ADVANCED METHODOLOGIES FOR THE MODELING OF METABOLIC PATHWAY SYSTEMS BASED ON TIME SERIES DATA

A Thesis
Presented to
The Academic Faculty

by

Sepideh Dolatshahi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2015

ADVANCED METHODOLOGIES FOR THE MODELING OF METABOLIC PATHWAY SYSTEMS BASED ON TIME SERIES DATA

Approved by:

Dr. Eberhard O. Voit
Wallace H. Coulter Dept. of Biomedical Engineering
Georgia Institute of Technology

Dr. Robert J. Butera, Co-Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Melissa L. Kemp
Wallace H. Coulter Dept. of Biomedical Engineering
Georgia Institute of Technology

Dr. Jeff S. Shamma
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Erik I. Verriest
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Brani Vidakovic
School of Systems and Industrial Engineering and Wallace H. Coulter Dept. of Biomedical Engineering
Georgia Institute of Technology

Date Approved: 14 July 2015

To mom and dad

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my mentor, Dr. Eberhard Voit. Dr. Voit introduced me to the field of systems biology and biochemical systems theory. He steadily taught me the most important scientific skill: the cognitive process in the form of conceptual or formal models to extract a storyline, which addresses real-world questions, from messy snapshots of data that illustrate select aspects of a biological phenomenon of interest. Dr. Voit's contribution to my thesis has been fundamental; from the central scientific ideas to little comments that made my writing and presentation clearer. Dr. Voit is a genuine role model, and I will leave him with much yet to learn.

Our collaboration with Dr. Brani Vidakovic has been very fruitful and led to developing the constrained iterative wavelet smoother for the task of pathway identification in systems biology. I still have records of our colorful pen-on-paper discussions. He has also been a great biostatistics teacher.

Many thanks to my biochemistry mentor and co-author on the *Lactococcus* project, Dr. Luis Fonseca. He has been continuously supportive and helpful beyond co-authorship and led me to the state of "biological literacy".

Lots of thanks to all my lab-mates at the Laboratory for Biological Systems Analysis. It is wonderful to belong to a group with such great academic integrity, cooperation, and social spirit. Thank you Po-Wei Chen, Ann Dam, Mojdeh Faraji, Luis Fonseca, Zhen Qi, and James Wade.

Special thanks to my committee members Dr. Robert Butera, Dr. Erik Verriest, Dr. Melissa Kemp, Dr. Jeff Shamma, and Dr. Brani Vidakovic for their support and constructive feedback along the way.

I am also grateful to Dr. Helena Santos and Dr. Ana Rute Neves of the Instituto de Tecnologia Química e Biológica (ITQB) at the New University of Lisbon (Portugal) for allowing me to use the *in vivo* NMR data presented here. The NMR spectrometers are part of The National NMR Facility, supported by Fundação para a Ciência e a Tecnologia (RECI/BBB-BQB/0230/2012). This work was funded in part by NSF grant MCB-0946595 (PI: EOV). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reect the views of the sponsoring institutions.

I have spent a total of six terrific years as a student at Georgia Institute of Technology and this has been a significant chapter in my life. I have had many great teachers and supporting friends who taught me the rigors of problem solving and stood by me when I needed help and emotional support.

Thank you mom and dad for always being infinitely and unconditionally loving and supportive. Thank you my precious sisters, Mehranoosh and Sussan, and my only brother, Behrang. I dedicate my thesis to my beautiful family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Metabolic pathways are series of enzyme-catalyzed chemical reactions that take place within a cell. These biochemical pathways can be quite elaborate and highly regulated with numerous positive or negative feedback or feed-forward mechanisms, which produce complex dynamical behaviors. Time series data have been more readily available in recent years as a result of the development of new measurement techniques. These techniques offer novel options for inferring the intricate regulatory structure of the metabolic pathways, analyzing the design and function of biological modules, as well as making predictions based on this analysis. The first objective of the proposed research is to advance mathematical methodologies for the study of metabolic and signaling pathways where time series data are available. The second objective is the application of these methodological advances toward a deeper understanding of the glycolytic pathway in the dairy bacterium Lactococcus lactis.

CHAPTER I

INTRODUCTION AND OBJECTIVES

1.1 Objective

Metabolic pathways are series of enzyme-catalyzed chemical reactions that take place within a cell. These biochemical pathways can be quite elaborate and highly regulated with numerous positive or negative feedback or feed-forward mechanisms, which produce complex dynamical behaviors. Time series data have been more readily available in recent years as a result of the development of new measurement techniques. These techniques offer novel options for inferring the intricate regulatory structure of the metabolic pathways, analyzing the design and function of biological modules, as well as making predictions based on this analysis. The first objective of the proposed research is to advance mathematical methodologies for the study of metabolic and signaling pathways where time series data are available. The second objective is the application of these methodological advances toward a deeper understanding of the glycolytic pathway in the dairy bacterium *Lactococcus lactis*.

1.2 Motivation and Significance of the Investigated Organism

Lactococcus lactis is a relatively simple lactic acid bacterium that has been serving as a model organism in molecular and biochemical studies for a long time. The rather simple carbohydrate metabolism of *L. lactis*, which is employed primarily for energy generation, makes it an attractive candidate to test metabolic engineering strategies. The genome of *L. lactis* has been fully sequenced, and a large number

of physiological, enzymatic, proteomic, transcriptomic, and microarray studies have been performed (*e.g.*, [5, 4, 38, 39]) leading to the availability of a large number of genetic tools, including food-grade cloning vectors and selection markers. Of special importance here, the organism has been serving as the primary test bed for novel non-invasive *in vivo* NMR techniques that generate time-series measurements of a variety of metabolites under many different conditions. Even though *L. lactis* has been the subject of intense research, a computational framework to integrate and interpret this information is missing.

L. lactis is of considerable practical significance in the food industry, being a crucial player in dairy fermentations and in the production of different cheeses, yoghurts, buttermilk, and other products. This organism metabolizes glucose homofermentatively to lactate, thereby providing an effective way of preserving the fermented products. In addition, traces of secondary metabolites can contribute to the flavor, texture and sometimes the nutritional value of the dairy products. The bacterium has also been used for the production pickled vegetables, bread, and beer and wine. Recently it has even been suggested for biofuel production [57]. An entirely different future application may be in therapeutics, if the organism can be genetically manipulated to survive low pH levels in the stomach, which would permit its potential use as a novel vehicle for the non-invasive delivery of vaccines and therapeutic proteins [65]. While the bacterium is a facultative aerobe, it typically lives in anaerobic conditions. Its preferred substrate is glucose, which is naturally available or unavailable in erratic time intervals. This uncertainty mandates that the organism is able to stop and start glucose uptake and glycolysis very quickly.

Genetic engineering of *L. lactis* strains cannot be achieved without a deep understanding of the metabolic pathways and the interdependent relationships among the different processes. A better understanding of cellular control and regulation can be achieved by developing kinetic-dynamic models that are to be based on reliable

time series concentrations of metabolite pools along the metabolic pathways. For this purpose, *in vivo* NMR is a technique of choice. NMR can provide unique information on intracellular metabolites in a non-invasive way. For this study, *in vivo* NMR data were provided by our collaborators , Drs. Helena Santos and Ana Rute Neves of the Instituto de Tecnologia Qumica e Biolgica (ITQB) at the New University of Lisbon (Portugal), who pioneered this type of experimentation.

1.3 Thesis Outline

The first goal of this thesis is the creation of methods to extract functional information from time series measurements of metabolite concentrations. The second goal is the application of these methods toward the analysis of the glycolytic pathway in the bacterium *Lactococcus lactis*. This information is to be integrated into functional models and subsequently utilized for explanations, predictions, manipulations, and the optimization of this pathway system.

In order to achieve the outlined goals, the results of the method development efforts as well as the biological insights are organized into the following Chapters:

Chapter 2 describes the tasks of model design, parameter estimation, and diagnostics for the glycolytic pathway in *L. lactis*. **Chapter 3** utilizes the model developed in Chapter 2, as well as novel computational techniques, to gain new insights into the complex regulation of the glycolytic pathway in *L. lactis*. These two chapters, together, offer a very detailed investigation and characterization of the control and pathway regulation of central carbon metabolism in the dairy bacterium *L. lactis*. Although this organism has been studied for several decades, its metabolism is still not well understood. The studies here fill significant gaps in this understanding.

Chapter 4 focuses on the complicated task of system identification, which may be considered the bottleneck of metabolic modeling. Over the past decades, great effort has been devoted to the sub-task of parameter estimation. This chapter addresses

the even greater challenge of inferring both the mathematical representation and parameterization of metabolic flux systems from time series of concentrations and is focused on extensions of the previously developed dynamic flux estimation (DFE). A specific issue burdening this inference method is the prevalent scenario that pathway systems contain fewer metabolite pools than reaction steps. This mismatch leads to an under-determined stoichiometric system of equations within the flux space at each time point. To alleviate this problem, the system will be subjected to optimization strategies that will ensure positivity and functional continuity of all fluxes, as well as other criteria of functional and biological validity and efficiency. These optimizations will lead to constraints within the admissible parameter space and provide very strong guidance for the following steps of parameter estimation. A subsequent section within this chapter addresses issues of incomplete data and advanced auxiliary methodologies for DFE.

Chapter 5 addresses the important task of data preprocessing by introducing a new constrained iterative wavelet smoother. Experimentally obtained time series data are always corrupted by noise to some degree. Most methods of data analysis and information retrieval require smoothing and noise reduction to be performed on raw data. This step is a crucial prerequisite for parameter estimation, but is currently not satisfactorily solved.

Finally, **Chapter 6** contains concluding remarks and suggestions for future research in this area.

CHAPTER II

MODEL DESIGN FOR THE GLYCOLYTIC PATHWAY OF *L. LACTIS* [1]

2.1 Introduction

The grand challenge of computational systems biology is the translation of biological systems into adequate mathematical models that permit analysis, prediction, manipulation, optimization, explanation, understanding, and the discovery of biological design and operating principles. Recent technological advancements have facilitated the generation of time series data that characterize the dynamics of genomic, proteomic, metabolic, and physiological responses. Of particular interest for biochemical pathway modeling and proteomics is the availability of relatively time-dense profiles of metabolites or proteins through measurement techniques such as mass spectrometry, nuclear magnetic resonance (NMR), protein kinase phosphorylation, or mass cytometry (CyTOF). These data contain valuable, but implicit, information about the structure and dynamics of the biological system under study. The availability of dynamic data enables the use of top-down modeling approaches. These approaches typically consist of minimizing the discrepancy between the measured data, *i.e.*, the time profiles, and the assumed model, which typically consists of a system of nonlinear ordinary differential equations (ODE) that are to be parameterized ([9]).

This chapter focuses on the task of model design and parameter estimation for the glycolytic pathway of *Lactococcus lactis*. In an effort to build a kinetic-dynamic model using time-series data, the most challenging steps are the identification of suitable mathematical formats for all flux representations and the estimation of their

[1] MUCH OF THIS MATERIAL HAS BEEN SUBMITTED FOR PUBLICATION.

unknown kinetic parameters. Central to these tasks is the methodology of dynamic flux estimation (DFE), which is augmented here with various other auxiliary techniques. The modeling effort culminates in a number of biological results and insights, which will be described in detail in Chapter 3.

2.2 Data

The time series data on which this study is based were collected through *in vivo* nuclear magnetic resonance (NMR) spectroscopy [47] from *L. lactis* cells that were initially starved and then offered a pulse of labeled glucose at time zero. These data contain extensive information about the structure, dynamics and regulation of the organisms metabolism, which in this method is essentially unadulterated by cell disruption, purification, centrifugation or other harsh experimental methods. The technical aspects of the non-invasive determination of the concentrations of the intracellular pools of intermediate metabolites using *in vivo* nuclear magnetic resonance spectroscopy (^{13}C- and ^{31}P-NMR) experiments was thoroughly discussed in [45, 46, 47]. A brief summary of the data follows below.

The measured time series include the external concentrations of glucose and of the end product lactate in the medium, along with several of the more abundant intermediate metabolites, namely glucose 6-phosphate (G6P), fructose 1,6-bisphosphate (FBP), phosphoenol pyruvate (PEP), 3-phosphoglycerate (3PGA). The time series reflect three experiments performed with 20, 40, or 80 mM of glucose input. Additional time series are available for some ubiquitous cofactors, such as NAD^+ and NADH, which are detectable with somewhat limited detection capability due to the required NMR acquisition time [46]. Finally, the level of NTP was measured using ^{31}P-NMR. The datasets used for modeling are summarized in Table 1.

In Experiment 1, 20 mM of [6-^{13}C] glucose was supplied to the cell suspension, and time series of concentrations were recorded for glucose, G6P, FBP, 3PGA, and

Table 1: Data Overview

Experiment	Condition	Technique	Metabolites Measured	Time
1: 20 mM glucose labeled on C_6	Anaerobic, pH 6.5	^{13}C-NMR	glc, G6P, FBP, 3PGA, lac	30 sec
2: 40 mM glucose labeled on C_1	Anaerobic, pH 6.5	^{13}C-NMR	glc, FBP, 3PGA, PEP, lac, NAD^+, NADH	2.2 min
		^{31}P-NMR	ATP, P_i	2.75 min
3: 80 mM glucose labeled on C_1	Anaerobic, pH 6.5	^{13}C-NMR	glc, FBP, 3PGA, PEP, lac, NAD^+, NADH	2.2 min
		^{31}P-NMR	ATP, P_i	2.75 min

lactate with a resolution of 30 seconds. In Experiments 2 and 3, cells were supplied beforehand with labeled [5-^{13}C] nicotinic acid, a precursor of NADH, thus ensuring this pool to be 100% labeled [46]. To start the *in vivo* NMR experiment, cells were supplied with 40 and 80 mM of [1-^{13}C] glucose, and time series of glucose, FBP, PEP, 3PGA, lactate, NAD^+ and NADH were recorded. The time resolution in these experiments was 2.2 minutes, which was needed to accommodate NAD^+ and NADH determination with a reasonable signal-to-noise ratio. In a separate, comparable experiment, cellular metabolism was investigated with ^{31}P-NMR and thus allowed the measurement of NTP (mostly ATP), P_i and pH with a time resolution of 2.75 min. Usage of [1-^{13}C] glucose prevents the determination of G6P in these datasets, due to the similarity in chemical shifts of [1-^{13}C] glucose and [1-^{13}C] G6P. Although the data described above exhibit experimental noise and are incomplete, with some metabolites or time points missing in each set, they are as good as a modeler can presently hope for.

Under the given experimental conditions, cells are not able to synthesize ATP (nor ADP) de novo, because they are suspended in phosphate buffer and supplied only with glucose, which is insufficient for *L. lactis* to grow. Once the culture is harvested and washed prior to the NMR experiment, the total amount of ATP+ADP remains constant. This fact enables us to infer ADP from the ATP data where these

are available (for Experiments 2 and 3), which is beneficial for the initial parameter estimation.

In all datasets, the concentrations of 3PGA and PEP are covariant, due to the fact that these two metabolites can be converted into 2PGA by two enzymatic steps (phosphoglycerate mutase and enolase). These reactions are fast and reversible with equilibria that favor 3PGA and PEP [11, 42, 51], thus maintaining 2PGA in a concentration range below the *in vivo* ^{13}C-NMR detection level. Since this covariance is maintained even during high glycolytic flux, we decided to aggregate these two pools into one dependent variable (X_4). Nonetheless, we are still able to calculate the concentration of each intermediate from the constant proportionality of $\tilde{0}.6444$ (3PGA/PEP).

The data seem to indicate an apparent absence of PEP and FBP at the beginning of the experiment. However, these metabolites are present, but just not labeled and therefore missed by the NMR detection. The concentrations of these metabolites were measured in a control experiment [45]. Furthermore, it seems to be a reasonable assumption that the cells re-enter a state of starvation at the end of the experiment and that the residual values of now labeled PEP and FBP constitute a state that is similar to the state at the beginning of the experiment.

2.3 Model Design

Generically, a mathematical model of a metabolic system consists of a system of ordinary differential equations with three components: (1) its stoichiometric matrix; (2) a vector of fluxes; and (3a) the functional forms of the fluxes and (3b) their corresponding parameter values. The stoichiometric matrix represents the essentially time-invariant wiring diagram of the pathway and shows which fluxes enter or leave each pool. It is often assumed to be known from biochemical experimentation, and the results of uncounted such studies are collected in databases like KEGG [37] and

MetaCyc [7]. Kinetic details and regulatory features are represented by parameters that are incorporated in appropriate functional forms representing the fluxes. The structure of the pathway is presented in Figure 1.

