PREDICTIVE MACROSCOPIC MODELING OF CHINESE HAMSTER OVARY CELLS IN FED-BATCH PROCESSES

Dissertation

zur Erlangung des Grades

des Doktors der Ingenieurwissenschaften

der Naturwissenschaftlich-Technischen Fakultät (NT)

von

Bassem Ben Yahia

Saarbrücken, Deutschland

2017

PREDICTIVE MACROSCOPIC MODELING OF CHINESE HAMSTER OVARY CELLS IN FED-BATCH PROCESSES

Dissertation

zur Erlangung des Grades

des Doktors der Ingenieurwissenschaften

der Naturwissenschaftlich-Technischen Fakultät (NT)

von

Bassem Ben Yahia

Saarbrücken, Deutschland

2017

Tag des Kolloquiums: 08. December 2017

Dekan: Prof. Dr. G. Kickelbick

Berichterstatter: Prof. Dr. C. Wittmann

Vorsitz: Prof. Dr. E. Heinzle

Prof. Dr. G.-W. Kohring

Akad. Mitarbeiter: Dr. B. Becker

“Don’t worry, life is easy.”

— Aaron

“Have no fear of perfection - you'll never reach it.”

— Salvador Dali

"I am never content until I have constructed a mechanical model of the subject I am studying. If I succeed in making one, I understand; otherwise I do not."

— William Thomson (Lord Kelvin)

To my mother

Acknowledgments

Starting this PhD thesis was already a big adventure: find a subject that interests me and is also valuable for UCB, identify an expert in the field that could supervise this thesis, find the funding, set up the contract... It was a long journey of almost 1 year with Laetitia Malphettes and Elmar Heinzle but it was worth it!

First, I can only be grateful for receiving the best PhD supervisor, Prof. Dr Elmar Heinzle, one can wish for. He was an inspiring mentor, a *sensei,* that always had time for my questions and for my new crazy ideas. Without his contributions of time and ideas, this thesis would not have been the same. I always had support from him when needed and I was, and I'm still impressed by the joy and enthusiasm that he has for his research that was contagious for me.

I would also particularly like to thank Laetitia Malphettes for the energy she has put into setting up this project with me. I am grateful for her trust, for her valuable guidance and for allowing me the space to pursue my research. Our animated and passionate discussions allowed me to redirect my thesis when needed and to always confront my ideas and hypothesis. I will always remember a stimulating sentence that she repeated a lot of time: "You must be more rigorous". I hope I improved as she expected.

I would like to acknowledge the support of Boris Gourevitch for his inspirational drive and for the time he dedicated to this thesis. Without his help, this thesis would not have existed. I was always impressed by his knowledge and his skills in programming and statistical analysis.

I would also like to thank Prof. Dr. Gert-Wieland Kohring for gladly consenting to reviewing this thesis and being the second supervisor. I want to thank the thesis committee in advance for their precious time and scientific input.

As I was waiting for the setting up of my contract, I spent almost 1 year in Germany in Saarbrücken where I started another adventure, trying to adapt to a new culture and a new language. I would like to thank Judith Wahrheit and Averina Nicolae for

all the practical information that enabled me to adapt well to my new life in Saarbrücken. I would especially like to thank Ghamdan Beshr for his help in finding a new apartment; without him I think I would still be looking for a new flat.

At UCB Pharma, I would like to thank all the team and all the people I met during my five years in this amazing group. It's a great team with an excellent work atmosphere. Special thanks to my friend and colleague Valentine Chevallier for accompanying me during these years both literally and figuratively (thanks for the carpooling!). I will miss our stimulating discussions in the car and all the fun we have had. Special thanks to Gregory Mathy, Guillaume Le Reverend and Boris Fessler: we started at the same time in the department in 2011 and I think we had a lot of fun together (and of course we worked a lot!). Also, thanks Jonathan Stern and Andrea McCann for your expertise in analytical methods and for our boardgame nights, Meriam Annani for her support in the lab, Carine Pereira for her help in developing the protocol for shake flasks, Thibault Lasgouzes for being a great friend (and for his constant complaints), Thomas Fontaine and Coralie Borrossi for being amazing host for our long and funny boardgame nights (with Thibault Lasgouzes too). I would like to acknowledge Mareike Harmsen who gave me the opportunity to start at UCB as a master student before the PhD thesis and all the great people in T_2 building in Braine L'Alleud.

For the non-scientific side of my thesis, I would also like to thank all of my friends especially Melanie, or I should say Dr. Melanie now, and "*la puissance du fond*" – Alexandre (*Vratefluch*), Pierre, Elie, Alice, Noura, Thibault (Titou) and Marine – with whom I had a lot of fun.

Last but not least, two women deserve my greatest gratitude for their love and support. First, Laura Sanchez for literally everything: for your support during this adventure and for your listening (I know that listening to me talking about bioreactors every night and every weekend was not easy for you). This thesis would not have existed without her. She was always my support in the moments when no one was there to answer my queries. Second, a huge thank you to my mother for her

unconditional love even if she still does not really understand what I'm working on. She has always been behind me in my periods of doubt. She is an example for me and I am sure that if she had the opportunity to pursue her studies, she would have been able to be a great scientist!

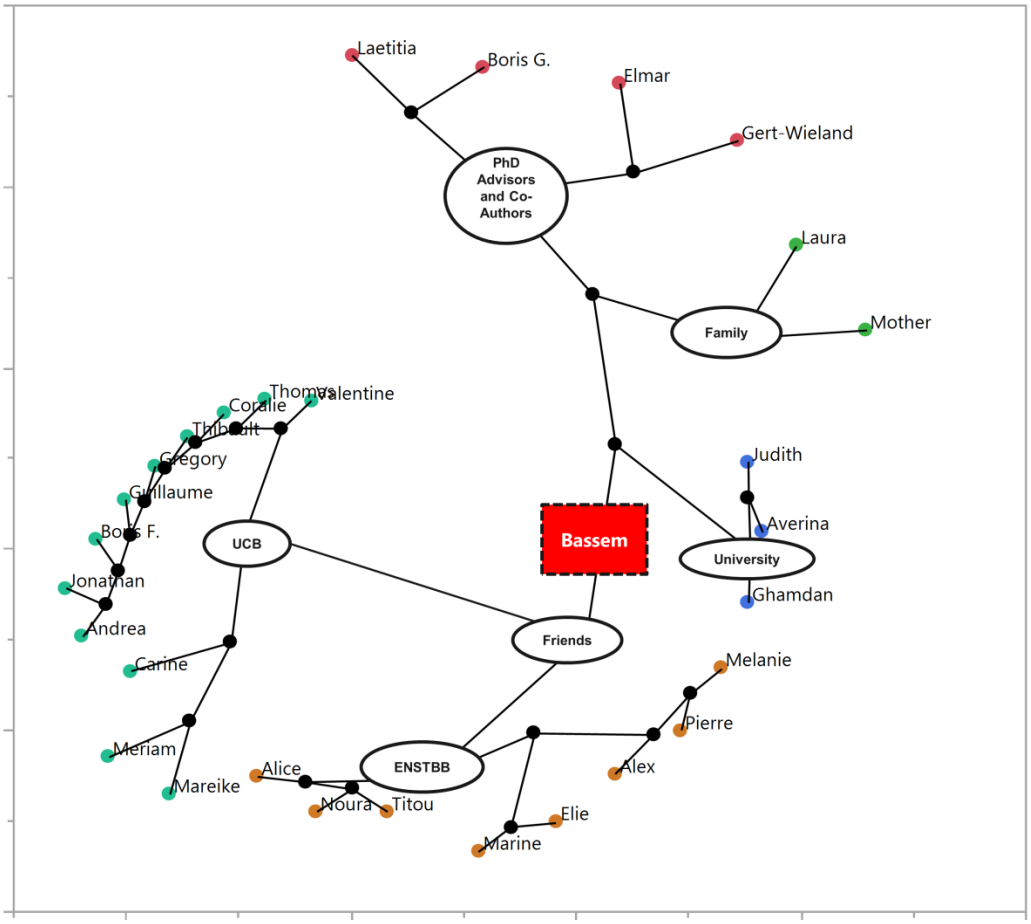


Figure 0.0. Thank you, Merci, Danke, شكرا

Zusammenfassung

Der Schwerpunkt dieser Dissertation liegt auf der systematischen Entwicklung Modellen für die Vorhersage des zellulären Stoffwechsels, des Wachstums und der Produktion von monoklonalen Antikörpern (mAb) in Kulturen von Chinesischen Hamster-Ovarzellen (CHO). Zunächst wurden mit segmentierter linearer Regression metabolischer Phasen identifiziert. Diese Identifizierung beruht auf der Annahme eines pseudo-stationären Zustands und somit, dass in einer Phase alle Raten linear miteinander korreliert waren. Die spezifischen Raten wurden aus den Zeitverläufen der Konzentrationen der Metabolite und des mAb sowie der Lebendzellzahl bestimmt. Durch die Korrelation konnten alle Raten über die Wachstumsrate im 2 L und im 2000 L Maßstab berechnet werden. Danach wurde ein kinetisches Modell des Wachstums der Zellen etabliert, was die Vorhersage aller Raten auch in fed-batch Kulturen erlaubt. Die Kinetik basiert auf der Monod-Kinetik modifiziert mit einer variablen maximalen spezifischen Wachstumsrate. Das kinetische Modell erlaubt eine rechnerische Optimierung der Substratzuführung für eine maximale Produktion. Damit wurde gezeigt, dass aus makroskopischen Daten, d.h. ohne intrazelluläre Messungen, wesentliche Informationen erhalten werden können, mit denen neue Experimente in einem industriellen Umfeld vorhergesagt werden können. Diese innovative und systematische Vorgangsweise eröffnet neue Perspektiven für die Reduzierung von Kosten und für eine Beschleunigung der Prozessentwicklung.

Abstract

This thesis focuses on developing a systematic modeling method that can capture the essential features for prediction of cell metabolism, growth and monoclonal antibody (mAb) production in Chinese Hamster Ovary (CHO) cells. In a first step all specific consumption rates are calculated based on time courses of extracellular metabolites, viable cell density and mAb. Then the metabolic phases within which the metabolic pseudo-steady state approximation is verified are identified. In a third step, all metabolic rates are expressed as a function of the specific growth rate within each metabolic phase.

We have applied this method to a set of small bioreactor data and have shown that the model obtained can predict specific conversion rates both small and also at large scale.

In the second part of this thesis, a kinetic model of the cell growth has been developed. Together with previously described methodology, this kinetic model results in a predictive metabolic model for each experimental cell growth data are not required. The kinetic model is based on Monod kinetics with a few modifications such as a varying the maximum specific growth rate as a function of the integral viable cell density. The full kinetic model can be used off line to design optimal feeding profiles. The results of this thesis demonstrate that rich knowledge can be derived from macroscopic data that can then be used to predict new production conditions in an industrial environment at small and large scale.

Extended abstract

The work performed in this thesis demonstrates that a systematic modeling approach can be applied to complex bioprocesses. This systematic methodology is based on material balances and metabolic networks. It provides simple and predictive dynamic models of Chinese Hamster Ovary (CHO) cell metabolism, cell growth and monoclonal antibody (mAb) production based on a limited subset of small scale data. This thesis is structured around three major parts that demonstrates how macroscopic modeling of the complex metabolism and cell growth of CHO cells leads to predictive models. It consists of chapters organized as successive steps building on each other and enabling *in silico* models of bioprocesses.

In the first part (**chapter 1**), a review of major modeling strategies and methods to understand and simulate the macroscopic behavior of mammalian cells is compiled. These strategies comprise two important steps: the first step is to identify stoichiometric relationships for the cultured cells connecting the extracellular inputs and outputs. In a second step, macroscopic kinetic models are introduced. These relationships together with bioreactor and metabolite balances provide a complete description of a system in the form of a set of differential equations. These can be used for the simulation of cell culture performance and further for optimization of production. The strategies described in this chapter were used to develop predictive models of CHO cells metabolism, cell growth and mAb production in **chapter 2** and **chapter 3**.

Chapter 2 focuses on the construction of an apparent simple model of CHO cell metabolism during biopharmaceutical production that can be applied across a wide range of cell culture conditions, and based on which further developments can be designed. This question was addressed through systematic application of the metabolic steady state concept. However, throughout the bioreactor production, the

cells adapt to the changing extracellular conditions resulting in the development of different metabolic phases and related metabolic shifts. This makes modeling of the cell metabolism for all phases of cultivation more difficult. In order to deal with this complexity, one can identify metabolic phases in which the metabolic pseudo steady state approximation is verified, divide the cell culture process into these phases and perform separate analyses. This method enables the utilization of simple mathematical procedures to model complex phenomena. Metabolic rates were computed based on time series of extracellular metabolites and product concentrations as well as viable cell density. For each metabolic phase, cells are assumed to be in pseudo-steady state and so the stoichiometric coefficients between the specific production rates of metabolites and the specific growth rate are defined as constant. The production rates of metabolites were computed and modeled as a function of the specific growth rate. First the total number of metabolic steady state phases and the location of the metabolic phase breakpoints were determined by recursive partitioning. For this, the smoothed derivative of the metabolic rates with respect to the growth rate were used followed by hierarchical clustering of the obtained partition. Piecewise regression, also called segmented linear regression, was then applied to the metabolic rates with the previously determined number of phases. This allowed identifying the growth rates at which the cells underwent a metabolic shift. The resulting model with piecewise linear relationships between metabolic rates and the growth rate did well describe cellular metabolism in the fed-batch cultures. Using the model structure and parameter values from a small scale cell culture (2 L) training dataset, it was possible to predict metabolic rates of new fed-batch cultures just by using the experimental specific growth rates. Such prediction was successful both at the laboratory scale with 2 L bioreactors and also at the production scale of 2000 L. The final mAb titer can also be predicted even if the cells are starved in some essential metabolites. This type of modeling demonstrates the feasibility of building a reliable and accurate macroscopic model and also provides a flexible framework to set a solid foundation for metabolic flux analysis and mechanistic type of modeling.

In the third part of this thesis (**chapter 3**), a systematic approach is described to establish dynamic predictive models of CHO cell growth during biopharmaceutical production. Cell growth, cell metabolism and monoclonal antibody (mAb) production are predicted by combining an empirical metabolic model with mixed Monod-inhibition type kinetics that were generalized to every possible external metabolite. We describe the maximum specific growth rate as a function of the integral viable cell density (IVCD). Moreover, the storage of metabolite in intracellular pools was taken into account and was assumed to influence cell growth. This is illustrated with fed-batch cultures of CHO cells producing a mAb. The impact of two identified and selected essential metabolites on cell growth and cell productivity was assessed and the macroscopic model was successfully used to predict the impact of new untested feeding strategies on cell growth and mAb production. The resulting model combining piecewise linear relationships between metabolic rates and the growth rate and Monod-inhibition type models for cell growth did well predict cell culture performance in fed-batch cultures even outside the range of experimental data used for establishing the model. The metabolic model obtained thanks to the methodologies presented can predict metabolic fluxes based on small scale data both at small and at large scale. It can also predict the cells' response to different feeding strategies at both scales. This is an important step towards reducing the number of bioreactor experiments required to control bioreactor production processes and moving towards *in silico* simulations of the impact of process parameters on process yields and cell metabolism.

The results of these different steps are discussed in the final conclusion and outlook (**chapter 5**) that summarizes the systematic methodologies developed, describes potential implications and proposes applications.

Abbreviations

Subscripts

c	Coupled to growth
nc	Not coupled to growth

Abbreviations

aa	Amino acid
ALAT	Alanine aminotransferase
Amm	Ammonium
BHK	Baby hamster kidney
CHO	Chinese hamster ovary
CI	Confidence interval
CQAs	Critical quality attributes
CV	Coefficient of variation
DGSM	Derivative based global sensitivity measures
EFM	Elementary flux mode analysis
FDA	Food and drug administration
FPMs	First principle models
Glc	Glucose
GRG	Generalized reduced gradient method
GSA	Global sensitivity analysis

HPLC	High performance liquid chromatography
IPOPT	Interior point primal dual line search algorithm
IVCD	Integral viable cell density
Lac	Lactate
LMA	Levenberg-Marquard algorithm
LOESS	Locally estimated scatterplot smoother
LOWESS	Locally weighted scatterplot smoother
mAb	Monoclonal antibody
MCMC	Markov chain Monte Carlo method
MFA	Metabolic flux analysis
NMPC	Non-linear model predictive controller
NN	Neural network
PCA	Principal component analysis
PE	Processing elements
PFA	Principal factor analysis
pO_2	Dissolved oxygen concentration
PQA	Product quality attribute
PSO	Particle swarm algorithm
QbD	Quality by design
QP	Quadratic programming
SDP	Semi-definite programming
SQP	Sequential quadratic programming

STR	Stirred tank glass bioreactor
TCA	Tricarboxylic acid cycle
TSD	Target seeding density
UCL	Upper control limit
VCD	Viable cell density
Vr	Bioreactor volume

Table of Contents

Aim and outline of the thesis

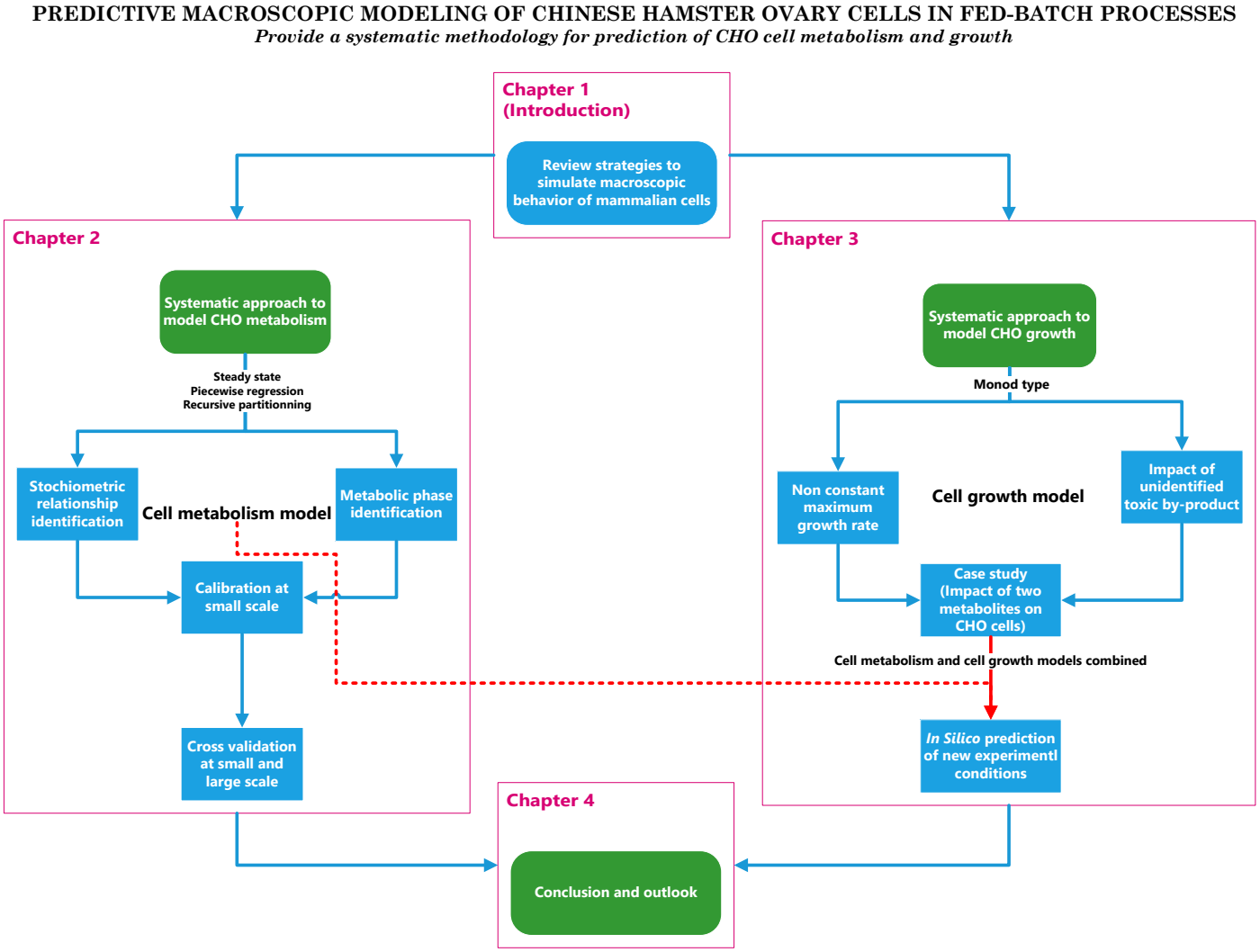
The aim of the presented thesis is to provide a systematic methodology for *in silico* prediction of Chinese Hamster Ovary (CHO) cell metabolism and growth which can be applied to complex bioprocesses in an industrial setting.. The work is structured into three major parts that build upon each other.

In the first part (**chapter 1**), we introduce and review major modeling strategies to predict the macroscopic behavior of mammalian cells. Here, the advantages of each strategy are discussed. This part ensures a smooth introduction to the modeling methodologies developed during the thesis and presented in the follow-up studies.

In the second part of the thesis (**chapter 2**), a systematic approach to model CHO metabolism during biopharmaceutical production across a wide range of cell culture conditions was described. Here, the metabolic steady state concept was applied to model the production rates of metabolites as a function of the specific growth rate. Two powerful methods, piecewise regression and recursive partitioning, were combined to identify metabolic shifts and stoichiometric relationships between production rates of metabolites and the specific growth rate. This model can be applied at various scales to increase its industrial applicability. This model served also as starting point to design further in-depth *in silico* model of CHO cells.

The third part of the work (**chapter 3**) presents the second building stone, i.e. the development of a systematic approach to model the dynamics of CHO cell growth during biopharmaceutical production. This leads to the development of a dynamic *in silico* macroscopic model of CHO cells. Here, a derivative of Monod type kinetics was developed, then the model was calibrated and combined to the metabolic model described in **chapter 2**. The model includes non-constant maximum specific growth rate that is dependent on an unidentified inhibitory by-product. Both models described in **chapter 2** and **chapter 3** constitute essential building blocks for the development of an *in silico* macroscopic model of CHO cells. The last section of the thesis presents the conclusion and the outlook of the thesis (**chapter 4**).

A schematic representation of the outline of the thesis is presented below.



Chapter 1

1 Macroscopic modeling of mammalian cell growth and metabolism

1.1. Abstract

We review major modeling strategies and methods to understand and simulate the macroscopic behavior of mammalian cells. These strategies comprise two important steps: the first step is to identify stoichiometric relationships for the cultured cells connecting the extracellular inputs and outputs. In a second step, macroscopic kinetic models are introduced. These relationships together with bioreactor and metabolite balances provide a complete description of a system in the form of a set of differential equations. These can be used for the simulation of cell culture performance and further for optimization of production.

This chapter was published as

Ben Yahia, B., Malphettes, L., Heinzle, E., 2015. Macroscopic modeling of mammalian cell growth and metabolism. Applied microbiology and biotechnology. 99, 7009-7024.

1.2. Introduction

Mammalian cell cultures are the major source of a number of biopharmaceutical products, including monoclonal antibodies (Sidoli, Mantalaris et al. 2004, Niklas and Heinzle 2012), viral vaccines (Vester, Rapp et al. 2010), and hormones (Nottorf, Hoera et al. 2007). Chinese hamster ovary (CHO) cells are widely used as an expression system for the synthesis of therapeutic glycosylated proteins (Palomares, Estrada-Moncada et al. 2004, Zhu 2012). Predicting the behavior of mammalian cells during cell culture processes under different culture conditions is highly desirable for both commercial and scientific reasons (Kell and Knowles 2006). In batch and fed-batch processes, the rate of overproduction of heterogeneous proteins by mammalian cells is limited by the decline in cell viability, by the depletion of required metabolites and substrates or by the accumulation of metabolic products and inhibitors. Therefore, it becomes imperative to identify the parameters which have a significant impact on cell viability and on protein production and understand their effects on the cellular phenotype. Moreover in 2004, the food and drug administration (FDA) proposed the "Quality by Design" (QbD) methodology to biopharmaceutical companies. The focus of this concept is that the quality, most important protein glycosylation, should be built into a product with a thorough understanding of the product itself and the process for its production (Tomba, Facco et al. 2013). Additionally, critical process parameters should be identified which have an impact on the critical quality attributes (CQAs) of the product (Kontoravdi, Asprey et al. 2007, Teixeira, Oliveira et al. 2009, Royle, Jimenez del Val et al. 2013).

Mammalian cell culture processes are complex (Stelling, Sauer et al. 2006), and numerous input parameters have to be identified to optimize growth and productivity (Nolan and Lee 2011, Sellick, Croxford et al. 2011, Nolan and Lee 2012). To understand biological mechanisms and to optimize production processes, rational design guided by experience is the most common method currently used. However, experiments are time consuming and expensive to perform, and generally generate noisy data. Mathematical models can help to characterize the different phenotypes

and the needs of mammalian cells (Sidoli, Mantalaris et al. 2004, Royle, Jimenez del Val et al. 2013). They can be used as a prediction tool in simulation and optimization (Wiechert 2002, Goudar, Biener et al. 2006). Mathematical models can also help to understand and identify mechanisms that cannot be easily identified only with experimental data and a pure statistical analysis of them. Therefore, modeling of metabolism has become highly desirable in the development process where the identification of the parameters impacting the cell culture processes and the prediction of the evolution of the processes are important. Identification of yield coefficients can be used for this purpose (Chen and Bastin 1996). This creates significant added value in terms of cost and time compared to methods that do not use models (Kessel 2011).

Compared to very detailed cellular models, the benefit of the use of macroscopic models is that it is much easier but yet very informative to analyze the cells as a black box or grey box rather than to take into account extended details of what happens inside the cell (Zamorano, Vande Wouwer et al. 2013). Analysis of intracellular metabolites necessary for setting up and tuning detailed kinetic models of metabolism is much more complex to perform than extracellular metabolite analysis and requires much more sophisticated techniques, particularly for suspended cells (Neermann and Wagner 1996, Wahrheit and Heinzle 2013). In addition, the number of model parameters in macroscopic models is significantly lower than the number of parameters in microscopic models. The identification of parameters is therefore more difficult for very detailed microscopic models.

A mathematical model can be used for different purposes (Ashyraliyev, Fomekong-Nanfack et al. 2009, Hu and Zhou 2012):

(1) To summarize a large volume of experimental data;

(2) To explore concepts and test hypotheses;

(3) To predict the behavior of the systems under non-tested conditions;

(4) To identify conditions for optimal performance of a process as defined by an objective function

The extrapolation power of a model cannot be predicted *a priori*. The probability that a model will allow prediction outside the originally observed region is, however, increasing if physically meaningful functions are used. In our review we emphasize the separation into a material balancing part, the so-called macroscopic reactions, and a kinetic part. The material balancing part, i.e. stoichiometry, provides a sound basis and must not be violated for keeping predictivity. The kinetic part relies very much on the characteristics of the rate determining processes, e.g. saturation kinetics of Michaelis-Menten type, allosteric kinetics of Hill-type or structure of feedback control loops in biological systems. The appropriate choice of the underlying types of mathematical functions is certainly a crucial point in this respect. For certain problems, e.g. metabolic network modeling as shown for CHO, the use of ensembles, i.e. sets of models with different structures and/or parameter values, seems useful for reducing prediction (Villaverde, Bongard et al. 2015).

In this review, we will present different types of models used in previous work to model the metabolism of suspension cells at the macroscopic level, i.e. to model extracellular outputs as function of extracellular inputs. This paper is organized as follows: (i) the first part introduces the types of models and the existing modeling frameworks (Mahadevan and Doyle 2003). (Mahadevan and Doyle) Then, different methods for identifying relevant parameters for creating a macroscopic metabolic model will be presented. Preliminary work has to be performed to reduce the number of parameters to study and to understand which parameters have a significant impact on the responses (Mahadevan and Doyle 2003). (Mahadevan and Doyle) In part three, different kinetic models are reviewed. Kinetic models are used after the selection of parameters and when the relationships between those parameters are defined. (iv) Model calibration and testing are reviewed and (v) applications to process control are described. (vi) Finally, main conclusions and an outlook are presented.

1.3. Types of models

There are different ways to classify models. The first distinguishes between empirical models, also called descriptive models, and mechanistic models. Empirical models use a pragmatic description of all the data with any suitable mathematical relationship. They only partially take into account the underlying phenomena or physical laws that govern the system behavior. Mechanistic models are based on theoretical foundations of systems and on known relationships. The predictions of the responses are based on biological, chemical and physical input of knowledge.

Another classification was proposed by Tsuchiya et al. (Tsuchiya, Fredrickson et al. 1966) and distinguishes deterministic models and probabilistic models. The first is based on continuous variables using differential equations. Reactions and interactions are represented as continuous processes (production, consumption, growth...) by corresponding mathematical functions. It is appropriate for systems composed of a relatively large number of cells, e.g. more than 10,000. This kind of model describes the population as average. Probabilistic or stochastic models use probability in the formulation of the model and are typically used for a population of only few cells or for molecular events with only small number of molecules, e.g. transcription. This allows representation of the variability of a population and a system. In cell culture the number of cells is usually very large (e.g. $>10^6$ cells/mL) allowing the preferential use of deterministic models.

Another classification distinguishes structured, non-structured, segregated and non-segregated models. Structured models take into account the cellular reactions within cells (Tsuchiya, Fredrickson et al. 1966, Harder and Roels 1982). Structured models can describe biological systems in great detail but are more difficult to set up. The number of parameters increases with the complexity of the model and with the number of intracellular reactions taken into account. In addition, despite the enormously increased knowledge about cellular process, there is still a significant lack of information about many steps, e.g. transport, control of enzymes activities and expression or post-transcriptional processing of proteins. Unstructured models are

easier to work with because they analyze the cells as a black or grey box. Intracellular reactions are not analyzed in detail. It is assumed, for example, that cell growth depends only on extracellular parameters. However, the extended and now easily excessible comprehensive knowledge about the biochemical reaction networks and its stoichiometry allows the incorporation of this information into macroscopic models. Macroscopic models are less accurate than structured models, but easier to set up and to apply. Segregated models, as opposite of non-segragated models, describe cellular behavior as a function of cell cycles or age of cells (Karra, Sager et al. 2010, Meshram, Naderi et al. 2013, García Münzer, Ivarsson et al. 2015, García Münzer, Kostoglou et al. 2015, Pisu, Concas et al. 2015). The vast majority of models is non-structured and non-segregated.

Neural networks are particularly useful to relate input and output variables to each other in complex systems with incomplete or even completely lacking knowledge of the systems structure and also in cases with incomplete measurements. Mechanistic knowledge can however be introduced by using hybrid models (van Can, te Braake et al. 1999, Oliveira 2004).

1.4. Identification of relevant input-output relationship

A general macroscopic reaction scheme of macroscopic reactions can be expressed as follow (Bastin and Dochain 1990):

$$\sum_{i\in R_k}\left(-\nu_{i,k}\right)\xi_i \xrightarrow{\varphi_k} \sum_{j\in P_k}\nu_{j,k}\xi_j \quad k\in[1,M] \tag{1.1}$$

where

- M is the number of reactions
- φ_k is the kth reaction rate;
- ξ_i and ξ_j are the ith and the jth component, respectively;
- $\nu_{i,k}$, $\nu_{j,k}$, are the corresponding stoichiometric coefficients;

- R_k is the kth set of reactant and catalyst indices;
- P_k is the kth set of product and catalyst indices.

This general reaction scheme represents a macroscopic-stoichiometric relationship. To set up such a macroscopic model, the important parameters, i.e. the relevant cellular inputs, ξ_i, and outputs, ξ_j, as well as the stoichiometric coefficients, $\nu_{i,k}, \nu_{j,k}$, relating the inputs to the outputs, have to be determined. This can start from the increasingly comprehensive knowledge of cellular reactions and transport or, as traditionally done, from purely empirical data. Ideally both types of information are combined as described below and indicated in Figure 1.1. This step is often the main bottleneck in the design of a macroscopic model for complex biotechnological processes.

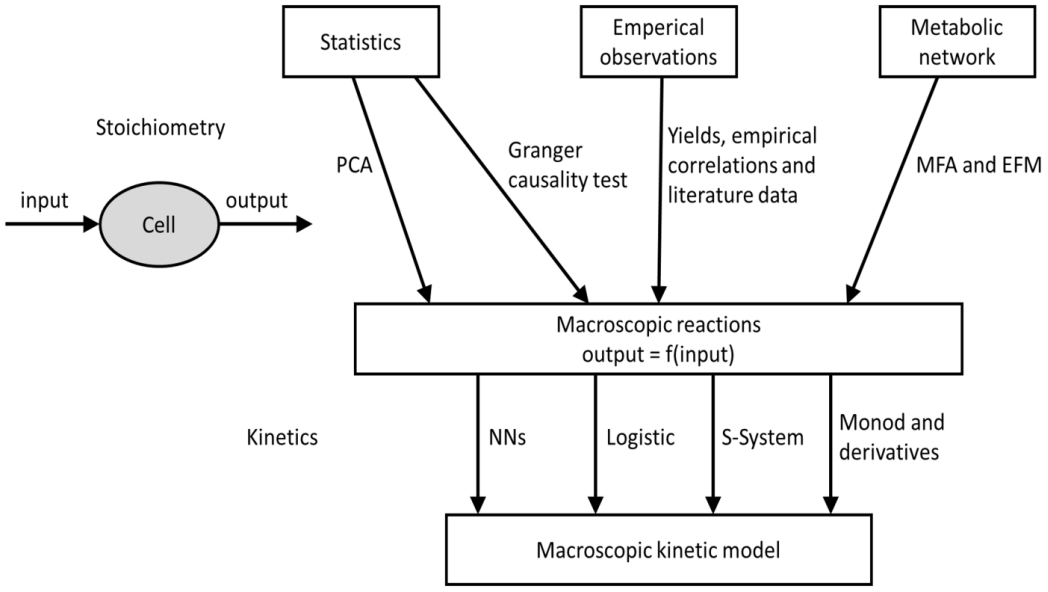


Figure 1.1. Methods to derive macroscopic kinetic models. In order to get a simulation and prediction model of the macroscopic cell behavior, first, the macroscopic reactions of the cell culture system have to be determined, i.e. the stoichiometry relating input and output of the cells. To do that, statistical methods, empirical observations and metabolic network based methods can be used. After that, the kinetics of the system have to be described and combined with the stoichiometric model. Finally the model is calibrated, usually using optimization based methods, and tested. **PCA**: **P**rincipal **C**omponent **A**nalysis; **MFA**: **M**etabolic **F**lux **A**nalysis; **EFM**: **E**lementary **F**lux **M**ode; **NNs**: **N**eural **N**etworks

1.4.1. Method based on expert reasoning

One possible approach to select significant parameters is based on expert reasoning and experimental observations. This approach measures correlations between the macroscopic outputs we want to model with the cell culture parameters, i.e. the macroscopic inputs, under different experimental conditions. A most popular method uses the concept of yield coefficients relating always two measured variables to each other, e.g. biomass to substrate or product to biomass (Dunn, Heinzle et al. 2003). Yield coefficients are frequently used to set up stoichiometric relationships to be applied in metabolic flux analysis using metabolite balancing (Niklas, Noor et al. 2009). It requires little thought about the actual detail of the system and uses most significant phenomena observed during experiments to define the extracellular parameters such as limiting nutrients or accumulation of side waste products. Typically, outputs/inputs taking into account in a macroscopic model with this kind of approach are biomass, glucose, glutamine, lactate and ammonia. For instance, Jang and Barford (Jang and Barford 2000) developed an unstructured model of growth and metabolism of a mouse murine hybridoma AFP-27 cell line producing an IgG1 antibody. They assumed that glucose, glutamine, lactate, and ammonia were growth limiting. Lactate and ammonia were considered as toxic products of catabolic reactions, which inhibit cell growth and can ultimately cause cell death In their model, even though they assumed that hybridoma cells can produce monoclonal antibodies until any of amino acid is depleted, they only considered glutamine as limiting amino acid. Moreover, based on the demonstration of Suzuki and Ollis (Suzuki and Ollis 1990), they considered the specific antibody production to be a function of the fraction of cells in G1 phases. Acosta et al. (Acosta, Sánchez et al. 2007) also assumed this link between specific growth rate and specific productivity in their model of IgG2a Mab production in hybridoma cells. Although glucose is generally important for cell growth, it was not found to be a limiting nutrient in another model (Bree, Dhurjati et al. 1988) that is, however, only relying on one batch experiment,

certainly a too limited observed experimental space for meaningful extrapolation. Lactate and ammonia are assumed to both inhibit and kill cells (Glacken, Adema et al. 1988, Batt and Kompala 1989, Ozturk, Riley et al. 1992) but the impact on specific antibody productivity was reported as not significant (Ozturk, Riley et al. 1992). Jang and Bradford (Jang and Barford 2000) and Dhir et al. (Dhir, Morrow et al. 2000) assumed that the lactate production was due to cellular consumption of glucose and glutamine. They assumed that the spontaneous degradation of glutamine was negligible It is however usually relevant but depends on the used medium and process duration (Glacken, Adema et al. 1988, Ozturk and Palsson 1990, Borchers, Freund et al. 2013). Amino acid depletion has been considered in another model developed by Liu et al. (Liu, Bi et al. 2008). Knowledge about metabolism and its control can be incorporated but usually not in a systematic manner. Meshram et al. (Meshram, Naderi et al. 2013) developed a macroscopic metabolic model and linked it to a model of apoptosis. A dynamic model of mAb synthesis and mAb glycosylation by hybridoma was described by Kontoravdi et al. (Kontoravdi, Asprey et al. 2007) using a structured model based on the work of Umaña and Bailey (Umaña and Bailey 1997). The availability of nutrients such as glucose or glutamine had an impact on protein glycosylation.

