}^{(0.5)}$ and another 25 % of the data between $\mathrm{Sc}_{\mathrm{P}i}^{(0.5)}$ and $\mathrm{Sc}_{\mathrm{P}i}^{(0.75)}$. Outliers are then defined as values that satisfy either

$$\mathrm{Sc}_{\mathrm{P}i}^{j} < \mathrm{Sc}_{\mathrm{P}i}^{(0.25)} - 1.5\left(\mathrm{Sc}_{\mathrm{P}i}^{(0.75)} - \mathrm{Sc}_{\mathrm{P}i}^{(0.25)}\right) \tag{5.1a}$$

or

$$\mathrm{Sc}_{\mathrm{P}i}^{j} > \mathrm{Sc}_{\mathrm{P}i}^{(0.75)} + 1.5\left(\mathrm{Sc}_{\mathrm{P}i}^{(0.75)} - \mathrm{Sc}_{\mathrm{P}i}^{(0.25)}\right) \quad . \tag{5.1b}$$

These values will not be considered for the calculation of the statistical properties.

The confidence interval of the (unknown) distribution of the set of scores $\{\mathrm{Sc}_{\mathrm{P}i}^{j}\}_{j=1}^{B}$ can then be determined by

$$\left[\mathrm{Sc}_{\mathrm{P}i}^{\mathrm{low}},\, \mathrm{Sc}_{\mathrm{P}i}^{\mathrm{up}}\right] = \left[\mathrm{Sc}_{\mathrm{P}i}^{(\alpha/2)},\, \mathrm{Sc}_{\mathrm{P}i}^{(1-\alpha/2)}\right] \quad , \tag{5.2}$$

where $(1-\alpha)\cdot 100\,\%$ of the data are found between the quantiles $\mathrm{Sc}_{\mathrm{P}i}^{(\alpha/2)}$ and $\mathrm{Sc}_{\mathrm{P}i}^{(1-\alpha/2)}$. The length $L_{\mathrm{P}i}$ of this confidence interval is defined by

$$L_{\mathrm{P}i} = \mathrm{Sc}_{\mathrm{P}i}^{\mathrm{up}} - \mathrm{Sc}_{\mathrm{P}i}^{\mathrm{low}} \quad . \tag{5.3}$$

As it is too expensive, time-consuming and even impossible to repeat the experiments sufficiently often, the proposed bootstrap method cannot be applied to experimental data. However, models describing growth and product formation of microorganisms can be used to test the method.

5.2 Case Studies

In this work, two models are used to analyze the uncertainty of the phenomena detected for unstructured models: the motivating example (Chapter 1) and a more complex, yet simple unstructured model (Appendix A). Furthermore, to analyze a more complex measurement situation, i.e., cell-intern measurements, a structured model is used.

At first, the phenomena detection approach is applied to the complete trend of simulated measurements, assuming that continuous measurements are possible (case A). Then, in-silico measurements are generated by only considering the time instants where measurements in the experiments are taken. The results of different smoothing and interpolation methods (see Chapter 3) are presented and for each measured variable, an adequate method is individually selected. The phenomena are detected with these data (case B). This gives an indication of the influence of measurement times on the phenomena detection. Afterwards, the measurements are assumed to be noisy: to each measurement $y_i(t_j)$, some normally distributed noise $\epsilon_i(t_j) \sim \mathcal{N}(0,\, \sigma_i^2(t_j))$ is added, the standard deviations $\sigma_i(t_j)$ will be shown below. In this manner, $B = 1000$ different sets of experimental data $\mathbf{Y}_i$ are generated. Here, the smoothing and interpolation methods chosen for case B are applied to the different data sets. The bootstrap method described in Section 5.1 is then applied. To calculate the confidence interval, $\alpha = 0.05$ is chosen, i.e., the confidence interval comprises 95 % of all values.

5.2.1 Motivating Example

At first, the phenomena detected with the continuous measurements (A) are compared to the expectations based on the model structure. Here, growth is limited by the substrate, and product formation is inhibited by the substrate. These two phenomena are expected to be proven. Likewise, the phenomenon *product formation is limited by the substrate* should be rejected since no such relationship is formulated by the model. The scores for these phenomena can be found in Table 5.1. The phenomena are (dis-) proven as expected. Additionally, with $|\mathrm{Sc}_{\mathrm{P}i}| > 0.6$, the considered phenomena can be accepted or rejected with reasonable certainty. As phenomena are detected or rejected with almost the same score in case B, a negative effect of sampling time of the measurements cannot be found here.

A possible impact of the measurement noise on the phenomena detection can now be found by analyzing the results of the bootstrap method. The noisy data are generated with the standard deviations specified in Table 5.2. In this case study, almost every set of randomly generated data sets is considered for the calculation of the statistic properties, i.e., only few samples lead to outliers or in only few samples, the phenom-

Table 5.1: Scores of the phenomena using simulations of the motivating example. Comparison between expected values and obtained values by (assumed) continuous and discrete measurements. The results of a bootstrap analysis are based on 1000 samples. The confidence interval is based on $\alpha = 0.05$.

	Exp.	Cont. meas.	Discrete measurements			
				Bootstrap		
Phenomenon Pi		Sc (A)	Sc (B)	Considered in %	$\mu_{\mathrm{P}i}$	$L_{\mathrm{P}i}$
The growth is limited by the substrate. (Figure 5.1(a))	+	0.75	0.75	98.4	0.38	0.88
The product formation is limited by the substrate. (Figure 5.1(b))	−	−0.67	−0.67	96.6	−0.52	0.42
The product formation is inhibited by the substrate. (Figure 5.1(c))	+	0.77	0.79	99.9	0.50	0.52

Table 5.2: Standard deviation of the measurement noise of the motivating example. For each measurement variable, the standard deviation is linearly approximated.

c_i	σ_i
c_X	$\sigma_\mathrm{X} = 0.25/12 \cdot c_\mathrm{X} + 0.05\,\mathrm{g/L}$
c_S	$\sigma_\mathrm{S} = 0.5/40 \cdot c_\mathrm{S} + 0.25\,\mathrm{g/L}$
c_P	$\sigma_\mathrm{P} = 0.05 \cdot c_\mathrm{P} + 0.01\,\mathrm{g/L}$

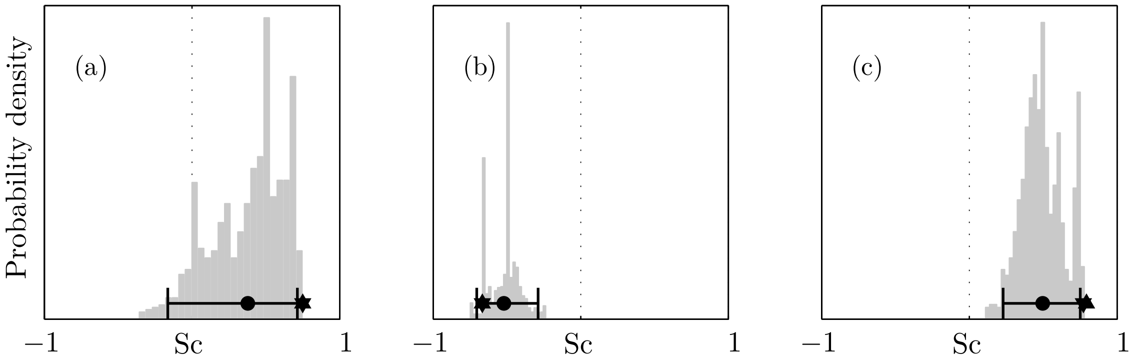


Figure 5.1: Distribution of the score Sc after a bootstrap analysis applied on the motivating example: (a) *The growth is limited by the substrate.* (b) *The product formation is limited by the substrate.* (c) *The product formation is inhibited by the substrate.* Additional symbols: mean $\mu_{\mathrm{P}i}$ of the bootstrap analysis (●), scores of the continuous simulation (A) (▼) and of the in-silico measurements (B) (▲), confidence interval (black line).

enon could not be analyzed. Comparing the mean values $\mu_{\mathrm{P}i}$ of the bootstrap method to the score values obtained in cases A and B, it can be seen that the phenomena detected with noisy data are less reliable than those detected with noiseless data. The mean values $\mu_{\mathrm{P}i}$ are closer to zero than Sc (A) and Sc (B). Here, the phenomenon *growth limited by the substrate* seems to be the most uncertain phenomenon since its confidence interval is twice the size of the other ones. This is confirmed in Figure 5.1 where more reliable phenomena show more narrow distributions ((b) and (c)) than the more uncertain one. Here, some data samples even reject that the growth is limited by the substrate. Besides other explanations, the measurement noise in combination with the chosen sampling time can lead to a wrong dynamic behavior compared to the noiseless simulation. An example can be found in Figure 5.2. Looking at the biomass measurements, the interpolation of noisy measurements (dash-dot line) shows an increasing behavior between 80 h and 90 h whereas the noiseless measurements are constant. Thus, the phenomenon detected by noiseless data is rejected in this case.

5.2.2 Unstructured Model with three Substrates

In this example with the experiments shown in Appendix A.2, it should be detected that the growth is limited by all of the substrates, that the product formation is limited by glucose and inhibited by phosphate, and that the substrates are consumed simultaneously. However, a simultaneous growth of biomass and product is not formulated by the model. As can be seen in Table 5.3, most expectations are satisfied well in case A, i.e., $|\mathrm{Sc}_{\mathrm{P}i}| \geq 0.5$ in most cases. However, the phenomena *growth limited by glucose* and *product formation limited by glucose* are only found with $|\mathrm{Sc}_{\mathrm{P}i}| < 0.2$, which cannot be regarded as highly certain. Here, with the experiments used to an-

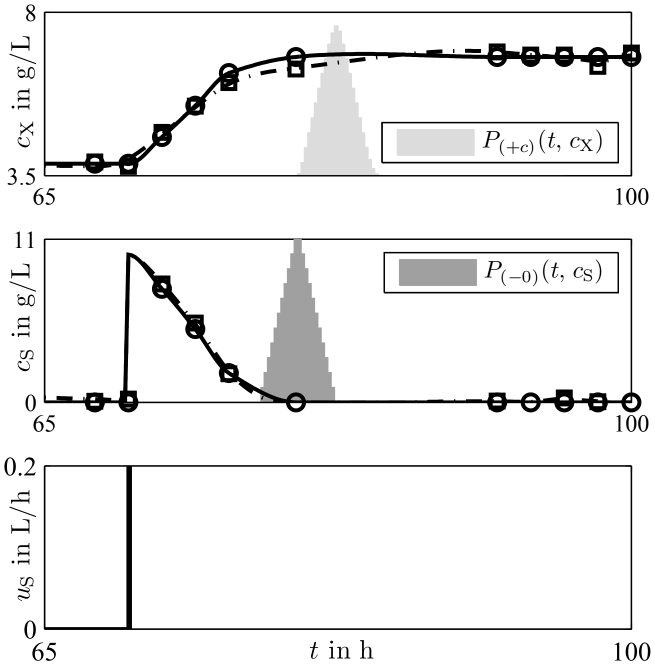


Figure 5.2: Differences in the phenomena detection between noisy and noiseless data. Considered phenomenon: *The growth is limited by the substrate.* Shown are the noiseless measurements (○) and their interpolations (solid line), as well as the noisy data (□) and their interpolations (dash-dot line). The shown landmark probabilities (different shades of gray) are calculated based on the noiseless measurements.

alyze the phenomena, the events to (dis-)prove these phenomena are rare and only a weak correlation can be detected.

In contrast to the motivating example, differences in the phenomena detection between the continuous (A) and the discrete measurement situation (B) can be found here. As is shown in Table 5.3, four phenomena cannot be evaluated, one is even falsely rejected. That means that the chosen sampling time can have an effect on the phenomena detection, i.e., there is a discrepancy between the occurrence of a phenomenon in an experiment and the chosen measuring time. A phenomenon that is detected in case A might not be considered in case B because the two required measuring times are missing. For example, the phenomenon *growth limited by phosphate* is detected twice in case A whereas it is not detected at all in case B. Another reason to explain the differences are interpolation errors, i.e., the differences between the simulation and the interpolation. Landmarks that occur in case A can be neglected by the interpolation in case B. An example is depicted in Figure 5.3. The phenomenon *product formation inhibited by phosphate* can be detected by the continuous measurements whereas it can only be rejected by the interpolation of the discrete measurements.

The lower percentages of considered bootstrap samples—generated with the standard deviations given in Table A.2—can thus be explained by the chosen sampling as well.

Table 5.3: Scores of selected phenomena using simulations of a more complex unstructured model. For more details see Table 5.1.

	Exp.	Cont. meas.	Discrete measurements			
				Bootstrap		
Phenomenon Pi		Sc (A)	Sc (B)	Considered in %	$\mu_{\mathrm{P}i}$	$L_{\mathrm{P}i}$
The growth is limited by ammonium.	+	0.67	0.32	80.8	0.39	0.59
The growth is limited by phosphate.	+	0.5	—	41.4	0.24	0.88
The growth is limited by glucose.	+	0.13	—	19.3	0.12	0.41
The product formation is limited by glucose. (Figure 5.4(a))	+	0.18	—	29.7	0.13	0.58
Ammonium and phosphate are consumed simultaneously. (Figure 5.4(b))	+	0.60	—	55.0	0.05	0.96
The product formation is inhibited by phosphate. (Figure 5.4(c))	+	0.54	−0.15	93.6	0.24	0.50
The biomass and the product grow simultaneously.	−	−0.32	−0.58	99.7	−0.41	0.75

The other results of the bootstrap analysis applied to this case study are consistent with those of the motivating example: in most cases, the mean values $\mu_{\mathrm{P}i}$ are closer to zero than Sc (A) or Sc (B). The sizes of the confidence intervals differ here, as well, leading to narrow or wide distributions (Figure 5.4). Unfortunately, a pattern to determine which phenomenon can be determined with less confidence cannot be discovered.

5.2.3 Structured Model for *Streptomyces tendae*

To test the uncertainty of phenomena detected for structured models, a model describing growth and production of an antibiotic by *S. tendae*, manually set up in King (1997) and shown in Appendix C.1, using the simulated experiments STdef1–STdef6, is exploited. The simulation data are shown in Appendix C.2. The results for some phenomena, which illustrate a cross-selection of all detected ones, are shown in Table 5.4. The bootstrap samples are taken with the standard deviations shown in Table C.2.

The most important phenomena that should be detected or rejected are:

- All measured cell-intern components (DNA, RNA, proteins) are not directly limited by any substrate.

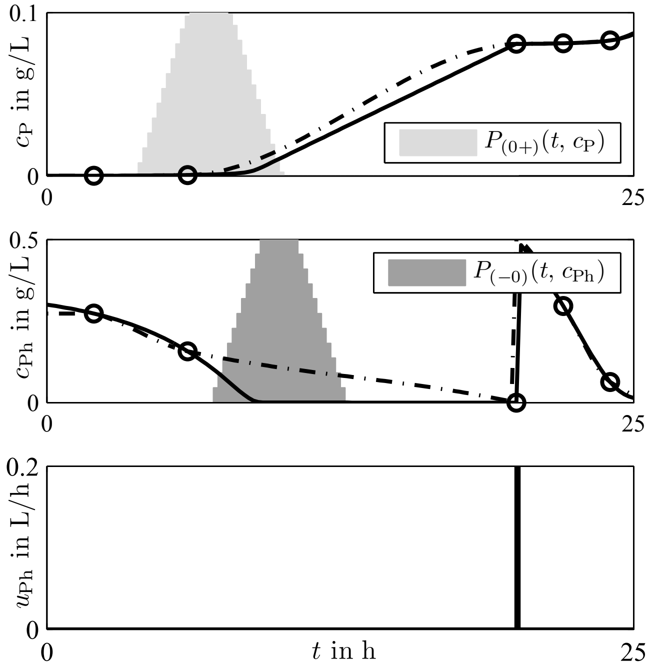


Figure 5.3: Differences in the phenomena detection between continuous and discrete measurements. Considered phenomenon: *The product formation is inhibited by phosphate.* Shown are the continuous simulation (solid line), the in-silico measurements (○), and their corresponding interpolations (dash-dot line). The shown landmark probabilities (different shades of gray) are calculated based on the continuous simulation.

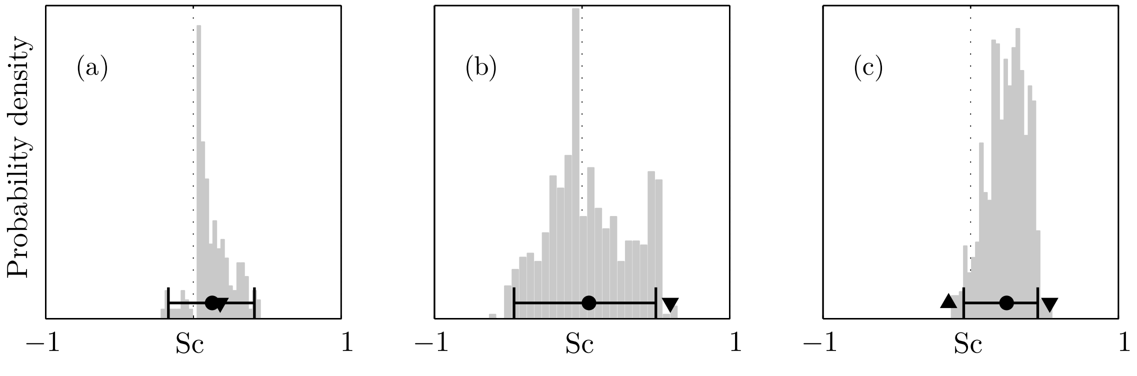


Figure 5.4: Distribution of the score Sc after a bootstrap analysis applied on a more complex unstructured model: (a) The product formation is limited by glucose. (b) Ammonium and phosphate are consumed simultaneously. (c) The product formation is inhibited by phosphate. For more details, see Figure 5.1.

- There are no direct limiting or inhibiting dependencies between the measured compartments.
- A direct limiting or inhibiting influence of the substrates or the measured compartments on the product nikkomycin (Nm) is not described by the model.
- The RNA and protein degradation rates are both inhibited by ammonium.
- There are two intermediate compartments that serve as precursors for DNA, RNA (nucleotides), and the proteins (amino acids).

In case A, only few phenomena are not (dis-)proven as expected (see Table 5.4). In fact, a direct inhibiting influence of the phosphate on the product formation can be found whereas the model does not formulate such a direct relationship. Instead, the nikkomycin building-up reaction is inhibited by amino acids (Aa) and nucleotides (Nu),

$$r_{\mathrm{Nm}}(t) = \left(\mu_{\mathrm{Nm1m}} \cdot \frac{K_{\mathrm{NmAa}}}{g_{\mathrm{Aa}}(t) + K_{\mathrm{NmAa}}} + \mu_{\mathrm{Nm2m}} \cdot \frac{K_{\mathrm{NmNu}}}{g_{\mathrm{Nu}}(t) + K_{\mathrm{NmNu}}}\right) \cdot g_{\mathrm{D}}(t) \quad .$$

However, the building-up reaction of the (unmeasured) nucleotides is, in turn, both limited and inhibited by phosphate,

$$r_{\mathrm{Nu}}(t) = \left(\mu_{\mathrm{Nu1m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Nu1}}} + \mu_{\mathrm{Nu2m}} \cdot \frac{K_{\mathrm{Nu2}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Nu2}}}\right) \cdot \frac{g_{\mathrm{Aa}}(t)}{g_{\mathrm{Aa}}(t) + K_{\mathrm{NuAa}}} \cdot g_{\mathrm{Pr}}(t) \quad .$$

This means that, although a direct relationship between phosphate and the product is not described by the model, an indirect relationship exists and is discovered by the phenomena detection.

Likewise, the ammonium-inhibited degradation reactions of RNA and the proteins are even rejected although such a relationship is clearly described by the model. Here, it seems possible that the experiments chosen for the phenomena detection do not stimulate the process in a way that these two phenomena can be detected properly. An example that illustrates this assumption is given in Figure 5.5. Here, the phenomenon *protein degradation inhibited by ammonium* is analyzed. As can be seen, the degradation reaction $r_{\mathrm{dPr}}(t)$ takes place after the depletion of ammonium and is thus inhibited by ammonium. However, the net growth rate is still positive because the protein synthesis rate $r_{\mathrm{Pr}}(t)$ is faster than the degradation. Therefore, the phenomenon cannot be proven here. Additionally, there is only one cultivation (STdef5) where for a very short period $c_{\mathrm{Pr}}(t)$ decreases.

Differences between the cases A and B can be detected here as well. Some phenomena detected in A are not considered in B because of the two missing, but required measurements. For instance, taking the phenomenon *product formation limited by ammonium* which is disproved three times in A but cannot be detected at all in B. Interpolation errors can cause different probabilities with which phenomena are detected. In both cases A and B, the phenomenon *product formation inhibited by phosphate* is proven five times and disproved four times. However, in B the found probabilities to

Table 5.4: Scores of selected phenomena using simulations of the dynamic model for *S. tendae* by King (1997). For more information, see Table 5.1.

	Exp.	Cont. meas.	Discrete measurements			
				Bootstrap		
Phenomenon Pi		Sc (A)	Sc (B)	Considered in %	$\mu_{\mathrm{P}i}$	$L_{\mathrm{P}i}$
The product formation is limited by ammonium.	–	−0.75	—	—	—	—
The product formation is limited by phosphate.	–	−0.85	−0.85	100.0	−0.02	1.18
The product formation is inhibited by ammonium. (Figure 5.6(a))	–	−0.03	0.09	97.8	0.17	0.33
The product formation is inhibited by phosphate.	–	0.23	0.09	98.4	0.39	0.52
The DNA formation is limited by phosphate.	–	−0.66	−0.63	100.0	−0.21	0.86
The RNA formation is limited by phosphate. (Figure 5.6(b))	–	−0.61	−0.63	99.0	−0.02	1.04
The protein formation is limited by phosphate.	–	−0.85	−0.84	98.2	−0.68	0.41
The DNA formation is inhibited by ammonium.	–	−0.22	−0.05	94.0	−0.15	0.38
The DNA formation is inhibited by phosphate.	–	−0.13	−0.34	93.0	−0.17	0.16
The RNA degradation is inhibited by ammonium. (Figure 5.6(c))	+	−0.06	−0.20	99.8	−0.05	0.58
The protein degradation is inhibited by ammonium.	+	−0.51	−0.55	99.4	−0.49	0.23
DNA formation despite lack of phosphate.	+	0.61	0.63	99.4	0.02	1.01
RNA formation despite lack of phosphate. (Figure 5.6(d))	+	0.66	0.25	99.0	0.33	1.01
Protein formation despite lack of phosphate.	+	0.83	0.86	91.6	0.60	0.46
The RNA formation is limited by DNA. (Figure 5.6(e))	–	—	—	49.8	−0.16	0.69
The protein formation is limited by DNA.	–	—	—	82.8	−0.32	0.61
The DNA formation is limited by RNA.	–	0.00	—	65.0	0.04	1.01
The protein formation is limited by RNA.	–	−0.45	—	94.2	−0.43	0.68
The product formation is limited by RNA. (Figure 5.6(f))	–	−0.05	—	56.8	−0.12	0.25

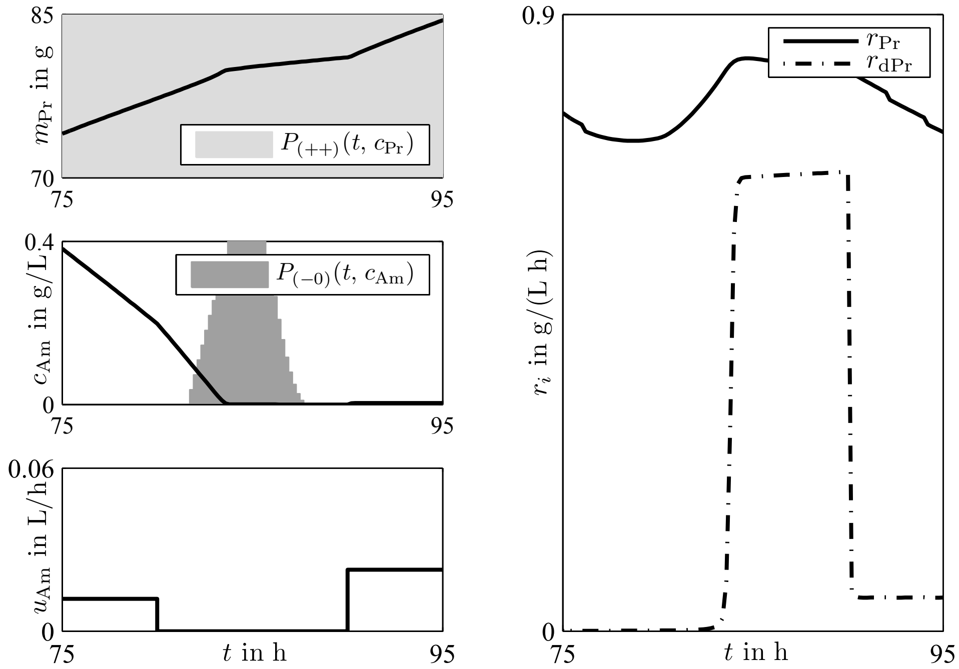


Figure 5.5: False rejection of a phenomenon. Considered phenomenon: *The protein degradation is inhibited by ammonium.* Shown are the continuous simulation and important landmark probabilities (different shades of gray) in an extract of the experiment STdef1. In the right column, the protein synthesis rate $r_{\mathrm{Pr}}(t)$ and the protein degradation rate $r_{\mathrm{dPr}}(t)$ are shown. Although the degradation is inhibited by ammonium, this effect is disproven by the simulation as the net growth is positive.

accept this phenomenon are lower than in A, leading to a lower score value. Moreover, interpolation errors might lead to detections and rejections of phenomena in case B different from those in A. For example, *RNA formation despite lack of phosphate* is proven five times and disproved once in A. However, in case B it is proven four times and disproved three times. In most cases, different scores in A and B cannot be explained by one of the aforementioned possible causes. Instead, a combination of all of them has to be taken into account.

As already discussed above, the impact of the measurement noise on the score is hard to predict. The distribution can be narrow or wide (see Figure 5.6(a)/(c)/(f) or (b)/(d)/(e), respectively, or compare the lengths $L_{\mathrm{P}i}$ of the confidence intervals). Due to the measurement noise and interpolation, errors occur which cause a shift between Sc (B) and $\mu_{\mathrm{P}i}$. Furthermore, phenomena are detected with noisy data that are neither discovered in case A nor in case B, e.g., *RNA formation limited by DNA*.

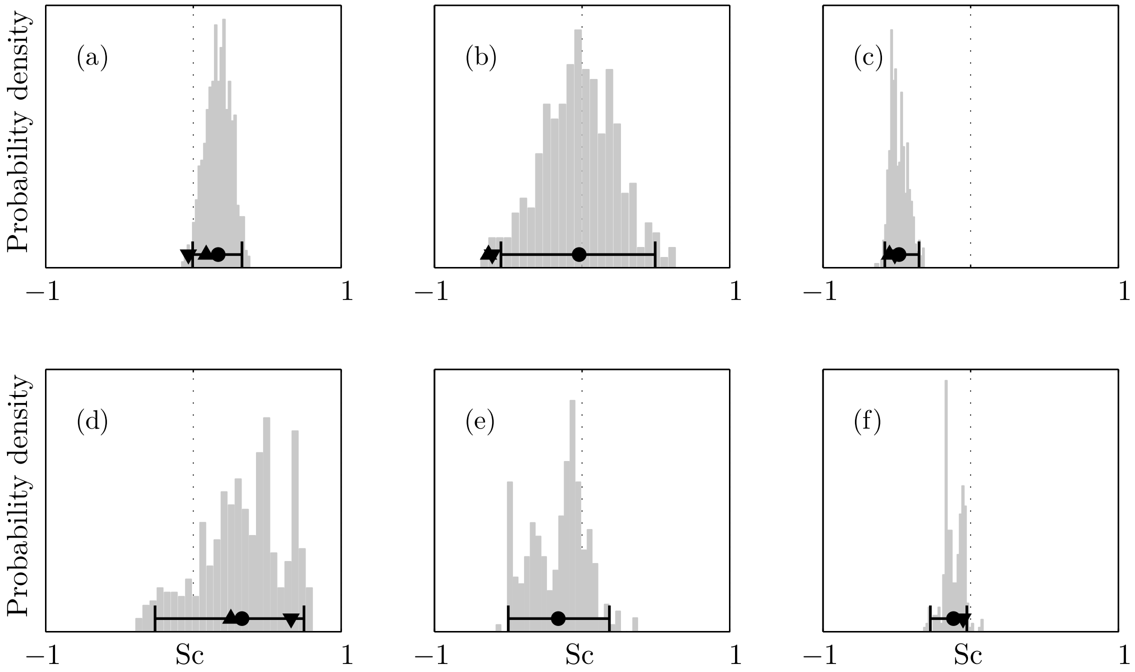


Figure 5.6: Distribution of the score after a bootstrap analysis applied on a model for *S. tendae*: (a) *The product formation is inhibited by ammonium.* (b) *The RNA formation is limited by phosphate.* (c) *The RNA degradation is inhibited by ammonium.* (d) *RNA formation despite lack of phosphate.* (e) *The RNA formation is limited by DNA.* (f) *The product formation is limited by RNA.* For more details, see Figure 5.1.

5.3 Summary

In a nutshell, a general statement about the uncertainty of the identified score cannot be made. Measuring time, measurement noise, interpolation, and the combination of all of them have an impact on $\mathrm{Sc}_{\mathrm{P}i}$ that is too complex to predict. Hence, when a detected phenomenon affects the model structure, the phenomenon should be thoroughly checked before a corresponding model component is added.

Despite these uncertainties, which are mainly caused by a difficult measuring situation, the value of automatic phenomena detection becomes evident when it is compared to what a human expert would detect. Both the expert and the computer face this measuring situation. However, in the case of many experiments a human modeler always runs the risk of overlooking some phenomena are overlooked or of ignoring some disapproving results when in other experiments seemingly obvious phenomena are found. This problem does not occur with automated phenomena detection since this rule-based approach finds and considers every possible phenomenon, independent of previously found results. Therefore, phenomena detection is an invaluable tool for modeling fed-batch cultivations and for speeding up the process of modeling.

Chapter 6

Proposing Model Structures

The detection of biological phenomena is supposed to facilitate and speed up finding a model describing a biological process. Based on the detected phenomena, a model structure has to be set up that can explain these phenomena. Here, a basic model structure is proposed initially that can be changed according to the phenomena. If assumptions inherent to this model are falsified by the detected phenomena, the model is changed. Likewise, the inclusion of additional components is guided by the attempt to model phenomena found.

6.1 Basic Model Structures

The measurement situation determines which kind of model can reasonably be built. With only biomass measurements at hand to describe growth, there is no virtue in starting with a structured model containing cell compartments and describing their interactions which can never be validated by the measurements. On the other hand, if more biotic measurement variables—apart from biomass—are available, implementing this additional information into the model can be beneficial. Therefore, depending on the measurement situation, different basic model structures are proposed initially, which is shown in the next paragraphs.

In all cases applied to the presented algorithm in this work, the measured substrates are ammonium, phosphate, and glucose for cultivations run in chemically defined media with the aforementioned nitrogen, phosphorus, and carbon sources. If cell-intern components are measured, these compartments comprise DNA, RNA, and proteins.

6.1.1 Describing Growth with Unstructured Models

If only the biomass X, the substrates S_j, and, if existing, a product P are measured, the initial model will be unstructured (see Section 2.1). The model (Eq. (2.25)),

$$\frac{\mathrm{d}\underline{m}(t)}{\mathrm{d}t} = \mathbf{K} \cdot \underline{r} \cdot V(t) + \mathbf{C}_{\mathrm{in}} \cdot \underline{u}^T - \underline{\nu} \quad .$$

comprises mass balances for each measured variable, one growth reaction r_{X}, and one product formation reaction r_{P}.

According to Eq. (2.11), it is assumed that each substrate is needed for growth,

$$\sum_j Y_{\mathrm{S}_j\mathrm{X}}\,\mathrm{S}_j \xrightarrow{r_\mathrm{X}} \mathrm{X} \quad ,$$

and has a limiting influence on the growth rate $r_\mathrm{X}(t)$, which is defined by

$$r_\mathrm{X}(t) = \mu_\mathrm{Xm} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot c_\mathrm{X}(t) \quad . \tag{6.1}$$

The maintenance rate $r_\mathrm{M}(t)$, included in $\underline{\nu}(t)$, is preselected according to Eq. (2.14),

$$r_\mathrm{M}(t) = \mu_\mathrm{Mm} \cdot \frac{c_{\mathrm{S_{carb}}}(t)}{c_{\mathrm{S_{carb}}}(t) + K_\mathrm{M}} \cdot c_\mathrm{X}(t) \quad .$$

6.1.2 Describing Growth with Structured Models

If, aside from the biomass, other cell compartments are also measured, an alternative structure will be proposed initially. This structure is influenced by King (1997), King and Büdenbender (1997), Kammerer and Gilles (2000), where starting from such a common basis, different strains could be described by a model. It is a simple structured model, compare with Eq. (2.47),

$$\frac{\mathrm{d}\underline{m}(t)}{\mathrm{d}t} = \mathbf{K} \cdot \underline{r}(t) \cdot V_\mathrm{X}(t) + \mathbf{C}_\mathrm{in} \cdot \underline{u}^T - \underline{\nu}(t) \quad ,$$

where the state variables are the masses of the measured compartments m_{C_i}, the remaining biomass

$$m_\mathrm{Xr} = m_\mathrm{X} - \sum_i m_{\mathrm{C}_i} \quad , \tag{6.2}$$

the measured substrates m_{S_j}, and, if existing, the measured product m_P.

It is assumed that there are as many compartment building-up reactions r_{C_i} as state variables introduced to describe compartments,

$$\sum_j Y_{\mathrm{S}_j\mathrm{C}_i}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{C}_i}} \mathrm{C}_i \quad . \tag{6.3}$$

Initially, all substrates S_j used in the experiments are included in this reaction step. Some of them may be excluded later. The building-up reaction rate r_{C_i} is calculated by

$$r_{\mathrm{C}_i}(t) = \mu_{\mathrm{C}_i\mathrm{m}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot g_{\mathrm{C}_k}(t) \quad , \tag{6.4}$$

where $g_{\mathrm{C}_k}(t)$ indicates the amount of a yet to be allocated compartment C_k to which the building-up reaction rate is proportional. By applying the biological knowledge of replication, transcription and translation introduced in Section 2.2, default values for C_k can be determined. To reflect replication and transcription, $r_\mathrm{D}(t)$ and $r_\mathrm{R}(t)$ are proportional to DNA each, i.e., $g_{\mathrm{C}_k}(t) = g_{\mathrm{D}}(t)$. Translation is considered by $r_{\mathrm{Pr}}(t)$, which is proportional to RNA ($g_{\mathrm{C}_k}(t) = g_{\mathrm{R}}(t)$). As the concept of the remaining biomass m_{Xr} does not bear any resemblance to a biological process but is the result of the lumping process, no preselection can be made beforehand with respect to $r_{\mathrm{Xr}}(t)$ and every compartment should be regarded as a possible option for C_k.

Degradation reactions r_{dC_i} exist for all compartments except for DNA and Xr,

$$\mathrm{C}_i \xrightarrow{r_{\mathrm{dC}_i}} Y_{\mathrm{C}_i\mathrm{Xr}}\,\mathrm{Xr} \quad , \tag{6.5}$$

with

$$r_{\mathrm{dC}_i}(t) = \mu_{\mathrm{dC}_i\mathrm{m}} \cdot g_{\mathrm{C}_i}(t) \quad . \tag{6.6}$$

While the assumption is certainly questionable that the product of degradation is always the remaining biomass, which is viewed as some sort of inactive biomass, it should be kept in mind that this assumption can be relaxed in what follows. Likewise, a degradation of DNA could be included. Here, the compartment degradation rate is proportional to the mass of the compartment itself. No regulation is considered here, assuming that the mechanism behind the degradation is just a simple dissolution or passive degradation.

6.1.3 Product Formation

With respect to the product formation, it is initially assumed that the product can only be built when all substrates are present, according to Eq. (2.17),

$$\sum_j Y_{\mathrm{S}_j\mathrm{P}}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{P}}} \mathrm{P} \quad .$$

In the unstructured model, the product formation rate $r_{\mathrm{P}}(t)$ is described by

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot c_{\mathrm{X}}(t) \quad , \tag{6.7}$$

and in the structured model by

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot g_{\mathrm{C}_k}(t) \quad . \tag{6.8}$$

By default, $r_{\mathrm{P}}(t)$ is proportional to the amount of DNA. Within the model development process, the default values for each C_k can be changed and possible options for ‘limit’ can be specified, see below.

6.2 Influence of Biological Phenomena on the Model Structure

Every detected phenomenon can now modify the basic structures presented. If a phenomenon falsifies an assumption that has been made to include a specific part of the basic model structure, this part will be deleted. Otherwise, if a phenomenon proves that an additional model component should be considered, corresponding parts will be added to the structure.

A note on simultaneous consumption/growth

When the *simultaneous consumption or growth of two substances* (phenomena P2, P15, P16, and P19) is detected, it is understood that these two substances under consideration have to appear together in at least one reaction of the network. However, the more complex the network will get, especially after including unmeasured components, the more unclear it becomes which reaction or reactions are affected by this phenomenon. Instead, several up to many combinations have to be tested. To avoid having too many possible model structures that are proposed at the end of this procedure, the aforementioned phenomena will not be considered here. They are left to the expert modeler who has to decide if and how the model structure has to be changed adequately.

6.2.1 Modifying the Basic Model Structures

Some phenomena only change some entries in the stoichiometric matrix or change the kinetic type. If an initially supposed limiting dependency on a substrate S_l is disproved, i.e., the phenomena P1, P4, or P9 (*growth, product formation, or compartment formation limited by a substrate*), this substrate is not considered in the corresponding reaction anymore, i.e., its entry in the stoichiometric matrix is deleted and the limiting dependency in the reaction rate disappears. When considering growth, for example, the growth reaction (2.11) then changes to

$$\sum_{j \neq l} Y_{\mathrm{S}_j\mathrm{X}}\, \mathrm{S}_j \xrightarrow{r_\mathrm{X}} \mathrm{X} \tag{6.9}$$

and the growth rate $r_\mathrm{X}(t)$ is then calculated by

$$r_\mathrm{X}(t) = \mu_{\mathrm{Xm}} \cdot \prod_{j \neq l} \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot c_\mathrm{X}(t) \quad . \tag{6.10}$$

Otherwise, if a limiting dependency on a component is proven that has not been considered yet in the basic structure, i.e., the phenomena P13 or P17 (*compartment formation or product formation limited by a compartment*), it will be included in the model structure. A corresponding entry in the stoichiometric matrix is added and in the building-up reaction, a limiting dependency on that component appears. For example, if the phenomenon P17: *product formation limited by compartment* C_l is proven, the product formation will be described by

$$\sum_j Y_{\mathrm{S}_j\mathrm{P}}\,\mathrm{S}_j + Y_{\mathrm{C}_l\mathrm{P}}\,\mathrm{C}_l \xrightarrow{r_\mathrm{P}} \mathrm{P} \quad , \tag{6.11}$$

where $r_\mathrm{P}(t)$ will be defined by

$$r_\mathrm{P}(t) = \mu_\mathrm{Pm} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot \mathrm{limit}(g_{\mathrm{C}_l}(t)) \cdot g_{\mathrm{C}_k}(t) \quad . \tag{6.12}$$

In addition, limiting dependencies should only occur if a compensated substrate vanishes at some time or is very low. If a substrate does not disappear or does not at least assume low values in any experiment, the limiting term in the reaction rate will also be deleted. It might be hard to be identified based on the experiments given.

Likewise, if the phenomena detection proves that there is an inhibiting dependency (phenomena P5, P10, P14, P18), the model structure will be modified accordingly. If, for instance, the phenomenon P5: *product formation inhibited by a substrate* S_l is detected, Eq. (2.17) will change to

$$\sum_{j\neq l} Y_{\mathrm{S}_j\mathrm{P}}\,\mathrm{S}_j \xrightarrow{r_\mathrm{P}} \mathrm{P} \quad , \tag{6.13}$$

the corresponding entry in $\mathbf{K}$ will be deleted and the production rate $r_\mathrm{P}(t)$ must contain an inhibiting function of the substrate S_l,

$$r_\mathrm{P}(t) = \mu_\mathrm{Pm} \cdot \prod_{j\neq l} \mathrm{limit}(c_{\mathrm{S}_j}) \cdot \mathrm{inhib}(c_{\mathrm{S}_l}) \cdot g_{\mathrm{C}_k}(t) \quad . \tag{6.14}$$

The same procedure will apply to degradation reactions, if the phenomenon P11 is proven.