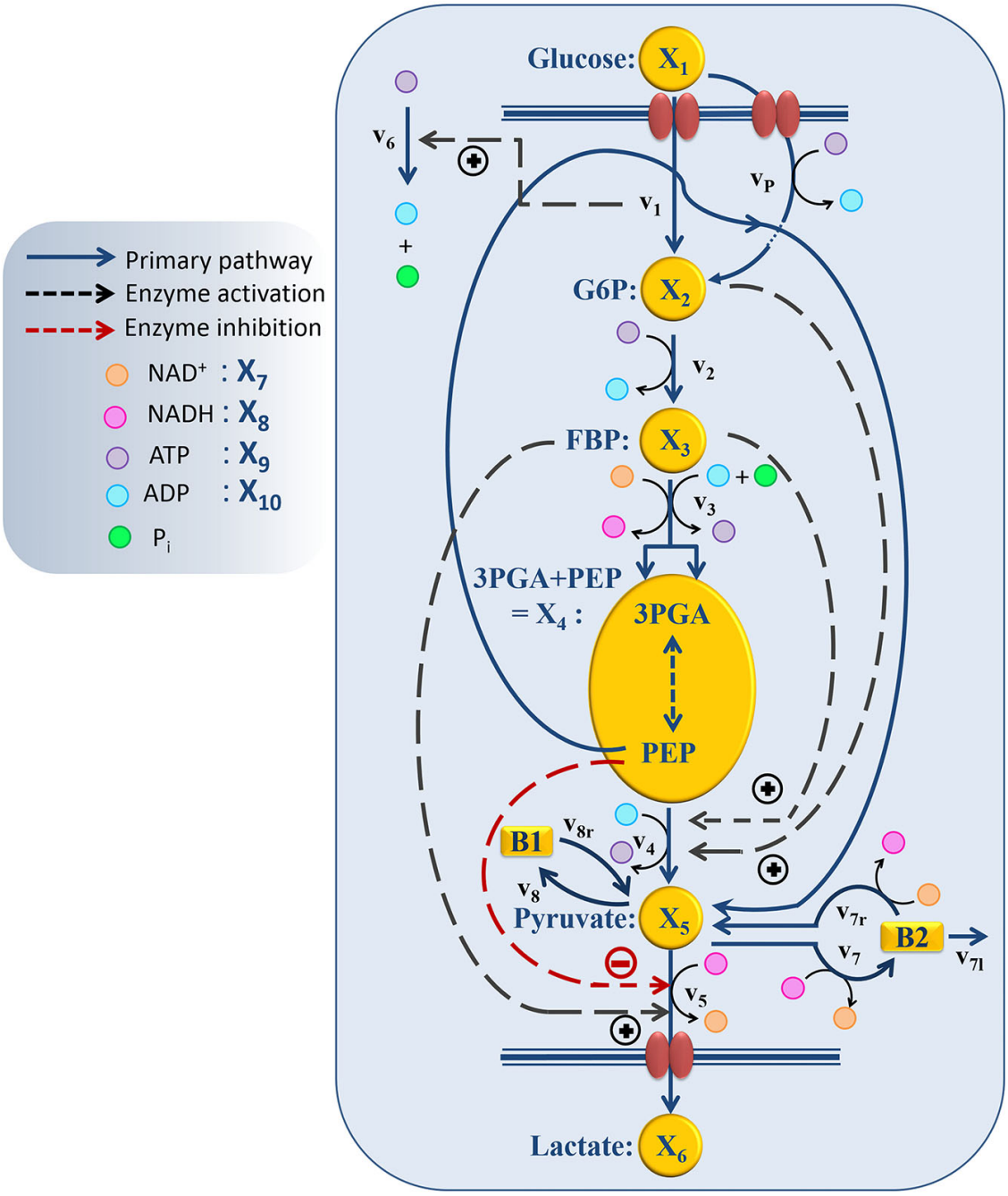

Figure 1: Model structure of the glycolytic pathway in *L. lactis*. Of particular importance are the PTS mechanism, which uses PEP for the initial phosphorylation of glucose, and several regulatory signals, indicated here with dashed arrows.

2.3.1 Dynamic Flux Estimation

The numerical characterization of an appropriate metabolic pathway model consists of the identification of mathematical formats for all process steps, and the estimation of optimal numerical parameter values in these process formulations. The determination of optimal representations for the processes in the model is by no means trivial, as no guidelines are available, but it is very important, because inadequate representations, even if they fit a target dataset, run the risk of error compensation among flux terms and of incurring problems during extrapolations [30, 31, 68].

We use for this important identification step an extension of Dynamic Flux Estimation (DFE) [31]. The main attraction of DFE is the fact that this method does not presuppose a functional form for any of the flux representations. This feature allows us to test in an objective manner whether particular functions, such as power-laws, Michaelis-Menten rate laws, or Hill functions, are capable of appropriately modeling a specific flux, or if other formulations should be considered. Importantly, careful analyses of all fluxes in this manner may suggest the existence of regulatory signals that had been missing from the assumed pathway structure. Such a suggestion corresponds to a novel hypothesis that is in principle testable with lab experiments and may lead to biological discoveries. An example is the case of glucose uptake, which is discussed later in section 3.2.

In addition to its diagnostic capacities, DFE allows for a much more efficient parameter estimation strategy in terms of computation cost associated with the integration of ODEs and global optimization. The parameters are estimated one flux at a time, thereby avoidingor at least reducingthe integration of ODEs.

DFE, as depicted in Figure 2, consists of two phases, of which the first is model-free and makes very few assumptions. It includes data preprocessing, time course smoothing and the estimation of slopes of the smoothed time courses. The ultimate result of this phase consists of numerical time series profiles of all fluxes; in other

words, one obtains plots of the fluxes against time or against contributing metabolites, but no functional formats. The second phase is dedicated to the mathematical characterization and parameterization of each process representation. This phase is still difficult but much simpler than the estimation of ODE systems, because it targets explicit functions of one or a few variables and with correspondingly few parameters.

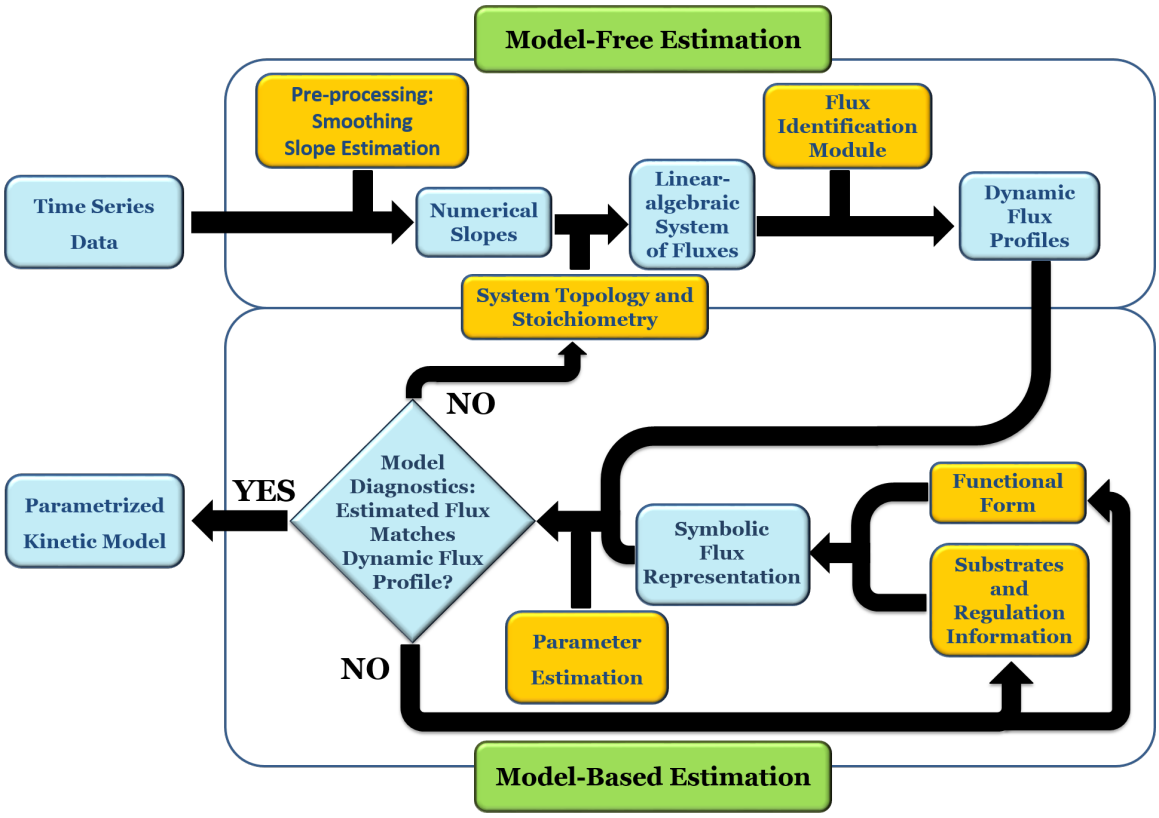

Figure 2: DFE is a model characterization strategy and consists of two phases. In the first, model-free estimation phase, it takes time series of concentration data as input and estimates the dynamic flux profiles, which in turn are used as input to phase 2, which consists of a model-based estimation. In this phase, functional forms and regulatory assumptions are incorporated and parameters are estimated for each flux separately.

As part of the first, model-free phase, DFE requires the estimation of slopes from the metabolic time course data. Several methods have been proposed for this purpose (*e.g.*., see discussion in [63]); as a further alternative, Chapter 5 introduces a smoothing algorithm based on the wavelet transforms that ensures conservation of

mass [23]. Once the data are smoothed, the differentials on the left-hand sides of the differential equations are replaced with the estimated slopes at many time points. Consequently, each differential equation is replaced with a set of algebraic equations, and each of these sets can be evaluated independently of all other sets. Thus, the entire system of equations is now purely algebraic [61, 71, 74]. The algebraic equations are solved and result in numerical time series of fluxes. This process is particularly straightforward if the number of independent fluxes equals the number of independent variables for which data exist. If the stoichiometric matrix is not full-rank, which is the situation here, a direct inversion is not possible. Instead, we look for a full-rank subset of the identifiable fluxes, which we can subsequently use as constraints for the identification of the remaining fluxes with auxiliary methods (see Chapter 4).

The second phase of DFE, shown in the lower panel of Figure 2, calls for the identification of functional forms of all fluxes. We start this phase by assuming a typical functional form for each process, such as a mass action, power law, Michaelis-Menten, or Hill function, and estimate the corresponding parameters by minimizing the error between the observed metabolite concentrations and the concentrations resulting from the assumed functional forms. The best fit, along with typical regression diagnostics, such as a run test for residuals [25], indicates whether the assumed functional form might be appropriate or is obviously wrong. No generic strategies exist at this point for selecting candidates or proving their optimality. For the special case of the power-law format, it is feasible to use a logarithmic transformation and test the appropriateness with diagnostic methods of multiple linear regression. In general, such a test is difficult.

While DFE has substantial advantages, it requires that the stoichiometric matrix of the system has full rank, which is frequently not the case. To extend the applicability of DFE beyond this case, auxiliary methods have been proposed for using additional information to make the stoichiometric matrix invertible (*e.g.*., [10, 36, 70, 76];

however, these methods are seldom general and often require very specific features in the datasets. As an alternative Chapter 4 describes a generic flux identification procedure for underdetermined systems.

2.3.2 Generic Issues of Parameter Estimation

Once functional forms for all processes are determined through DFE and appropriate additional assumptions and settings, the model is to be fitted per optimization of parameter values. This optimization generically calls for minimizing a cost function, calculated as the sum of squared differences between the experimental data points and simulated values, summed over all time points and metabolites, and over all three data sets simultaneously. Although straightforward in principle, this optimization is almost always difficult [9].

In our specific case, several factors render the estimation of system parameters particularly challenging. First, the regulatory structure and reaction mechanisms of the model are a priori unknown, and their identification is by itself non-trivial, even if aided by DFE. Second, the error surface is embedded within a large parameter space of 44 dimensions. This surface is complicated, and one has to expect numerous local minima that severely confound steepest descent, evolutionary, or randomized optimization techniques that are typically used for parameter estimation. Third, datasets for some of the intermediate metabolites are not available, due to experimental limitations and, in particular, of the NMR-technique, which include higher detection limits for the various metabolites than would be desirable. Also, the ^{13}C- and ^{31}P-NMR experiments, by their nature, must be executed separately and therefore both exclude some of the dependent variables. Missing data include time series for pyruvate for all three datasets, G6P for datasets 2 and 3, and ATP, NAD^+, and NADH for dataset 1, as well as F6P, G3P, DHAP, 1,3GBP, 2PGA, and ADP for all

datasets. Fourth, stiffness and other numerical issues associated with the model equations can lead to substantial algorithmic difficulties while integrating the differential equations. While these issues may not even be present in the ultimate, optimized model, they are frequently encountered when an estimation algorithm determines inopportune combinations of settings during its scanning of a high-dimensional parameter space [71]. Finally, even if a good fit has been determined, one cannot be sure of its uniqueness and must assume that the system might have a certain degree of sloppiness [33, 34, 58, 64, 53]. Many of these issues can be addressed by employing the methodologies of Chapter 4 as an extension of DFE. In the present case, incorporating three datasets in conjunction with DFE can alleviate the identifiability issues and help reveal regulatory information about the pathway.

2.3.3 Functional Formats of Fluxes: Biochemical Systems Theory

Biochemical systems theory (BST) is a mathematical and computational framework of ordinary differential equations (ODEs), which was originally developed for modeling and simulating biochemical pathways but has been widely applied to other biological systems ever since [69, 55, 75]. BST is considered canonical, which implies that the construction of the system of ODEs, its analysis, and its diagnosis follow relatively strict, well-structured guidelines. The power-law representation of each reaction is the key ingredient of BST. It constitutes a multi-variate, linear approximation in a logarithmic space and expresses a process as a product of power-law functions of all variables that directly affect the process [54, 75]. It has been shown that these power-law models are highly effective representations of biochemical kinetics [55, 67]. Power-laws offer the flexibility of non-integer kinetic orders, which enable a representation of situations commonly found in real biological systems. Additional support for the richness of power-laws in presenting complex nonlinear dynamic behavior comes from work showing that essentially any set of continuous nonlinear differential equations

can be recast equivalently as a power-law system [56].

A generic reaction v_k within the BST formalism is represented as $v_k = \alpha_k \prod_{j=1}^{n} X_j^{h_{kj}}$, $k = 1, 2, \cdots, m$, where the rate constant α_k and the kinetic orders h_{kj} are fundamental characteristics. The rate constant describes the turn-over of the process, while the kinetic orders quantify the strengths with which reactants and regulators affect the process. These parameters need to be estimated to provide a full description of the metabolic pathway under study.