Such empirical procedures can be a valuable tool for understanding metabolic processes as well as for process design and optimization. They are used to design a macroscopic model and select the extracellular parameters which have an impact on the response defined. Nevertheless, very little real understanding of the cell culture process is obtained with this kind of procedure.

1.4.2. Method based on statistical tools

A large number of variables can be identified and quantified due to the recent development of high-resolution and high-throughput analytical techniques (Martin, Reynolds et al. 2014, Steinhoff, Ivarsson et al. 2014). In this context, it becomes more complex to select the significant input only with an empirical approach and based on

expert judgment. Moreover, the relations of variables are generally dynamic and involve temporal dependencies.

To deal with these challenges, multivariate data analysis methods, e.g. principal component analysis (PCA), can be used as a statistical tool to select parameters. PCA is a multivariate analysis method based on eigenvalue analysis, which is actually the projection of original data onto a new set of axes, i.e., the principal components. PCA has been introduced by Pearson (Pearson 1901) and Hotelling (Hotelling 1933) to describe the variation of multivariate data in terms of a set of uncorrelated variables. It is used to reduce a high-dimensional dataset into fewer dimensions while retaining important information. Starting out with high-dimensional noisy experimental data, one can reduce the dimensionality and even remove pure random errors by determination of significant factors (Malinowski 1991). Using significant factor analysis followed by rotation, a stoichiometric model with only two independent, physically meaningful reactions were identified for *Bacillus subtilis* batch culture (Saner, Heinzle et al. 1992). Xing et al. (Xing, Li et al. 2008) used a methodology based on principal factor analysis (PFA) to identify threshold values of repressing metabolites, i.e. ammonium, lactate, osmolality and carbon dioxide levels, on CHO growth and protein quality (glycosylation properties) but also to select significant inputs. PFA was applied by rotating principal components obtained by PCA and seeks physically meaningful linear combinations of variables. In their study, Xing et al. determined that ammonia and glucose negatively contributed to cell growth. Lactate and osmolality were positively correlated to cell growth and pCO_2 levels can reduce protein quality above a defined threshold. Multivariate analysis methods can be a powerful tool to determine the macroscopic stoichiometry of a biological system that cannot easily be determined by intuition. However, it becomes more complex to evaluate correlations and to apply this kind of statistical method with time-series data with varying number of metabolic phases, particularly in fed-batch cultures.

Another possibility to deal with this complexity is to use time series data analysis such as the Granger causality test. The Granger causality test is a statistical

hypothesis test used to determine causality among parameters. It was developed by Clive Granger (1934-2009), a British economist (Granger 1969). This test has recently been used to analyze transcriptomics and metabolomics profiles (Sriyudthsak, Shiraishi et al. 2013). Siryudthsak et al. introduced this test to evaluate causality among metabolites. Direct relationships between two metabolites were evaluated using the bivariate Granger causality test. This method has not yet been used to develop macroscopic metabolic reactions and to select the significant input parameter but it is expected to be applied in the future.

Statistical tools are useful when the underlying phenomena are too complex to resolve manually, such as multivariate data or temporal data. The two statistical methods presented above can help to structure problems, to reduce the dimensionality of the problem, to select relevant input and output parameters and to develop a macroscopic stoichiometric model.

1.4.3. Method based on metabolic network knowledge

The central idea is that the macroscopic behavior of cellular metabolism is the result of a combination of intracellular microscopic reactions that are more and more easily accessible via public databases. Metabolic networks are represented as a system of metabolite balance equations based on stoichiometric reactions. The general goal is to identify a minimal set of macroscopic reactions that can then build a sound basis for a macroscopic model.

1.4.4. Network construction

Metabolic network models of the central metabolism of mammalian cells have been built from the available genomic and biochemical information. Multiple databases can be used as resource for metabolic network reconstruction. As an example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa, Goto et al. 2014)and the BioCyc database collection (Caspi, Altman et al. 2014) are important databases that can be used to reconstruct a metabolic network. A number of studies have proposed metabolic networks of central metabolisms (Ahn and Antoniewicz 2012, Antoniewicz 2013, Zamorano, Vande Wouwer et al. 2013, Nicolae,

Wahrheit et al. 2014). To set up stoichiometric macroscopic relationships of cell metabolism, the main difficulty is the size of the metabolic network which can make the decomposition into external macroscopic reactions complex (Rügen, Bockmayr et al. 2012). To overcome this problem, metabolic networks can be reduced and simplified using computed fluxes in order to detect and remove insignificant pathways.

1.4.5. Metabolic flux analysis

Metabolic flux analysis (MFA) using metabolite balancing, first applied for microorganisms (Aiba and Matsuoka 1978), has been widely applied to mammalian cells. Metabolite balancing is a powerful method to quantify the manifestation of a phenotype (Varma and Palsson 1994, Goudar, Biener et al. 2006, Goudar, Konstantinov et al. 2009, Quek, Dietmair et al. 2010, Niklas, Schräder et al. 2011, Sengupta, Rose et al. 2011, Ahn and Antoniewicz 2012, Niklas and Heinzle 2012, Antoniewicz 2013, Grafahrend-Belau, Junker et al. 2013, Klein, Heinzle et al. 2013, Wahrheit, Niklas et al. 2014). Metabolite balancing is based on the assumption that accumulation of intracellular metabolites is insignificant compared to the extracellular fluxes in batch and fed-batch cultures (Niklas and Heinzle 2012). This assumption is valid for small concentration of intracellular metabolites which is usually fulfilled but may deviate to a certain extent for highly concentrated metabolites, e.g. of the TCA cycle (Rehberg, Rath et al. 2014). Based on this quasi-steady state assumption, we can say that the sum of influxes and effluxes of an internal metabolite of a metabolic network is equal to zero.

$$\boldsymbol{S} \cdot \boldsymbol{v} = \mathbf{0} \qquad (1.2)$$

where S is a stoichiometric matrix, based on a defined metabolic network, with each row corresponding to a balanced internal metabolite and each column corresponding to a flux in the flux vector, v.

We can then split equation 1.2 to have on one side, the fluxes that are experimentally measured (substrates, products, biomass), v_m, and on the other side, fluxes that will be calculated by MFA, v_c.:

$$\boldsymbol{S_m \cdot v_m = -S_c \cdot v_c} \quad (1.3)$$

S_m and S_c are the stoichiometric matrices associated to v_m and v_c respectively. If S_c is a square matrix of full rank, the fluxes are calculated by:

$$\boldsymbol{v_c = -(S_c)^{-1} \cdot S_m \cdot v_m} \quad (1.4)$$

The uptake and production rates of metabolites are such measurable external fluxes that can be related to the specific growth rate, μ, by yield coefficients $\boldsymbol{Y_{Met/Bio}}$:

$$\boldsymbol{v_{m,i} = \mu \cdot Y_{Met/Bio}} \quad (1.5)$$

Monte-Carlo simulation can be used to get a more precise and realistic estimation of the standard deviation of the calculated fluxes. A dynamic metabolic flux analysis can also be performed in order to have the profile of the intracellular flux over time (Niklas, Schräder et al. 2011, Wahrheit, Nicolae et al. 2014). When metabolite balancing is performed, reactions with insignificant fluxes can be identified and then deleted from the metabolic network to simplify it.

1.4.6. Elementary flux mode analysis (EFM)

EFM analysis can then be applied on a metabolic network as defined in equation 1.2. EFM analysis is the calculation of independent, minimal biochemical pathways in a metabolic network at steady-state, which are thermodynamically and stoichiometrically possible taking into account the irreversibility or the reversibility of the reactions (Schuster, Dandekar et al. 1999). There is a distinction between external and internal metabolites. A 'flux mode' is a steady-state flux distribution in which the proportions of fluxes are fixed and it is called 'elementary' if it is not decomposable. Various software can be used for this purpose such as COPASI (Hoops,

Sahle et al. 2006), Metatool (Schuster and Schuster 1993), efmtool (Terzer and Stelling 2008) or CellNetAnalyser (Klamt and von Kamp 2011).

To perform EFM, the stoichiometric matrix based on a metabolic network is used, and the convex basis vectors are computed using equation 1.2, taking into account the thermodynamic feasibility constraints (Schuster, Dandekar et al. 1999). Any possible flux distribution v can be expressed as a non-negative linear combination of a set of elementary flux vectors e_i which represent the not decomposable metabolic paths between the substrates and the final products:

$$\boldsymbol{v = \omega_1 e_1 + \omega_2 e_2 + \cdots + \omega_p e_p} \qquad \boldsymbol{\omega_i \geq 0} \tag{1.6}$$

The non-negative matrix E with column vectors e_i satisfies $S \cdot E = 0$. E constitutes the admissible flux space also known as the convex polyhedral cone (Gagneur and Klamt 2004). However, a critical issue in EFM is the calculation of these elementary flux vectors because of dramatically increasing computational demands with increasing network size. Based on the matrix E, a set of macroscopic reactions of the extracellular metabolites can be derived (Provost and Bastin 2004, Provost, Bastin et al. 2006, Gao, Gorenflo et al. 2007, Dorka, Fischer et al. 2009, Baughman, Huang et al. 2010, Niu, Amribt et al. 2013, Zamorano, Vande Wouwer et al. 2013). Examples of the stoichiometric matrix E are presented in Table 1.1. A methodology was proposed by Junger et al. (Jungers, Zamorano et al. 2011) to compute minimal elementary decompositions of metabolic flux vectors. Later Zamorano et al. (Zamorano, Vande Wouwer et al. 2013) showed that this method allows the estimation of metabolic fluxes even with an underdetermined mass balance system where data are not sufficient to uniquely define these fluxes. This provides also an excellent basis for setting up macroscopic models.

The output of these approaches are stoichiometric macroscopic relationships of cell metabolism based on a metabolic network and on biological and biochemical knowledge that provide a necessary input for kinetic macroscopic models.

1.5. Macroscopic kinetic models

After a first screening to select input parameters and to set up the stoichiometric macroscopic reactions, the macroscopic kinetic reactions can be developed. Different types of kinetics are available and this section will present some of the most important ones.

1.5.1. Monod model and its derivatives

For modeling of mammalian cell culture kinetics, the Monod equation and derivations of it are most frequently applied. These kinetics with slight modifications are capable to simulate different types of characteristics like saturation, inhibition and limitation by substrates and other components.

For Monod kinetics the growth is defined as:

$$\boldsymbol{\mu = \mu_{max}\left[\prod \frac{S_i}{S_i + K_{Si}}\right]} \tag{1.7}$$

Where μ is the specific growth rate; S_i and K_{Si} are the corresponding substrate concentration and half-saturation constant, respectively. μ_{max} is the maximum specific growth rate. To incorporate inhibitory effects a corresponding term is added to the denominator. In the case of balanced growth all other rates can be related to μ by yield coefficients (Equation 1.5).

Table 1.1 Stoichiometric matrices of macroscopic reaction networks for CHO cell lines

	e1	e2a	e2b	e3	e4	e5	e6	e7	e8	e9	e10
Glucose	-1	-1	-1	-1	0	0	0	-0.0508	0	0	0
Gln	0	-2	0	0	0	0	-1	-0.0577	-0.0104	-1	0
Lac	2	0	2	2	0	0	0	0	0	0	0
Glu	0	2	-2	-2	-1	0	1	-0.0016	-0.0107	1	-1
Asn	0	0	0	0	0	-1	1	-0.006	-0.0072	0	0
Asp	0	0	0	0	0	1	-1	-0.0201	-0.0082	0	0
Ala	0	2	2	0	0	0	0	-0.0133	-0.011	0	0
Pro	0	0	0	0	1	0	0	-0.008	-0.0148	0	0
BM	0	0	0	0	0	0	0	1	0	0	0
Mab	0	0	0	0	0	0	0	0	1	0	0

(Dorka, Fischer et al. 2009)

	e1	e2	e3	e4	e5	e6	e7
Glc	-1	-1	0	0	0	-1	-1
Gln	0	0	-1	-1	-1	-3	-2
Lac	2	0	0	1	0	0	0
NH_3	0	0	1	2	2	1	1
Ala	0	0	1	0	0	0	0
CO_2	0	6	2	2	5	2	2
Nucl	0	0	0	0	0	1	1

(Provost and Bastin 2004)

	e1	e2	e3	e4	e5	e6	e7	e8	e9
Glucose	-1	-1	-1	0	0	0	-0.0508	0	0
Gln	0	0	0	0	0	-1	-0.0577	-0.0104	-1
Lac	2	2	2	0	0	0	0	0	0
NH_3	0	0	0	0	1	0	0	0	1
Glu	0	-2	-2	-1	0	1	-0.0016	-0.0107	1
Asn	0	0	0	0	-1	1	-0.006	-0.0072	0
Asp	0	0	2	0	1	-1	-0.0201	-0.0082	0
Ala	0	2	0	0	0	0	-0.0133	-0.011	0
Pro	0	0	0	1	0	0	-0.0081	-0.0148	0
CO2	0	2	6	0	0	0	0	0	0
BM	0	0	0	0	0	0	1	0	0
Mab	0	0	0	0	0	0	0	1	0

(Gao, Gorenflo et al. 2007)

	e1	e2	e3	e4	e5	e6	e7	e8	e9
Glc	-1	-1	-1	0	0	0	-0.0508	0	0
Gln	0	0	0	0	0	-1	-0.0577	-0.0104	-1
Lac	2	2	2	0	0	0	0	0	0
NH_3	0	0	0	0	1	0	0	0	1
Glu	0	-2	-2	-1	0	1	-0.0016	-0.0107	1
Asn	0	0	0	0	-1	1	-0.006	-0.0072	0
Asp	0	0	2	0	1	-1	-0.0201	-0.0082	0
Ala	0	0	0	1	0	0	-0.0081	-0.0148	0
Pro	0	2	0	0	0	0	0	0	0
BM	0	0	0	0	0	0	1	0	0
Mab	0	0	0	0	0	0	0	1	0

(Baughman, Huang et al. 2010)

Glc: Glucose; Lac: Lactate; BM: Biomass; Mab: Monoclonal antibody; Nucl: Nucleotides; amino acids are specified using the standard three-letter code.

There are two methods for estimation of the specific growth rate, μ, and the associated Monod constant, K_{Si}. One is the steady-state measurement of growth and the limiting substrate concentration in continuous culture at different dilution rates. An alternative method is the measurement of growth rate at different substrate concentrations in batch culture (Banerjee 1993). To estimate the specific growth rate, μ, the associated Monod constant, K_{Si}, were arbitrarily set to small values to obtain balanced growth (Provost, Bastin et al. 2006, Dorka, Fischer et al. 2009). This seems well justified for batch cultures but will not allow to transfer such a model to continuous or fed-batch processes without readjustment of these constants. Monod-type models are widely used, but it is often difficult to define which formulation is the best to characterize the cell behavior (Bastin and Dochain 1990). Furthermore, the finding the optimal formulation of this kind of model and estimating model parameters can be time-consuming. Table 1.2 presents kinetic growth models used in the literature to describe growth of different organisms. Generally, only the growth rate is described by a Monod-type model, the other components, products and substrates, are then described by simple mass balance equations (Sainz, Pizarro et al. 2003, Baughman, Huang et al. 2010, Xing, Bishop et al. 2010, Borchers, Freund et al. 2013) also called first principle models (FPMs) . To describe the relationship between the variation of substrates and products with the cell number, the mass balance equations are defined as a set of ordinary differential equations (ODEs) based on biological knowledge and taking into account the inner structure of the cells. Often the specific consumption/production rates are assumed to be proportional to the specific growth rate during the process but this is not always the case. Monod type models can also be used to describe other specific consumption/production rates of metabolites independent of growth (Provost and Bastin 2004, Gao, Gorenflo et al. 2007, Dorka, Fischer et al. 2009, Baughman, Huang et al. 2010). Batt and Kompala (Glacken, Adema et al. 1988, Batt and Kompala 1989, Ozturk, Riley et al. 1992) described a four-compartment structured model to describe growth of hybridoma and monoclonal antibody productions using Monod and Haldane type kinetic models

Table 1.2 Kinetic models for mammalian cell growth

Kinetic parameters	Cells	References	Growth equations
μ_{max}= 0.125 h^{-1} k_{gln}= 0.8 mM k_{lac}= 8 mM k_{amm}= 1.05 mM	Hybridoma	(Bree, Dhurjati et al. 1988)	$\mu = \mu_{max} * \frac{\boldsymbol{Gln}}{k_{gln} + \boldsymbol{Gln}} * \frac{k_{lac}}{k_{lac} + \boldsymbol{Lac}} * \frac{k_{amm}}{k_{amm} + \boldsymbol{Amm}}$
μ_{max}= 0.055 h^{-1} $(k_s)_0$= 26.5 β= 0.21 k_{gln}= 0.15 mM k_{amm}= 26 mM^2	Hybridoma	(Glacken, Adema et al. 1988)	$\mu = \mu_{max} * \frac{\boldsymbol{Ser} * \boldsymbol{Gln}}{(\boldsymbol{Ser} + (k_s)_0 * X^{-\beta}) * (k_{gln} + \boldsymbol{Gln}) * (1 + \frac{\boldsymbol{Amm}^2}{k_{amm}})}$
μ_{max}= 0.053 h^{-1} k_{serum}= 0.0139 v/v	Hybridoma	(Ozturk and Palsson 1991)	$\mu = \mu_{max} * \frac{\boldsymbol{Serum}}{k_{serum} + \boldsymbol{Serum}}$
μ_{max}= 0.045 h^{-1} k_{glc}= 1 mM k_{gln}= 0.3 mM	Hybridoma	(de Tremblay, Perrier et al. 1992)	$\mu = \mu_{max} * \frac{\boldsymbol{Glc}}{k_{glc} + \boldsymbol{Glc}} * \frac{\boldsymbol{Gln}}{k_{gln} + \boldsymbol{Gln}}$
μ_{max}= 0.036 h^{-1} k_{gln}= 0.06 mM	Hybridoma	(Pörtner, Schilling et al. 1996)	$\mu = \mu_{max} * \frac{\boldsymbol{Gln}}{k_{gln} + \boldsymbol{Gln}}$
μ_{max}= 0.689 h^{-1} k_{glc}= 4.79 mM k_{gln}= 0.032 mM k_{lac}= 0.67 mM k_{rd}= 0.019 h^{-1} k_A= 0.275 h^{-1}	Hybridoma	(Dhir, Morrow et al. 2000)	$\mu = \mu_{max} * \frac{\boldsymbol{Glc}}{k_{glc} + \boldsymbol{Glc}} * \frac{\boldsymbol{Gln}}{k_{gln} + \boldsymbol{Gln}} * \frac{k_{lac}}{k_{lac} + \boldsymbol{Lac}} - k_{rd} \frac{\boldsymbol{Lac}}{k_{dLac} + \boldsymbol{Lac}} - k_A * \boldsymbol{Amm}$
μ_{max}= 0.065 h^{-1} k_{glc}= 0.75 mM k_{gln}= 0.075 mM k_{lac}= 90 mM k_{amm}= 15 mM	Hybridoma	(Jang and Barford 2000)	$\mu = \mu_{max} * \frac{\boldsymbol{Glc}}{k_{glc} + \boldsymbol{Glc}} * \frac{\boldsymbol{Gln}}{k_{gln} + \boldsymbol{Gln}} * \frac{k_{lac}}{k_{lac} + \boldsymbol{Lac}} * \frac{k_{amm}}{k_{amm} + \boldsymbol{Amm}}$
μ_{max}= 0.028 h^{-1} k_{glc}= 0.084 mM k_{gln}= 0.047 mM k_{lac}= 43 mM k_{amm}= 6.51 mM	CHO	(Xing, Bishop et al. 2010)	
μ_{max}= 0.0190 h^{-1} k_{glc}= 1.45 mM	AGE1.HN	(Borchers, Freund et al. 2013)	$\mu = \mu_{max} * \frac{\boldsymbol{Glc}}{k_{glc} + \boldsymbol{Glc}} *$

Glc: Glucose; Gln: Glutamine; Lac: Lactate; Amm : Ammonium.
*kinetic model of growth in 500mL stirred tank reactors

1.5.2. Logistic equation

Verhulst (Vogels, Zoeckler et al. 1975) developed the first logistic equation to describe population growth based on the work of Thomas Malthus. Verhulst added an extra

term, *K*, called the *overall saturation constant* to the first model of Malthus to represent the resistance to growth up to a certain limit value of biomass concentration as shown in the equation describing *logistic growth*.

$$\frac{dX(t)}{dt} = \boldsymbol{r} \cdot \boldsymbol{X(t)} \cdot (\mathbf{1} - \frac{\boldsymbol{X(t)}}{\boldsymbol{K}}) \tag{1.8}$$

This model does not take into account the death of cells, either by necrosis or apoptosis, observed in mammalian cell processes. Therefore cell growth and death have been taken into account in an alternative formulation, the so-called four-parameter-generalized-logistic-equation which can describe cell density profiles in batch and fed-batch cultures (Jolicoeur and Pontier 1989).

$$\boldsymbol{X(t)} = \frac{\boldsymbol{A}}{\boldsymbol{e^{Bt} + Ce^{-Dt}}} \tag{1.9}$$

Where *X(t)* is the cell density at time t. *A* is related to the initial value of *X* while *B* and *C* correspond to the maximum death and growth rate, respectively. Goudar (Goudar 2012) applied such logistic modeling in batch and fed-batch cultures of mammalian cells. To describe the cell culture process system, besides equation 1.9 two types of equations were used for the formation of products and for substrate consumption.

The second type of equation used was the *logistic growth equation* to describe monotonously increasing quantities of product concentrations, *P*, such as lactate and ammonium

$$\boldsymbol{P(t)} = \frac{\boldsymbol{A}}{\mathbf{1} + \boldsymbol{Ce^{-Dt}}} \tag{1.10}$$

Finally the logistic *decline equation* has been used to describe monotonously decreasing quantities of nutrient concentration, *N*, such as glucose or glutamine concentration.

$$N(t) = \frac{A}{e^{Bt+C}} \tag{1.11}$$

To get robust logistic modeling, initial estimation of parameters using linearization have been successfully used. More complex equations can be used. For instance, Acosta et al. (Acosta, Sánchez et al. 2007) use two asymmetric logistic equations for growth and nutrients and products. Logistic equations have been successfully used in a variety of applications to describe the dynamic of population growth, most of them involved bacterial growth (Gibson, Bratchell et al. 1987, Tsoularis and Wallace 2002) but also mammalian cell growth (Goudar, Joeris et al. 2005, Goudar, Konstantinov et al. 2009, Goudar 2012, Goudar 2012). This kind of models are particularly useful if the matrix *S* from equation 1.2 is not known.

The main differences between the logistic equation and the Monod model are that the logistic equation uses fewer parameters compared to the Monod model and that it does not require knowledge about limiting substrates. That makes the computational step from logistic approach simpler than classical approaches but seems less suited for extrapolation and not suited to incorporate additional information on metabolism.

1.5.3. Neural networks and hybrid models

Neural Networks (NNs) are computational models of black-box type. They are used to model a wide spectrum of problems. NNs are an interconnected network structure composed of a set of processing elements (PEs) (Price and Shmulevich 2007). Giving some input, computations are made using the transfer functions of the network to estimate the output. The network is composed of different layers: the input layer, the hidden layers and the output layer. The PEs are composed of transfer functions

(polynomial, hyperbolic, kernel, …) and the significance of the connection is called the weight.

Marique et al. (Marique, Cherlet et al. 2001) used a NN to simulate nonlinear kinetics of CHO strains. For the transfer function, a classical sigmoid function was applied. Biomass, glucose, glutamine, lactate and ammonia concentrations represented output and input layers. A model with CHO K1 of those five variables was obtained by using only one hidden layer. Moreover, the same NN has been used to predict the behavior of another cell line (CHO TF70R) by adjusting the time scale. As described above, mechanistic knowledge is not needed to create NNs. Nevertheless, hybrid neural networks are more used since a decade, combining non-parametric functions such as NNs and parametric functions based on cell culture process knowledge (van Can, te Braake et al. 1999, Vande Wouwer, Renotte et al. 2004, Laursen, Webb et al. 2007). Laursen et al. combined material balances to estimate accumulation rates of biomass, product and metabolites in a bacterial fed-batch culture combined with a NN for each variable. Vande Wouver et al. (Vande Wouwer, Renotte et al. 2004) used several hybrid NNs to describe CHO batch cultures. A set of NNs for the calculation of the reaction rates was combined with material balances of a bioreactor (Chen, Bernard et al. 2000). Teixeira et al. (Teixeira, Alves et al. 2007) used EFM to reduce the metabolic network of a recombinant Baby Hamster Kidney (BHK-21A) cell line producing a glycoprotein (IgG1-IL2) to a minimal set of macroscopic reactions which then served as a basis for a hybrid NN model. A three-layered backpropagation neural network was used as a non-parametric function to describe the kinetics of the system. By using this hybrid model in a fed-batch process, they were able to increase the final productivity of IgG1-IL2 by 10% (Teixeira, Alves et al. 2007). Graefe et al. (Graefe, Bogaerts et al. 1999) applied a serial hybrid model to CHO-K1 by combining mass balances and neural network kinetics. A convincing prediction of components concentrations in the stirred tank bioreactor was achieved.

Hybrid models exploit the advantages of parametric models (“grey box model”) and of non-parametric models (“Black box models”) and overcome the limits of each used

individually. For complex problems, this kind of methodology provides a good benefit/cost ratio.

To conclude, logistic models and neural network models do not or only partially consider the underlying physical, biological phenomena. Nevertheless, they are less difficult to develop than mechanistic models. There parameters are hardly physically interpretable in contrast to mechanistic models that take into account the underlying phenomena including mass balances which supports biological understanding. Moreover, mechanistic type models are generally more suited for extrapolation outside the experimentally explored space. For very complex systems with limited mechanistic knowledge available, logistic models and NNs can be useful due to the lower number of parameters to identify. Hybrid models take the advantages of both approaches, the mechanistic approach and the empirical/semi-empirical approach, e.g. by improving model extrapolation compared to a pure NN (Van Can, Te Braake et al. 1998) but require complex optimization tools to calibrate them.

1.6. Model calibration and testing

Before starting to identify model parameters, it is important to identify and remove outliers. Outliers can increase the level of variance of the model parameters (Yang, Martin et al. 2011), can reduce the model performance by biasing parameter estimates and can lead to false conclusion. Outliers are often due to fault, biological deviations or human/instrumental errors. For instance Borchers et al. (Borchers, Freund et al. 2013) defined an outlier detection approach for AGE1.HN cell line based on a model (model generic approach) by introducing an additional pessimistic bound (relative error). Then, they identified model parameters and performed a reachability analysis. The outliers were then selected by comparing the reachable state sets with the measurements data. There are many other possible methods to identify outliers like the locally estimated scatterplot smoother (LOESS) (Sriyudthsak, Shiraishi et al. 2013) or using splines (Laursen, Webb et al. 2007) but most of them depend on the context of the experiment, the equipment performed and the variables analyzed.

Kinetic parameters are usually determined by fitting the model to the experimental data. Parameter estimation is an optimization problem, in which an objective or cost function characterizing the deviation of a model prediction from the experimental data is minimized by adjusting model parameters. Typically least squares or the maximum likelihood functions are applied. Together with the usually non-linear differential equations of the model, non-convex problems result that are hard to solve but powerful algorithms and mathematical tools have been developed to treat them. These were successfully applied to macroscopic models of mammalian cells. For instance Borchers et al. (Borchers, Freund et al. 2013) used a semi-definite programming (SDP) algorithm to solve a polynomial function by reformulating and relaxing the non-convex constraint problem into a convex optimization problem, whereas Baughman et al. (Baughman, Huang et al. 2010) used a simple discretization scheme combined with a filtered interior point primal dual line search algorithm (IPOPT) to identify global optima for the non-convex problem.

The choice of optimization algorithms depends on the type of optimization problem, the number of parameters and variables, the constraints, the model but also software availability, e.g. gOPT from gPROMs (Kontoravdi, Asprey et al. 2007), ADMIT toolbox (Streif, Savchenko et al. 2012, Borchers, Freund et al. 2013) and MATLAB (MathWorks, Natick, MA) (Sainz, Pizarro et al. 2003, Vande Wouwer, Renotte et al. 2004, Teixeira, Alves et al. 2007).

Goudar et al. (Goudar, Konstantinov et al. 2009) compared the simplex method, the generalized reduced gradient method (GRG) and the Levenberg-Marquard algorithm (LMA) for nonlinear parameter estimation of logistic model parameters of batch and fed-batch mammalian cell culture. The simplex method and GRG methods resulted in a better fit than LMA. LMA was also used by Vande Wouwer et al. (Vande Wouwer, Renotte et al. 2004) for batch CHO cell culture. LMA was applied in the training process for hybrid models of bioprocesses (Graefe, Bogaerts et al. 1999). Dorka et al. (Dorka, Fischer et al. 2009) successfully identified Monod-type parameters of hybridoma cell culture during exponential phase by using quadratic programming

(QP). For the post-exponential phase, the maximal rates and the half saturation constants were calibrated using a Markov chain Monte Carlo method (MCMC) using a Metropolis-Hasting algorithm. More examples of applications of optimization algorithms for macroscopic modeling of mammalian cells can be found in a number of studies, e.g. the method of Powell (Glacken, Adema et al. 1988), MCMC (Xing, Bishop et al. 2010), linear programming (Sainz, Pizarro et al. 2003), the particle swarm algorithm (PSO) (Selişteanu, Şendrescu et al. 2015), sequential quadratic programming (SQP) (Kontoravdi, Asprey et al. 2007) and a quasi-Newton method (Teixeira, Alves et al. 2007). It is not possible to *a priori* recommend any single algorithm as the superior method as it is problem-dependent.

Major problems of parameter estimation in non-linear systems are the potential existence of multiple local minima and over fitting. Additionally, models have to be assessed for their predictive power and their robustness against perturbations.

Model validation is one of the most critical parts of the modeling process. We can identify two ways to evaluate the quality of a model: one called direct validation compares the model prediction with the same experimental data as used to estimate the parameters (Goudar, Konstantinov et al. 2009, Selişteanu, Şendrescu et al. 2015). The second method uses an independent new data set to validate or invalidate the model (cross validation). For instance Xing et al. (Xing, Bishop et al. 2010) identified the parameters on three independent sampling trains with different initial parameters and then used two types of validation. The first one validated the model by applying the model to different cell cultures to assess the applicability of the model. Secondly, they applied the model to a perturbed system to assess the accuracy of the model. For hybrid neural models, two data sets, one training/calibration data set to identify hybrid model parameters and one validation data set to assess the model quality are usually used (Oliveira 2004, Vande Wouwer, Renotte et al. 2004, Teixeira, Alves et al. 2007).

A more complex methodology was used by Borchers et al. (Borchers, Freund et al. 2013) for the invalidation of models and for parameter estimation. Their set-based

method builds on a semi-definite programming relaxation and outer-bounding techniques supported by the ADMIT toolbox (Streif, Savchenko et al. 2012).

Another important method to assess the quality of a model is to perform sensitivity analysis that can provide valuable information regarding the importance of parameters on the model output and on the possible impact of variability of the input on the output. For instance, one can evaluated the largest possible variation of the parameters which does not lead to rejection of the model (Borchers, Freund et al. 2013). Baughman et al. (Baughman, Huang et al. 2010) quantified the impact of the linear discretization on the parameters and on the numerical error. Moreover, the impact of possible measurement variability on model estimate has been performed by using Monte Carlo simulation using normal distribution (Baughman, Huang et al. 2010). Global sensitivity analysis (GSA) has the advantage of evaluating the effect of a factor while all other factors are varied simultaneously (Kiparissides, Rodriguez-Fernandez et al. 2008). For instance Kontoravdi et al. (Kontoravdi, Asprey et al. 2007) used the Sobol' global sensitivity method to assess the sensitivity of the parameters of dynamic hybridoma model and, based on the same case study, Kiparissides et al. (Kiparissides, Rodriguez-Fernandez et al. 2008) evaluated the performance of the Sobol' method and derivative based global sensitivity measures (DGSM) as a GSA method. The DSGM method was identified as more useful than the Sobol' method due to the lower computational requirement while producing the same quality of results.

1.7. Application of models for control of processes

An important application for industrial production is the use of macroscopic models for the control of production processes. Generally, models applied for control should be simple and variations of process conditions and cell characteristics can be taken into account by adapting the model parameters online. It is most straightforward to use a stoichiometric model together with dynamic material balances to estimate the state of a culture. A feeding strategy can be determined in a fed-batch process based on the model and on defining an objective to reach. For instance, Haas et al. (Haas,

Lane et al. 2001) used an open-loop-feedback-optimal controller to maintain glucose and glutamine at low levels in a culture of a hybridoma cell line producing an IgG antibody. This controller was based on a Monod-type model of growth in which the parameters and the state are estimated, and then an optimization part calculates an optimal feeding profile. Teixeira et al. (Teixeira, Alves et al. 2007) also used a controller with an hybrid model on a culture of a BKH cell line producing an IgG1-IL2. The glucose and glutamine feeding rate was optimized to maximize the total amount of antibody produced at the end of the experiment. Finally, Craven et al. (Craven, Whelan et al. 2014) used a non-linear model predictive controller (NMPC) in a CHO cell line to control the glucose concentration. The kinetic models used were of Monod type.

1.8. Conclusion and outlook

As was described in this review, setting up of macroscopic models is carried out in primarily two steps (Figure 1.1) After identification of the stoichiometric part of a model, kinetics for growth and metabolite conversion are defined to yield relatively simple yet useful combined models. As in most other cases of modeling, macroscopic modeling of mammalian cell cultures is an iterative process of setting up a model, calibrating, validating and testing it, designing and performing new experiments and revising the model.

Macroscopic modeling of metabolism can be used in many applications to accelerate cell line selection, medium optimization, feeding strategy development, and other bioprocess development activities. By using macroscopic models, it is possible to understand what the significant parameters are that have an impact on the cell culture process and then predict how the process will evolve if one parameter is changed. Having predictive models of cell culture processes can be a powerful tool to help identifying the critical process parameters which have an impact on CQAs, e.g. glycosylation, and to optimize process performance with respect to a defined objective function. Therefore, macroscopic models are more and more used by

biopharmaceutical companies. They also assist in fulfilling requirements of the QbD methodology by providing a handle to further improve CQAs.