6.2.2 Extending the Basic Unstructured Model

The detection of other phenomena will change the whole structure of the model. If a storage $\mathrm{S}_l\mathrm{St}$ of a substrate S_l is detected, i.e., the phenomena P7 or P8 are proven, the basic unstructured model will be affected. A new state $m_{\mathrm{S}_l\mathrm{St}}$, describing the dynamic behavior of the storage, has to be established. Since the storage is seen as a compartment of the biomass, m_X itself is not a state anymore but the part of the biomass without the storage (active biomass),

$$m_\mathrm{Xa} = m_\mathrm{X} - \sum_j m_{\mathrm{S}_j\mathrm{St}} \quad . \tag{6.15}$$

The assumed growth reaction (2.11) is replaced by Eq. (2.28),

$$\sum_j Y_{\mathrm{S}_j\mathrm{Xa}}\,\mathrm{S}_j \xrightarrow{r_{\mathrm{Xa}}} \mathrm{Xa} \quad ,$$

where $r_{\mathrm{Xa}}(t)$ is defined by

$$r_{\mathrm{Xa}}(t) = \mu_{\mathrm{Xam}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot g_{\mathrm{Xa}}(t) \quad . \tag{6.16}$$

The product formation rate $r_{\mathrm{P}}(t)$ (Eq. (6.7)) is modified by

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \prod_j \mathrm{limit}(c_{\mathrm{S}_j}(t)) \cdot g_{\mathrm{Xa}}(t) \quad . \tag{6.17}$$

Then, two additional reactions for storage synthesis and storage degradation have to be introduced. The reaction scheme

$$\mathrm{S}_l \underset{r_{\mathrm{dS}_l\mathrm{St}}}{\overset{r_{\mathrm{S}_l\mathrm{St}}}{\rightleftarrows}} \mathrm{S}_l\mathrm{St} \tag{6.18}$$

is assumed. Here, the substrate S_l is stored by the storage $\mathrm{S}_l\mathrm{St}$ which releases the substrate into the medium when necessary. Once being released, the substrate is then consumed according to Eq. (6.16). As an alternative to this assumption, a more complex growth reaction has to be defined that additionally depends on the storage $\mathrm{S}_l\mathrm{St}$.

Depending on the storage type that is detected, the storage synthesis rate $r_{\mathrm{S}_l\mathrm{St}}(t)$ is calculated differently. If Storage A (P7, storage of the vanished substrate S_l) is proven, $r_{\mathrm{S}_l\mathrm{St}}(t)$ is supposed to be described by

$$r_{\mathrm{S}_l\mathrm{St}}(t) = \mu_{\mathrm{S}_l\mathrm{Stm}} \cdot \mathrm{limit}(c_{\mathrm{S}_l}(t)) \cdot g_{\mathrm{Xa}}(t) \quad . \tag{6.19}$$

In the case of Storage B (P8), substrate S_l is stored while another substrate S_0 vanishes and substrate S_c is constant. It seems possible that the storage synthesis depends on the amount of the vanishing substrate S_0. For instance, the storage synthesis might not take place until S_0 is depleted. That is, an inhibiting dependency on this substrate should not be neglected. The influence of S_c on the storage synthesis is unclear. Therefore, the synthesis rate is supposed to be calculated as

$$r_{\mathrm{S}_l\mathrm{St}}(t) = \mu_{\mathrm{S}_l\mathrm{Stm}} \cdot \mathrm{limit}(c_{\mathrm{S}_l}(t)) \cdot \mathrm{inhib}^\star(c_{\mathrm{S}_0}(t)) \cdot f_{\mathrm{S}_l\mathrm{St}}(c_{\mathrm{S}_c}(t)) \cdot g_{\mathrm{Xa}}(t) \quad , \tag{6.20}$$

where $\mathrm{inhib}^\star(c_{\mathrm{S}_0}(t))$ contains the possibility that the synthesis rate might either be inhibited by the substrate S_0 or is independent of this substrate, i.e., $\mathrm{inhib}^\star = \{\mathrm{inhib} \cup 1\}$, and $f_{\mathrm{S}_l\mathrm{St}}(c_{\mathrm{S}_c}(t))$ stands for any dependency (limiting, inhibiting, none).

The storage degradation rate $r_{\mathrm{dS}_l\mathrm{St}}(t)$ is defined by

$$r_{\mathrm{dS}_l\mathrm{St}}(t) = \mu_{\mathrm{dS}_l\mathrm{Stm}} \cdot f_{\mathrm{dS}_l\mathrm{St}}(\underline{c}_{\mathrm{S}}(t)) \cdot g_{\mathrm{S}_l\mathrm{St}}(t) \quad . \tag{6.21}$$

The term $f_{\mathrm{dS}_l\mathrm{St}}(\underline{c}_\mathrm{S}(t))$ describes possible influences of substrates on the degradation rate. Assuming that the degradation is a simple dissolution, $f_{\mathrm{dS}_l\mathrm{St}}(\underline{c}_\mathrm{S}(t)) = 1$. However, since the division of the cell into active biomass and storages is a very simple approach to describe the processes, it might be necessary to include more complex, i.e., regulated degradation rates, as well. Therefore, the following definition of $f_{\mathrm{dS}_l\mathrm{St}}(\underline{c}_\mathrm{S}(t))$ is considered:

$$f_{\mathrm{dS}_l\mathrm{St}}(\underline{c}_\mathrm{S}(t)) = \mathrm{inhib}^\star(c_{\mathrm{S}_l}) \cdot \prod_{j \neq l} \mathrm{limit}^\star(c_{\mathrm{S}_j}(t)) \quad , \tag{6.22}$$

where $\mathrm{limit}^\star = \{\mathrm{limit} \cup 1\}$ describes either a limiting dependency on substrate S_j or no dependency at all. This description incorporates two assumptions:

- The amount of substrate S_l—which will be stored in $\mathrm{S}_l\mathrm{St}$—can inhibit the degradation of the storage. As long as there is a considerable amount of substrate S_l, there is no need to empty the storage.

- To continue growing, the storage degradation will take place if no substrate S_l is present. In a defined medium, the cell also needs the other essential substrates $\mathrm{S}_{j \neq l}$ to grow. If one of these substrates is missing, the storage does not need to be emptied. Therefore, the storage degradation only takes place as long as the substrates $\mathrm{S}_{j \neq l}$ are present.

However, a simple dissolution is still possible. The parameter identification, parameter validation and model discrimination steps will tell, which assumption fits the measurements best.

At last, the metabolism is now described by $\mathbf{K} \cdot \underline{r}(t) \cdot V_\mathrm{X}(t)$ instead of $\mathbf{K} \cdot \underline{r}(t) \cdot V(t)$.

6.2.3 Extending the Basic Structured Model

If a phenomenon indicates that an intermediate compartment C_i^* should be introduced as precursor of C_i (P12), this compartment C_i^* has to be established in the structure. A new state $m_{\mathrm{C}_i^*}$, separated from m_{Xr}, has to be added to the existing model and additional reactions have to be formulated. In addition to the compartment building-up reaction (6.3) of C_i, the following reaction scheme is introduced,

$$\sum_l Y_{\mathrm{S}_l\mathrm{C}_i^*}\, \mathrm{S}_l \xrightarrow{r_{\mathrm{C}_i^*}} \mathrm{C}_i^* \xrightarrow{r_{\mathrm{C}_i^*\mathrm{C}_i}} \mathrm{C}_i \quad , \tag{6.23}$$

which takes into account that, alternatively, C_i can be built by the newly introduced intermediate compartment C_i^*. The reaction rates $r_{\mathrm{C}_i^*}(t)$ and $r_{\mathrm{C}_i^*\mathrm{C}_i}(t)$ are calculated as

$$r_{\mathrm{C}_i^*}(t) = \mu_{\mathrm{C}_i^*\mathrm{m}} \cdot \prod_l \mathrm{limit}(c_{\mathrm{S}_l}(t)) \cdot g_{\mathrm{C}_i}(t) \tag{6.24}$$

and

$$r_{\mathrm{C}_i^*\mathrm{C}_i}(t) = \mu_{\mathrm{C}_i^*\mathrm{C}_i\mathrm{m}} \cdot \mathrm{limit}(g_{\mathrm{C}_i^*}(t)) \cdot g_{\mathrm{C}_i}(t) \quad , \tag{6.25}$$

If an intermediate compartment is added as a precursor to RNA, the reactions (6.24) and (6.25) are proportional to the amount of DNA, i.e., $g_{\mathrm{C}_i}(t) = g_{\mathrm{D}}(t)$ in this case. The substrates S_l represent all the substrates for which the phenomenon could be proven. The corresponding degradation reaction (6.5) will also be changed to

$$\mathrm{C}_i \xrightarrow{r_{\mathrm{dC}_i}} Y_{\mathrm{C}_i\mathrm{C}_i^*}\,\mathrm{C}_i^* + Y_{\mathrm{C}_i\mathrm{Xr}}\,\mathrm{Xr} \quad . \tag{6.26}$$

The degradation rate $r_{\mathrm{dC}_i}(t)$ (Eq. (6.6)) is modified, it might now be inhibited by C_i^*,

$$r_{\mathrm{dC}_i}(t) = \mu_{\mathrm{dC}_i\mathrm{m}} \cdot \mathrm{inhib}^\star(g_{\mathrm{C}_i^*}(t)) \cdot g_{\mathrm{C}_i}(t) \quad , \tag{6.27}$$

relaxing the assumption of a simple dissolution and allowing for a regulated degradation of C_i.

By default, the reactions (6.24), (6.25), and (6.27) are proportional to the amount of C_i to describe exponential growth, for which C_i^* acts as a precursor. A final decision is made by the user to whom an updated overview of the reaction network is presented, and who can change these proportionalities.

Only one intermediate compartment C_i^* can be assigned to a measured compartment C_i. DNA and RNA share the same C_i^*, considering that both share the same precursor (nucleotides) and that their C_i^* could take on the role of that precursor. It has to be emphasized again that the reaction scheme and the (intermediate) compartments do not necessarily describe the real metabolism at a microscopic scale but rather are lumped states combining several cellular functions and dynamics. The reason for these simplifications is the aim to get a manageable process model for process control application while giving significantly more flexibility for the dynamic description than with the often used unstructured models. Any uncertainty introduced can then be compensated for by closed-loop control methods.

After changing $\underline{m}$ and $\underline{r}$, $\mathbf{K}$ has to be adapted as well.

6.2.4 Structure Probabilities

The changes caused by the phenomena take place according to the corresponding values of the score Sc. If a phenomenon results in a deletion of a part, this change is done for all existing structure proposals if $\mathrm{Sc} \in [-1, -2/3)$. Alternative model structures to the already existing ones are built if $\mathrm{Sc} \in [-2/3, -0.2)$. If a phenomenon results in adding something new, this change is done for all existing structure proposals if $\mathrm{Sc} \in (2/3, 1]$. Alternative model structures are built if $\mathrm{Sc} \in (0.2, 2/3]$. Other values will not change anything. After evaluating each analyzed phenomenon Pi, a probability for the changed structures $\mathcal{S}_k$ is calculated by

$$P^i_{\mathcal{S}_k} = \frac{|\mathrm{Sc}| + 1}{2} \cdot P^{i-1}_{\mathcal{S}_k} \tag{6.28a}$$

and accordingly the probabilities of all unchanged structures $\mathcal{S}_l$ have to be reduced,

$$P^i_{\mathcal{S}_l} = \frac{1 - |\mathrm{Sc}|}{2} \cdot P^{i-1}_{\mathcal{S}_l} \quad , \tag{6.28b}$$

where the initial model starts with $P^0_{\mathcal{S}_{\mathrm{init}}} = 1$. The proposed structures are ordered, with priority given to those with a high structure probability.

At the end of this automatic procedure, several up to many model structures are proposed to explain the measurements of the experiments. Each structure still describes a whole model family, as for the associated models the possible kinetics used to calculate the reaction rates need yet to be fixed. All of these models and their corresponding parameter files are coded automatically by RapOpt (Violet et al., 2009) in a MATLAB m-file and coded and compiled in C. Subsequently, a parameter identification of these models has to be done to get the values of the yield coefficients and kinetic parameters.

6.3 Parameter Identification

To identify the optimal model parameters $\underline{\widehat{\theta}}$ that match the corresponding model best, a cost function Φ_{PI} has to be minimized, taking into account some constraints:

$$\left[\underline{\widehat{\theta}},\, \underline{\widehat{x}}_0\right] = \arg\min_{\underline{\theta},\, \underline{x}_0} \left(\Phi_{\mathrm{PI}}(\underline{\theta},\, \underline{x}_0)\right) \tag{6.29a}$$

$$\text{s.t.} \quad \frac{\mathrm{d}\underline{x}(t)}{\mathrm{d}t} = \underline{f}(\underline{x}(t),\, \underline{u}(t),\, t,\, \underline{\theta}), \quad \underline{x}(t=0) = \underline{x}_0 \tag{6.29b}$$

$$\underline{y}(t_k) = \underline{h}(\underline{x}(t_k),\, t_k,\, \underline{\theta}) \tag{6.29c}$$

$$\underline{\theta}_{\min} \le \underline{\theta} \le \underline{\theta}_{\max} \quad . \tag{6.29d}$$

Eqs. (6.29b) and (6.29c) are a general description of the model, where $\underline{x}(t)$ is the vector of state variables, $\underline{u}(t)$ constitutes the input variables, and $\underline{y}(t)$ is the vector of measured variables. The differential equations are given by $\underline{f}(\cdot)$ and the measurement equations by $\underline{h}(\cdot)$. To account for states that cannot be measured, optimal initial values $\underline{\widehat{x}}_0$ can also be calculated. The parameter values are usually bounded by minimal and maximal values $\underline{\theta}_{\min}$ and $\underline{\theta}_{\max}$, as described by Eq. (6.29d).

In this work, a weighted least squares (WLS) approach for the cost function is chosen,

$$\Phi_{\mathrm{PI}} = \sum_{j=1}^{N_{\mathrm{Exp}}} \sum_{i=1}^{N_j} \Big(\underline{y}(t_i) - \underline{h}(\underline{x}(t_i),\, t_i,\, \underline{\theta})\Big)^T \mathbf{C}^{-1}_{\underline{y}(t_i)} \Big(\underline{y}(t_i) - \underline{h}(\underline{x}(t_i),\, t_i,\, \underline{\theta})\Big) \quad , \tag{6.30}$$

where N_{Exp} is the number of experiments used for the parameter identification and N_j is the total number of time points where measurements are taken in the j-th experiment. The covariance matrix of the measurement noise, $\mathbf{C}_{\underline{y}(t_i)}$, is assumed to be known,

$$\mathbf{C}_{\underline{y}(t_i)} = \begin{pmatrix} (\sigma_1(t_i))^2 & 0 & \cdots & 0 \\ 0 & (\sigma_2(t_i))^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\sigma_q(t_i))^2 \end{pmatrix} \quad . \tag{6.31}$$

The standard deviations of the measurement error, $\sigma_j(t_i)$, are determined by linear approximations

$$\sigma_j(t_i) = a_j \cdot c_j(t_i) + b_j \quad , \tag{6.32}$$

the parameters a_j and b_j are specified in Tables 5.2, A.2, B.2, C.2, and D.2.

The minimization problem (6.29a)–(6.29d) is solved by algorithms based on sequential quadratic programming (SQP), e.g., SNOPT by TOMLAB®.

6.4 Model Selection and Model Discrimination

If there are several potential model candidates $\mathcal{M}_i$, $i = 1, \ldots, N_{\text{Mod}}$, that can describe the experimental data, the most plausible model has to be found out. A widely used approach that accounts for both goodness-of-fit and model complexity is Akaike's Information Criterion (AIC) (Akaike, 1974, Burnham and Anderson, 2002). If a least squares approach is used for the cost functional Φ_{PI}, the AIC is calculated by

$$\text{AIC} = n \ln \left(\frac{\Phi_{\text{PI}}}{n} \right) + 2K^\star \quad , \tag{6.33}$$

where n is the size of experimental data used for the parameter identification and $K^\star = K + 1$ is the number of identified model parameters (K) plus one. If the sample size is small in relation to the number of identified parameters ($n/K^\star < 40$), Burnham and Anderson (2002) suggest using a corrected AIC instead,

$$\text{AIC}_c = \text{AIC} + \frac{2K^\star(K^\star + 1)}{n - K^\star - 1} = \text{AIC} + \frac{2(K+1)(K+2)}{n - K} \quad . \tag{6.34}$$

To compare the different potential model candidates, the AIC differences

$$\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}} \quad . \tag{6.35}$$

can be calculated. The model candidates can be ranked accordingly, prioritizing those with low Δ_i values. The relative likelihood of a model can be expressed by the Akaike weight

$$w_{\mathcal{M}_i} = \frac{\exp\left(-\Delta_i/2\right)}{\sum\limits_{j=1}^{m} \exp\left(-\Delta_j/2\right)} = P_{\mathcal{M}_i} \quad , \tag{6.36}$$

which is equivalent to a probability value $P_{\mathcal{M}_i}$ of model candidate $\mathcal{M}_i$ (Burnham and Anderson, 2002, Schenkendorf and Mangold, 2013).

Besides using already conducted experiments to select a plausible model among multiple candidates, future experiments can be planned to discriminate between these candidates, as well. For this purpose, an adequate stimulus of the process has to be established that can expose the differences between the model candidates. For fed-batch experiments, optimal flow rates $\widehat{\underline{u}}(t)$ have to be identified that lead to different predicted measurements $\underline{y}_{\mathcal{M}_i}(t_k)$. After having conducted the experiment, the predictions can be compared to the measurements and the best model can be detected.

To calculate $\widehat{\underline{u}}(t)$, a cost functional Φ_{MD} has to be minimized:

$$\widehat{\underline{u}}(t) = \arg\min_{\underline{u}(t)} \left(\Phi_{\mathrm{MD}}(\underline{u}(t))\right) \tag{6.37a}$$

$$\text{s.t.} \quad \frac{\mathrm{d}\underline{x}_{\mathcal{M}_i}(t)}{\mathrm{d}t} = \underline{f}_{\mathcal{M}_i}(\underline{x}_{\mathcal{M}_i}(t),\, \underline{u}(t),\, t,\, \underline{\theta}_{\mathcal{M}_i}), \quad \underline{x}_{\mathcal{M}_i}(t=0) = \underline{x}_{\mathcal{M}_i,0} \tag{6.37b}$$

$$\underline{y}_{\mathcal{M}_i}(t_k) = \underline{h}_{\mathcal{M}_i}(\underline{x}_{\mathcal{M}_i}(t_k),\, t_k,\, \underline{\theta}_{\mathcal{M}_i}) \tag{6.37c}$$

$$0 \le \underline{u}(t) \le \underline{u}_{\max} \quad , \tag{6.37d}$$

with $i = 1, \ldots, N_{\mathrm{Mod}}$. In this work, the flow rate $\underline{u}(t)$ is specified as a zero-order hold, and the sampling interval of this input is given. Due to technical constraints, the flow rates are bounded, see Eq. (6.37d).

To get different model outcomes $\underline{y}_{\mathcal{M}_i}(t_k)$, the cost functional Φ_{MD} has to consider all possible differences

$$\Delta_{\mu\nu,j}(t_k) = \left| y_{\mathcal{M}_\mu,j}(t_k) - y_{\mathcal{M}_\nu,j}(t_k) \right| \tag{6.38}$$

between two models $\mathcal{M}_\mu$ and $\mathcal{M}_\nu$, $\mu, \nu = 1, \ldots, N_{\mathrm{Mod}}$, $\nu \neq \mu$, for all measurement variables y_j, $j = 1, \ldots, q$, at all sampling times t_k, $k = 1, \ldots, N$. Differences $\Delta_{\mu\nu,j}(t_k) > 0$ have to be rewarded. If two or more model candidates show the same outcome, i.e., $\Delta_{\mu\nu,j}(t_k) = 0$, this has to be penalized. Furthermore, the measurement noise should be considered, as well. Terziev (2014) developed a cost functional Φ_{MD} that satisfies these requirements:

$$\Phi_{\mathrm{MD}} = -\sum_{i=1}^{N}\sum_{j=1}^{q}\left(\left(\sum_{\mu=1}^{N_{\mathrm{Mod}}} \sigma_{\mathcal{M}_\mu,j}(t_i)\right)^{-1}\left(\prod_{\mu=1}^{N_{\mathrm{Mod}}}\prod_{\nu=\mu+1}^{N_{\mathrm{Mod}}} \Delta_{\mu\nu,j}(t_k)\right)^{\frac{1}{N_{\mathrm{Comb}}}}\right) \quad , \tag{6.39}$$

with $N_{\mathrm{Comb}} = \binom{N_{\mathrm{Mod}}}{2}$. The standard deviation of the measurement error of model $\mathcal{M}_\mu$, $\sigma_{\mathcal{M}_\mu,j}(t_i)$, is linearly approximated according to Eq. (6.32),

$$\sigma_j(t_i) = a_j \cdot c_{\mathcal{M}_\mu,j}(t_i) + b_j \quad . \tag{6.40}$$

The minimization problem (6.37a)–(6.37d) is solved by SQP algorithms as well.

Chapter 7

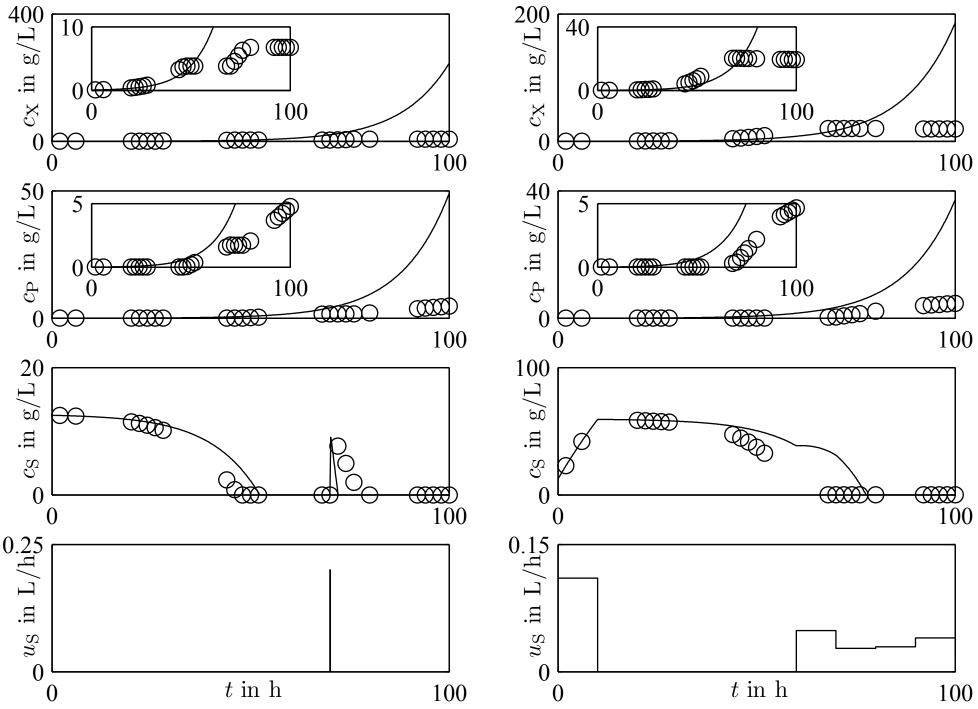
Automated Detection of Model Deficiencies

So far, it has been shown how detected biological phenomena can be used to propose model structures that might describe the dynamic behavior of the measurements. As described in Chapter 6, the score $\mathrm{Sc}_{\mathrm{P}i}$ of a detected phenomena determines if alternative model structures are proposed that consider the phenomenon-related model parts or if the changes are applied to all existing structures. In the first case, the previous existing structures remain unchanged. In the second case, no alternatives need to be proposed (see Section 6.2.4). In the worst case, i.e., $0.2 < |\mathrm{Sc}_{\mathrm{P}i}| < 2/3$, for every phenomenon P$i$ an alternative structure is built and the number of model structures to be investigated is doubled. Then, every structure is a general description for a family of several model candidates. The number of these candidates is determined by the number of kinetic laws used to replace the 'limit' and 'inhib' terms in the model structures. It is obvious that the presented approach to identify convenient models might propose too many model candidates that are all subjected to a parameter identification step—a procedure that costs time and other useful resources. Moreover, identifying many model candidates does not necessarily mean that a model is found which describes the measurements and their underlying dynamics sufficiently well.

In this chapter, an alternative to the aforementioned 'broad' approach is drafted. Here, only few models are considered initially. After the parameter identification step, the simulations are compared to the measurements. Based on the differences between the measurements and the simulations, model deficiencies are able to be identified and to be eliminated.

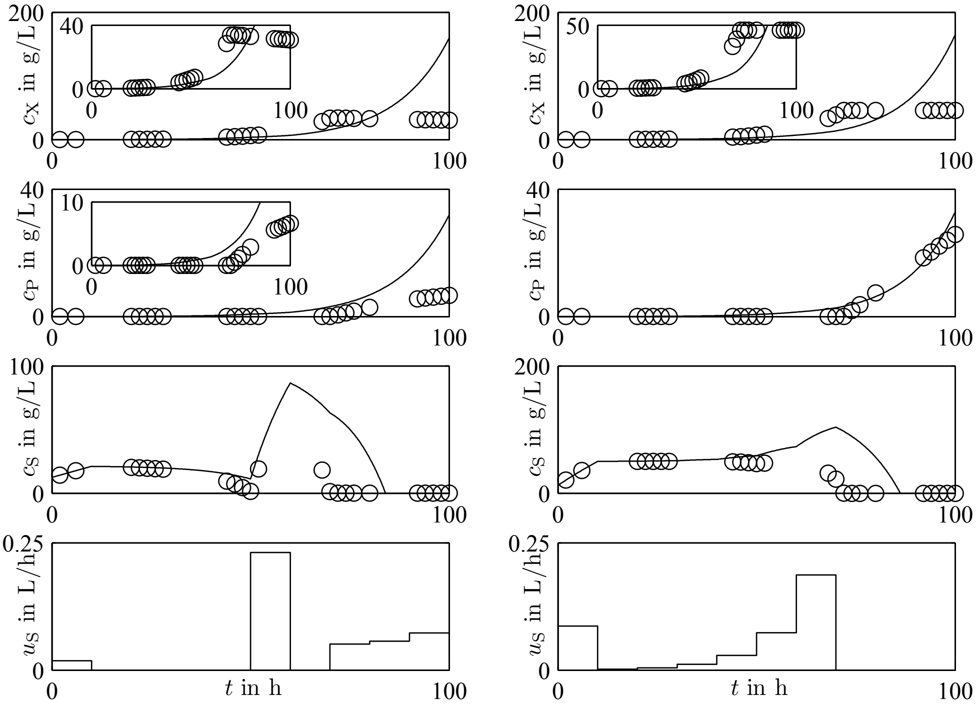
7.1 Detecting Model Deficiencies and Proposing Improvements

To detect model deficiencies and, hence, being able to propose model improvements, it is necessary to look at the measurements, to compare them to the simulations, to notice any different behavior, and to understand the causes for these differences. For example, looking at Figure 7.1 where the measurements of the experiments ME1–ME4 are compared to simulations of a model that is assumed describing the process. As can be seen, the measurements and the simulations match badly. From these data, an experienced human modeler will conclude that growth reaction lacks a limiting influence



(a) ME1

(b) ME2



(c) ME3

(d) ME4

Figure 7.1: Comparison between measurements and simulations of an assumed model. The inset plots show the same information with a differently scaled concentration axis.

and he or she will change the model accordingly. But how can this deficiency—and model deficiencies in general—be automatically detected?

At first, the differences between the measurements and the simulations need to be listed. Questions such as the following two need to be considered. Do events occur in the measurements that are not explained by the model? Does the model show a behavior that cannot be found in the measurements? The deficiencies identified can now be ranked by the time they occurred. Before an event at the end of an experiment can be considered for model improvement, all preceding deficiencies need to be resolved.

Having determined the differences, the question arises regarding any modifications needed to improve the model. This depends on the type of the deficiency found. If the model lacks a limiting term in the growth reaction, this term has to be introduced into the model. If the model features an inhibiting term that cannot be found in the measurements, it needs to be removed. However, due to different model complexities and nonlinearities in the model, simple rules describing what part of the model needs to be changed when a certain behavior can be observed are generally not adaptable.

Now, instead of comparing only simulations and measurements, the information about the detected phenomena in the measurements is used. Here, important events, correlations, and the times they occur are already listed. They serve as expectations that should be satisfied by the simulated model. Therefore, the phenomena detection is applied to the simulated data of the model under consideration as well. The results can then be easily compared. Likewise, as soon as it becomes clear which phenomena are not considered by the model, or which phenomena are inherent to the model that cannot be found in the measurements, conclusions can be drawn quickly regarding which model part has to be changed to improve the model. Moreover, taking the phenomena detected on the basis of the measurements as the guideline, no model parts will be added that would lead to phenomena not being proven by the measurements. Likewise, model parts that are necessary according to the measurements cannot be deleted.

The detected deficiencies are then ranked to define an order regarding which deficiency should be first corrected. Since, usually, several experiments are used for the detection of the phenomena and the steps for the parameter identification, it is possible that, in each experiment, another deficiency appears first. So, unlike as explained above, other criteria than the first appearance are used. The first measure is the percentage of how often a phenomenon inherent to the measurements is not proven, the second criterion is the percentage for how often a phenomenon is wrongly proven or rejected. Then, according to the allocated rank of the detected deficiency, a model part eliminating this deficiency is to be added or removed and the parameters are identified again. An improvement can then be checked either by comparing the goodness-of-fit values or equivalents before and after the change or by testing the new model for deficiencies and comparing these results with each other.

The procedure should be applied as long as deficiencies can be detected and improvements can be suggested. A general flow chart is shown in Figure 7.2.

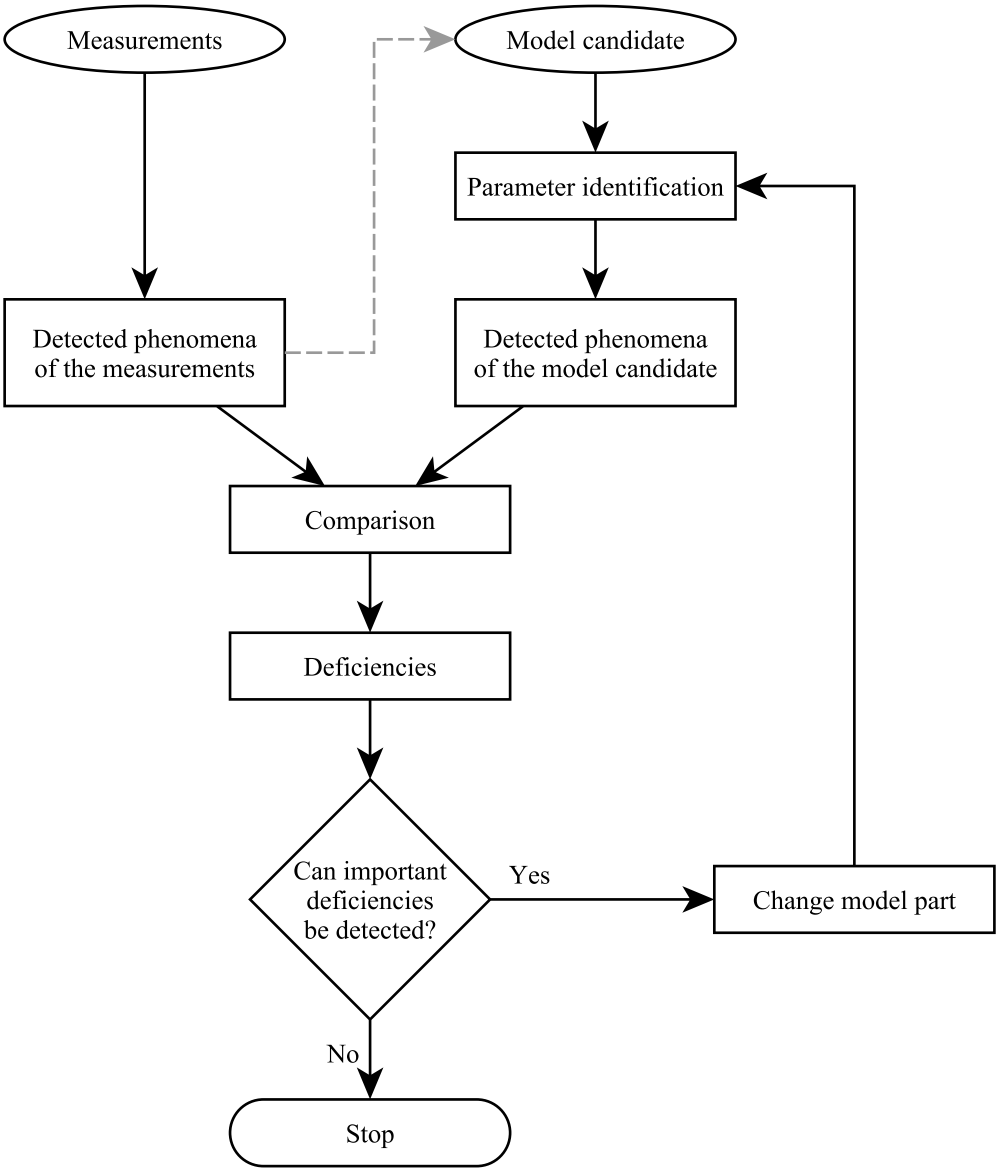

Measurements
Model candidate
Parameter identification
Detected phenomena of the measurements
Detected phenomena of the model candidate
Comparison
Deficiencies
Can important deficiencies be detected?
Yes
Change model part
No
Stop

Figure 7.2: Flow chart of the model deficiency detection

7.2 Case Studies

The presented deficiency detection is tested based on two models: the motivating example (Chapter 1) and a more complex, yet simple unstructured model (Appendix A). Each model possesses n specific characteristics (like limiting or inhibiting dependencies), and alternative models are generated where one up to n of these characteristics are neglected—considering every possible combination. This leads to a model family that comprises the basic model and $\sum\limits_{i=1}^{n} \binom{n}{i}$ alternative models. Each of the alternative models is then tested for deficiencies and an improved model is proposed. Within the whole model family, there will always be two models k and l that only differ in one specific characteristic, i.e., in model k, an influence is missing that exists in model l, the rest is the same. So, when model k is tested for deficiencies, the corresponding model l should be proposed as an improvement.

7.2.1 Motivating Example

All models derived from the motivating example can be described by the model structure

$\mathcal{S}_{\mathrm{ME}}$:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{X}}(t) \\ m_{\mathrm{S}}(t) \\ m_{\mathrm{P}}(t) \end{pmatrix} &= \begin{pmatrix} 0 \\ c_{\mathrm{S,\,in}} \cdot u_{\mathrm{S}}(t) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ r_{\mathrm{M}}(t) V(t) \\ 0 \end{pmatrix} + V(t) \begin{pmatrix} 1 & 0 \\ -Y_{\mathrm{SX}} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_{\mathrm{X}}(t) \\ r_{\mathrm{P}}(t) \end{pmatrix} \\
\frac{\mathrm{d}V(t)}{\mathrm{d}t} &= u_{\mathrm{S}}(t) \\
r_{\mathrm{X}}(t) &= \mu_{\mathrm{Xm}} \cdot \mu_{\mathrm{XS}}(t) \cdot c_{\mathrm{X}}(t) \\
r_{\mathrm{P}}(t) &= \mu_{\mathrm{Pm}} \cdot \mu_{\mathrm{PS}}(t) \cdot c_{\mathrm{X}}(t) \\
r_{\mathrm{M}}(t) &= \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{S}}(t)}{c_{\mathrm{S}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \\
\underline{y}(t) &= \begin{pmatrix} c_{\mathrm{X}}(t) \\ c_{\mathrm{S}}(t) \\ c_{\mathrm{P}}(t) \end{pmatrix} \quad .
\end{aligned}
$$

In the case of the motivating example (in the following model 1), the specific reaction rates are defined by

$$
\mu_{\mathrm{XS}}(t) = \frac{c_{\mathrm{S}}(t)}{c_{\mathrm{S}}(t) + K_{\mathrm{XS}}}
$$

Table 7.1: Model candidates of $\mathcal{S}_{\mathrm{ME}}$: allocation between model number and omitted model parts

Model	Omitted
1	—
2	$\mu_{\mathrm{PS}}(t)$
3	$\mu_{\mathrm{XS}}(t)$
4	$\mu_{\mathrm{PS}}(t)$, $\mu_{\mathrm{XS}}(t)$

and

$$\mu_{\mathrm{PS}}(t) = \frac{K_{\mathrm{PS}}}{c_{\mathrm{S}}(t) + K_{\mathrm{PS}}} \quad .$$

Three other model candidates are generated by omitting either $\mu_{\mathrm{XS}}(t)$ or $\mu_{\mathrm{PS}}(t)$ or both, i.e., they are set to one—neglecting the limiting or inhibiting dependency, respectively. The allocation between the model numbers and the omitted model parts can be found in Table 7.1. After a parameter identification step, each of the alternative models 2, 3, and 4 is tested for deficiencies, and corresponding model improvements are proposed. As can be seen in Table 7.1, model 1 and the models 2 and 3 differ only in one component, either $\mu_{\mathrm{XS}}(t)$ or $\mu_{\mathrm{PS}}(t)$ is missing. And models 2 and 3 and model 4 differ only in one component as well. If the deficiencies are detected correctly, models 2 and 3 should be proposed as an improvement to 4. Likewise, model 1 has to be found as an improvement to models 2 and 3.

Starting with model 4, the similarities and differences between the phenomena inherent to the measurements and the phenomena detected on basis of the simulations (see Figure 7.1) are shown in Table 7.2. With respect to the differences (right column), the types of entries that can be found in this table are:

- “not proven,” indicating that a measurement-inherent phenomenon in a specific experiment at a certain time cannot be found in the simulations, and
- “wrongly disproved,” meaning that a phenomenon is disproved by the simulations but cannot be analyzed on the basis of the measurements. For instance, when substrate-limited growth is wrongly rejected, a situation in the simulations occurs where the substrate is vanishing and the biomass is still growing. However, in the measurements, the substrate does not vanish and the phenomenon can neither be proved or disproved based on the data. “Wrongly disproved” does not mean that a phenomenon is found in the measurements and rejected in the simulations.

It is obvious that the model is not or hardly able to reproduce the phenomena *growth limited by substrate* or *product formation inhibited by substrate.* Here, the phenomenon

Table 7.2: Phenomena detection of the simulations of the assumed model: similarities with and differences to the phenomena found in the measurements

	Similarities	Differences
	Growth limited by substrate	
ME1		Not proven[a] at $t = 46.0\,\mathrm{h}$ Wrongly disproved[b] at $t = 51.8\,\mathrm{h}$ Wrongly disproved[b] at $t = 70.2\,\mathrm{h}$ Not proven[a] at $t = 80.6\,\mathrm{h}$
ME4		Not proven[a] at $t = 70.0\,\mathrm{h}$ Wrongly disproved[b] at $t = 84.2\,\mathrm{h}$
	Product formation inhibited by substrate	
ME1	Proven at $t = 46.0\,\mathrm{h}$	
ME2		Not proven[a] at $t = 61.6\,\mathrm{h}$
ME3		Not proven[a] at $t = 69.8\,\mathrm{h}$
ME4		Not proven[a] at $t = 71.8\,\mathrm{h}$

[a] Phenomena inherent to the measurements are not proven by the simulations.

[b] Phenomena are disproved by the simulations but cannot be analyzed on the basis of the measurements.

growth limited by substrate is not proven in 100 % of all cases detected in the measurements. Furthermore, this phenomenon is wrongly disproved three times, meaning that in the simulations, the biomass continues growing after the substrate has depleted (see also Figure 7.1). This depletion cannot be found in the measurements. However, this behavior is consistent to model 4 since $\mu_{\mathrm{XS}}(t)$ is set to one and therefore no limiting influence of the substrate on the biomass exists. Not proving the phenomenon *growth limited by substrate* and—to a lesser extent—wrongly disproving it are both indicators for this missing dependency.

The phenomenon *product formation inhibited by substrate* is not proven in 75 % of the cases detected in the measurements.

To improve the model behavior, either

$$\mu_{\mathrm{XS}}(t) = \mathrm{limit}(c_{\mathrm{S}}(t))$$

or

$$\mu_{\mathrm{PS}}(t) = \mathrm{inhib}(c_{\mathrm{S}}(t))$$

have to be introduced—leading to model 2 or 3, respectively. Since a lack of $\mu_{\mathrm{XS}}(t)$ is detected with a higher percentage, this specific reaction rate should be inserted first.