2.3.4 Details of the Mathematical Representation of the *L. lactis* Model

Our default for the functional forms of fluxes in the *L. lactis* model is a product of power-law functions, so that the model is mostly in the generalized mass action (GMA) format within the framework of BST. However, we allow for deviations from this format if they are suggested by the DFE analysis. The result is presented in Eqs. 1 and 2. In addition to the typical BST parameters, namely rate constants and kinetic orders, the equations contain the intracellular and extracellular volumes Vin and Vout. The extracellular volume is 50 ml in all three experiments, while the intracellular volume is calculated from the measured biomass in each case, using a conversion value of $2.9 \mu L/mg$ of protein for the intracellular volume [49]. The differential equations need to account for these volume differences, because the amounts of biomass, M_B, differ among the experiments. B_1 and B_2 in Eqs. 2h and 2i are temporary buffers.

$$\dot{X}_1 = -\frac{v_1}{V_{out}} - \frac{v_P}{V_{out}} \tag{1a}$$

$$\dot{X}_2 = \frac{v_1}{V_{in}} + \frac{v_P}{V_{in}} - \frac{v_2}{V_{in}} \tag{1b}$$

$$\dot{X}_3 = \frac{v_2}{V_{in}} - \frac{v_3}{V_{in}} \tag{1c}$$

$$\dot{X}_4 = 2\frac{v_3}{V_{in}} - \frac{v_1}{V_{in}} - \frac{v_4}{V_{in}} \tag{1d}$$

$$\dot{X}_5 = \frac{v_4}{V_{in}} + \frac{v_1}{V_{in}} + \frac{v_5}{V_{in}} - \frac{v_7 - v_{7r}}{V_{in}} - \frac{v_8 - v_{8r}}{V_{in}} \tag{1e}$$

$$\dot{X}_6 = \frac{v_5}{V_{out}} \tag{1f}$$

$$\dot{X}_7 = \frac{v_5}{V_{in}} - 2\frac{v_3}{V_{in}} + \frac{v_7 - v_{7r}}{V_{in}} \tag{1g}$$

$$\dot{X}_8 = -\frac{v_5}{V_{in}} + 2\frac{v_3}{V_{in}} - \frac{v_7 - v_{7r}}{V_{in}} \tag{1h}$$

$$\dot{X}_9 = 2\frac{v_3}{V_{in}} - \frac{v_2}{V_{in}} + \frac{v_4}{V_{in}} - \frac{v_6}{V_{in}} - \frac{v_P}{V_{in}} \tag{1i}$$

$$\dot{X}_{10} = -2\frac{v_3}{V_{in}} + \frac{v_2}{V_{in}} - \frac{v_4}{V_{in}} + \frac{v_6}{V_{in}} + \frac{v_P}{V_{in}} \tag{1j}$$

$$\dot{B}_1 = \frac{v_7 - v_{7r}}{V_{in}} - \frac{v_{7l}}{V_{in}} \tag{1k}$$

$$\dot{B}_2 = \frac{v_8 - v_{8r}}{V_{in}} \tag{1l}$$

While power-law functions appear to be adequate for most flux terms, they are in general not defined for variables with values of zero, which do appear in our system toward the end of the experiments. It is therefore beneficial to introduce switches in the equations that set flux terms equal to zero before a variable becomes zero. The equations below 2.a-2.i contain the model that includes such switches. These switches are implemented by multiplying these fluxes by $(X_1 > 10^{-4})$, so that when glucose drops below the concentration of 10^{-4} mM, these fluxes become zero. Only two switches were considered for the fluxes out of glucose, namely the PTS flux (v_1) and the permease flux (v_P).

$$v_1 = \alpha_1 M_B X_1^{h_{1,1}} (\frac{X_4}{X_4 + K_{M4}})(1 + X_4)^{-h_{1,4}}(1 - e^{-\frac{t}{T}})X_7^{h_{1,7}} \tag{2a}$$

$$v_P = \alpha_0 M_B X_1^{h_{0,1}} X_9^{h_{0,9}} \tag{2b}$$

$$v_2 = \alpha_2 M_B X_2^{h_{2,2}} X_9^{h_{2,9}} \tag{2c}$$

$$v_3 = \alpha_3 M_B X_3^{h_{3,3}} X_7^{h_{3,7}} X_{10}^{h_{3,10}} \tag{2d}$$

$$v_4 = \alpha_4 M_B X_4^{h_{4,4}} X_{10}^{h_{4,10}} X_2^{h_{4,2}} X_3^{h_{4,3}} \tag{2e}$$

$$v_5 = \alpha_5 M_B X_5^{h_{5,5}} X_8^{h_{5,8}} X_3^{h_{5,3}} (1 + X_4)^{-h_{5,4}} \tag{2f}$$

$$v_6 = \alpha_6 M_B X_9^{h_{6,9}} v_1^{h_{6,1}} \tag{2g}$$

$$v_7 = \alpha_7 M_B X_5^{h_{7,5}} X_8^{h_{7,8}}, v_{7r} = \alpha_{7r} M_B B_2^{h_{7r,2}} X_7^{h_{7r,7}}, v_{7l} = \alpha_{7l} M_B B_2^{h_{7l,2}} \tag{2h}$$

$$v_8 = \alpha_8 M_B X_9^{h_{8,9}}, v_{8r} = \alpha_{8r} M_B B_1^{h_{8r,1}} \tag{2i}$$

The ODE solver utilized is ode15s in MATLAB, which is used for stiff ODE systems.

Detailed analyses during the second phase of DFE suggest that the PEP: Carbohydrate Phosphotransferase System (v_{PTS}), is a function not only of its substrates glucose and PEP, but that it also needs to be regulated by additional effectors. A detailed analysis in Chapter 3 (Section 3.2) leads to the specific conclusion that NAD^+ is a potential activator, that 3PGA is an inhibitor, or that both effectors are present. These regulatory terms are included in the model (Eq. 2a) as $X_7^{h_{1,7}}$ and $(1 + X_4)^{-h_{1,4}}$ respectively.

2.3.4.1 Modeling the First Two Minutes of Glucose Uptake

The initial rate of glucose uptake increases with decreasing substrate, *i.e.* glucose, availability in the first two minutes. This phenomenon is counterintuitive, as more substrate seems to suggest higher uptake. Section 3.2 will discuss various reasons that could be responsible for this observation. Nonetheless, in order to model this so-far unexplained observation, a black box module was used. Namely, instead of ignoring

the first time period and starting the model at time 2 min, we multiply a term of the form $(1 - e^{-t/T})$ to the modeled PTS flux. This term exponentially approaches 1 and loses its effect after a short period of time. For instance, at $3T$ minutes, it is equal to $1 - e^{-3} \approx 0.9502$. Thus, the term is ineffective for most of the experimental period. Different T's were allowed and fitted for different experiments.

2.4 Results

2.4.1 General Features of the Model

Based on literature information, we established an initial model diagram, determined fluxes with methods of DFE, and subsequently performed parameter estimation. The fully parameterized model was diagnosed with standard methods of stability, sensitivity, and robustness analysis, and subsequently used for representative simulations.

The analysis led to several slight amendments of the original diagram (Fig. 1). First, we observed that the actually measured total carbon mass decreases over time for all three datasets, with about 7% -10% of the mass being unaccounted (see Fig. 4A), in spite of the fact that, outside glycolysis, the organism is metabolically inactive under the given conditions [47]. The loss is biologically not very significant, but it is immediately inconsistent with the model structure of the initial diagram, where glucose is the only input substrate and lactate is the only output. The most reasonable option for remedying this inconsistency is the addition of a minor efflux out of one or more metabolite pools. A good candidate is pyruvate, because several reactions could use pyruvate as a substrate and thereby be responsible for the diversion of material [29]. Because these effluxes potentially affect the NAD^+/NADH balance, we included two types of leakage from pyruvate, one with and one without the consumption of NADH. Optimization determined the magnitudes of these very-low-capacity fluxes. We also considered alternative locations of leakage, such as PEP and G6P, but did not find them beneficial. Other amendments are discussed in the following sections.

The dots in Figure 4A show the total carbon mass calculated from the measured data, and the lines show the calculated amounts resulting from the simulated model with leakage terms.

The model contains twelve dependent variables. Six represent the main metabolites glucose, G6P, FBP, the aggregated pool of 3PGA and PEP, pyruvate, and lactate, four variables represent the cofactors ATP, ADP, NAD^+, and NADH, and the remaining two represent temporary buffers [75]. Eq. 1 of the previous section presents ordinary differential equations describing the system connectivity, while Eq. 2 shows the functional representations. The model is more complicated than earlier models of glycolysis in *L. lactis*, as it addresses the pathway dynamics under anaerobic conditions. In contrast to aerobic conditions, the organism cannot easily recycle NAD^+ under anaerobic conditions, and the balance between NAD^+ and NADH therefore changes dynamically. These changes must be expected to affect the dynamics of glycolysis and are therefore considered important for the model.

2.4.2 Fully Parameterized Model and Model fits

Methods of DFE, combined with numerous parameter estimation techniques, led to the following set of parameter values, which lead to simultaneous fits for the three available experiments.

$$
\begin{aligned}
P = [&\alpha_1 \quad h_{1,1} \quad K_{M4} \quad h_{1,4} \quad h_{1,7} \cdots \\
&\alpha_0 \quad h_{0,1} \quad h_{0,9} \cdots \\
&\alpha_2 \quad h_{2,2} \quad h_{2,9} \cdots \\
&\alpha_3 \quad h_{3,3} \quad h_{3,7} \quad h_{3,10} \cdots \\
&\alpha_4 \quad h_{4,4} \quad h_{4,10} \quad h_{4,2} \quad h_{4,3} \cdots \\
&\alpha_5 \quad h_{5,5} \quad h_{5,8} \quad h_{5,3} \quad h_{5,4} \cdots \\
&\alpha_6 \quad h_{6,9} \quad h_{6,1} \cdots \\
&\alpha_7 \quad h_{7,5} \quad h_{7,8} \cdots \\
&\alpha_{7r} \quad h_{7r,2} \quad h_{7r,7} \cdots \\
&\alpha_{7l} \quad h_{7l,2} \cdots \\
&\alpha_8 \quad h_{6,9} \cdots \\
&\alpha_{8r} \qquad h_{8r,1}];
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
P = [&12.5290 \quad 0.1422 \quad -0.2547 \quad 0.2250 \quad 0.6469 \cdots \\
&1.1208 \quad 1.9844 \quad 0.2742 \cdots \\
&0.2403 \quad 1.0079 \quad 0.7114 \quad 0.0603 \cdots \\
&0.1600 \quad 0.6192 \quad 0.8661 \quad 0.3715 \quad 0.7479 \cdots \\
&0.6522 \quad 3.0000 \quad -0.5013 \quad 0.4867 \quad 0.8638 \cdots \\
&0.3432 \quad 1.2558 \quad 0.4101 \cdots \\
&0.1228 \quad 0.1720 \quad 0.0252 \cdots \\
&0.3293 \quad 1.3664 \quad 1.4756 \cdots \\
&0.1504 \quad 0.3026 \quad 0.1657 \cdots \\
&0.0061 \quad 0.0764 \cdots \\
&1.0528 \quad 0.4075 \cdots \\
&0.1950 \quad 0.5500];
\end{aligned}
\tag{4}
$$

In addition, the model uses the following initial values for the dependent variables and settings:

$$Y0_{80} = [80; 0.1; 4; 14.8; 0.1; 0.1; 5.74; 0.1; 0.1; 8.815; 0.1; 0.1];$$

$$Y0_{40} = [40; 0.1; 4; 14.8; 0.1; 0.1; 5.74; 0.1; 0.1; 8.815; 0.1; 0.1];$$

$$Y0_{20} = [20; 0.1; 4; 14.8; 0.1; 0.1; 5.74; 0.1; 0.1; 8.815; 0.1; 0.1];$$

$$T_{20} = 0.7124\ min \quad T_{40} = 0.7450\ min \quad T_{80} = 1.6661\ min$$

$$M_{B20} = 13.92\ mg\,protein/ml$$

$$M_{B40} = 17.11\ mg\,protein/ml$$

$$M_{B80} = 19.53\ mg\,protein/ml$$

Equation 3 shows the parameter vector. These parameter settings lead to good fits of the data, which are displayed in Panels A-C of Fig. 3. The capacity of a single parameter set to capture different conditions is important, because it significantly increases the predictive power of the model. Furthermore, since this parameter set was derived from DFE, its extrapolation reliability is increased, because the risk of error compensation among flux terms within the same or in different equations is greatly reduced [31].

Nonetheless, in reality, different natural systems obviously exhibit a certain degree of variability. We allowed the common parameter set, which fits the three experiments simultaneously, to vary slightly in order to account for this variability among the different cell populations (Eq. 6-7). The results are shown in Panels D-F of Figure 3. These model instantiations used the following parameter sets for the 20, 40, and 80 mM experiment. A parameter-by-parameter comparison demonstrates how close the three sets are.

$$
\begin{aligned}
P_{20} = [&14.3646 \quad 0.0722 \quad -0.0858 \quad 0.1703 \quad 0.4805 \cdots \\
&1.0071 \quad 1.4924 \quad 0.1773 \cdots \\
&0.2032 \quad 1.0207 \quad 1.0117 \quad 0.1416 \cdots \\
&0.1888 \quad 0.6645 \quad 0.7712 \quad 0.6590 \quad 0.7766 \cdots \\
&0.5633 \quad 3.0000 \quad -0.6826 \quad 0.4946 \quad 1.0296 \cdots \\
&0.3627 \quad 1.1649 \quad 0.3731 \cdots \\
&0.0335 \quad 0.7391 \quad 0.1038 \cdots \\
&0.3761 \quad 1.9952 \quad 0.9710 \cdots \\
&0.1119 \quad 0.5980 \quad 0.7281 \cdots \\
&0.0282 \quad 0.1696 \cdots \\
&1.1080 \quad 0.5652 \cdots \\
&0.1998 \quad 0.6500]; \quad , \quad T_{20} = 0.7391 \; min
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
P_{40} = [&11.6912 \quad 0.1790 \quad -0.3313 \quad 0.3439 \quad 0.4664 \cdots \\
&0.9740 \quad 1.7155 \quad 0.4049 \cdots \\
&0.2059 \quad 0.9700 \quad 0.8164 \quad 0.1410 \cdots \\
&0.1830 \quad 0.4572 \quad 0.7985 \quad 0.4100 \quad 0.9091 \cdots \\
&0.5838 \quad 2.9162 \quad -0.6379 \quad 0.3987 \quad 0.8822 \cdots \\
&0.3335 \quad 1.1774 \quad 0.3556 \cdots \\
&0.0231 \quad 0.8343 \quad 0.1026 \cdots \\
&0.3624 \quad 2.0246 \quad 1.2299 \cdots \\
&0.0702 \quad 0.5273 \quad 0.5884 \cdots \\
&0.0238 \quad 0.2526 \cdots \\
&1.0248 \quad 0.2501 \cdots \\
&0.2021 \quad 0.6321]; \quad , \quad T_{40} = 0.5990 \; min
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
P_{80} = [&12.4598 \quad 0.1422 \quad -0.2547 \quad 0.2250 \quad 0.7444 \cdots \\
&1.2357 \quad 1.9157 \quad 0.2879 \cdots \\
&0.2403 \quad 1.0079 \quad 0.7090 \quad 0.0603 \cdots \\
&0.1600 \quad 0.6110 \quad 0.8661 \quad 0.3715 \quad 0.7479 \cdots \\
&0.6522 \quad 2.9076 \quad -0.5013 \quad 0.4635 \quad 0.8638 \cdots \\
&0.3432 \quad 1.2558 \quad 0.4101 \cdots \\
&0.2000 \quad 0.1720 \quad 0.0252 \cdots \\
&0.3293 \quad 1.4836 \quad 1.4756 \cdots \\
&0.1475 \quad 0.3104 \quad 0.1740 \cdots \\
&0.0061 \quad 0.0764 \cdots \\
&1.0528 \quad 0.4717 \cdots \\
&0.1950 \quad 0.5500]; \quad , \quad T_{80} = 1.5985 \; min
\end{aligned}
\tag{7}
$$

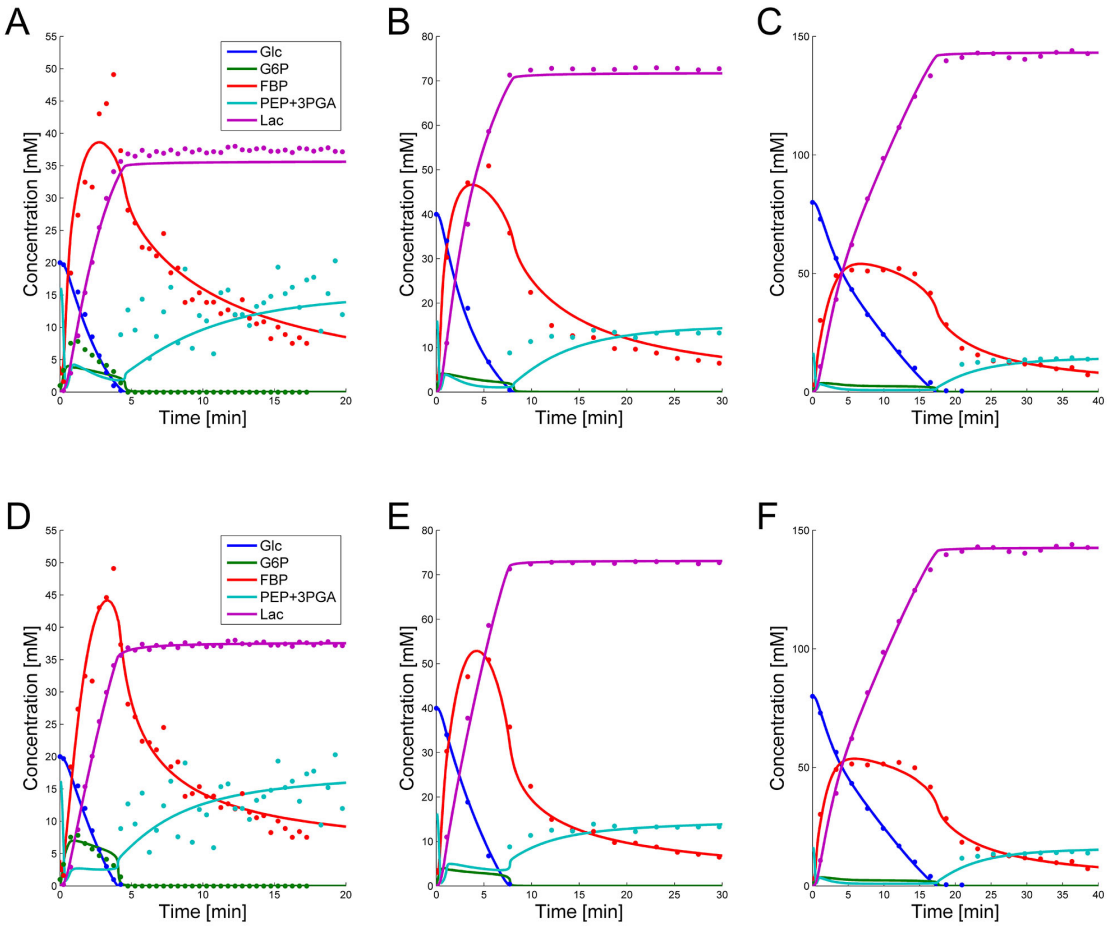

Figure 3: Simulation results for glucose, lactate, G6P, FBP and PEP+3PGA, superimposed on the corresponding data, for Experiment 1 (Panel A), Experiment 2 (Panel B), and Experiment 3 (Panel C). Note the different Y-scales. Accounting for modest variability among cell populations, the common parameter set for Panels A-C was allowed to vary slightly among experiments. The resulting fits are depicted in Panels D, E, and F.