For fed-batch cell culture processes, macroscopic models can be applied to predict the time courses of metabolites which have significant impact on the cell culture process or to estimate process rates of interest and then control feeding rates based on model prediction and using an appropriate objective function.

Different types of models can be used to select variables and determine the macroscopic reactions of the system and then, different kinetic models can be applied to simulate and predict the macroscopic behavior of the cells. All types of combinations of those model can be applied; for example, a stoichiometric model using the EFM method combined with a logistic kinetic equation or empirical stoichiometric relationships identified with a PCA combined with a Monod-type kinetic equations and so on. Stoichiometry derived by the EFM method are used together with NNs to result in so-called hybrid neural network models. Such kind of hybrid models are more of grey-box rather than black-box type (van Can, te Braake et al. 1999, Vande Wouwer, Renotte et al. 2004, Laursen, Webb et al. 2007). The choice of the model depends on the aim of the study but also on the complexity of the system we want to simulate and understand. As summary of strategies used to develop macroscopic models with mammalian cells from various studies is presented on Table 1.3.

Table 1.3 Applied strategies to develop macroscopic models for mammalian cells

input-output relationship methodology	Kinetic model	Model variables	Parameter estimation	Cell line	Reference
-	Logistic	Cell density NH_3 Lactate Glucose Gln	LMA Simplex GRG	CHO BHK Hybridoma	(Goudar, Konstantinov et al. 2009)
Metabolic network	Hybrid model (Serial and Parallel)	Cell density Lactate Glucose Gln	LMA	CHO	(Vande Wouwer, Renotte et al. 2004)
Metabolic network +EFM	Monod type	Cell density NH_3 Lactate Gln Asp Asn Pro Mab	Quadratic programing (exponential phase) MCMC method (post exponential phase)	Hybridoma	(Gao, Gorenflo et al. 2007)
Expert reasoning	Monod type Inhibition type	Cell density Glucose CO_2 Gln Lactate NH_3 Mab	Literature data	Hybridoma	(Glacken, Adema et al. 1988, Batt and Kompala 1989, Ozturk, Riley et al. 1992)
Expert reasoning	Monod type	Cell density Gln NH_3 Mab	Powel method	Hybridoma CRL-1606	(Glacken, Adema et al. 1988)
Expert reasoning	Monod type / First principle	Cell density Glucose Gln Lactate NH_3 Mab	MCMC method	CHO	(Xing, Bishop et al. 2010)
PFA	Monod type Canonical	Cell density Glucose CO_2 Gln Lactate NH_3 Mab	Monte Carlo simulation And Canonical algorithm	CHO	(Xing, Li et al. 2008)
Metabolic network	First principle	Cell density Glucose NH_3 Ethanol Glycerol	Linear programming	Yeast	(Sainz, Pizarro et al. 2003)
Expert reasoning	Hybrid serial	Cell density Glucose Ethanol	Large scale SQP	Baker's yeast	(Oliveira 2004)
Expert reasoning (from Gao et al.)	Monod-type	Cell density NH_3 Lactate Gln Asp Asn Pro Mab	PSO	Hybridoma 130-8F	(Selişteanu, Șendrescu et al. 2015)
Metabolic network + EFM	Hybrid model	Cell density Glucose Gln Lactate NH3 Ala Mab	Quasi-newton algorithm with conjugate gradient line search	BHK-21A	(Teixeira, Alves et al. 2007)
Expert reasoning	First principle	Cell density Glucose Gln Lactate NH_3	SDP	AGE1.HN	(Borchers, Freund et al. 2013)
Expert reasoning	Monod-type / First principle	Cell density Glucose Gln Lactate NH_3 Mab Glycosylation	SQP	Hybridoma 14-4-4S	(Kontoravdi, Asprey et al. 2007)
Expert reasoning (From Gao et al.)	Monod-type	Cell density NH_3 Lactate Gln Asp Asn Pro Mab	IPOPT	Hybridoma 130-8F	(Baughman, Huang et al. 2010)

Chapter 2

2 Segmented linear modeling of CHO fed-batch culture and its application to large scale production

2.1. Abstract

We describe a systematic approach to model CHO metabolism during biopharmaceutical production across a wide range of cell culture conditions. To this end, we applied the metabolic steady state concept. We analyzed and modeled the production rates of metabolites as a function of the specific growth rate. First the total number of metabolic steady state phases and the location of the breakpoints were determined by recursive partitioning. For this the smoothed derivative of the metabolic rates with respect to the growth rate were used followed by hierarchical clustering of the obtained partition. We then applied a piecewise regression to the metabolic rates with the previously determined number of phases. This allowed identifying the growth rates at which the cells underwent a metabolic shift. The resulting model with piecewise linear relationships between metabolic rates and the growth rate did well describe cellular metabolism in the fed-batch cultures. Using the model structure and parameter values from a small scale cell culture (2 L) training dataset, it was possible to predict metabolic rates of new fed-batch cultures just using the experimental specific growth rates. Such prediction was successful both at the laboratory scale with 2 L bioreactors but also at the production scale of 2000 L. This type of modeling provides a flexible framework to set a solid foundation for metabolic flux analysis and mechanistic type of modeling.

This chapter was published as

Ben Yahia, B., Gourevitch, B., Malphettes, L., Heinzle, E., 2017. Segmented linear modeling of CHO fed-batch culture and its application to large scale production. Biotechnology and Bioengineering. 114(4):785-797.

2.2. Introduction

Fed-batch cultivation of Chinese hamster ovary (CHO) cells is a widely used technology for the production of therapeutic glycosylated proteins (Sidoli, Mantalaris et al. 2004, Niklas and Heinzle 2012, Tsang, Wang et al. 2014, Tescione, Lambropoulos et al. 2015). So far, process development of mammalian cells producing monoclonal antibodies (mAb) and other biopharmaceuticals has been largely done by designing and performing experiments in an empirical manner. In the attempt to understand the mechanism determining such production in-depth, systems biology methods, i.e. genomic, transcriptomic, proteomic and metabolomic analyses are increasingly applied together with associated modeling. For process development mainly two methods were already used to get a better understanding of cellular metabolism as a basis for process optimization. On the one side, mechanistic metabolic modeling (Ashyraliyev, Fomekong-Nanfack et al. 2009, Hu and Zhou 2012) is used to describe the physiological behavior of cells and further to optimize cultivation and production (Dorka, Fischer et al. 2009, Nolan and Lee 2011). On the other side, metabolic flux analysis (Niklas and Heinzle 2012) quantifies the intracellular fluxes and therefore provides a better understanding of cellular physiology (Provost and Bastin 2004, Provost, Bastin et al. 2006, Dorka, Fischer et al. 2009, Jungers, Zamorano et al. 2011, Naderi, Meshram et al. 2011, Nolan and Lee 2011, Amribt, Niu et al. 2013, Meshram, Naderi et al. 2013, Zamorano, Vande Wouwer et al. 2013). During cultivation, cells adapt to the extracellular environment that changes due to the successive consumption and depletion of substrates and the accumulation of waste byproducts. Different metabolic phases linked by metabolic shifts are the results of this (Wahrheit, Nicolae et al. 2014, Wahrheit, Niklas et al. 2014, Mulukutla, Yongky et al. 2015). This makes mechanistic modeling of the metabolism for the duration of a whole bioproduction difficult. Metabolic flux analysis requiring a metabolic steady state is then only applicable for each metabolic phase individually. However, metabolic phases with metabolic steady state have to be identified first (Provost and Bastin 2004, Provost, Bastin et al. 2006, Niklas and

Heinzle 2012). Usually, cell cultivation is divided into phases based on the growth profile (Altamirano, Illanes et al. 2001, Altamirano, Illanes et al. 2006, Niklas, Schräder et al. 2011, Wahrheit, Nicolae et al. 2014). This procedure is performed manually from visual inspection of cell growth (Dean and Reddy 2013, Fan, Jimenez Del Val et al. 2015) but can also be based on non-linear models such as Neural Network (Simon, Karim et al. 1998) or on a structural approach (Borchers, Freund et al. 2013). However, these methods focus on growth phases and may thus miss metabolic shifts that are only seen by observing the yield coefficients between metabolite consumption/production and cell growth. Identification of growth phases based on growth profiles is even more difficult in fed-batch cultures with their varying conditions. To overcome this problem, the concept of metabolic-steady state has been applied and extended. Under such conditions, intracellular fluxes or, at least, flux ratios remain constant. Moreover, biomass yields on substrates as well as on all precursor molecules are constant; that can be proven by the identification of linear correlations between metabolic rates (Deshpande, Yang et al. 2009).

The aim of the prevailing work is to provide a systematic methodology for identifying metabolic phases and for simulating the evolution of cell metabolism based on the relationship between external metabolite rates and the specific growth rate. For that purpose, segmented linear regression, also called piecewise regression, was used (McGee and Carleton 1970, Muggeo 2003, Toms and Lesperance 2003). In segmented regression models two or more regression lines are joining at unknown points, called breakpoints. Using the growth rate as a criterion to identify metabolic phases and predict cell metabolism provides the unique possibility to compare various states of growth. Finally, this new methodology can be used without any assumed metabolic network model. This method is illustrated with the example of a Chinese hamster ovary (CHO) cells cultivated in fed-batch production at 2 L scale that was used to establish and calibrate the piecewise model. The model was then validated for its applicability for scaling up to a production scale bioreactor of 2000 L.

2.3. Modeling and theoretical aspects

2.3.1. General Representation and Metabolic Steady-State Assumption

Metabolic phases are defined by a metabolic steady state such that intracellular metabolite concentrations remain constant within a phase (Provost, Bastin et al. 2006). If all intracellular concentrations remain constant, then metabolic fluxes as well as yield coefficients are constant (Deshpande, Yang et al. 2009). Moreover, the consumption of substrates can be separated into a part associated with growth and into one not consumed in association with growth, e.g. for maintenance purposes or for the synthesis of products in a non-growth associated manner (Pirt 1965, Pirt 1982). A metabolic-steady state can in principle be reached in any cultivation including batch and fed-batch processes where extracellular concentrations vary. We illustrate our approach with a simple example of cells in a fed-batch bioreactor, consuming substrates M_i and producing biomass X and products M_j (Figure 2.1).

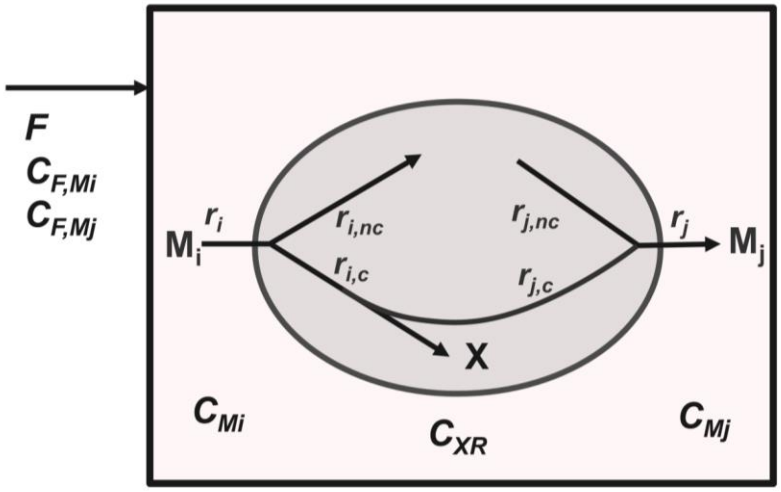


Figure 2.1. Interactions between cells and bioreactor. Schematic representation of major fluxes into cells in a bioreactor during a fed batch process. Substrates, represented by M_i, are consumed with specific rates r_i either associated with growth, with specific rate $r_{i,c}$, or independent of growth with specific rate $r_{i,nc}$. Part of the growth associated consumption of M_i is converted to products M_j with a specific rate $r_{j,c}$. A fraction of M_j is produced not correlated with growth with specific rate $r_{j,nc}$. Substrate M_i and product M_j are also added by feeding them with a volume rate F.

In this context, both substrates and products are metabolites denoted as M. The extracellular substrate M_i can be consumed either in a growth associated manner - and hence the specific consumption rate of M_i is proportional to the specific growth rate μ - or independent of growth. Part of the substrate is directly incorporated into the biomass or consumed for the synthesis of it, some is converted to products or used for maintenance purposes. Similarly, product formation can either be coupled with growth (Luedeking and Piret 1959), characterized by rate $r_{j,c}$, or independent from growth, described by $r_{j,nc}$. A mass balance of substrate M_i in the reactor is described by equation 2.1:

$$\frac{d(V_R.C_{Mi})}{dt} = F.C_{F,Mi} + r_i.C_{XR}V_R \qquad (2.1)$$

The substrate consumption rate, r_i is split into two parts as shown in Figure, a growth associated one, $r_{i,c}$ and one not correlated with growth, $r_{i,nc}$.

$$r_i = r_{i,c} + r_{i,nc} = -Y_{Mi/X}.\mu + r_{i,nc} \qquad (2.2)$$

For product formation we get

$$r_j = r_{j,c} + r_{j,nc} = Y_{Mj/X}.\mu + r_{j,nc} \qquad (2.3)$$

With variables: $C_{F,Mi}$ = concentration of substrate M_i in the feed (mol/L); C_{Mi} = concentration of substrate M_i in the bioreactor (mol/L); C_{XR} = viable cell density (cell/L); V_R = reactor volume (L); $Y_{Mi/X}$ = biomass yield coefficient (mol M_i / cell); μ = specific growth rate (1/day); r_i, r_j = specific rates of formation of M_i and M_j (mol M /

(cell·day); $\boldsymbol{F}$ = feed flow rate (L/day); X indicates biomass associated variables and indices *c* and *nc* indicate processes coupled and not coupled to growth, respectively.

The rates of substrate consumption, r_i, and product formation, r_j, are both calculated from experimental data by rearranging Equation 2.1:

$$\boldsymbol{r_i} = \left(-\boldsymbol{F}.\boldsymbol{C_{F,Mi}} + \frac{\boldsymbol{d(V_R.C_{Mi})}}{\boldsymbol{dt}}\right).\frac{\boldsymbol{1}}{\boldsymbol{C_{XR}.V_R}} \qquad (2.4)$$

The derivative was computed by dividing the change of total quantity of metabolite M_i in a defined period divided by the length of the respective period.

For products, index j is used. In matrix notations, Equation 2.2 and 2.3 read

$$\begin{bmatrix}\boldsymbol{r_i}\\ \boldsymbol{r_j}\end{bmatrix} = \boldsymbol{R} = \boldsymbol{A}.\boldsymbol{\mu} + \boldsymbol{B} = \begin{bmatrix}\boldsymbol{a_i}\\ \boldsymbol{a_j}\end{bmatrix}.\boldsymbol{\mu} + \begin{bmatrix}\boldsymbol{b_i}\\ \boldsymbol{b_j}\end{bmatrix} \qquad (2.5)$$

where $\boldsymbol{R}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors with $\boldsymbol{A}$ and $\boldsymbol{B}$ constant within each metabolic phase.

As a practical consequence, the specific production rate of each metabolite can be expressed as a function of the growth rate by using joined linear sub-models corresponding to distinct metabolic phases. The breakpoints between each sub-model correspond to metabolic shifts.

2.3.2. Data Cleaning and Outlier Identification

As experimental data contain errors that may corrupt the conclusions, outliers must be identified. In particular, at low viable cell density during culture startup, computed specific production rates are inherently noisy and make interpretation of cell metabolism difficult. We thus identified outliers on the first days of production. For each day from day 0 to day 2, data of all experiments was pooled since the conditions were identical until the feed addition started on fay 3 then principal components analysis (PCA) was performed on all the specific production rates of all metabolites for all experiments to reduce the dimensionality of the data (Bersimis,

Panaretos et al. 2005). The multidimensional distance from a sample point to its sample mean was then estimated using the T^2 Hoteling distance (Mason 1997, Bersimis, Panaretos et al. 2005). Values that fall outside an upper control limit (UCL_{T^2}) are defined as outliers assuming the data follows a multidimensional normal distribution. The UCL on the T2 distance is defined as:

$$\boldsymbol{UCL_{T^2} = \frac{(n-1)^2}{n}\beta_{[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}]}} \qquad (2.6)$$

where

n = number of observations

p = number of variables

$\beta_{[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}]}$ = (1–α^{th}) quantile of a Beta ($\frac{p}{2}$;$\frac{n-p-1}{2}$) distribution

This scheme is iterated until no more outliers are identified (Figure 2.2).

On the top of that outlier identification with PCA between day 0 and day 2, two constraints were added for later data points:

1. All the data with viability below 50%, that were usually only observed in later culture phases, were removed from the dataset as low cell viability can also lead to biased and incorrect estimation of the specific production rates of metabolites.
2. Data points with depletion of metabolites during a time period of interest were also removed from the dataset as the computation of the specific production rate of a metabolite would be underestimated in these cases.

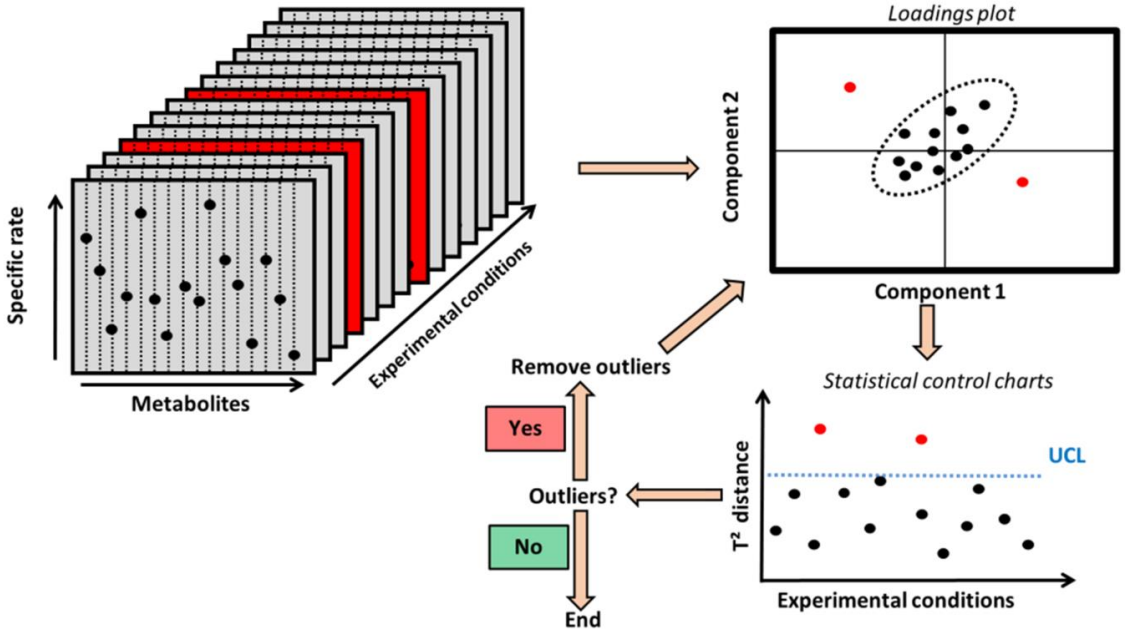


Figure 2.2. Schematic representation of the data cleaning process. A principal component analysis (PCA) is performed, for each day from day 0 to 2 separately, on a pool of the specific production rates of all metabolites . The T^2 Hotelling distance **(Mason 1997, Bersimis, Panaretos et al. 2005)** is then computed by assuming a multi-normal distribution for the data. A simple statistical process control (SPC) **(Mason 1997, Bersimis, Panaretos et al. 2005)** is then used on these T^2 values to identify possible outliers. This scheme is repeated until no more outlier is identified. UCL: upper control limit. This methodology has been used on day 0, 1 and 2 of the cell cultivation since the experimental conditions are similar before the feeding starts on day 3.

2.3.3. Identification of the Number of Metabolic Phases

To avoid over fitting with the segmented regression, the number of phases has to be determined. Based on the metabolic-steady state paradigm, we assume that the vectors A and B (Equation 2.5) are constant within a metabolic phase. We can estimate coefficients A for all metabolites for the whole cell culture process by taking the derivative of metabolic rates with respect to the specific growth rate, μ:

$$\frac{dR}{d\mu} = \boldsymbol{A} \qquad (2.7)$$

Vector A is assumed constant within each metabolic phase. As the derivative can amplify possible biological and analytical errors, the specific production rates were, preliminarily to deriving, smoothed as a function of the specific growth rate with the linear Locally Weighted Scatterplot Smoother (LOWESS) method (Cleveland 1979) by using SAS software JMP 11 ©. The LOWESS method represents non-parametric statistics that do not require any specific model. We used a "tricube" function (Cleveland 1979) as a weight function and for each fitted value, a fraction of the data points of 0.5 was used for the computation. The weight function W is defined as:

$$\boldsymbol{W(x) = \begin{cases} (1 - |x|^3)^3 & \textit{for } |x| < 1 \\ 0 & \textit{otherwise} \end{cases}} \qquad (2.8)$$

The derivatives of the LOWESS function is then also computed with JMP 11 ©. The recursive partitioning (Gaudard, Ramsey et al. 2006) is then used on the smoothed derivatives defined in Equation 2.7: the data is successively partitioned according to a splitting value for a given factor. The splitting value is the one that maximize the – log(p-value), also called logworth, of the chi-square test measuring how different data is between the two partitions. The purpose of partitioning is to split all derivatives $dR/d\mu$ of each metabolite (Equation 2.7), as a function of the specific growth rate and then to determine the number of breakpoints. On the dataset of all breakpoints for all metabolites, we apply a hierarchical clustering to identify similar set of breakpoint values (Mojena 1977, Murtagh 1983, Szekely and Rizzo 2005) (Figure 2.3). Each observation starts in its own cluster, and at each step the clustering process calculates the distance between each cluster, and combines the two clusters that are closest together (agglomerative procedure). The agglomerative procedure is Ward's method (Murtagh 1983). The linkage distance is defined as the "cost" in between-class sum of square to join the clusters. The final number of clusters selected is chosen as

the first "knee" point in the linkage function, i.e. the peak in the second-order derivative of the linkage distance function. The outcome is a first estimation of the breakpoint value of each metabolic phase and the total number of metabolic phases.

2.3.4. Segmented Linear Regression

Linear segments between all specific metabolic rates with the specific growth rate, μ, were identified using segmented linear regression analysis (McGee and Carleton 1970, Toms and Lesperance 2003, Ryan, Porth et al. 2007). It is a regression model composed of a sequence of joined linear sub-models. If we define r_k as the observed specific production rate of metabolite M_k and μ as the growth rate, we have, for $\boldsymbol{n}\geq 2$ metabolic phases and $\boldsymbol{n-1}$ breakpoints $\boldsymbol{B_{P_s}}$ such that $s\in\{1,\ldots,n-1\}$:

$$\boldsymbol{r_k(\mu) = a_{k,1} * \mu + b_{k,1} + \sum_{s=1}^{n-1} a_{k,s+1} * \left(\mu - B_{P_j}\right) * u_s}$$

$$\boldsymbol{u_s = \begin{cases} 1 & for\ \mu \leq B_{P_s} \\ 0 & for\ \mu > B_{P_s} \end{cases} \quad 1 \leq s \leq n-1} \qquad (2.9)$$

$\boldsymbol{b_{i,1}}$, $\boldsymbol{a_{i,1}}$, $\boldsymbol{a_{i,j+1}}$ and $\boldsymbol{B_{P_j}}$ constant coefficients of metabolite M_i, for $j\in\{1,\ldots,n-1\}$.

This expression allows the regression function to be continuous at the breakpoint (Ryan, Porth et al. 2007). Amino acid limitation in the medium can lead to its depletion that would impact mAb and protein synthesis (Kilberg, Shan et al. 2009, Gramer 2014). As an extra constraint imposed to our model, as soon as an essential amino acid is depleted, the specific mAb productivity predicted by the model is set to zero. We consider tryptophan, histidine, isoleucine, methionine, threonine, phenylalanine, valine, tyrosine, leucine, lysine, glutamine, arginine and cysteine as essential amino acids (Hu and Zhou 2012).

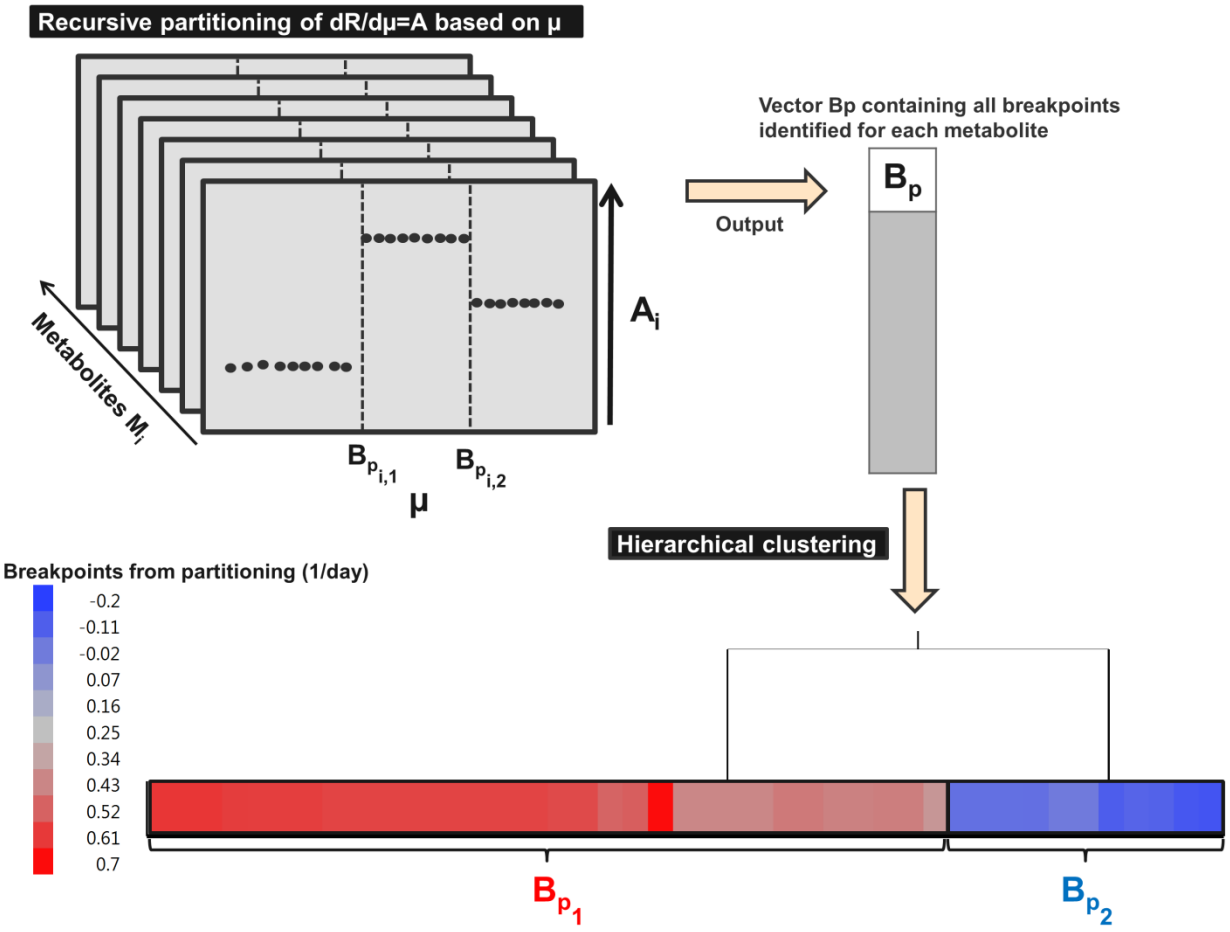


Figure 2.3. Identification of the number of metabolic phase breakpoints. Hierarchical clustering was performed on the vector containing the breakpoint growth rate values identified from recursive partitioning. Each observation/breakpoint starts in its own cluster, and at each step the clustering process calculates the distance between each other cluster, and combines the two clusters that are closest together (Agglomerative procedure) (Murtagh 1983). The agglomerative procedure uses the Ward's method to calculate the distance between each cluster. The objective was to identify the number of distinct metabolic phase breakpoints required to calibrate the segmented regression model (Figure 2.4). Two groups of breakpoints were identified, which correspond to two metabolic phase breakpoints B_{P1} and B_{P2}. B_{P1}= 0.54 ± 0.06 day^{-1}; B_{P2}=-0.08 ± 0.06 day^{-1}.

2.3.5. Parameter Estimation

For each metabolite, based on the number *n* of metabolic phases determined by the hierarchical clustering, *n* models were set up from zero to *n-1* breakpoints. For each model, the estimation of model parameters with the best fit was selected using a least-

square minimization (Wagner, Soumerai et al. 2002). To assess if the addition of a breakpoint makes the model prediction statistically superior to a model with a lower number of breakpoint, an F-test was performed at 95% confidence level. To alleviate the selection of too close breakpoints, a criterion has been added to the model:

$$|B_{P1} - BP_2| > 0.2\ day^{-1} \quad (2.10)$$

A summary of the methodology is presented in Figure 2.4. All estimations were carried out using EXCEL (Microsoft) for primary data treatment and Matlab Release 2013a (The Mathworks, Natick, MA, USA) for further calculations unless otherwise stated. MATLAB scripts are supplied as Supplementary Material to automatically carry out segmented linear regression using a supplied data set.

2.4. Material and Methods

2.4.1. Cell Line, Cell Cultivation, Sampling and Rate Estimations

A CHO-DG44 cell line was used. The cells were cultivated in a proprietary chemically defined serum free medium in 2 L stirred tank glass bioreactor (STR) with supply towers (C-DCUII, Sartorius Stedim Biotech) controlled by a multi-fermentation control system (MFCS, Sartorius Stedim Biotech). The reactors were equipped with a 3-segment blade impeller (elephant ear impeller).

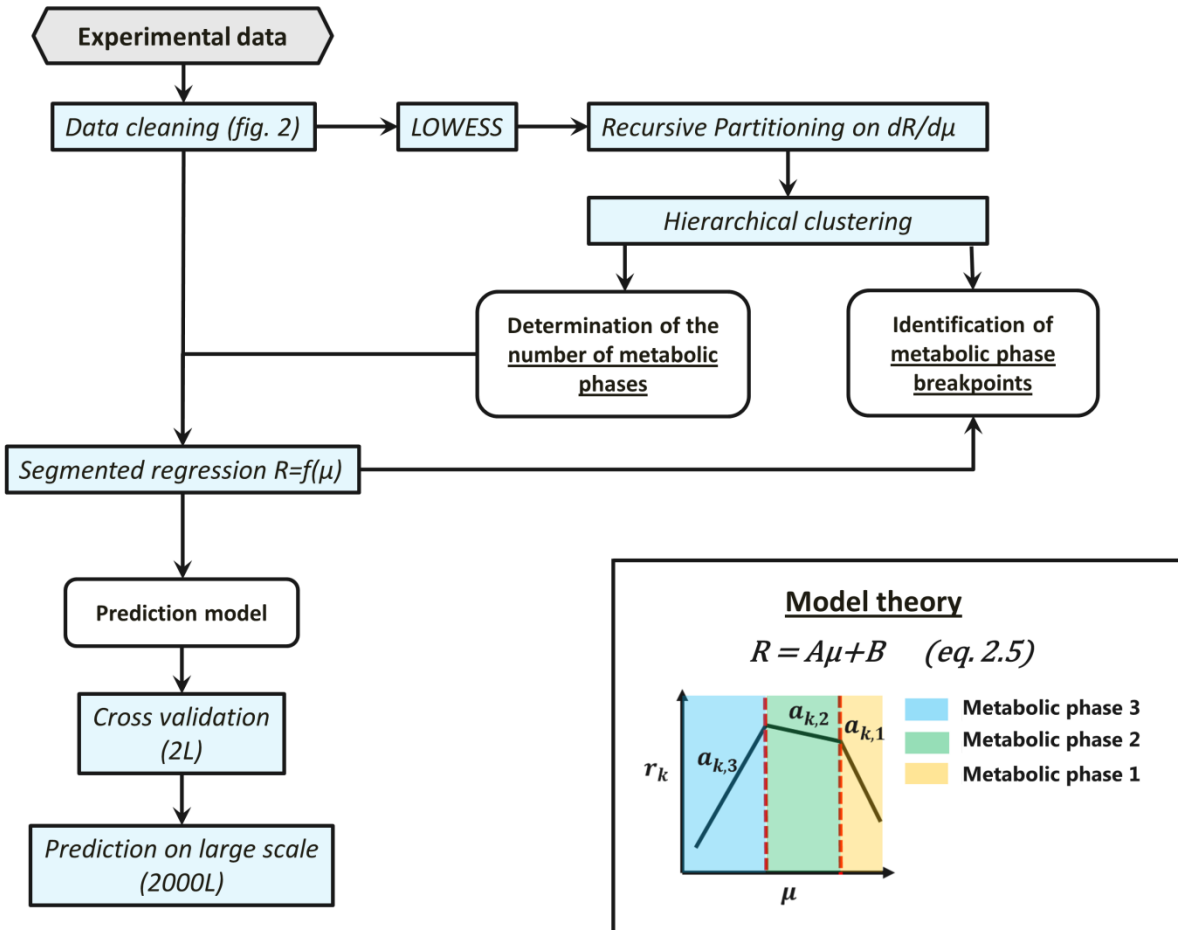


Figure 2.4. Developed methodology to identify and characterize metabolic phases. Experimental data are first cleaned using the methodology presented in Figure 2.2 and additionally by removing data with a viability below 50% or a depletion of metabolites during a measurement interval. The number of metabolic phases during the cell culture process are determined by differentiating the smoothed (LOWESS) reaction rates of all metabolites with respect to the growth rate ($dR/d\mu$). Recursive partitioning is then applied on those derivatives to get a vector of possible metabolic phase breakpoints. Hierarchical clustering is then applied on this vector of possible breakpoints to define the number of final metabolic phases (clusters). Knowing the number of metabolic phases, the segmented regression can then be calibrated on the calibration dataset for each metabolite and validated on the cross validation dataset of the 2 L bioreactor and also of the 2000 L bioreactor.

The cultivation start volume was adapted to ensure an optimal cultivation end volume. The production bioreactors were seeded at similar target seeding density (TSD). The pH was controlled at a value of 7 with a dead band of 0.2. Dissolved oxygen

concentration (pO2) was set to 40% air saturation. To control pO2, air, nitrogen and oxygen were sparged into the culture using a cascade controller with a predefined mixture profile. The temperature was controlled at 36.8 °C.

The culture was operated in fed-batch mode for 14 days. During the feeding phase, the monoclonal antibody (mAb) is secreted into the medium. Samples were drawn daily to determine total and viable cell number, viability, off-line pH, partial pressure of CO_2, pCO_2, osmolality, glucose-lactate, amino acid and mAb concentrations (stored at -80 °C). Antifoam was added manually on demand every day to control the build-up of foam. 72 hours after inoculation, continuous nutrient feeding , i.e. constant feed rate specific for each day, was started with a predetermined rate using a proprietary chemically defined concentrated feed. In addition to that proprietary chemically defined concentrated feed addition,. a glucose solution of 500g/L was added as a bolus to the culture when the glucose concentration dropped below 6 g/L but only from day 6 onwards so that in the experimental conditions tested, glucose was never depleted at any time during the culture.. Samples for the amino acid analysis were taken before the feed addition. The extracellular concentrations after feeding were computed based on the feed composition information. Specific growth rate, μ, was computed for each experimental condition separately as the slope of the linear trend line obtained by plotting $\ln(C_{XR}.V_R)$ against time (Clarke, Doolan et al. 2011, Chin, Chin et al. 2015).

2.4.2. Experimental Conditions

Various feed compositions of amino acid were tested in small scale bioreactors (2 L) for a total of 29 experimental conditions. We varied the concentration of three different amino acids (*aa1*, *aa2*, *aa3*) contained in our feed as described in the Supplementary Table S2.1.