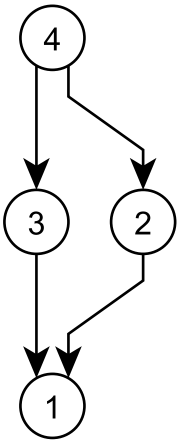


Figure 7.3: Deficiency detection with $\mathcal{S}_{\mathrm{ME}}$: It is shown which model is proposed after the deficiency detection.

The other models are also tested for deficiencies and the proposed improvements are recorded. Figure 7.3 is a summary of how the proposed changes in the model lead to other model candidates. It is evident that in both cases, model 1 is suggested as an improvement. In this case study, all deficiencies can be detected, model 1 is eventually proposed by the algorithm as the ultimate model, even if the starting point is a model lacking important components in the reaction rates.

7.2.2 Unstructured Model with three Substrates

The approach is now tested using data generated by the more complex, yet simple unstructured model given in Appendix A. The models are generated based on the model structure

$\mathcal{S}_{\mathrm{UM3S}}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{X}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \\ m_{\mathrm{P}}(t) \end{pmatrix} = \begin{pmatrix} 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t)V(t) \\ 0 \end{pmatrix} + V(t)\begin{pmatrix} 1 & 0 \\ -Y_{\mathrm{AmX}} & 0 \\ -Y_{\mathrm{PhX}} & 0 \\ -Y_{\mathrm{GcX}} & -Y_{\mathrm{GcP}} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} r_{\mathrm{X}}(t) \\ r_{\mathrm{P}}(t) \end{pmatrix}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t)$$

$$r_{\mathrm{X}}(t) = \mu_{\mathrm{Xm}} \cdot \mu_{\mathrm{XAm}}(t) \cdot \mu_{\mathrm{XPh}}(t) \cdot \mu_{\mathrm{XGc}}(t) \cdot c_{\mathrm{X}}(t)$$

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \mu_{\mathrm{PPh}}(t) \cdot \mu_{\mathrm{PGc}}(t) \cdot c_{\mathrm{X}}(t)$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{S}}(t)}{c_{\mathrm{S}}(t) + K_{\mathrm{MS}}} \cdot c_{\mathrm{X}}(t)$$

$$\underline{y}(t) = \begin{pmatrix} c_{\mathrm{X}}(t) \\ c_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc}}(t) \\ c_{\mathrm{P}}(t) \end{pmatrix} \quad ,$$

---

where the specific growth rates in the initial model (model 1) are given by

$$\begin{aligned} \mu_{\mathrm{XAm}}(t) &= \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{XAm}}} \\ \mu_{\mathrm{XPh}}(t) &= \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XPh}}} \\ \mu_{\mathrm{XGc}}(t) &= \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XGc}}} \\ \mu_{\mathrm{PPh}}(t) &= \frac{K_{\mathrm{PPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PPh}}} \\ \mu_{\mathrm{PGc}}(t) &= \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{PGc}}} \quad . \end{aligned}$$

First version

At first, 31 other model candidates are generated by omitting different specific growth rates $\mu_i(t)$ and all possible combinations thereof. Furthermore, if an omitted $\mu_{ij}(t)$ is given by the Michaelis–Menten law ($\mu_{\mathrm{XAm}}(t)$, $\mu_{\mathrm{XPh}}(t)$, $\mu_{\mathrm{XGc}}(t)$, $\mu_{\mathrm{PGc}}(t)$), the corresponding yield coefficient Y_{ji} is set to zero as well. That way, the dynamic behavior of the particular substrate on the one hand and of the biomass or the product on the other hand are completely uncoupled. The resulting models can be seen as model proposals if the phenomena *growth or product formation limited by a substrate* is not found or even neglected. The case where the Y_{ji} are not set to zero are considered below. An overview of the models and the omitted model parts is given in Table 7.3.

After the parameter identification, each model is tested for model deficiencies and possible improvements. As there will always be two models k and l that only differ in one specific characteristic, every proposed improvement should lead to a model already generated that has to be tested next. However, since the deficiency detection is based on phenomena inherent to the measurements, a deficiency will be hard to find if the corresponding phenomenon cannot be detected on the basis of the measurements. Here, three phenomena are not detected: the phosphate-limited growth, the glucose-limited growth, and the glucose-limited product formation cannot be found in the measurements. The lack of these corresponding specific growth rates $\mu_{\mathrm{XPh}}(t)$, $\mu_{\mathrm{XGc}}(t)$, and $\mu_{\mathrm{PGc}}(t)$ will therefore be hard to find.

Table 7.3: Model candidates of $\mathcal{S}_{\mathrm{UM3S}}$, first version: allocation between model number and omitted model parts

	Model number															
Omitted	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mu_{\mathrm{PGc}}(t)$		×		×		×		×		×		×		×		×
Y_{GcP}		×		×		×		×		×		×		×		×
$\mu_{\mathrm{PPh}}(t)$			×	×			×	×			×	×			×	×
$\mu_{\mathrm{XGc}}(t)$					×	×	×	×					×	×	×	×
Y_{GcX}					×	×	×	×					×	×	×	×
$\mu_{\mathrm{XPh}}(t)$									×	×	×	×	×	×	×	×
Y_{PhX}									×	×	×	×	×	×	×	×

	Model number															
Omitted	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$\mu_{\mathrm{PGc}}(t)$		×		×		×		×		×		×		×		×
Y_{GcP}		×		×		×		×		×		×		×		×
$\mu_{\mathrm{PPh}}(t)$			×	×			×	×			×	×			×	×
$\mu_{\mathrm{XGc}}(t)$					×	×	×	×					×	×	×	×
Y_{GcX}					×	×	×	×					×	×	×	×
$\mu_{\mathrm{XPh}}(t)$									×	×	×	×	×	×	×	×
Y_{PhX}									×	×	×	×	×	×	×	×
$\mu_{\mathrm{XAm}}(t)$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Y_{AmX}	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

The result of the deficiency detection is depicted in Figure 7.4. Black arrows show which model improvements are detected by the algorithm. Existing deficiencies that are not detected by the algorithm are indicated by gray arrows. Dashed arrows mean that the phenomenon necessary to detect a specific deficiency cannot be detected in the measurements. Here, 51 possible model improvements are detected, whereas 29 improvements cannot be found. Table 7.4 shows how often the absence of an individual specific growth rate can be found. As already mentioned above, some deficiencies are hard to detect because the necessary phenomenon have not been found in the measurements. This means that, depending on the starting point, model 1 cannot be reached in all cases. However, the deficiency detection is successful when the corresponding phenomenon is at hand. Here, in all 32 possible cases, the correct model improvements are proposed. Furthermore, in 19 cases, improvements can be found although the corresponding phenomena are not inherent to the measurements. In these cases, limiting dependencies on glucose are wrongly rejected. This means that in the simulations, situations occur where glucose is vanishing and the biomass or the product are still growing whereas, in the measurements, glucose does not deplete and the phenomena can neither be proven nor rejected. However, since the rejections of these limiting dependencies show a behavior that cannot be found in the measurements, the

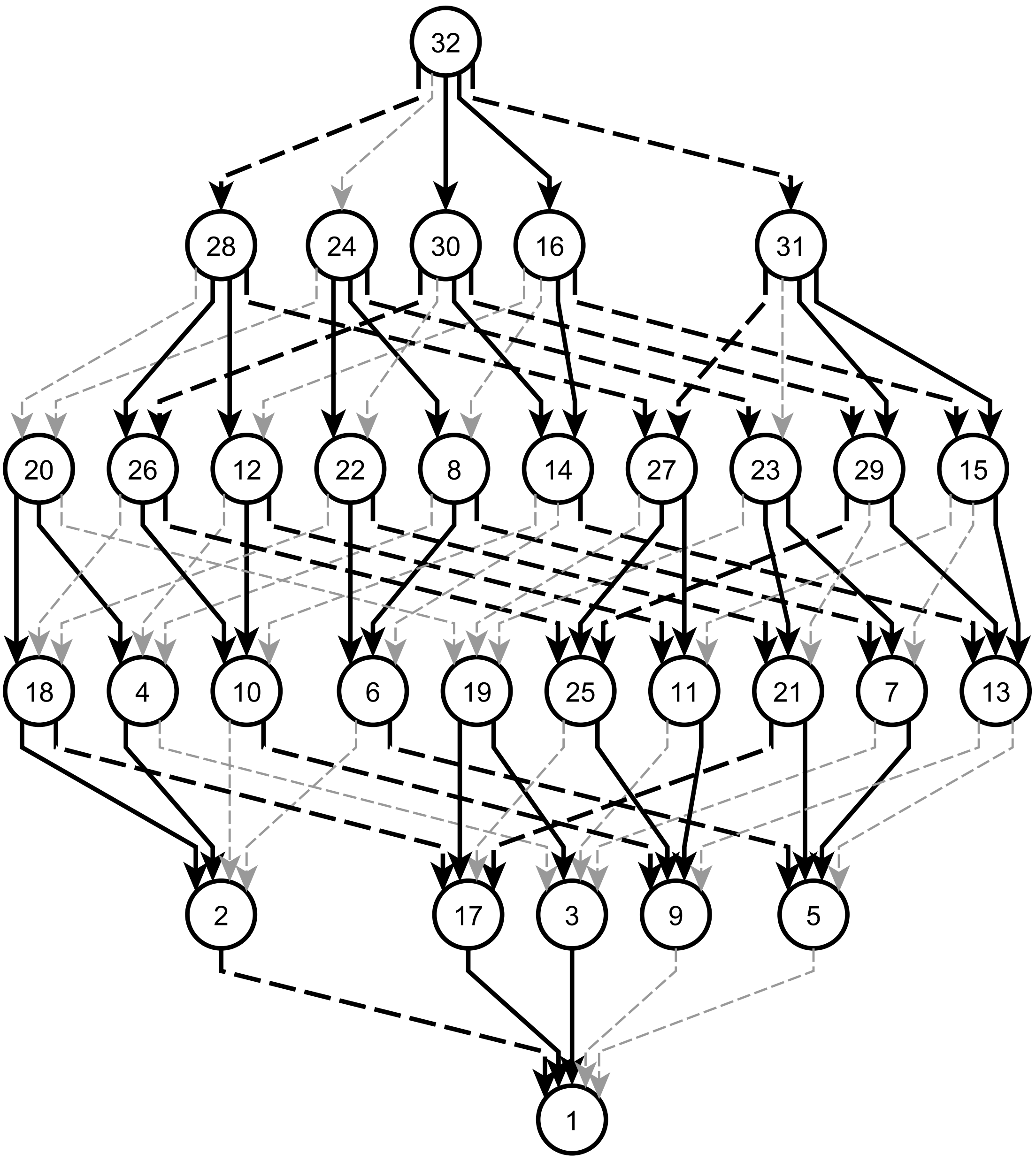


Figure 7.4: Deficiency detection with $\mathcal{S}_{\mathrm{UM3S}}$, first version: black arrows show which models are proposed after the deficiency detection (51), gray arrows indicate which possible model improvements are not detected (29). Solid arrows mean that the phenomenon necessary to detect a specific deficiency can be found in the measurements, dashed arrows show the absence of this phenomenon.

inclusion of these dependencies into the model is proposed as an improvement and will be tested. Taking for example model 2 which lacks $\mu_{\mathrm{PGc}}(t)$. The phenomenon *product formation limited by glucose* cannot be tested by the measurements. However, it is

Table 7.4: Deficiency detection with $\mathcal{S}_{\mathrm{UM3S}}$, first version: Specific growth rates $\mu_i(t)$ and how often a lack thereof is detected.

In the measurements, the necessary phenomenon is			
detected		not detected	
		$\mu_{\mathrm{XPh}}(t)$	0/16
$\mu_{\mathrm{XAm}}(t)$	16/16	$\mu_{\mathrm{XGc}}(t)$	5/16
$\mu_{\mathrm{PPh}}(t)$	16/16	$\mu_{\mathrm{PGc}}(t)$	14/16
	32/32		19/48

wrongly rejected by the simulations. To eliminate this difference, $\mu_{\mathrm{PGc}}(t)$ is included into the model, hoping that this false rejection will not occur in the next iteration.

Second version

Here, 31 alternative models to (the initial) model 1 are generated by omitting all possible combinations of specific growth rates only. In contrast to the first version, the yield coefficients remain in the different model candidates. An overview of the models and the omitted specific growth rates can be found in Table 7.5.

The same procedure as in the first version is applied to the models here: they are tested for model deficiencies, leading to proposals of other models within this model family

Table 7.5: Model candidates of $\mathcal{S}_{\mathrm{UM3S}}$, second version: allocation between model number and omitted growth rate

	Model number															
Omitted	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mu_{\mathrm{PGc}}(t)$		×		×		×		×		×		×		×		×
$\mu_{\mathrm{PPh}}(t)$			×	×			×	×			×	×			×	×
$\mu_{\mathrm{XGc}}(t)$					×	×	×	×					×	×	×	×
$\mu_{\mathrm{XPh}}(t)$									×	×	×	×	×	×	×	×

	Model number															
Omitted	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$\mu_{\mathrm{PGc}}(t)$		×		×		×		×		×		×		×		×
$\mu_{\mathrm{PPh}}(t)$			×	×			×	×			×	×			×	×
$\mu_{\mathrm{XGc}}(t)$					×	×	×	×					×	×	×	×
$\mu_{\mathrm{XPh}}(t)$									×	×	×	×	×	×	×	×
$\mu_{\mathrm{XAm}}(t)$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

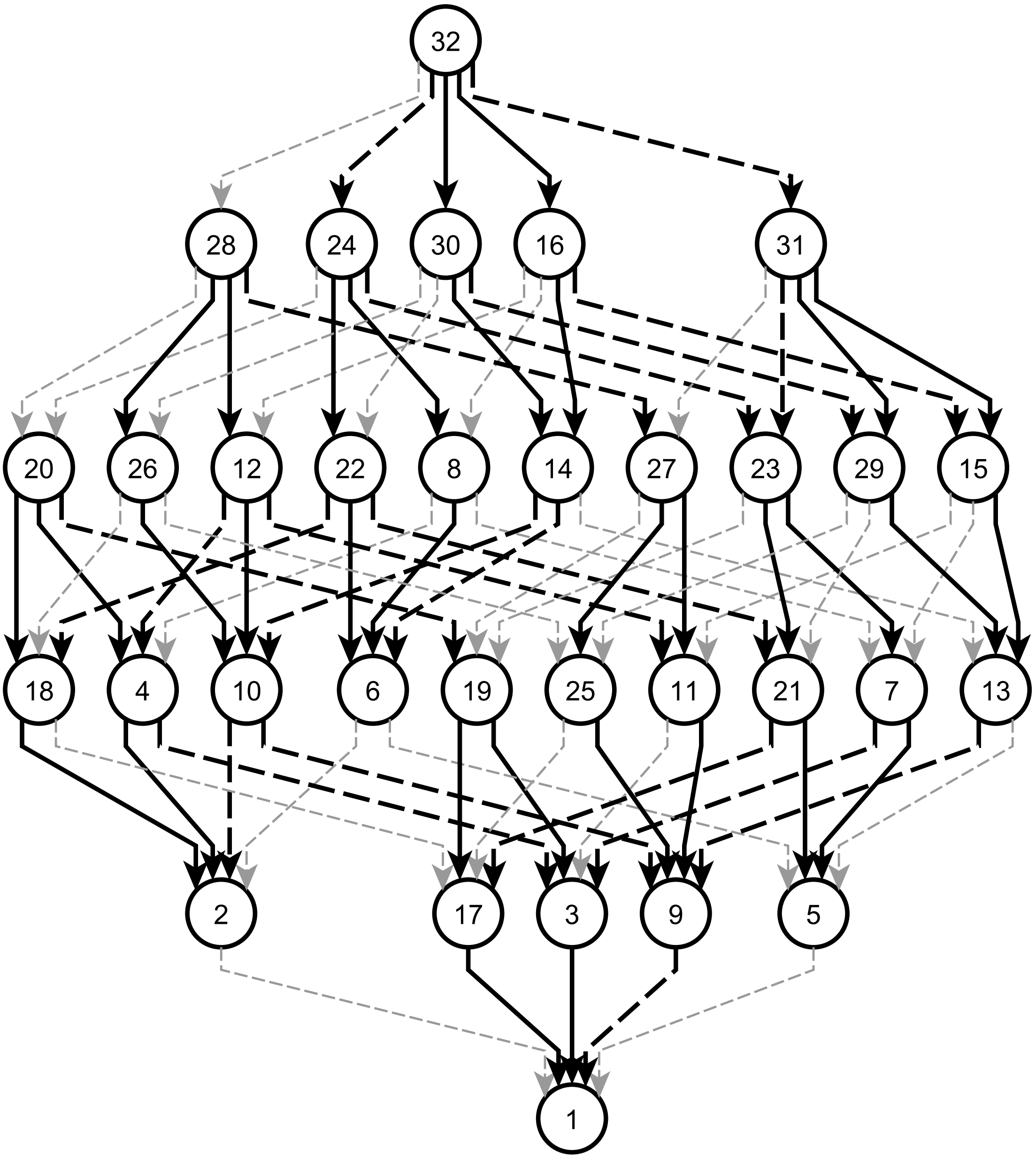


Figure 7.5: Deficiency detection with $\mathcal{S}_{\text{UM3S}}$, second version: 53 possible improvements are found, 27 deficiencies are not detected. For more information, see Figure 7.4.

that can be seen as an improvement. The result is shown in Figure 7.5, Table 7.6 gives a more detailed overview of the individual μ_i and how often a lack thereof can be detected. Similar to the first version, many possible improvements are not detected, leading to dead ends other than model 1. However, when the necessary phenomena to detect a deficiency have been found in the measurements, this deficiency is always detected. Additionally, 21 other deficiencies are detected although the corresponding

Table 7.6: Deficiency detection with $\mathcal{S}_{\mathrm{UM3S}}$, second version: Specific growth rates $\mu_i(t)$ and how often a lack thereof is detected.

In the measurements, the necessary phenomenon is			
detected		not detected	
		$\mu_{\mathrm{XPh}}(t)$	6/16
$\mu_{\mathrm{XAm}}(t)$	16/16	$\mu_{\mathrm{XGc}}(t)$	5/16
$\mu_{\mathrm{PPh}}(t)$	16/16	$\mu_{\mathrm{PGc}}(t)$	10/16
	32/32		21/48

phenomena are not found by the measurements. Here, as mentioned above, limiting dependencies are wrongly rejected by the simulations and the deficiency algorithm proposes to include these dependencies into the model. In comparison to the first version, the found improvements are not the same.

7.3 Summary

As can be seen, the presented approach to detect model deficiencies automatically seems promising. Differences between the measurements and the model simulations are listed and ranked in a way so that plausible improvements can be proposed automatically. It is understood that the quality of the phenomena detection determines the success of the deficiency detection. If a specific correlation cannot be found by the phenomena detection, it will be hard to detect a lack of this correlation in the model. Nonetheless, the case studies used to test the presented approach show that deficiencies are found when the phenomena are at hand. Moreover, in some cases, models can be improved although a correlation has not been found in the measurements.

However, it has to be pointed out that, at this stage, the model improvement has not yet been completely automated. The deficiencies are detected but an automatic generation of an improved model is still to be done.

Chapter 8

Application to Experimental Data

The presented methodology to detect biological phenomena is now applied to data from fed-batch experiments of three different organisms—*Paenibacillus polymyxa*, *Streptomyces tendae*, and *Streptomyces griseus*—which were all cultivated in a defined minimal medium. In all cases, the concentrations of the biomass, of the compartments DNA, RNA, and proteins, of the substrates ammonium, phosphate, and glucose, and of the corresponding product of interest are measured. For all experimental data under consideration, the results obtained by different smoothing and interpolation methods are presented to the user who selects the method that is considered most adequate. The parameter values given in Table 8.1 are used to detect the phenomena. In some experiments, $\Delta\dot{c}_{\mathrm{Tol}}$ for the corresponding product is manually modified. For *P. polymyxa*, models are built automatically using both simple and more advanced measurement situations, i.e., without or with the cell-intern measurements, respectively. The small-size structured models are then used for a model-discriminating trajectory planning. For the two *Streptomyces* strains, the focus is on developing medium-size structured models including cell-intern measurements automatically. The measurement situation without the cell-intern measurements is neglected. Unfortunately, new experiments with these strains cannot be conducted for reasons addressed below.

In each application, the best model is tested for deficiencies as presented in Chapter 7. However, since possible model improvements cannot be coded automatically yet, the detection of deficiencies will be stopped after one iteration.

Table 8.1: Parameter values used for the phenomena detection

Parameter	Value
Δt_{Intp}	$0.2\,\mathrm{h}$
Δc_{Tol}	$1/20 \cdot \sigma(c(t))$
$\Delta\dot{c}_{\mathrm{Tol}}$	$\min(1/12 \cdot \sigma(\dot{c}(t)),\, 5 \times 10^{-3})$
Δt_{Ep}	$2\,\mathrm{h}$
Δt_{Lm}	$8\,\mathrm{h}$
Δt_{Phen}	$4\,\mathrm{h}$

8.1 *Paenibacillus polymyxa*

The gram-positive, spore-forming bacterium *Paenibacillus polymyxa* has great biotechnological potential in different industrial processes (Lal and Tabacchioni, 2009). It produces a wide variety of secondary metabolites, including antibiotic compounds (Rosado and Seldin, 1993, Piuri et al., 1998, He et al., 2007). In this work, the production of macrolactins is considered—a group of antibiotics that possess a wide range of pharmacological activities, e.g., significant antiviral activities against the *Herpes simplex* virus or the human immunodeficiency virus (HIV) (Xue et al., 2008, Lu et al., 2008).

Seven different fed-batch experiments are chosen to obtain information about the biological phenomena. The complete list can be found in Appendix B, Table B.1. Based on the different scores, models with different complexity are developed.

8.1.1 Models Derived from a Simple Measurement Situation

At first, only the measurements of the biomass, the substrates, and the product are considered for the model development. In Table 8.2, the phenomena are shown that influence the basic unstructured model. According to their score Sc, the following changes are done as described in Section 6.2:

- The assumed limiting influences of ammonium, phosphate, and glucose on the product formation do not exist.
- An inhibiting effect of both ammonium and phosphate on the product formation is considered in alternative model structures.
- A limiting influence of phosphate on the growth is neglected. However, as a storage for phosphate is detected, it remains in the growth reaction and an additional state describing the dynamic behavior of the storage is included into the model. Alternatively, models are considered, as well, that do not consider the (A-type) storage for phosphate but a B-type storage for the two other substrates ammonium and glucose.
- A term describing the biomass degradation is included into the model.

As a result, 320 different model structures are proposed. The five most likely ones according to an evaluation with Eqs. (6.28a) and (6.28b) are considered to find suitable model candidates. The ‘limit’ terms in the reaction network are substituted for the Michaelis–Menten law, the ‘inhib’ terms are replaced with the Jerusalimski–Engamberdiev law. Testing all regulatory possibilities for the different storage synthesis and storage degradation rates (Eqs. (6.20)–(6.22)) in the considered model structures lead to 2320 different model candidates. Their parameters are automatically

Table 8.2: Phenomena changing the basic unstructured model for *P. polymyxa*

Phenomenon	Sc
The growth is limited by phosphate.	−0.69
The product formation is limited by ammonium.	−0.70
The product formation is limited by phosphate.	−0.66
The product formation is limited by glucose.	−0.67
The product formation is inhibited by ammonium.	0.30
The product formation is inhibited by phosphate.	0.64
Storage A for phosphate	0.69
Degradation of biomass	0.67

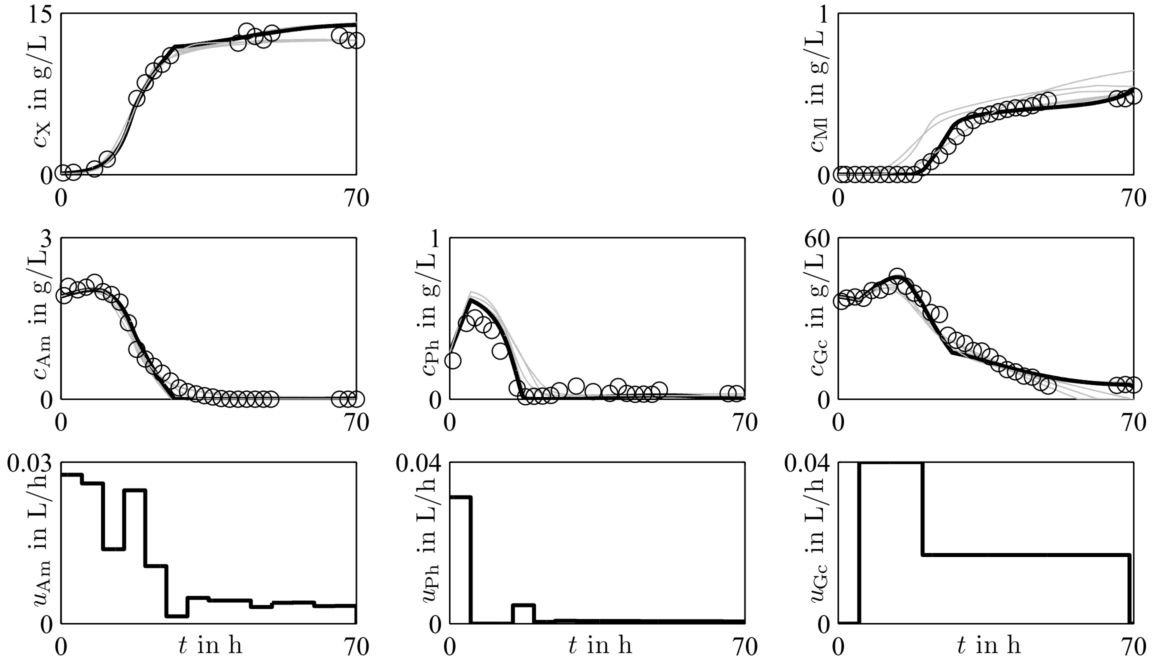
identified (Section 6.3) based on four fed-batch experiments. In addition to the parameters, the unknown and experiment-specific initial values for the unmeasured storages are identified for each experiment. The models are then ordered by their AIC_c value.

The simulations of the 13 best identified models are compared to the measurements of two experiments in Figure 8.1. The best identified model is indicated by the black solid line. Additional identification experiments are shown in Appendix B.3. As can be seen, most models describe the measurements equally well. However, some models show deficiencies, especially in the description of the dynamics of macrolactin.

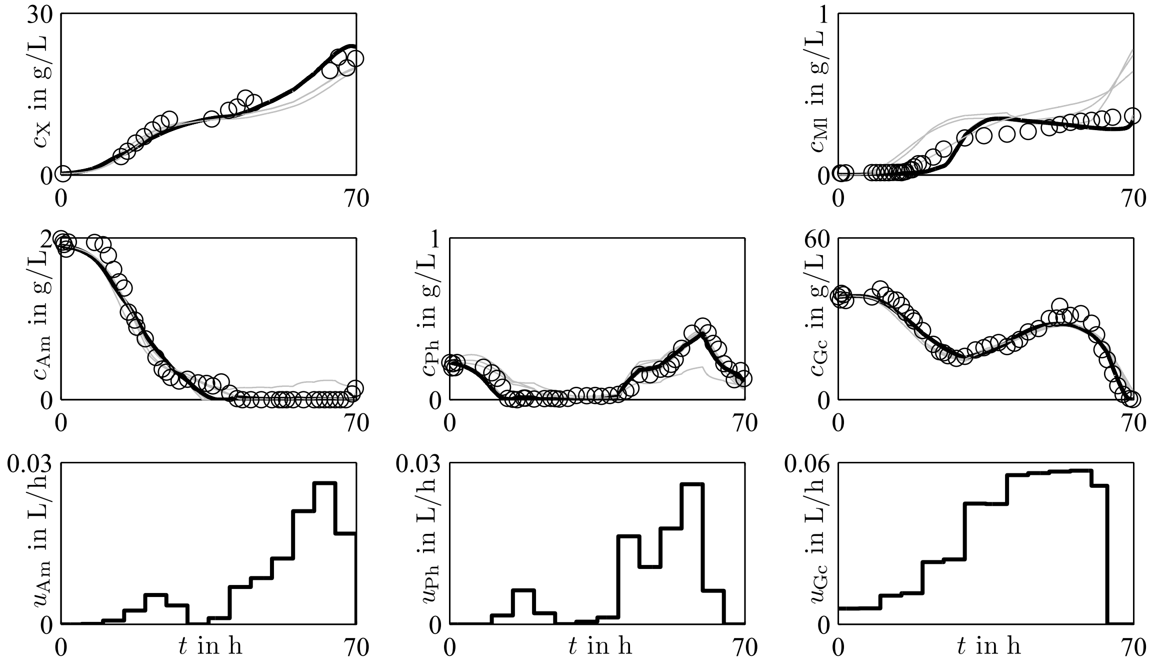
Then, three experiments, which were not used for the identification, are used for a validation, i.e., the models are tested for their ability to predict the measurements. However, it has to be pointed out that, similarly to the identification experiments, the initial values for the storages are unknown for the validation experiments as well. Before a validation takes place, the validation experiments are used to identify these initial values. In Figure 8.2, the comparison between the predictions and the actual measurements can be seen. For the third validation experiment, see Appendix B.4. The best identified and the best validated model are not the same. They are highlighted by the black solid and the black dashed line, respectively. In most cases, the predictions and the measurements match well. However, as it is the case in the identification, shortcomings in describing the product macrolactin can be observed. Macrolactin is overestimated at the end of the fermentation, especially in Figure 8.2(b). Furthermore, the glucose dynamics at the end of this experiment are not described well.

Model-discriminating experiment

A model-discriminating experiment is planned according to Section 6.4, using 13 model candidates. It is run for 100 hours and the sampling time is specified in advance. The flow rates for ammonium, phosphate, and glucose can be changed stepwise every ten

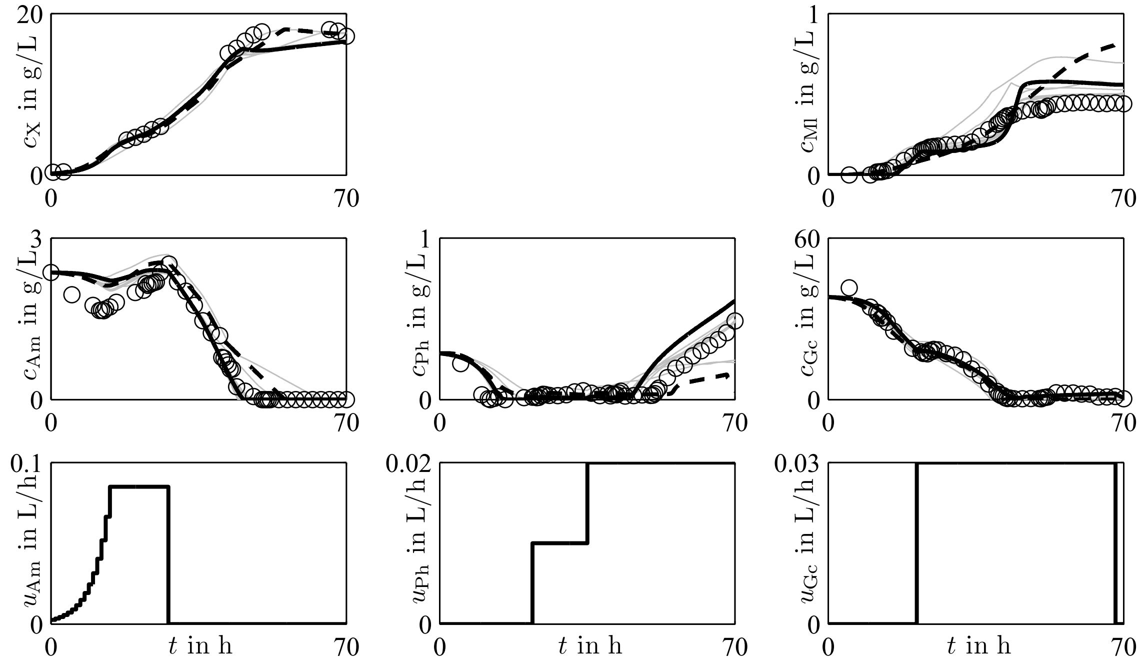


(a) Identified experiment PPdef11

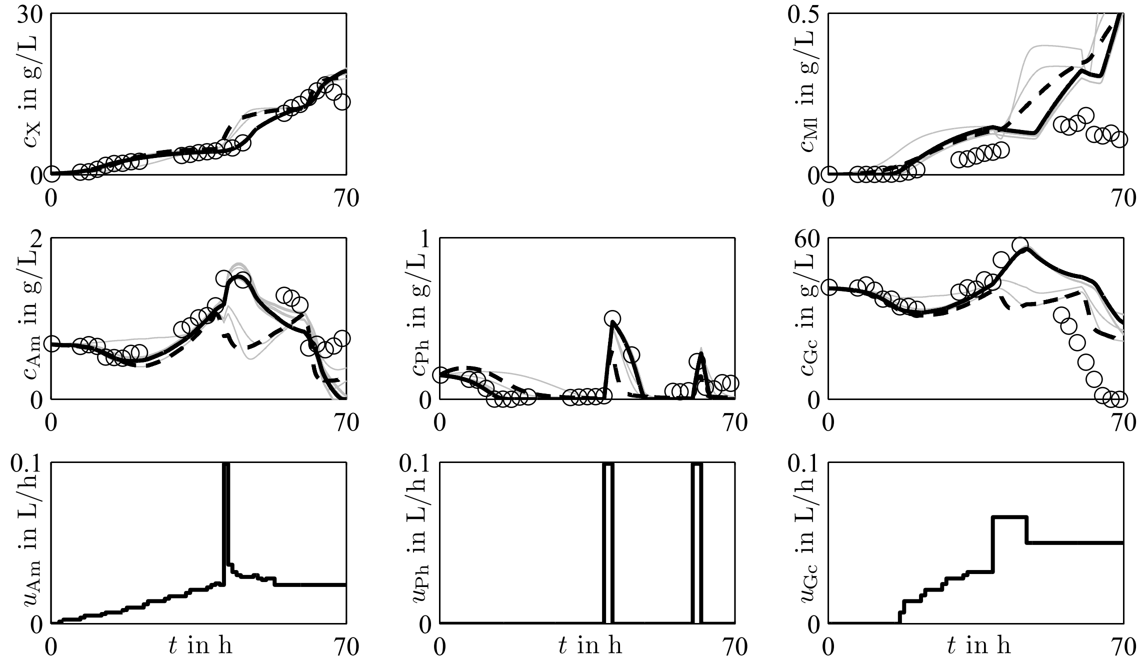


(b) Identified experiment PPdef12

Figure 8.1: Identified experiments for *P. polymyxa.* The simulations based on the 13 best identified model candidates are shown as solid lines, the black line displays the best identified model. Circles indicate the measurements.



(a) Validation experiment PPdef9



(b) Validation experiment PPdef17

Figure 8.2: Validation experiments for *P. polymyxa*. The dashed line indicates the best validated model.

hours. The initial values for biomass, the substrates, and macrolactin are given as well. Concerning the initial values for the (not measured) storages, one should refrain from specifying values that are shared by each model. Instead, model-specific initial values for these storages are specified. For this purpose, the (identified) initial values from the identified and validated experiments are taken and the mean value is calculated for each model that serves as the initial value for the model-discriminating experiment.

Then, the experiment is conducted as calculated by the optimization algorithm. Unfortunately, due to some technical problems during the fermentation, it could not be run exactly as planned. Figure 8.3 shows the experimental measurements and the simulations of the models considered, using the flow rates that were actually used in this experiment. As can be seen, the growth of the biomass and the consumption of glucose can be described equally well by any model chosen. Concerning ammonium, many models are able to mimic the dynamic behavior of the measurements, but an exact prediction is not achieved by any chosen model. The phosphate measurements show an unexpected behavior that can hardly be seen in the simulations. As for the macrolactin, most models overestimate its formation. However, better results might be achieved if other initial values for the storages were used.

Nevertheless, the best model (indicated by the black line; coninciding with the best validated model, for details see Appendix B.5) is able to predict the most important dynamics and can therefore be used as a solid basis of the modeling process toward

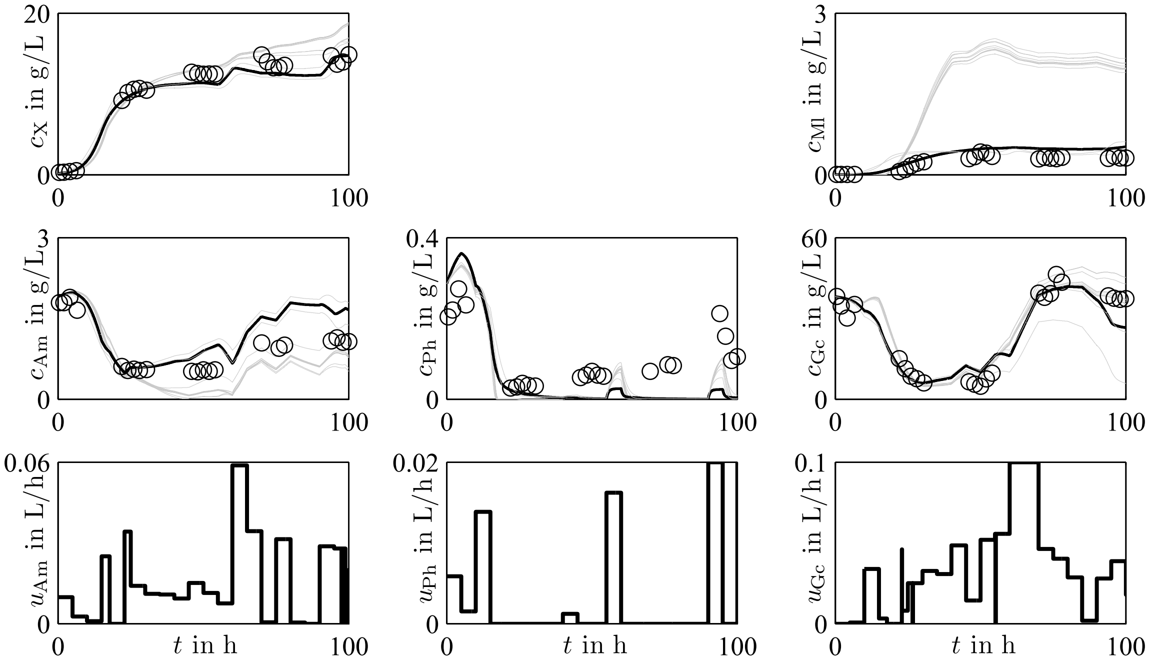


Figure 8.3: Results of the model-discriminating trajectory planning. The simulation of the model with the highest probability is indicated by the black line. The gray lines show the other model simulations. The circles show the measurements.

process models for control. Alternatively, the whole cycle consisting of phenomena detection, model structure proposals, and parameter identification is restarted with this additional experiment.

Automated detection of model deficiencies

Now, the best model is tested for deficiencies, the most important of which are:

- an inhibiting effect of phosphate on the product formation is not found by the simulations to the same extent as by the measurements;
- the necessary phosphate storage for growth is not found at all in the simulations.

These detected deficiencies cannot be used to suggest model improvements. The best model already considers both the phosphate-inhibited product formation and the phosphate storage. The result of the parameter identification might explain why these deficiencies are nevertheless detected. Taking, for instance, the phosphate-inhibited formation of macrolaction which is considered by

$$\mu_{\mathrm{MlPh}}(t) = \frac{K_{\mathrm{MlPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{MlPh}}} \quad .$$

The estimated parameter value is $K_{\mathrm{MlPh}} = 42.001\,\mathrm{g/L}$. With phosphate concentrations $0 \leq c_{\mathrm{Ph}}(t) \leq 0.5\,\mathrm{g/L}$, this inhibiting effect on the product formation is lost. Macrolactin starts growing immediately, as can be seen in Figure 8.4, making it impossible to detect the corresponding phenomenon in the simulations.

Thus, although a model improvement, i.e., a change in the model structure, cannot be suggested on the basis of the detected deficiencies, they might indicate where the parameter identification estimates parameter values that are not reasonable.

8.1.2 Models Considering Cell-Intern Measurements

Now, the cell-intern measurements DNA, RNA, and proteins are considered as well and more detailed structured models are proposed. The phenomena influencing the basic structured model are shown in Table 8.3. The basic model structure is modified as follows:

- The assumed limiting dependency on phosphate in the protein formation is omitted in all model structure proposals. Additionally, alternative structures are developed that neglect the limiting influences of ammonium on the DNA formation and of glucose on the RNA formation as well.
- Model structures are proposed that include precursors to the proteins or to RNA (and thus DNA) or both.
- Alternative model structures are developed that consider degradation reactions for DNA or the product or both.