2.4.3 Simulation Results for Secondary Metabolites

NAD^+, NADH, and ATP data are available only for Experiments 2 and 3. The simulation results for these metabolites as well as pyruvate are shown in Figures 4B and 4C. Although the fits are not as good as those for the main metabolites, they capture the trends and timing. Also, their concentrations are comparatively very small. For instance, no pyruvate was detected in the experiments. Therefore, the simulation results need to be below the detection limit of each experiment, possibly except for the first few minutes where unlabeled PEP and 3PGA are converted into unlabeled pyruvate that remains undetected in NMR experiments. The simulation results show a spike in the beginning; shortly afterwards, pyruvate decreases to below 2 mM for all experiments and stays below the detection limit until it approaches zero toward the end of the experiment.

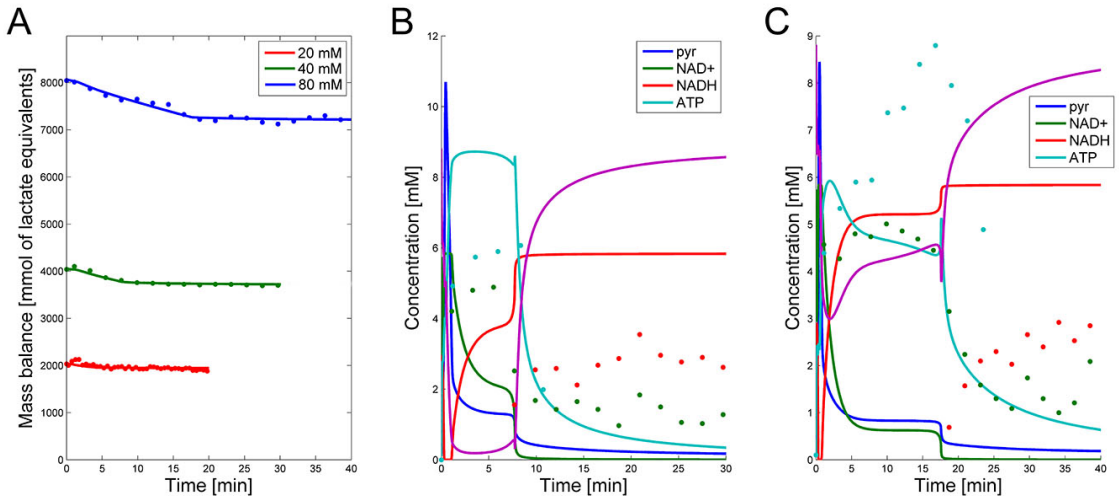

Figure 4: (A) Mass balance in mmol of lactate equivalents *vs.* time, calculated by taking into account the appropriate stoichiometry and volume conversions. The mass is plotted for the three datasets. (B) Data of NAD^+, NADH and ATP are shown as dots. Simulation results for NAD^+, NADH and ATP are superimposed. No data are available for pyruvate; simulation results are shown in dark blue.

2.5 Conclusions

To shed light on the surprisingly complicated glycolytic control system, a dynamic mathematical model was devised. This model accounts for the key glycolytic metabolites, as well as the dynamics of such cofactors as NAD^+, NADH, ATP, and ADP. As with many modeling studies, the most difficult step of model development was the estimation of parameter values. In this case, this estimation was based on experimental time series of glycolytic intermediates from three experiments with different substrate availability. The technical difficulties of the estimation process were directly related to the high dimensionality of the parameter space, the enormous complexity of the landscape of residual errors between model and data, and the often ignored fact that we do not really know what functions are best suited to represent each process within a biological system. To address these issues, a combination of mathematical and computational techniques were developed, including a custom-tailored Monte Carlo algorithm, different optimization techniques and, most importantly, methods of dynamic flux estimation (DFE), which enabled me to reduce the admissible parameter space and prevent flux terms from compensating errors in their representations.

For the first time, a single model fits all available metabolic time courses reasonably well with the same parameter set. Because this model reflects three independent datasets, one might expect that it has a higher extrapolation potential than earlier models that were based on single datasets. The next chapter will discuss key aspects of the control of the pathway, which is crucially important for survival.

CHAPTER III

NEW INSIGHTS INTO THE COMPLEX REGULATION OF THE GLYCOLYTIC PATHWAY IN *LACTOCOCCUS LACTIS* [1]

3.1 Introduction

The dairy bacterium *Lactococcus lactis* has to master a complicated task. It must control its essentially linear glycolytic pathway in such a fashion that, when the glucose substrate runs out, it retains enough phosphoenolpyruvate and fructose-1,6-bisphosphate to be able to restart glycolysis as soon as new glucose becomes available. Although glycolysis is arguably the best-studied metabolic pathway, its details in *L. lactis* are still unclear, and it is, in particular, not understood how the bacterium manages the stop-and-start task. The primary purpose of this chapter is a clarification of some of the details of the governing processes. The efforts described in Chapter 2 resulted in a fully kinetic, dynamic model, which constitutes a crucial prerequisite for the analysis in this chapter. This analysis offers a good example demonstrating how computational modeling can add genuine value to wet lab experimentation by rendering it possible to convert data, which provide snapshots of reality, into dynamic storylines that explain the strategies, which organisms employ to survive.

In contrast to previous models, which captured the dynamics of single datasets (*e.g.*, [35, 72, 32], Chapter 2 described a single aggregate model that combines three data sets for different input glucose concentrations. This single model enables the

[1] MUCH OF THIS MATERIAL HAS BEEN SUBMITTED FOR PUBLICATION.

assessment of several intriguing observations, which to date had not been explained convincingly and are discussed in this chapter.

For instance, if the glucose availability in the medium is increased after a period of starvation, one would expect a corresponding increase in the peak level of the downstream metabolite fructose-1,6-bisphosphate (FBP). However, the data clearly show that the peak concentration is largely independent of the external glucose concentrations. Specifically, comparing the different series of experimental results with increasing glucose concentrations, the FBP accumulation shows a progressively more noticeable plateau, whose duration, but not height, varies with substrate availability (Fig. 5). For instance, a previous model for aerobic conditions [72] was able to fit glycolytic data for a single glucose concentration of 20 mM, but extrapolating the model toward different glucose inputs, such as 40 or 80 mM, predicted almost a doubling or quadrupling of the FBP peak height, which is in clear contrast to the experimental measurements under anaerobic conditions.

While the earlier models had addressed the behavior of the organism under aerobic conditions [72], we study here the glycolytic pathway under its preferred anaerobic conditions. In this situation, no oxygen is available for the NADH-oxidase (NOX) reaction and, as a consequence, the concentrations of NAD^+ and NADH are exclusively affected by the glyceraldehyde phosphate dehydrogenase (GAPDH) and lactate dehydrogenase (LDH) steps of the glycolytic pathway. This restriction renders the NAD^+ and NADH concentrations pivotal for the timely shut-down of the pathway, as we will discuss later. Although we focus primarily on anaerobic operation, it is beneficial to compare the time courses under the two conditions (Fig. 6), because the comparison helps explain differences in the response strategies that *L. lactis* applies under different experimental conditions.

Figure 6 depicts the time profiles for the concentrations of glucose, FBP, PEP, 3PGA, end products lactate and acetate, as well as ubiquitous cofactors NAD^+ and

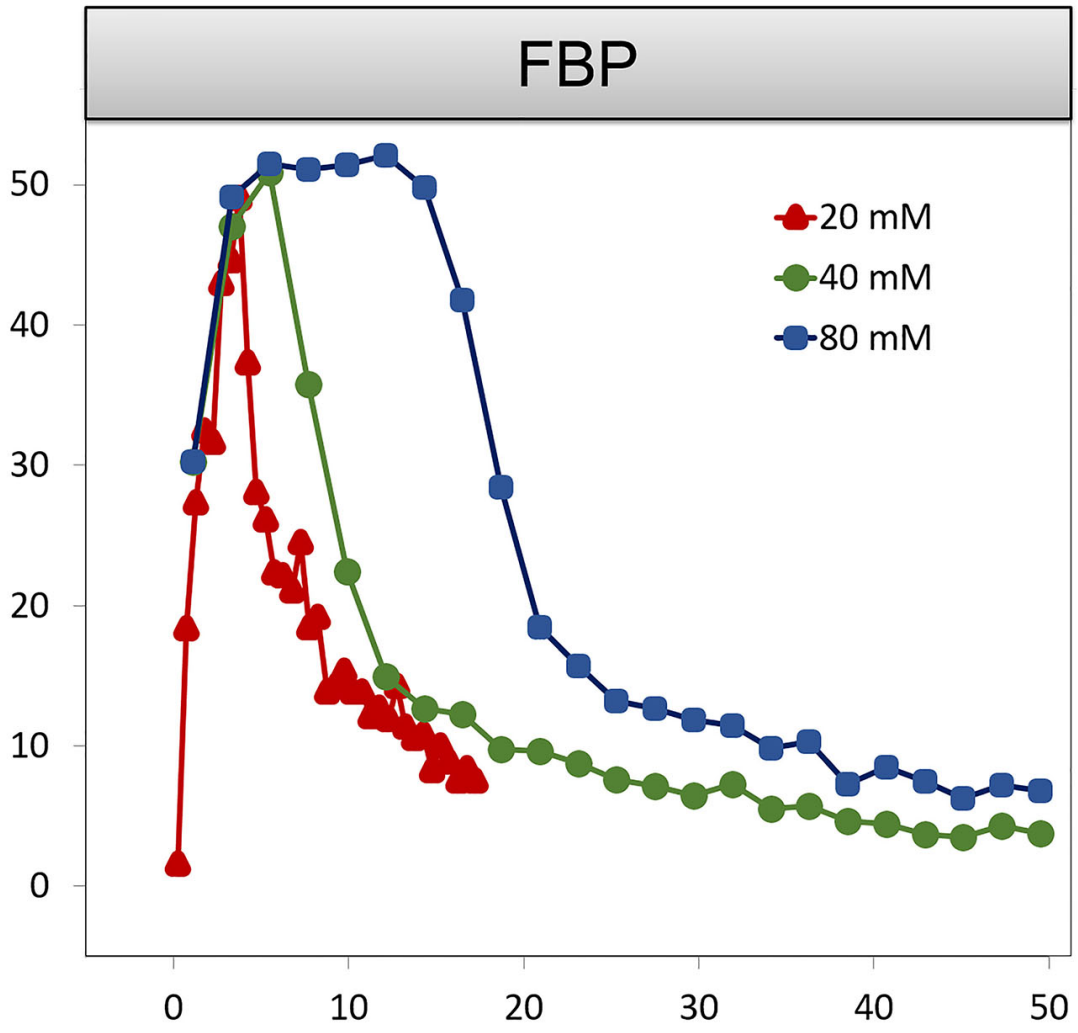

Figure 5: FBP data for three glucose concentrations in the medium under anaerobic conditions. The peak level seems to be independent of the available substrate concentrations. One also notes that FBP does not vanish completely and instead maintains some residual concentration.

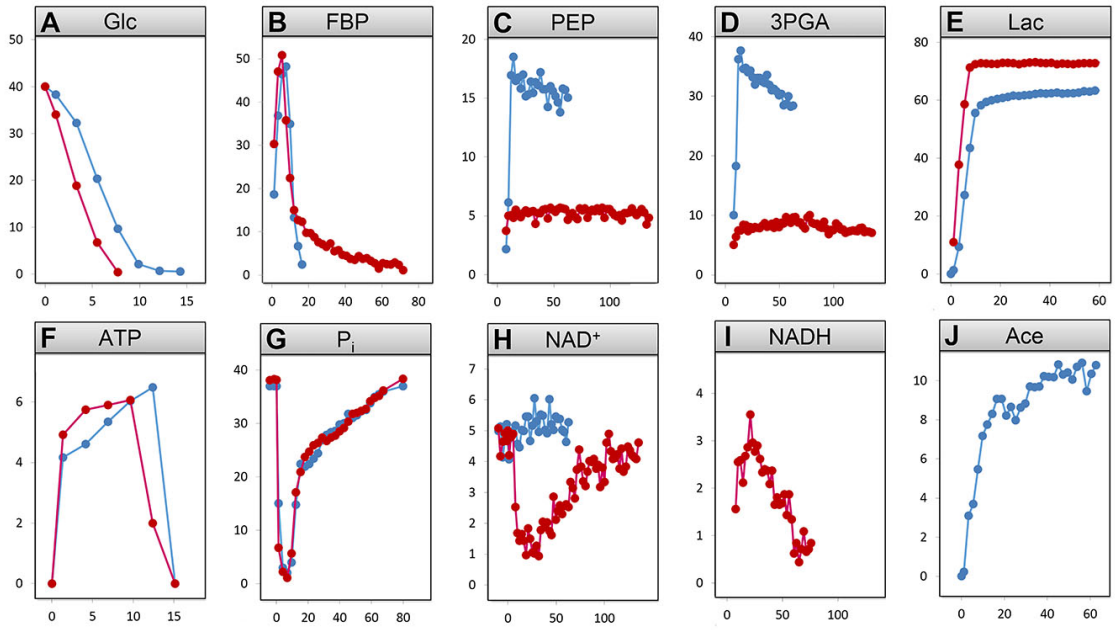

Figure 6: Comparison of the measured concentrations (mmol/l) of glucose, FBP, PEP, 3PGA, lactate, acetate, ATP, P_i, NAD^+ and NADH under aerobic (blue) *vs.* anaerobic (red) conditions. In both experiments, 40 mM of glucose was provided to the cells at time zero.

NADH, inorganic phosphate (P_i), and ATP under the two conditions. The two datasets were acquired after giving 40 mM of glucose to starving *L. lactis* cells at time zero. A comparison of the time profiles under the two conditions reveals some differences that are obvious, while others are rather subtle. For instance, NAD^+ drops sharply only under anaerobic conditions when glucose is depleted (Panel I). Under both conditions, PEP and 3PGA are closely correlated due to a fast equilibrium between the two metabolites. Nonetheless, the two exhibit concentrations of much lower magnitudes in the anaerobic experiment (Panels C and D). Lactate accumulates to somewhat lower amounts under aerobic conditions, due to channeling of some pyruvate toward acetate (Panels E and F). FBP becomes entirely depleted under aerobic condition, while it retains a residual concentration under anaerobic conditions (Panel B). This difference may appear to be minor, but our analysis later in this chapter will identify this difference as very important. Glucose is depleted faster under anaerobic

conditions (Panel A), and we will address this issue as well.

This study uses the computational model from Chapter 2 without changes. The model will be explored to decipher the complex coordination of regulatory signals of the pathway under both aerobic and anaerobic conditions and explain in detail the complex chains of events required to stop and restart glycolysis, which are distinct under aerobic *vs.* anaerobic conditions.

3.2 The PEP: Carbohydrate Phosphotransferase System (PTS)

PTS drives the uptake and consumption of glucose. The process consists of two tightly coupled components, namely the import of glucose and its immediate conversion into G6P. For purposes of structure identification and parameter estimation, the dynamics of this complex step can be derived directly with a simple form of DFE. This computation just requires estimation of the slope of the glucose consumption profile, which is not difficult as this profile is essentially error-free.