The volume of feed added per bioreactor volume in the STR was the same in each condition. pH, temperature, stirrer speed and all the bioreactor parameters were controlled at the same value. The TSD was also the same for all experiments. Three 2000 L production runs with similar experimental conditions, i.e. with identical feed

addition profiles and chemically defined feed composition (Table S2.1), were also performed. The temperature set point, pH set point, pO2 set point, target seeding density, medium formulation, nutrient feed formulation, feeding strategy, and culture duration were the same as the 2 L bioreactor scale.

2.4.3. Analytical Methods

Cells were counted by using a VI-CELL® XR (Beckman-Coulter, Inc., Brea, CA) automated cell counting device that applied the trypan blue exclusion method. Glucose and lactate levels in the culture medium were determined using a NOVA 400 BioProfile automated analyzer (Nova Biomedical, Waltham, MA). A model 2020 freezing-point osmometer (Advanced Instruments, Inc., Norwood, MA) was used for osmolality determination. Offline gas and pH measurements were performed with a BioProfile pHOx® blood gas analyzer (Nova Biomedical Corporation, Waltham, MA). Product titer analysis was performed with a ForteBio Octet model analyzer (ForteBio, Inc., Menlo Park, CA) or protein A high performance liquid chromatography (HPLC) with cell culture supernatant samples which were stored at -80°C prior to analysis. Amino acids were analyzed by reversed-phase UPLC (Waters AccQ·Tagultra method) after ultra-filtration using Amicon Ultra-0.5 mL centrifugal filters (Merck Millipore, Billerica, MA). Statistical analysis were performed using SAS software JMP 11 ©. Matlab Release 2013a (The Mathworks, Natick, MA, USA) was used to calibrate the segmented linear model.

2.5. Results and Discussion

The growth profile of all experimental conditions is depicted in Figure 2.5a. The growth behavior varies with the experimental conditions. For further analysis, we computed the specific production rates, r_i, of glucose, lactate, ammonia, all amino acids and the mAb.

2.5.1. Data Cleaning and Determination of the Number of Metabolic Phases

We applied the outlier identification methodology based on the PCA on our dataset for days 0, 1 and 2 separately. From a total of 404 data points, 86 data points are from days 0, 1 and 2. From these 47 outliers were identified by using PCA and removed from the dataset. Based on the two extra constraints added to our data cleaning procedure from day 3 to day 14, data points with depletion of metabolites and/or with viability lower than 50% throughout the cell culture production were also removed from the dataset resulting in a total of 215 remaining data points.

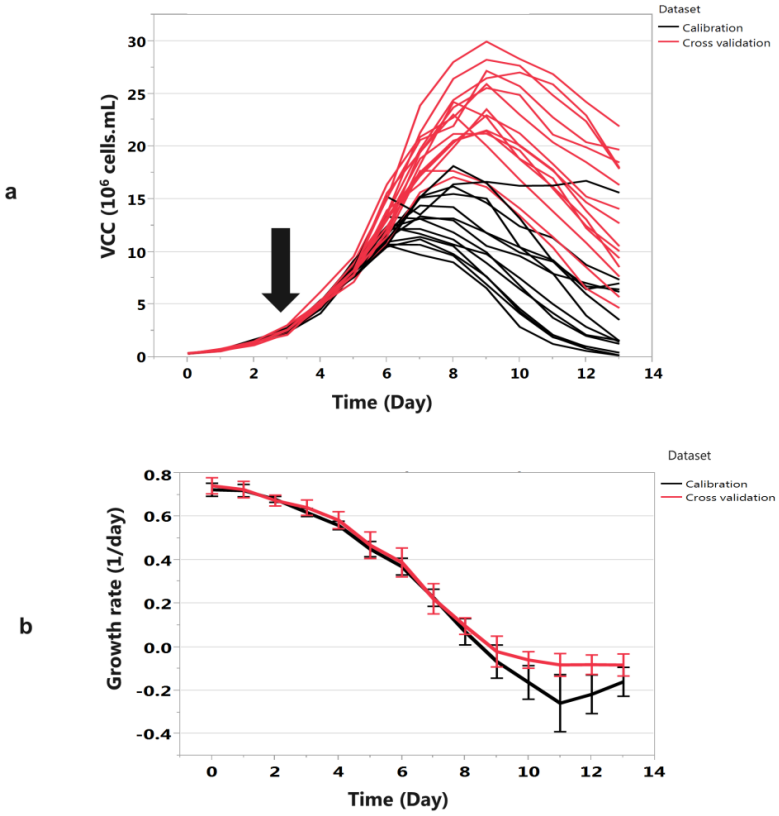


Figure 2.5. Experimental viable cell count and time course of specific growth rates. (a) Growth profiles of CHO-DG44 for 29 experimental conditions (see supplementary Table S2.1) with various cell growth behaviors. The cells were cultivated in a 2 L bioreactor operated in fed-batch mode for 14 days. Black arrow: Start of nutrient feeding with a predetermined rate. (b) Specific growth rate of the 29 experimental conditions after data cleaning (Figure 2.2) for the calibration dataset and the cross validation dataset.

This cleaned dataset was partitioned into two datasets based on the growth rate profiles of each experimental condition: 115 data points (calibration dataset) with a wide range of experimental conditions as specified in the Supplementary Table S2.1 and 100 data points (cross validation dataset) with similar experimental conditions. The cross validation dataset contains experimental conditions within the design space of the calibration dataset. From the calibration dataset, two distinct clusters, which correspond to the breakpoints, were identified based on recursive partitioning and hierarchical clustering as described above (Figure 2.3): one (B_{p1}) at a growth rate of 0.54 ± 0.06 day^{-1} and a second one (B_{p2}) at -0.08 ± 0.06 day^{-1}. The first breakpoint was identified for all metabolites. The second breakpoint was only identified for proline, valine, leucine, methionine, tyrosine, threonine, cysteine, asparagine, lysine, glutamate, lactate and mAb. Hence, to avoid over fitting during the calibration of the segmented linear regression, the maximum number of breakpoints to identify was set to two.

2.5.2. Calibration of the Prediction Model Using the Segmented Regression Model

Segmented regression was applied on the cleaned calibration dataset, containing specific growth rates and not smoothed specific production rates of metabolites and mAb (Figure 2.6a). The identification of the metabolic phase breakpoints and the calibration of models were performed separately for each metabolite. Estimated model parameters are listed in Table 2.1. Twelve metabolites, i.e ammonium, glycine, alanine, methionine, serine, asparagine, glutamine, arginine, aspartate, glutamate, glucose and lactate, are impacted by metabolic phases. These metabolites are linked to glucose/glutamine metabolism and cell proliferation which confirms results of Rehberg et al.(Rehberg, Rath et al. 2014). Most models for these metabolites include only one breakpoint, i.e. only two metabolic phases were identified. Only glutamate, methionine and lactate have a specific production rate profile divided into three metabolic phases. Based on an F-test, twelve metabolites are better fitted with a

simple linear regression model and not impacted by metabolic phases: proline, isoleucine, leucine, lysine, valine, phenylalanine, cysteine, tyrosine, tryptophan, threonine, histidine and mAb (Figure 2.6a).

2.5.3. Suitability of the Segmented Model to Identify Metabolic Phases

Breakpoints were identified for 12 metabolites. For the 12 metabolites that have significant breakpoints, the breakpoints B_{p1} and/or B_{p2} identified share similar values with a relative precision of breakpoint identification close to 5% for both breakpoints which supports the suitability of the method. The relative precision is defined here as the standard deviation divided by the range of the possible growth rate values.

The first breakpoint between phases P1 and P2 was reached when the initially high specific growth rate, μ, decreased below a value of 0.58 ± 0.09 day^{-1}, the second one between phases P2 and P3 when μ fell below -0.18 ± 0.09 day^{-1} (Figure b). The segmented linear regression identified breakpoint B_{p1} for only 11 metabolites and breakpoint B_{p2} for 4 metabolites (Table 2.1) which is less than those identified with the combination of recursive partitioning and hierarchical clustering. This may be explained by the sensitivity of the second method to outliers: the LOWESS regression may create additional trends that only exist due to possible outliers and the derivative computation can amplify the noise. Moreover, no F-test is performed with the combination of recursive partitioning and hierarchical clustering which could possibly identify non-significant metabolites. Nevertheless, breakpoints identified by both methods are similar which proves the reliability of the segmented linear regression to identify metabolic phases. As a conclusion from our work, the application of segmented regression is sufficient to identify relevant breakpoints and parameter. Therefore, hierarchical clustering is not necessary which simplifies the overall procedure.

Table 2.1. Segmented model coefficients. For each metabolites and for each metabolic phase, the value of coefficient *a* and *b* from Equation 2.5 is presented. The coefficients were also identified with the cross validation dataset and presented in the brackets. Red names correspond to metabolites that are impacted by the three metabolic phases. Bold names are metabolites that are impacted by two metabolic phases. Glc – glucose; Lac – lactate, mAb – monoclonal antibody

	a (10^{-09} mmol/cell)			*b* (10^{-09} mmol/[cell.day])		
	P1	P2	P3	P1	P2	P3
NH4+	**12.45** (9.90)	**-0.16** (-0.18)		**-7.81** (-6.10)	**0.04** (0.04)	
Gly	**2.04** (2.12)	**-0.002** (0.08)		**-1.07**(-1.15)	**0.02** (0.02)	
Ala	**5.15** (4.78)	**-0.11** (0.02)		**-3.03** (-2.84)	**0.02** (-0.02)	
Pro		**-0.05** (-0.10)			**-0.06** (-0.04)	
Val		**-0.06** (-0.13)			**-0.09** (-0.06)	
Leu		**-0.09** (-0.14)			**-0.12** (-0.08)	
Ile		**-0.07** (-0.03)			**-0.06** (-0.04)	
Met	**-0.46** (-0.57)	**-0.03** (-0.006)	**-0.24** (N/A)	**0.28** (0.36)	**0.01** (0.0044)	**-0.04 (N/A)**
Phe		**-0.04** (-0.05)			**-0.03** (-0.02)	
Tyr		**-0.00009** (-0.05)			**-0.06** (-0.03)	
Trp		**-0.009** (-0.012)			**-0.02** (-0.01)	
Ser	**-1.47** (-1.87)	**-0.08** (-0.17)		**0.48** (0.14)	**-0.18** (-0.14)	
Thr		**-0.04** (-0.09)			**-0.07** (-0.05)	
Cys		**0.01** (-0.06)			**-0.07** (-0.05)	
Asn	**-0.44** (-0.49)		**0.78** (N/A)	**-0.25** (-0.17)		**-0.11** (N/A)
Gln	**-12.39** (-12.17)	**-0.03** (-0.06)		**7.42** (7.38)	**0.007** (0.007)	
Lys		**-0.10** (-0.20)			**-0.10** (-0.06)	
His		**-0.03** (-0.05)			**-0.01** (-0.009)	
Arg	**-3.14** (-0.72)	**-0.08** (-0.09)		**2.19** (0.38)	**-0.03** (-0.03)	
Asp	**1.97** (1.49)	**-0.02** (-0.02)		**-1.25** (-0.95)	**-0.05** (-0.05)	
Glu	**1.84** (1.31)	**0.07** (0.05)	**1.20** (N/A)	**-1.19** (-0.85)	**-0.09** (-0.09)	**0.18** (N/A)
Glc	**-18.34** (-18.58)	**-0.04** (-0.21)		**9.61** (10.00)	**-0.98** (-0.88)	
Lac	**63.41**(56.79)	**-0.05** (0.15)	**-17.66** (N/A)	**-39.42** (-35.51)	**0.03** (-0.06)	**-4.86** (N/A)
mAb		**-0.0006** (-0.0002)			**0.0007** (0.006)	

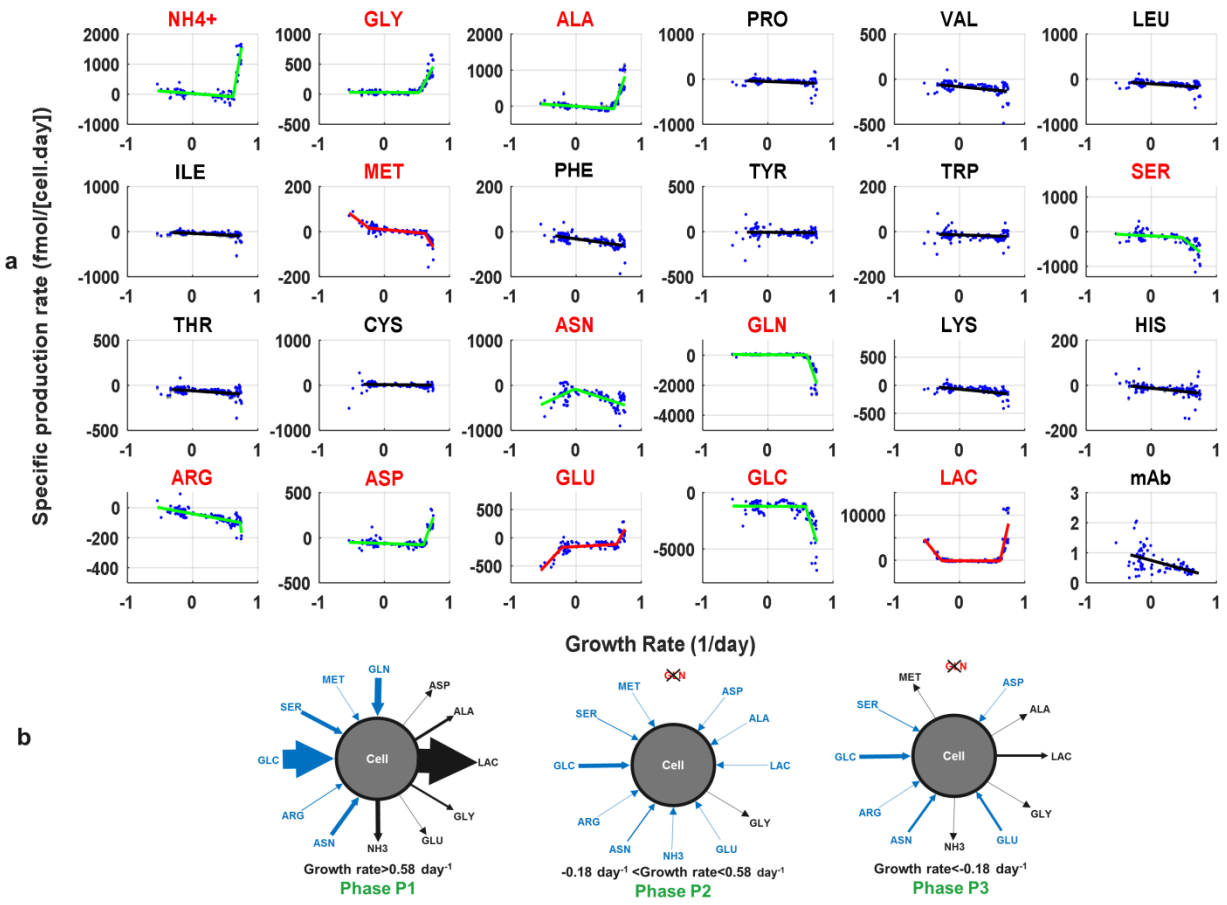


Figure 2.6. Segmented regression of specific rates as a function of the growth rate. To identify metabolic phases, segmented regression was used (Figure 2.4). Data of the 2 L bioreactor calibration dataset were used. Three models were set up for each metabolite, from zero up to two breakpoints. To assess whether the addition of a breakpoint makes the model prediction statistically superior to a model with a lower number of breakpoints, an F-test was performed with 95% confidence level. (a**)** Segmented regression models are presented. Red names correspond to metabolites that are impacted by metabolic phases, i.e. which show better prediction with one to two breakpoints. When the segmented regression model is characterized by a red line, three metabolic phases were identified. Identification of two metabolic phases are characterized by a green line. (b) The twelve metabolites that were significantly impacted by metabolic phases are presented for each metabolic phase: P1, P2 and P3. Blue arrow: net uptake; Dark arrow: net secretion. The widths of the arrows are proportional to the average specific production rate values for the defined metabolic phase.

The first metabolic phase P1, defined by a high growth rate, is characterized by a high production of ammonium, glycine, alanine, lactate, glutamate and aspartate and a high consumption of methionine, asparagine, arginine, serine, glucose and glutamine. This is a common metabolic profile observed in literature with suspension CHO cells also called *overflow metabolism* (Niu, Amribt et al. 2013, Wahrheit 2015) or *exponential phase* (Dorka, Fischer et al. 2009, Amribt, Niu et al. 2013, Meshram, Naderi et al. 2013). All the other amino acids are also consumed during this first metabolic phase but the rates are lower than those of the twelve metabolites impacted by metabolic phases. During phase P1, lactate is a byproduct of the glycolysis and high excretion of alanine is due to the conversion of pyruvate to alanine via the alanine aminotransferase (ALAT) serving as a nitrogen sink. Glycine synthesis is a result of high serine uptake which is generally observed for mammalian cells and can be linked to nucleotide synthesis and to cell proliferation (Narkewicz, Sauls et al. 1996). Glutamine is taken up as a carbon source for the tricarboxylic acid cycle (TCA) cycle and hydrolyzed into glutamate and ammonium. Asparagine is partly converted into aspartate which can then be converted into oxaloacetate. In the second metabolic phase P2 (Figure 2.6b), the growth rate further decreased and some metabolites, i.e. ammonium, alanine, lactate, glutamate and aspartate, started being consumed. This metabolic phase is usually called *balanced metabolism* (Wahrheit, Niklas et al. 2014, Wahrheit 2015) or *transition phase* (Provost, Bastin et al. 2006, Naderi, Meshram et al. 2011, Zamorano, Vande Wouwer et al. 2013). Overall, the rates of all metabolites were lower than in phase P1. Alanine and lactate, accumulated during the phase P1, were converted back to pyruvate, which is a major characteristic of CHO cell metabolism. The last metabolic phase P3 (Figure 2.6b) is characterized by an accumulation of methionine, an increase of the consumption of glutamate and asparagine, and an overproduction of lactate. For the other eight metabolites that were impacted by the first metabolic shift, no break of slope could be observed between phases P2 and P3. The growth rate is negative for that metabolic phase, also

called in the literature the *maintenance phase* (Yu, Hu et al. 2011, Wahrheit 2015) or *death phase* (Provost, Bastin et al. 2006, Zamorano, Vande Wouwer et al. 2013).

Usually, different growth behaviors can be observed during process production, making it quite difficult to compare their metabolic characteristics by only using time. Using the growth rate rather than time allows better identification of metabolic phases and better comparison of their characteristics between various experimental conditions particularly in fed-batch cultivation.

2.5.4. Validation of the Model at Small Scale (2 L)

Metabolic profiles of the cross validation dataset were predicted (Figure 2.7) from the experimental growth rate of the cross validation dataset and parameters of segmented regression model identified using the calibration dataset (Table 2.1). Since the growth rates were overall similar in the cross validation dataset, the data can be presented as a function of time. The model prediction closely followed the experimental trends throughout the cell culture process. The only discrepancy occurred in the late stages, on day 13, for 18 metabolites. For 4 metabolites, i.e. tyrosine, tryptophan, cysteine and glucose, the prediction is also out of range for day 12.

2.5.5. Prediction of the specific production rate in large scale (2000 L)

To estimate the transferability of the segmented regression model to a larger scale cultivation, specific production rates of metabolites of each 2000 L experiment were predicted (Figure 2.8) by the model estimated on the 2 L bioreactor calibration set (Table 2.1). As the growth rates were similar for the triplicate experimental conditions, the data could be presented together as a function of time. The prediction model is able to track the experimental trends of almost all metabolites over the entire culture period. Overall 88% of the predictions fall into a 3 standard deviations interval. Two amino acids, named *aa1* and *aa2*, were depleted throughout the cell culture process but the model was able to predict their concentrations (Figure 2.9b).

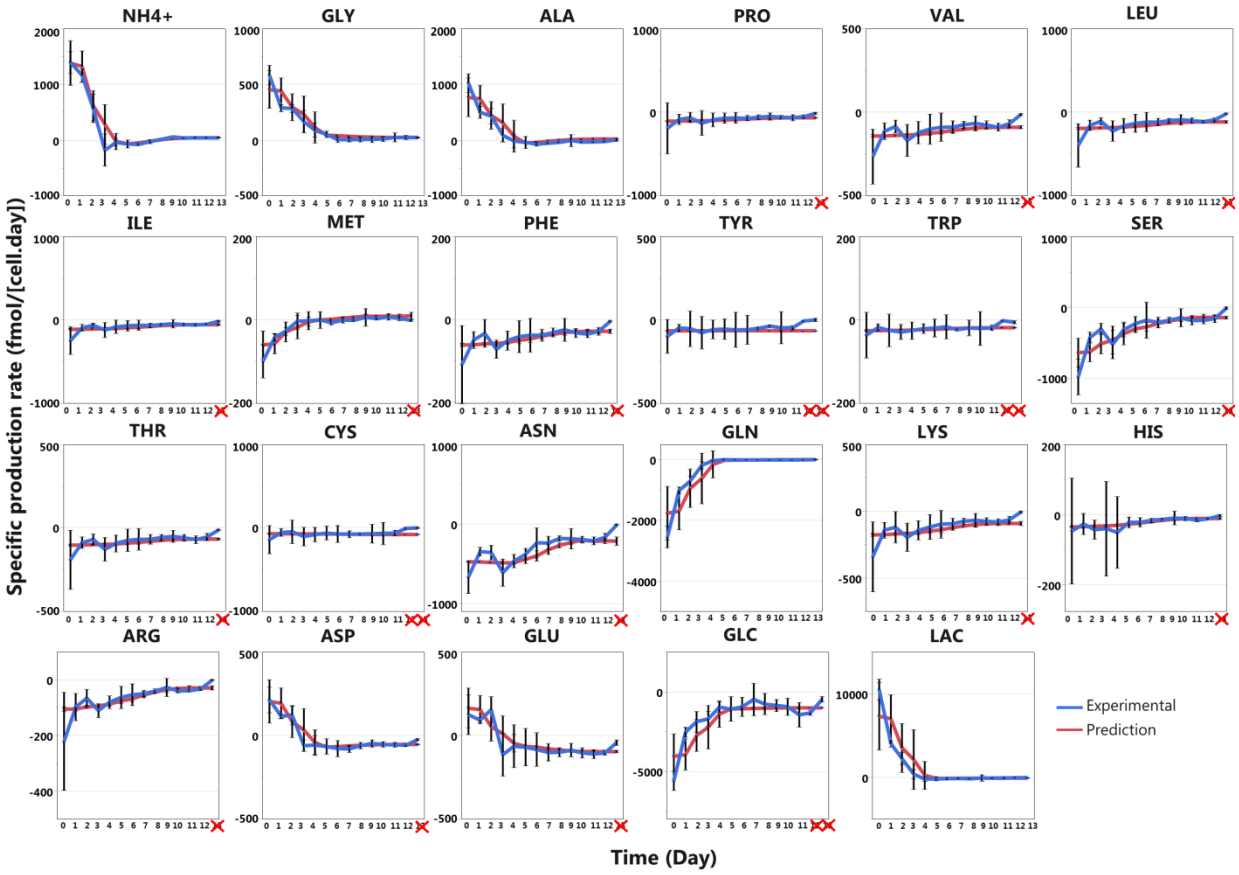


Figure 2.7. Cross validation of the segmented regression models for each metabolite in 2 L bioreactor runs. The specific production rates of all metabolites for the validation dataset are presented as a function of time (day) as the specific growth rate profiles were similar (Figure 2.5). The error bars correspond to 3 standard deviations. The prediction (red line) is based on the segmented regression of the calibration dataset (Figure 2.6) and used the experimental specific growth rate profile (Figure 2.5b) to estimate specific production rates. The prediction variability is due to the growth rate value variability and the error bars presented for the prediction correspond also to the 3 standard deviation. When the prediction was out of the 3 standard deviations range, the corresponding day is crossed in red.

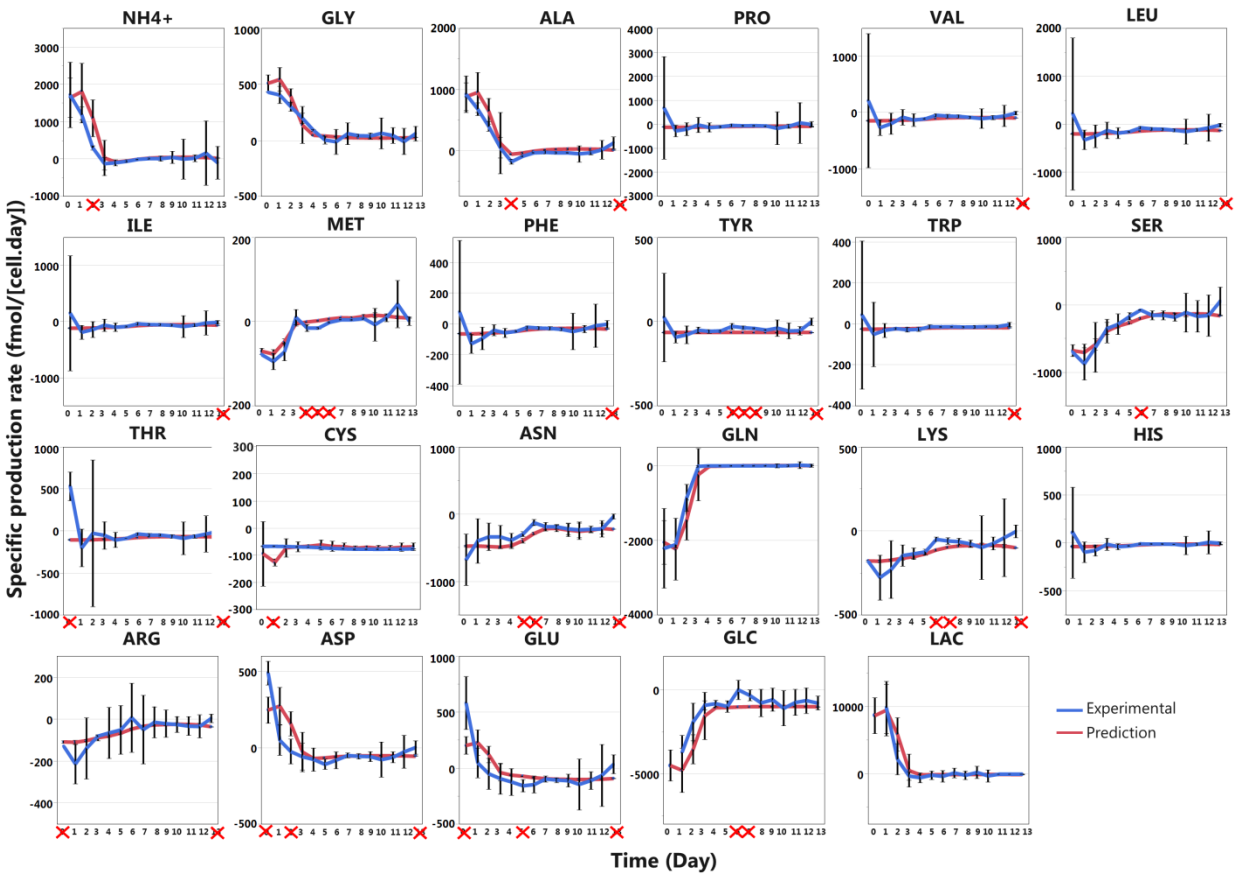


Figure 2.8. Validation of the segmented regression models for each metabolite at 2000 L bioreactor scale. The specific production rates of each metabolites for three 2000 L bioreactors runs with the same experimental condition are presented as a function of time (day) as the specific growth rate profiles are similar. The error bars correspond to 3 standard deviations. The prediction model (red line) is based on the segmented regression of the calibration dataset (Figure 6) and is based on the experimental specific growth rate profile (Figure 5b). The prediction variability is due to the growth rate value variability and the error bars presented for the prediction correspond also to the 3 standard deviation. When the prediction was out of the 3 standard deviations range, the corresponding day is crossed in red.

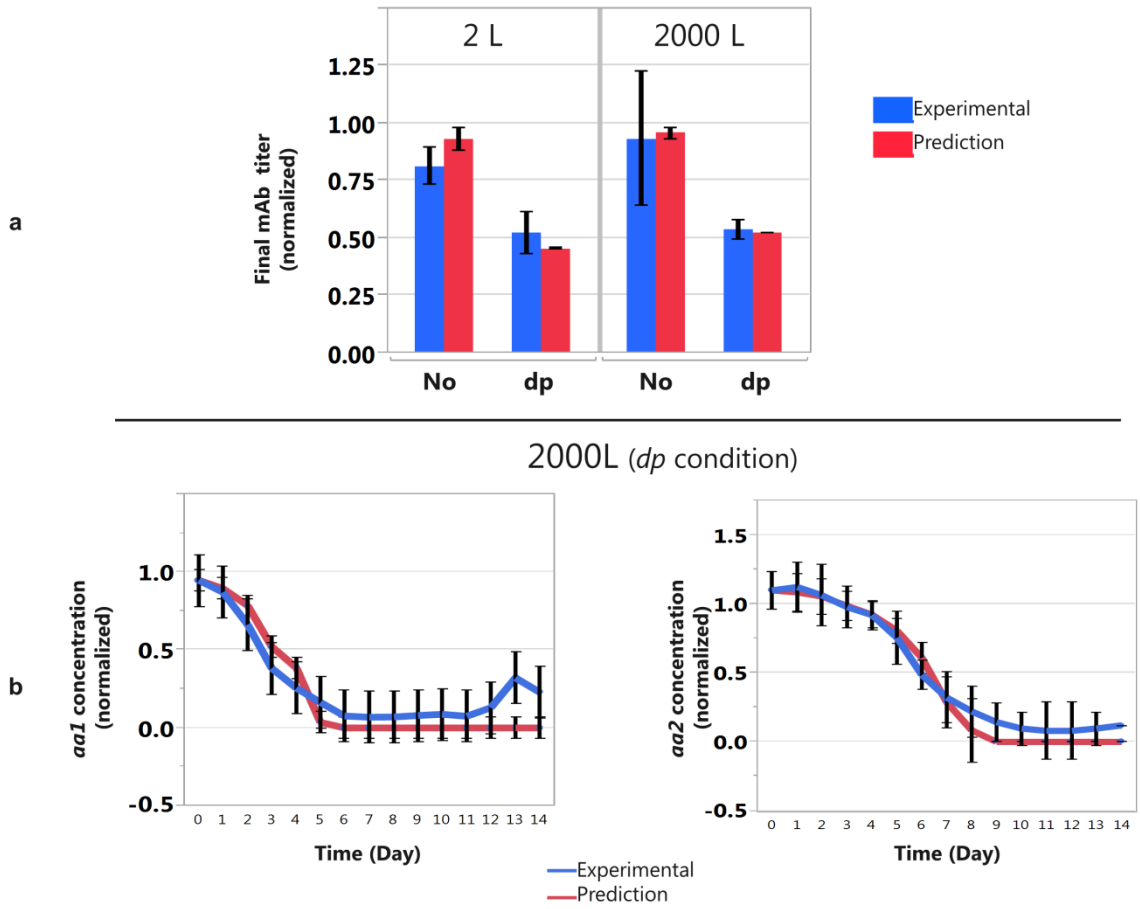


Figure 2.9. Prediction of the final mAb titer and of aa1 and aa2 concentrations. (a) Experimental and predicted final mAb titers for the calibration dataset (2 L) and for large scale bioreactor runs (2000 L). The prediction is based on the segmented regression model of the specific productivity. The calibration dataset (2 L) and the large scale bioreactor runs (2000 L) are divided into 2 subsets: one without depletion of any metabolite during the cell culture process (No), and the other with depletions of *aa1* and *aa2*, two amino acids, during the cell culture process (dp). The error bars correspond to 3 standard deviations for both prediction and experimental data. (b) Prediction of *aa1* and *aa2* concentrations before feeding during the whole cell culture process in triplicate experiments in 2000 L bioreactors (dp). The concentrations were normalized to the initial concentrations.

2.5.6. Accuracy of the Segmented Model for Prediction of Metabolite Profiles (2 L and 2000 L)

Cross validation of the segmented models showed good results for all metabolites. The ability to predict the experimental metabolite profiles in large scale experiments reinforces the validity of the model and justifies the initial assumption of linear correlation of specific rates of metabolites with the specific growth rate. The accuracy of our model with two to a maximum of six parameters to estimate for each metabolite is remarkable. For twelve metabolites, even if a simple linear model was used, the model has provided accurate results. Moreover, adding more parameters to the model may lead to over fitting. From this point of view, the model can be used for accurate prediction of the specific production rates of metabolites and so, of the metabolite concentrations, requiring only the experimental specific growth rate of the prevailing experiment. The established model would also allow online estimation of metabolic rates on the bases of online measured biomass parameters, e.g. viable cell count. It could therefore potentially also be applied for on line optimization of feeding profiles.

2.5.7. Accuracy of the segmented model for prediction of final mAb titers (2 L and 2000 L)

The model was used to predict the final mAb titer. Results of prediction model are presented in Figure 2.9a. The final mAb titer prediction in 2 L is accurate and within the range of ±3 standard deviations even for conditions with a depletion of metabolites throughout the cell culture process. Similarly to the small scale, the 2000 L triplicate experimental conditions previously used to compare the metabolite prediction profiles were depleted in some metabolites, i.e. *aa1* and *aa2*, during the cell cultivation. Hence, the prediction of the final mAb titer in 2000 L is also compared to other duplicates experimental conditions in 2000 L bioreactors runs without metabolites limitations throughout the cell culture process. The final mAb titer prediction is accurate and within the ±3 standards deviations even with metabolite

limitations throughout the cell culture process (Figure 2.9a). Our model is able to predict mAb titer decrease due to essential amino acid depletion.

2.5.8. Prediction outside calibration experimental conditions.

The segmented regression methodology was used to identified coefficients a and b from equation 2.10 with the cross validation dataset. The objective was to compare both coefficients identified with the cross validation dataset and with the calibration dataset. Results are presented in Table 2.1 in the brackets and in supplementary material (Figure S2.1). As expected, the breakpoints B_{p2} was not identified with the cross validation dataset as the growth rates minimum values were higher than B_{p2}. For coefficients a and b within metabolic phase P1 and P2, coefficients are similar which prove the predictability and applicability of the methodology. As the cross validation dataset only contained 4 different experimental conditions, we can conclude from this work that the presented method does not require a wide range of different experimental conditions in order to set up a robust and predictive model. The question of the minimum of data points and experiments needed is difficult to answer, since it depends on the quality of the measured data.

2.6. Conclusion

In summary, we propose an accurate predictive model of external metabolite rates which requires few parameters to estimate and seems very robust. The final titer can also be predicted even if the cells are starved in some metabolites. It should also be highlighted that an entire and complex metabolic network is not needed in order to achieve the macroscopic modeling which makes it simpler and, possibly, easily adaptable to other cell clones and cell lines. We also presented a systematic methodology to identify metabolic phases that allows comparing various experimental conditions with different growth behavior. It provides an excellent basis for later metabolic flux analysis and for later dynamic mechanistic modeling. We

showed that the metabolites that are more impacted by metabolic shift are those linked to glucose/glutamine metabolism and cell proliferation.

Chapter 3

3 Predictive Macroscopic Modeling of Cell Growth, Metabolism and Monoclonal Antibody Production: Case Study of a CHO Fed-batch Production

3.1. Abstract

We describe a systematic approach to establish predictive models of CHO cell growth during biopharmaceutical production. Cell growth, cell metabolism and monoclonal antibody (mAb) production are predicted by combining an empirical metabolic model with mixed Monod-inhibition type kinetic that we generalized to every possible external metabolite. We describe the maximum specific growth rate as a function of the integral viable cell density (IVCD). Moreover, we also take into account the accumulation of intracellular metabolite pools that can influence cell growth. This is illustrated with fed-batch cultures of Chinese Hamster Ovary (CHO) cells producing a mAb. The impact of two selected essential metabolites on cell growth and cell productivity was assessed and the macroscopic model was successfully used to predict the impact of new untested feeding strategies on cell growth and mAb production. The resulting model combining piecewise linear relationships between metabolic rates and the growth rate and Monod-inhibition type models for cell growth did well predict cell culture performance in fed-batch cultures even outside the range of experimental data used for establishing the model.