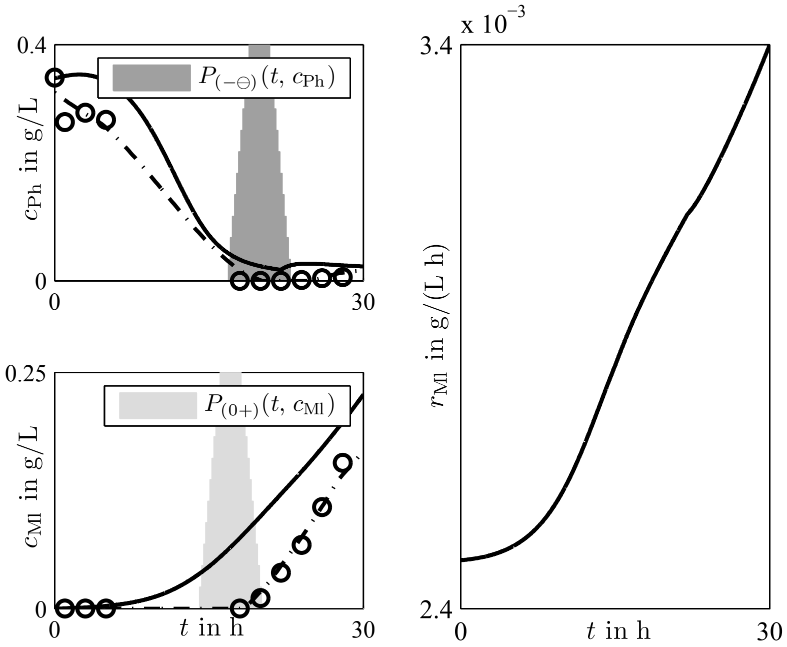


Figure 8.4: Deficiency detection with model of *P. polymyxa*. The measurement-inherent phenomenon *product formation inhibited by phosphate* cannot be found by the simulations. Shown are the concentrations of phosphate and macrolactin as well as the reaction rate r_{Ml}. Measurements are indicated by circles, their interpolations by the dash-dot lines, and the solid lines show the model simulations. The shown landmark probabilities (shades of gray) belong to the measurement interpolations.

The three most likely model structures are used to find suitable model candidates. As in the case above, the ‘limit’ terms are replaced with the Michaelis–Menten law,

Table 8.3: Phenomena changing the basic medium-size structured model for *P. polymyxa*

Phenomenon	Sc
The DNA formation is limited by ammonium.	−0.46
The protein formation is limited by phosphate.	−0.73
The RNA formation is limited by glucose.	−0.50
Intermediate compartment between ammonium and proteins	0.22
Intermediate compartment between phosphate and RNA	0.58
Intermediate compartment between phosphate and proteins	0.58
Degradation of DNA	0.52
Degradation of the product	0.27

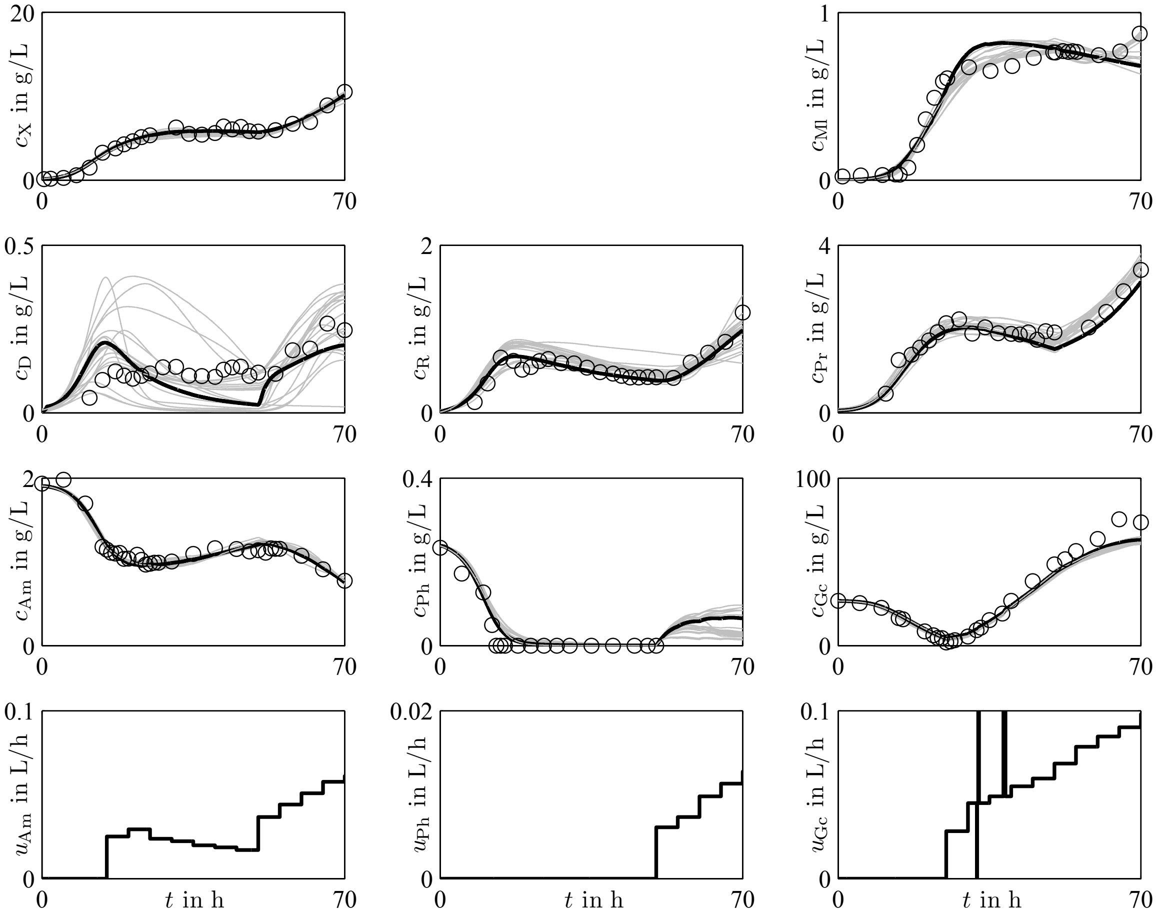


Figure 8.5: Identified experiment PPdef13 for *P. polymyxa* with cell-intern measurements. The black line shows the simulations of the best identified model.

the 'inhib' terms are substituted for the Jerusalimski–Engamberdiev law. Regarding DNA, RNA, proteins, and the remaining biomass Xr as possible options for C_k in the building-up reaction r_{Xr} (see Eq. (6.4)) and testing both Eqs. (6.6) and (6.27) to calculate the compartment degradation rates r_{dCi} lead to 96 model candidates.

Two fed-batch experiments are used for the estimation of the model parameters and of the initial values of the cell compartments. In Figure 8.5, the measurements of the identified experiment PPdef13 are compared with the best models. The second identified experiment can be found in Appendix B.6. The best model is highlighted by the black solid line. Here, most of the measured variables can be described by the models. However, the models are not able to mimic the dynamics of DNA.

Two experiments are used for a validation, where the initial values of the cell-intern compartments are identified beforehand. In Figure 8.6, the predictions and the measurements are compared to each other. The second validation experiment is shown in Appendix B.7. Here, the best identified and the best validated model are different, they are indicated by the black solid and the black dashed line, respectively (for the

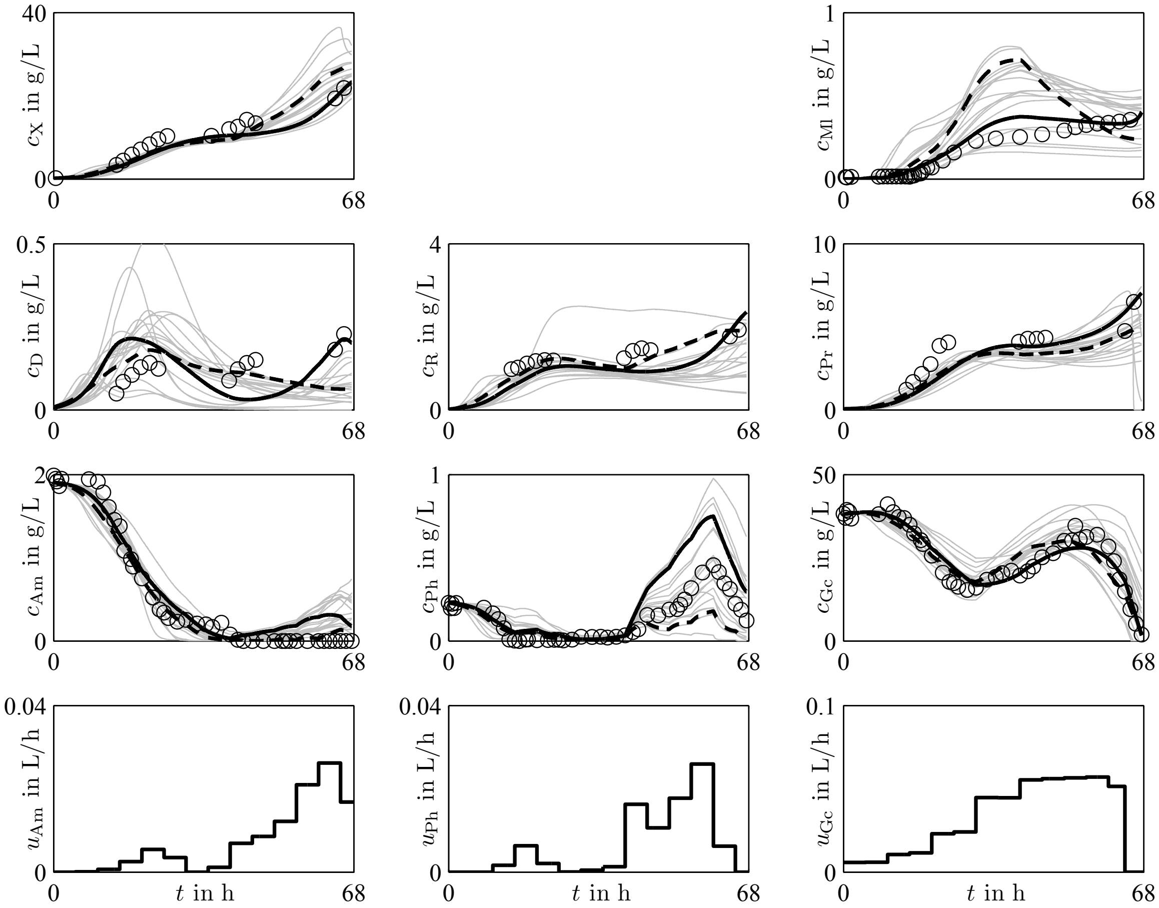


Figure 8.6: Validation experiment PPdef12 for *P. polymyxa* with cell-intern measurements. The dashed line shows the best validated model.

best validated model, see Appendix B.8). Most of the models predict the dynamics of biomass and the substrates correctly. Regarding the product macrolactin, some models are not able to describe its behavior properly. The aforementioned shortcoming in describing the DNA correctly can also be seen in the validation experiment. Furthermore, only few models are able to correctly predict the dynamics of RNA and proteins.

Automated detection of model deficiencies

Based on the best validated model, important proposals for model improvements are:

- An inhibiting dependency of ammonium on the RNA formation might be considered.
- A limiting effect of RNA on the protein formation might be missing.

Two additional model candidates are built manually with each model considering one of the listed improvements. Unfortunately, the changes conducted do not lead to models that describe the measurements better.

8.2 *Streptomyces tendae*

Streptomyces strains are gram-positive bacteria that form a mycelium. They are able to produce a large variety of secondary metabolites (Roubos, 2002) and provide more than half of medically important antimicrobial and antitumor agents (Liu et al., 2013). As an example, the strain *Streptomyces tendae* is considered which produces nikkomycin.

Here, six different fed-batch fermentations are chosen to obtain information about the biological phenomena. The most important phenomena detected and their score values can be seen in Table 8.4. The complete list is given in Appendix C, Table C.3. The influence of the phenomena on the proposal of possible model candidates is as follows:

- In the product formation reaction, the limiting dependencies on ammonium and phosphate are neglected and hence, all corresponding components in the basic model structure will be deleted.

Table 8.4: Important phenomena for the medium-size structured model for *S. tendae*

Phenomenon	Sc
The product formation is limited by ammonium.	−0.75
The product formation is limited by phosphate.	−0.78
The product formation is inhibited by ammonium.	0.56
The product formation is inhibited by phosphate.	0.66
Storage A for ammonium	0.50
Storage A for phosphate	0.50
The DNA formation is limited by ammonium.	−0.50
The RNA formation is limited by ammonium.	−0.50
The protein formation is limited by ammonium.	−0.50
The DNA formation is limited by phosphate.	−0.65
The RNA formation is limited by phosphate.	−0.24
The protein formation is limited by phosphate.	−0.38
Intermediate compartment between phosphate and DNA	0.58
Intermediate compartment between phosphate and RNA	0.25
Intermediate compartment between phosphate and proteins	0.67
Degradation of DNA	0.50
Degradation of the product	0.50

- Alternative model structures are proposed that include an inhibiting dependency on either ammonium or phosphate or both in the product formation.
- Alternative model structures will neglect a direct limiting dependency of the compartment building-up reactions on ammonium or phosphate. However, alternative model structure are proposed that consider a precursor (intermediate compartment) for both protein and DNA/RNA which is built up on phosphate.
- Simple relationships between DNA, RNA, and proteins like limiting or inhibiting dependencies cannot be found and will therefore not be considered in any model structure, i.e., Eq. (6.4) will not be extended by limiting or inhibiting kinetic expressions with respect to these compartments.
- A relationship between the product and DNA, RNA, or the proteins cannot be found either, i.e., r_{P} is independent of these compartments.

The six most likely model structures are considered to find suitable model candidates. Again, the 'limit' terms in the reaction network are substituted for the Michaelis–Menten law and the 'inhib' term is replaced with the Jerusalimski–Engamberdiev law. DNA, RNA, proteins and the remaining biomass Xr are regarded as possible options for C_k in the building-up reaction r_{Xr} (6.4). To calculate the compartment degradation rates $r_{\mathrm{dC}i}$, both Eqs. (6.6) and (6.27) are tested. This leads to 152 model candidates. The parameters and the experiment-related initial values of the compartments are identified based on four experimental runs. Finally, the models are ordered by their AIC_c values.

The simulations of the 10 best identified models are compared to real data in Figure 8.7. For the other identified experiments, see Appendix C.5. As can be seen, many of the simulations barely differ from each other and can explain most of the measurements equally well. However, the models are not able to mimic the phosphate consumption correctly. This becomes obvious when phosphate is depleted in the experiments and it starts being fed. In the simulations, the phosphate concentration increases whereas it cannot be measured in the experiments.

Subsequently, two experimental runs—not used for the identification—are used for validation, where the initial values of the cell-intern states are estimated. In Figure 8.8, the comparison between the predictions and the actual measurements can be seen. The second experiment is given in Appendix C.6. It is obvious that some dynamic aspects in the reaction network are yet to be considered by the model candidates. In addition to the aformentioned shortcomings regarding phosphate, DNA and nikkomycin measurements show some characteristics, as well, that cannot be described by the simulations. However, the models are able to explain the dynamic of the other measurements well.

Unfortunately, further experiments to improve the model quality cannot be conducted as the strain used, *S. tendae* Tü 901/8c, had over the years lost its capability to produce nikkomycin. Therefore, new data could not be compared to the old data used here.

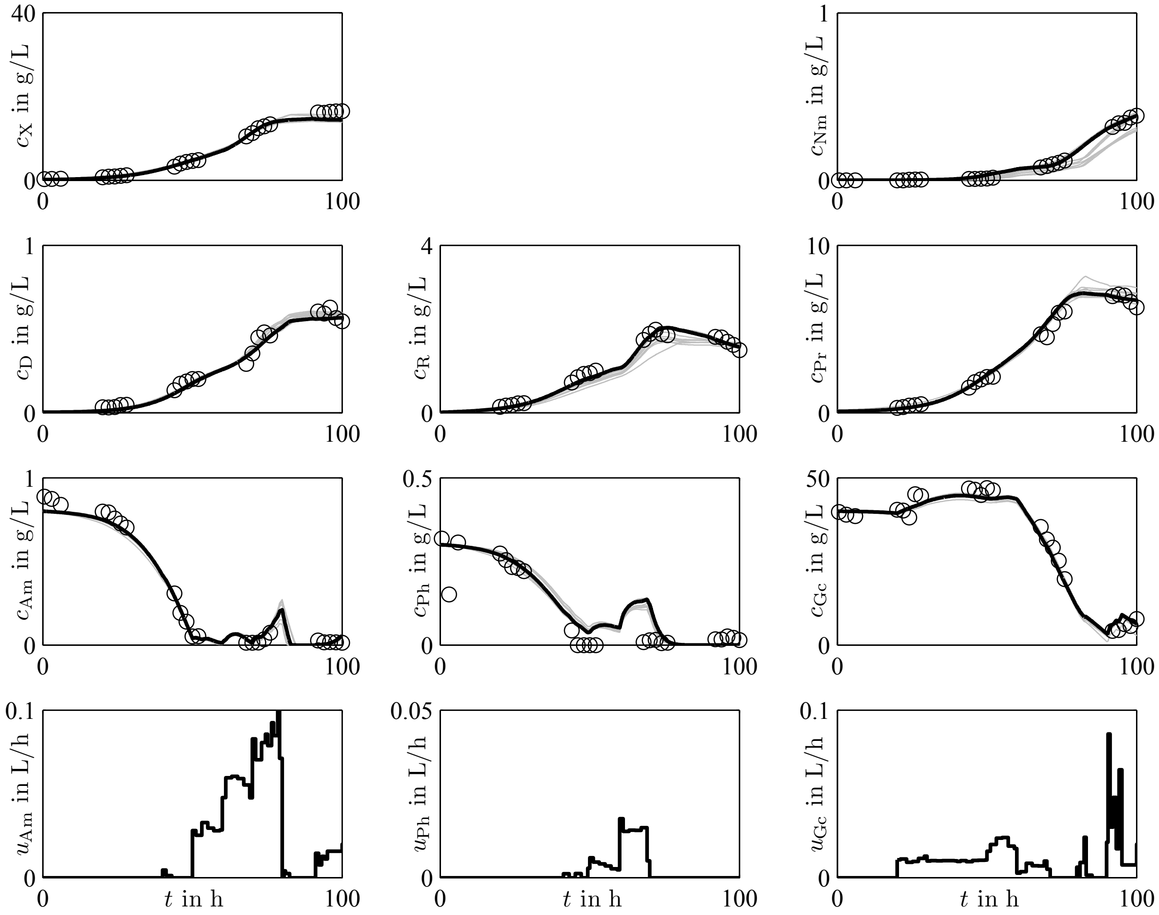


Figure 8.7: Identified experiment STdef2 for *S. tendae*. The simulations based on the 10 best identified model candidates are shown as solid lines, circles indicate the measurements. The feeding rates are the result of an on-line trajectory planning.

Automated detection of model deficiencies

The best validated model is tested for model deficiencies. The two most important differences between the measurements and simulations are listed below.

- Based on the measurements, the phenomenon *DNA formation limited by ammonium* is rejected with $\text{Sc} = -0.50$, whereas the simulations accept it with $\text{Sc} = 0.50$.
- The same applies to the phenomenon *protein formation limited by ammonium.*

To account for these deficiencies, the simplest approach is to eliminate the limiting effect of ammonium in both $r_{\mathrm{D}}(t)$ and $r_{\mathrm{Pr}}(t)$. However, since both phenomena are rejected by the measurements, model structures already exist that neglect those influences and do not perform better than the best model. Moreover, since the additionally added compartments D* and Pr* are not built up on ammonium—the measurement-

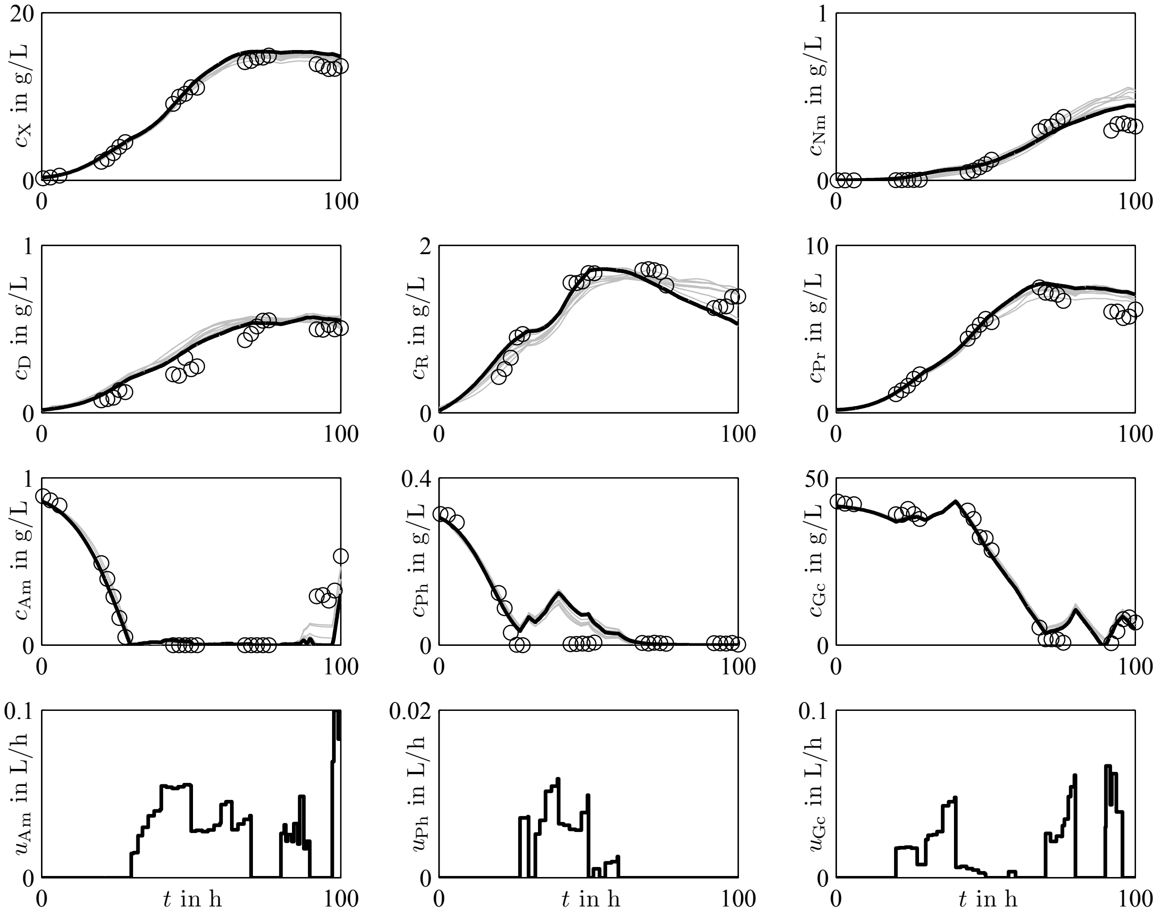


Figure 8.8: Validation experiment STdef3 for *S. tendae* using as well the 10 best model candidates

inherent phenomena do not show such relationships—DNA and the proteins are not built up on ammonium at all. Taking the biological knowledge (see Section 2.2) into account, this does not make any sense. Therefore, the simplest approach is not used to improve the model. Instead, other modifications are tried.

In Figure 8.9(a), the relevant part of the biological network of the best validated model can be seen. Two model candidates (Figure 8.9(b) and 8.9(c)) try to account for the false detection of the ammonium-limited DNA formation. Here, different combinations of Eqs. (8.1)–(8.3) are manually changed. For example, model STBV$_b$ neglects the limiting influence of ammonium in the DNA building-up reaction $r_D(t)$ and adds ammonium in the formation of D*. Additionally to these changes, model STBV$_c$ eliminates a direct mass flow from ammonium to DNA by modifying Eq. (8.1). Correspondingly, two models (Figure 8.9(d) and 8.9(e)) try to compensate for the false detection of the ammonium-limited protein formation. At last, two models are manually generated that try to account for both shortcomings and are a combination of the aforementioned models.

$$\mathrm{Am} + \mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{D}}} \mathrm{D} \tag{8.1}$$

$$r_{\mathrm{D}} = r_{\mathrm{D}}(c_{\mathrm{Am}}(t)) \tag{8.2}$$

$$\mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.3}$$

$$\mathrm{Am} + \mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{Pr}}} \mathrm{Pr} \tag{8.4}$$

$$r_{\mathrm{Pr}} = r_{\mathrm{Pr}}(c_{\mathrm{Am}}(t)) \tag{8.5}$$

$$\mathrm{Ph} \rightarrow \mathrm{Pr}^{*} \rightarrow \mathrm{Pr} \tag{8.6}$$

(a) Best validated model STBV

$$r_{\mathrm{D}} \neq r_{\mathrm{D}}(c_{\mathrm{Am}}(t)) \tag{8.2b}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.3b}$$

(b) Modifications for model STBV$_b$

$$\mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{D}}} \mathrm{D} \tag{8.1c}$$

$$r_{\mathrm{D}} \neq r_{\mathrm{D}}(c_{\mathrm{Am}}(t)) \tag{8.2c}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.3c}$$

(c) Modifications for model STBV$_c$

$$r_{\mathrm{Pr}} \neq r_{\mathrm{Pr}}(c_{\mathrm{Am}}(t)) \tag{8.5d}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.6d}$$

(d) Modifications for model STBV$_d$

$$\mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{Pr}}} \mathrm{Pr} \tag{8.4e}$$

$$r_{\mathrm{Pr}} \neq r_{\mathrm{Pr}}(c_{\mathrm{Am}}(t)) \tag{8.5e}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.6e}$$

(e) Modifications for model STBV$_e$

$$r_{\mathrm{D}} \neq r_{\mathrm{D}}(c_{\mathrm{Am}}(t)) \tag{8.2f}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.3f}$$

$$r_{\mathrm{Pr}} \neq r_{\mathrm{Pr}}(c_{\mathrm{Am}}(t)) \tag{8.5f}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.6f}$$

(f) Modifications for model STBV$_f$

$$\mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{D}}} \mathrm{D} \tag{8.1g}$$

$$r_{\mathrm{D}} \neq r_{\mathrm{D}}(c_{\mathrm{Am}}(t)) \tag{8.2g}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.3g}$$

$$\mathrm{Ph} + \mathrm{Gc} \xrightarrow{r_{\mathrm{Pr}}} \mathrm{Pr} \tag{8.4g}$$

$$r_{\mathrm{Pr}} \neq r_{\mathrm{Pr}}(c_{\mathrm{Am}}(t)) \tag{8.5g}$$

$$\mathrm{Am} + \mathrm{Ph} \rightarrow \mathrm{D}^{*} \rightarrow \mathrm{D} \tag{8.6g}$$

(g) Modifications for model STBV$_g$

Figure 8.9: Modifications to the best validated model of *S. tendae* to account for detected deficiencies. (a) Relevant excerpt from the biological network of the best validated model. (b)–(g) Different modifications that lead to six alternative model candidates.

These six model candidates are subjected to a parameter identification step and a validation step, using the same identification and validation experiments as mentioned above. Unfortunately, the models do not describe the measurements better.

8.3 *Streptomyces griseus*

As a second example for *Streptomyces* strains, the bacterium *Streptomyces griseus* is considered that produces streptomycin—the first antibiotic used against tuberculosis.

The biological phenomena are detected based on seven different fed-batch fermentations. The phenomena that influenced the model building and their score values are shown in Table 8.5. The complete list is given in Appendix D, Table D.1. The following changes of the basic structured model are carried out:

- For the product formation, alternative model structures are proposed that neglect the limiting dependency on phosphate.
- Alternative structures are proposed that include an inhibiting dependency on phosphate.
- The assumed limiting influences of ammonium and phosphate on the protein formation are neglected in alternative model structures.
- Precursors to the proteins are considered in alternative model structures. Additionally and despite the insufficient score ($\mathrm{Sc} < 0.2$), a precursor to RNA (and thus DNA) is introduced in alternative model structures as well.
- DNA and product degradation reactions are introduced.

Table 8.5: Phenomena changing the basic medium-size structured model for *S. griseus*

Phenomenon	Sc
The product formation is limited by phosphate.	−0.20
The product formation is inhibited by phosphate.	0.24
The protein formation is limited by ammonium.	−0.47
The protein formation is limited by phosphate.	−0.46
Intermediate compartment between phosphate and RNA	0.15
Intermediate compartment between phosphate and proteins	0.52
Degradation of DNA	0.64
Degradation of the product	0.66

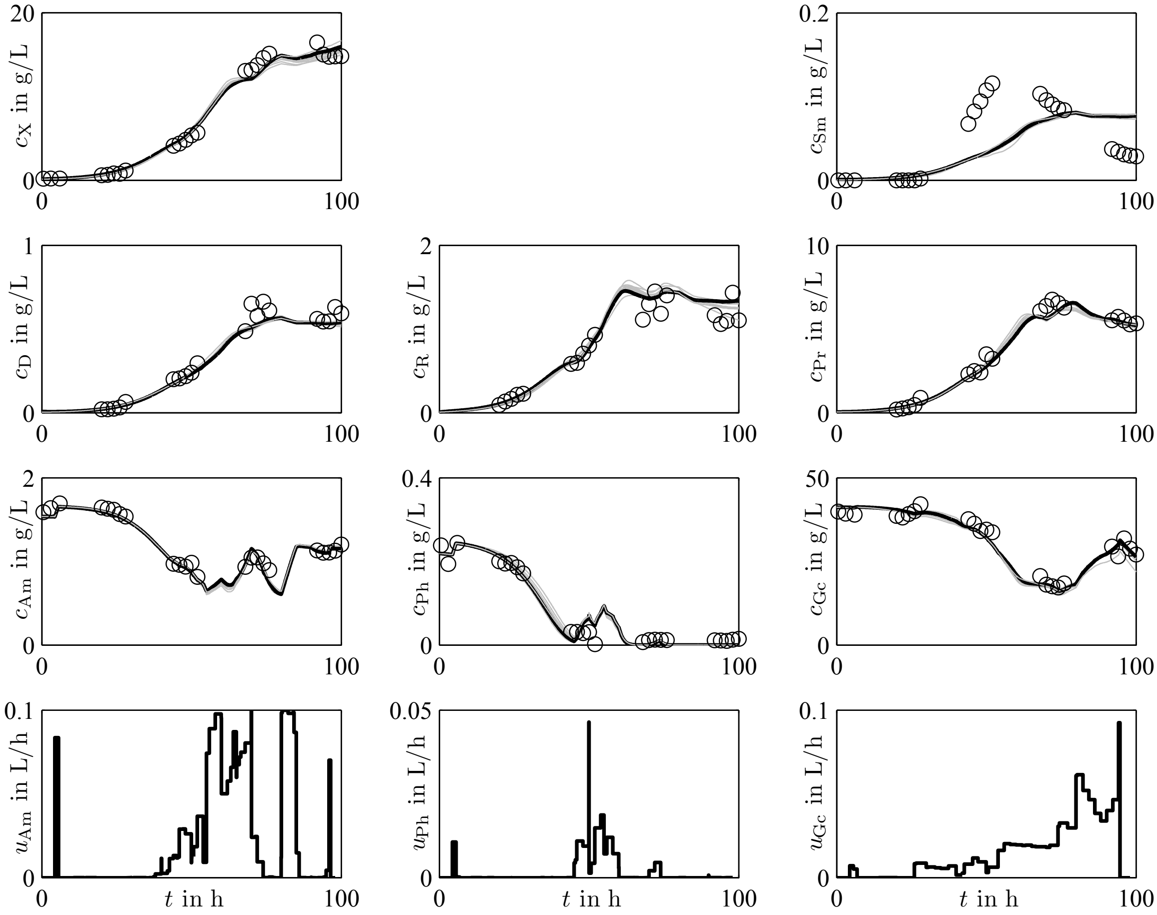


Figure 8.10: Identified experiment SGdef32 for *S. griseus* with cell-intern measurements

Five model structures are taken to find suitable model candidates. As in the other applications above, the ‘limit’ terms in the reaction network are substituted for the Michaelis–Menten law, the ‘inhib’ term is replaced with the Jerusalimski–Engamberdiev law. Possible options for C_k in the building-up reaction r_{Xr} (6.4) are DNA, RNA, proteins and the remaining biomass Xr. Eqs. (6.6) and (6.27) are both tested to calculate the compartment degradation rates r_{dCi}. In summary, 120 model candidates are tested. The parameters and the initial values of the compartments for each experiment are identified using four fed-batch experiments. Afterwards, the models are ordered according to their AIC_c values.

The simulations of the best identified models are compared to real data in Figure 8.10. The other identified experiments are shown in Appendix D.3. As can be seen, the identified models describe the dynamic behavior of most of the measured variables well. However, the models are not able to mimic the phosphate dynamics correctly. As it is the case with the identified models for *S. tendae*, the consumption of phosphate cannot be described after phosphate has depleted and is started being fed again. Furthermore, RNA shows some dynamics that cannot be imitated by the model candidates and

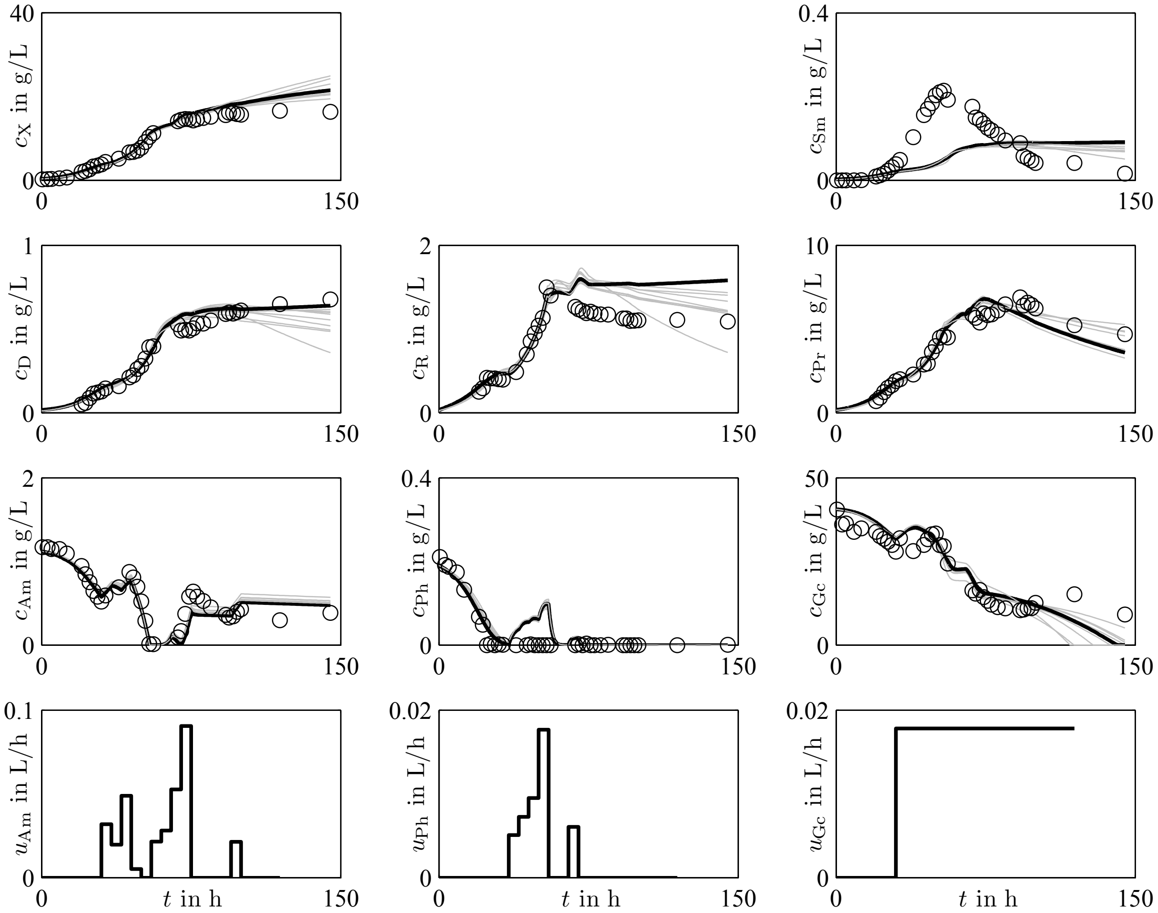


Figure 8.11: Validation experiment SGdef25 for *S. griseus* using the best model candidates

the dynamics of the product streptomycin cannot be described well by any model candidate.

Three experimental runs are used for validation, where the initial values of the cell-intern states are estimated. In Figure 8.11, the comparison between the predictions and actual measurements can be seen. Additional validation experiments can be found in Appendix D.4. The shortcomings that have already been noted in the identified experiments are also visible in the validation experiments. However, most of the measurements can be satisfactorily predicted.

To improve the quality of the model(s) further experiments need to be conducted. Unfortunately, this cannot be done because the strain of *S. griseus* used lost its capability to produce streptomycin.

Automated detection of model deficiencies

The best validated model is tested for model deficiencies. The most important ones are the following:

- the limiting dependency of ammonium on the DNA formation cannot be found in the simulations;
- an inhibiting effect of phosphate on the streptomycin production is not found by the simulations to the same extent as by the measurements.

The first deficiency cannot improve the model since a limiting dependency

$$\mu_{\mathrm{DAm}}(t) = \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{DAm}}}$$

has already been implemented into the structure. However, it indicates that an unreasonable parameter value was estimated. In fact, with $K_{\mathrm{DAm}} = 1 \times 10^{-4}\,\mathrm{g/L}$, the limiting effect is lost.

The second deficiency listed here is used to modify the best validated model manually. An inhibiting effect of phosphate on the product formation has not yet been part of the best validated model and is, thus, included. After the parameter identification and validation steps, this new model shows almost equivalent results, but the AIC_c value cannot be improved.

8.4 Summary

The results achieved with the experimental data of the different strains used above are ambivalent. The proposed models show shortcomings in describing several measurements correctly. The predictions of the validation experiments specifically show that these models are still in need of improvement. Especially the description of the phosphate consumption needs to be revisited since the same erratic behavior can be seen in the simulations of two different strains. However, it has to be emphasized again that all these models are automatically derived. Despite some differences from the measurements, these automatically built models significantly speed up the modeling process and can be used as a solid basis of the modeling process toward process models for control.

Regarding the detection of model deficiencies, the presented approach is not able to propose changes that will improve the best models—despite the visible differences between measurements and simulations. In fact, many improvements are proposed that have already been considered in the model structure, i.e., the corresponding model part already exists. However, this may indicate that unreasonable values are assigned to the corresponding parameters and the parameter identification should be revisited.

Chapter 9

Conclusion

This work showed how the tedious task of developing process models of different complexities can be facilitated by algorithms that automatically discover biological phenomena and propose, from this basis, model structures or detect model deficiencies automatically. For this purpose, necessary information needed to be extracted from the measurements. At first, two smoothing techniques were introduced that can lessen the effect of measurement noise. Then, to get a time-continuous format of the measurements, alternative interpolation methods to conventional cubic splines were shown that resulted in less wiggly data and resembled lines manually drawn by a human expert. Afterwards, to account for the effect of feeding and sampling on the dynamics of the measurements, an approach was presented that compensated the data for feeding and sampling, thus making it possible to consider data from fed-batch experiments for modeling. To reveal the qualitative behavior of the measurements automatically, the proposal of Cheung and Stephanopoulos (1990) was fundamentally extended. Probability values for the different episodes were introduced that consider the effect of measurement noise and interpolation errors. Lastly, measurements were tested for changes in their qualitative behavior and probability values were allocated to the different transitions.

A library of rules to automatically (dis-)prove biological phenomena was established. The approach by King et al. (2002) was extended by phenomena considering cell-intern measurements. Moreover, probabilities were allocated to each phenomenon and an approach was shown for how the results of different experiments could be merged. Then, it was demonstrated how model structures of different complexities could be automatically proposed and modified, based on biological knowledge and the phenomena detected. Since many model candidates were proposed, Akaike's Information Criterion was applied to select the best model. Additionally, it was shown how an experiment could be planned to discriminate between several model candidates. Furthermore, an approach was presented that might help to find model deficiencies by comparing measurement-inherent phenomena with phenomena of the simulations of a specific model.