Intriguingly, the rate of glucose import rapidly increases during the first few minutes. This observation is structurally incompatible with any function whose rate monotonically increases with substrate availability, which is the case for typical Michaelis-Menten and power-law functions. The reason is that the glucose concentration is the highest at the beginning, which would suggest the highest uptake rate in the first minutes. Galazzo and Bailey [27] suggested that G6P could be an inhibitor of this step in yeast. But even if this were true in *L. lactis*, the initial rise could not be explained, because the concentration of G6P is initially very small.

Several options are available to address the initial brief rise in PTS. First, one could attempt to identify the true mechanisms leading to the initial increase in glucose uptake, which corresponds to the sigmoidal shape of the glucose concentration curve. For instance, one could explain this observation with the fact that the experimental

set-up causes slight delays, which could affect the uptake profile. Also, it has been argued that the cells might recover from starvation with some variability, which could be due to cell-to-cell variation in the speed of glucose uptake or slight differences in glucose availability to individual cells, or could be caused by the process of mixing of glucose throughout the medium or other extraneous factors. Indeed, it was shown with simulations that a narrowly distributed uptake profile can directly convert the monotonic trend in glucose consumption, as predicted by the model structure, into a sigmoidal trend, as it is observed [72]. Second, one could try to identify a "black-box" fitting function without being constrained by mechanistic considerations. Third, one could use the data directly as ("off-line") inputs instead of representing them functionally in the model [43]. As presented in Chapter 2, We decided on a hybrid option, where we use a black-box adjustment function for the first two minutes (see under Section 2.3.4.1) and dynamically model glucose uptake afterwards, starting at t = 2 min.

It is possible during the first two minutes that insufficient amounts of FBP accumulate, thereby limiting the material flux into the pool of PEP and 3PGA. As a consequence, the initial PEP concentration might not be able to produce the high level of v_1 that we directly, and in a model-free manner, computed through numerical differentiation of the glucose consumption profile. This short-term discrepancy can be resolved when we account for a small auxiliary permease flux (v_P), which has been observed in this strain of *L. lactis* [8], but was never used in earlier models.

There is a more significant issue with the shape of the glucose uptake profile until glucose is depleted. The issue might not be apparent simply by looking at the glucose concentration curves. In fact, this delicate but significant problem was first diagnosed and resolved with DFE techniques as outlined below.

The PTS flux (PTSf) can be assessed directly from the data, without a model or particular assumptions. Under the simple supposition that PTSf is a function in

the strict mathematical sense, *i.e.*, that it has a unique value for each argument, it is easy to show that the sole dependence of PTSf on glucose and PEP is insufficient. Namely, the three graphs in Figures 7A and B should overlap if they were functions of only glucose or PEP. However, they clearly do not.

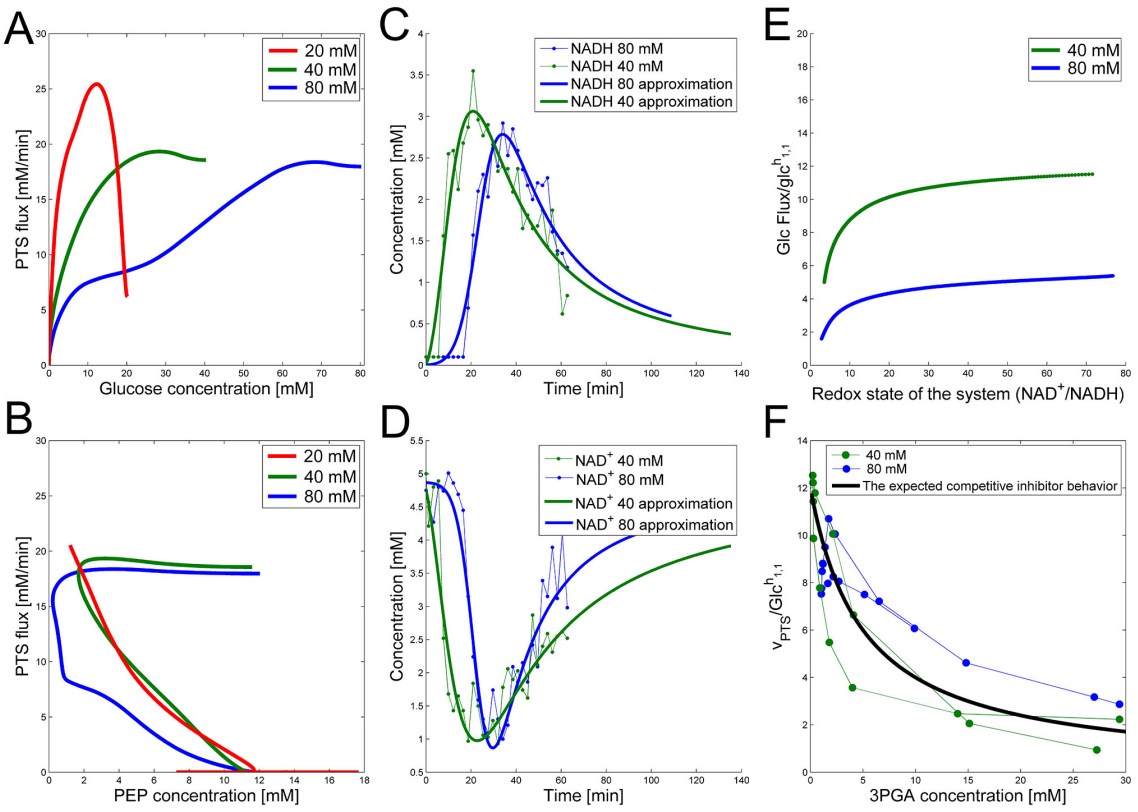

Figure 7: (A) PTSf *vs.* glucose concentration for Experiments 1 (blue), 2 (green), and 3 (red), assuming a constant dependence of PTSf on PEP. If PTSf were a true function of glucose, the plots would overlap. However, they clearly do not, thus demonstrating that PTSf is not a function solely of glucose. (B) PTSf plotted against PEP concentrations for Experiments 1 (blue), 2 (green) and 3 (red), assuming a constant dependence of PTSf on glucose. PTSf decreases with increasing concentrations of PEP. (C) NADH curves are smoothed with GS-functions and subsequently used for the analysis of PTSf. (D) NAD^+ curves indirectly smoothed by a GS function. (E) Plot of $\mathbf{PTSFglc^{h_{1,1}}}$ *vs.* NAD^+/NADH. (F) $\mathbf{PTSFglc^{h_{1,1}}}$ *vs.* 3PGA concentrations for duplicate experiments under aerobic conditions and for the same initial concentration of glucose (20 mM) are shown is blue and green. The dots show the measured data points and the thin lines connecting them are plotted to show the time adjacency of these points. The thick black line is a fitted $\frac{\mathbf{a}}{\mathbf{1+b.[3PGA]}}$ to these points, which exhibits a similar trend.

The first step of this assessment is the smoothing and numerical differentiation

of the glucose time courses from the three experiments. If one assumes that the process is saturated with respect to glucose, which is available in high concentrations, one may plot PTSf against the concentration of inferred PEP. One should note that *in vivo* NMR could not detect the concentration of PEP at the beginning of the experiment, until PEP starts to accumulate as glucose is being depleted. Therefore, the corresponding data for PEP are missing. The reason is that PEP is initially unlabeled and/or below the detection limit of *in vivo* NMR.

Neves [45] measured phosphorylated glycolytic metabolites in *L. lactis* MG5267 with *in vivo* NMR experiments, as well as using perchloric acid extracts obtained during the metabolism of glucose under anaerobic conditions. These measurements complement the NMR experiments and give a more comprehensive picture of the dynamics of PEP under anaerobic conditions in this organism. It was observed that the concentration of PEP in the beginning of the experiment was essentially equal to its final concentration. Within one minute into the experiment, the PEP concentration dropped to 1.2 mM and stayed low until glucose started to get depleted, upon which PEP started to accumulate.

L. lactis MG1363, which we model in this paper, shows a very similar dynamics to MG5267. The measured PEP dynamics [45] can thus be used to infer missing data points for PEP in our experiments. Figure 7B was plotted using the inferred PEP data. This plot, for the three experiments, shows that PTSf is not solely a function of PEP, because the trajectories from the three experiments would have to overlap, but they do not. Even if we attribute this discrepancy to noise in the data and possible inference errors, PTSf depends on PEP in such a fashion that a decrease in PEP results in a higher flux, which is not to be expected from a substrate.

Plots of PTSf after subtracting the expected effect of glucose (taken from literature; not shown here) also showed a decreasing behavior with increasing PEP. A possible explanation is that the dependence on PEP is rather weak and that simply a

minimal amount of PEP is required to keep the PTS flux running. Expressed within the conceptual framework of Michaelis-Menten kinetics, PTSf appears to have a very low Michaelis constant (K_m) for PEP. If so, the dependence of PTSf on PEP is essentially constant, and PTSf depends almost exclusively on glucose. However, plots of PTSf against glucose for the three datasets (Fig. 7A) demonstrate that PTSf is not a function only of glucose either because, again, the plots would have to overlap, but do not. The next sections describe potential means of resolving the issue.

3.2.1 NAD^+ May Regulate the PTS Flux

The results described above lead to the conclusion that PTSf depends on an additional variable. If we assume that PTSf depends on glucose in the form of a power-law function, it is beneficial to plot $\frac{PTSflux}{glucose^{h_{1,1}}}$ against all metabolites in the system, one at a time, and thus to determine if any of them can lead to overlapping trajectories. If so, it would make PTSf a true mathematical function of glucose and the candidate metabolite. In our case, this analysis was done for different representative values of $h_{1,1}$.

As the most pertinent illustration, consider NAD^+ as the candidate metabolite. We smoothed the NADH trajectories, using a so-called GS-function [44] which seems to offer good representations for Experiments 1 and 2, for which data are available (Fig. 7C). This rather unbiased black-box choice also naturally replaces low concentration values below the detection limit with non-zero values. Under anaerobic conditions, the NAD^+ dynamics is closely linked to NADH, because their sum is essentially constant. Using this fact, which is suggested by the biology of the system under anaerobic conditions, NAD^+ is easily inferred from NADH; the results are shown in Figure 7D. We then plotted $\frac{PTSflux}{glucose^{h_{1,1}}}$ *vs.* $\frac{NAD^+}{NADH}$, using a kinetic order $h_{1,1} = 0.14$ for glucose, which was the result of optimization. The results are depicted

inFigure 7E. With an appropriate scaling factor, they exhibit close to overlapping trajectories, which means that the representation of PTSf becomes a true mathematical function of glucose and $\frac{NAD^+}{NADH}$. In fact, screening all other metabolites in the system, we identified NAD^+, or alternatively the redox state of the system $\frac{NAD^+}{NADH}$, as one of only two feasible metabolites, the other one being 3PGA.

From a biological point of view, an effect of NAD^+ might be plausible, as the level of NAD^+/NADH is directly linked to the redox state of the system. While this conclusion is drawn directly from the data and without a specific mathematical model, the potential regulatory effect is a pure prediction that will require experimental validation. Nonetheless, according to the literature [19, 20], NAD^+ has been observed to activate the PTS system in *E. coli* by modulating the activity of the ATP-dependent Enzyme I-Kinase (EI-K), which reversibly phosphorylates Enzyme I at its active site histidine; thus ATP and EI-K can replace PEP. Whether the same mechanism is active in *L. lactis* is not known, but the ratios of NAD^+/NADH in datasets 2 and 3 start ramping down while FBP is being depleted, and the proposed NAD^+ regulation is a valid candidate for eliminating the discrepancy in PTSf between the experimental data and the model. Since our argument is purely mathematical, any variable outside the system that is strongly correlated with NAD^+ could be substituted as well. Whatever the case may be, the dynamics of PTSf involves more than glucose and PEP and needs to be further investigated with experimental means.

3.2.2 Possible inhibition by 3PGA

A compelling feature of DFE is its ability to predict the shape of the dynamic trend of a potential inhibitor throughout the experimental time period. The analysis of FBP (see below: Section 3.3.1) requires that glucose uptake follows a saturating function, such as a Michaelis-Menten function, with very low K_m (Fig. 8B). However, such a setting is not compatible with observations regarding the glucose uptake profile (Fig.

8A). Specifically, we can compare the glucose dynamics in the three datasets (Fig. 8A) with the glucose dynamics in the model. Here, we use a saturated Michaelis-Menten rate law with $K_m = 0.013\ mM$, as reported in [8], and a V_{max} that was set to the maximum flux among the three experiments and all time points, but does not actually affect the shape of the dynamic profile (Fig. 8B). The comparison allows us to infer the dynamic trend of the posited regulator for the three experiments (Fig. 8C). One candidate thus identified as exhibiting the correct trend is 3PGA. While the analysis does not provide proof that 3PGA affects glucose uptake, it is a reasonable candidate, because the molecule is structurally very similar to PEP. In addition, 3PGA has been observed to be a potent competitive inhibitor for the PTS reaction in *E. coli* and *Salmonella typhimurium* strains [52].

3.2.2.1 PTS Flux and 3PGA Dynamics under Aerobic Conditions

This section investigates the hypothesis that 3PGA could be an inhibitor of PTSf under aerobic conditions. So far, I showed that under anaerobic conditions, and after subtracting the effect of glucose from the PTS flux as characterized with parameters from the literature [8], the inferred effect is U-shaped over time (Figure 8C). This shape is similar to the dynamics of 3PGA.

To demonstrate the feasibility of 3PGA as an inhibitor, I use duplicate experimental data for 20 mM glucose input under aerobic conditions. Superimposing the plot of $\mathbf{PTSFglc^{h_{1,1}}}$ *vs.* 3PGA concentration and the plot of a typical inhibitor function, namely constant1/(1+constant2[3PGA]) *vs.* 3PGA concentration, demonstrates that the two have similar shapes, which indicates that 3PGA could well be an inhibitor (see Figure 7F).

Support for the role of 3PGA as potential inhibitor is indirectly provided by data in Table 2, which were excerpted from Table 1 of [46]. Enzyme activities are expressed in micromoles per minute per milligram of protein and are means ± standard

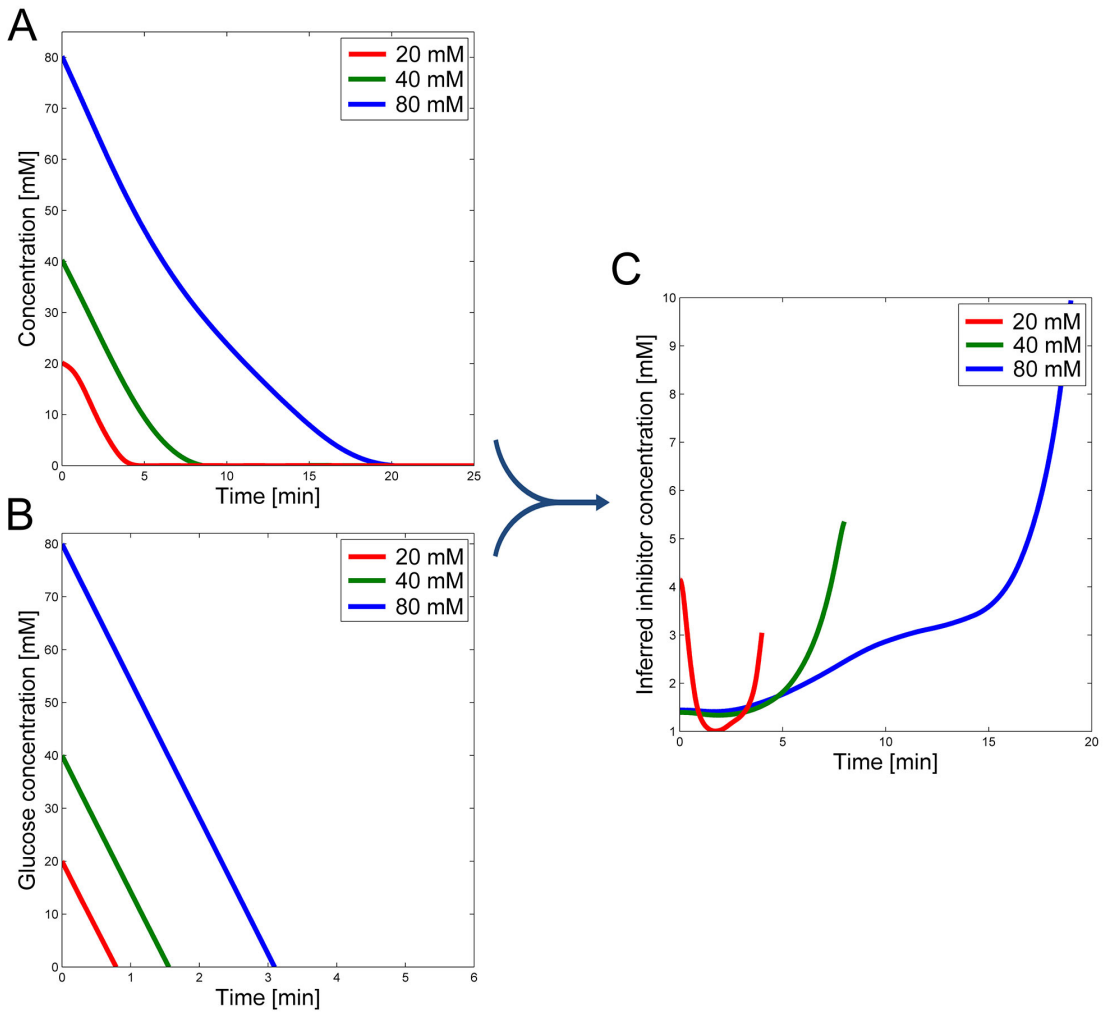

Figure 8: A: Glucose concentrations for the three experiments *vs.* time. B: Assuming a saturating Michaelis-Menten function with low K_m =0.013 mM for the PTS flux [8], the computed glucose concentrations are shown as functions of time. C: DFE analysis of the trends in panels A and B permits the prediction of the time trend of a postulated inhibitor.