This chapter is in preparation for submission

Ben Yahia, B., Malphettes, L., Heinzle, E. Predictive Macroscopic Modeling of Cell Growth, Metabolism and Monoclonal Antibody Production: Case Study of a CHO Fed-batch Production.

3.2. Introduction

Determination of the optimal nutrients conditions throughout the bioreactor production step is essential to reach high productivity and product quality. Currently it relies on numerous time-consuming and costly experiments. Predictive models could help reduce the need for expensive experiments and accelerate bioprocess development. Various attempts have been made in the past to characterize these processes by mathematical models (Ben Yahia et al., 2015; Klein et al., 2015; Niu et al., 2013; Nolan and Lee, 2012). However, throughout the bioreactor production, the cells adapt to the changing extracellular conditions resulting in the development of different metabolic phases and related metabolic shifts (Ben Yahia et al., 2016; Ben Yahia et al., 2015; Dean and Reddy, 2013; Farzan et al., 2016; Nicolae et al., 2014; Niklas et al., 2013; Wahrheit, 2015; Wahrheit et al., 2014a; Wahrheit et al., 2014b). The metabolic and regulatory processes underlying such metabolic shifts remain largely unknown. Integrating principles developed in the metabolic engineering field might improve such modeling. We previously developed a systematic approach to identify metabolic phases and obtain an accurate stoichiometric model of cells metabolism during CHO fed-batch culture (Ben Yahia et al., 2016). We also demonstrated that the model developed at small scale remains correct upon scale up. This purely stoichiometric model does not require any a priori metabolic network information and describes metabolic changes as a function of the specific growth rate that can be obtained experimentally. Nevertheless, it also does not take into account any inhibitory impact of high metabolite concentrations or of metabolite depletion on cell culture performances. Moreover, a predictive kinetic model of the cell growth and associated metabolic activities is needed as the input of the previously developed model in order to develop an in silico predictive model. In this paper we provide a generalized methodology for prediction of cell growth based on the Monod type kinetics combined with substrate inhibition building on previous work (Amribt et al., 2013; Ben Yahia et al., 2015; Farzan et al., 2016; Nolan and Lee, 2012). This model is linked to the previously developed stoichiometric model connecting growth,

metabolism and production of mAb. We apply it to a case study of a fed-batch cell culture production of a mAb with CHO cells. Such macroscopic models rely on the category of so-called "unstructured models" (Ben Yahia et al., 2015; Borchers et al., 2013; Dhir et al., 2000; García Münzer et al., 2015). A difference from the common Monod type model is the assumption of non-constant maximum specific growth rate which is assumed to be function of the integral viable cell density (IVCD). We also provide insights into the predictive power of simple model structures with regard to the impact of untested feeding strategies on recombinant protein production performance in fed-batch culture.

3.3. Modeling and theoretical aspects

3.3.1. Step 1 : Calibration of the maximum specific growth rate observed

For modeling of mammalian cell culture kinetics, the Monod equation and derivatives are most frequently applied (Amribt et al., 2013; Ben Yahia et al., 2015; Farzan et al., 2016; Nolan and Lee, 2012). Indeed when slightly modified, this formalism enables the prediction of a wide range of characteristics such as saturation, inhibition, and limitation by substrates and other components. Hence we apply this approach to predict cell growth without using any complex parameter calibration.

The growth rate is defined as a function of the maximum specific growth rate, μ_{max}, and of selected metabolite concentrations to simulate limitation and inhibition effects. However, generally μ_{max} is defined as constant (Amribt et al., 2013; Ben Yahia et al., 2015; Farzan et al., 2016; Nolan and Lee, 2012). This does however not describe the observed decrease of the maximum specific growth rate throughout a cell culture production even if optimal experimental conditions are maintained (Borchers et al., 2013; Lee, 2002). We propose the following equation to simulate the maximum specific growth rate which may be caused by the inhibitory effect of an undefined component produced by the cells (Mulukutla et al.):

$$\begin{cases} \mu_{max(obs)} = \mu_{max} - \alpha_I * I^{\beta_I} \\ I = a * \int_0^t VCD * dt = a * IVCD \end{cases} \quad (3.1)$$

where: μ_{max} - maximum specific growth rate (1/day), $\mu_{max(obs)}$ - maximum specific growth rate observed (1/day), *IVCD* - cumulative integral viable cell density (cell.day/mL), *VCD* - Viable cell density (cell/mL), *I* - undefined inhibitory component produced by cells (mg/mL), α_I - constant inhibitory rate of component I ($mL^{\beta I}$ /[$mg^{\beta I}$.day]), β_I - constant index for inhibitory effect, a - constant specific production rate of inhibitory component I (mg/[cell.day]).

As we don't exactly know which inhibitory component is produced by the cells, the IVCD was used as an indicator:

$$\mu_{max_{(obs)}} = \mu_{max} - \alpha * IVCD^{\beta_I} \quad (3.2)$$

where α - constant inhibitory rate of IVCD equal to $a^*\alpha_I$ ($mL^{\beta I}$/[$mg^{\beta I-1}$.cell.day^2]).

3.3.2. Step 2 : Calibration of the generalized model of specific growth rate

The depletion of any essential metabolite but also high concentrations of such metabolites limit cell growth. Here we consider for instance the amino acids tryptophan, histidine, isoleucine, methionine, threonine, phenylalanine, valine, tyrosine, leucine, lysine, glutamine, arginine and cysteine as essential metabolites (Hu and Zhou, 2012). The generalized model of specific growth rate μ is expressed as follows based on Monod type equations:

$$\boldsymbol{\mu} = \begin{cases} \boldsymbol{\mu_{max(obs)} * \prod_{i=1}^{n} \frac{M_i}{M_i + K_i} * \prod_{i=1}^{n} \frac{KI_i}{KI_i + M_i}} & \boldsymbol{if\ \mu_{max(obs)} > 0} \\ \boldsymbol{-\mu_D} & \boldsymbol{else} \end{cases} \quad (3.3)$$

Where, n – number of metabolites, M_i - concentration of essential metabolite i (mg/mL), K_i - half saturation constant of metabolite i (mg/mL), KI_i - inhibition constant of metabolite i (mg/mL) and μ_D - specific death rate (day^{-1}).

When the cells reach an IVCD that leads to a negative maximum specific growth rate observed defined by equation 3.2, the specific growth rate is then defined as equal to the specific death rate μ_D which is assumed to be constant.

3.3.3.Step 3 : Calibration of the cell metabolism and the specific productivity model

As the specific growth rate depends on the concentration of essential metabolites M_i, the production rates of these metabolites have to be predicted within each metabolic phase in which the pseudo steady state approximation is verified. The specific conversion rates of these metabolites can be expressed as follows (Ben Yahia et al., 2017):

$$\boldsymbol{r_{Mi} = a_i * \mu + b_i} \quad \boldsymbol{\forall i \in [1, n]} \quad (3.4)$$

Where: r_{Mi} - specific production rate of metabolite i (mg/[cell.day]), a_i - parameter for metabolite i production dependent on growth ($mg/cell$) and b_i - parameter for metabolite i production independent on growth (mg/[cell.day]). a_i and b_i are constant within each metabolic phase identified according to the method from Ben Yahia et al. (Ben Yahia et al., 2016)

Similarly, in each metabolic phase, the specific productivity of the mAb is defined as linearly dependent on the specific growth rate as follow:

$$\boldsymbol{Q_{p\,max} = a_{mAb} * \mu + b_{mAB}} \quad (3.5)$$

Where: $Q_{p\ max}$ - specific production rate of mAb (mg/[cell.day]), a_{mAb} - parameter for mAb production dependent on growth (mg/cell) and b_{mAb} - parameter for mAb production independent on growth (mg/cell). a_i and b_i are constant within each metabolic phase identified by Ben Yahia et al. (Ben Yahia et al., 2016).

The parameters of these equations can be identified by segmented linear regression (Ben Yahia et al., 2016)

Metabolites limitation in the medium can lead to their depletion. Such depletions can in turn impact mAb and protein synthesis (Gramer, 2014; Kilberg et al., 2009). Moreover, high concentration of essential metabolites can also lead to a decrease in specific productivity. To express these impacts, the specific productivity observed is defined as follows based on Monod type equations:

$$\boldsymbol{Q_p = Q_{p\,max} \prod_{i=1}^{n} \frac{M_i}{M_i + K'_i} \prod_{i=1}^{n} \frac{KI'_i}{KI'_i + M_i}} \quad (3.6)$$

Where: $Q_{p\ max}$ - specific production rate of mAb predicted by equation 3.5 (mg/[cell.day]), M_i - concentration of essential metabolite i (mg/mL), K'_i - half saturation constant of metabolite $i$and KI'_I - Inhibition constant of metabolite i (mg/mL).

3.3.4. Step 4 : Prediction of an accumulation of intracellular metabolite

Modeling of storage components can also be taken into account to extend kinetic models as described in (Richelle and Bogaerts, 2015) and by the case study presented in this paper. If the cell growth model prediction does not fit with experimental data for conditions with depletion of a specific metabolite M_i, the model is updated by taking into account the hypothetic accumulation of intracellular metabolites related to metabolite M_i. We hypothesized that M_i is used for cell growth and mAb production and that M_i related components, called metabolite M_{xi}, accumulate as intracellular pool that can be used for growth when extracellular M_i is depleted. The accumulation rate of the intracellular M_{xi} pool is assumed to be equal to the constant parameter b_i from equation 3.4 minus the rate of production of the mAb adjusted with mass fraction of M_i in the mAb sequence. The part of M_i used for cell growth is defined as $a_i * \mu$. When M_i is depleted, the specific productivity is equal to zero, as described in equation 3.6, and also r_{Mi} (Figure 3.1). This is formulated as follow:

$$\begin{aligned} \boldsymbol{r_{Mxi,pool}} &= \boldsymbol{r_{Mi} - a_i * \mu - Q_p * x_{Mi}} \\ &= \boldsymbol{b_i - Q_p * x_{Mi}} \end{aligned} \qquad (3.7)$$

Where

x_{Mi} : mass fraction of metabolite M_i in mAb (mg/mg)

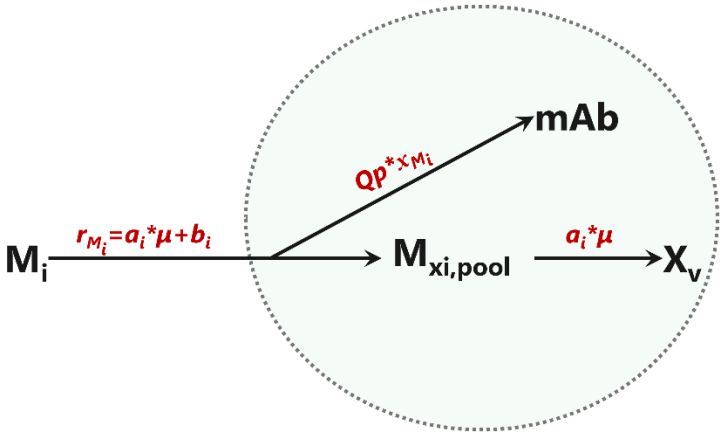


Figure 3.1 Representation of the accumulation of intracellular Mxi pool in a cell.

Metabolite i (Mi) is directly used for mAb production and partly converted into a metabolite Mxi denoted as Mxi,pool that can accumulate and further be used for cell growth. When extracellular Mi is depleted, the Mxi pool can still be consumed to allow cell growth. The accumulation of the intracellular pool is estimated using the segmented linear relationship between the specific production rate of Mi, rMi, and the specific growth rate, rMi=ai*μ+bi. The part not correlated to the cell growth, bi, is defined equal to the specific production rate of mAb adjusted to mass fraction of Mi in the mAb, i.e. xMi, and the rate of accumulation of Mxi pool. The part correlated to the cell growth, ai*μ, is the part of Mi eventually used for cell growth.

Equation 3.3 is then modified to include the impact of intracellular M_{xi} pool on cell growth as follows:

$$\boldsymbol{\mu} = \begin{cases} \boldsymbol{\mu_{max(obs)}} * \frac{\boldsymbol{M_{xi\,pool}}}{\boldsymbol{M_{xi\,pool}+K_i}} * \prod_{i=1}^{n-1} \frac{\boldsymbol{M_i}}{\boldsymbol{M_i+K_i}} * \prod_{i=1}^{n} \frac{\boldsymbol{KI_i}}{\boldsymbol{KI_i+M_i}} & \boldsymbol{if\ \mu_{max(obs)} > 0} \\ \boldsymbol{-\mu_D} & \boldsymbol{else} \end{cases} \quad (3.8)$$

Where: $M_{xj\ pool}$ - intracellular metabolite M_{xj} pool related to metabolite M_j. It is expressed as mass of M_{xj} per bioreactor volume (mg/mL).

A summary of the macroscopic modeling methodology is presented in Figure 3.2.

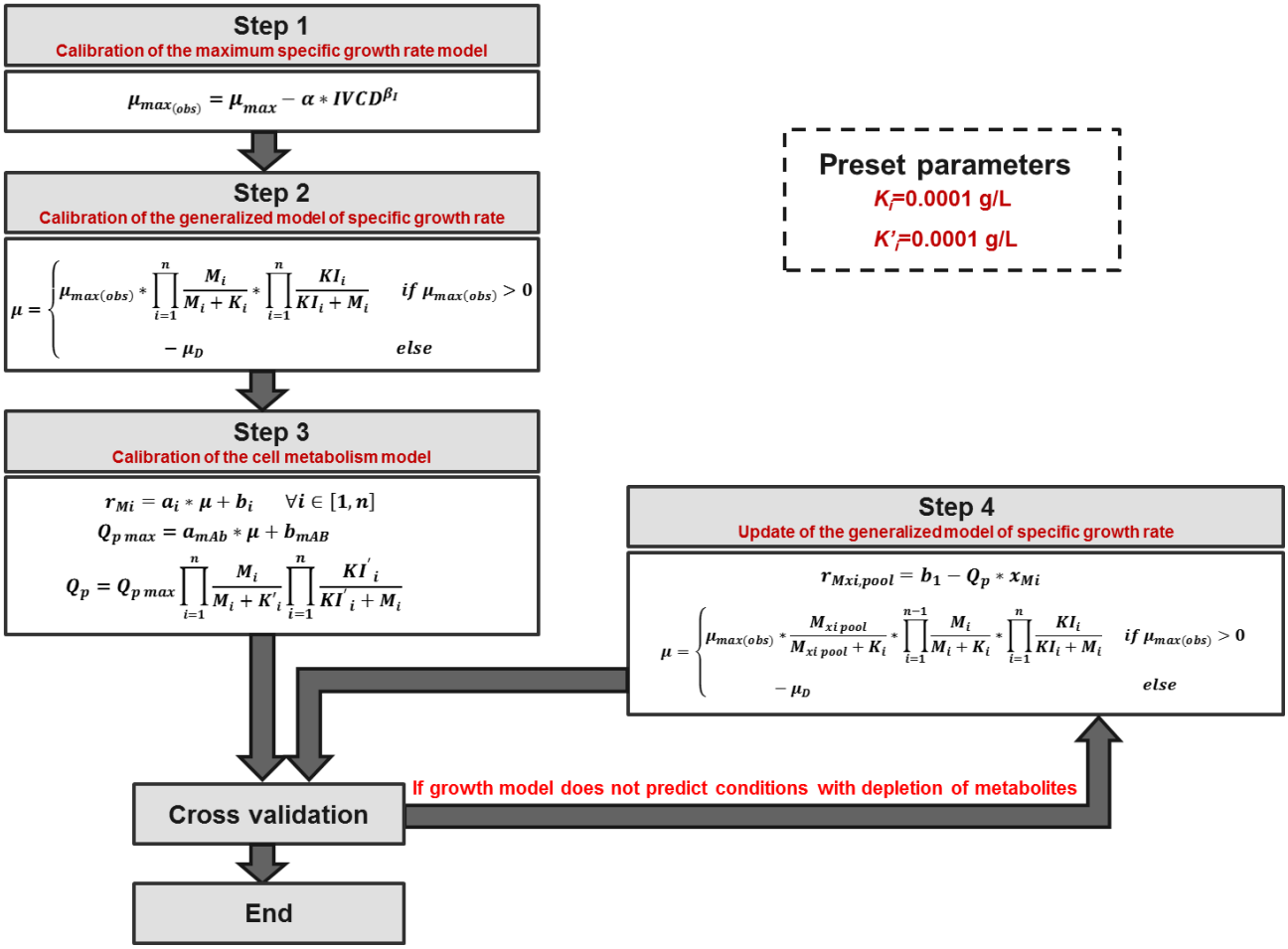


Figure 3.2 Macroscopic modeling methodology. The methodology is separated into 3 mandatory steps and one facultative step. First the maximum specific growth rate model is calibrated with a control conditions with no depletion of metabolites and no inhibitory metabolite concentrations (based on literature). The generalized growth model from equation 3.3 is calibrated in step 2 and finally the cell metabolism model from equation 3.4 and 5 are calibrated using segmented linear regression (Ben Yahia et al., 2017) in step 3. Finally the cell growth, mAb titer and metabolite concentrations are compared to experimental data. If the growth prediction does not fit with experimental data for conditions with depletion of metabolites, the model is updated as described in step 4.

3.4. Case Study

3.4.1. Objective

The case study presented in this paper focuses on the fed-batch production of a mAb produced by Chinese Hamster Ovary (CHO) cells. The critical impact on growth and production of two metabolites present in the bioreactor during the production step were identified in previous experiments to have the largest influence on growth and mAb titer. In particular cell growth and mAb titer were reduced when the concentration of M_1 and/or M_2 in the feed was doubled compared to the condition with the original concentration of M_1 and M_2 in the feed (see supplementary Figure S3.1). The goal of this paper is to predict is the prediction of the impact of these two metabolites on cell growth, metabolism and mAb production thanks to a dynamic model that will enable scalable in silico optimization of the process.

3.4.2. Macroscopic model

Both metabolites M_1 and M_2 are known to highly impact cell growth. Based on the model theory presented in the paper, it is assumed that the depletion of these metabolites stop the cell growth and that high concentrations of those metabolites also inhibit cell growth. Starting from equation 3.3, the model is expressed as follows based on Monod type equations (*Step 2*):

$$\boldsymbol{\mu} = \begin{cases} \boldsymbol{\mu_{max(obs)} * \frac{M_1}{M_1+K_1} * \frac{M_2}{M_2+K_2} * \frac{KI_1}{KI_1+M_1} * \frac{KI_2}{KI_2+M_2}} & \boldsymbol{if\ \mu_{max(obs)} > 0} \\ \boldsymbol{-\mu_D} & \boldsymbol{else} \end{cases} \quad (3.9)$$

With $\mu_{max\,(obs)}$ defined in equation 3.2 (*Step 1*) and parameters K_1 and K_2 are preset to a low value (0.0001 mg/mL). KI_1 and KI_2 have to be identified using experimental data.

As the specific growth rate is depending on the concentrations of M_1 and M_2, the production rate of both metabolites can be predicted. The specific production rate of

any metabolite is defined by equation 3.4 using a segmented linear model (*Step 3*). In this case study, the relationship between both metabolites and the specific growth rate was found to be constant throughout the whole cultivation and we have then:

$$\begin{cases} r_{M1} = a_1 * \mu + b_1 \\ r_{M2} = a_2 * \mu + b_2 \end{cases} \quad (3.10)$$

With a_1, a_2, b_1 and b_2 parameters identified as described in Ben Yahia et al. (Ben Yahia et al., 2016).

Based on equation 3.6 we also have:

$$Q_p = Q_{p\,max} * \frac{M_1}{M_1 + K'_1} * \frac{M_2}{M_2 + K'_2} * \frac{KI'_1}{KI'_1 + M_1} * \frac{KI'_2}{KI'_2 + M_2} \quad (3.11)$$

3.4.3. Material and methods

A genetically modified mAb production cell line (CHO-DG44) was used. The cells were cultivated in 2 L stirred tank glass bioreactor (STR) with supply towers (C-DCUII, Sartorius Stedim Biotech) controlled by a multi-fermentation control system (MFCS, Sartorius Stedim Biotech) where pH, temperature, stirring speed, gas sparging and feed addition were controlled.

The culture was operated in fed-batch mode for 14 days. During the feeding phase, the monoclonal antibody (mAb) is secreted into the medium. Samples were drawn daily to determine total and viable cell number, viability, off-line pH, the partial pressure of CO_2, pCO_2, osmolality, glucose lactate, amino acid and mAb concentrations (stored at -80 °C). Samples for the amino acid analysis were taken before the feed addition. The extracellular concentrations after feeding were computed based on the feed composition information. The specific growth rate, μ, was computed for each experimental condition separately as the slope of the linear trend

line obtained by plotting ln($VCD.V_R$) against time (Chin et al., 2015; Clarke et al., 2011), where V_R represents the bioreactor volume.

3.4.4. Experimental conditions

To identify the inhibitory parameters of the Monod type equations (Equations 3.9 and 3.11), we assessed the impact of high concentrations of those metabolites M_1 and M_2. For that purpose, five experimental conditions with different bolus additions on day 3 and one control condition (experimental condition 1) were designed (Table 3.1). The total quantity of M_1 and M_2 added throughout the cell culture production was the same for each experiment as the feed addition was adapted when bolus addition were performed. To adapt the feeding strategy, the metabolite added as a bolus on day 3 was not added for the next days of production until the control condition reached the same total quantity added of the metabolite. Day 3 was selected based on preliminary screening experiments: cultures were run in a fed-batch mode in 250 mL shake flasks for 8 days, each with a bolus addition of medium with a high medium concentration of M_1 of 0.8 g/L at different time points (supplementary Figure S3.2).

The model calibrated with this data set was also cross validated with a wide range of historical experimental data where the composition of both metabolites M_1 and M_2 in the feed was varied (Table 3.2).

Table 3.1 Experimental conditions to identify inhibitory parameters. The maximum medium concentration of two different metabolites (M_1, M_2) were varied. Bolus additions were performed on day 3 of production in order to reach the maximum concentration depicted in the table.

Experimental ID	Experimental condition	Maximum concentration of M_1 (mg/mL)	Maximum concentration of M_2 (mg/mL)
1	1	0.06	0.2
2	2	0.06	0.6
3	3	0.9	0.6
4	4	0.48	0.38
5	5	0.9	0.38
6	6	0.9	0.2
7	1	0.06	0.2
8	4	0.48	0.38

Table 3.2 Experimental conditions to cross validate the model. The concentration of two different metabolites (M_1, M_2) contained in our feed were varied. They are presented as percentage of the maximum concentration tested. The model of cell growth (Equation 3.10), cell metabolism (Equation 3.8) and mAb production (Equation 3.11) were used to predict cell culture performances of the experimental conditions.

Experimental ID	Experimental condition	M_1 (%)	M_2 (%)
9	7	100	50
10	8	10	50
11	9	50	100
12	10	50	10
13	11	10	10
14	12	50	50
15	13	25	25
16	14	20	50
17	15	10	30
18	16	10	50
19	17	30	50
20	18	30	10
21	19	50	30

3.4.5. Parameter identification

For each steps of the modeling methodology presented in Figure 3.2, parameters were identified as described below.

3.4.5.1. Parameter estimation for Step 1

Parameters of equation 3.2 were identified by analyzing one control condition, i.e. experimental condition 1 (Table 3.1), using the Gauss-Newton method. We assumed that for that control condition, the specific growth rate observed daily corresponds to the maximum growth rate possible as no depletion of metabolites M_1 and M_2 was observed and also M_1 and M_2 maximum concentrations do not exceed inhibitory concentration described in literature.

3.4.5.2. Parameter estimation for Step 2

The specific death rate was identified by taking the mean of the negative specific growth rate values observed at the end of the culture of experimental condition 1 (Table 3.1). Half saturation constants of equation 3.9 are preset to 0.00001 mg/mL, i.e. small enough to be effective on the growth rate and only close to depletion of the respective metabolites (Provost et al., 2006). In order to identify inhibitory parameters of equation 3.9, a response surface design of experiment (DoE) was established (Table 3.1). The input parameters were the maximum concentration of M_1 and M_2 on day 3. The inhibitory parameters were identified using the Gauss-Newton method by assuming that the specific growth rate of the control condition, i.e. experimental conditions 1 (Table 3.1), are the maximum possible between day 3 and day 4. .

3.4.5.3. Parameter estimation for Step 3

Parameters from equations 3.10 and 3.5 were previously identified using segmented linear regression (Ben Yahia et al., 2016). Inhibitory parameters of equation 3.11 were identified based on experimental conditions described in Table 3.1 in the manner than the identification of inhibitory parameters of the generalized cell growth model (Equation 9). Half saturation constants of equation 3.11 are also preset to 0.00001 mg/mL

3.5. Results

Experimental conditions presented in Table 3.1 were performed and the impact on growth profiles is shown in Figure 3.3.

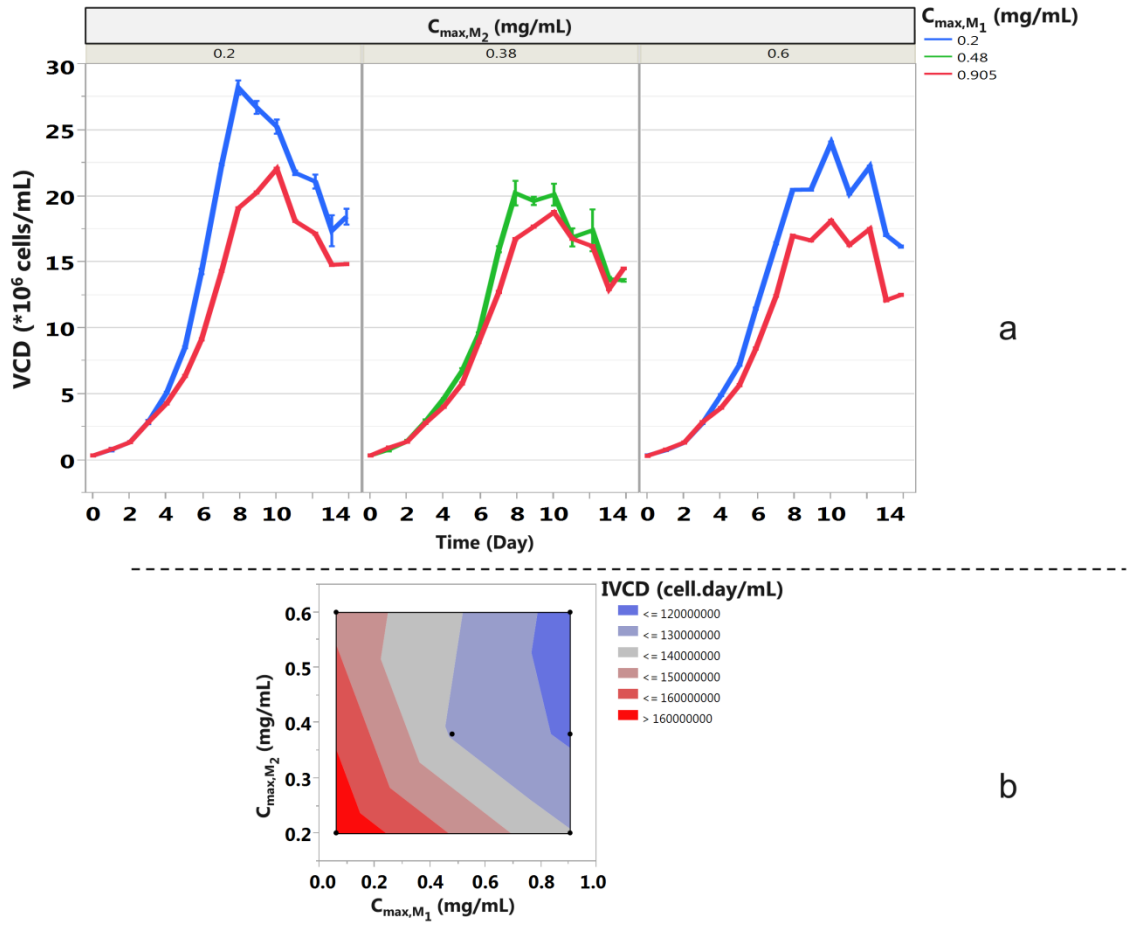


Figure 3.3 Impact of high concentration of M_1 and M_2 on cell growth. In (a), the impact of M_1 and M_2 maximum concentration reached throughout the cell culture production, defined as $C_{max,M1}$ and $C_{max,M2}$ respectively, on cell growth is depicted. The viable cell density profile is presented. Error bars correspond to one standard deviation. In (b), a contour plot of the effect of M_1 and M_2 maximum concentration throughout the cell culture production on the final integral viable cell density (IVCD) is depicted.

3.5.1. Modeling of the maximum specific growth rate observed (Step 1)

The control experiments, i.e. the experimental condition 1 in Table 3.1, was used to calibrate the model of the maximum specific growth rate as a function of the IVCD

(Equation 3.2). Parameters μ_{max}, α and β_I were identified: 0.73 day^{-1}, 2.19*10^{-6} mL$^{0.69}$/[mg$^{-0.31}$.cell.day^2] and 0.69 respectively (Figure 3.4b).

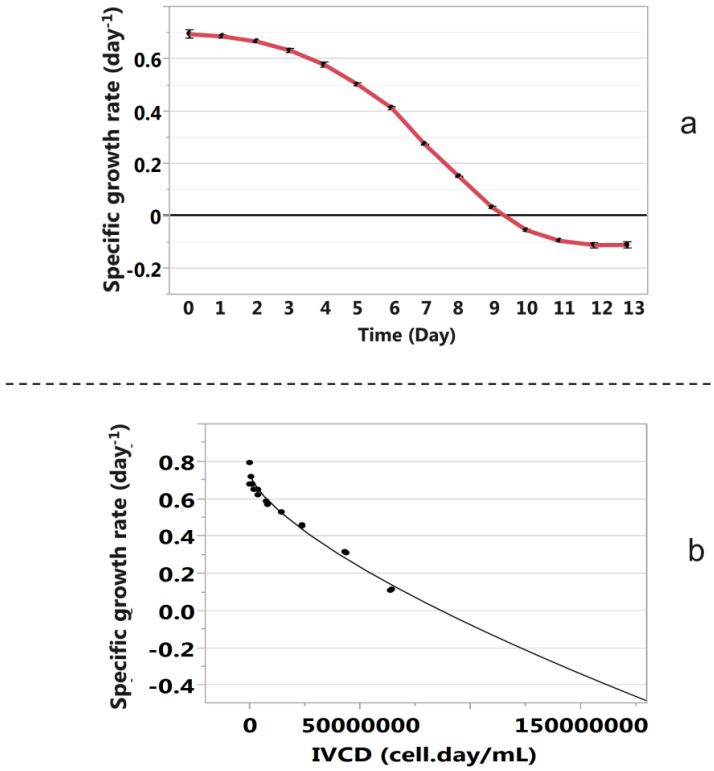


Figure 3.4 Specific growth rate profile for control conditions. In (a), the specific growth rate profiles of control experimental conditions defined by experimental ID 1 in Table 3.1 are depicted as a function of time. In (b), the specific growth rate of control experimental conditions defined by experimental ID 1 in Table 3.1 is expressed as a function of the integral viable cell density IVCD. The maximum specific growth rate is assumed to be inhibited by IVCD with an effect due to possible inhibitory components produced by the cells. Parameters α and β_I (Equation 3.2) were identified: 2.19*10^{-6} mL$^{0.69}$/[mg$^{-0.31}$.cell.day^2] and 0.69, respectively.

3.5.2. Identification of cell growth inhibitory parameters of M1 and M2 (Step 2)

Bolus addition were performed on day 3 in order to reach the maximum concentrations listed in Table 3.1. Before day 3, the experimental conditions were the

same in all experiments. The cell growth behavior is impacted by maximum concentration of M_1 and M_2 reached (Figure 3.3, Figure 3.5a).

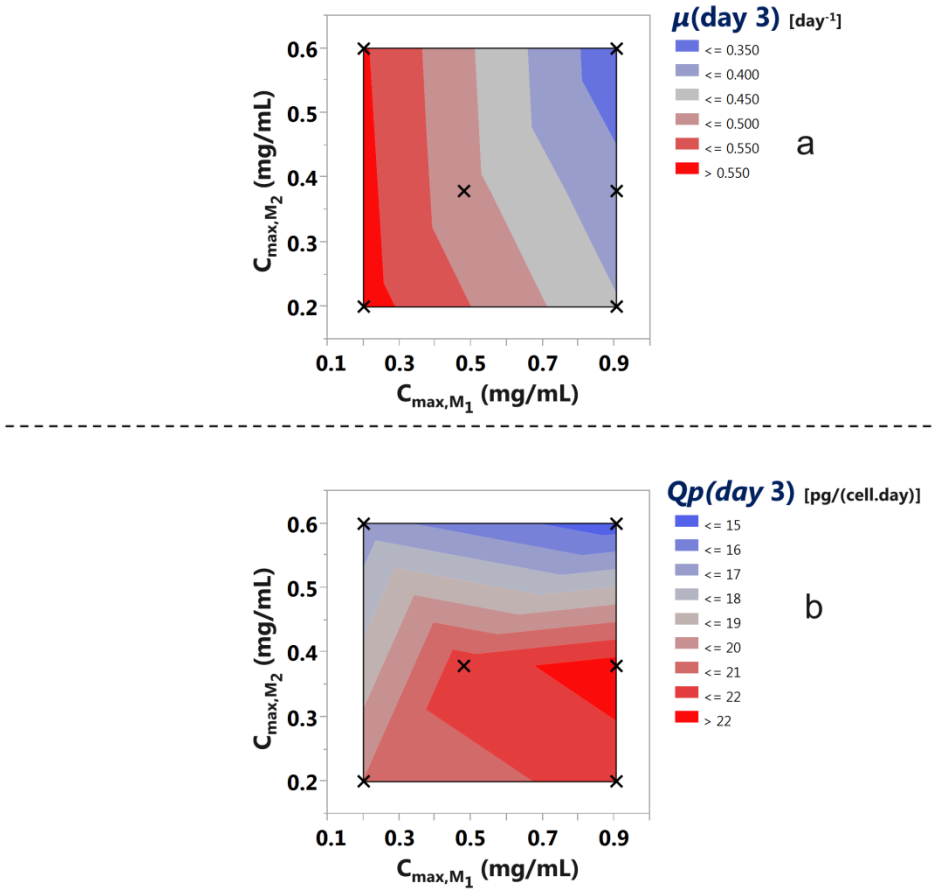


Figure 3.5 Dependence of specific growth rate and specific productivity on maximum concentrations of M_1 and M_2. In (a), the maximum concentration on day 3 of M_1 and M_2, defined as $C_{max,M1}$ and $C_{max,M2}$ respectively, on the specific growth rate between day 3 and day 4 (μ(day 3)). In (b), $C_{max,M1}$ and $C_{max,M2}$ on the specific mAb productivity between day 3 and day 4 (Qp(day3)).

The death of cells follows first order kinetics (Equation 3.9) that is shown by the constant specific death rate observed after day 9 of the cultures for experimental condition 1 (Figure 3.4a). The specific death rate was identified as 0.1 day^{-1}. Parameters KI from equation 3.9 were identified by setting the control conditions, i.e. experimental conditions 1 in Table 3.1, as the maximum specific growth rate between

day 3 and day 4, i.e. 0.55 day^{-1}. For each experimental conditions we defined the specific growth rate on day 3 as:

$$\mu_{(day3)} = 0.55 * \frac{KI_1}{KI_1+M_1} * \frac{KI_2}{KI_2+M_2} \quad (3.12)$$

Parameters were identified using Gauss-Newton algorithm. KI_1 and KI_2 were identified as 1.29 mg/mL and 14.3 mg/mL respectively.