In different case studies, it could be shown that phenomena detection is sensitive towards measurement noise, measuring time, interpolation, etc. The identification and validation results for the strains used showed that the models developed still needed to be improved. However, in many cases they were able to describe the dynamics satisfactorily and could even be used for a trajectory planning. Thus, the approach to propose models automatically could speed up the modeling process significantly. Regarding the automated detection of model deficiencies, it became clear that the approach still needed to be improved. Even though the used case studies delivered promising results, the application to experimental data showed otherwise. Here, actual improvements could not be proposed although differences between the measurements and the simulations were obvious.

Outlook

As can be seen, the goal to automatically develop good process models has successfully been accomplished. However, this work only represents a small step and there is still room for improvement. Some possibilities for improvement are mentioned, but the list is not exhaustive. Initially, the algorithm presented in this work was only applied to organisms cultivated in chemically defined minimal media and needs to be adapted to organisms cultivated in complex media. Then, the measured variables that have been considered for phenomena detection are the biomass, substrates, cell-intern components, and products. Usually, only discrete measurements exist for these substances. Continuous measurements like oxygen or carbon dioxide in the exhaust gas can also be integrated and might even show behavior that cannot be found in the aforementioned substances. Regarding the model discrimination, it is understood that a trajectory planned off-line might not be sufficient to distinguish between several models where some components cannot be measured and initial values are unknown. Instead, more sophisticated methods should be applied. For example, Schenkendorf and Mangold (2013) present an approach that combines model discrimination and state estimation, thus planning the trajectory on-line. However, their method has to be modified since—in contrast to the authors' assumption—not all state variables in the models used can be measured, and the measurements are usually not available immediately. With regards to the medium-size structured models, the proposed structure might be revisited. Especially when it comes to the integration of intermediate compartments, alternative stoichiometric and regulatory concepts could lead to better results. Choosing other laws than those proposed by Monod or Jerusalimski and Engamberdiev to replace the 'limit' and 'inhib' terms might improve the models or can be used to further adjust them. However, to avoid the proposal of too many model candidates, the automated detection of model deficiencies and model improvements should be developed further. If a method can be found that successfully determines deficiencies and proposes improvements, the modeling process may be sped up more significantly.

Appendix A

Unstructured Model with three Substrates

A.1 Dynamic Model

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{X}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \\ m_{\mathrm{P}}(t) \end{pmatrix} = \begin{pmatrix} 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t)V(t) \\ 0 \end{pmatrix} + V(t)\begin{pmatrix} 1 & 0 \\ -Y_{\mathrm{AmX}} & 0 \\ -Y_{\mathrm{PhX}} & 0 \\ -Y_{\mathrm{GcX}} & -Y_{\mathrm{GcP}} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} r_{\mathrm{X}}(t) \\ r_{\mathrm{P}}(t) \end{pmatrix} \tag{A.1}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{S}}(t) \tag{A.2}$$

$$r_{\mathrm{X}}(t) = \mu_{\mathrm{Xm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{XAm}}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XPh}}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XGc}}} \cdot c_{\mathrm{X}}(t) \tag{A.3}$$

$$r_{\mathrm{P}}(t) = \mu_{\mathrm{Pm}} \cdot \frac{K_{\mathrm{PPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PPh}}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XGc}}} \cdot c_{\mathrm{X}}(t) \tag{A.4}$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \tag{A.5}$$

Table A.1: Parameter values of the unstructured model

	Parameter	Value	Unit
r_{X}	μ_{Xm}	0.30	1/h
	K_{XAm}	0.20	g/L
	K_{XPh}	0.01	g/L
	K_{XGc}	1.00	g/L
r_{P}	μ_{Pm}	0.01	1/h
	K_{PPh}	0.01	g/L
	K_{PGc}	7.00	g/L

	Parameter	Value	Unit
r_{M}	μ_{Mm}	0.05	1/h
	K_{MS}	0.01	g/L
	Y_{AmX}	1.0	g/g
	Y_{PhX}	0.4	g/g
	Y_{GcX}	10.0	g/g
	Y_{GcP}	1.0	g/g

$$\underline{y}(t) = \begin{pmatrix} c_{\mathrm{X}}(t) \\ c_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc}}(t) \\ c_{\mathrm{P}}(t) \end{pmatrix} \tag{A.6}$$

A.2 Simulation Data

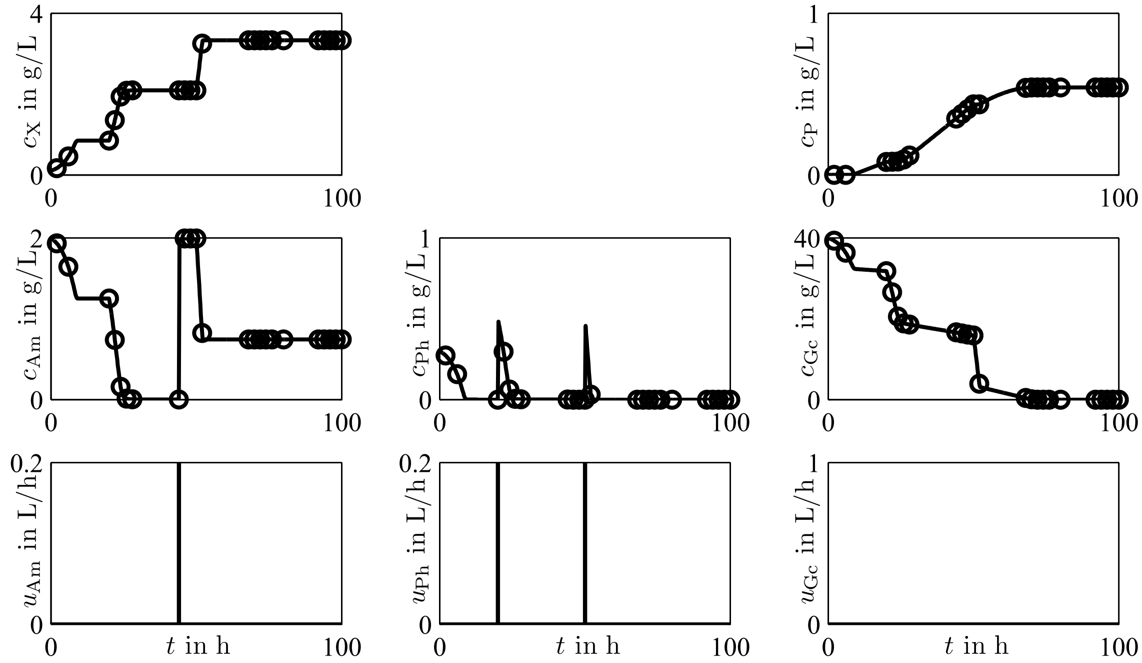


Figure A.1: UM3S1

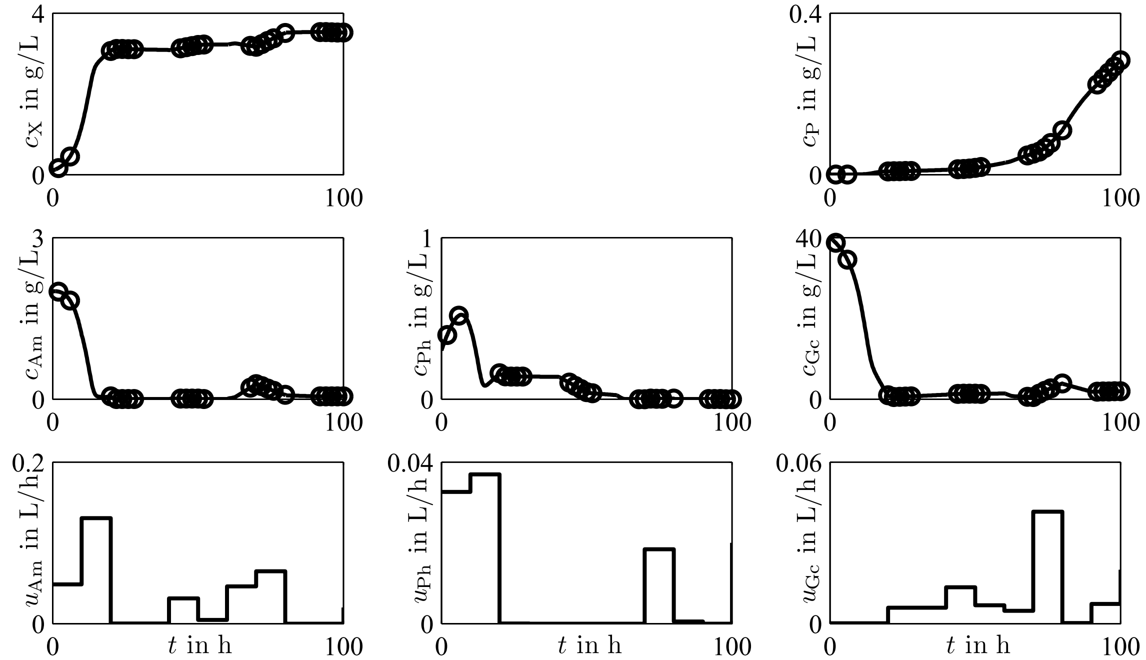


Figure A.2: UM3S2

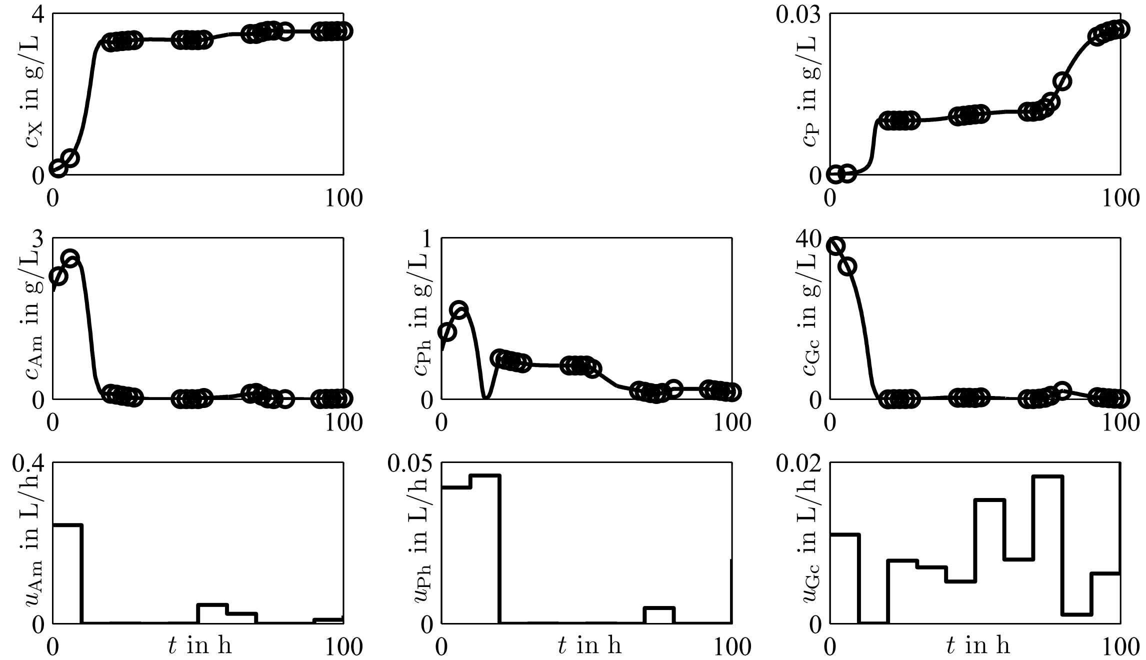


Figure A.3: UM3S3

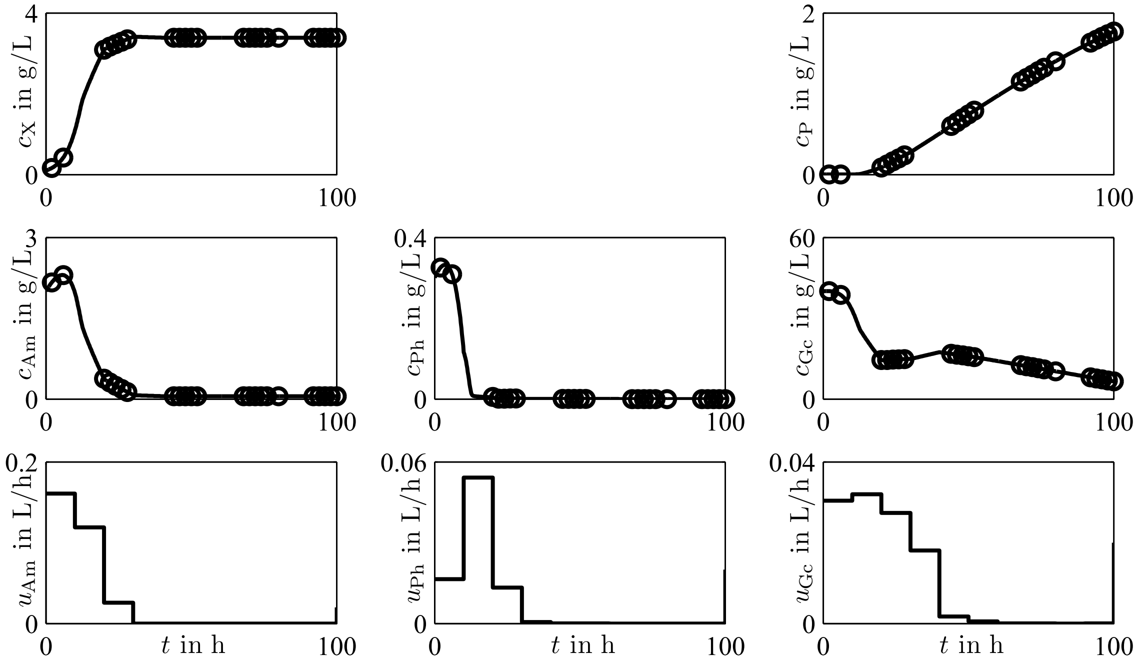


Figure A.4: UM3S4

A.3 Standard Deviation of the Measurement Noise

Table A.2: Standard deviation of the measurement noise of the unstructured model with three substrates. For each measurement variable, the standard deviation is linearly approximated.

c_i	σ_i
c_X	$\sigma_X = 0.009 \cdot c_X + 0.022\,\text{g/L}$
c_{Am}	$\sigma_{Am} = 0.024 \cdot c_S + 0.009\,\text{g/L}$
c_{Ph}	$\sigma_{Ph} = 0.018 \cdot c_S + 0.007\,\text{g/L}$
c_{Gc}	$\sigma_{Gc} = 0.0125 \cdot c_S + 0.25\,\text{g/L}$
c_P	$\sigma_P = 0.01 \cdot c_P + 0.001\,\text{g/L}$

Appendix B

Supplementary Material for *Paenibacillus polymyxa*

B.1 Results of the Phenomena Detection

Table B.1: Complete list of detected biological phenomena with measurements of *P. polymyxa*

Phenomenon	Sc
The growth is limited by ammonium.	0.06
The growth is limited by phosphate.	−0.69
Ammonium and phosphate are consumed simultaneously.	−0.49
Ammonium and glucose are consumed simultaneously.	−0.38
Phosphate and glucose are consumed simultaneously.	−0.72
The biomass and the product grow simultaneously.	−0.70
The product formation is limited by ammonium.	−0.70
The product formation is limited by phosphate.	−0.66
The product formation is limited by glucose.	−0.67
The product formation is inhibited by ammonium.	0.30
The product formation is inhibited by phosphate.	0.64
The product formation is inhibited by glucose.	0.14
Storage A for ammonium	−0.06
Storage A for phosphate	0.69
Storage B for phosphate	−0.33
Storage B for glucose	−0.33
The DNA formation is limited by ammonium.	−0.46
The RNA formation is limited by ammonium.	0.01
The protein formation is limited by ammonium.	0.31
The DNA formation is limited by phosphate.	−0.04
The RNA formation is limited by phosphate.	0.01
The protein formation is limited by phosphate.	−0.73
The RNA formation is limited by glucose.	−0.50
The protein formation is limited by glucose.	0.50

To be continued on next page

Table B.1: Detected biological phenomena of *P. polymyxa* – continued

Phenomenon	Sc
The DNA formation is inhibited by ammonium.	−0.08
The RNA formation is inhibited by ammonium.	0.08
The DNA formation is inhibited by phosphate.	−0.01
The DNA formation is inhibited by glucose.	−0.03
The DNA degradation is inhibited by ammonium.	−0.64
The RNA degradation is inhibited by ammonium.	0.12
The protein degradation is inhibited by ammonium.	−0.39
The DNA degradation is inhibited by phosphate.	−0.70
The RNA degradation is inhibited by phosphate.	−0.58
The protein degradation is inhibited by ammonium.	−0.21
The DNA degradation is inhibited by glucose.	−0.23
The RNA degradation is inhibited by glucose.	−0.49
The protein degradation is inhibited by glucose.	0.02
Intermediate compartment between ammonium and DNA	0.11
Intermediate compartment between ammonium and RNA	−0.12
Intermediate compartment between ammonium and proteins	0.22
Intermediate compartment between phosphate and DNA	−0.18
Intermediate compartment between phosphate and RNA	0.58
Intermediate compartment between phosphate and proteins	0.58
The protein formation is limited by DNA.	−0.56
The protein formation is limited by RNA.	0.19
The DNA formation is inhibited by RNA.	−0.06
The DNA formation is inhibited by proteins.	−0.04
The RNA formation is inhibited by proteins.	0.05
DNA and RNA grow simultaneously.	−0.38
DNA and proteins grow simultaneously.	−0.07
RNA and proteins grow simultaneously.	−0.17
DNA and RNA are degraded simultaneously.	−0.48
DNA and proteins are degraded simultaneously.	−0.66
RNA and proteins are degraded simultaneously.	−0.66
The product formation is limited by DNA.	−0.14
The product formation is limited by RNA.	−0.29
The product formation is limited by proteins.	−0.03
The product and DNA grow simultaneously.	−0.78
The product and RNA grow simultaneously.	−0.71
The product and proteins grow simultaneously.	−0.84
Degradation of biomass	0.67
Degradation of DNA	0.52
Degradation of the product	0.27

B.2 Standard Deviation of the Measurement Noise

Table B.2: Standard deviation of the noise for measurements of *P. polymyxa.* For each measurement variable, the standard deviation is linearly approximated.

c_i	σ_i
c_{X}	$\sigma_{\mathrm{X}} = 0.01 \cdot c_{\mathrm{X}} + 0.05\,\mathrm{g/L}$
c_{Am}	$\sigma_{\mathrm{Am}} = 0.015/2 \cdot c_{\mathrm{Am}} + 0.003\,\mathrm{g/L}$
c_{Ph}	$\sigma_{\mathrm{Ph}} = 0.011/0.6 \cdot c_{\mathrm{Ph}} + 0.007\,\mathrm{g/L}$
c_{Gc}	$\sigma_{\mathrm{Gc}} = 0.5/40 \cdot c_{\mathrm{Gc}} + 0.25\,\mathrm{g/L}$
c_{D}	$\sigma_{\mathrm{D}} = 0.024/12 \cdot c_{\mathrm{X}} + 0.006\,\mathrm{g/L}$
c_{R}	$\sigma_{\mathrm{R}} = 0.04\,\mathrm{g/L}$
c_{Pr}	$\sigma_{\mathrm{Pr}} = 0.12/14 \cdot c_{\mathrm{X}} + 0.02\,\mathrm{g/L}$
c_{Ml}	$\sigma_{\mathrm{Ml}} = 0.01\,\mathrm{g/L}$

B.3 Additional Identified Experiments for the Small-Size Structured Model

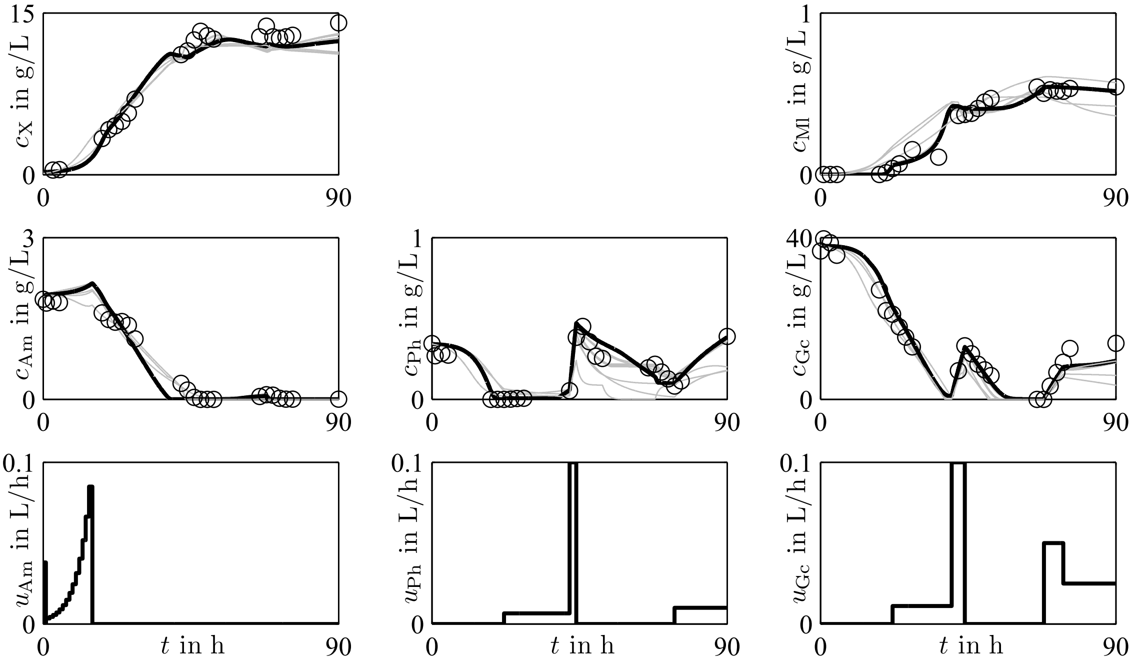


Figure B.1: Identified experiment PPdef8 for *P. polymyxa*

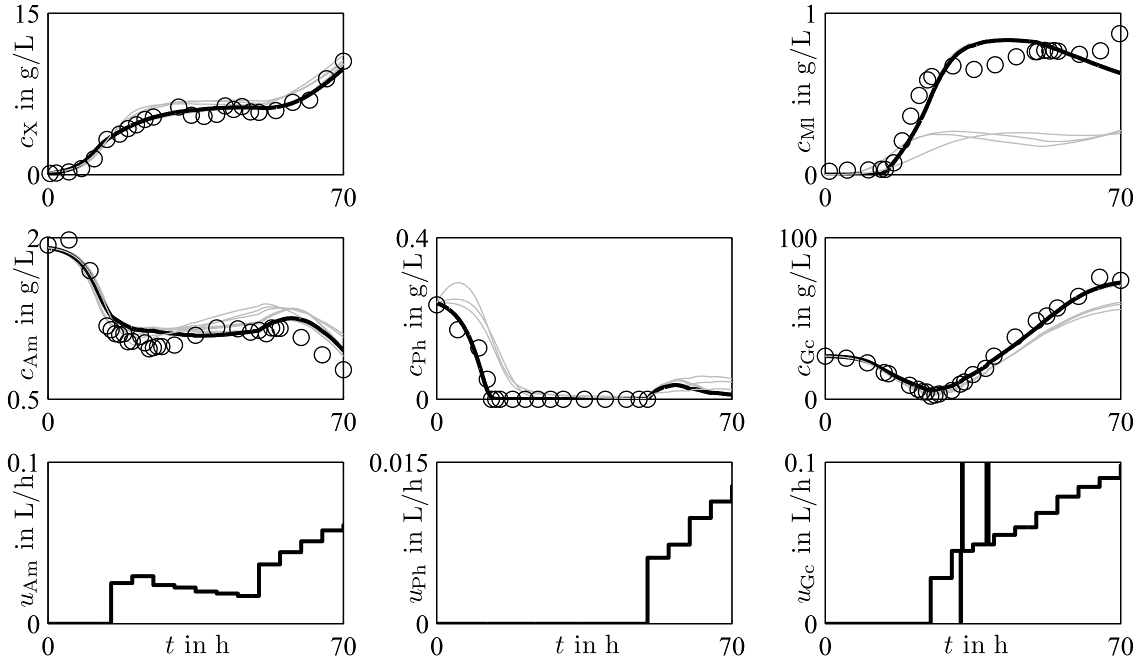


Figure B.2: Identified experiment PPdef13 for *P. polymyxa*

B.4 Additional Validation Experiment for the Small-Size Structured Model

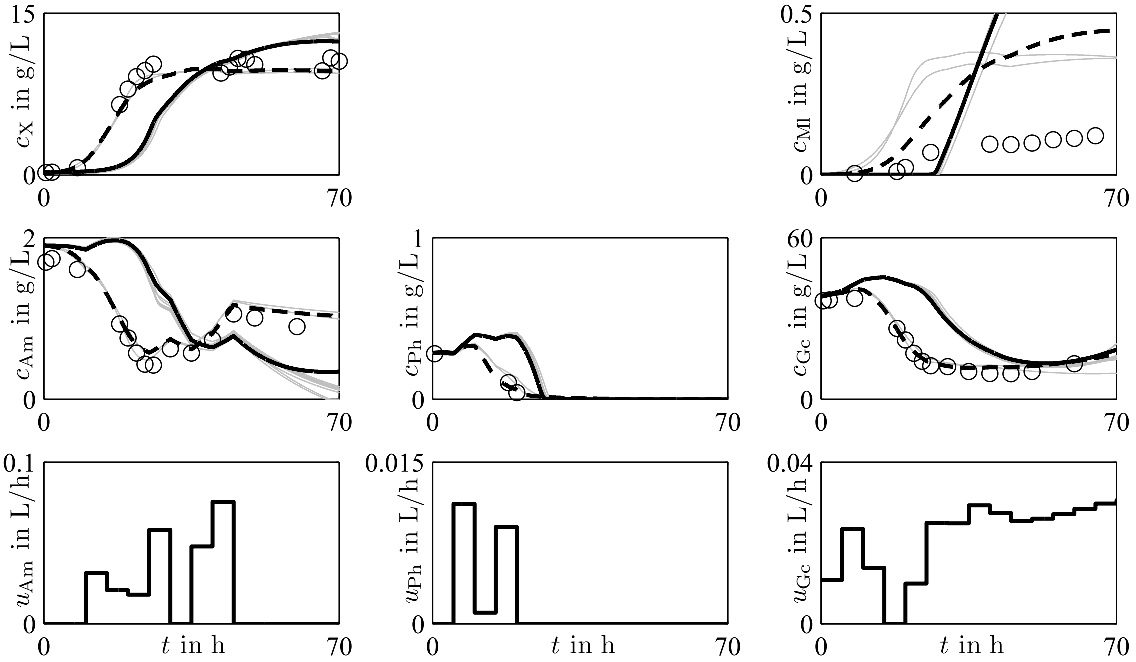


Figure B.3: Validation experiment PPdef15 for *P. polymyxa*

B.5 Best Validated Small-Size Structured Model

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{Xa}}(t) \\ m_{\mathrm{PhSt}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t) V(t) \end{pmatrix} + \\
& \quad V_{\mathrm{X}}(t) \begin{pmatrix} Y_{\mathrm{Xa}} & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 \\ -Y_{\mathrm{PhXa}} & 0 & -1 & 1 \\ -Y_{\mathrm{GcXa}} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} r_{\mathrm{Xa}}(t) \\ r_{\mathrm{dXa}}(t) \\ r_{\mathrm{PhSt}}(t) \\ r_{\mathrm{dPhSt}}(t) \end{pmatrix}
\end{aligned} \tag{B.1}
$$

$$
\frac{\mathrm{d}m_{\mathrm{Ml}}(t)}{\mathrm{d}t} = r_{\mathrm{Ml}}(t) \cdot V_{\mathrm{X}}(t) - r_{\mathrm{dMl}}(t) \cdot V(t) \tag{B.2}
$$

$$
\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t) \tag{B.3}
$$

Table B.3: Parameter values of the best validated small-size structured model for *P. polymyxa*

	Parameter	Value	Unit
r_{Xa}	μ_{Xam}	0.0955	1/h
	K_{XaAm}	0.0530	g/L
	K_{XaPh}	0.269	g/L
	K_{XaGc}	1×10^{-4}	g/L
r_{dXa}	μ_{dXam}	0	1/h
r_{PhSt}	μ_{PhStm}	1.879	1/h
	K_{PhSt}	29.249	g/L
r_{dPhSt}	μ_{dPhStm}	0.161	1/h
	K_{dPhSt}	5.40×10^{-4}	g/L

	Parameter	Value	Unit
r_{Ml}	μ_{Mlm}	3.26×10^{-3}	1/h
	K_{MlAm}	40.698	g/L
	K_{MlPh}	42.001	g/L
	K_{MlGc}	100	g/L
r_{dMl}	μ_{dMlm}	0.0615	1/h
r_{M}	μ_{Mm}	0.0799	1/h
	K_{M}	0.01	g/L
	Y_{Xa}	5.414	g/g
	Y_{PhXa}	0.247	g/g
	Y_{GcXa}	15.890	g/g

$$r_{\mathrm{Xa}}(t) = \mu_{\mathrm{Xam}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{XaAm}}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XaPh}}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XaGc}}} \cdot g_{\mathrm{Xa}}(t) \tag{B.4}$$

$$r_{\mathrm{dXa}}(t) = \mu_{\mathrm{dXam}} \cdot g_{\mathrm{Xa}}(t) \tag{B.5}$$

$$r_{\mathrm{PhSt}}(t) = \mu_{\mathrm{PhStm}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PhSt}}} \cdot g_{\mathrm{Xa}}(t) \tag{B.6}$$

$$r_{\mathrm{dPhSt}}(t) = \mu_{\mathrm{dPhStm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{dPhSt}}} \cdot g_{\mathrm{PhSt}}(t) \tag{B.7}$$

$$r_{\mathrm{Ml}}(t) = \mu_{\mathrm{Mlm}} \cdot \frac{K_{\mathrm{MlAm}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{MlAm}}} \cdot \frac{K_{\mathrm{MlPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{MlPh}}} \cdot \frac{K_{\mathrm{MlGc}}}{c_{\mathrm{Gc}}(t) + K_{\mathrm{MlGc}}} \cdot g_{\mathrm{Xa}}(t) \tag{B.8}$$

$$r_{\mathrm{dMl}}(t) = \mu_{\mathrm{dMlm}} \cdot c_{\mathrm{Ml}}(t) \tag{B.9}$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \tag{B.10}$$

$$\underline{y}(t) = \begin{pmatrix} c_{\mathrm{X}}(t) \\ c_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc}}(t) \\ c_{\mathrm{Ml}}(t) \end{pmatrix} \tag{B.11}$$

B.6 Additional Identified Experiment for the Medium-Size Structured Model

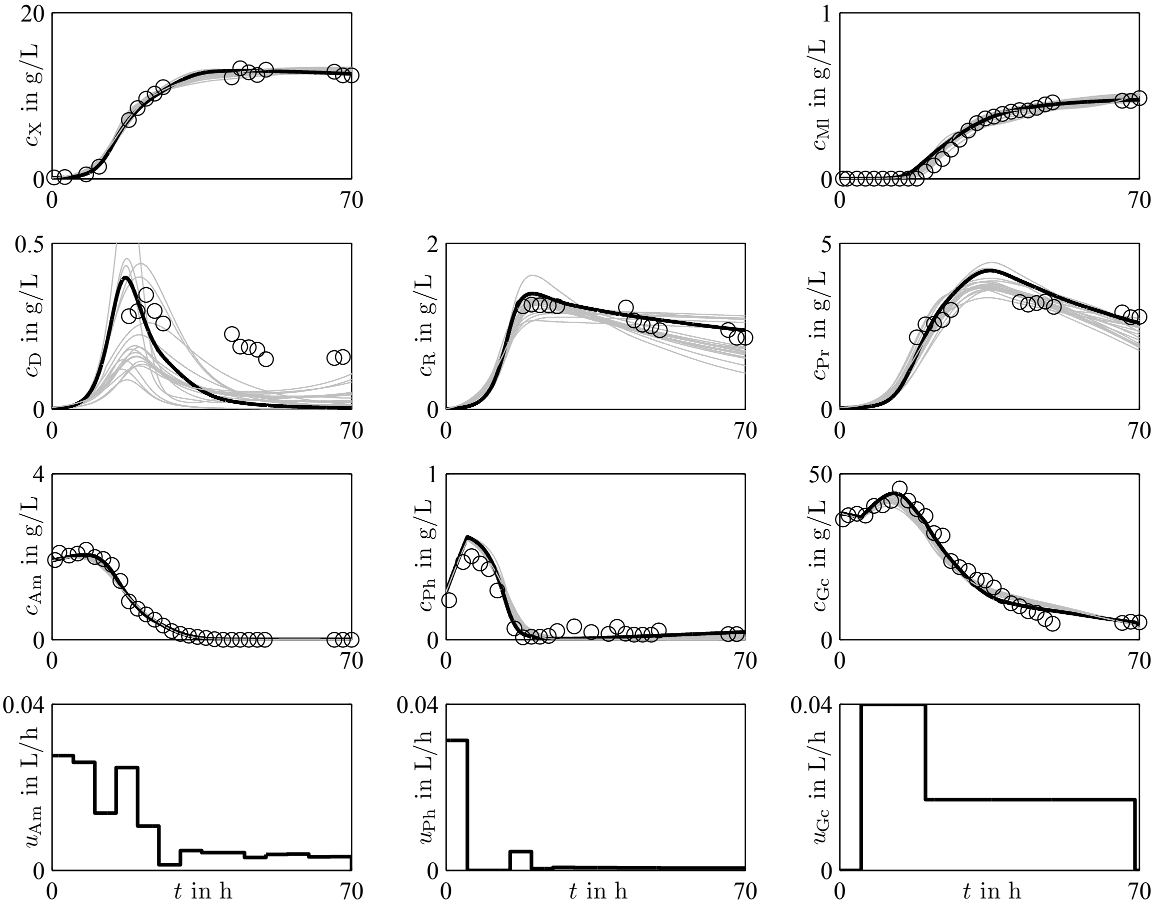


Figure B.4: Identified experiment PPdef11 for *P. polymyxa* with cell-intern measurements

B.7 Additional Validation Experiment for the Medium-Size Structured Model

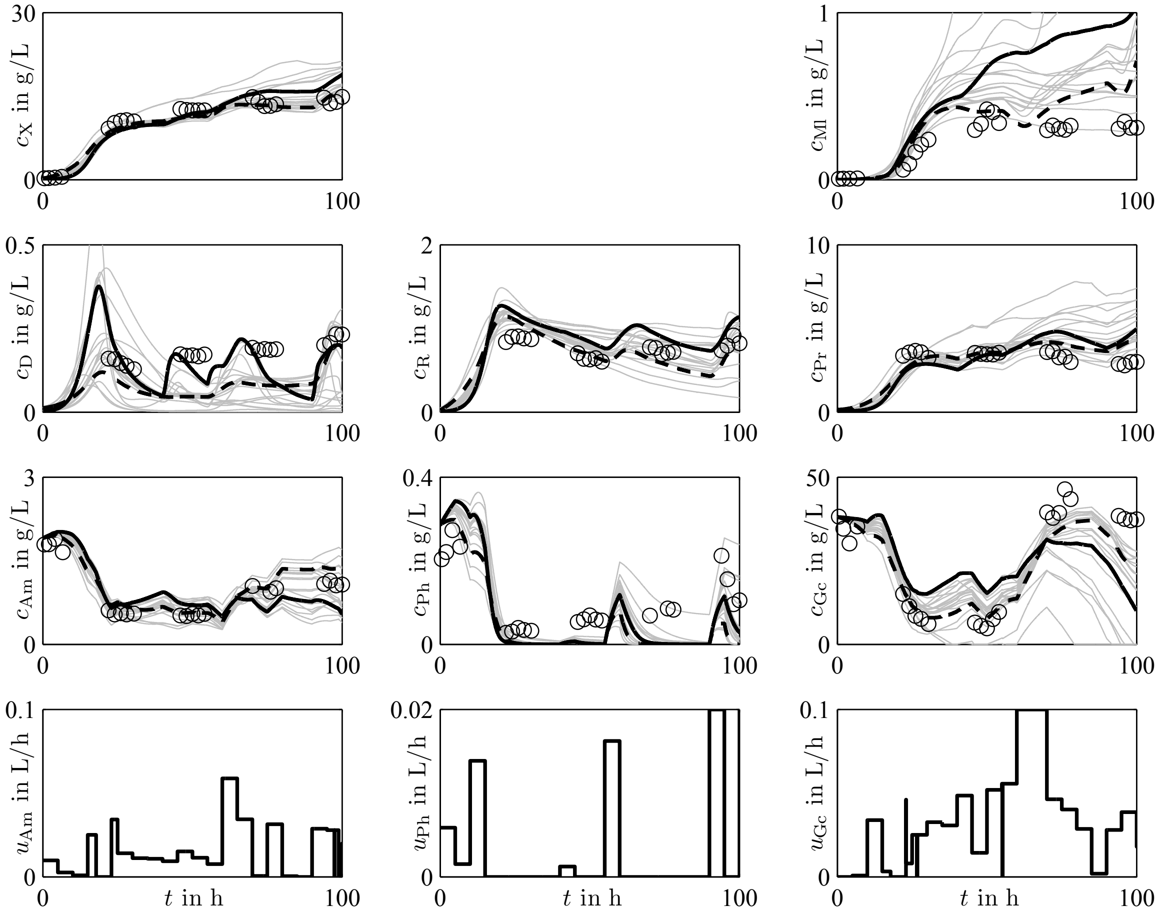


Figure B.5: Validation experiment PPdef19 for *P. polymyxa* with cell-intern measurements

B.8 Best Validated Medium-Size Structured Model

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{D}}(t) \\ m_{\mathrm{R}}(t) \\ m_{\mathrm{Pr}}(t) \\ m_{\mathrm{D^*}}(t) \\ m_{\mathrm{Pr^*}}(t) \\ m_{\mathrm{Xr}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix} = &\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t)V(t) \end{pmatrix} + \\
&V_{\mathrm{X}}(t)\begin{pmatrix} Y_{\mathrm{D}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & Y_{\mathrm{RD^*}} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & Y_{\mathrm{RXr}} \\ -1 & 0 & 0 & -Y_{\mathrm{AmR}} & 0 & 0 \\ -Y_{\mathrm{PhD}} & 0 & 0 & -Y_{\mathrm{PhR}} & 0 & 0 \\ -Y_{\mathrm{GcD}} & 0 & 0 & -Y_{\mathrm{GcR}} & 0 & 0 \end{pmatrix}\begin{pmatrix} r_{\mathrm{D}}(t) \\ r_{\mathrm{D^*D}}(t) \\ r_{\mathrm{dD}}(t) \\ r_{\mathrm{R}}(t) \\ r_{\mathrm{D^*R}}(t) \\ r_{\mathrm{dR}}(t) \end{pmatrix} + \\
&V_{\mathrm{X}}(t)\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ Y_{\mathrm{Pr}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & Y_{\mathrm{PrPr^*}} & 0 & 1 & 0 \\ 0 & 0 & Y_{\mathrm{PrXr}} & 0 & 0 & 1 \\ -1 & 0 & 0 & -Y_{\mathrm{AmD^*}} & 0 & 0 \\ -Y_{\mathrm{PhPr}} & 0 & 0 & -Y_{\mathrm{PhD^*}} & -Y_{\mathrm{PhPr^*}} & -Y_{\mathrm{PhXr}} \\ -Y_{\mathrm{GcPr}} & 0 & 0 & 0 & 0 & -Y_{\mathrm{GcXr}} \end{pmatrix}\begin{pmatrix} r_{\mathrm{Pr}}(t) \\ r_{\mathrm{Pr^*Pr}}(t) \\ r_{\mathrm{dPr}}(t) \\ r_{\mathrm{D^*}}(t) \\ r_{\mathrm{Pr^*}}(t) \\ r_{\mathrm{Xr}}(t) \end{pmatrix}
\end{aligned}
\tag{B.12}
$$

$$\frac{\mathrm{d}m_{\mathrm{Ml}}(t)}{\mathrm{d}t} = r_{\mathrm{Ml}}(t) \cdot V_{\mathrm{X}}(t) - r_{\mathrm{dMl}}(t) \cdot V(t) \tag{B.13}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t) \tag{B.14}$$

$$r_{\mathrm{D}}(t) = \mu_{\mathrm{Dm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{DAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{DPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{DGc}}} \cdot g_{\mathrm{D}}(t) \tag{B.15}$$

$$r_{\mathrm{D^*D}}(t) = \mu_{\mathrm{D^*Dm}} \cdot \frac{g_{\mathrm{D^*}}(t)}{g_{\mathrm{D^*}}(t) + K_{\mathrm{DD^*}}} \cdot g_{\mathrm{D}}(t) \tag{B.16}$$

$$r_{\mathrm{dD}}(t) = \mu_{\mathrm{dDm}} \cdot g_{\mathrm{D}}(t) \tag{B.17}$$

$$r_{\mathrm{R}}(t) = \mu_{\mathrm{Rm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{RAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{RPh}}} \cdot g_{\mathrm{D}}(t) \tag{B.18}$$

$$r_{\mathrm{D^*R}}(t) = \mu_{\mathrm{D^*Rm}} \cdot \frac{g_{\mathrm{D^*}}(t)}{g_{\mathrm{D^*}}(t) + K_{\mathrm{RD^*}}} \cdot g_{\mathrm{D}}(t) \tag{B.19}$$