Table 2: Comparison of enzyme specific activities in crude extracts of MG1363 cells grown under anaerobic and aerobic conditions.

Strain	Condition	Total NADH oxidase activity ($\mu mol.min^{-1}.mg$ of $protein^{-1}$)	Glucose consumption rate of ($\mu mol.min^{-1}.mg$ of $protein^{-1}$)	Concentration (mM) of 3PGA
MG1363	*Anaerobic*	0.07 ± 0.01	0.41 ± 0.01	8 ± 2
	Aerobic	0.22 ± 0.01	0.25 ± 0.03	33 ± 3
NOX^+	*Anaerobic*	16.5 ± 0.23	0.35 ± 0.02	10 ± 2
	Aerobic	17.0 ± 0.50	0.21 ± 0.02	36 ± 3
NOX^-	*Anaerobic*	0.03 ± 0.00	0.37 ± 0.03	3 ± 0.4
	Aerobic	0.03 ± 0.00	0.23 ± 0.04	25 ± 3

deviations ($n \geq 4$). The table shows that there is a negative correlation between glucose consumption rate and the concentration of 3PGA at the end of the experiment, which is assumed to be similar to its concentration at time zero where PTSf is nonzero. This observation applies to different strains of *L. lactis*, including wild type MG1363 grown under anaerobic and aerobic conditions, as well as strains where NADH oxidase is knocked out (NOX^-) or over-expressed (NOX^+). The assumption of 3PGA as an inhibitor of PTSf also appears to be reasonable judging by additional data, which were obtained *in vitro* in perchloric acid extracts during the metabolism of glucose under anaerobic conditions [45]. The data were obtained from *L. lactis* MG5267 and only for the initial glucose concentration of 20 mM. Similar concentrations for both 3PGA and PEP in the beginning and end of the experiment were reported. Of course, correlation does not prove causation, but these data are at the very least consistent with our hypothesis.

Since trends in both NADH and 3PGA show strong resemblance to the predicted trend of a putative inhibitor, I allowed both as potential candidates in the model until targeted experimental confirmation or refutation is available. To illustrate this conclusion, the dynamics of the hypothesized inhibitor in Figure 8C is replotted

in Figure 9, where it is overlaid with the actual trends in 3PGA (Panel A) and NADH (Panel B). As these panels indicate, there is good consistency between the hypothesized and postulated inhibitors.

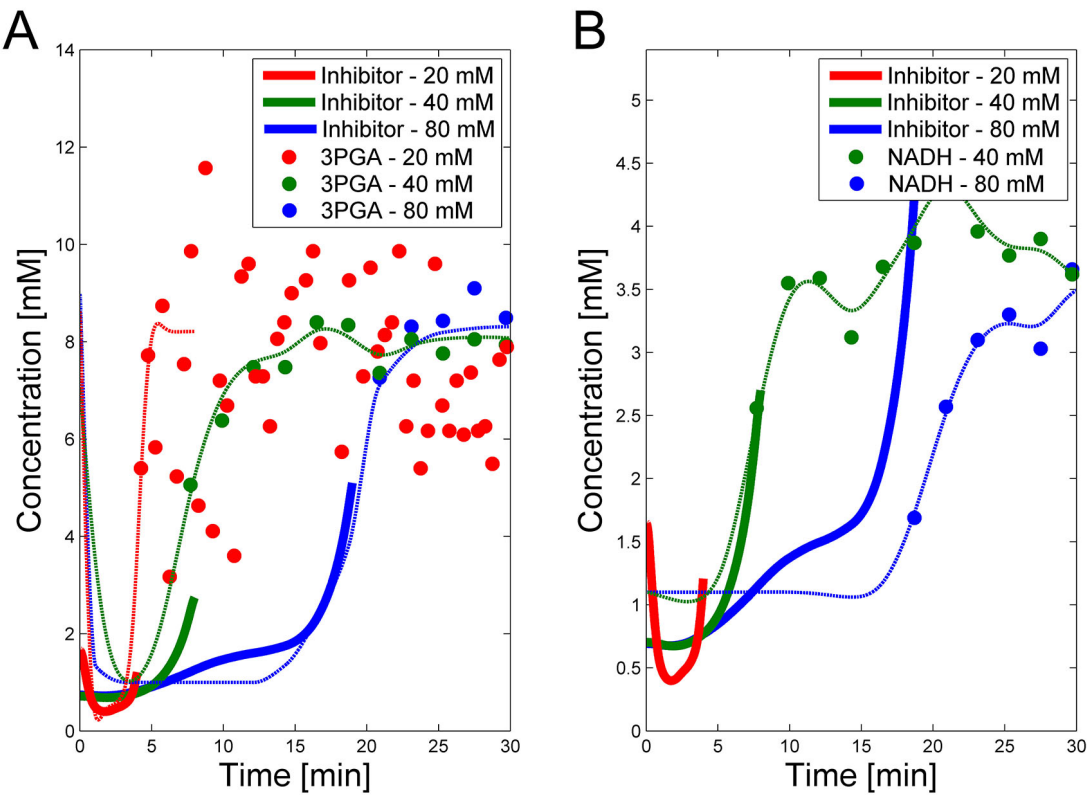

Figure 9: Panel A depicts the postulated inhibitor concentration trends (heavy lines; see Fig. 8C) *vs.* time for the three experiments, superimposed on measurements of 3PGA, scaled by 1.5 to emphasize the similarity. The dashed lines show the smoothed trends for 3PGA using the fact that 3PGA is non-zero in the beginning. Panel B shows the same putative trend in inhibitor but superimposed on NADH data scaled by 5. NADH data are only available for Experiments 2 and 3. Note that PTSf is zero once glucose is depleted, so that the inhibitor trend can only be inferred for the time points where glucose concentrations are nonzero. For these important time periods, both candidates seem feasible.

3.3 Quasi-steady FBP peak

An intriguing observation among the three experiments is that FBP reaches more or less the same peak level as long as a certain minimal level of glucose is available to the system (Fig. 5). This observation has been puzzling for a long time but can now be

explained with a detailed analysis of the model structure and the quantitative features of its fluxes. The mathematical and biological details of the explanation are discussed below. In a nutshell, the shapes of the peaks are driven by saturation of the PTS flux (v_1), while the subsequent production flux of FBP (v_3) operates substantially below saturation. The proposed relationship between v_1 and v_3 yields strong mathematical constraints that restrict the admissible parameter space. In particular, the analysis reveals that glucose must have a low kinetic order in the PTS flux.

3.3.1 Analysis of FBP Dynamics

To rationalize the counterintuitive peak dynamics of FBP, let us at first consider a slightly simplified scenario where G6P is omitted as an explicit intermediate between glucose (X_1) and FBP (X_2) and where the PTS flux v_1 solely depends on X_1 and v_2 on X_2. The dependence on NADH is not of importance here. This simplified diagram is shown in Figure 10A.

Figure 10B visualizes the dynamics of fluxes of accumulation (v_1) and consumption (v_2) of FBP relative to one another for the representative case of Experiment 3 with 80 mM of initial glucose. The horizontal top and bottom axes for the two curves are different and color-coded with blue (glucose) and black (FBP). In the first phase of the experiment, glucose is abundant and FBP accumulates (shown with a curved green upward arrow) while glucose is being consumed at a more or less constant rate (shown with the horizontal green arrow indicating that glucose is consumed at the same time as FBP accumulates). The intersection of the curves represents a quasi-steady state for FBP with a concentration of about 50 mM. At this state, accumulation and consumption of FBP are in balance: $v_1 = v_2$. As the experiment proceeds, FBP remains constant, whereas glucose is being consumed from about 50 mM down to about 10 mM (color-coded with straight orange arrow). Since the value of v_1 is essentially constant, FBP remains at its quasi-steady state during this period. As

soon as the glucose concentration drops below about 10 mM, FBP becomes depleted while glucose is being used up as the two red arrows indicate. Panel C shows the same three phases on a concentration *vs.* time plot of glucose and FBP for further clarification.

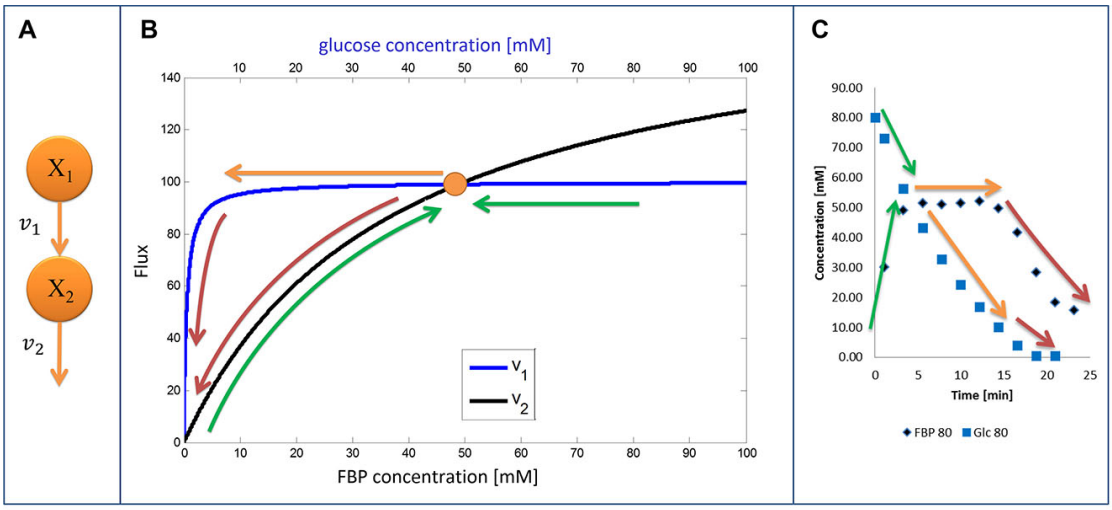

Figure 10: (A) Simplified diagram of glucose (X_1) conversion into FBP (X_2). Panel (B) depicts the relative shapes of v_1 and v_2 as functions of their substrate concentrations. Colored arrows show different phases of the dynamic behavior if FBP. Green: FBP accumulation (first 5 minutes);Orange: FBP constancy at peak level ($\tilde{1}$5 minutes); Red: FBP depletion. Panel (C) shows the same phases on a concentration *vs.* time plot. The color coding is consistent between B and C.

The concepts outlined above were implemented in a simple model. We start by representing the flux terms with Michaelis-Menten rate laws. In order to obtain the same peak for different amounts of glucose, v_1 needs to be saturated and equal to V_{max1} for glucose with a value between about 10 to 80 mM for Experiment 1 and between about 5 to 40 mM for Experiment 2. These numerical settings permit the quasi-steady state and extended peak for FBP if v_3 has a high $\mathrm{V_{max}}$ and a low K_m. In fact, these values must be such that for the peak FBP concentration (about 50 mM), v_2 equals v_1 for the appropriate glucose concentrations.

These quantitative considerations can be converted into constraints for the model

parameters. Specifically, the equation $v_2(FBP = 50) = v_1(5 \leq glucose \leq 50)$ mandates the following: First, K_{m1} must be such that v_1 is saturated for high glucose values (above about 5mM); this ensures a similar FBP peak for all pertinent glucose concentrations in the three experiments. Second, the FBP accumulation and consumption fluxes v_1 and v_2 need to intersect. This requires $K_{m2} > K_{m1}$ and $V_{max2} > V_{max1}$, and the values must be such that v_2 is equal to v_1 when FBP is at the observed quasi-steady state of about 50 mM. A solution to these constraints is:

$$V_{max1} = \frac{V_{max2} \cdot (FBPpeakconc.)}{k_{m2} + (FBPpeakconc.)}$$

An additional constraint can be derived from the speed with which glucose is depleted. This constraint pertains to V_{max1}. Taken together, we obtain the following set of conditions:

$$k_{m2} > k_{m1} \tag{8a}$$

$$V_{max2} > V_{max1} \tag{8b}$$

$$V_{max1} = \frac{V_{max2} \cdot (FBPpeakconc.)}{k_{m2} + (FBPpeakconc.)} \tag{8c}$$

$$V_{max1} \approx 4 \tag{8d}$$

Figure 11A shows the glucose and FBP dynamics resulting from a representative example for which the inferred sets of constraints hold. The concentration curves are very similar to the measurement data (Fig 5).

The relationships between fluxes are not dependent on the choice of the Michaelis-Menten framework. To reproduce similar conditions for power-law representations, we must simply require a low kinetic order and rate constant for glucose that results in the correct dynamics for v_1. The result is a similar flux value for different glucose concentrations between 5 and 50 mM. Furthermore, v_2 needs to have a higher kinetic order so that the two flux curves intersect. Rate constants need to be calculated

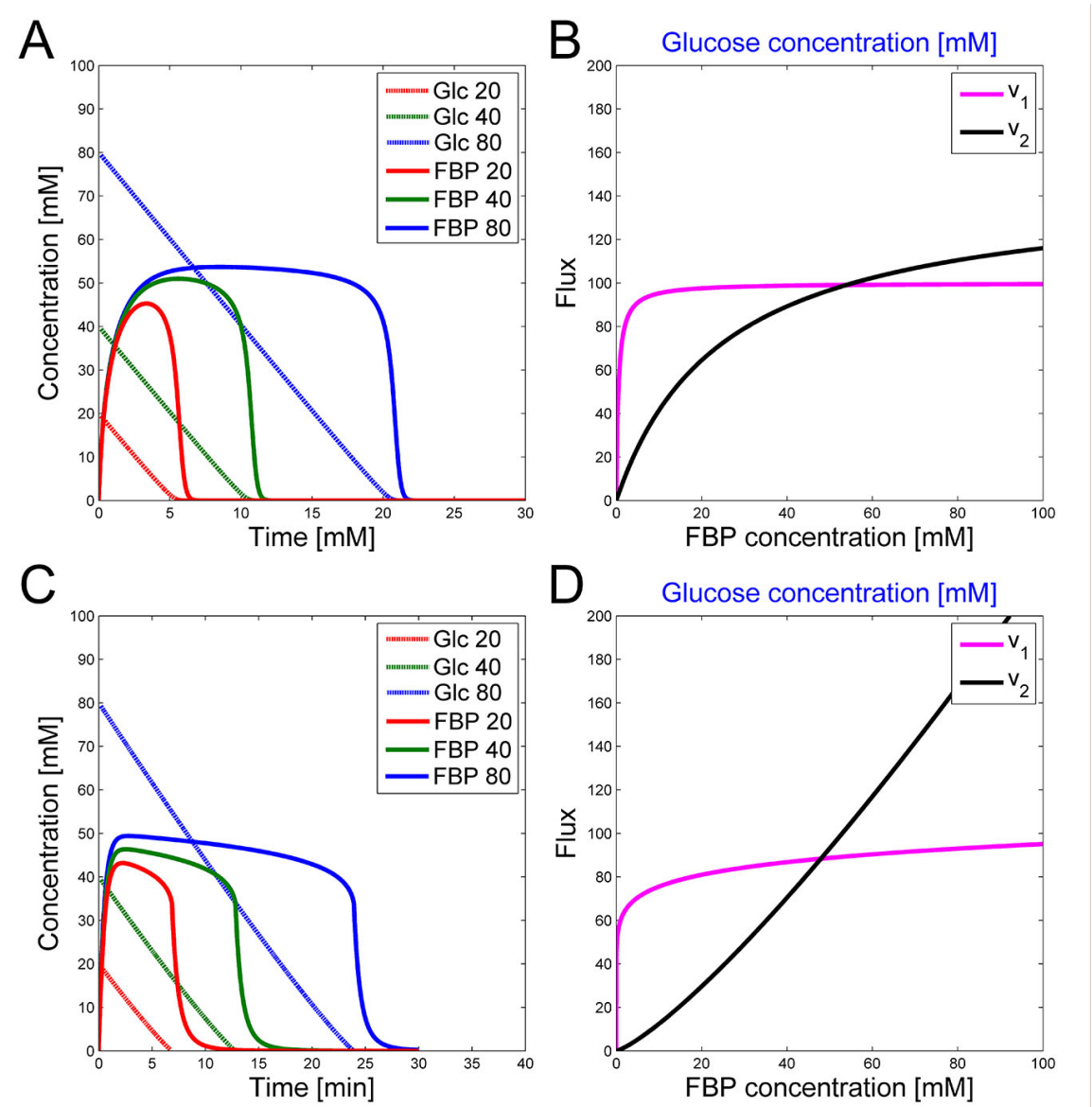

Figure 11: (A) Computed concentrations of glucose (blue) and FBP (green) for the 20, 40, and 80 mM of initial glucose, using the constraints on parameters in a Michaelis-Menten formulation with representative values of $V_{max1} = 4$, $V_{max2} = 145$, $K_{m1} = 0.5$, $K_{m2} = 25$ and an input to output volume ratio of 25. (B) Corresponding glucose and FBP trends *vs.* time for the Michaelis-Menten formulation in (A). (C) Plots of flux *vs.* concentration for the v_1 and v_2 fluxes in power-law format. v_1 has a low kinetic order. (D) Corresponding glucose and FBP trends *vs.* time for the power-law formula in (C). Experiment 1 with the highest amount of glucose input results in the most extended FBP peak.

such that $v_1 = v_2$ for the peak FBP concentration of about 50 mM. Similar to the Michaelis-Menten case, the kinetic order and rate constant for v_1 need to be set such that they satisfy the speed for glucose depletion. Results for a representative, feasible parameter set with power-law functions are illustrated with the fluxes in Figure 11C. Figure 11D shows the corresponding glucose and FBP concentrations for the three experiments with different initial glucose concentrations.