3.5.3. Identification of specific productivity inhibitory parameters of M1 and M2 (Step 3)

The specific mAb productivity on day 3 is also impacted by maximum concentration of M_1 and M_2 reached (Figure 3.5b).

High concentration of M_1 seems to lead to a decrease of specific productivity of mAb (Figure 3.5b). Parameters KI_1' and KI_2' from equation 3.10 were identified by setting the specific production rate at control conditions, i.e. experimental conditions 1 defined in Table 3.1, as the maximum specific productivity on day 3. For each experimental condition we defined the specific productivity of mAb on day 3 as:

$$Q_{p\,(day3)} = Q_{p\,max_{(day3)}} * \frac{KI'_1}{KI'_1+M_1} * \frac{KI'_2}{KI'_2+M_2} \quad (3.13)$$

Parameters were identified using Gauss-Newton algorithm. KI'_1 and KI'_2 were initially set to 72300mg/mL and 2.12 mg/mL respectively. As KI'_1 is very high, high concentrations of M_1 are not inhibitory for specific productivity which confirms the observation of the contour plot in Figure 3.5b. Therefore, equation 3.11 can then be simplified to:

$$Q_p = Q_{p\,max} * \frac{M_1}{M_1+K'_1} * \frac{M_2}{M_2+K'_2} * \frac{KI'_2}{KI'_2+M_2} \quad (3.14)$$

$Q_{p\ max}$ is defined by equation 3.5. The parameters a_{mab} and b_{mab}, but also a_1,a_2, b_1 and b_2, were earlier identified using segmented linear regression (Ben Yahia et al., 2016).

3.5.4. Accumulation of intracellular metabolite (Step 4)

Based on a first assessment of the predictive capacity of the calibrated cell growth model, the prediction of cell growth does not fit with experimental data for some experimental conditions with depletion of metabolite M_1 (data not shown). As presented on step 4 of our modeling methodology (Figure 3.2), the model was updated based on equation 3.7. This is formulated as follow

$$\begin{aligned} \boldsymbol{r_{Mx1,pool}} &= \boldsymbol{r_{M1} - a_1 * \mu - Q_p * x_{M1}} \\ &= \boldsymbol{b_1 - Q_p * x_{M1}} \end{aligned} \quad (3.15)$$

Where

x_{M1} : mass fraction of M_1 in mAb (mg/mg)

In order to verify the hypothesis of an accumulation of a M_{x1} pool, a special experiment was designed. Three experimental conditions with triplicates were performed with various concentration of M_1 in the feed until day 5, i.e. 50%, 25% and 10%, where 100% corresponds to the highest concentration of M_1 in the feed tested. Then the cells were transferred into a medium in which M_1 is absent. After that transfer, only a feed without M_1 was used. The objective was to see if the cells could grow without M_1 and, if this was the case, if they grow better when they were fed with a higher concentration of M_1 previously. Experimental conditions are summarized in Table 3.3.

Experiments at conditions depicted in Table 3.3 were performed. The IVCD is depicted in Figure 3.6 for each experimental condition. Before media exchange, all experiments showed similar IVCD. After media exchange, increasing the percentage of M_1 in the feed before media exchange led to higher IVCD which supports the hypothesis of a possible intracellular accumulation of a metabolite M_{x1} related to metabolite M_1.

Table 3.1 Experimental conditions to test the hypothesis of an intracellular accumulation of a metabolite related to metabolite M_1. The concentration of metabolite M_1 contained in our feed were varied from day 3 to day 5 included. They are presented as percentage of the maximum concentration tested in experimental conditions presented in Table 3.2. On day 5, the cells were transferred into the inoculation media deprived of metabolite M_1 (red arrow in Figure 3.6). After day 5 the cells were only fed with a feed not containing any metabolite M_1.

Experimental ID	Experimental condition	M_1 (%)
22	20	50
23	20	50
24	20	50
25	21	25
26	21	25
27	21	25
28	22	10
29	22	10
30	22	10

At experimental conditions with 25% and 10% M_1 in the feed, M_1 was depleted (data not shown) before media exchange but the duration of depletion was shorter using 25% M_1 in the feed. We assume that for condition with 10% M_1 in the feed, the intracellular pool was highly consumed before media exchange which led to an even lower IVCD compared to 25% M_1 in the feed. Finally, at the experimental condition with 50% M_1 in the feed before media exchange, metabolite M_1 was never depleted before media exchange. Therefore the intracellular metabolite M_{x1} pool is assumed higher than at other experimental conditions. As a consequence of this the resulting final IVCD was highest at 50% M_1 (Figure 3.6).

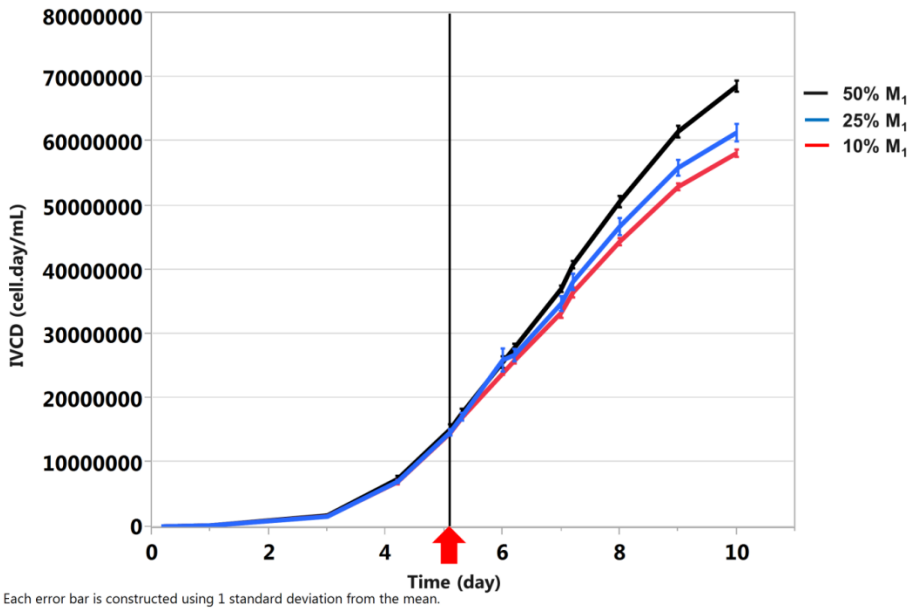


Figure 3.6 Impact of percentage of M_1 in the feed on cell growth after media deprived in M_1 exchange. The integral viable cell density (ICVD) profiles is presented for three experimental conditions with various composition of M_1 in the feed depicted in Table 3.3. On day 5, media was exchanged with a media depleted on metabolite M_1 (red arrow and vertical line).

A summary of the model representation and the corresponding parameters identified for this case study are presented in Table 3.4.

3.5.5. Comparison with experimental data (Cross validation)

The model was used to predict the impact of various concentrations of M_1 and M_2 in the feed on cell growth, mAb titer and M_1 and M_2 concentrations using a similar feeding strategy. Experimental conditions presented in Table 3.2 were not used to calibrate the model. Results are presented in Figures 3.7, 3.8, 3.9 and 3.10. We can observe that the cell growth predicted fits well with experimental data (Figure 3.7). High concentrations and also low concentrations of M_1 in the feed lead to a reduced cell growth as predicted by the model. Experimental mAb titers from day 12 to day 14 were compared with predicted ones (Figure 3.8). A correlation was observed (r^2=0.846) with a small offset for low concentration. The model can predict final mAb titer and identify experimental conditions with higher production.

Table 3.4 Model representations and corresponding parameters identified to predict the impact of two essential metabolites (M_1 and M_2). Model parameters where identified using data from experimental conditions presented in Table 3.1. The parameters of the metabolic model where identified by Ben Yahia et al. **(Ben Yahia, Gourevitch et al. 2016)**.

Model	Model parameter values
$\mu = \begin{cases} \mu_{max(obs)} * \frac{M_{x1\,pool}}{M_{x1\,pool} + K_1} * \frac{M_2}{M_2 + K_2} * \frac{KI_1}{KI_1 + M_1} * \frac{KI_2}{KI_2 + M_2} & if\ \mu_{max(obs)} > 0 \\ -\mu_D & else \end{cases}$	K_1 = 0.0001 mg/mL K_2 = 0.0001 mg/mL KI_1 = 1.29 mg/mL KI_2 = 14.32 mg/mL Parameter identification from Table 3.1
$\mu_{max(obs)} = \mu_{max} - \alpha * IVCC^{\beta_I}$	μ_{max} = 0.73 day^{-1} α = 2.19*10^{-6} $mL^{0.69}$/[$mg^{-0.31}$.cell.day^2] β_I = 0.69 Parameter identification from Table 3.1
$Q_p = Q_{p\,max} * \frac{M_1}{M_1 + K'_1} * \frac{M_2}{M_2 + K'_2} * \frac{KI'_2}{KI'_2 + M_2}$	K'_1 = 0.0001 mg/mL K'_2 = 0.0001 mg/mL KI'_2 = 2.12 mg/mL Parameter identification from Table 3.1
$Q_{p\,max} = a_{mAb} * \mu + b_{mAB}$	Parameter values taken from (Ben Yahia, Gourevitch et al. 2016)
$r_{M1} = a_1 * \mu + b_1$ $r_{M2} = a_2 * \mu + b_2$	Parameter values taken from (Ben Yahia, Gourevitch et al. 2016)

Finally, the concentrations of metabolite M_1 and M_2 were compared to those predicted (Figures 3.9 and 3.10). The model prediction closely followed the experimental trends throughout the cell culture process (Figures 3.9 and 3.10). Based on the metabolite profile (data not shown) we can observe that for experimental condition 7 (Table 3.1), M_1 was depleted from day 6 to day 14 but no impact on the cell growth profile was observed which tend to supports our hypothesis that a metabolite pool related to M_1 is present within the cells and is used when needed. For the experimental conditions with 10% M_1, a depletion is also occurring during the cell culture process but the

influence of M_1 on cell growth is well predicted by taking into account the possible accumulation of an intracellular M_{x1} pool.

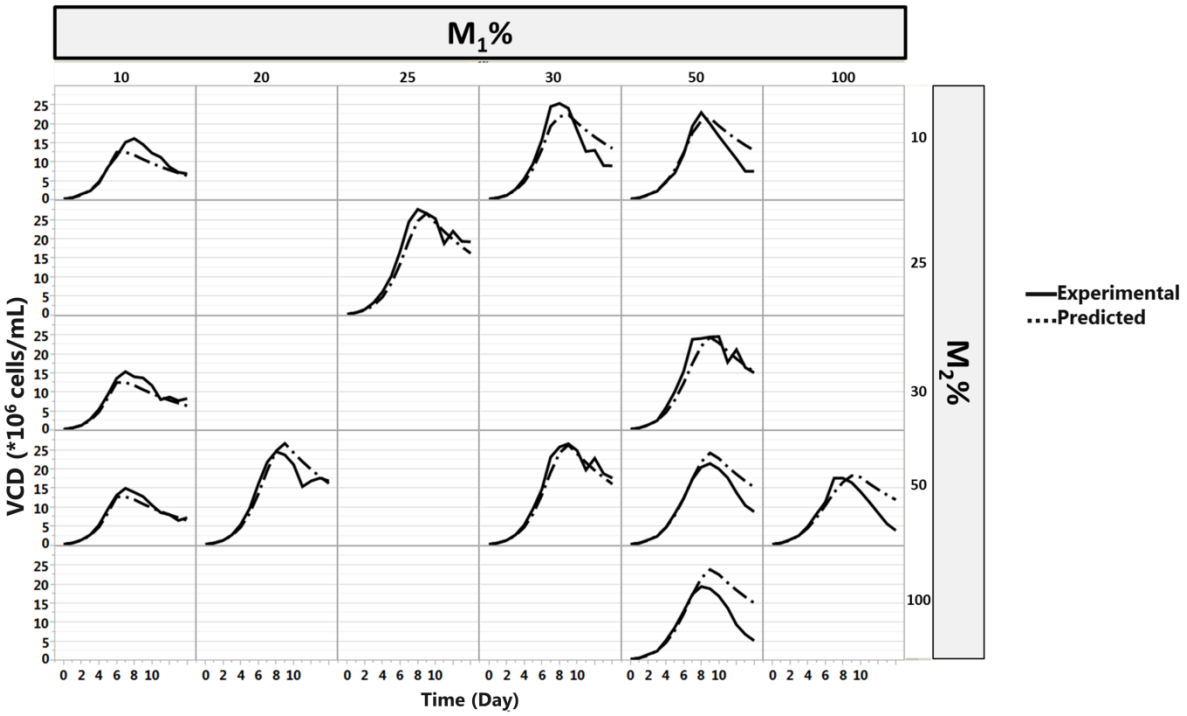


Figure 3.7 Comparison of experimental with predicted VCD. The model developed in the present research (Equations 3.2 and 3.10) was used to predict the cell growth for experimental conditions at various M_1 and M_2 concentrations in the feed, expressed as a percentage of the maximum concentration tested (Table 3.2). The experimental conditions were not used to calibrate the model. The experimental cell growth profile (solid) is compared to predicted cell growth profile (dashed).

To summarize, the model can be used to extrapolate to untested conditions and predict cell growth, metabolite concentrations of targeted metabolites and the mAb titer which supports the suitability of the method. One can highlight that even for some experimental conditions (Table 3.2), where concentrations of M_2 were out of the range tested (Table 3.1), the prediction of the mAb titer, cell growth and metabolite concentrations were still accurate. This proves the predictive power of the model when metabolite concentrations out of the range tested before occur.

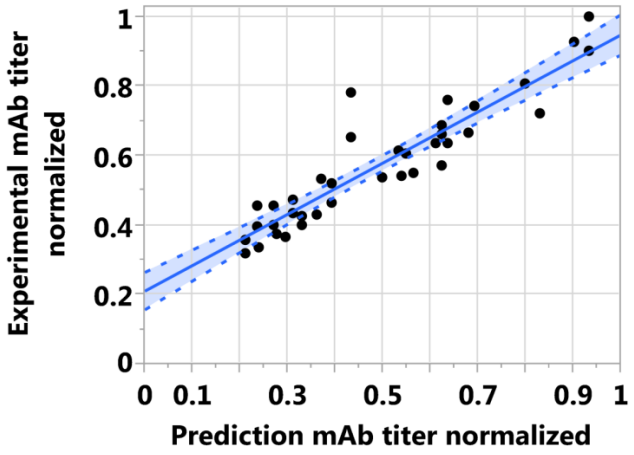


Figure 3.8 Comparison of experimental with predicted mAb titers (from day 12 to day 14). The model developed in the present research was used to predict the mAb production for experimental conditions with various M_1 and M_2 concentrations in the feed, expressed as a percentage of the maximum concentration tested (Table 3.2). The experimental conditions in Table 3.1 were not used to calibrate the model. The experimental mAb titers from day 12 to day 14 are compared to the corresponding predicted mAb titer. The dashed line corresponds to 95% confidence interval. (R^2adjusted=0.846)

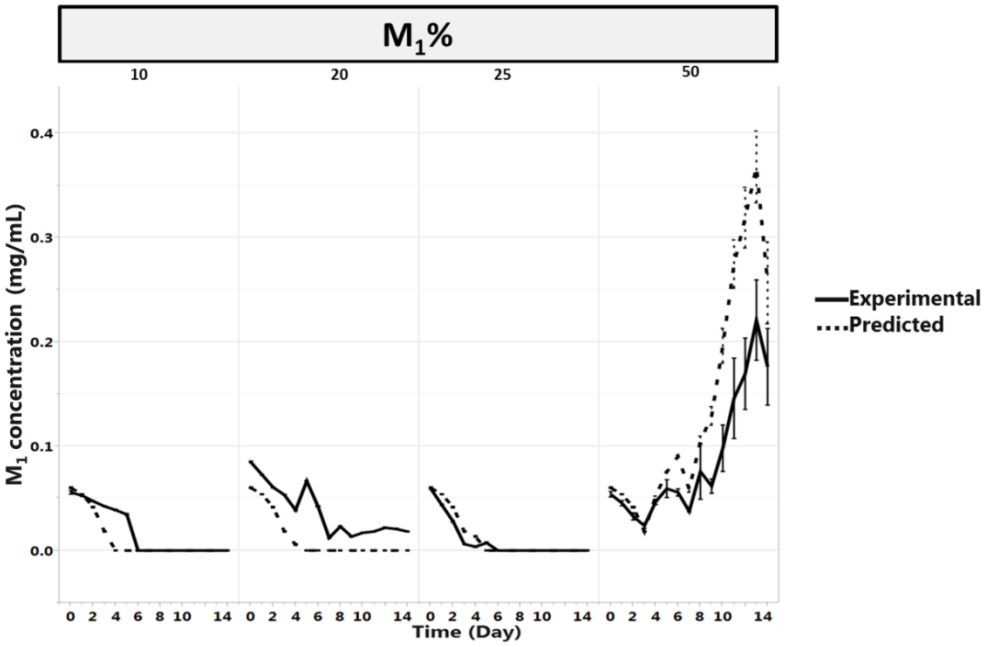


Figure 3.9 Comparison of experimental with predicted M_1 concentrations. The model developed in the present research (Equation 3.8) was used to predict the M_1 concentrations for experimental conditions with various M_1 concentrations in the feed, expressed as a percentage of the maximum concentration tested (Table 3.2). The experimental conditions were not used to calibrate the model. The experimental profile of M_1 (solid) is compared to predicted M_1 concentrations (dashed). Experimental metabolite concentrations were only available for experimental ID 2, 3, 4, 5, 6, 7 and 8.

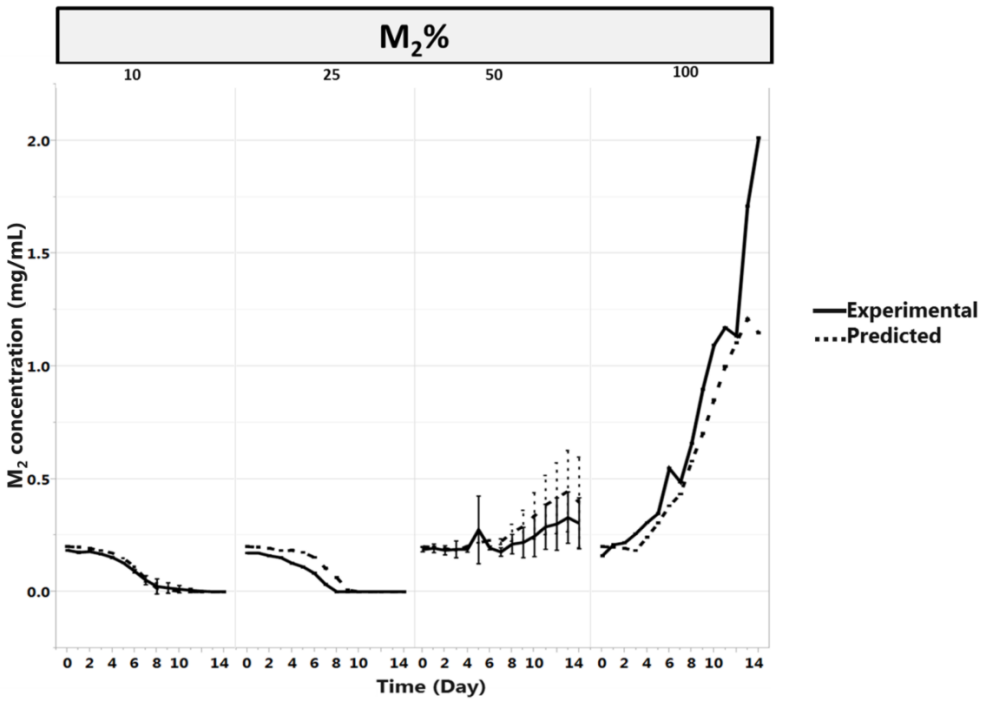


Figure 3.10 Comparison of experimental with predicted M_2 concentrations. The model developed in the present research (Equation 3.8) was used to predict the M_2 concentrations for experimental conditions with various M_2 concentrations in the feed, expressed as a percentage of the maximum concentration tested (Table 3.2). The experimental conditions were not used to calibrate the model. The experimental profile of M_2 (solid) is compared to predicted M_2 concentrations (dashed). Experimental metabolites concentrations were only available for experimental ID 2, 3, 4, 5, 6, 7 and 8.

3.6. Conclusion

By combining a stoichiometric metabolic model based on segmented linear regression presented in a previous paper (Ben Yahia, Gourevitch et al. 2016) and the cell growth model based on Monod type kinetics extended with substrate inhibition presented in this paper, it was possible to develop a simple macroscopic model of cell growth, metabolite concentrations and mAb titer. In total ten kinetic parameters were identified to calibrate the metabolic model, the specific productivity model and the cell growth model, respectively. A major difference compared to usual Monod type cell growth model is that the maximum specific growth rate is not assumed constant but

defined as a function of the IVCD due to possible inhibitory component accumulation. Moreover, another difference from the common Monod type model, is the assumption of accumulation of a specific metabolite in an intracellular pool that can be used for growth when extracellular metabolites are depleted. We highlight that the simple model represented in Figure 3.1 is not the only one able to explain the experimental observations. To estimate this metabolite accumulation, the segmented linear model was successfully used. By using this estimation of metabolite pool accumulation, we were able to predict under which conditions the depletion of metabolite M_1 would impact cell growth. This approach is simple and does not require any complex analysis of intracellular metabolites. Finally, only few experimental conditions are needed to set up the model: i.e. only eight fed-batch cell cultures were performed with only six different experimental conditions. The model can be used to predict cell growth of untested experimental conditions with various feeding strategy of targeted metabolites. In our case study, the model was successfully used to describe the influence of the concentrations of two metabolites (M_1 and M_2) in the feed. If a simple model of the influence of these two metabolites on product quality attributes (PQA) would be developed, this macroscopic model would be able to predict *in silico* the impact of new experimental feeding strategies on cell culture performance and PQA. This is a first step towards reducing the number of bioreactor experiments required to control fed-batch processes for monoclonal antibody production and moving towards *in silico* simulations of the impact of process parameters on product yields and cell metabolism. This model can therefore be further used to predict the best feeding strategy in order to get a high mAb titer and most likely also good PQAs. We strongly believe that this modeling methodology can be applied and extended to any essential medium component. We also emphasize that presented simple model structures can be accurate and predictive.

3.7. Acknowledgements

We gratefully acknowledge the excellent experimental support by the following collaborators of UCB Pharma S.A.: Chevallier Valentine, Annani Meriam and Baddour Yasser.

Chapter 4
4 Conclusion and outlook

Chinese Ovary Hamster (CHO) are an indispensable tool for biotechnological production of biologics which is a multi-million business. Recently, the pharmaceutical industry is increasingly focusing on early drug development which comes with increasing constraints to accelerate process development, reduce costs and demonstrate a deep understanding of cell culture processes. However, the problem with cells *in vivo* is their enormous complexity. Despite the fact that CHO can be cultivated in various types of bioreactors applying sophisticated feeding strategies, we are still not able to characterize and control end to end all the behavior of these cells. Present industrial practices in developing cell culture processes still rely to a large extent on statistical planning of experiments and large series of time-consuming culture development. Modern systems biology promises modeling of such processes on the basis of a system-wide understanding of cellular processes but, as cellular metabolism is composed of thousands of regulated metabolic reactions combined with complex production processes involving fed-batch cultures, the development of complex predictive models is very difficult and time consuming. Moreover, even with our large and increasing knowledge of metabolic networks, the present knowledge on these and on their impact on the production of biologics is still incomplete. This may be one of the reasons that truly predictive models based on detailed system biological model have to be shown to become successful. Nevertheless, as Voit described it in his book (Voit 2000), "*the search for "exact laws" is futile*" i.e. that all models are based on approximations that can be modified, improved and even rejected. From an industrial point of view, applicable predictive models are highly demanded in industry for the purpose of process optimization and control in order to reduce costs and to accelerate process development. To reach to this objective, the focus of the presented thesis was to develop a systematic methodology for metabolic and cell growth modeling that can be easily applied in an

industrial environment. The models developed are believed to identify the essence of the biological mechanism, which makes them relatively simple.

In the present work, simple modeling tools are presented that can be applied to CHO cells to study and predict fed-batch cell culture performances. The simplicity of the modeling approach comes from the low number of parameters to identify but also to the mathematical tools used which are well known and established in the scientific community. In **chapter 1** the steps are introduced that are commonly used in the literature to set up macroscopic models of cell growth and metabolism in mammalian cells. It was identified that there are generally three main steps: the first one is the identification of the input-output relationship. Then a kinetic model is defined in order to link the inputs with the outputs. Finally the model is calibrated and the model parameters are identified. This kind of formalism was applied during the journey to identify and develop a systematic procedure to generate predictive models.

The first step of the work was to develop a predictive model of cell metabolism which was presented in **chapter 2** and was based on a segmented linear model. The objective was first to establish stoichiometric empirical models using metabolite specific rates and knowledge from metabolic engineering field on the principle of metabolic network, e.g. the metabolic steady state paradigm. By using this paradigm, metabolic phases can be identified which simplifies the development of predictive models. The bioproduction process was split into metabolic phases within each the stoichiometric relationships between the specific conversion rates of all metabolites and specific growth rate are constant. A systematic procedure to set up this kind of model is presented. Following the formalism presented in **chapter 1**, the input and outputs selected were the specific growth rate and the specific conversion rates of metabolites, i.e. amino acids, ammonium, glucose, lactate and mAb, respectively. First of all, the specific metabolite rates were estimated from experimental data, i.e. time series of metabolites and product concentrations as well as viable cell number, including fed-batch cultures which required high quality experimental data as the

computation of specific rate is highly sensitive to noise and outliers. To this end, a pre-cleaning of data, i.e. the specific rates computed, was also developed and applied in order to identify and remove possible outliers from the analysis. The model used to identify the metabolic phases and the stoichiometric relationship between the specific rates of metabolites and the specific growth rates was a segmented linear model in order to take into account the metabolic phases that the cell undergo during the cell culture production. To the best of our knowledge, this is the first study that presents a simple segmented linear model structure to describe mAb production but also amino acids, glucose, lactate and ammonium metabolism. The experimental growth rates of small scale production processes (2 L) but also large scale (2000 L) were used to predict the specific metabolic rates. The results show that the model prediction of specific rates of metabolites based on only one variable, i.e. the experimental specific growth rate, was comparable to experimental data at small (2 L) but also large scale (2000 L). Surprisingly, beside the simplicity of the model structure, the accuracy but also the predictability and scalability of the models have been successfully proven during this thesis. An entire and complex metabolic network model is not needed for this model which makes it easily accessible. Moreover, the model was able to predict the impact of the depletion of essential metabolites on the specific productivity. The successfully applied modeling structure supports the initial assumption that even if the inner cell metabolism is complex, the specific production rates of each metabolites can be represented with an eventually simple linear model. In addition, we observed that the metabolites that are impacted by metabolic phases are the one linked to glycolysis and TCA cycle metabolism and cell growth. This information can be used to fully adapt the feeding strategy of those specific metabolites as a function of the metabolic phase. This kind of methodology can contribute to accelerate process development by using rational and systematic development.

The model was accurate and predictive at small but also large scale, but still needed experimental specific growth rate data to fully predict cell metabolism. In a next step in **chapter 3** a fully predictive cell growth model was established by incorporating growth kinetics for the identified phases in order to improve our modeling procedure and get a complete *in silico* model of the bioprocesses. Following the formalism presented in **chapter 1**, the inputs and outputs selected for that step were all the essential metabolites concentrations and the specific growth rate, respectively. The kinetic model developed was based on Monod-type structure but with some modifications. The main change was the use of a non-constant maximum specific growth rate. The logic behind this extension is based on the observation that CHO cells, even if having the characteristics of proliferative/cancer cells, cannot grow indefinitely in fed-batch bioreactors. The hypothesis proposed was that there is a possible accumulation of an inhibitory by-product which is continuously produced by cells. As the cell culture medium is never replaced, the cells cannot grow indefinitely. The other observation supporting this hypothesis, is the fact that cells can grow for months in perfusion mode hypothetically due to the continuous dilution of the inhibitory component. In order to test the developed modeling procedure on a case study with CHO cells, two essential metabolites driving growth kinetics were identified based on historical data. With this simple model structure and with a minimum number of data to calibrate the model, a wide range of new experimental conditions and cell culture performances was predicted. Moreover, in our case study, in order to predict the cell growth, the impact of intracellular metabolites related to M_1 that accumulate in the cell had to be integrated in the model structure. Experimental data supports that hypothesis. To predict the accumulation of those intracellular metabolites pool, the segmented linear model of cell metabolism was successfully used. Therefore, the cell metabolism model (**chapter 2**) can also be extended and used to predict the accumulation of metabolites in intracellular pools but the model prediction of the intracellular metabolite pools was not compared to experimental intracellular pool sizes.

The cell growth model (**chapter 3**) combined with the linear piecewise regression model of cell metabolism (**chapter 2**) allows us to get a complete *in silico* prediction of the impact of untested feeding strategies on cell culture performances. To summarize the modeling procedure, the piecewise linear model of cell metabolism is used to predict the specific rates of metabolites as a function of the specific growth rate and the stoichiometric coefficients identified for each metabolic phase. Then the metabolic concentrations are computed based on the predicted specific rates and on the feeding strategy. Finally the specific growth rate is predicted using the Monod type structure and the metabolic concentrations predicted in previous step. This cycle combining the piecewise linear model and the kinetic model of growth allows to predict *in silico* a complete fed-batch bioprocess. A simple model to predict product quality is also in development and, if combined to the cell metabolism and growth models, can be used to fully predict the impact of new experimental conditions on the main variables such as growth, metabolite concentrations, titer and even PQA.

In summary, the key highlights of the methodologies described in this thesis are that systematic modeling methodologies were developed and validated that capture the essential features for prediction. They are relatively easy to implement but are accurate enough for prediction and for model based optimization of cell culture bioprocesses. This thesis demonstrates what rich knowledge can be derived from high quality macroscopic experimental data, and then used to predict new experiments.

However, the model structures developed in this thesis do not take into account any physicochemical parameter such as the temperature, oxidative stress or pH which make the current model prediction only applicable to similar physicochemical conditions where only the feeding strategy is modified. Nevertheless, it is envisioned that the systematic procedure presented in this thesis is portable to other cell lines but can also be combined with other models that incorporate the influence of pH, osmolality or any other physicochemical parameters. The models can also be improved by incorporating a detailed description of intracellular processes, e.g.

control of metabolic activity, a more detailed model of protein synthesis, folding and secretion, metabolite transports into the cells but also into mitochondria or the incorporation of other omics data such as of proteomics and metabolomics.

The spirit of this thesis was to develop a systematic procedure to get eventually simple predictive models of complex processes - cell metabolism, cell growth and product formation, but also of product quality attributes. The ambitious goal was driving the development of a unique platform for *in silico* process optimization. This plays an important role in refining and developing hypotheses that can then be tested experimentally which will eventually lead to the reduction of cost and wet lab work but also allow the acceleration of process development.

5 References

Acosta, M. L., A. Sánchez, F. García, A. Contreras and E. Molina (2007). "Analysis of kinetic, stoichiometry and regulation of glucose and glutamine metabolism in hybridoma batch cultures using logistic equations." Cytotechnology **54**(3): 189-200.

Ahn, W. S. and M. R. Antoniewicz (2012). "Towards dynamic metabolic flux analysis in CHO cell cultures." Biotechnol J **7**(1): 61-74.

Altamirano, C., A. Illanes, S. Becerra, J. J. Cairó and F. Gòdia (2006). "Considerations on the lactate consumption by CHO cells in the presence of galactose." J Biotechnol **125**(4): 547-556.

Altamirano, C., A. Illanes, A. Casablancas, X. Gámez, J. J. Cairó and C. Gòdia (2001). "Analysis of CHO Cells Metabolic Redistribution in a Glutamate-Based Defined Medium in Continuous Culture." Biotechnol Prog **17**(6): 1032-1041.

Amribt, Z., H. Niu and P. Bogaerts (2013). "Macroscopic modelling of overflow metabolism and model based optimization of hybridoma cell fed-batch cultures." Biochem Eng J **70**(0): 196-209.

Antoniewicz, M. R. (2013). "Dynamic metabolic flux analysis --tools for probing transient states of metabolic networks." Curr Opin Biotechnol **24**(6): 973-978.

Ashyraliyev, M., Y. Fomekong-Nanfack, J. A. Kaandorp and J. G. Blom (2009). "Systems biology: parameter estimation for biochemical models." FEBS J **276**(4): 886-902.

Banerjee, U. C. (1993). "Evaluation of different bio-kinetic parameters of Curvularia lunata at different environmental conditions." Biotechnol Tech **7**(9): 635-638.

Bastin, G. and D. Dochain (1990). On-line estimation and adaptive control of bioreactors, Elsevier.

Batt, B. C. and D. S. Kompala (1989). "A structured kinetic modeling framework for the dynamics of hybridoma growth and monoclonal antibody production in continuous suspension cultures." Biotechnol Bioeng **34**(4): 515-531.

Baughman, A. C., X. Huang, S. T. Sharfstein and L. L. Martin (2010). "On the dynamic modeling of mammalian cell metabolism and mAb production." Comput Chem Eng **34**(2): 210-222.

Ben Yahia, B., B. Gourevitch, L. Malphettes and E. Heinzle (2016). "Segmented linear modelling of CHO fed-batch culture and its application to large scale production." Biotechnol Bioeng **114**(4): 785-797.

Ben Yahia, B., L. Malphettes and E. Heinzle (2015). "Macroscopic modeling of mammalian cell growth and metabolism." Appl Microbiol Biotechnol **99**(17): 7009-7024.

Bersimis, S., J. Panaretos, S. Psarakis and L.-M.-U. M. Volkswirtschaftliche Fakultät (2005). "Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry." Transport Res B-Meth **81**(P1): 78-102.

Borchers, S., S. Freund, A. Rath, S. Streif, U. Reichl and R. Findeisen (2013). "Identification of Growth Phases and Influencing Factors in Cultivations with AGE1.HN Cells Using Set-Based Methods." PLoS ONE **8**(8): e68124.

Bree, M. A., P. Dhurjati, R. F. Geoghegan and B. Robnett (1988). "Kinetic modelling of hybridoma cell growth and immunoglobulin production in a large-scale suspension culture." Biotechnol Bioeng **32**(8): 1067-1072.

Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang and P. D. Karp (2014). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." Nucleic Acids Res **42**(D1): D459-D471.

Chen, L. and G. Bastin (1996). "Structural identifiability of the yield coefficients in bioprocess models when the reaction rates are unknown." Math Biosci **132**(1): 35-67.

Chen, L., O. Bernard, G. Bastin and P. Angelov (2000). "Hybrid modelling of biotechnological processes using neural networks." Control Eng Pract **8**(7): 821-827.

Chin, C. L., H. K. Chin, C. S. Chin, E. T. Lai and S. K. Ng (2015). "Engineering selection stringency on expression vector for the production of recombinant human alpha1-antitrypsin using Chinese Hamster ovary cells." BMC Biotechnol **15**: 44.

Clarke, C., P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard and M. Clynes (2011). "Predicting cell-specific productivity from CHO gene expression." J Biotechnol **151**(2): 159-165.

Cleveland, W. S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." J Am Stat Assoc **74**(368): 829-836.

Craven, S., J. Whelan and B. Glennon (2014). "Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller." J Process Contr **24**(4): 344-357.

de Tremblay, M., M. Perrier, C. Chavarie and J. Archambault (1992). "Optimization of fed-batch culture of hybridoma cells using dynamic programming: single and multi feed cases." Bioprocess Eng **7**(5): 229-234.

Dean, J. and P. Reddy (2013). "Metabolic analysis of antibody producing CHO cells in fed-batch production." Biotechnol Bioeng **110**(6): 1735-1747.

Deshpande, R., T. H. Yang and E. Heinzle (2009). "Towards a metabolic and isotopic steady state in CHO batch cultures for reliable isotope-based metabolic profiling." Biotechnol J **4**(2): 247-263.