Table B.4: Parameter values of the best validated medium-size structured model for *P. polymyxa*

	Parameter	Value	Unit
r_{D}	μ_{Dm}	0.0565	1/h
	K_{DAm}	0.665	g/L
	K_{DPh}	0.0597	g/L
	K_{DGc}	1.26×10^{-4}	g/L
$r_{\mathrm{D^*D}}$	$\mu_{\mathrm{D^*Dm}}$	0.537	1/h
	$K_{\mathrm{DD^*}}$	0.434	g/L
r_{dD}	μ_{dDm}	0.231	1/h
r_{R}	μ_{Rm}	2.987	1/h
	K_{RAm}	1×10^{-4}	g/L
	K_{RPh}	0.227	g/L
$r_{\mathrm{D^*R}}$	$\mu_{\mathrm{D^*Rm}}$	9.28×10^{-4}	1/h
	$K_{\mathrm{RD^*}}$	1×10^{-4}	g/L
r_{dR}	μ_{dRm}	0.0164	1/h
	K_{dRAm}	100	g/L
	$K_{\mathrm{dRD^*}}$	100	g/L
r_{Pr}	μ_{Prm}	0.0804	1/h
	K_{PrAm}	0.179	g/L
	K_{PrPh}	1×10^{-4}	g/L
	K_{PrGc}	0.683	g/L
$r_{\mathrm{Pr^*Pr}}$	$\mu_{\mathrm{Pr^*Prm}}$	1.616	1/h
	$K_{\mathrm{PrPr^*}}$	1×10^{-4}	g/L
r_{dPr}	μ_{dPrm}	20	1/h
	K_{dPrGc}	8.26×10^{-3}	g/L
	$K_{\mathrm{dPrPr^*}}$	0.0177	g/L
$r_{\mathrm{D^*}}$	$\mu_{\mathrm{D^*m}}$	0.0263	1/h
	$K_{\mathrm{D^*Am}}$	1×10^{-4}	g/L
	$K_{\mathrm{D^*Ph}}$	1.853	g/L

	Parameter	Value	Unit
$r_{\mathrm{Pr^*}}$	$\mu_{\mathrm{Pr^*m}}$	0.203	1/h
	$K_{\mathrm{Pr^*Ph}}$	0.956	g/L
r_{Xr}	μ_{Xrm}	3.702	1/h
	K_{XrPh}	0.810	g/L
	K_{XrGc}	50	g/L
r_{Ml}	μ_{Mlm}	0.646	1/h
	K_{MlAm}	100	g/L
	K_{MlPh}	0.0232	g/L
	K_{MlGc}	100	g/L
r_{dMl}	μ_{dMlm}	0.0557	1/h
r_{M}	μ_{Mm}	0.0678	1/h
	K_{M}	0.01	g/L
	Y_{D}	8.957	g/g
	Y_{Pr}	2.599	g/g
	$Y_{\mathrm{RD^*}}$	7.403	g/g
	$Y_{\mathrm{PrPr^*}}$	0.0106	g/g
	Y_{RXr}	0.152	g/g
	Y_{PrXr}	0.0779	g/g
	Y_{AmR}	0.282	g/g
	$Y_{\mathrm{AmD^*}}$	12.253	g/g
	Y_{PhD}	0.163	g/g
	Y_{PhR}	0.162	g/g
	Y_{PhPr}	0	g/g
	$Y_{\mathrm{PhD^*}}$	0.248	g/g
	$Y_{\mathrm{PhPr^*}}$	1.25×10^{-4}	g/g
	Y_{PhXr}	0.0821	g/g
	Y_{GcD}	0	g/g
	Y_{GcR}	0	g/g
	Y_{GcPr}	18.433	g/g
	Y_{GcXr}	2.304	g/g

$$r_{\mathrm{dR}}(t) = \mu_{\mathrm{dRm}} \cdot \frac{K_{\mathrm{dRAm}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{dRAm}}} \frac{K_{\mathrm{dRD^*}}}{g_{\mathrm{D^*}}(t) + K_{\mathrm{dRD^*}}} \cdot g_{\mathrm{R}}(t) \tag{B.20}$$

$$r_{\mathrm{Pr}}(t) = \mu_{\mathrm{Prm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{PrAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{PrGc}}} \cdot g_{\mathrm{R}}(t) \tag{B.21}$$

$$r_{\mathrm{Pr^*Pr}}(t) = \mu_{\mathrm{Pr^*Prm}} \cdot \frac{g_{\mathrm{Pr^*}}(t)}{g_{\mathrm{Pr^*}}(t) + K_{\mathrm{PrPr^*}}} \cdot g_{\mathrm{Pr}}(t) \tag{B.22}$$

$$r_{\mathrm{dPr}}(t) = \mu_{\mathrm{dPrm}} \cdot \frac{K_{\mathrm{dPrGc}}}{c_{\mathrm{Gc}}(t) + K_{\mathrm{dPrGc}}} \frac{K_{\mathrm{dPrPr^*}}}{g_{\mathrm{Pr^*}}(t) + K_{\mathrm{dPrPr^*}}} \cdot g_{\mathrm{Pr}}(t) \tag{B.23}$$

$$r_{\mathrm{D^*}}(t) = \mu_{\mathrm{D^*m}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{D^*Am}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{D^*Ph}}} \cdot g_{\mathrm{D}}(t) \tag{B.24}$$

$$r_{\mathrm{Pr^*}}(t) = \mu_{\mathrm{Pr^*m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Pr^*Ph}}} \cdot g_{\mathrm{Pr}}(t) \tag{B.25}$$

$$r_{\mathrm{Xr}}(t) = \mu_{\mathrm{Xrm}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XrGc}}} \cdot g_{\mathrm{Pr}}(t) \tag{B.26}$$

$$r_{\mathrm{Ml}}(t) = \mu_{\mathrm{Mlm}} \cdot \frac{K_{\mathrm{MlAm}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{MlAm}}} \frac{K_{\mathrm{MlPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{MlPh}}} \frac{K_{\mathrm{MlGc}}}{c_{\mathrm{Gc}}(t) + K_{\mathrm{MlGc}}} \cdot g_{\mathrm{D}}(t) \tag{B.27}$$

$$r_{\mathrm{dMl}}(t) = \mu_{\mathrm{dMlm}} \cdot c_{\mathrm{Ml}}(t) \tag{B.28}$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot c_{\mathrm{X}}(t) \tag{B.29}$$

Appendix C

Supplementary Material for *Streptomyces tendae*

C.1 Dynamic Model

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{D}}(t) \\ m_{\mathrm{R}}(t) \\ m_{\mathrm{Pr}}(t) \\ m_{\mathrm{Aa}}(t) \\ m_{\mathrm{Nu}}(t) \\ m_{\mathrm{U}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_{\mathrm{Am,\,in}} u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} u_{\mathrm{Gc}}(t) \end{pmatrix}}_{\text{Inlet}} \underbrace{- \frac{Q_{\mathrm{out}}(t)}{V(t)} \begin{pmatrix} m_{\mathrm{D}}(t) \\ m_{\mathrm{R}}(t) \\ m_{\mathrm{Pr}}(t) \\ m_{\mathrm{Aa}}(t) \\ m_{\mathrm{Nu}}(t) \\ m_{\mathrm{U}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix}}_{\text{Outlet}} - \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t) V_{\mathrm{X}}(t) \end{pmatrix}}_{\text{Maintenance}} + \\
\underbrace{V_{\mathrm{X}}(t) \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 & -Y_{\mathrm{AaNu}} & -Y_{\mathrm{AaU}} & 0 & Y_{\mathrm{AaPr}} \\ -1 & -1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 - Y_{\mathrm{AaPr}} \\ 0 & 0 & 0 & -Y_{\mathrm{AmAa}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -Y_{\mathrm{PhNu}} & 0 & 0 & 0 \\ 0 & 0 & 0 & -Y_{\mathrm{GcAa}} & -Y_{\mathrm{GcNu}} & Y_{\mathrm{AaU}} - 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} r_{\mathrm{D}}(t) \\ r_{\mathrm{R}}(t) \\ r_{\mathrm{Pr}}(t) \\ r_{\mathrm{Aa}}(t) \\ r_{\mathrm{Nu}}(t) \\ r_{\mathrm{U}}(t) \\ r_{\mathrm{dR}}(t) \\ r_{\mathrm{dPr}}(t) \end{pmatrix}}_{\text{Metabolism}}
\end{aligned}
\tag{C.1}
$$

Replication and DNA activity

$$
r_{\mathrm{D}}(t) = \mu_{\mathrm{Dm}} \cdot \frac{g_{\mathrm{Aa}}(t)}{g_{\mathrm{Aa}}(t) + K_{\mathrm{PrAa}}} \cdot \frac{g_{\mathrm{Nu}}(t)}{g_{\mathrm{Nu}}(t) + K_{\mathrm{PrNu}}} \cdot g_{\mathrm{R}}(t) \cdot g_{\mathrm{D}}(t) \cdot \phi(t) \tag{C.2}
$$

$$
\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = -\mu_{\phi\mathrm{m}} \cdot \frac{K_{\phi\mathrm{Aa}}}{g_{\mathrm{Aa}}(t) + K_{\phi\mathrm{Aa}}} \cdot \phi(t), \quad \phi(t_0) = 1 \tag{C.3}
$$

Table C.1: Parameter values of the dynamic model of *S. tendae* according to Majer (1997)

	Parameter	Value	Unit
r_{D}	μ_{Dm}	0.705	l/(gh)
	K_{DAa}	0.0164	g/L
r_{ϕ}	$\mu_{\phi\mathrm{m}}$	0.0237	1/h
	$K_{\phi\mathrm{Aa}}$	1.23×10^{-4}	g/L
r_{R}	μ_{Rm}	1.888	1/h
	K_{RNu}	0.0183	g/L
	K_{RG}	11 263	1
	K_{G}	1.38×10^{-6}	g/L
r_{dR}	μ_{dRm}	0.0154	1/h
	K_{dR}	0.607	g/L
r_{Pr}	μ_{Prm}	0.43	1/h
	K_{PrAa}	0.013	g/L
	K_{PrNu}	5.56×10^{-3}	g/L
r_{dPr}	μ_{dPrm}	9×10^{-3}	1/h
	K_{dPr}	2.9×10^{-4}	g/L
r_{Aa}	μ_{Aa1m}	0.02	1/h
	K_{Aa1}	4.14×10^{-3}	g/L
	μ_{Aa2m}	6.4×10^{-2}	1/h
	K_{Aa2}	0.606	g/L
	K_{AaGc}	1.1418	g/L
	μ_{AaNu}	2.07	1
	K_{AaNu}	2.48×10^{-3}	g/L
	K_{AaPh}	0.856	$\mathrm{L}^3/\mathrm{g}^3$

	Parameter	Value	Unit
r_{Nu}	μ_{Nu1m}	0.0334	1/h
	K_{Nu1}	0.0774	g/L
	μ_{Nu2m}	0.606	1/h
	K_{Nu2}	9.31×10^{-3}	g/L
	K_{NuAa}	4.36×10^{-3}	g/L
r_{U}	μ_{Um}	0.018	1/h
	K_{UAa}	0.1	g/L
	K_{UI}	0.241	g/L
	K_{UNu}	1.85	1
r_{M}	μ_{Mm}	0.0353	1/h
	K_{M}	0.6	g/L
r_{Nm}	μ_{Nm1m}	3.77×10^{-3}	1/h
	K_{NmAa}	0.368	g/L
	μ_{Nm2m}	0.03	1/h
	K_{NmNu}	0.165	g/L
r_{dNm}	μ_{dNmm}	2.74×10^{8}	1/h
	K_{dNm1}	0.787	1
	K_{dNm2}	8830	K
	Y_{AaNu}	0.0183	g/g
	Y_{AaU}	0.175	g/g
	Y_{AaPr}	0.157	g/g
	Y_{AmAa}	0.226	g/g
	Y_{PhNu}	0.313	g/g
	Y_{GcAa}	2.143	g/g
	Y_{GcNu}	1.096	g/g

Transcription and RNA degradation

$$r_{\mathrm{R}}(t) = \mu_{\mathrm{Rm}} \cdot \frac{g_{\mathrm{Nu}}(t)}{g_{\mathrm{Nu}}(t) + K_{\mathrm{RNu}} + K_{\mathrm{RG}} \cdot \dfrac{K_{\mathrm{G}}}{g_{\mathrm{Aa}}(t) + K_{\mathrm{G}}} \cdot g_{\mathrm{R}}(t)} \cdot g_{\mathrm{D}}(t) \tag{C.4}$$

$$r_{\mathrm{dR}}(t) = \mu_{\mathrm{dRm}} \cdot \frac{K_{\mathrm{dR}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{dR}}} \cdot g_{\mathrm{R}}(t) \tag{C.5}$$

Translation and protein degradation

$$r_{\mathrm{Pr}}(t) = \mu_{\mathrm{Prm}} \cdot \frac{g_{\mathrm{Aa}}(t)}{g_{\mathrm{Aa}}(t) + K_{\mathrm{PrAa}}} \cdot \frac{g_{\mathrm{Nu}}(t)}{g_{\mathrm{Nu}}(t) + K_{\mathrm{PrNu}}} \cdot g_{\mathrm{R}}(t) \tag{C.6}$$

$$r_{\mathrm{dPr}}(t) = \mu_{\mathrm{dPrm}} \cdot \frac{K_{\mathrm{dPr}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{dPr}}} \cdot g_{\mathrm{Pr}}(t) \tag{C.7}$$

Production of amino acids

$$\begin{aligned} r_{\mathrm{Aa}}(t) = {} & \left(\mu_{\mathrm{Aa1m}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{Aa1}}} + \mu_{\mathrm{Aa2m}} \cdot \frac{K_{\mathrm{Aa2}}}{c_{\mathrm{Am}}(t) + K_{\mathrm{Aa2}}}\right) \cdot \frac{c_{\mathrm{C}}(t)}{c_{\mathrm{C}}(t) + K_{\mathrm{AaC}}} \\ & \cdot \left(1 + \mu_{\mathrm{AaNu}} \cdot \frac{g_{\mathrm{Nu}}(t)}{g_{\mathrm{Nu}}(t) + K_{\mathrm{AaNu}}}\right) \cdot \frac{1}{1 + (c_{\mathrm{Ph}}(t))^3 \cdot K_{\mathrm{AaPh}}} \cdot g_{\mathrm{Pr}}(t), \quad c_{\mathrm{Am}}(t) > 0 \end{aligned} \tag{C.8}$$

Nucleotide synthesis

$$\begin{aligned} r_{\mathrm{Nu}}(t) = {} & \left(\mu_{\mathrm{Nu1m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Nu1}}} + \mu_{\mathrm{Nu2m}} \cdot \frac{K_{\mathrm{Nu2}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Nu2}}}\right) \\ & \cdot \frac{g_{\mathrm{Aa}}(t)}{g_{\mathrm{Aa}}(t) + K_{\mathrm{NuAa}}} \cdot g_{\mathrm{Pr}}(t), \quad c_{\mathrm{Ph}}(t) > 0 \end{aligned} \tag{C.9}$$

Structural elements

$$r_{\mathrm{U}}(t) = \mu_{\mathrm{Um}} \cdot \frac{g_{\mathrm{Aa}}(t)}{g_{\mathrm{Aa}}(t) + K_{\mathrm{UAa}} + \dfrac{(g_{\mathrm{Aa}}(t))^2}{K_{\mathrm{UI}}} + K_{\mathrm{UNu}} \cdot g_{\mathrm{Nu}}(t)} \cdot g_{\mathrm{Pr}}(t) \tag{C.10}$$

Maintenance

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{M}}} \cdot g_{\mathrm{X}}(t) \tag{C.11}$$

Secondary metabolite

$$r_{\mathrm{Nm}}(t) = \left(\mu_{\mathrm{Nm1m}} \cdot \frac{K_{\mathrm{NmAa}}}{g_{\mathrm{Aa}}(t) + K_{\mathrm{NmAa}}} + \mu_{\mathrm{Nm2m}} \cdot \frac{K_{\mathrm{NmNu}}}{g_{\mathrm{Nu}}(t) + K_{\mathrm{NmNu}}}\right) \cdot g_{\mathrm{D}}(t) \tag{C.12}$$

$$r_{\mathrm{dNm}}(t) = \mu_{\mathrm{dNmm}} \cdot \exp\left(K_{\mathrm{dNm1}} \cdot \mathrm{pH}(t) - \frac{K_{\mathrm{dNm2}}}{T(t)}\right) \cdot c_{\mathrm{Nm}}(t) \tag{C.13}$$

$$\frac{\mathrm{d}m_{\mathrm{Nm}}(t)}{\mathrm{d}t} = V_{\mathrm{X}}(t) \cdot r_{\mathrm{Nm}}(t) - V(t) \cdot r_{\mathrm{dNm}}(t) \tag{C.14}$$

Reactor volume

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t) - Q_{\mathrm{out}}(t) + \sum_j Q_j(t) \tag{C.15}$$

C.2 Simulation Data

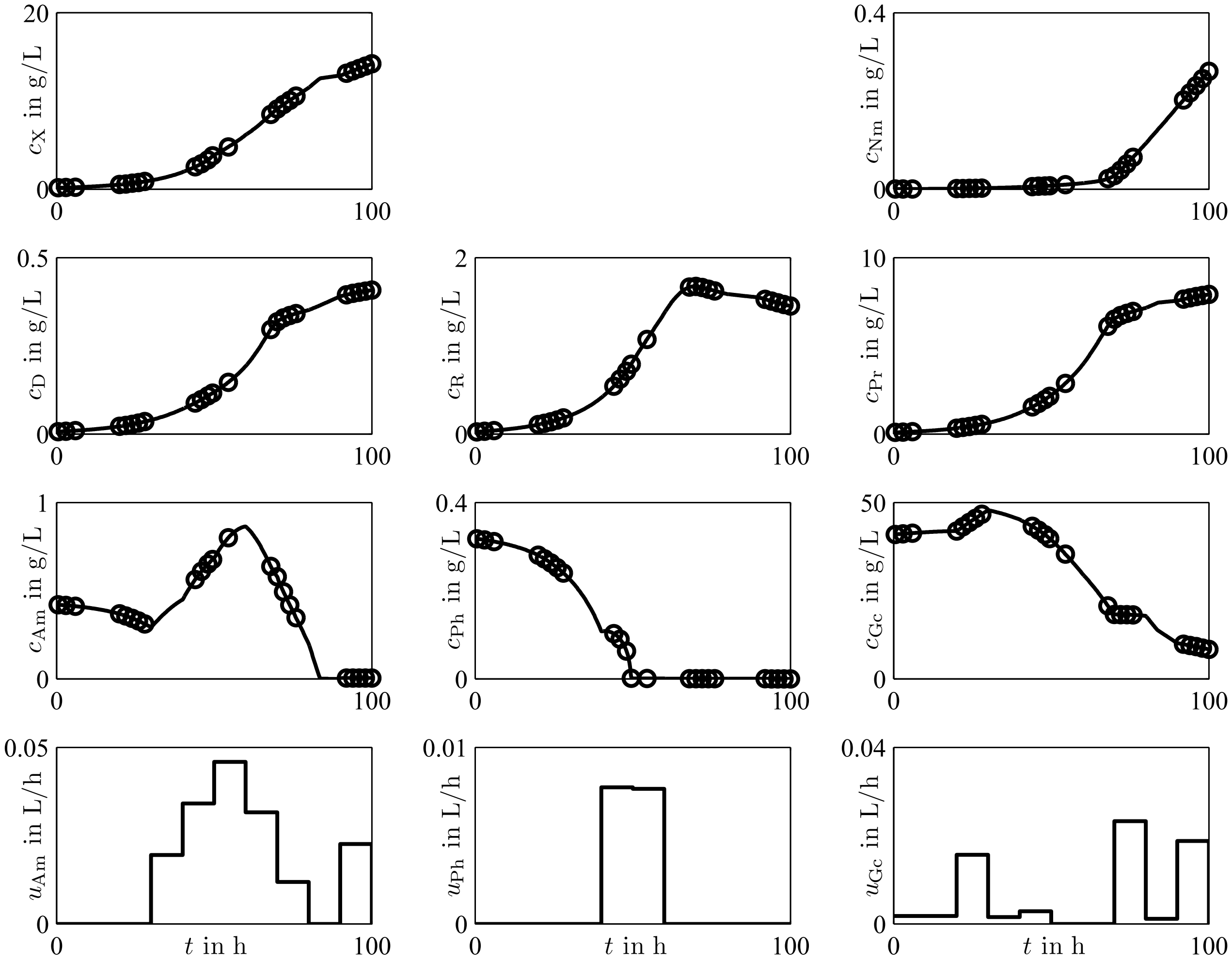


Figure C.1: Simulation of STdef1

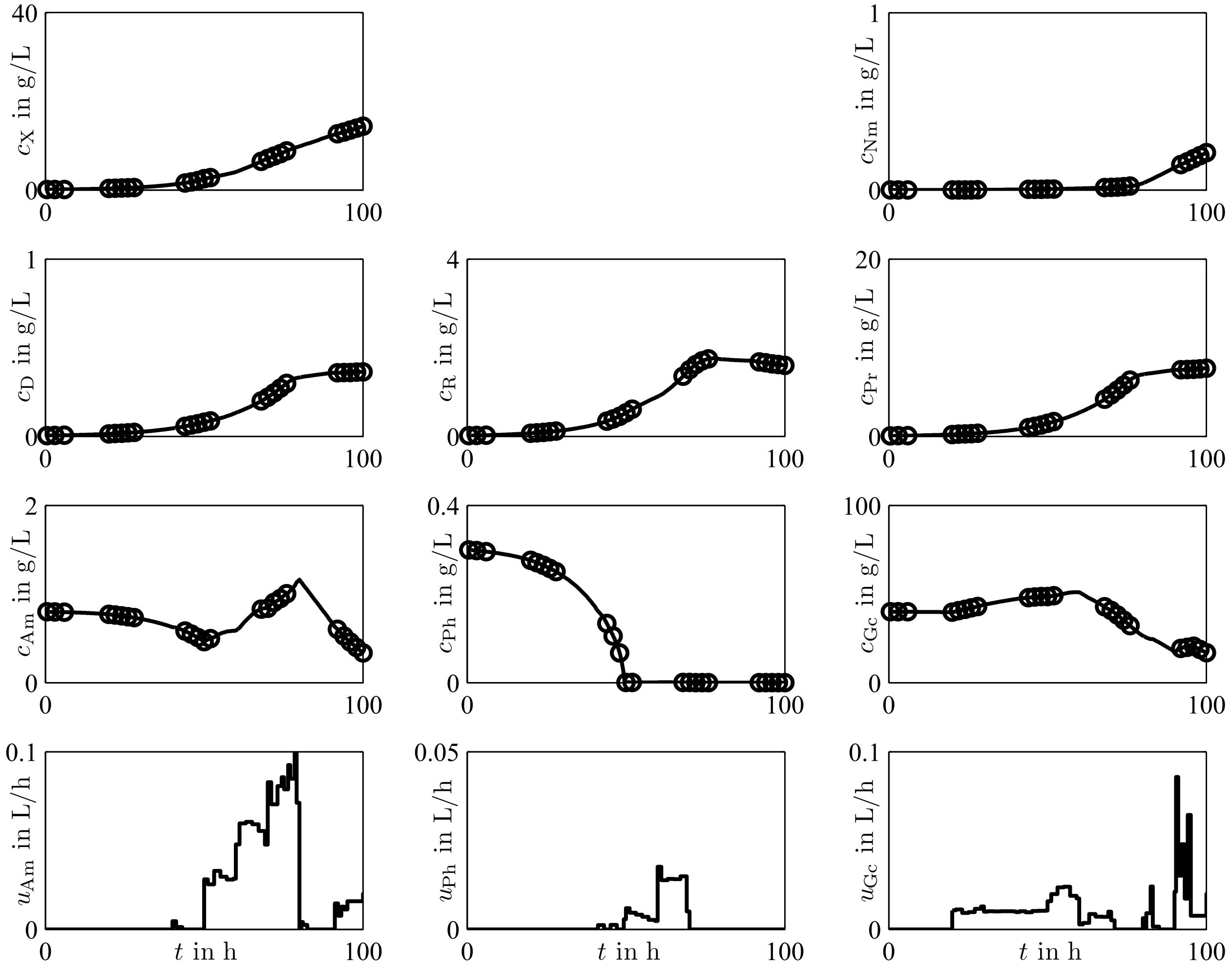


Figure C.2: Simulation of STdef2

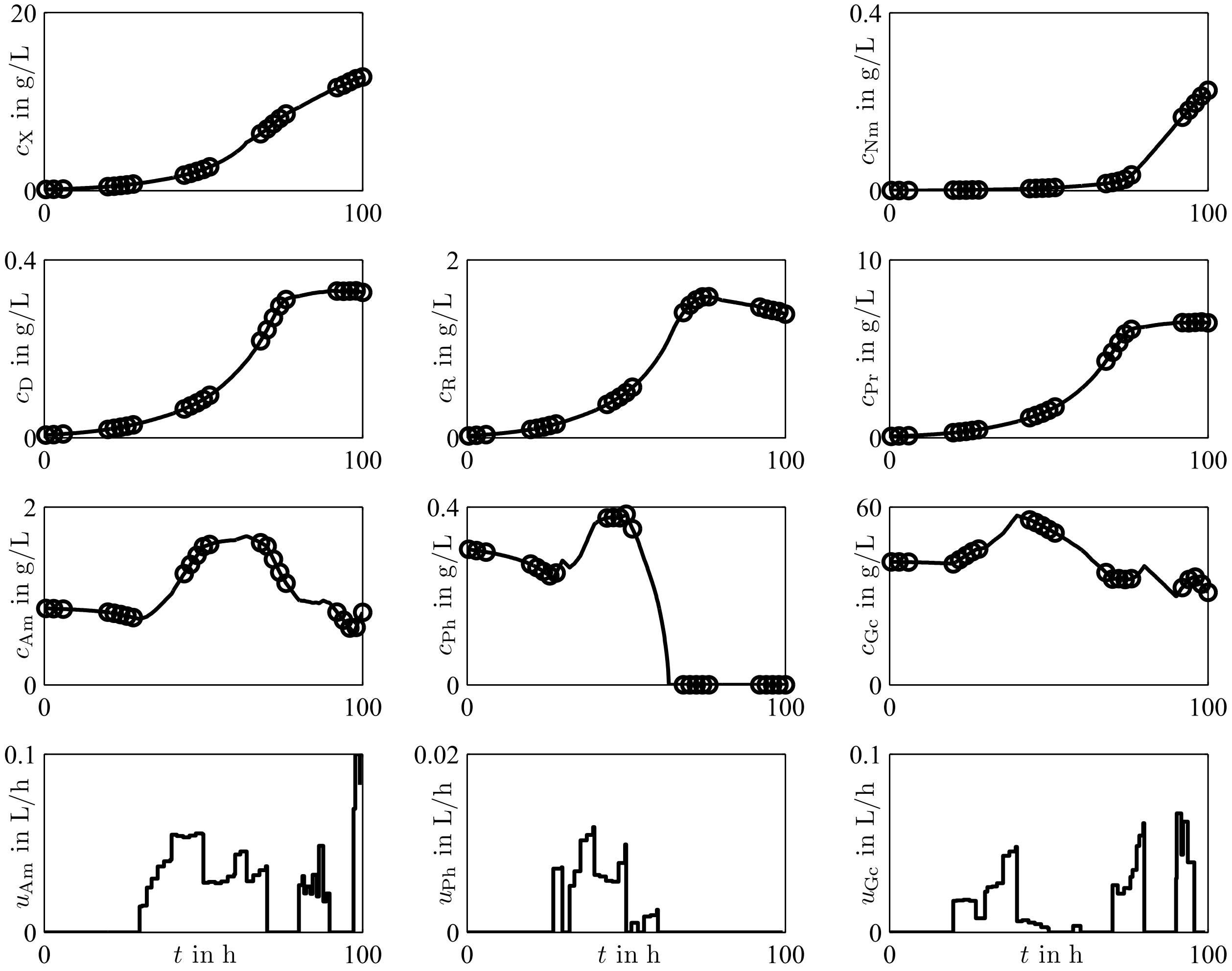


Figure C.3: Simulation of STdef3

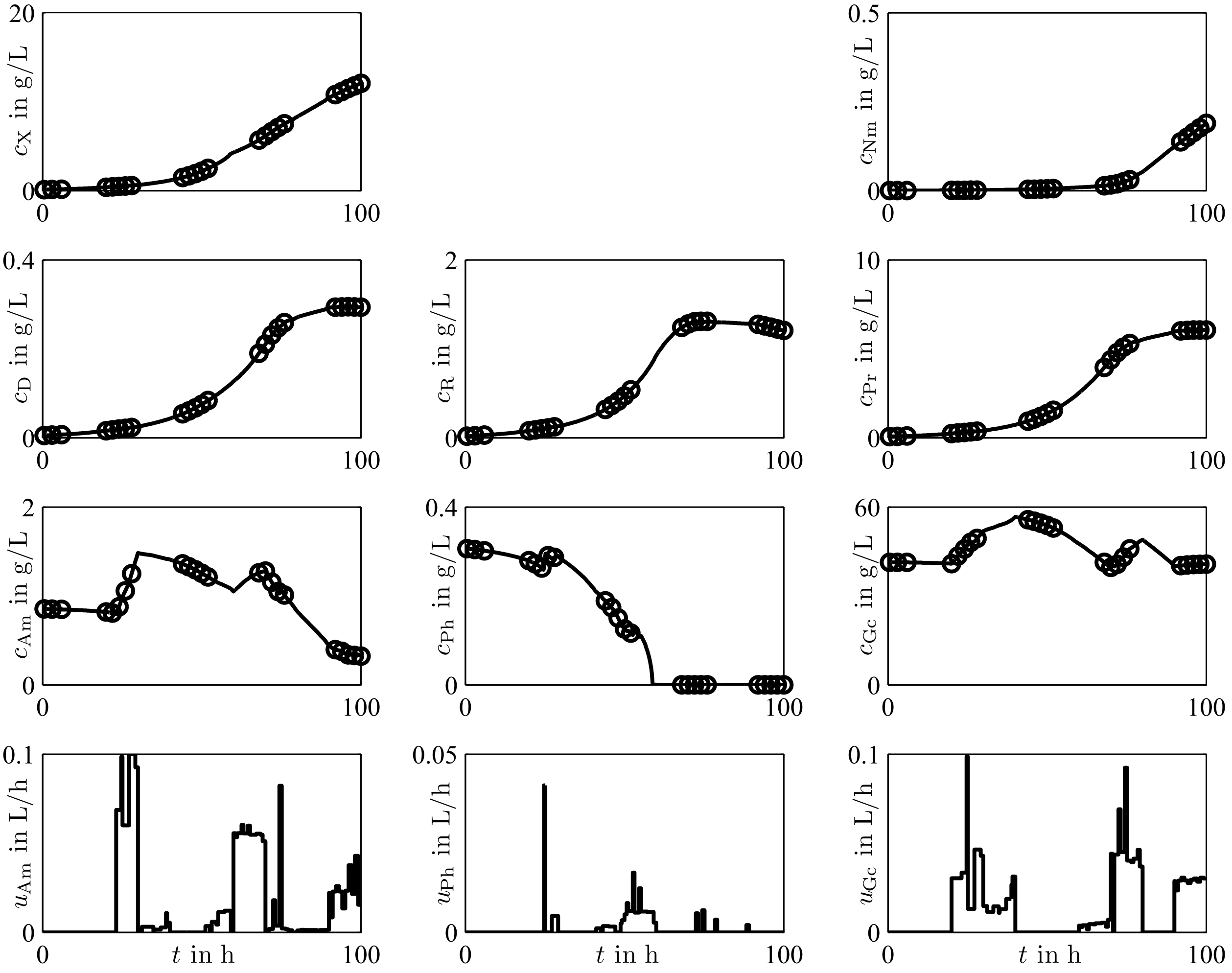


Figure C.4: Simulation of STdef4

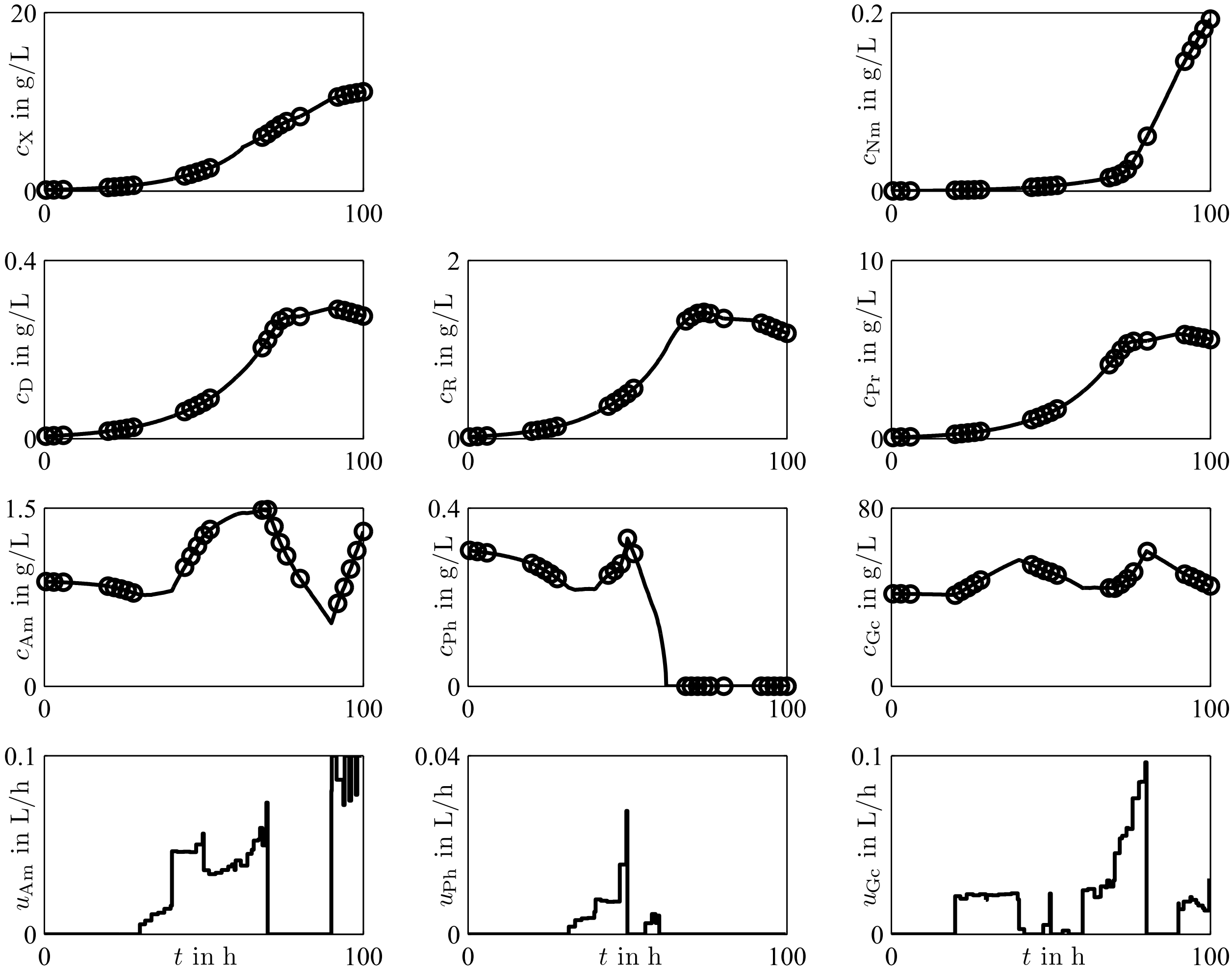


Figure C.5: Simulation of STdef5

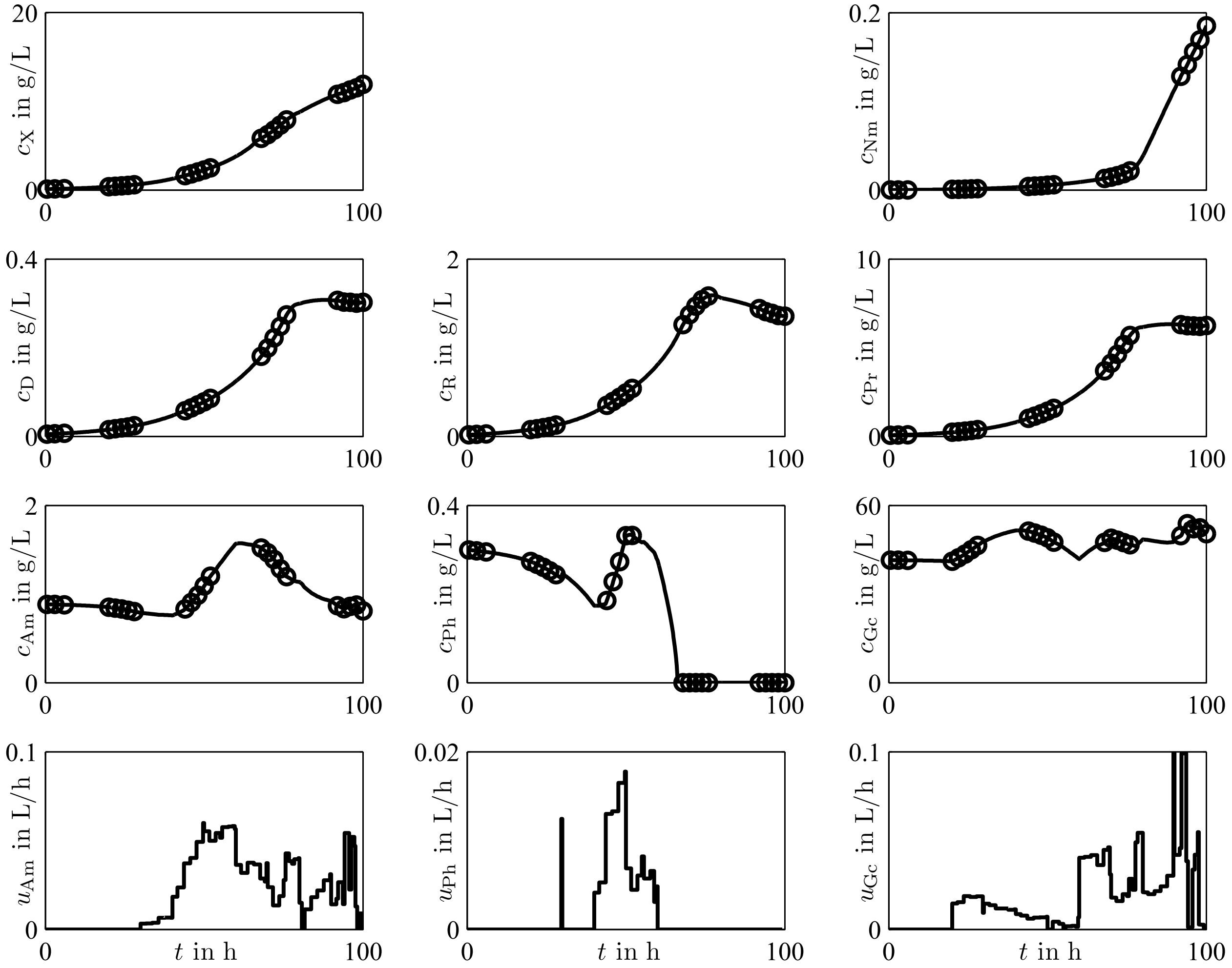


Figure C.6: Simulation of STdef6

C.3 Standard Deviation of the Measurement Noise

Table C.2: Standard deviation of the noise for measurements of *S. tendae*. For each measurement variable, the standard deviation is linearly approximated.