3.3.1.1 Including G6P in the analysis

The question arises of whether the above considerations are affected by the existence of G6P as an intermediate between glucose and FBP. The short answer is *no*.

Unfortunately, the [1-^{13}C] NMR experiments did not permit measurements of G6P for the three experiments. However, it is likely that the G6P dynamics is similar to the FBP dynamics, although with a lower peak value or a very low saturation threshold as it is observed for Experiment 1 with initial glucose of 20 mM.

Supposing that v_3 is a function of FBP only, we find: (i) at times when FBP is constant, v_3 is constant as well; and (ii) FBP being constant requires that $v_2 = v_3$. (i) and (ii) are possible in the following two scenarios only when G6P and FBP are constant simultaneously, or when v_2 is saturated for the amount of G6P during that time period, and therefore has a very low K_m.

Furthermore, no matter whether G6P is constant or has a low K_m , $v_1 = v_2 = v_3$ must hold at the FBP peak, and v_1 must thus be saturated for glucose values between about 5 and 50 mM as reasoned before. Thus, the earlier conclusions regarding the FBP dynamics hold, whether or not G6P is explicitly presented.

3.4 Regulation of Glycolysis

The model contains four feedforward mechanisms that were identified in the literature as potential controllers of the glycolytic pathway in *L. lactis* [13, 14, 17, 15, 16, 28, 45, 66]. Model simulations with and without these regulatory signals confirm and explain

the importance of their roles. Highlights of this part of the analysis are reported in the following sections.

3.4.1 G6P and FBP activate PK

The critical enzyme pyruvate kinase (PK) is known to be regulated by G6P and FBP [13, 14, 17, 16, 60]. The feedforward activation of PK by G6P and FBP enables the accumulation of 3PGA and PEP when glucose becomes depleted (see Fig. 1 of Chapter 2). Although the accumulation of intermediate metabolites in linear pathways is generally considered disadvantageous, it here is obligatory for the cell to maintain a relatively high concentration of PEP, because it allows the organism to restart glycolysis through the PTS as soon as new glucose becomes available after starvation [66]. Also, the accumulation of G6P and its activation of PK use PEP, which indirectly leads to the production of ATP, which in turn is needed for the conversion of G6P into FBP. The activation of PK furthermore leads to the fast production of pyruvate and then lactate, which is advantageous for the organism, as it sours the medium and prevents or at least impedes the growth of competing species. In an analysis of the pathway under aerobic conditions, it was previously decided to use FBP as the sole, representative activator, because measurements of G6P were scarce, and the activation by FBP was sufficient [66]. Indeed, under aerobic conditions and a relatively low glucose supply (20 mM), FBP vanishes completely, together with G6P, and is therefore able to shut down PK effectively. In anaerobic experiments, by contrast, FBP is never completely exhausted (Fig. 5), but retains a residual concentration, even long after glucose is depleted. Thus, FBP cannot complete shut off pyruvate production. Because most upper glycolytic intermediates are activators of PK [45], we used the corresponding model variables, FBP and G6P, both as regulators. This choice is effective but raises the secondary question of how and why the cell retains FBP. We discuss this question in a later section.

3.4.2 PEP inhibits LDH

Under aerobic conditions, the concentrations of NAD^+ and NADH remain essentially constant, and conversions between the two are rapid [48]. However, in the absence of oxygen, oxidation of NADH to NAD^+ by NADH oxidases (NOXs) does not occur, because the step requires oxygen. Meanwhile, NAD^+ is being consumed by the GAPDH reaction. Thus, in contrast to aerobic oxidation, the cell under anaerobic conditions critically depends on the recycling of NADH to NAD^+ through the lactate dehydrogenase (LDH) reaction, which now plays a limiting role.

3.4.3 Ready to Respond

The limiting role of LDH, together with a reduced activation of PK, leads to a chain of events that ultimately stops glycolysis in a ready-to-respond state where the organism is perfectly positioned to restart glycolysis immediately once glucose becomes available. This chain of events is depicted in Figure 12. In detail, our model suggests the following. Very shortly after glucose runs out, the G6P concentration decreases to zero. As a consequence, G6P and FBP no longer activate PK, which causes PEP to accumulate to a relatively high concentration. The persistently high concentration of PEP leads to strong inhibition of LDH. As a consequence, the production of NAD^+ ceases, NAD^+ depletes, the redox state is altered, and the GAPDH reaction (v_3) can no longer proceed without the cofactor. The result of this process is a residual amount of FBP, even after external glucose is exhausted, which is not observed under aerobic conditions. One notes that the retention of sufficient residual amounts of FBP and PEP at the end of the experiment requires well-coordinated regulation, because the v_3 and PK fluxes must be shut down at just the right time.

3.4.4 Restarting glycolysis after starvation

As explained elsewhere [72], PEP is needed to provide the phosphate moiety for the phosphorylation of glucose. Also, since the experiment starts without ATP, as shown

in Experiments 2 and 3, some initial amount of ATP is needed for the PFK reaction to go forward and thus to start glycolysis. With the initial influx through PTS, G6P and F6P start to accumulate, which leads to the activation of PK. This in turn provides the initial ATP needed for PFK to proceed. By retaining PEP, the cell is ready to take up glucose once it becomes available after starvation, both under aerobic and anaerobic conditions. However, specifically under anaerobic conditions, once glucose is depleted, some FBP is retained (Fig. 6B).

The residual amount of FBP is the result of v_3 shutting off due to lack of NAD^+ recycling by v_5, as discussed above. This component of the shut-down strategy contrasts the aerobic conditions where v_3 is not entirely turned off and all of the remaining FBP is converted into PEP (Fig. 6C). Retaining some FBP under anaerobic conditions furthermore provides the cells with a higher amount of ATP through the phosphoglycerate kinase reaction, which could be used as a secondary mechanism to restart glycolysis. Indeed, this mechanism might explain why glucose uptake is faster under anaerobic conditions (Fig. 6A). Additionally, the lower 3PGA concentration at the beginning of the experiment may contribute to the faster anaerobic metabolism of glucose.

3.4.5 ATP dynamics is indirectly affected by glucose availability

DFE allows us to infer the dynamics of flux v_6, which converts ATP into ADP and Pi. Because the first phase of DFE is model-free and assumption-free, this flux estimate is only minimally biased, if at all. Interestingly, the results indicate that no monotonic function that depends exclusively on ATP as its substrate, including power-law and Michaelis-Menten functions, can provide a good fit to the observed dynamics. In turn, this analysis strongly suggests that another variable must modulate this flux. The only variable in the model that has the right time profile to resolve the problem is glucose. Of course, glucose is located outside the cell, and thus unavailable to regulate

intracellular processes directly. However, it is likely that cells have glucose sensors that could allow the modulation of ATP usage. A candidate for this role is the component EIIA-P of the PTS system. Since, the EIIA-P concentration is dependent on the flux of glucose uptake and not on the external glucose, we included v_1 as the regulator for ATP usage (v_6). This data-driven hypothesis, which was obtained independently from the model, is very interesting in its biological implication, as it suggests that some or all of the ATP consuming processes in the cell are regulated by glucose availability. If so, the cells appear to be able to selectively shut off some of these ATP utilizing processes when glucose is scarce. To the best of our knowledge, this type of regulation has not yet been reported and seems worthy of further experimental investigation.

3.4.6 Comparison of aerobic and anaerobic operation of glycolysis

The model facilitates an explanation of the differences in the dynamic profiles of the metabolites of the pathway under aerobic *vs.* anaerobic conditions. The main difference affects Step 4 in the two panels of Figure 12. Under aerobic conditions, NAD^+ recycling is not a limitation for the GAPDH reaction, because oxygen is available for NADH oxidase to produce sufficient quantities of NAD^+. As a consequence, the NAD^+ concentration remains more or less constant and certainly never drops much. Therefore, v_3 is not shut off under aerobic conditions, and the cell continues to consume FBP until it is entirely used up. The consumption of FBP, in turn, produces more PEP and 3PGA, thus explaining the difference in Panels C and D of Figure 6. An additional possible explanation for the stronger accumulation of PEP and 3PGA under aerobic conditions appears to be the timing of the pathway events, which is explicitly visible in the speed of glucose uptake.

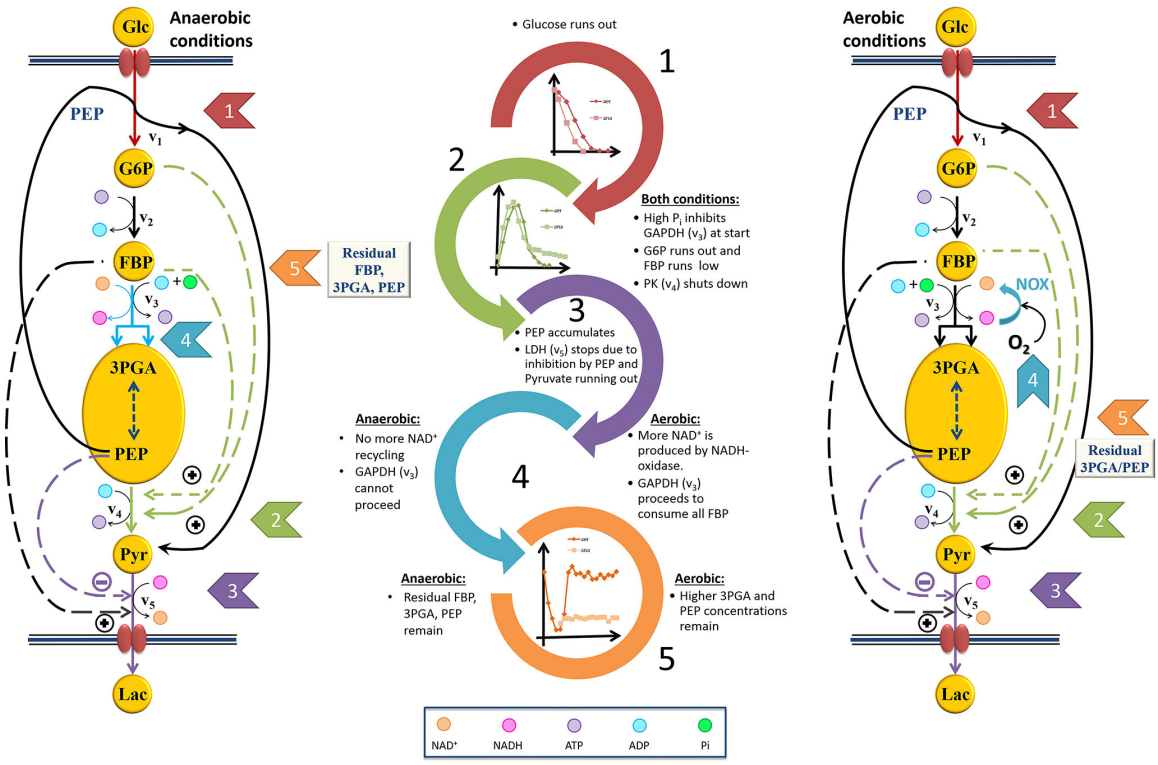

Figure 12: Cascades of events resulting in a well-coordinated system shut-down under anaerobic and aerobic conditions. The shut-down leads to different ready-to-respond states under the two conditions. The chain of events upon glucose depletion rationalizes why some residual FBP is left at the end of the experiment (see Fig. 5), but only under anaerobic conditions. In both conditions, the cells retain some 3PGA and PEP when glucose runs out. Some fluxes, which are not directly pertinent to the shut-down process, are omitted from this figure.

3.5 Discussion

High-throughput dynamic data contain very valuable information about the function and regulation of cellular systems *in vivo*. This is especially the case for time series of concentrations that can be generated with *in vivo* NMR techniques. Here, we analyzed such data in the context of intricate survival strategies with which *L. lactis* stops and restarts glycolysis when glucose substrate is about to run out or becomes available again.

For the first time, the model explains several observations that were made over the past several decades. In particular, it captures the unusual and puzzling FBP accumulation dynamics across different glucose inputs and offers an explanation for the differences in the dynamic behavior of the pathway under aerobic *vs.* anaerobic conditions. These differences are related to the fact that the dynamics of NAD^+ and NADH under anaerobic conditions is only affected by the activity of the glycolytic pathway and the lactate dehydrogenase step, which makes the control task more difficult than for aerobic conditions, where NADH is easily oxidized. The model also posits new, experimentally testable hypotheses regarding the regulation of critical steps in the pathway. Importantly, it reveals in detail the finely tuned timing of events that lead to an orderly shutdown of the pathway when glucose in the medium becomes depleted and to a metabolic resting state in which the cells are perfectly positioned to utilize new glucose as soon as it becomes available.

Detailed model analysis revealed several differences in metabolic time trends between operation under aerobic and anaerobic conditions. Some of these differences are quite subtle but, surprisingly, emerged as crucial. For example, FBP is consumed more slowly under anaerobic conditions and retained at a residual concentration for a long time. This observation is not new, but was by and large ignored, as it was seen as coincidence or experimental noise. We demonstrate here, arguably for the first time, why this residual amount is necessary: It directly affects the activity of the

PK reaction, as well as the generation of ATP and is therefore very beneficial, if not obligatory, for restarting glycolysis after a period of starvation.

A second set of observed differences between aerobic and anaerobic operation is a faster consumption of glucose and a faster accumulation of lactate in the latter case. Also, NAD^+ dips down at about the same time when PEP and 3PGA peak, because NADH oxidase is not operational without oxygen, while it stays more or less constant under aerobic conditions. This effect is an important contributor to the control of glycolysis after starvation. In particular, the difference in NAD^+ constitutes a limitation for the GAPDH reaction, which leads to the accumulation of FBP under anaerobic, but not under aerobic conditions, where FBP is quickly depleted when no more glucose is available. The result is a higher concentration of PEP and PGA.

The model provides a likely explanation for the slower glucose consumption under aerobic conditions. Namely, glucose uptake must depend on more factors than its two substrates, glucose and PEP. Detailed analysis identified 3PGA, as well as the state of the redox system, as viable candidates for regulating this step. At this point, the role of these components as inhibitors is purely speculative. Their inhibiting effect has been documented in other bacteria, but it is yet to be confirmed with laboratory experiments in *L. lactis*.