Dhir, S., K. J. Morrow, R. R. Rhinehart and T. Wiesner (2000). "Dynamic optimization of hybridoma growth in a fed-batch bioreactor." Biotechnol Bioeng **67**(2): 197-205.

Dorka, P., C. Fischer, H. Budman and J. Scharer (2009). "Metabolic flux-based modeling of mAb production during batch and fed-batch operations." Bioproc Biosyst Eng **32**(2): 183-196.

Dunn, I. J., E. Heinzle, J. Ingham and J. E. Přenosil (2003). Biological Reaction Engineering. Dynamic Modelling Fundamentals with Simulation Exercises, Wiley-VCH, Weinheim

Fan, Y., I. Jimenez Del Val, C. Müller, A. M. Lund, J. W. Sen, S. K. Rasmussen, C. Kontoravdi, D. Baycin-Hizal, M. J. Betenbaugh, D. Weilguny and M. R. Andersen (2015). "A multi-pronged investigation into the effect of glucose starvation and culture duration on fed-batch CHO cell culture." Biotechnol Bioeng **112**(10): 2172-2184.

Farzan, P., B. Mistry and M. G. Ierapetritou (2016). "Review of the important challenges and opportunities related to modeling of mammalian cell bioreactors." AIChE J: DOI: 10.1002/aic.15442.

Gagneur, J. and S. Klamt (2004). "Computation of elementary modes: a unifying framework and the new binary approach." BMC Bioinformatics **5**(1): 175.

Gao, J., V. M. Gorenflo, J. M. Scharer and H. M. Budman (2007). "Dynamic metabolic modeling for a MAB bioprocess." Biotechnol Prog **23**(1): 168-181.

García Münzer, D. G., M. Ivarsson, C. Usaku, T. Habicher, M. Soos, M. Morbidelli, E. N. Pistikopoulos and A. Mantalaris (2015). "An unstructured model of metabolic and temperature dependent cell cycle arrest in hybridoma batch and fed-batch cultures." Biochem Eng J **93**(0): 260-273.

García Münzer, D. G., M. Kostoglou, M. C. Georgiadis, E. N. Pistikopoulos and A. Mantalaris (2015). "Cyclin and DNA Distributed Cell Cycle Model for GS-NS0 Cells." PLoS Comput Biol **11**(2): e1004062.

Gaudard, M., P. Ramsey and M. Stephens (2006). "Interactive data mining and design of experiments: The JMP partition and custom design platforms." CaryNorth Carolina: North Haven Group, LLC.

Gibson, A. M., N. Bratchell and T. A. Roberts (1987). "The effect of sodium chloride and temperature on the rate and extent of growth of Clostridium botulinum type A in pasteurized pork slurry." J Appl Bacteriol. **62**(6): 479-490.

Glacken, M. W., E. Adema and A. J. Sinskey (1988). "Mathematical descriptions of hybridoma culture kinetics: I. Initial metabolic rates." Biotechnol Bioeng **32**(4): 491-506.

Goudar, C. T. (2012). "Analyzing the dynamics of cell growth and protein production in mammalian cell fed-batch systems using logistic equations." J Ind Microbiol Biotechnol **39**(7): 1061-1071.

Goudar, C. T. (2012). "Computer programs for modeling mammalian cell batch and fed-batch cultures using logistic equations." Cytotechnology **64**(4): 465-475.

Goudar, C. T., R. Biener, C. Zhang, J. Michaels, J. Piret and K. Konstantinov (2006). Towards Industrial Application of Quasi Real-Time Metabolic Flux Analysis for Mammalian Cell Culture. Cell Culture Engineering. W.-S. Hu, Springer Berlin Heidelberg. **101:** 99-118.

Goudar, C. T., K. Joeris, K. B. Konstantinov and J. M. Piret (2005). "Logistic Equations Effectively Model Mammalian Cell Batch and Fed-Batch Kinetics by Logically Constraining the Fit." Biotechnol Prog **21**(4): 1109-1118.

Goudar, C. T., K. B. Konstantinov and J. M. Piret (2009). "Robust parameter estimation during logistic modeling of batch and fed-batch culture kinetics." Biotechnol Prog **25**(3): 801-806.

Graefe, J., P. Bogaerts, J. Castillo, M. Cherlet, J. Wérenne, P. Marenbach and R. Hanus (1999). "A new training method for hybrid models of bioprocesses." Bioproc Biosyst Eng **21**(5): 423-429.

Grafahrend-Belau, E., A. Junker, A. Eschenröder, J. Müller, F. Schreiber and B. H. Junker (2013). "Multiscale Metabolic Modeling: Dynamic Flux Balance Analysis on a Whole-Plant Scale." Plant Physiol **163**(2): 637-647.

Gramer, M. J. (2014). "Product quality considerations for mammalian cell culture process development and manufacturing." Adv Biochem Eng Biotechnol **139**: 123-166.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." Econometrica **37**(3): 424-438.

Haas, V. C., P. Lane, M. Hoffmann, B. Frahm, J.-O. Schwabe, R. Pörtner and A. Munack (2001). Model-Based Control of Hybridoma Cell Cultures. 8th IFAC international conference, Québec City, Pergamon Press.

Harder, A. and J. A. Roels (1982). Application of simple structured models in bioengineering. Microbes and Engineering Aspects, Springer Berlin Heidelberg. **21:** 55-107.

Hoops, S., S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes and U. Kummer (2006). "COPASI—a COmplex PAthway SImulator." Bioinformatics **22**(24): 3067-3074.

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables Into Principal Components." J Educ Psychol. **24**: 417-441.

Hu, W. S. and W. Zhou (2012). Cell Culture Bioprocess Engineering, Wei-Shou Hu. **1:** 162-166.

Jang, J. D. and J. P. Barford (2000). "An unstructured kinetic model of macromolecular metabolism in batch and fed-batch cultures of hybridoma cells producing monoclonal antibody." Biochem Eng J **4**(2): 153-168.

Jolicoeur, P. and J. Pontier (1989). "Population growth and decline: a four-parameter generalization of the logistic curve." J Theor Biol **141**(4): 563-571.

Jungers, R. M., F. Zamorano, V. D. Blondel, A. V. Wouwer and G. Bastin (2011). "Fast computation of minimal elementary decompositions of metabolic flux vectors." Automatica **47**(6): 1255-1259.

Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2014). "Data, information, knowledge and principle: back to metabolism in KEGG." Nucleic Acids Res **42**(D1): D199-D205.

Karra, S., B. Sager and M. N. Karim (2010). "Multi-Scale Modeling of Heterogeneities in Mammalian Cell Culture Processes." Ind Eng Chem Res **49**(17): 7990-8006.

Kell, D. B. and J. D. Knowles (2006). The Role of Modeling in Systems Biology. Syst Model Cell Biol. Z. Szallasi, J. Stelling and V. Periwal, The MIT Press**:** 3-18.

Kessel, M. (2011). "The problems with today's pharmaceutical business[mdash]an outsider's view." Nat Biotechnol **29**(1): 27-33.

Kilberg, M. S., J. Shan and N. Su (2009). "ATF4-dependent transcription mediates signaling of amino acid limitation." Trends Endocrinol Metab **20**(9): 436-443.

Kiparissides, A., M. Rodriguez-Fernandez, S. Kucherenko, A. Mantalaris and E. Pistikopoulos (2008). Application of global sensitivity analysis to biological models. Computer Aided Chemical Engineering. B. Bertrand and J. Xavier, Elsevier. **Volume 25:** 689-694.

Klamt, S. and A. von Kamp (2011). "An application programming interface for CellNetAnalyzer." Biosystems **105**(2): 162-168.

Klein, T., E. Heinzle and K. Schneider (2013). "Metabolic fluxes in Schizosaccharomyces pombe grown on glucose and mixtures of glycerol and acetate." Appl Microbiol Biotechnol **97**(11): 5013-5026.

Klein, T., J. Niklas and E. Heinzle (2015). "Engineering the supply chain for protein production/secretion in yeasts and mammalian cells." J Ind Microbiol Biotechnol **42**(3): 453-464.

Kontoravdi, C., S. P. Asprey, E. N. Pistikopoulos and A. Mantalaris (2007). "Development of a dynamic model of monoclonal antibody production and glycosylation for product quality monitoring." Comput Chem Eng **31**(5–6): 392-400.

Laursen, S. Ö., D. Webb and W. F. Ramirez (2007). "Dynamic hybrid neural network model of an industrial fed-batch fermentation process to produce foreign protein." Comput Chem Eng **31**(3): 163-170.

Lee, E. Y. (2002). "Kinetic analysis of the effect of cell density on hybridoma cell growth in batch culture." Biotechnol Bioproc E **7**(2): 117-120.

Liu, Y.-H., J.-X. Bi, A.-P. Zeng and J.-Q. Yuan (2008). "A simple kinetic model for myeloma cell culture with consideration of lysine limitation." Bioproc Biosyst Eng **31**(6): 569-577.

Luedeking, R. and E. L. Piret (1959). "A kinetic study of the lactic acid fermentation. Batch process at controlled pH." J Biochem Microbiol Technol Eng **1**(4): 393-412.

Mahadevan, R. and F. J. Doyle, III (2003). "On-Line Optimization of Recombinant Product in a Fed-Batch Bioreactor." Biotechnol Prog **19**(2): 639-646.

Malinowski, E. R. (1991). Factor Analysis in Chemistry, 2nd Edition, Wiley-Interscience.

Marique, T., M. Cherlet, V. Hendrick, F. Godia, G. Kretzmer and J. Wérenne (2001). "A general artificial neural network for the modelization of culture kinetics of different CHO strains." Cytotechnology **36**(1-3): 55-60.

Martin, H. J., J. C. Reynolds, S. Riazanskaia and C. L. P. Thomas (2014). "High throughput volatile fatty acid skin metabolite profiling by thermal desorption secondary electrospray ionisation mass spectrometry." Analyst **139**(17): 4279-4286.

Mason, R. L. (1997). "A Practical Approach for Interpreting Multivariate T2 Control Chart Signals." J Qual Technol **29**(4): 396-406.

McGee, V. E. and W. T. Carleton (1970). "Piecewise Regression." J Am Stat Assoc **65**(331): 1109-1124.

Meshram, M., S. Naderi, B. McConkey, B. Ingalls, J. Scharer and H. Budman (2013). "Modeling the coupled extracellular and intracellular environments in mammalian cell culture." Metab Eng **19**(0): 57-68.

Mojena, R. (1977). "Hierarchical grouping methods and stopping rules: an evaluation." Comput J **20**(4): 359-363.

Muggeo, V. M. (2003). "Estimating regression models with unknown break-points." Stat Med **22**(19): 3055-3071.

Mulukutla, B. C., J. Kale, T. Kalomeris, M. Jacobs and G. W. Hiller "Identification and control of novel growth inhibitors in fed-batch cultures of Chinese hamster ovary cells." Biotechnology and Bioengineering: n/a-n/a.

Mulukutla, B. C., A. Yongky, S. Grimm, P. Daoutidis and W. S. Hu (2015). "Multiplicity of steady states in glycolysis and shift of metabolic state in cultured mammalian cells." PLoS One **10**(3): e0121561.

Murtagh, F. (1983). "A Survey of Recent Advances in Hierarchical Clustering Algorithms." Comput J **26**(4): 354-359.

Naderi, S., M. Meshram, C. Wei, B. McConkey, B. Ingalls, H. Budman and J. Scharer (2011). "Development of a mathematical model for evaluating the dynamics of normal and apoptotic Chinese hamster ovary cells." Biotechnol Prog **27**(5): 1197-1205.

Narkewicz, M. R., S. D. Sauls, S. S. Tjoa, C. Teng and P. V. Fennessey (1996). "Evidence for intracellular partitioning of serine and glycine metabolism in Chinese hamster ovary cells." Biochem J **313 (Pt 3)**: 991-996.

Neermann, J. and R. Wagner (1996). "Comparative analysis of glucose and glutamine metabolism in transformed mammalian cell lines, insect and primary liver cells." J Cell Physiol **166**(1): 152-169.

Nicolae, A., J. Wahrheit, J. Bahnemann, A.-P. Zeng and E. Heinzle (2014). "Non-stationary 13C metabolic flux analysis of Chinese hamster ovary cells in batch culture using extracellular labeling highlights metabolic reversibility and compartmentation." BMC Syst Biol **8**(1): 50.

Niklas, J. and E. Heinzle (2012). "Metabolic flux analysis in systems biology of Mammalian cells." Adv Biochem Eng Biotechnol **127**: 109 - 132.

Niklas, J., F. Noor and E. Heinzle (2009). "Effects of drugs in subtoxic concentrations on the metabolic fluxes in human hepatoma cell line Hep G2." Toxicol Appl Pharmacol **240**(3): 327-336.

Niklas, J., C. Priesnitz, T. Rose, V. Sandig and E. Heinzle (2013). "Metabolism and metabolic burden by α1-antitrypsin production in human AGE1.HN cells." Metab Eng **16**: 103-114.

Niklas, J., E. Schräder, V. Sandig, T. Noll and E. Heinzle (2011). "Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1.HN using time resolved metabolic flux analysis." Bioproc Biosyst Eng **34**(5): 533-545.

Niu, H., Z. Amribt, P. Fickers, W. Tan and P. Bogaerts (2013). "Metabolic pathway analysis and reduction for mammalian cell cultures—Towards macroscopic modeling." Chem Eng Sci **102**(0): 461-473.

Nolan, R. P. and K. Lee (2011). "Dynamic model of CHO cell metabolism." Metab Eng **13**(1): 108-124.

Nolan, R. P. and K. Lee (2012). "Dynamic model for CHO cell engineering." J Biotechnol **158**(1–2): 24-33.

Nottorf, T., W. Hoera, H. Buentemeyer, S. Siwiora-Brenke, A. Loa and J. Lehmann (2007). Production of Human Growth Hormone in a Mammalian Cell High Density Perfusion Process. Cell Technology for Cell Products. R. Smith, Springer Netherlands. **3:** 789-793.

Oliveira, R. (2004). "Combining first principles modelling and artificial neural networks: a general framework." Comput Chem Eng **28**(5): 755-766.

Ozturk, S. S. and B. O. Palsson (1990). "Chemical Decomposition of Glutamine in Cell Culture Media: Effect of Media Type, pH, and Serum Concentration." Biotechnol Prog **6**(2): 121-128.

Ozturk, S. S. and B. O. Palsson (1991). "Growth, metabolic, and antibody production kinetics of hybridoma cell culture: 2. Effects of serum concentration, dissolved oxygen concentration, and medium pH in a batch reactor." Biotechnol Prog **7**(6): 481-494.

Ozturk, S. S., M. R. Riley and B. O. Palsson (1992). "Effects of ammonia and lactate on hybridoma growth, metabolism, and antibody production." Biotechnol Bioeng **39**(4): 418-431.

Palomares, L. A., S. Estrada-Moncada and O. Ramírez (2004). Production of Recombinant Proteins. Recombinant Gene Expression : Reviews and Protocols. P. Balbás and A. Lorence, Humana Press. **267:** 15-51.

Pearson, K. (1901). "{On lines and planes of closest fit to systems of points in space}." Philos Mag **2**(6): 559-572.

Pirt, S. J. (1965). "The maintenance energy of bacteria in growing cultures." Proc R Soc Lond B Biol Sci **163**(991): 224-231.

Pirt, S. J. (1982). "Maintenance energy: a general model for energy-limited and energy-sufficient growth." Arch Microbiol **133**(4): 300-302.

Pisu, M., A. Concas and G. Cao (2015). "A novel quantitative model of cell cycle progression based on cyclin-dependent kinases activity and population balances." Comput Biol Chem **55**(0): 1-13.

Pörtner, R., A. Schilling, I. Lüdemann and H. Märkl (1996). "High density fed-batch cultures for hybridoma cells performed with the aid of a kinetic model." Bioprocess Eng **15**(3): 117-124.

Provost, A. and G. Bastin (2004). "Dynamic metabolic modelling under the balanced growth condition." J Process Contr **14**(7): 717-728.

Provost, A., G. Bastin, S. N. Agathos and Y. J. Schneider (2006). "Metabolic design of macroscopic bioreaction models: application to Chinese hamster ovary cells." Bioproc Biosyst Eng **29**(5-6): 349-366.

Quek, L.-E., S. Dietmair, J. O. Krömer and L. K. Nielsen (2010). "Metabolic flux analysis in mammalian cell culture." Metab Eng **12**(2): 161-171.

Rehberg, M., A. Rath, J. B. Ritter, Y. Genzel and U. Reichl (2014). "Changes in intracellular metabolite pools during growth of adherent MDCK cells in two different media." Appl Microbiol Biotechnol **98**(1): 385-397.

Richelle, A. and P. Bogaerts (2015). "Macroscopic Modelling of Intracellular Reserve Carbohydrates Production during Baker's Yeast Cultures." IFAC-PapersOnLine **48**(1): 731-736.

Royle, K. E., I. Jimenez del Val and C. Kontoravdi (2013). "Integration of models and experimentation to optimise the production of potential biotherapeutics." Drug Discov Today **18**(23–24): 1250-1255.

Rügen, M., A. Bockmayr, J. Legrand and G. Cogne (2012). "Network reduction in metabolic pathway analysis: Elucidation of the key pathways involved in the photoautotrophic growth of the green alga Chlamydomonas reinhardtii." Metab Eng **14**(4): 458-467.

Ryan, S. E., L. S. Porth and S. Rocky Mountain Research (2007). A tutorial on the piecewise regression approach applied to bedload transport data. Fort Collins, CO, U.S. Dept. of Agriculture, Forest Service, Rocky Mountain Research Station.

Sainz, J., F. Pizarro, J. R. Pérez-Correa and E. Agosin (2003). "Modeling of yeast metabolism and process dynamics in batch fermentation." Biotechnol Bioeng **81**(7): 818-828.

Saner, U., E. Heinzle and D. Bonvin (1992). Computation of stochiometric models for bioprocess. 2nd IFAC Symposium on Modeling and Control of Biotechnical Processes. I. M. a. C. o. B. Processes. Keystone, Colorado, IFAC Modeling and Control of Biotechnical Processes.

Schuster, R. and S. Schuster (1993). "Refined algorithm and computer program for calculating all non–negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed." Comput Appl Biosci **9**(1): 79-85.

Schuster, S., T. Dandekar and D. A. Fell (1999). "Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering." Trends Biotechnol **17**(2): 53-60.

Selişteanu, D., D. Şendrescu, V. Georgeanu and M. Roman (2015). "Mammalian Cell Culture Process for Monoclonal Antibody Production: Nonlinear Modelling and Parameter Estimation." BioMed Res Int **2015**: 16.

Sellick, C. A., A. S. Croxford, A. R. Maqsood, G. Stephens, H. V. Westerhoff, R. Goodacre and A. J. Dickson (2011). "Metabolite profiling of recombinant CHO cells: Designing tailored feeding regimes that enhance recombinant antibody production." Biotechnol Bioeng **108**(12): 3025-3031.

Sengupta, N., S. T. Rose and J. A. Morgan (2011). "Metabolic flux analysis of CHO cell metabolism in the late non-growth phase." Biotechnol Bioeng **108**(1): 82-92.

Sidoli, F. R., A. Mantalaris and S. P. Asprey (2004). "Modelling of Mammalian Cells and Cell Culture Processes." Cytotechnology **44**(1-2): 27-46.

Simon, L., M. N. Karim and A. Schreiweis (1998). "Prediction and classification of different phases in a fermentation using neural networks." Biotechnol Tech **12**(4): 301-304.

Sriyudthsak, K., F. Shiraishi and M. Y. Hirai (2013). "Identification of a Metabolic Reaction Network from Time-Series Data of Metabolite Concentrations." PLoS ONE **8**(1): e51212.

Steinhoff, R. F., M. Ivarsson, T. Habicher, T. K. Villiger, J. Boertz, J. Krismer, S. R. Fagerer, M. Soos, M. Morbidelli, M. Pabst and R. Zenobi (2014). "High-throughput nucleoside phosphate monitoring in mammalian cell fed-batch cultivation using quantitative matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." Biotechnol J.

Stelling, J., U. Sauer, F. J. Doyle and J. Doyle (2006). Complexity and Robustness of Cellular Systems. Syst Model Cell Biol, The MIT Press**:** 19-40.

Streif, S., A. Savchenko, P. Rumschinski, S. Borchers and R. Findeisen (2012). "ADMIT: a toolbox for guaranteed model invalidation, estimation and qualitative–quantitative modeling." Bioinformatics **28**(9): 1290-1291.

Suzuki, E. and D. F. Ollis (1990). "Enhanced Antibody Production at Slowed Growth Rates: Experimental Demonstration and a Simple Structured Model." Biotechnol Prog **6**(3): 231-236.

Szekely, G. J. and M. L. Rizzo (2005). "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method." J Classif **22**(2): 151-183.

Teixeira, A., C. Alves, P. Alves, M. Carrondo and R. Oliveira (2007). "Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control." BMC Bioinformatics **8**(1): 30.

Teixeira, A. P., R. Oliveira, P. M. Alves and M. J. T. Carrondo (2009). "Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative." Biotechnol Adv **27**(6): 726-732.

Terzer, M. and J. Stelling (2008). "Large-scale computation of elementary flux modes with bit pattern trees." Bioinformatics **24**(19): 2229-2235.

Tescione, L., J. Lambropoulos, M. R. Paranandi, H. Makagiansar and T. Ryll (2015). "Application of bioreactor design principles and multivariate analysis for development of cell culture scale down models." Biotechnol Bioeng **112**(1): 84-97.

Tomba, E., P. Facco, F. Bezzo and M. Barolo (2013). "Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review." Int J Pharm **457**(1): 283-297.

Toms, J. D. and M. L. Lesperance (2003). "Piecewise regression: A tool for identifying ecological thresholds." Ecology **84**(8): 2034-2041.

Tsang, V. L., A. X. Wang, H. Yusuf-Makagiansar and T. Ryll (2014). "Development of a scale down cell culture model using multivariate analysis as a qualification tool." Biotechnol Prog **30**(1): 152-160.

Tsoularis, A. and J. Wallace (2002). "Analysis of logistic growth models." Math Biosci **179**(1): 21-55.

Tsuchiya, H. M., A. G. Fredrickson and R. Aris (1966). Dynamics of Microbial Cell Populations. Adv Chem Eng. J. W. H. Thomas B. Drew and V. Theodore, Academic Press. **Volume 6:** 125-206.

Umaña, P. and J. E. Bailey (1997). "A mathematical model of N-linked glycoform biosynthesis." Biotechnol Bioeng **55**(6): 890-908.

van Can, H. J. L., H. A. B. te Braake, A. Bijman, C. Hellinga, K. C. A. M. Luyben and J. J. Heijnen (1999). "An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances: Part II." Biotechnol Bioeng **62**(6): 666-680.

Van Can, H. J. L., H. A. B. Te Braake, S. Dubbelman, C. Hellinga, K. C. A. M. Luyben and J. J. Heijnen (1998). "Understanding and applying the extrapolation properties of serial gray-box models." AIChE J **44**(5): 1071-1089.

Vande Wouwer, A., C. Renotte and P. Bogaerts (2004). "Biological reaction modeling using radial basis function networks." Comput Chem Eng **28**(11): 2157-2164.

Varma, A. and B. O. Palsson (1994). "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use." Nat Biotechnol **12**(10): 994-998.

Vester, D., E. Rapp, S. Kluge, Y. Genzel and U. Reichl (2010). "Virus–host cell interactions in vaccine production cell lines infected with different human influenza A virus variants: A proteomic approach." J Proteomics **73**(9): 1656-1669.

Villaverde, A. F., S. Bongard, K. Mauch, D. Müller, E. Balsa-Canto, J. Schmid and J. R. Banga (2015). "A consensus approach for estimating the predictive accuracy of dynamic models in biology." Comput Methods Programs Biomed **119**(1): 17-28.

Vogels, M., R. Zoeckler, D. Stasiw and L. Cerny (1975). "P. F. Verhulst's “notice sur la loi que la populations suit dans son accroissement” from correspondence mathematique et physique. Ghent, vol. X, 1838." J Biol Phys **3**(4): 183-192.

Voit, E. O. (2000). Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists, Cambridge University Press.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." J Clin Pharm Ther **27**(4): 299-309.

Wahrheit, J. (2015). Metabolic dynamics and compartmentation in the central metabolism of Chinese hamster ovary cells, Saarländische Universitäts- und Landesbibliothek.

Wahrheit, J. and E. Heinzle (2013). "Sampling and quenching of CHO suspension cells for the analysis of intracellular metabolites." BMC Proc **7**(6): 1-2.

Wahrheit, J., A. Nicolae and E. Heinzle (2014). "Dynamics of growth and metabolism controlled by glutamine availability in Chinese hamster ovary cells." Appl Microbiol Biotechnol **98**(4): 1771-1783.

Wahrheit, J., J. Niklas and E. Heinzle (2014). "Metabolic control at the cytosol–mitochondria interface in different growth phases of CHO cells." Metab Eng **23**(0): 9-21.

Wiechert, W. (2002). "Modeling and simulation: tools for metabolic engineering." J Biotechnol **94**(1): 37-63.

Xing, Z., N. Bishop, K. Leister and Z. J. Li (2010). "Modeling kinetics of a large-scale fed-batch CHO cell culture by Markov chain Monte Carlo method." Biotechnol Prog **26**(1): 208-219.

Xing, Z., Z. Li, V. Chow and S. S. Lee (2008). "Identifying Inhibitory Threshold Values of Repressing Metabolites in CHO Cell Culture Using Multivariate Analysis Methods." Biotechnol Prog **24**(3): 675-683.

Yang, A., E. Martin and J. Morris (2011). "Identification of semi-parametric hybrid process models." Comput Chem Eng **35**(1): 63-70.

Yu, M., Z. Hu, E. Pacis, N. Vijayasankaran, A. Shen and F. Li (2011). "Understanding the intracellular effect of enhanced nutrient feeding toward high titer antibody production process." Biotechnol Bioeng **108**(5): 1078-1088.

Zamorano, F., A. Vande Wouwer, R. M. Jungers and G. Bastin (2013). "Dynamic metabolic models of CHO cell cultures through minimal sets of elementary flux modes." J Biotechnol **164**(3): 409-422.

Zhu, J. (2012). "Mammalian cell protein expression for biopharmaceutical production." Biotechnol Adv **30**(5): 1158-1170.

6 Author Contributions

Bassem Ben Yahia was the contributor to the experimental design and executed the experiments, the results of which are presented in all chapters. He also developed the modeling methodologies, analyzed and interpreted the data of all chapters, drafted and wrote the manuscripts (**chapter 2** and **chapter 3**) and the mini-review (**chapter 1**) and wrote the R script in the Supplementary Material.

Boris Gourevitch was involved in the manuscript presented in **chapter 2**. He wrote the Matlab script in the Supplementary Material, provided help with the data analysis and critically revised the manuscript.

Meriam Annani and Valentine Chevallier performed the experiments presented in Table 3.1

Yasser Baddour performed the experiments presented in Table 3.3

Laetitia Malphettes and Elmar Heinzle supervised the thesis, provided help with the data analysis and the development of the modeling methodologies, critically revised and finalized the manuscripts.

7 Supplementary Material

SUPPLEMENTARY MISCELLANEOUS RESEARCH

Minimum number of dataset needed to calibrate the segmented linear model

During the development methodology presented in **chapter 3**, one of the question that was raised was the minimum number of data needed to calibrate the cell metabolism model.

The feasibility of such method with less number of datapoints has been assessed and the minimum number of dataset needed to calibrate this model has been identified but data were not presented in this thesis. The methodology used is described in Figure S1.

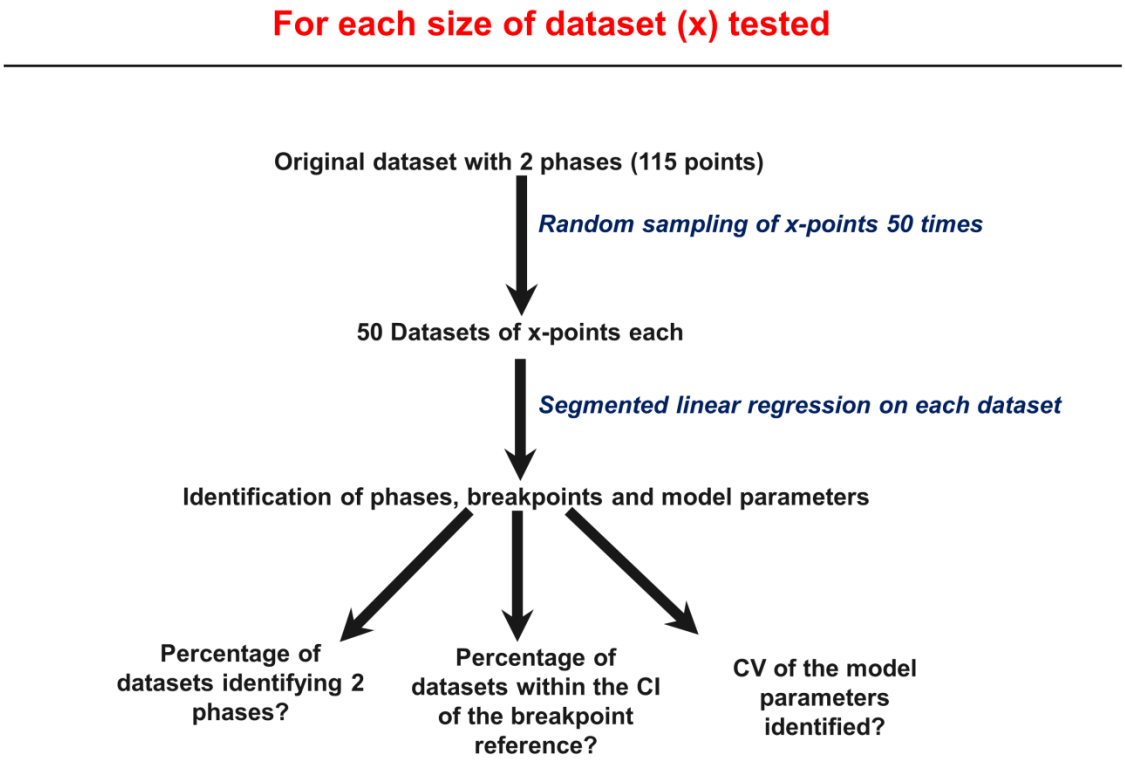


Figure S1 Methodology used to identify minimum number of data points necessary for accurate segmented linear modeling of cell metabolism. CI: confidence interval; CV: coefficient of variation.

The same calibration dataset used in **chapter 3** (Table 3.1) was used for a total of 115 datapoints. We focused on the specific production rate of glycine that has been randomly selected for the presented analysis from the set of metabolites that have shown more than one metabolic phase. From this dataset, subsets of 20, 30, 50, 80 and 100 datapoints were randomly extracted 50 times each ones. Then, for each subset, we analyzed the percentage of datasets identifying two metabolic phases, the percentage of datasets identifying breakpoints within the confidence interval (CI) of the breakpoint reference and the coefficient of variation (CV) of the model parameters identified. The CI of the breakpoint reference was defined from the initial dataset of 115 datapoints of the first breakpoint BP_1 and corresponds to the interval [0.507;0.567]. Results are presented in Figure S2. In our case, with the amino acid analytical method and measurement analysis equipment used, 80 data points were defined as sufficient in order to identify accurately metabolic phases and model parameters.

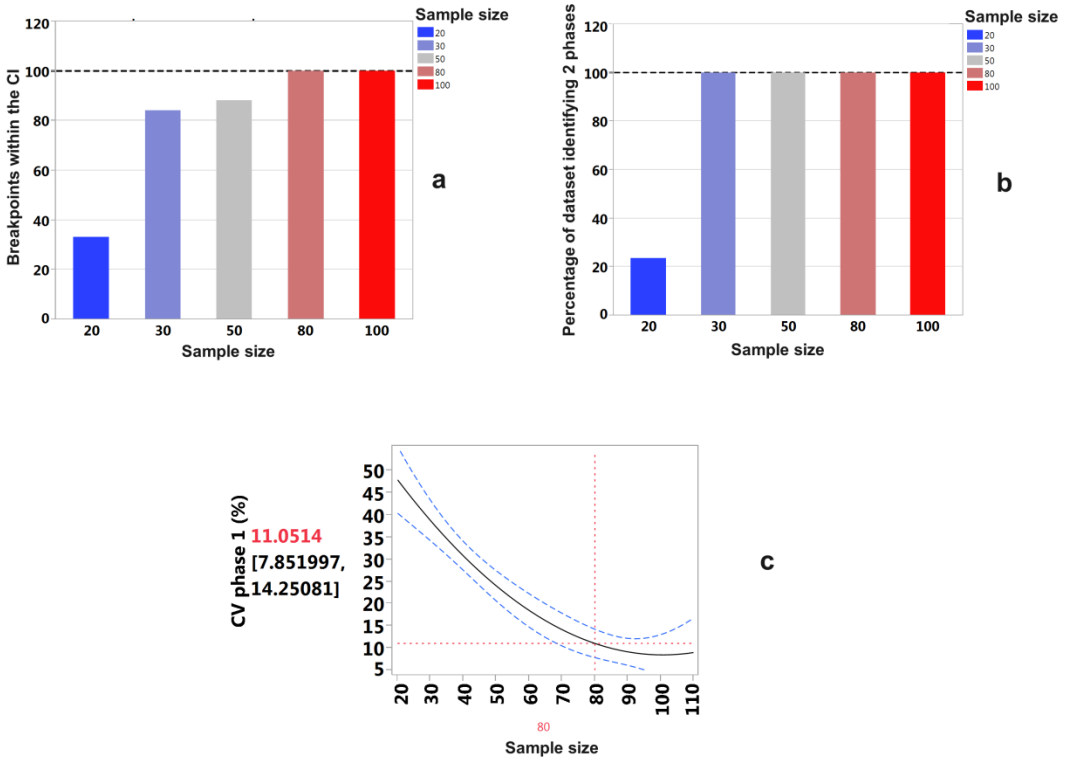


Figure S2 Accuracy of the segmented linear model with increasing dataset size. In a the percentage of breakpoints value identified for each dataset and within the confidence of interval of the initial reference dataset i.e. [0.507;0.567] are depicted. In b, the percentage of datasets identifying two metabolic phases is depicted. Two metabolic phases correspond to the maximum number of phases identified with glycine with the initial reference dataset. In c, the coefficient of variation (CV) of parameters in phase P1 (cf. Chapter 2) of the segmented linear model is depicted as function of the number of datapoints used.

This information can contribute to improve biopharmaceutical production as the number of datapoints used to calibrate the metabolic model developed in this thesis can be reduced. It can speed up model development with another cell clone, another cell culture process and increase the information that can be drawn from experimental data.

Product quality attributes macroscopic predictive model

In order to get the most complete i*n silico* model that can provide guidance on how to improve cell culture processes for industrial purpose, a prediction model of product quality attributes (PQA) can be developed in order to improve the modeling methodology presented in this thesis. For that purpose, a simple and empiric linear model of PQA was developed ad multivariate data analysis was used to identified the metabolites that are linked to PQA (data not shown). The linear model developed was used to predict new experimental condition (Figure S3). Results show promising results.