c_i	σ_i
c_{X}	$\sigma_{\mathrm{X}} = 0.25/12 \cdot c_{\mathrm{X}} + 0.05\,\mathrm{g/L}$
c_{Am}	$\sigma_{\mathrm{Am}} = 0.015/2 \cdot c_{\mathrm{Am}} + 0.003\,\mathrm{g/L}$
c_{Ph}	$\sigma_{\mathrm{Ph}} = 0.011/0.6 \cdot c_{\mathrm{Ph}} + 0.007\,\mathrm{g/L}$
c_{Gc}	$\sigma_{\mathrm{Gc}} = 0.5/40 \cdot c_{\mathrm{Gc}} + 0.25\,\mathrm{g/L}$
c_{D}	$\sigma_{\mathrm{D}} = 0.024/12 \cdot c_{\mathrm{X}} + 0.006\,\mathrm{g/L}$
c_{R}	$\sigma_{\mathrm{R}} = 0.04\,\mathrm{g/L}$
c_{Pr}	$\sigma_{\mathrm{Pr}} = 0.12/14 \cdot c_{\mathrm{X}} + 0.02\,\mathrm{g/L}$
c_{Nm}	$\sigma_{\mathrm{Nm}} = 0.05\,\mathrm{g/L}$

C.4 Results of the Phenomena Detection

Table C.3: Complete list of detected biological phenomena with measurements of *S. tendae*

Phenomenon	Sc
The growth is limited by phosphate.	−0.76
Ammonium and phosphate are consumed simultaneously.	−0.50
Ammonium and glucose are consumed simultaneously.	−0.67
Phosphate and glucose are consumed simultaneously.	−0.81
The product formation is limited by ammonium.	−0.75
The product formation is limited by phosphate.	−0.78
The product formation is inhibited by ammonium.	0.56
The product formation is inhibited by phosphate.	0.66
The product formation is inhibited by glucose.	0.07
Storage A for ammonium	0.50
Storage A for phosphate	0.50
The DNA formation is limited by ammonium.	−0.50
The RNA formation is limited by ammonium.	−0.50
The protein formation is limited by ammonium.	−0.50
The DNA formation is limited by phosphate.	−0.65
The RNA formation is limited by phosphate.	−0.24
The protein formation is limited by phosphate.	−0.38

To be continued on next page

Table C.3: Detected biological phenomena of *S. tendae* – continued

Phenomenon	Sc
The DNA formation is inhibited by phosphate.	−0.04
The RNA formation is inhibited by phosphate.	0.03
The DNA formation is inhibited by glucose.	−0.05
The DNA degradation is inhibited by ammonium.	−0.72
The RNA degradation is inhibited by ammonium.	−0.76
The protein degradation is inhibited by ammonium	−0.67
The DNA degradation is inhibited by phosphate.	−0.37
The RNA degradation is inhibited by phosphate.	−0.64
The protein degradation is inhibited by phosphate.	−0.60
The DNA degradation is inhibited by glucose.	−0.43
The RNA degradation is inhibited by glucose.	−0.23
The protein degradation is inhibited by glucose.	−0.34
Intermediate compartment between ammonium and DNA	−0.50
Intermediate compartment between phosphate and DNA	0.58
Intermediate compartment between phosphate and RNA	0.25
Intermediate compartment between phosphate and proteins	0.67
The RNA formation is limited by DNA.	−0.03
The protein formation is limited by RNA.	−0.50
The DNA formation is limited by proteins.	0.05
DNA and RNA grow simultaneously.	−0.63
DNA and proteins grow simultaneously.	−0.25
RNA and proteins grow simultaneously.	−0.86
The product formation is limited by RNA.	−0.15
The product formation is inhibited by DNA.	0.06
The product and DNA grow simultaneously.	−0.43
The product and RNA grow simultaneously.	−0.92
The product and proteins grow simultaneously.	−0.64
Degradation of biomass	0.50
Degradation of DNA	0.50
Degradation of the product	0.50

C.5 Additional Identified Experiments

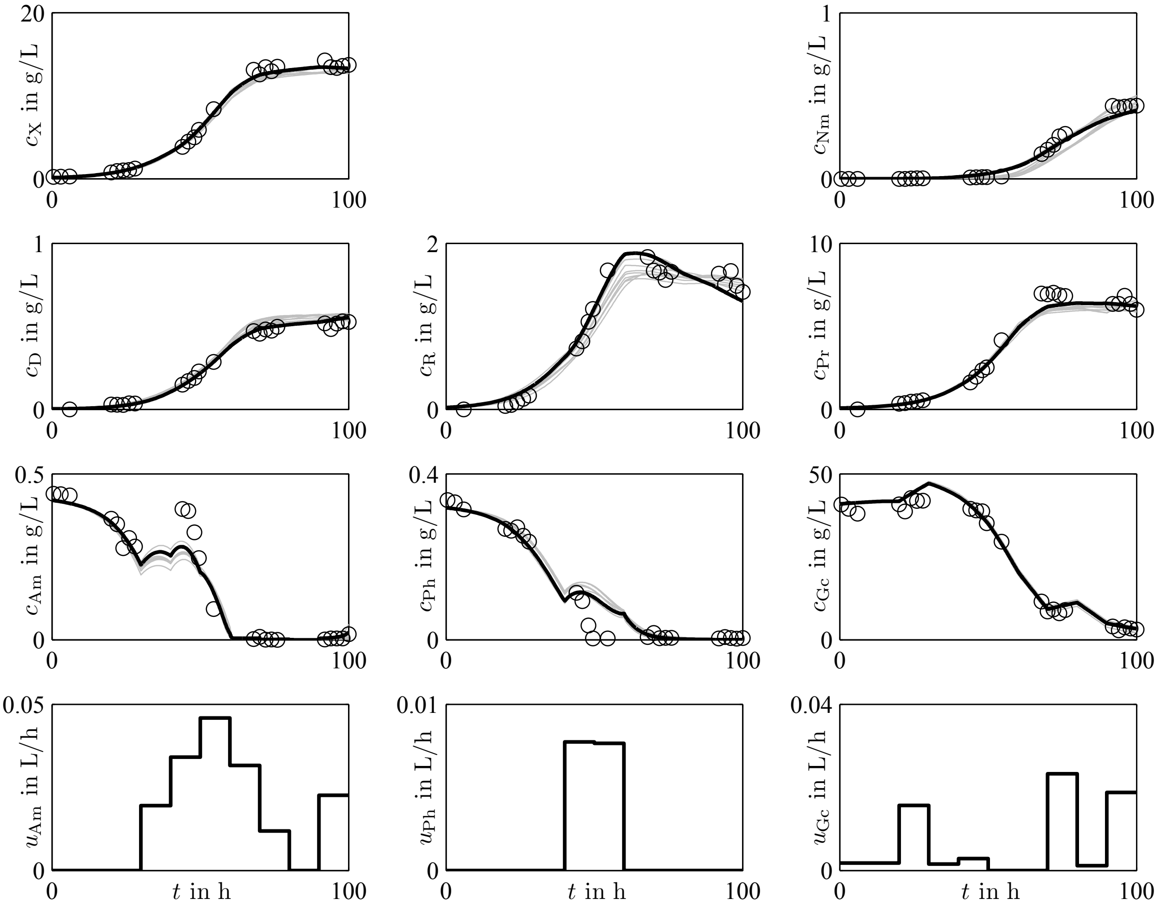


Figure C.7: Identified experiment STdef1 for *S. tendae*

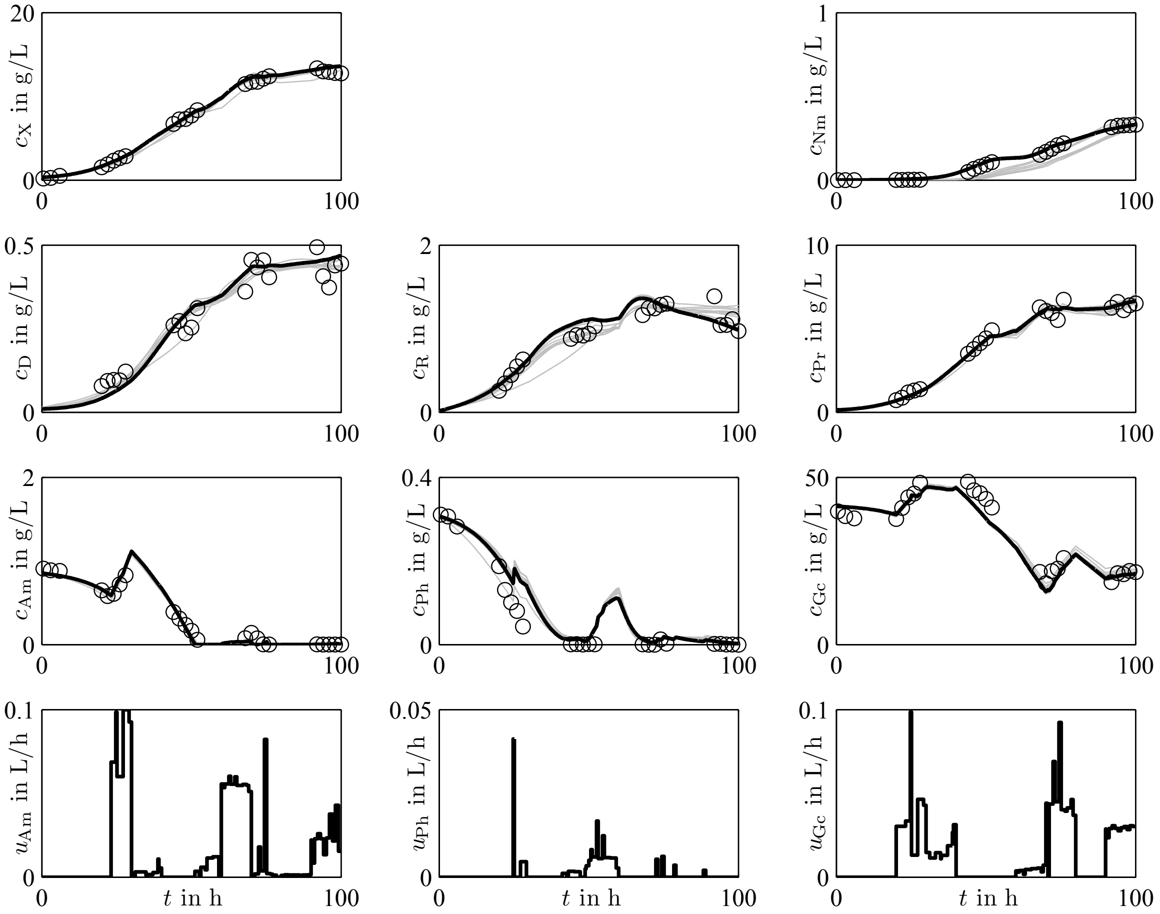


Figure C.8: Identified experiment STdef4 for *S. tendae*

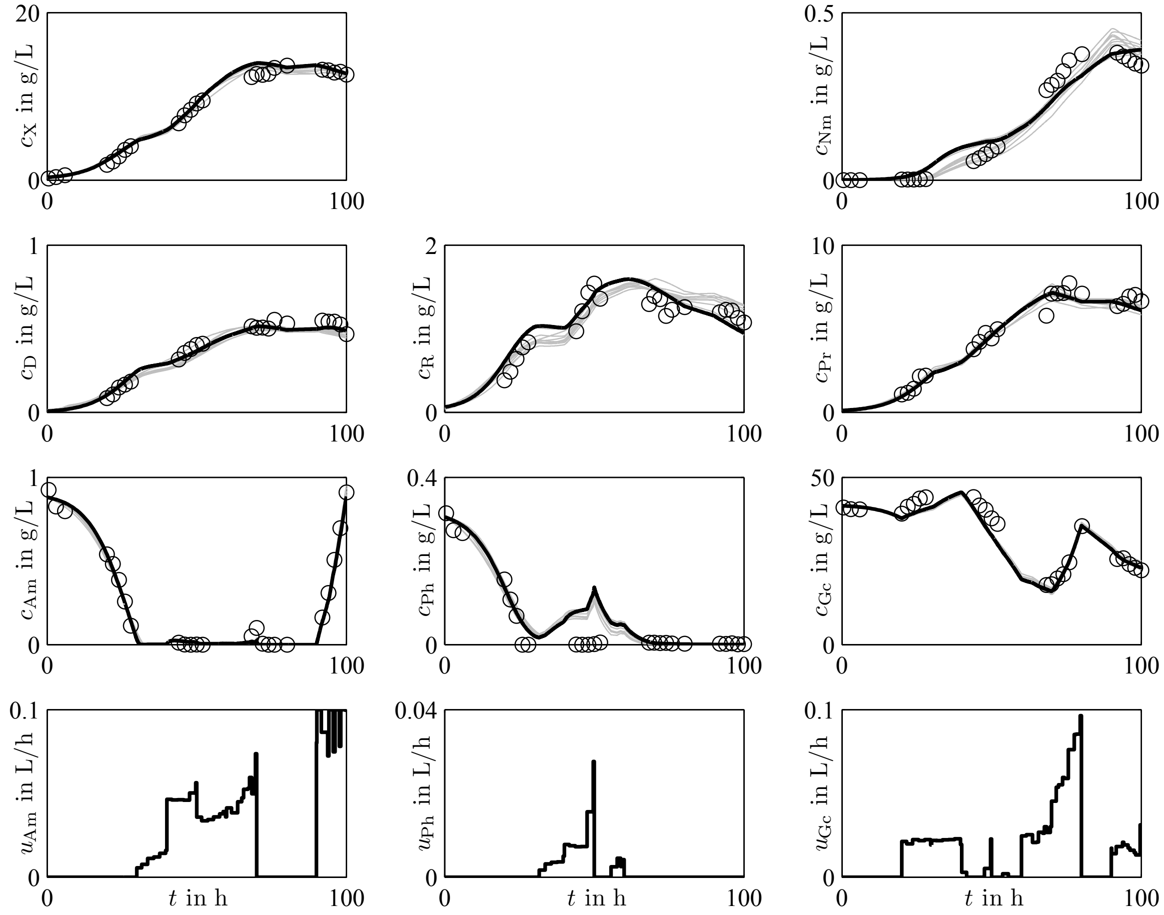


Figure C.9: Identified experiment STdef5 for *S. tendae*

C.6 Additional Validation Experiment

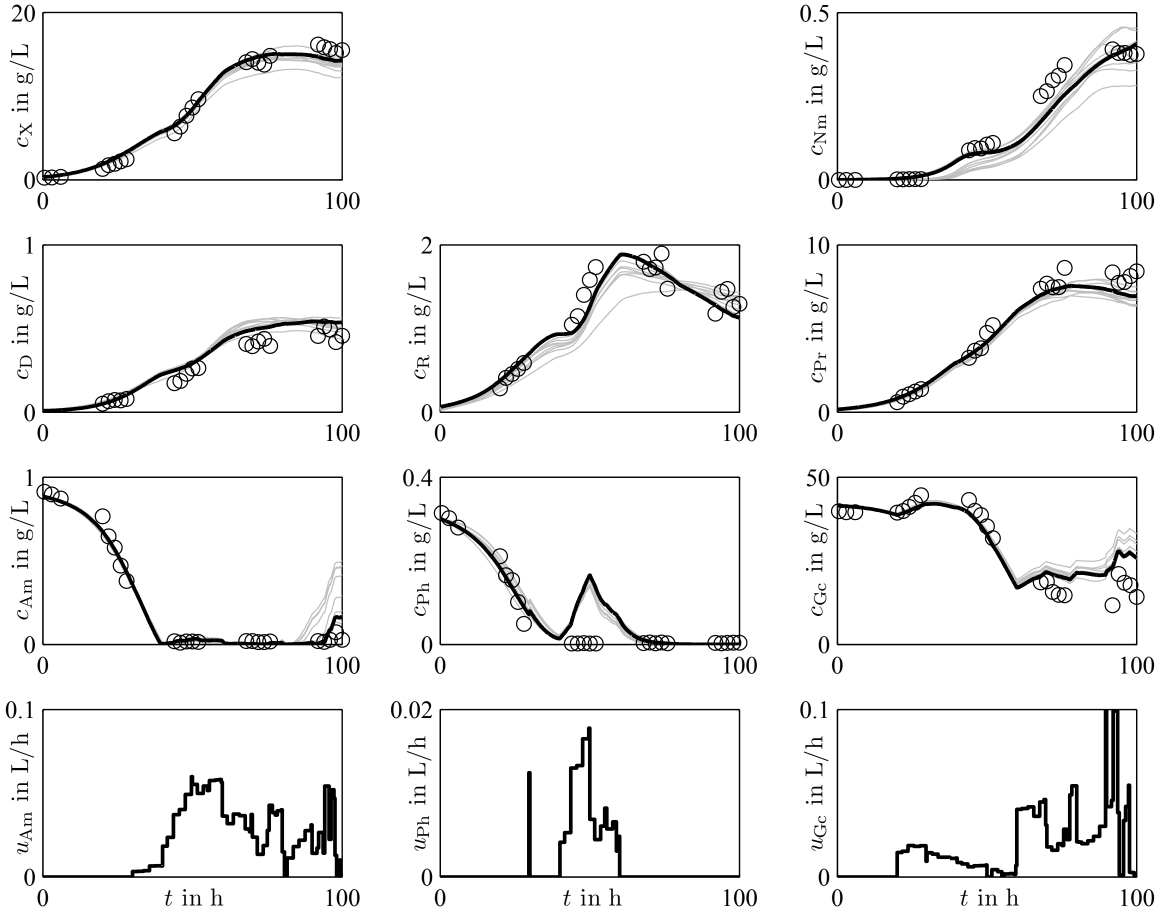


Figure C.10: Validation experiment STdef6 for *S. tendae*

C.7 Best Validated Model

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{D}}(t) \\ m_{\mathrm{R}}(t) \\ m_{\mathrm{Pr}}(t) \\ m_{\mathrm{D}^*}(t) \\ m_{\mathrm{Pr}^*}(t) \\ m_{\mathrm{Xr}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix} = & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t)V(t) \end{pmatrix} + \\
& V_{\mathrm{X}}(t)\begin{pmatrix} Y_{\mathrm{D}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & Y_{\mathrm{R}} & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & Y_{\mathrm{RD}^*} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & Y_{\mathrm{RXr}} \\ -1 & 0 & 0 & -1 & 0 & 0 \\ -Y_{\mathrm{PhD}} & 0 & 0 & -Y_{\mathrm{PhR}} & 0 & 0 \\ -Y_{\mathrm{GcD}} & 0 & 0 & -Y_{\mathrm{GcR}} & 0 & 0 \end{pmatrix}\begin{pmatrix} r_{\mathrm{D}}(t) \\ r_{\mathrm{D^*D}}(t) \\ r_{\mathrm{dD}}(t) \\ r_{\mathrm{R}}(t) \\ r_{\mathrm{D^*R}}(t) \\ r_{\mathrm{dR}}(t) \end{pmatrix} + \\
& V_{\mathrm{X}}(t)\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ Y_{\mathrm{Pr}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & Y_{\mathrm{PrPr}^*} & 0 & 1 & 0 \\ 0 & 0 & Y_{\mathrm{PrXr}} & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -Y_{\mathrm{PhPr}} & 0 & 0 & -Y_{\mathrm{PhD}^*} & -Y_{\mathrm{PhPr}^*} & -Y_{\mathrm{PhXr}} \\ -Y_{\mathrm{GcPr}} & 0 & 0 & 0 & 0 & -Y_{\mathrm{GcXr}} \end{pmatrix}\begin{pmatrix} r_{\mathrm{Pr}}(t) \\ r_{\mathrm{Pr^*Pr}}(t) \\ r_{\mathrm{dPr}}(t) \\ r_{\mathrm{D}^*}(t) \\ r_{\mathrm{Pr}^*}(t) \\ r_{\mathrm{Xr}}(t) \end{pmatrix}
\end{aligned} \tag{C.16}
$$

$$\frac{\mathrm{d}m_{\mathrm{Nm}}(t)}{\mathrm{d}t} = r_{\mathrm{Nm}}(t) \cdot V_{\mathrm{X}}(t) + r_{\mathrm{dNm}}(t) \cdot V(t) \tag{C.17}$$

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t) \tag{C.18}$$

$$r_{\mathrm{D}}(t) = \mu_{\mathrm{Dm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{DAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{DPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{DGc}}} \cdot g_{\mathrm{D}}(t) \tag{C.19}$$

$$r_{\mathrm{D^*D}}(t) = \mu_{\mathrm{D^*Dm}} \cdot \frac{g_{\mathrm{D}^*}(t)}{g_{\mathrm{D}^*}(t) + K_{\mathrm{DD}^*}} \cdot g_{\mathrm{D}}(t) \tag{C.20}$$

$$r_{\mathrm{dD}}(t) = \mu_{\mathrm{dDm}} \cdot g_{\mathrm{D}}(t) \tag{C.21}$$

$$r_{\mathrm{R}}(t) = \mu_{\mathrm{Rm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{RAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{RPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{RGc}}} \cdot g_{\mathrm{D}}(t) \tag{C.22}$$

$$r_{\mathrm{D^*R}}(t) = \mu_{\mathrm{D^*Rm}} \cdot \frac{g_{\mathrm{D}^*}(t)}{g_{\mathrm{D}^*}(t) + K_{\mathrm{RD}^*}} \cdot g_{\mathrm{D}}(t) \tag{C.23}$$

Table C.4: Parameter values of the best validated dynamic model of *S. tendae*

	Parameter	Value	Unit
r_{D}	μ_{Dm}	0.174	1/h
	K_{DAm}	6.47×10^{-3}	g/L
	K_{DPh}	1×10^{-4}	g/L
	K_{DGc}	4.36×10^{-3}	g/L
$r_{\mathrm{D^*D}}$	$\mu_{\mathrm{D^*Dm}}$	0.150	1/h
	$K_{\mathrm{DD^*}}$	0.0582	g/L
r_{dD}	μ_{dDm}	2.68×10^{-3}	1/h
r_{R}	μ_{Rm}	0.718	1/h
	K_{RAm}	0.0272	g/L
	K_{RPh}	0.497	g/L
	K_{RGc}	5.79×10^{-4}	g/L
$r_{\mathrm{D^*R}}$	$\mu_{\mathrm{D^*Rm}}$	0.176	1/h
	$K_{\mathrm{RD^*}}$	3.82×10^{-3}	g/L
r_{dR}	μ_{dRm}	0.0305	1/h
r_{Pr}	μ_{Prm}	0.0151	1/h
	K_{PrAm}	1.61×10^{-3}	g/L
	K_{PrPh}	2.02×10^{-3}	g/L
	K_{PrGc}	9.522	g/L
$r_{\mathrm{Pr^*Pr}}$	$\mu_{\mathrm{Pr^*Prm}}$	0.0519	1/h
	$K_{\mathrm{PrPr^*}}$	0.204	g/L
r_{dPr}	μ_{dPrm}	0.0178	1/h
$r_{\mathrm{D^*}}$	$\mu_{\mathrm{D^*m}}$	0.237	1/h
	$K_{\mathrm{D^*Ph}}$	1.255	g/L
$r_{\mathrm{Pr^*}}$	$\mu_{\mathrm{Pr^*m}}$	0	1/h
	$K_{\mathrm{Pr^*Ph}}$	0.435	g/L

	Parameter	Value	Unit
r_{Xr}	μ_{Xrm}	0.104	1/h
	K_{XrPh}	0.331	g/L
	K_{XrGc}	0.0712	g/L
r_{Nm}	μ_{Nmm}	0.0435	1/h
	K_{NmPh}	0.0870	g/L
r_{dNm}	μ_{dNmm}	0.0428	1/h
r_{M}	μ_{Mm}	0.0512	1/h
	K_{M}	0.01	g/L
	Y_{D}	0.141	g/g
	Y_{R}	3.624	g/g
	Y_{Pr}	15.447	g/g
	$Y_{\mathrm{RD^*}}$	0.734	g/g
	$Y_{\mathrm{PrPr^*}}$	1.142	g/g
	Y_{RXr}	0.627	g/g
	Y_{PrXr}	0	g/g
	$Y_{\mathrm{AmD^*}}$	0	g/g
	$Y_{\mathrm{AmPr^*}}$	0.0753	g/g
	Y_{PhD}	0	g/g
	Y_{PhR}	0.639	g/g
	Y_{PhPr}	0	g/g
	$Y_{\mathrm{PhD^*}}$	0.418	g/g
	$Y_{\mathrm{PhPr^*}}$	0.0975	g/g
	Y_{PhXr}	0	g/g
	Y_{GcD}	10.473	g/g
	Y_{GcR}	3.518	g/g
	Y_{GcPr}	3.862	g/g
	Y_{GcXr}	2.510	g/g

$$r_{\mathrm{dR}}(t) = \mu_{\mathrm{dRm}} \cdot g_{\mathrm{R}}(t) \tag{C.24}$$

$$r_{\mathrm{Pr}}(t) = \mu_{\mathrm{Prm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{PrAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{PrGc}}} \cdot g_{\mathrm{R}}(t) \tag{C.25}$$

$$r_{\mathrm{Pr^*Pr}}(t) = \mu_{\mathrm{Pr^*Prm}} \cdot \frac{g_{\mathrm{Pr^*}}(t)}{g_{\mathrm{Pr^*}}(t) + K_{\mathrm{PrPr^*}}} \cdot g_{\mathrm{Pr}}(t) \tag{C.26}$$

$$r_{\mathrm{dPr}}(t) = \mu_{\mathrm{dPrm}} \cdot g_{\mathrm{Pr}}(t) \tag{C.27}$$

$$r_{\mathrm{D^*}}(t) = \mu_{\mathrm{D^*m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{D^*Ph}}} \cdot g_{\mathrm{D}}(t) \tag{C.28}$$

$$r_{\mathrm{Pr}^*}(t) = \mu_{\mathrm{Pr}^*\mathrm{m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Pr}^*\mathrm{Ph}}} \cdot g_{\mathrm{Pr}}(t) \tag{C.29}$$

$$r_{\mathrm{Xr}}(t) = \mu_{\mathrm{Xrm}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XrGc}}} \cdot g_{\mathrm{Pr}}(t) \tag{C.30}$$

$$r_{\mathrm{Nm}}(t) = \mu_{\mathrm{Nmm}} \cdot \frac{K_{\mathrm{NmPh}}}{c_{\mathrm{Ph}}(t) + K_{\mathrm{NmPh}}} \cdot g_{\mathrm{D}}(t) \tag{C.31}$$

$$r_{\mathrm{dNm}}(t) = \mu_{\mathrm{dNmm}} \cdot c_{\mathrm{Ni}}(t) \tag{C.32}$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \tag{C.33}$$

Appendix D

Supplementary Material for *Streptomyces griseus*

D.1 Results of the Phenomena Detection

Table D.1: Complete list of detected biological phenomena with measurements of *S. griseus*

Phenomenon	Sc
The growth is limited by ammonium.	−0.67
The growth is limited by phosphate.	−0.35
Ammonium and phosphate are consumed simultaneously.	−0.62
Ammonium and glucose are consumed simultaneously.	−0.43
Phosphate and glucose are consumed simultaneously.	−0.67
The product formation is limited by phosphate.	−0.20
The product formation is limited by glucose.	−0.23
The product formation is inhibited by ammonium.	0.07
The product formation is inhibited by phosphate.	0.24
The product formation is inhibited by glucose.	−0.18
Storage A for phosphate	0.62
Storage B for ammonium	−0.03
Storage B for glucose	−0.03
The DNA formation is limited by ammonium.	0.53
The protein formation is limited by ammonium.	−0.12
The DNA formation is limited by phosphate.	0.14
The RNA formation is limited by phosphate.	0.15
The protein formation is limited by phosphate.	−0.40
The DNA formation is inhibited by ammonium.	0.01
The DNA formation is inhibited by phosphate.	−0.16
The RNA formation is inhibited by phosphate.	0.09
The protein formation is inhibited by phosphate.	0.03
The DNA formation is inhibited by glucose.	0.00
The RNA formation is inhibited by glucose.	0.09

To be continued on next page

Table D.1: Detected biological phenomena of *S. griseus* – continued

Phenomenon	Sc
The protein formation is inhibited by glucose.	0.03
The DNA degradation is inhibited by ammonium.	−0.36
The RNA degradation is inhibited by ammonium.	−0.03
The protein degradation is inhibited by ammonium	−0.13
The DNA degradation is inhibited by phosphate.	−0.64
The RNA degradation is inhibited by phosphate.	−0.40
The protein degradation is inhibited by phosphate.	−0.51
The DNA degradation is inhibited by glucose.	−0.26
The RNA degradation is inhibited by glucose.	0.07
The protein degradation is inhibited by glucose.	−0.23
Intermediate compartment between ammonium and DNA	−0.30
Intermediate compartment between ammonium and RNA	−0.22
Intermediate compartment between phosphate and DNA	−0.16
Intermediate compartment between phosphate and RNA	0.15
Intermediate compartment between phosphate and proteins	0.52
The protein formation is limited by DNA.	−0.10
The protein formation is limited by RNA.	−0.54
The DNA formation is inhibited by RNA.	0.04
The DNA formation is inhibited by the proteins.	0.05
DNA and RNA grow simultaneously.	0.32
DNA and proteins grow simultaneously.	0.17
RNA and proteins grow simultaneously.	−0.41
DNA and proteins are degraded simultaneously.	−0.14
RNA and proteins are degraded simultaneously.	−0.50
The product formation is inhibited by DNA.	0.04
The product formation is inhibited by RNA.	−0.19
The product formation is inhibited by proteins.	0.07
The product and DNA grow simultaneously.	−0.51
The product and RNA grow simultaneously.	−0.78
The product and proteins grow simultaneously.	0.00
Degradation of biomass	0.50
Degradation of DNA	0.64
Degradation of the product	0.66

D.2 Standard Deviation of the Measurement Noise

Table D.2: Standard deviation of the noise for measurements of *S. griseus*. For each measurement variable, the standard deviation is linearly approximated.

c_i	σ_i
c_{X}	$\sigma_{\mathrm{X}} = 0.25/12 \cdot c_{\mathrm{X}} + 0.05\,\mathrm{g/L}$
c_{Am}	$\sigma_{\mathrm{Am}} = 0.015/2 \cdot c_{\mathrm{Am}} + 0.003\,\mathrm{g/L}$
c_{Ph}	$\sigma_{\mathrm{Ph}} = 0.011/0.6 \cdot c_{\mathrm{Ph}} + 0.007\,\mathrm{g/L}$
c_{Gc}	$\sigma_{\mathrm{Gc}} = 0.5/40 \cdot c_{\mathrm{Gc}} + 0.25\,\mathrm{g/L}$
c_{D}	$\sigma_{\mathrm{D}} = 0.024/12 \cdot c_{\mathrm{X}} + 0.006\,\mathrm{g/L}$
c_{R}	$\sigma_{\mathrm{R}} = 0.04\,\mathrm{g/L}$
c_{Pr}	$\sigma_{\mathrm{Pr}} = 0.12/14 \cdot c_{\mathrm{X}} + 0.02\,\mathrm{g/L}$
c_{Sm}	$\sigma_{\mathrm{Sm}} = 0.01\,\mathrm{g/L}$