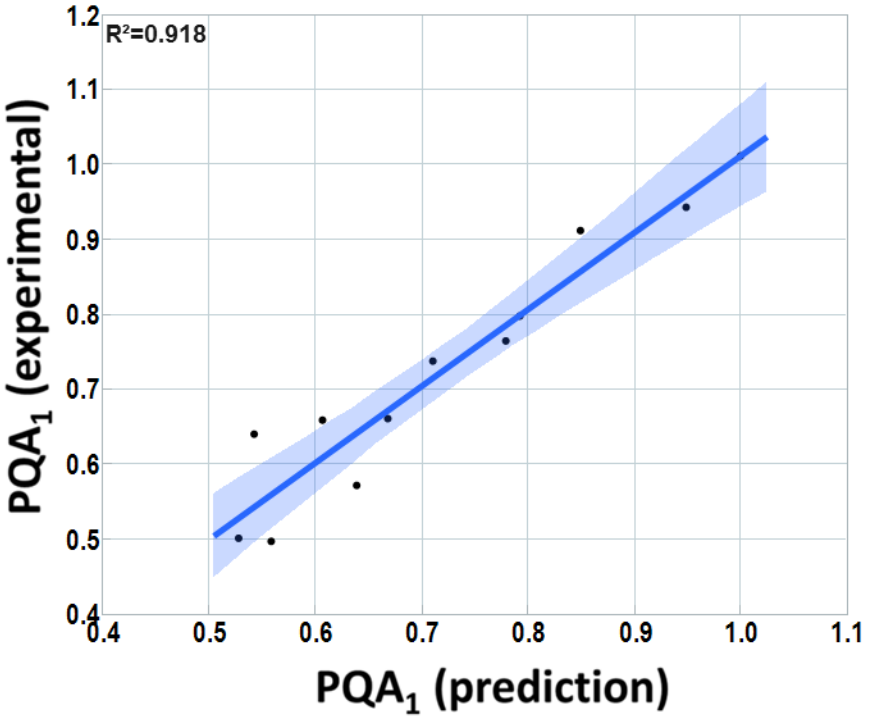


Figure S3 Comparison of one experimental product quality attribute 1 (PQA_1) with the model prediction. The data were normalized with the highest value reached. PQAs can refer to aggregates, charge variants, misincorporation, glycosylation profiles, drug color, deamination, amino acid oxidations, adducts, glycations and many more.

SUPPLEMENTARY FIGURES

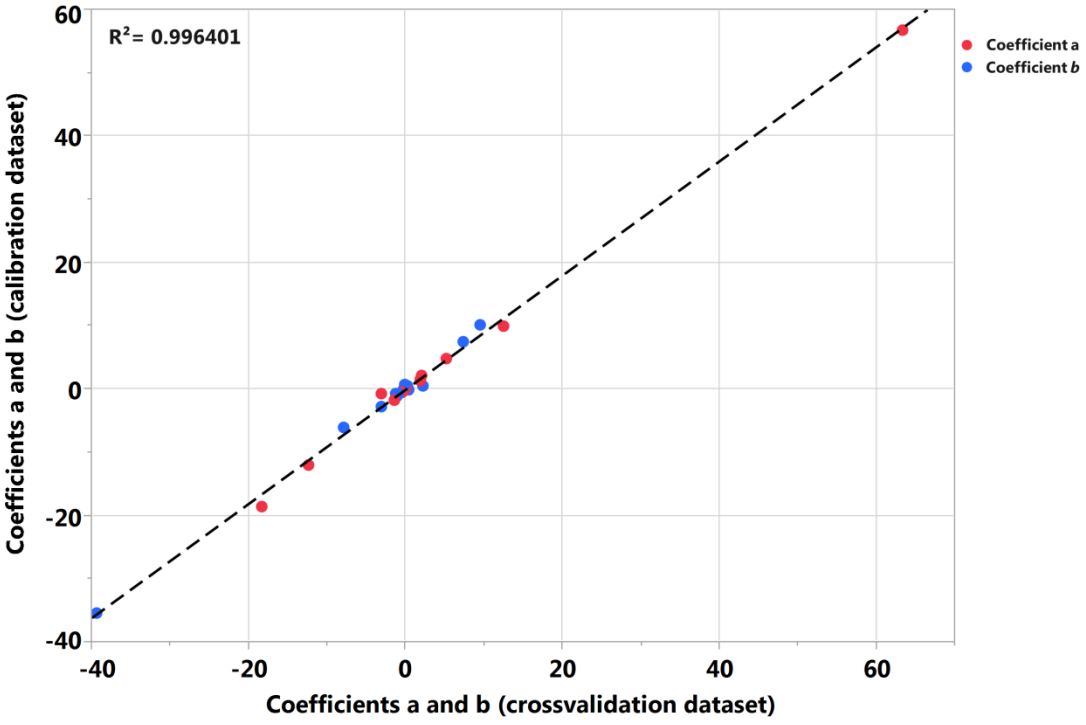


Figure S2.1. Comparison of segmented model coefficients. For each metabolite and for each metabolic phase, the values of coefficients a and b from Equation 5 were identified with the calibration dataset and the cross validation dataset.

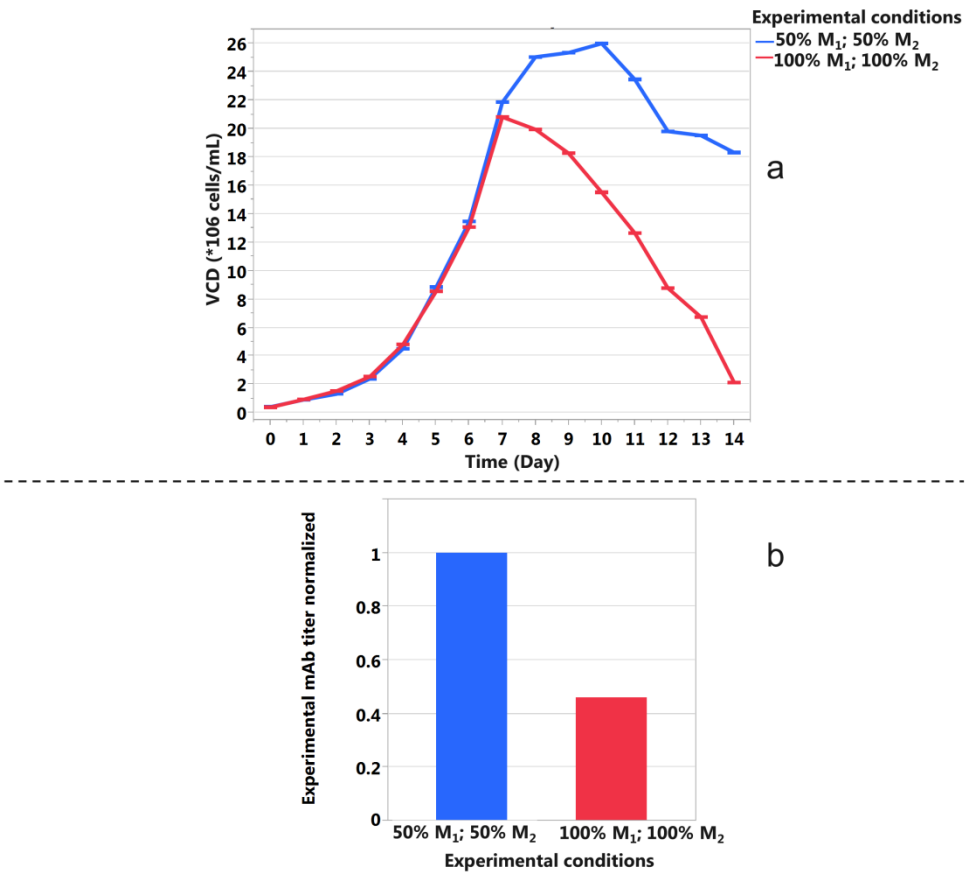


Figure. S3.1. Impact of metabolite M_1 and M_2 on cell growth and mAb titer. Two experimental conditions performed in 2 L bioreactors vessel are depicted. The concentration of two different metabolites (M_1, M_2) contained in the feed were varied. They are presented as percentage of the maximum concentration tested. **a** - viable cell density (VCD); **b** - mAb titer normalized to the highest titer reached.

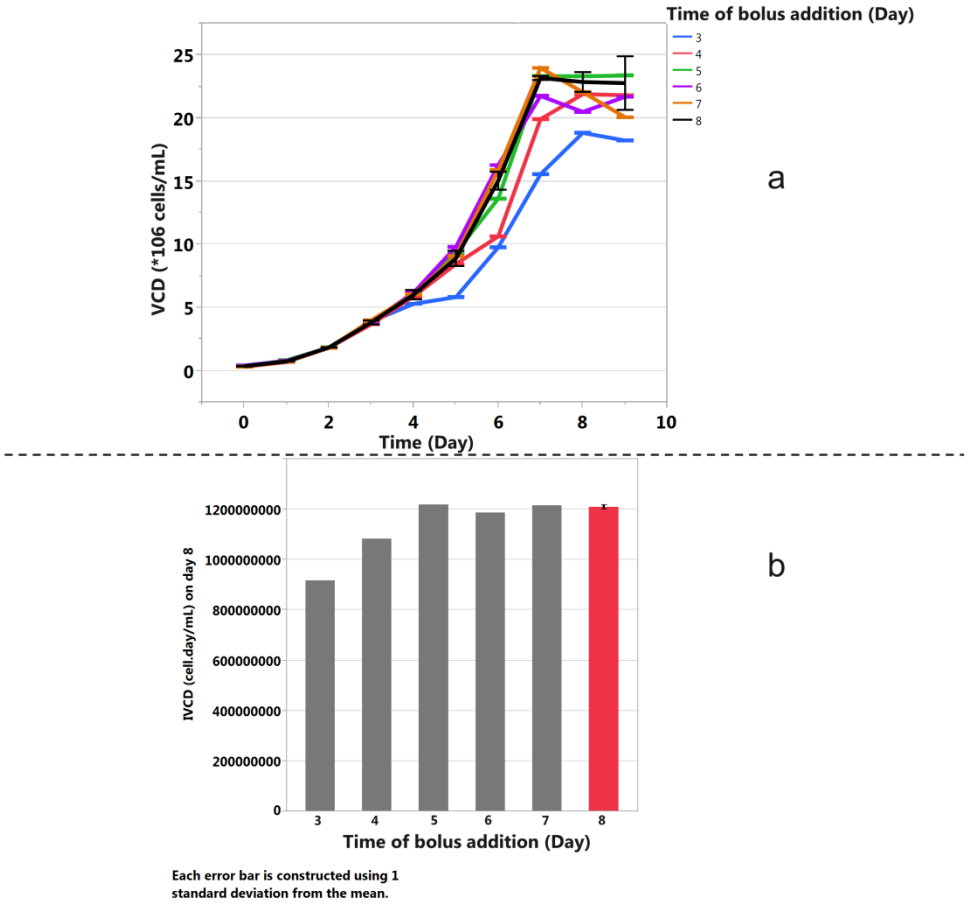


Figure S3.2. Impact of the day of bolus addition of inhibitory metabolite M_1 on cell growth. Fed-batch cultures were performed in 250 mL shake flasks for 8 days. In five different experiments, a bolus addition of metabolite M_1 was performed on day 3, 4, 5, 6 or 7. The concentration of M_1 in the cell culture medium after bolus addition was 0.8 g/L. Three experiments were made in a batch mode without addition of M_1 (black curve in part **a** and red bar in part **b** of the figure). **a** - viable cell density (VCD); **b** - integral viable cell density (IVCD) on day 8.

SUPPLEMENTARY TABLES

Supplementary Table S2.1. Experimental conditions. The concentration of three different amino acids (aa_1, aa_2, aa_3) contained in the feed that were varied. They are presented as percentage of the maximum concentration tested

*Control condition

Vessel	Experimental ID	Condition	aa1 (%)	aa2 (%)	aa3 (%)	Dataset
2 L	01	01	25	50	25	Cross validation
2 L	02		25	50	25	Cross validation
2 L	03		25	50	25	Cross validation
2 L	04	02	10	10	10	Calibration
2 L	05		10	10	10	Calibration
2 L	06	03	10	10	100	Calibration
2 L	07	04	10	50	50	Calibration
2 L	08		10	50	50	Calibration
2 L	09	05	10	100	10	Calibration
2 L	10	06	10	100	100	Calibration
2 L	11	07	50	10	50	Calibration
2 L	12		50	10	50	Calibration
2 L	13	08	50	50	10	Cross validation
2 L	14		50	50	10	Cross validation
2 L*	15	09	50	50	50	Cross validation
2 L*	16		50	50	50	Cross validation
2 L*	17		50	50	50	Cross validation
2 L*	18		50	50	50	Cross validation
2 L*	19		50	50	50	Cross validation
2 L*	20		50	50	50	Cross validation
2 L*	21		50	50	50	Cross validation
2 L*	22		50	50	50	Cross validation
2 L	23	10	100	10	10	Calibration
2 L	24	11	100	10	100	Calibration
2 L	25	12	100	50	50	Cross validation
2 L	26		100	50	50	Cross validation
2 L	27	13	100	100	10	Calibration
2 L	28	14	100	100	100	Calibration
2 L	29	15	12.5	12.5	12.5	Calibration
2000 L	30	01	25	50	25	Scale up
2000 L	31		25	50	25	Scale up
2000 L	32		25	50	25	Scale up

SUPPLEMENTARY SCRIPTS

MATLAB scripts (Ben_Yahia_et_al_segmented_regression.m and seg_reg.m)

The two MATLAB scripts appended allow the segmented regression as used in the publication.

It runs on Matlab Release 2013a (The Mathworks, Natick, MA, USA) and higher versions and requires the Statistical Toolbox supplied by The Mathworks.

Ben_Yahia_et_al_segmented_regression.m

```
%Script of segmented regression:
%this script generates simulated data of seven metabolites production
rates as a function of one growth rate variable:
%three metabolites with 2 breakpoints,
%two metabolites with 1 breakpoint, one metabolite with no breakpoint
and
%a last one completely random
%The simulated data is stored in an xls file which shows the typical
%structure of data used by the computations performed here. This xls
file
%can be used as a template for your data (in that case clear the first
cell of the script then run the script)
%The script loads data from the xls file then computes breakpoints and
%segmented regression models using the subroutine seg_reg
%Results are then displayed
%% Simulate data and write it into an xls file
Metab={'Abc','Def','Ghi','Klm','Nop','Qrs','Tuv'};%variable names
nData=3*100;%n points to simulate

Dataset=rand(nData,length(Metab))-0.5;
range_data=[-0.6 0.8];
gr=diff(range_data)*rand(nData,1)+range_data(1);

%Data with two breakpoints Dataset(:,1:3)
bp1=[-0.28 -0.25 -0.22];
bp2=[0.6 0.63 0.66];
slopes=[-5 1 5;-5 -1 5;5 -5 5];
cst0=[5 5 -5];

Dataset(:,1:3)=Dataset(:,1:3)+repmat(cst0,nData,1);
```

```
bounds=[repmat(range_data(1),3,1) bp1' bp2' repmat(range_data(2),3,1)];
slopes_temp=[slopes(:,1) diff(slopes,[],2)];
for i=1:length(bp1),
    for bo=1:3,
        index=gr>=bounds(i,bo);
        Dataset(index,i)=Dataset(index,i)+(gr(index)-
bounds(i,bo))*slopes_temp(i,bo);
    end;
end;



%Data with one breakpoint Dataset(:,4:5)
bp1=[0.6 0.63];
slopes=[1 10;5 -10];
cst0=[1 1];

Dataset(:,4:5)=Dataset(:,4:5)+repmat(cst0,nData,1);
bounds=[repmat(range_data(1),2,1) bp1' repmat(range_data(2),2,1)];
slopes_temp=[slopes(:,1) diff(slopes,[],2)];
for i=(1:length(bp1)),
    for bo=1:2,
        index=gr>=bounds(i,bo);
        Dataset(index,i+3)=Dataset(index,i+3)+(gr(index)-
bounds(i,bo))*slopes_temp(i,bo);
    end;
end;

%Linear data with no breakpoint Dataset(:,6)
slopes=5;
cst0=1;
Dataset(:,6)=Dataset(:,6)+gr*slopes+cst0;

%Random data is Dataset(:,7)

%write to excel table (you can fill this table with your own data with
the same formatting)
xlswrite('DATASIM_SEG.xls',{Metab{:},'gr'},'A1:H1');
xlswrite('DATASIM_SEG.xls',[Dataset gr],'A2:H301');
```

```
%% Load data
clear
[Dataset,Metab]=xlsread('DATASIM_SEG.xls');
gr=Dataset(:,end); % Growth rate variable
Dataset(:,end)=[];
nvar=size(Dataset,2);
for i=1:nvar,
    Metab_mini{i}=Metab{i}(1:3);%shorter variable names
end;
n=length(gr);%size of data

%% Computations of segmented regression

seg=seg_reg(Dataset,gr);
%seg is a structure of size nvar which contains all information about
%segmented regression

%% Display data and best model found for each metabolite
figure('position',[240          49        1000         500]);
for i=1:nvar,
    subplot(2,ceil(nvar/2),i);
    hold on

    plot(gr,Dataset(:,i),'.','color',[0 0 1]);%scattergram data

    %plot best model
    grsort=sort(gr);%trick: sort data and connect points with lines to
show segemnted regression, easier than plotting three lines
    %identify which was the best model, choose a color for each type of
model found
    if seg(i).model_choice==2,%2BP is the best model
        %build regression matrix
        mat_reg=[ones(n,1)                                              grsort
double(grsort>=seg(i).values_2bp(1)).*(grsort-seg(i).values_2bp(1))
double(grsort>=seg(i).values_2bp(2)).*(grsort-seg(i).values_2bp(2))];
        col='r';
        suff=' - 2BP';
    elseif seg(i).model_choice==1,%1BP is the best model
        mat_reg=[ones(n,1)                                              grsort
double(grsort>=seg(i).values_1bp(1)).*(grsort-seg(i).values_1bp(1))];
        col='g';
        suff=' - 1BP';
    elseif seg(i).model_choice==0,%LIN is the best model
```

```
         mat_reg=[ones(n,1) grsort];
        col='c';
        suff=' - LIN';
    else%no model is good enough (Fisher test)
        col='k';
        suff=' - NO';
    end;
    %Is the segmented regression model found a good one ? Condition
R2>0.5,
    %indicated by a "*" in the title
    if seg(i).r2>0.5&~isnan(seg(i).model_choice),
        good_model='*';
    else
        good_model='';
    end;
    %if there was a regression model, show the lines
    if ~isnan(seg(i).model_choice),
        ypred=mat_reg*seg(i).beta;
        plot(grsort,ypred,col,'linewidth',2);
    end;
    title([Metab_mini{i} suff good_model],'color',col);

    if i==1,
        ylabel('Production Rate');
        xlabel('Growth Rate');
    end;
    grid on
end;

%% set of BP found by best 2BP segmented regression
%(even if that was not the final selected model)
disp('BP values for the best 2BP segmented regression model');
for i=1:nvar,
    fprintf('%s \t %0.2f \t %0.2f \n',Metab_mini{i},seg(i).values_2bp)
end;
```

seg_reg.m

```
function [seg,grsteps]=seg_reg(Dataset,gr)
%function seg=seg_reg(Dataset,gr)
```

```
%computes segmented regression model up to 2 breakpoints
%
%INPUT:
%     Dataset       : matrix observations * variables
%     gr            : variable to segment (growth rate in the paper)
%
%OUTPUT:
%     seg is a structure with the followig fields:
%     fstat_lin     : fisher stats for the linear model
%     values_2bp    : Breakpoints (BP) values of the 2 BP model
%     fstat_2bp     : fisher stats for the 2 BP model
%     values_1bp    : BP value of the 1 BP model
%     fstat_1bp     : fisher stats for the 1 BP model
%     fish2_1       : fisher stat value for the 2 BP vs 1 BP model
%     pvfish2_1     : fisher stat pvalue for the 2 BP vs 1 BP model
%     fish1_0       : fisher stat value for the 1 BP vs 0 BP model
%     pvfish1_0     : fisher stat pvalue for the 1 BP vs 0 BP model
%     model_choice  : number of BP in the best model
%     beta          : coefficient values in the best model
%     slopes        : slope values in the best model
%     r2            : r2 of the best model
%     ssr           : Residual Sum of Square of the best model
%     n             : number of observations in the data
%     statsreglin   : stats for the linear model (no BP)
%     statsregseg   : stats for all tested segmented models

%maximum number of breakpoints = 2 (Based on previous data analysis)
nvar=size(Dataset,2);

%create a set of possible breakpoints
rangebp=prctile(gr,[5 95]);%growth rate range
number_steps=50;%number of possible values for breakpoints
grsteps=linspace(rangebp(1),rangebp(2),number_steps);%sample         50
breakpoints over the range of growth rate
pt_middle=median(gr);

%loop on all available variables
for i=1:nvar,

    idx=~isnan(Dataset(:,i));%select non-missing data
    n=sum(idx);

    %estimate classic linear regression (0BP)
```

```
    statsreglin(i)                                                    =
regstats(Dataset(idx,i),gr(idx),'linear',{'beta','covb','tstat','rsqua
re','r','fstat'});
    SSR_ref(i)=sum(statsreglin(i).r.^2);%reference sum of square
    R2_ref(i)=statsreglin(i).rsquare;%reference R^2 statistics
    seg(i).fstat_lin=statsreglin(i).fstat;


    %we now compute the regression model for each pair of possible
breakpoints
    c1=0;
    for bp1=grsteps,
        c1=c1+1;
        c2=0;
        for bp2=grsteps,
            c2=c2+1;
            ok=0;
            if  (bp1<pt_middle&&bp2>pt_middle&&bp2-bp1>0.2),%  We  put
each breakpoint on each side of growth rate mediane. Morever, the minimum
distance between two breakpoints has been set to 0.2 day^-1 except if
breakpoints are the same (case of one breakpoint)
                mat_reg=[gr                    double(gr>=bp1).*(gr-bp1)
double(gr>=bp2).*(gr-bp2)];%regression matrix
                ok=1;
            elseif (bp1==bp2)
                 mat_reg=[gr     double(gr>=bp1).*(gr-bp1)];%regression
matrix
                 ok=1;
            end;
            if ok,
                statsregseg(i,c1,c2)                                  =
regstats(Dataset(idx,i),mat_reg(idx,:),'linear',{'beta','rsquare','r',
'fstat'});
                ssrbp{i}(c1,c2)=sum(statsregseg(i,c1,c2).r.^2);%sum  of
residual square
            end;
        end;
    end;


    %which bp are the best ? (case of two breakpoints)
    ssrbp{i}(ssrbp{i}==0)=NaN;
    [mintemp,c1min]=nanmin(ssrbp{i});
    [~,c2min]=nanmin(mintemp);
    c1min=c1min(c2min(1));
    seg(i).values_2bp=grsteps([c1min  c2min]);%store  BP  values  in  a
structure seg
```

```
    seg(i).fstat_2bp=statsregseg(i,c1min,c2min).fstat;%store     fisher
stats for the whole model

    %case with only one BP
    [~,c1min1bp]=nanmin(diag(ssrbp{i}));
    seg(i).values_1bp=grsteps(c1min1bp);
    seg(i).fstat_1bp=statsregseg(i,c1min1bp,c1min1bp).fstat;

    %comparison two BP vs one BP: Fisher statistics and pvalue
    seg(i).fish2_1=(ssrbp{i}(c1min,c2min)-
ssrbp{i}(c1min1bp,c1min1bp))/(n-3-1-(n-2-
1))/ssrbp{i}(c1min1bp,c1min1bp)*(n-2-1);%same bp so it is two segments
instead of three
    seg(i).pvfish2_1=1-fcdf(seg(i).fish2_1,abs(n-3-1-(n-2-1)),n-2-1);

    %comparison of one BP with simple linear regression
    seg(i).fish1_0=(ssrbp{i}(c1min1bp,c1min1bp)-SSR_ref(i))/(n-2-1-(n-
2))/SSR_ref(i)*(n-2);
    seg(i).pvfish1_0=1-fcdf(seg(i).fish1_0,abs((n-2-1-(n-2))),(n-2));

    %final segmented model - store results
    if
seg(i).pvfish2_1<0.05/nvar&&seg(i).fstat_2bp.pval<0.05/nvar,%best model
is 2BP. threshold is "bonferronized"
        seg(i).model_choice=2;%number of breakpoints
        seg(i).beta=statsregseg(i,c1min,c2min).beta;

seg(i).slopes=[cumsum(statsregseg(i,c1min,c2min).beta(2:end))'];%slope
s in a segemented regression
        seg(i).r2=statsregseg(i,c1min,c2min).rsquare;
        seg(i).ssr=ssrbp{i}(c1min,c2min);%residuals sum of square
    elseif
seg(i).pvfish1_0<0.05/nvar&&seg(i).fstat_1bp.pval<0.05/nvar,%best model
is 1BP
        seg(i).model_choice=1;
        seg(i).beta=statsregseg(i,c1min1bp,c1min1bp).beta;

seg(i).slopes=[cumsum(statsregseg(i,c1min1bp,c1min1bp).beta(2:end))'];
%slopes in a segemented regression
        seg(i).r2=statsregseg(i,c1min1bp,c1min1bp).rsquare;
        seg(i).ssr=ssrbp{i}(c1min1bp,c1min1bp);%residuals sum of square
    elseif seg(i).fstat_lin.pval<0.05/nvar,%best model is Linear (0BP)
        seg(i).model_choice=0;
        seg(i).beta=statsreglin(i).beta;
```

```
        seg(i).slopes=seg(i).beta(end);%slopes    in    a    segemented
regression
        seg(i).r2=R2_ref(i);
        seg(i).ssr=SSR_ref(i);%residuals sum of square
    else
        seg(i).model_choice=NaN;
        seg(i).beta=NaN;
        seg(i).slopes=NaN;%slopes in a segemented regression
        seg(i).r2=NaN;
        seg(i).ssr=NaN;%residuals sum of square
    end;
    seg(i).n=n;
    seg(i).statsreglin=statsreglin(i);
    seg(i).statsregseg=squeeze(statsregseg(i,:,:));
end;
disp('Computations done')
```

R scripts

The R scripts appended allow the segmented regression in R with Shiny, an open source R package that provides a web application.

It runs on R 3.2.2 (Copyright (c) 2015 The R Foundation for Statistical Computing) or RStudio v1.0 (Copyright (c) 2015 The R Foundation for Statistical Computing) and higher versions. The data should be structured like for the MATLAB script but the xls file should be converted into csv file.

```
library(shiny)
library(segmented)
library(data.table)
library(ggplot2)
#script to perform a segmented linear regression with Shiny(R)
ui <- fluidPage(
  titlePanel("Segmented linear regression"),
  sidebarLayout(
    #define the inputs and the outputs
    sidebarPanel(
      fileInput('file1', 'Choose CSV
File',accept=c('text/csv','text/semicolon-separated-
values,text/plain','.csv')),
      hr(),
      uiOutput("varselect"),
      uiOutput("varselect2"),
      uiOutput("breakpoint"),
      fluidRow(
        textOutput("txt"),
        textOutput("text"),
        tableOutput("tabest"),
        h3(textOutput("Information")),
        tableOutput("Breakpoint")
      )),
    mainPanel(
      fluidRow(
        p(strong("(2016)"), ("Methodologie developped by"),em(" Bassem Ben
Yahia et al."),("bassem.benyahia@ucb.com")),
        titlePanel("Plots"),
        column(6,plotOutput("hist")),
        column(6,h5(plotOutput("res")))
      ),
      fluidRow(
        column(6,tableOutput("tabrquared")),
        column(6,tableOutput("tab"))
      ))))

server <- function(input, output) {
  Dataset<-reactive({
```

```
    infile<-input$file1
    #if no dataset loaded, return NULL
    if (is.null(infile)){
      return(NULL)
    }
    #if datased available, read it and save it in Dataset
    read.csv(infile$datapath,";",h=T)
  })
  #Give a value to all output (if Dataset available)
  output$varselect<-renderUI({
    infile<-input$file1
    if (is.null(infile)){
      #No dataset available
      h4("No variable: please choose a CSV file",style="color:red")
    }else{
      #Dataset available: define the variables presents (variable2 and
variable)
      #then define the new input (selectinput and breakpoints selection)
      cols<-names(Dataset())
      output$varselect2<-renderUI({selectInput("variable2", "X
variable:",choices=cols)})
      output$breakpoint<-
renderUI({numericInput("bp","Breakpoints",0,min=0,max=3,step=1)})
      selectInput("variable", "Y variable:",choices=cols)
    }
  })
  #Plot and perform the statistical analysis
  output$hist<-renderPlot({
    infile<-input$file1
    if (is.null(infile)){
      points(NULL)
    }else{
      Dati<-Dataset()
      x<-as.numeric(unlist(Dati[input$variable2]))
      y<-as.numeric(unlist(Dati[input$variable]))
      variable<-input$variable
      variable2<-input$variable2
      if(input$bp==0){

plot(x,y,xlab=variable2,ylab=variable,pch=17,col="black");abline(lm(y~x),lwd=
5)
        output$tab<-renderTable({anova(lm(y~x))})
        output$res<-
renderPlot({plot(x,residuals(lm(y~x)),xlab=input$variable2,ylab="redisuals",p
ch=18);abline(0,0,lwd=5,lty=3,col="blue")})
        h3(output$Information<-renderText({"Summary"}))
        r.table<-
data.frame(data.frame(summary(lm(y~x))$adj.r.squared),data.frame(summary(lm(y
~x))$r.squared))
        colnames(r.table)<-c("R.squared","Adj.R.squared")
        output$tabrquared<-renderTable({r.table})
        emptytab<-data.frame("")
        colnames(emptytab)<-""
        rownames(emptytab)<-""
        output$tabest<-renderTable({emptytab})
      }else{
```

```
        if (input$bp==1){
          o.seg<-
segmented(lm(y~x),seg.Z=~x,control=seg.control(display=FALSE,it.max = 60))

plot(o.seg,conf.level=0.95,shade=TRUE,xlab=variable2,ylab=variable,col="green
",lwd=5);points(x,y,pch=17)
          output$tab<-renderTable({anova(o.seg,lm(y~x))})
          output$res<-
renderPlot({plot(x,residuals(o.seg),xlab=input$variable2,ylab="redisuals",pch
=18);abline(0,0,lwd=5,lty=3,col="blue")})
          output$Information<-renderText({"Breakpoint"})
          breakpointtable<-data.frame(summary(o.seg)$psi)
          rownames(breakpointtable)<-"Breakpoint"
          output$Breakpoint<-renderTable({breakpointtable})
          t.slope<-data.frame(slope(o.seg))
          interc<-data.frame(intercept(o.seg))
          t.slope<-t.slope[1]
          colnames(t.slope)<-"Est."
          n.table<-rbind(t.slope,interc)
          row.names(n.table)<-c("b2","b1","a2","a1")
          output$tabest<-renderTable({n.table},rownames=TRUE)
          r.table<-
data.frame(data.frame(summary(o.seg)$adj.r.squared),data.frame(summary(o.seg)
$r.squared))
          colnames(r.table)<-c("R.squared","Adj.R.squared")
          output$tabrquared<-renderTable({r.table})

        }else{
          if (input$bp==2){
            o.seg<-
segmented(lm(y~x),seg.Z=~x,psi=c(qnorm(0.5,mean(x),sd(x)),qnorm(0.85,mean(x),
sd(x))),control=seg.control(display=FALSE,it.max = 60))

plot(o.seg,conf.level=0.95,shade=TRUE,xlab=variable2,ylab=variable,col="red",
lwd=5);points(x,y,pch=17)
            output$tab<-
renderTable({anova(o.seg,segmented((lm(y~x)),seg.Z=~x))})
            output$res<-
renderPlot({plot(x,residuals(o.seg),xlab=input$variable2,ylab="redisuals",pch
=18);abline(0,0,lwd=5,lty=3,col="blue")})
            output$Information<-renderText({"Breakpoints"})
            breakpointtable<-data.frame(summary(o.seg)$psi)
            rownames(breakpointtable)<-c("Breakpoint 1","Breakpoint 2")
            output$Breakpoint<-renderTable({breakpointtable})
            t.slope<-data.frame(slope(o.seg))
            interc<-data.frame(intercept(o.seg))
            t.slope<-t.slope[1]
            colnames(t.slope)<-"Est."
            n.table<-rbind(t.slope,interc)
            row.names(n.table)<-c("b3","b2","b1","a3","a2","a1")
            output$tabest<-renderTable({n.table},rownames = TRUE)
            r.table<-
data.frame(data.frame(summary(o.seg)$adj.r.squared),data.frame(summary(o.seg)
$r.squared))
            colnames(r.table)<-c("R.squared","Adj.R.squared")
            output$tabrquared<-renderTable({r.table})
```

```
          }else{
            o.seg<-
segmented(lm(y~x),seg.Z=~x,psi=c(qnorm(0.5,mean(x),sd(x)),qnorm(0.85,mean(x),
sd(x)),qnorm(0.85,mean(x),sd(x))/2),control=seg.control(display=FALSE,it.max
= 60))

plot(o.seg,conf.level=0.95,shade=TRUE,xlab=variable2,ylab=variable,col="viole
t",lwd=5);points(x,y,pch=17)
            output$tab<-
renderTable({anova(o.seg,segmented(lm(y~x),seg.Z=~x,psi=c(qnorm(0.5,mean(x),s
d(x)),qnorm(0.85,mean(x),sd(x)))))})
            output$res<-
renderPlot({plot(x,residuals(o.seg),xlab=input$variable2,ylab="redisuals",pch
=18);abline(0,0,lwd=5,lty=3,col="blue")})
            output$Information<-renderText({"Breakpoints"})
            breakpointtable<-data.frame(summary(o.seg)$psi)
            rownames(breakpointtable)<-c("Breakpoint 1","Breakpoint 2","
Breakpoint 3")
            output$Breakpoint<-renderTable({breakpointtable})
            t.slope<-data.frame(slope(o.seg))
            interc<-data.frame(intercept(o.seg))
            t.slope<-t.slope[1]
            colnames(t.slope)<-"Est."
            n.table<-rbind(t.slope,interc)
            row.names(n.table)<-c("b4","b3","b2","b1","a4","a3","a2","a1")
            output$tabest<-renderTable({n.table},rownames=TRUE)
            r.table<-
data.frame(data.frame(summary(o.seg)$adj.r.squared),data.frame(summary(o.seg)
$r.squared))
            colnames(r.table)<-c("R.squared","Adj.R.squared")
            output$tabrquared<-renderTable({r.table})
          }
        }}}
  })
}
shinyApp(ui = ui, server = server)
```

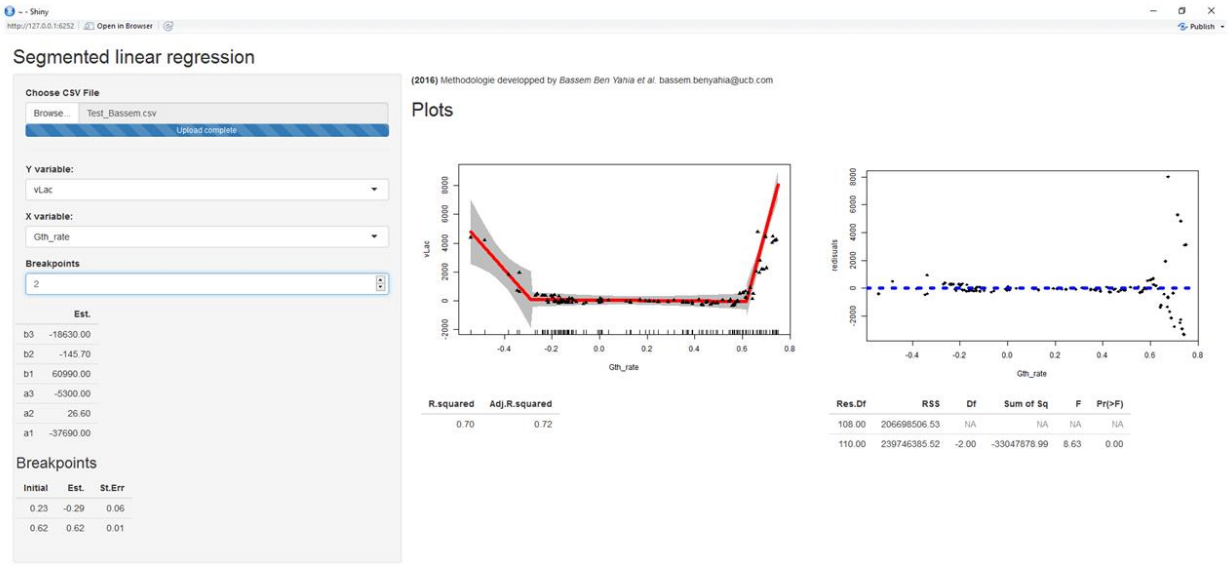


Figure S4 Screenshot of an example of application of the R script