D.3 Additional Identified Experiments

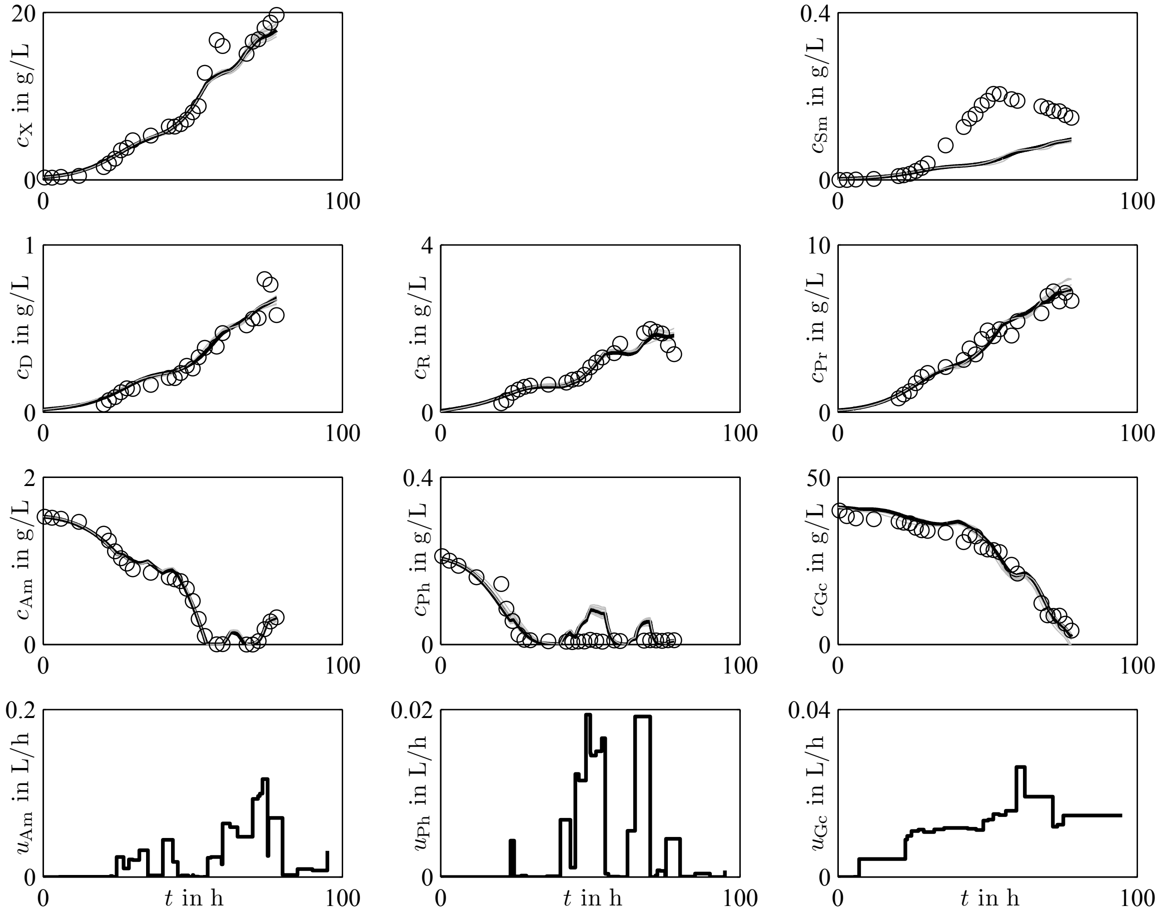


Figure D.1: Identified experiment SGdef30 for *S. griseus*

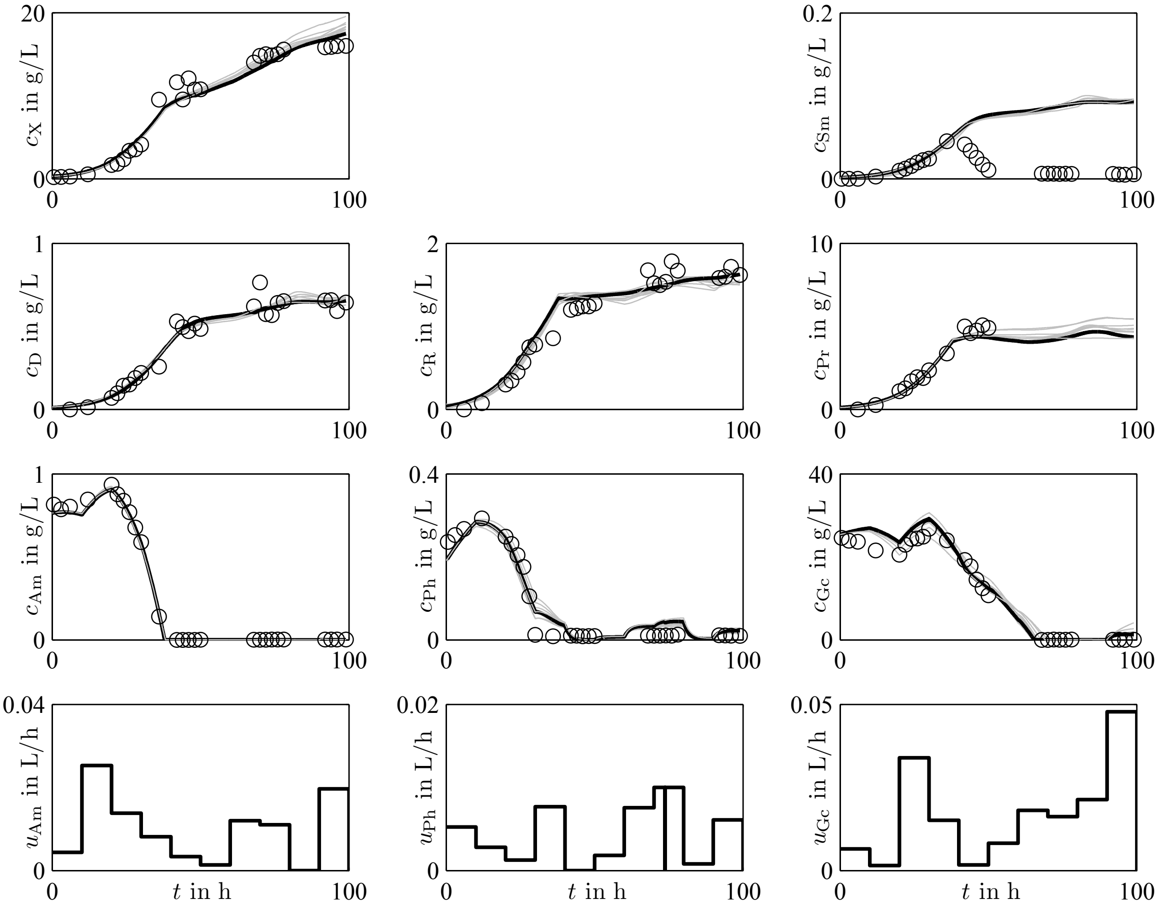


Figure D.2: Identified experiment SGdef31 for *S. griseus*

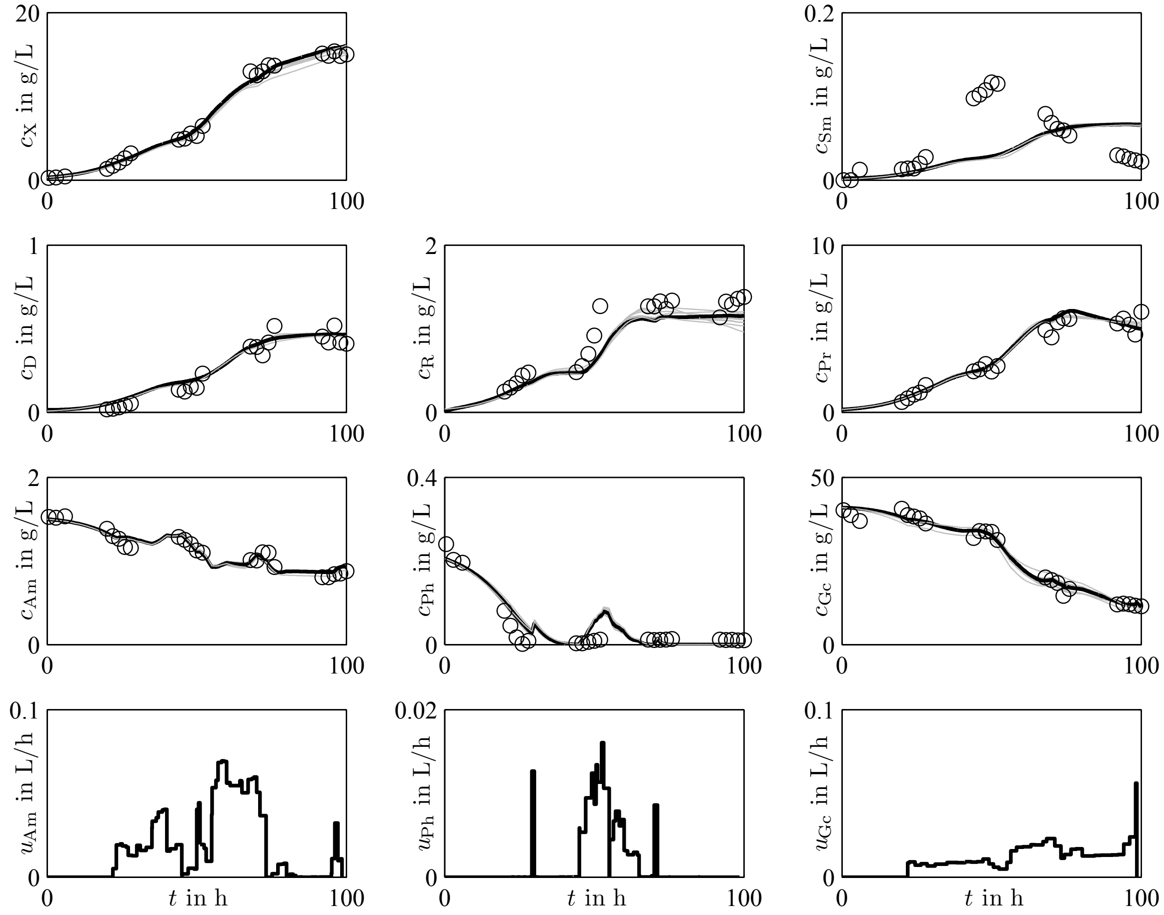


Figure D.3: Identified experiment SGdef33 for *S. griseus*

D.4 Additional Validation Experiments

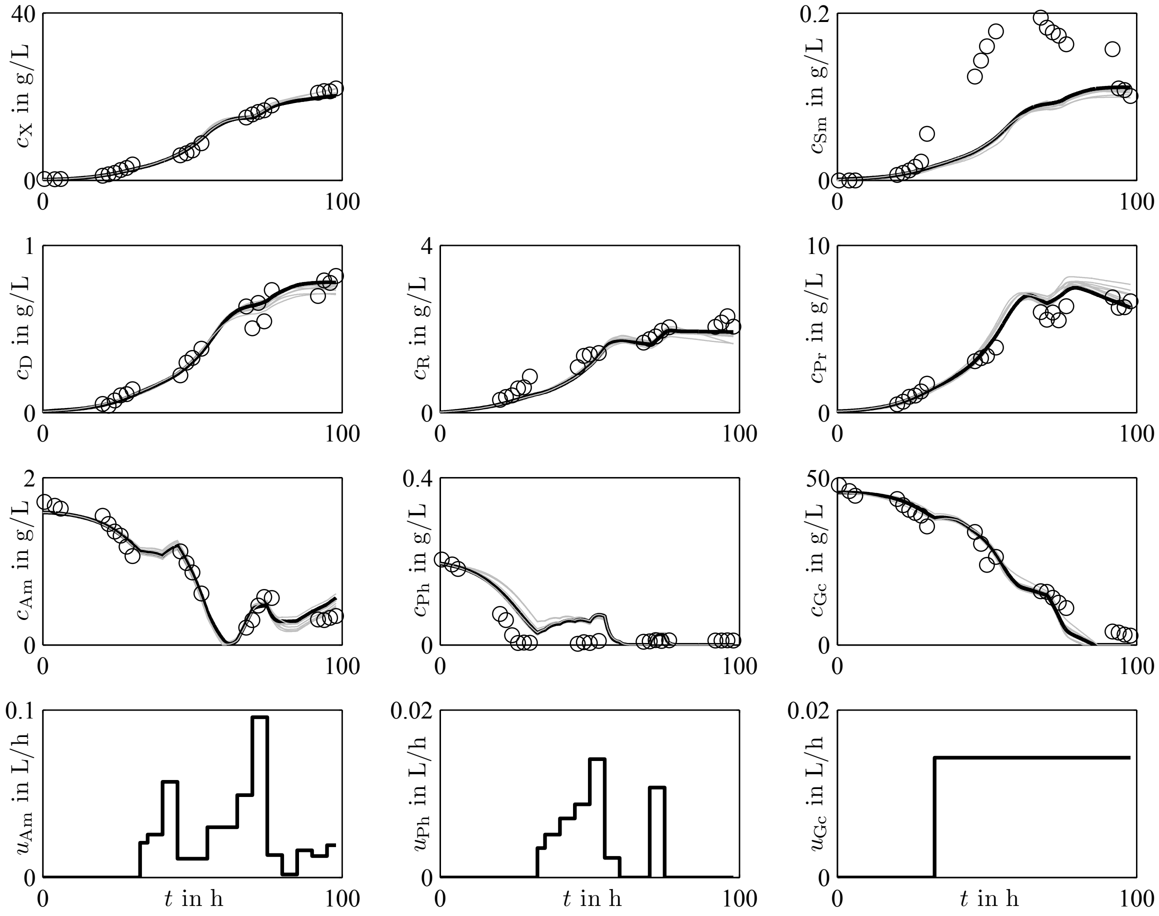


Figure D.4: Validation experiment SGdef29 for *S. griseus*

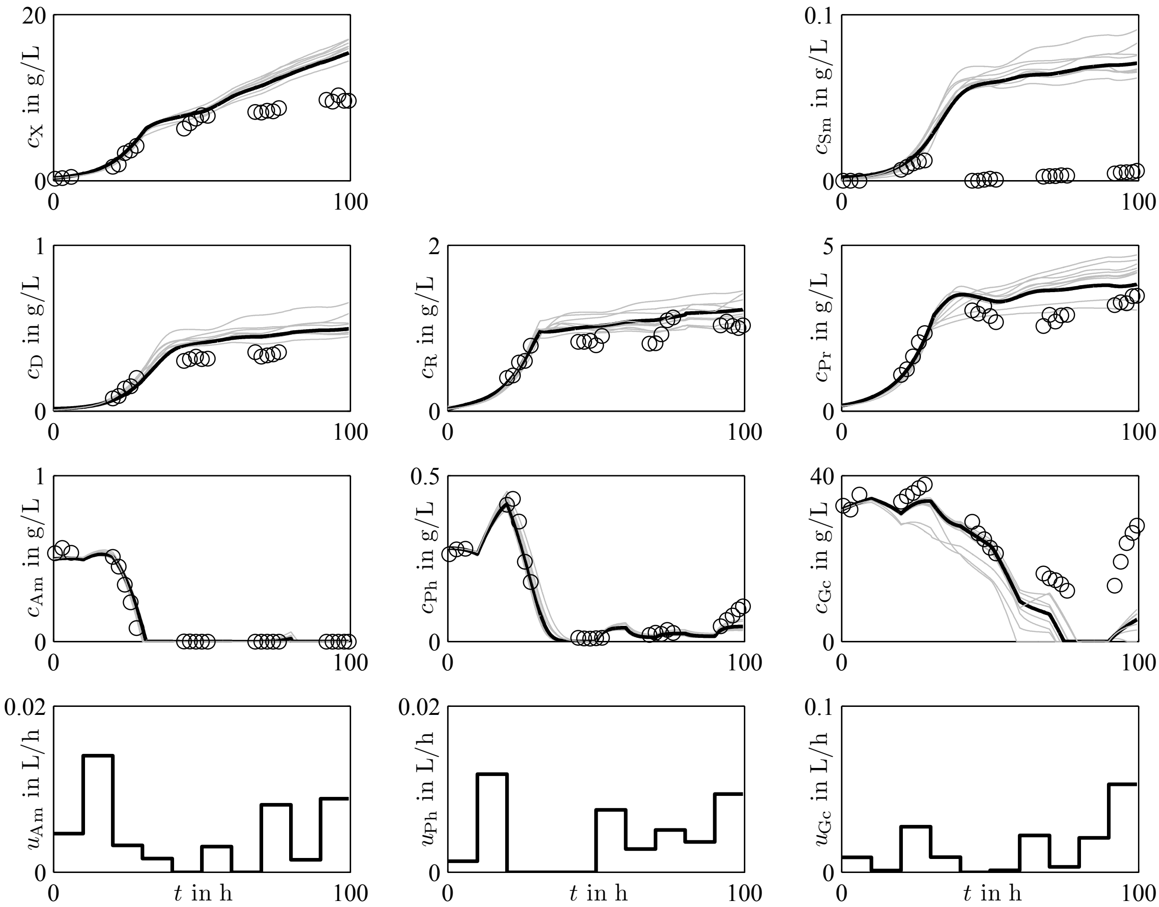


Figure D.5: Validation experiment SGdef34 for *S. griseus*

D.5 Best Validated Structured Model

$$
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_{\mathrm{D}}(t) \\ m_{\mathrm{R}}(t) \\ m_{\mathrm{Pr}}(t) \\ m_{\mathrm{D}^*}(t) \\ m_{\mathrm{Pr}^*}(t) \\ m_{\mathrm{Xr}}(t) \\ m_{\mathrm{Am}}(t) \\ m_{\mathrm{Ph}}(t) \\ m_{\mathrm{Gc}}(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_{\mathrm{Am,\,in}} \cdot u_{\mathrm{Am}}(t) \\ c_{\mathrm{Ph,\,in}} \cdot u_{\mathrm{Ph}}(t) \\ c_{\mathrm{Gc,\,in}} \cdot u_{\mathrm{Gc}}(t) \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ r_{\mathrm{M}}(t)V(t) \end{pmatrix} +
$$

$$
V_{\mathrm{X}}(t)\begin{pmatrix} Y_{\mathrm{D}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & Y_{\mathrm{R}} & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & Y_{\mathrm{RD}^*} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & Y_{\mathrm{RXr}} \\ -1 & 0 & 0 & -1 & 0 & 0 \\ -Y_{\mathrm{PhD}} & 0 & 0 & -Y_{\mathrm{PhR}} & 0 & 0 \\ -Y_{\mathrm{GcD}} & 0 & 0 & -Y_{\mathrm{GcR}} & 0 & 0 \end{pmatrix}\begin{pmatrix} r_{\mathrm{D}}(t) \\ r_{\mathrm{D}^*\mathrm{D}}(t) \\ r_{\mathrm{dD}}(t) \\ r_{\mathrm{R}}(t) \\ r_{\mathrm{D}^*\mathrm{R}}(t) \\ r_{\mathrm{dR}}(t) \end{pmatrix} + \tag{D.1}
$$

$$
V_{\mathrm{X}}(t)\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ Y_{\mathrm{Pr}} & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & Y_{\mathrm{PrPr}^*} & 0 & 1 & 0 \\ 0 & 0 & Y_{\mathrm{PrXr}} & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -Y_{\mathrm{PhPr}} & 0 & 0 & -Y_{\mathrm{PhD}^*} & -Y_{\mathrm{PhPr}^*} & -Y_{\mathrm{PhXr}} \\ -Y_{\mathrm{GcPr}} & 0 & 0 & 0 & 0 & -Y_{\mathrm{GcXr}} \end{pmatrix}\begin{pmatrix} r_{\mathrm{Pr}}(t) \\ r_{\mathrm{Pr}^*\mathrm{Pr}}(t) \\ r_{\mathrm{dPr}}(t) \\ r_{\mathrm{D}^*}(t) \\ r_{\mathrm{Pr}^*}(t) \\ r_{\mathrm{Xr}}(t) \end{pmatrix}
$$

$$
\frac{\mathrm{d}m_{\mathrm{Sm}}(t)}{\mathrm{d}t} = r_{\mathrm{Sm}}(t) \cdot V_{\mathrm{X}}(t) - r_{\mathrm{Sm}}(t) \cdot V(t) \tag{D.2}
$$

$$
\frac{\mathrm{d}V(t)}{\mathrm{d}t} = u_{\mathrm{Am}}(t) + u_{\mathrm{Ph}}(t) + u_{\mathrm{Gc}}(t) \tag{D.3}
$$

$$
r_{\mathrm{D}}(t) = \mu_{\mathrm{Dm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{DAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{DPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{DGc}}} \cdot g_{\mathrm{D}}(t) \tag{D.4}
$$

$$
r_{\mathrm{D}^*\mathrm{D}}(t) = \mu_{\mathrm{D}^*\mathrm{Dm}} \cdot \frac{g_{\mathrm{D}^*}(t)}{g_{\mathrm{D}^*}(t) + K_{\mathrm{DD}^*}} \cdot g_{\mathrm{D}}(t) \tag{D.5}
$$

$$
r_{\mathrm{dD}}(t) = \mu_{\mathrm{dDm}} \cdot g_{\mathrm{D}}(t) \tag{D.6}
$$

$$
r_{\mathrm{R}}(t) = \mu_{\mathrm{Rm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{RAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{RPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{RGc}}} \cdot g_{\mathrm{D}}(t) \tag{D.7}
$$

$$
r_{\mathrm{D}^*\mathrm{R}}(t) = \mu_{\mathrm{D}^*\mathrm{Rm}} \cdot \frac{g_{\mathrm{D}^*}(t)}{g_{\mathrm{D}^*}(t) + K_{\mathrm{RD}^*}} \cdot g_{\mathrm{D}}(t) \tag{D.8}
$$

Table D.3: Parameter values of the best validated model for *S. griseus*

	Parameter	Value	Unit
r_{D}	μ_{Dm}	0.118	1/h
	K_{DAm}	1×10^{-4}	g/L
	K_{DPh}	5.91×10^{-3}	g/L
	K_{DGc}	1.03×10^{-4}	g/L
$r_{\mathrm{D^*D}}$	$\mu_{\mathrm{D^*Dm}}$	0.328	1/h
	$K_{\mathrm{DD^*}}$	0.0487	g/L
r_{dD}	μ_{dDm}	0.0774	1/h
r_{R}	μ_{Rm}	0.104	1/h
	K_{RAm}	3.89×10^{-4}	g/L
	K_{RPh}	0.121	g/L
	K_{RGc}	3.56×10^{-4}	g/L
$r_{\mathrm{D^*R}}$	$\mu_{\mathrm{D^*Rm}}$	0.199	1/h
	$K_{\mathrm{RD^*}}$	1×10^{-4}	g/L
r_{dR}	μ_{dRm}	0.105	1/h
	K_{dR}	0.0440	g/L
r_{Pr}	μ_{Prm}	0.0535	1/h
	K_{PrAm}	2.28×10^{-3}	g/L
	K_{PrPh}	1×10^{-4}	g/L
	K_{PrGc}	7.404	g/L
$r_{\mathrm{Pr^*Pr}}$	$\mu_{\mathrm{Pr^*Prm}}$	0.0945	1/h
	$K_{\mathrm{PrPr^*}}$	0.0729	g/L
r_{dPr}	μ_{dPrm}	0.0167	1/h
$r_{\mathrm{D^*}}$	$\mu_{\mathrm{D^*m}}$	0.0822	1/h
	$K_{\mathrm{D^*Ph}}$	0.393	g/L

	Parameter	Value	Unit
$r_{\mathrm{Pr^*}}$	$\mu_{\mathrm{Pr^*m}}$	0.119	1/h
	$K_{\mathrm{Pr^*Ph}}$	0.168	g/L
r_{Xr}	μ_{Xrm}	0.0522	1/h
	K_{XrPh}	0.0128	g/L
	K_{XrGc}	1.18×10^{-4}	g/L
r_{Sm}	μ_{Smm}	1.372	1/h
r_{dSm}	μ_{dSmm}	9.602	1/h
r_{M}	μ_{Mm}	0.0587	1/h
	K_{M}	0.01	g/L
	Y_{D}	0.136	g/g
	Y_{R}	7.769	g/g
	Y_{Pr}	3.596	g/g
	$Y_{\mathrm{RD^*}}$	1.458	g/g
	$Y_{\mathrm{PrPr^*}}$	0.560	g/g
	Y_{RXr}	0.0403	g/g
	Y_{PrXr}	1.659	g/g
	Y_{PhD}	0	g/g
	Y_{PhR}	0	g/g
	Y_{PhPr}	0	g/g
	$Y_{\mathrm{PhD^*}}$	0	g/g
	$Y_{\mathrm{PhPr^*}}$	0.240	g/g
	Y_{PhXr}	0	g/g
	Y_{GcD}	1.13×10^{-3}	g/g
	Y_{GcR}	0.122	g/g
	Y_{GcPr}	0	g/g
	Y_{GcXr}	22.402	g/g

$$r_{\mathrm{dR}}(t) = \mu_{\mathrm{dRm}} \cdot \frac{K_{\mathrm{dR}}}{g_{\mathrm{D^*}}(t) + K_{\mathrm{dR}}} \cdot g_{\mathrm{R}}(t) \tag{D.9}$$

$$r_{\mathrm{Pr}}(t) = \mu_{\mathrm{Prm}} \cdot \frac{c_{\mathrm{Am}}(t)}{c_{\mathrm{Am}}(t) + K_{\mathrm{PrAm}}} \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{PrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{PrGc}}} \cdot g_{\mathrm{R}}(t) \tag{D.10}$$

$$r_{\mathrm{Pr^*Pr}}(t) = \mu_{\mathrm{Pr^*Prm}} \cdot \frac{g_{\mathrm{Pr^*}}(t)}{g_{\mathrm{Pr^*}}(t) + K_{\mathrm{PrPr^*}}} \cdot g_{\mathrm{Pr}}(t) \tag{D.11}$$

$$r_{\mathrm{dPr}}(t) = \mu_{\mathrm{dPrm}} \cdot g_{\mathrm{Pr}}(t) \tag{D.12}$$

$$r_{\mathrm{D^*}}(t) = \mu_{\mathrm{D^*m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{D^*Ph}}} \cdot g_{\mathrm{R}}(t) \tag{D.13}$$

$$r_{\mathrm{Pr^*}}(t) = \mu_{\mathrm{Pr^*m}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{Pr^*Ph}}} \cdot g_{\mathrm{Pr}}(t) \tag{D.14}$$

$$r_{\mathrm{Xr}}(t) = \mu_{\mathrm{Xrm}} \cdot \frac{c_{\mathrm{Ph}}(t)}{c_{\mathrm{Ph}}(t) + K_{\mathrm{XrPh}}} \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{XrGc}}} \cdot g_{\mathrm{R}}(t) \tag{D.15}$$

$$r_{\mathrm{Sm}}(t) = \mu_{\mathrm{Smm}} \cdot g_{\mathrm{D}}(t) \tag{D.16}$$

$$r_{\mathrm{dSm}}(t) = \mu_{\mathrm{dSmm}} \cdot c_{\mathrm{Sm}}(t) \tag{D.17}$$

$$r_{\mathrm{M}}(t) = \mu_{\mathrm{Mm}} \cdot \frac{c_{\mathrm{Gc}}(t)}{c_{\mathrm{Gc}}(t) + K_{\mathrm{M}}} \cdot c_{\mathrm{X}}(t) \tag{D.18}$$

Bibliography

Aiba, S., Shoda, M., and Nagatani, M. Kinetics of product inhibition in alcohol fermentation. *Biotechnology and Bioengineering*, 10:845–864, 1968. DOI: 10.1002/(SICI)1097-0290(20000320)67:6<671::AID-BIT6>3.0.CO;2-W.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC 19(6):716–723, 1974. DOI: 10.1109/TAC.1974.1100705.

Akima, H. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the Association for Computing Machinery*, 17(4):589–602, 1970. DOI: 10.1145/321607.321609.

Allenby, N. E. E., Laing, E., Bucca, G., and Kierzek, A. M. Diverse control of metabolism and other cellular processes in *Streptomyces coelicolor* by the PhoP transcription factor: genome-wide identification of *in vivo* targets. *Nucleic Acids Research*, 40(19):9543–9556, 2012. DOI: 10.1093/nar/gks766.

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., and Barabási, A.-L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427:839–843, 2004. DOI: 10.1038/nature02289.

Alves, R., Antunes, F., and Salvador, A. Tools for kinetic modeling of biochemical networks. *Nature Biotechnology*, 24(6):667–672, 2006. DOI: 10.1038/nbt0606-667.

Bailey, J. E. and Ollis, D. F. *Biochemical Engineering Fundamentals*. Chemical Engineering Series. McGraw-Hill, second edition, 1986.

Bajpai, R. K. and Reuß, M. A mechanistic model for penicillin production. *Journal of Chemical Technology and Biotechnology*, 30(1):332–344, 1980. DOI: 10.1002/jctb.503300140.

Bapat, P. M., Bhartiya, S., Venkatesh, K. V., and Wangikar, P. P. Structured kinetic model to represent the utilization of multiple substrates in complex media during rifamycin b fermentation. *Biotechnology and Bioengineering*, 93(4):779–790, 2006. DOI: 10.1002/bit.20767.

Bernard, O. and Bastin, G. On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Mathematical Biosciences*, 193:51–77, 2005. DOI: 10.1016/j.mbs.2004.10.004.

biotechnologie.de. The German Biotechnology Sector 2013. Online document, 2013. URL `http://www.biotechnologie.de/BIO/Redaktion/PDF/de/umfrage/2013-umfrage,property=pdf,bereich=bio,sprache=de,rwb=true.pdf`.

Bisswanger, H. *Enzyme Kinetics: Principles and Methods*. Wiley-VCH, second edition, 2008. DOI: 10.1002/9783527622023.

de Boor, C. *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Springer, revised edition, 2001.

Büdenbender, C. *Modellentwicklung und Trajektorienplanung für Fed-Batch-Fermentationen mit komplexen Nährmedien*. PhD thesis, Technische Universität Berlin, 2004.

Burnham, K. P. and Anderson, D. R. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2002.

Çelik, E., Çalık, P., and Oliver, S. G. A structured kinetic model for recombinant protein production by Mut^{+} strain of *Pichia pastoris*. *Chemical Engineering Science*, 64:5028–5035, 2009. DOI: 10.1016/j.ces.2009.08.009.

Cheung, J. T.-Y. and Stephanopoulos, G. Representation of process trends—Part I. A formal representation framework. *Computers and Chemical Engineering*, 14(4/5): 495–510, 1990. DOI: 10.1016/0098-1354(90)87023-I.

Chmiel, H., editor. *Bioprozesstechnik*. Spektrum Akademischer Verlag, 2006.

Cho, D.-Y., Cho, K.-H., and Zhang, B.-T. Identification of biochemical networks by S-tree based genetic programming. *Bioinformatics*, 22(13):1631–1640, 2006. DOI: 10.1093/bioinformatics/btl122.

Clementschitsch, F. and Bayer, K. Improvement of bioprocess monitoring: development of novel concepts. *Microbial Cell Factories*, 5:19, 2006. DOI: 10.1186/1475-2859-5-19.

Clewley, R. Hybrid models and biological model reduction with PyDSTool. *PLoS Computational Biology*, 8(8):e1002628, 2012. DOI: 10.1371/journal.pcbi.1002628.

Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics*. Portland Press, London, third edition, 2004.

Crampin, E. J., Schnell, S., and McSharry, P. E. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics and Molecular Biology*, 86(1):77–112, 2004. DOI: 10.1016/j.pbiomolbio.2004.04.002.

Currie, J. N. The citric acid fermentation of *Aspergillus niger*. *Journal of Biological Chemistry*, 31:15–37, 1917. URL `http://www.jbc.org/content/31/1/15.short`.

Demain, A. L. Microbial biotechnology. *Trends in Biotechnology*, 18(1):26–31, 2000a. DOI: 10.1016/S0167-7799(99)01400-6.

Demain, A. L. Small bugs, big business: The economic power of the microbe. *Biotechnology Advances*, 18(6):499–514, 2000b. DOI: 10.1016/S0734-9750(00)00049-5.

Dochain, D., editor. *Bioprocess Control*. Control Systems, Robotics and Manufacturing Series. ISTE, London, 2008. DOI: 10.1002/9780470611128.

Droste, P., Miebach, S., Niedenführ, S., Wiechert, W., and Nöh, K. Visualizing multi-omics data in metabolic networks with the software Omix—a case study. *Biosystems*, 105(2):154–161, 2011. DOI: 10.1016/j.biosystems.2011.04.003.

Dunn, I. J., Heinzle, E., Ingham, J., and Přenosil, J. E. *Biological Reaction Engineering: Dynamic Modeling Fundamentals with Simulation Examples*. Wiley-VCH, second edition, 2003. DOI: 10.1002/3527603050.

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979. DOI: 10.1214/aos/1176344552.

Eilers, P. H. C. A perfect smoother. *Analytical Chemistry*, 75(14):3631–3636, 2003. DOI: 10.1021/ac034173t.

Epanechnikov, V. A. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Application*, 14(1):153–158, 1969. DOI: 10.1137/1114019.

Flöttmann, M., Schaber, J., Hoops, S., Klipp, E., and Mendes, P. ModelMage: A tool for automatic model generation, selection and management. *Genome Informatics*, 20:52–63, 2008. DOI: 10.11234/gi1990.20.52.

Fraenkel, G. S. The raison d'être of secondary plant substances. *Science*, 129(3361): 1466–1470, 1959. DOI: 10.1126/science.129.3361.1466.

Freyer, S., Graefe, J., Heinzel, M., and Marenbach, P. Evolutionary generation and refinement of mathematical process models. In *Eufit '98, 6th European Congress on Intelligent Techniques and Soft Computing, ELITE – European Laboratory for Intelligent Techniques Engineering*, volume 3, pages 1471–1475, Aachen, Germany, 1998. URL `http://www1.rtr.tu-darmstadt.de/pdf/freyer-1998.pdf`.

Fried, J. and Zietz, S. Curve fitting by spline and Akima methods: Possibility of interpolation error and its suppression. *Physics in Medicine and Biology*, 18(4): 550–558, 1973. DOI: 10.1088/0031-9155/18/4/306.

Gaden, E. L., Jr. Fermentation process kinetics. *Journal of Biochemical and Microbiological Technology and Engineering*, 1(4):413–429, 1959. DOI: 10.1002/jbmte.390010407.

Gavrilescu, M. and Chisti, Y. Biotechnology—a sustainable alternative for chemical industry. *Biotechnology Advances*, 23(7–8):471–499, 2005. DOI: 10.1016/j.biotechadv.2005.03.004.

Gombert, A. K. and Nielsen, J. Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11(2):180–186, 2000. DOI: 10.1016/S0958-1669(00)00079-3.

Gostner, R., Baldacci, B., Morine, M. J., and Priami, C. Graphical modeling tools for systems biology. *ACM Computing Surveys*, 47(2):Article 16, 2014. DOI: 10.1145/2633461.

Haefner, J. W. *Modeling Biological Systems: Principles and Applications*. Springer, second edition, 2005. DOI: 10.1007/b106568.

Haunschild, M. D., Freisleben, B., Takors, R., and Wiechert, W. Investigating the dynamic behavior of biochemical networks using model families. *Bioinformatics*, 21:1617–1625, 2005. DOI: 10.1093/bioinformatics/bti225.

He, Z., Kisla, D., Zhang, L., Yuan, C., Green-Church, K. B., and Yousef, A. E. Isolation and identification of a *Paenibacillus polymyxa* strain that coproduces a novel lantibiotic and polymyxin. *Applied and Environmental Microbiology*, 73(1): 168–178, 2007. DOI: 10.1128/AEM.02023-06.

Heine, T. *Modellgestützte Überwachung und Führung von Fed-Batch-Prozessen zur Antibiotikaproduktion.* PhD thesis, Technische Universität Berlin, 2004. URL `http://opus.kobv.de/tuberlin/volltexte/2004/764/pdf/heine_thomas.pdf`.

Herwig, C. Prozess Analytische Technologie in der Biotechnologie. *Chemie Ingenieur Technik*, 82(4):405–414, 2010. DOI: 10.1002/cite.200900136.

Hilberg, D. Akima-Interpolation. Noch besser als das Spline-Verfahren. *c't*, 6:206–214, 1989.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. COPASI—a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006. DOI: 10.1093/bioinformatics/btl485.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kintano, H., and the rest of the SBML forum:, Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. DOI: 10.1093/bioinformatics/btg015.

Hulhoven, X., Vande Wouver, A., and Bogaerts, P. On a systematic procedure for the predetermination of macroscopic reaction schemes. *Bioprocess and Biosystems Engineering*, 27:283–291, 2005. DOI: 10.1007/s00449-005-0406-4.

Jerusalimski, N. D. and Engamberdiev, N. B. *Conitinuous Cultivation of Microorganisms.* Academic Press, New York, 1969.

Joshi, M., Seidel-Morgenstern, A., and Kremling, A. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447–455, 2006. DOI: 10.1016/j.ymben.2006.04.003.

Junker, B. H. and Wang, H. Y. Bioprocess monitoring and computer control: Key roots of the current PAT initiative. *Biotechnology and Bioengineering*, 95(2):226–261, 2006. DOI: 10.1002/bit.21087.

Kammerer, C. and Gilles, E. D. Modeling secondary metabolite production of *Actinomyces*. In *11th International Biotechnology Symposium: Biotechnology 2000*, Berlin, Germany, 2000.

Kawohl, M., Heine, T., and King, R. Model based estimation and optimal control of fed-batch fermentation processes for the production of antibiotics. *Chemical Engineering and Processing: Process Intensification*, 46(11):1223–1241, 2007. DOI: 10.1016/j.cep.2006.06.023.

King, R. A structured mathematical model for a class of organisms: 1. Development of a model for *Streptomyces tendae* and application of model-based control. *Journal of Biotechnology*, 52:219–234, 1997. DOI: 10.1016/S0168-1656(96)01647-1.

King, R. and Büdenbender, C. A structured mathematical model for a class of organisms: 2. Application of the model to other strains. *Journal of Biotechnology*, 52: 235–244, 1997. DOI: 10.1016/S0168-1656(96)01648-3.

King, R., Leifheit, J., and Freyer, S. Automatic identification of mathematical models of chemical and biochemical reaction systems. In *CHISA 2002*, pages 495–510, Prague, Czech Republic, 2002.

Lal, S. and Tabacchioni, S. Ecology and biotechnological potential of *Paenibacillus polymyxa*: a minireview. *Indian Journal of Microbiology*, 49(1):2–10, 2009. DOI: 10.1007/s12088-009-0008-y.

Liu, G., Chater, K. F., Chandra, G., Niu, G., and Tan, H. Molecular regulation of antibiotic biosynthesis in *Streptomyces*. *Microbiology and Molecular Biology Reviews*, 77(1):112–143, 2013. DOI: 10.1128/MMBR.00054-12.

Loew, L. M. and Schaff, J. C. The virtual cell: a software environment for computational cell biology. *Trends in Biotechnology*, 19(10):401–406, 2001. DOI: 10.1016/S0167-7799(01)01740-1.

Lu, X. L., Xu, Q. Z., Liu, X. Y., Cao, X., Ni, K. Y., and Jiao, B. H. Marine drugs – macrolactins. *Chemistry & Biodiversity*, 5(9):1669–1674, 2008. DOI: 10.1002/cbdv.200890155.

Majer, P. *Parameterschätzung, Versuchsplanung und Trajektorienoptimierung für verfahrenstechnische Prozesse*. PhD thesis, Universität Stuttgart, 1997.

Mangold, M., Angeles-Palacios, O., Ginkel, M., Waschler, R., Kienle, A., and Gilles, E. D. Computer Aided Modeling of Chemical and Biological Systems – Methods, Tools, and Applications. *Industrial & Engineering Chemistry Research*, 44(8):2579–2591, 2005. DOI: 10.1021/ie0496434.

Marenbach, P., Bettenhausen, K. D., Freyer, S., Nieken, U., and Rettenmaier, H. Data-driven structured modeling of a biotechnological fed-batch fermentation by means of genetic programming. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 211(5):325–332, 1997. DOI: 10.1243/0959651971539858.

Marquardt, W. Model-based experimental analysis of kinetic phenomena in multiphase reactive systems. *Chemical Engineering Research & Design*, 83(A6):561–573, 2005. DOI: 10.1205/cherd.05086.

Martín, J. F., Sola-Landa, A., Santos-Beneit, F., and Rodríguez-García, A. Network mechanisms of phosphate control of primary and secondary metabolism. In Dyson, P., editor, Streptomyces*: Molecular Biology and Biotechnology*, pages 137–149. Caister Academic Press, Norwich, UK, 2011.

Michaelis, L. and Menten, M. L. Die Kinetik der Invertinwirkung. *Biochemische Zeitschrift*, 49:334–369, 1913. DOI: 10.1021/bi201284u.

Mirschel, S., Steinmetz, K., Rempel, M., Ginkel, M., and Gilles, E. D. ProMoT: modular modeling for systems biology. *Bioinformatics*, 25(5):687–689, 2009. DOI: 10.1093/bioinformatics.

Monod, J. The growth of bacterial cultures. *Annual Review of Microbiology*, 3:371–394, 1949. DOI: 10.1146/annurev.mi.03.100149.002103.

Montgomery, D. C., Runger, G. C., and Hubele, N. F. *Engineering Statistics.* Wiley, New York, 2001.

Moser, H. *The dynamics of bacterial populations maintained in the chemostat.* Carnegie Institution of Washington, Washington, 1958.

Mundry, C. and Kuhn, K.-P. Modelling and parameter identification for batch fermentations with *Streptomyces tendae* under phosphate limitation. *Applied Microbiology and Biotechnology*, 35:306–311, 1991. DOI: 10.1007/BF00172717.

Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach.* Sinauer, Sunderland, MA, 1990.

Nielsen, J., Nikolajsen, K., and Villadsen, J. Structured modeling of a microbial system: I. A theoretical study of lactic acid fermentation. *Biotechnology and Bioengineering*, 38(1):1–10, 1991a. DOI: 10.1002/bit.260380102.

Nielsen, J., Nikolajsen, K., and Villadsen, J. Structured modeling of a microbial system: II. Experimental verification of a structured lactic acid fermentation model. *Biotechnology and Bioengineering*, 38(1):11–23, 1991b. DOI: 10.1002/bit.260380103.

Nielsen, J., Pedersen, A. G., Strudsholm, K., and Villadsen, J. Modeling fermentations with recombinant microorganisms: Formulation of a structured model. *Biotechnology and Bioengineering*, 37(9):802–808, 1991c. DOI: 10.1002/bit.260370903.

Nikolajsen, K., Nielsen, J., and Villadsen, J. Structured modeling of a microbial system: III. Growth on mixed substrates. *Biotechnology and Bioengineering*, 38(1): 24–29, 1991. DOI: 10.1002/bit.260380104.

Nocon, J., Steiger, M. G., Pfeffer, M., Sohn, S. B., Kim, T. Y., Maurer, M., Rußmayer, H., Pflügel, S., Ask, M., Haberhauer-Troyer, C., Ortmayr, K., Hann, S., Koellensperger, G., Gasser, B., Lee, S. Y., and Mattanovich, D. Model based engineering of *Pichia pastoris* central metabolism enhances recombinant protein production. *Metabolic Engineering*, 24:129–138, 2014. DOI: 10.1016/j.ymben.2014.05.011.

Paul, G. C., Syddall, M. T., Kent, C. A., and Thomas, C. R. A structured model for penicillin production on mixed substrates. *Biochemical Engineering Journal*, 2(1): 11–21, 1998. DOI: 10.1016/S1369-703X(98)00012-6.

Piuri, M., Sanchez-Rivas, C., and Ruzal, S. M. A novel antimicrobial activity of a *Paenibacillus polymyxa* strain isolated from regional fermented sausages. *Letters in Applied Microbiology*, 27(1):9–13, 1998. DOI: 10.1046/j.1472-765X.1998.00374.x.

Rehbock, C., Beutel, S., Brückerhoff, T., Hitzmann, B., Riechers, D., Rudolph, G., Stahl, F., Scheper, T., and Friehs, K. Bioprozessanalytik. *Chemie Ingenieur Technik*, 80(3):267–286, 2008. DOI: 10.1002/cite.200700164.

Rodriguez-Fernandez, M. and Banga, J. R. SensSB: a software toolbox for the development and sensitivity analysis of systems biology models. *Bioinformatics*, 26(13): 1675–1676, 2010. DOI: 10.1093/bioinformatics/btq242.

Rosado, A. S. and Seldin, L. Production of a potentially novel anti-microbial substance by *Bacillus polymyxa*. *World Journal of Microbiology and Biotechnology*, 9:521–528, 1993. DOI: 10.1007/BF00386287.

Ross, B. J. The evolution of higher-level biochemical reaction models. *Genetic Programming and Evolvable Machines*, 13(1):3–31, 2012. DOI: 10.1007/s10710-011-9144-3.

Roubos, J. A. *Bioprocess modeling and optimization: Fed-batch clavulanic acid production by Streptomyces clavuligerus*. PhD thesis, TU Delft, 2002.

Sanchez, S. and Demain, A. L. Metabolic regulation of fermentation processes. *Enzyme and Microbial Technology*, 31(7):895–9060, 2002. DOI: 10.1016/S0141-0229(02)00172-2.

Schaber, J., Liebermeister, W., and Klipp, E. Nested uncertainties in biochemical models. *IET Systems Biology*, 3(1):1–9, 2009. DOI: 10.1049/iet-syb:20070042.

Schaber, J., Flöttmann, M., Li, J., Tiger, C.-F., Hohman, S., and Klipp, E. Automated ensemble modeling with *modelMaGe*: Analyzing feedback mechanisms in the Sho1 branch of the HOG pathway. *PLoS one*, 6:e14791, 2011. DOI: 10.1371/journal.pone.0014791.

Schenkendorf, R. and Mangold, M. Online model selection approach based on Unscented Kalman Filtering. *Journal of Process Control*, 23(1):44–57, 2013. DOI: 10.1016/j.jprocont.2012.10.009.

Schmidt, H. and Jirstrand, M. Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics*, 22(4):514–515, 2006. DOI: 10.1093/bioinformatics/bti799.

Schügerl, K. Progress in monitoring, modeling and control of bioprocesses during the last 20 years. *Journal of Biotechnology*, 85(2):149–173, 2001. DOI: 10.1016/S0168-1656(00)00361-8.

Shonkwiler, R. W. and Herod, J. *Mathematical Biology: An Introduction with Maple and Matlab.* Undergraduate Texts in Mathematics. Springer, New York, second edition, 2009. DOI: 10.1007/978-0-387-70984-0.

Simonoff, J. S. *Smoothing Methods in Statistics.* Springer Series in Statistics. Springer, New York, 1996.

Sin, G., Ödman, P., Petersen, N., Eliasson Lantz, A., and Gernaey, K. V. Matrix notation for efficient development of *First-Principles* models within PAT applications: Integrated modeling of antibiotic production with *Streptomyces coelicolor*. *Biotechnology and Bioengineering*, 101(1):153–171, 2008. DOI: 10.1002/bit.21869.

Sugimoto, M., Kikuchi, S., and Tomita, M. Reverse engineering of biochemical equations from time-course data by means of genetic programming. *BioSystems*, 80(2): 155–164, 2005. DOI: 10.1016/j.biosystems.2004.11.003.

Tang, S., Chen, J., and Zhang, Z. Structured models for recombinant human interleukin-11 fermentation. *Biochemical Engineering Journal*, 35(2):21–217, 2007. DOI: 10.1016/j.bej.2007.01.016.

Terziev, S. Offline- und online-modelldiskriminierende Versuchsplanung biologischer Prozesse. Bachelor's thesis, Technische Universität Berlin, 2014.

U.S. Food and Drug Administration. PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance, 2004. URL `http://www.fda.gov/downloads/Drugs/Guidances/ucm070305.pdf`.

Vemuri, G. N. and Aristidou, A. A. Metabolic engineering in the -omics era: Elucidating and modulating regulatory networks. *Microbiology and Molecular Biology Reviews*, 69(2):197–216, 2005. DOI: 10.1128/MMBR.69.2.197-216.2005.

Violet, N., Rossner, N., Heine, T., and King, R. RapOpt − An automation tool for production-orientated run-to-run model evolution. In *6th Vienna International Conference on Mathematical Modelling (MATHMOD 2009)*, volume 6, pages 2339–2346, Vienna, Austria, February 2009.

Wahl, S. A., Haunschild, M. D., Oldiges, M., and Wiechert, W. Unravelling the regulatory structure of biochemical networks using stimulus response experiments and large-scale model selection. *Systems Biology, IEE Proceedings*, 153(4):275–285, 2006. DOI: 10.1049/ip-syb:20050089.

Whittaker, E. T. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922. DOI: 10.1017/S001309150000359X.

Wiechert, W. Modeling and simulation: tools for metabolic engineering. *Journal of Biotechnology*, 94(1):37–63, 2002. DOI: 10.1016/S0168-1656(01)00418-7.

Xue, C., Tian, L., Xu, M., Deng, Z., and Lin, W. A new 24-membered lactone and a new polyene δ-lactone from the marine bacterium *Bacillus marinus*. *The Journal of Antibiotics*, 61:668–674, 2008. DOI: 10.1038/ja.2008.94.

Yarmush, M. L. and Banta, S. Metabolic engineering: Advances in modeling and intervention in health and disease. *Annual Review of Biomedical Engineering*, 5: 349–381, 2003. DOI: 10.1146/annurev.bioeng.5.031003.163247.

Yeh, K. C. and Small, R. D. Pharmacokinetic evaluation of stable piecewise cubic polynomials as numerical-integration functions. *Journal of Pharmacokinetics and Biopharmaceutics*, 17(6):721–740, 1989. DOI: 10.1007/BF01062126.

Own Publications

Herold, S. and King, R. Automatic identification of structured process models based on biological phenomena detected in (fed-)batch experiments. *Bioprocess and Biosystems Engineering*, 37(7):1289–1304, 2014. DOI: 10.1007/s00449-013-1100-6.

Herold, S., Heine, T., and King, R. An automated approach to build process models by detecting biological phenomena in (fed-)batch experiments. In *11th IFAC Symposium on Computer Application in Biotechnology*, volume 11, pages 138–143, Leuven, Belgium, July 2010. DOI: 10.3182/20100707-3-BE-2012.0012.

ISBN 978-3-7375-2957-0