Taken together, the details of the dynamics of glycolysis in *L. lactis* portray an intriguing and finely tuned system of signals regulating the pathway. The structure of this control system emerged through a suitable representation in a mathematical model, based on time series data.

CHAPTER IV

EXTENSIONS OF DFE

As introduced in Chapter 2, DFE was proposed as a minimally biased identification method for metabolic pathway models [31]. Among other features, this method tends to prevent the compensation of errors among flux terms. A hallmark feature of DFE is that it does not presuppose a functional form for the flux representations. This feature allows us to test in an objective manner if particular functions, such as power-laws or Michaelis-Menten rate laws, are capable of appropriately modeling each specific flux, or if other formulations should be considered. In particular, analyses of DFE results may suggest the likely existence of regulatory signals that had been missing from the assumed pathway structure. Such a suggestion corresponds to a novel hypothesis that is testable with further experiments and may lead to biological discoveries, as was demonstrated in Chapter 3.

The main drawback of DFE is the fact that it applies directly only to systems that contain as many independent fluxes as metabolites; in other words, when the stoichiometric matrix has full rank. If this property does not hold, DFE cannot be executed directly. To circumvent this obstacle, auxiliary methods for independently determining some of the independent fluxes in certain scenarios have been proposed, but they are only effective in specific situations and often cumbersome (*e.g.*, [36, 76, 10]).

This chapter consists of two main sections. The first develops extensions to the model-free phase of DFE for underdetermined pathway systems, while the second section suggests strategies for dealing with missing data and proposes mixed parameter estimation strategies when DFE is only partially applicable. This second section

involves the second, model-based phase of DFE and use the *Lactococcus* model as a representative example.

4.1 Characterization of Metabolic Fluxes from Time Series Data

The purpose of this section is to assess DFE for underdetermined pathway systems, describe the space of feasible solutions, and suggest systematic ways to select from among these solutions the most likely candidates. These should consist of non-negative fluxes that are continuous over time and satisfy certain other conditions, which will be discussed later. Along with the exploration of this space, whose dimension equals the degrees of freedom (DOF) of the problem at hand, useful strategies will be introduced to visualize feasible candidate sets. Initially, no information about the functional forms and the contributing metabolites and modulators of each flux is assumed to be available. Later on, minimal generic features of metabolic fluxes are used as constraints to improve the results. It is noted, though, that, even with these constraints, the solutions are not necessarily unique. Finally, solutions in the form of point-wise numerically defined fluxes will be suggested that are appropriate, if not optimal, according to certain criteria of biological reasonableness.

4.1.1 Mathematical Formulation of the Problem

A dynamic representation of a metabolic pathway system is formulated in Eq. 9 in general matrix and vector notation:

$$\frac{d\mathbf{X}}{dt} = \dot{\mathbf{X}} = A \cdot \mathbf{v} \tag{9}$$

Here, $\mathbf{X}$ denotes a vector of n metabolite concentrations and $\mathbf{v}$ is a vector of m fluxes, *i.e.* reaction rates, while A is the stoichiometric matrix. The vectors change with time, and the functional forms governing the fluxes are functions of their substrates and regulators. They are in general unknown or based on assumptions

that might or might not hold under the given experimental conditions or *in vivo*. Moreover, in certain cases, regulators and cofactors are yet to be discovered and are therefore falsely omitted. This uncertainty is the reason to use minimal assumptions while executing the task of inferring flux profiles from metabolic time series data. At the same time, DFE provides us in this phase with the option of testing and challenging prior assumptions and possibly discovering missing regulatory effects.

Assuming that data smoothing and slope estimation had been conducted at each time point t_i, we replace the left-hand side of Eq. 9 with the vector of slopes at time t_i, which we call $\mathbf{b}(t_i)$. Eq. 9 can thus be written as a set of algebraic equations. Specifically, suppose that $\mathbf{b}(t) = [\dot{X}_1(t), \cdots, \dot{X}_n(t)]^T$ is the vector of slopes of dependent variables at time t and A is the $n \times m$ stoichiometric matrix, which is time independent. Then we obtain directly the linear algebraic system:

$$A \cdot \mathbf{v}(t) = \mathbf{b}(t) \tag{10}$$

At a steady state, or when the numerical values of the derivatives are known, Eq. 10 has a solution that can be computed for every time point by matrix inversion, if the system has full rank. However, most metabolic systems are under-determined, so that a unique solution does not exist.

We can thus distinguish three situations. (1) When the system has maximal rank, the solution is obtained with the regular inverse, so that $\mathbf{v}(t_i) = A^{-1}\mathbf{b}(t_i)$ is the solution of the system of equations. (2) When the system is over-determined and has more equations than unknowns ($m < n$), the Moore-Penrose pseudo-inverse A^+ of matrix A minimizes the sum of squared errors, $\underset{\mathbf{v}}{\operatorname{argmin}} \|A\mathbf{v}(t_i) - \mathbf{b}(t_i)\| = A^+\mathbf{b}(t_i)$. This solution is equivalent to the result of linear regression. Finally, (3), the case of under-determined systems ($m > n$) is the most common situation in metabolic modeling, because most pathway systems contain more reaction steps than metabolites. This common occurrence makes the under-determined case particularly important for the model-free phase of DFE and suggests that we investigate if the

pseudo-inverse solution $\mathbf{v}(t_i) = A^+\mathbf{b}(t_i)$ constitutes a biologically feasible, or even optimal, solution.

Pseudo-inverses have been used to solve under-determined systems for a long time. They are characterized by the minimum L^2-norm within a one- or higher-dimensional space of admissible solutions, *i.e.* $\underset{\mathbf{v}}{\operatorname{argmin}}\|A\mathbf{v}(t_i)-\mathbf{b}(t_i)\|$. While the best solution, in terms of the smallest norm, is guaranteed by the pseudo-inverse, the resulting fluxes are not necessarily positive, and there is no guarantee that they are smooth over time and biologically meaningful, let alone optimal. In fact, experience shows that minimum-norm solutions often include negative values, which are not biologically feasible as flux values. The issue of under-determined systems in DFE has been known since the inception of the method, and characterizability analysis, based on pseudo-inverses, was introduced as an a priori check for the applicability of DFE given a particular pathway system [70].

4.1.2 Compact Representation: Gamma-space and Gamma-trajectory

In order to characterize the space of admissible flux sets $\mathbf{v}(t) = [v_1(t), \cdots, v_m(t)]^T, t \in [0, \inf)$ in an efficient manner, we need a more compact representation. For pathways with m fluxes and n dependent variables, where $m > n$, let d be the number of degrees of freedom (DOF): $d \geq m - n$. At each time point t, the space of solutions satisfying Eq. 10 can be written as:

$$\mathbf{v}(t) = A^+\mathbf{b}(t) + (I - A^+A)\mathbf{w}(t) = A^+\mathbf{b}(t) + null(A)\boldsymbol{\gamma}(t) \tag{11}$$

Here, $A^+ = A^T(AA^T)^{-1}$ is the Moore-Penrose pseudo-inverse and $A^+\mathbf{b}(t)$ is the minimum-norm flux set at time t, which is not necessarily non-negative and in fact often results in one or more negative fluxes for some time points. Furthermore, if $\mathbf{w}(t_i)$ is a vector of m arbitrary, real-valued elements, then the complete solution $\mathbf{v}(t_i) = A^+\mathbf{b}(t_i) + (I - A^+A)\mathbf{w}(t_i)$ represents all possible solutions, and spans the null space of the stoichiometric matrix A.

Expressed differently, the columns of $null(A) = [\mathbf{vec}_1, \mathbf{vec}_2, \cdots, \mathbf{vec}_d]$ span the null space of A, and $\boldsymbol{\gamma}(t) = [\gamma_1(t), \gamma_2(t), \cdots, \gamma_d(t)]^T$ is the corresponding vector of coefficients at time t. Each feasible solution of Eq. 10 at time t can thus be uniquely represented by $\boldsymbol{\gamma}(t)$. This representation allows us to explore the d-dimensional "Gamma-space instead of the feasible subset of the m-dimensional space of fluxes, whose visual representation is much more challenging.

For each time point t, the Gamma coefficients for a feasible flux set $\mathbf{v}(t)$ can be calculated by finding coefficients that satisfy $\mathbf{v_{null}}(t) = null(A)\,\boldsymbol{\gamma}(t) = [v_1(t), \cdots, v_m(t)]^T - A^{+}\mathbf{b}(t)$. This equation can be assessed by projecting $\mathbf{v_{null}}(t)$ onto the vectors $\mathbf{vec}_1$, $\mathbf{vec}_2$, $\cdots$, $\mathbf{vec}_d$, which together span the null space of A. The vector $[\gamma_1(t), \gamma_2(t), \cdots, \gamma_d(t)]^T$ constitutes a point in the d-dimensional Gamma-space, representing time point t. Over time, these points constitute a trajectory, which I call the "Gamma-trajectory". Each Gamma-trajectory represents a flux set traversing all time points, as long as this trajectory corresponds exclusively to non-negative fluxes.

The illustration example of Figure 13A shows a simple network consisting of two dependent variables and four fluxes. Suppose that metabolite concentrations $X_1(t)$ and $X_2(t)$ have been measured every 0.5 minutes between 0 and 15 minutes. Finding the slopes of the concentration trends directly yields $b_1(t)$ and $b_2(t)$ (Figure 13B). The feasible space of solutions, in terms of fluxes, is a two-dimensional plane within a 4-dimensional space, which is difficult to visualize directly. Figure 13C shows some representative flux solutions. Even though they are very different, and several of them have in fact little similarity to the fluxes in the model used to generate the "data" (black curves in Fig. 13C), all these fluxes satisfy Eq. 12 exactly. The corresponding Gamma-trajectories are depicted in Panel D of Figure 13. The fluxes and Gamma-trajectory with which the concentration data were originally generated are shown in black in Panels C and D.

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \\ v_4(t) \end{bmatrix} = \begin{bmatrix} b_1(t) \\ b_2(t) \end{bmatrix} \tag{12}$$

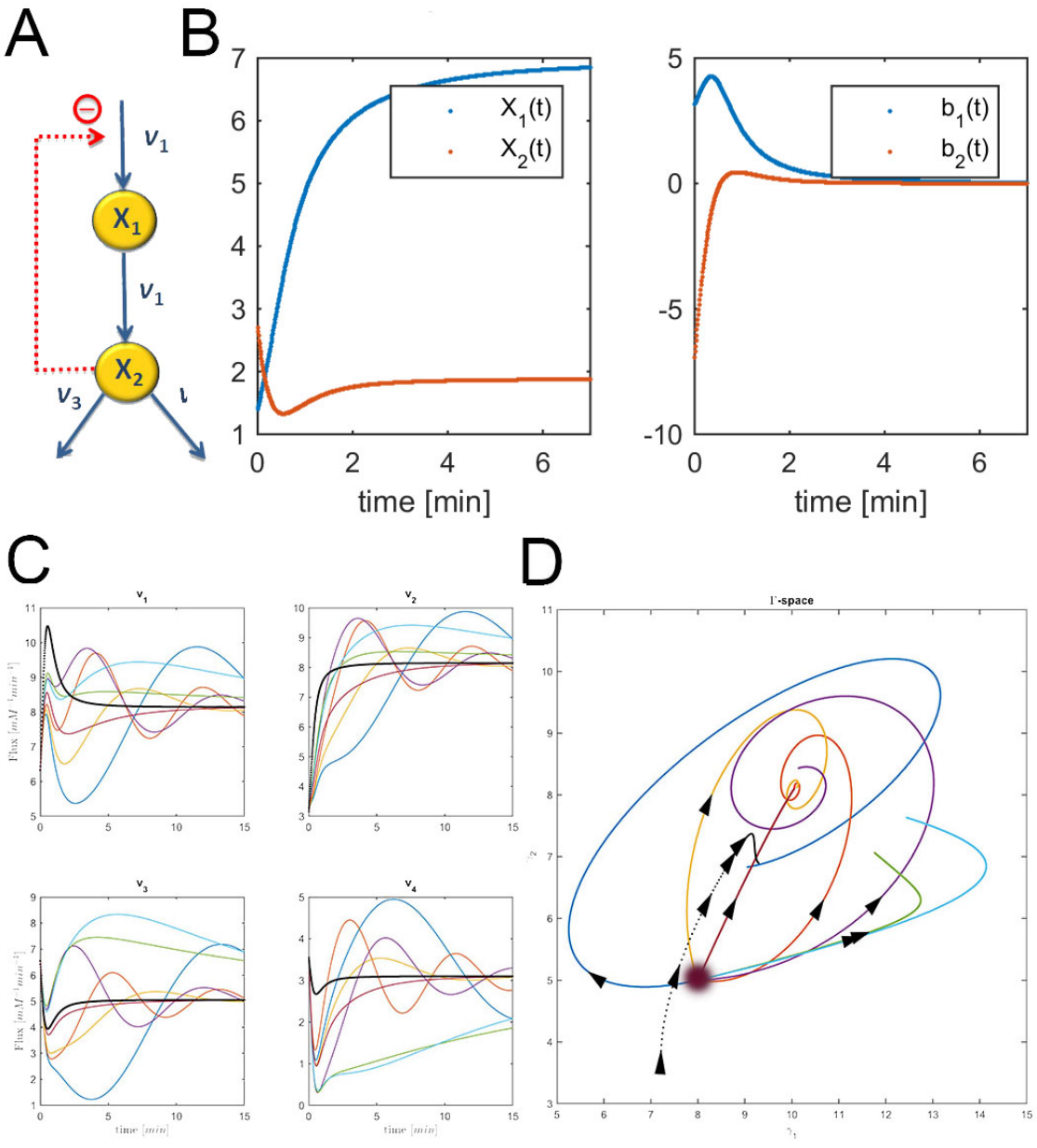

Figure 13: Illustration example used to demonstrate the core concepts of the flux

characterization procedure. The pathway has a simple structure as depicted in Panel (A). Panel (B) shows $X_1(t)$ and $X_2(t)$ on the left and the slopes of $X_1(t)$ and $X_2(t)$ estimated from noise-free measurements on the right. Panel (C) shows 7 examples of flux sets exactly satisfying Eq. 12; for this illustration, all start at the same point, shown with a magenta circle, as the original flux set ($\mathbf{v}(0) = [6.3271, 3.1588, 6.5486, 3.5486]$ corresponding to $\boldsymbol{\gamma}(0)^T = [8, 5]$). The thicker black curves are the fluxes with which the original data were produced. The corresponding Gamma-trajectories are depicted with the same color code in Panel (D).

The solutions in Figure 13 are among the infinitely many admissible solutions generated by the following procedure. A subset of these solutions can be generated in the following manner. Starting at some initial point in the Gamma-space, a phase-plane trajectory is computed according to a linear state-space model $\dot{\boldsymbol{\gamma}}(t) = B\boldsymbol{\gamma}(t)$. One should note that this is certainly not the only strategy for creating flux sets, but it is among the simplest options that lead to continuous fluxes. A Monte-Carlo approach is utilized, in which a 2×2 matrix B is randomly generated, but where only those B are retained that have negative real eigenvalues and result in non-negative fluxes for all time points. These resulting trajectories yield many different dynamical characteristics for the fluxes. Figure 13C shows some feasible solutions for fluxes v_1 through v_4 in multiple colors in thin lines, superimposed on the flux of the actual model, from which the concentration data were generated. These fluxes are shifted in Panel C, so that their initial values match, in order to facilitate easier comparisons. Interestingly, the resulting fluxes can possess behaviors ranging from simple shoulder curves to oscillatory responses.

4.1.3 Admissible Subset of Gamma-space: the Subspace of Non-negative Fluxes

For each time point t, let us determine the set of $\boldsymbol{\gamma}$'s for which the corresponding vector $\mathbf{v}(t)$ consists of non-negative values for all fluxes and all times. According to

Eq. 11, the feasible space given by $\mathbf{v}(t) = A^{+}\mathbf{b}(t) + null(A)\boldsymbol{\gamma} > 0$, is an intersection of m half-spaces characterized by Eq. 13:

$$A^{+}(i,:)\mathbf{b}(t) + \gamma_1 vec_{1,i} + \cdots + \gamma_d vec_{d,i} \geq 0, \quad i = 1, 2, \cdots, m \tag{13}$$

Here, $A^{+}(i,:)$ denotes the i^{th} row of the m by n Moore-Penrose pseudo-inverse matrix. The inequalities are linear and thus constitute a bounded or unbounded polytope.

4.1.4 Formulating the Problem as an Optimization Task

A biologically relevant constraint for the selection of meaningful flux profiles is the overall minimization of flux magnitudes, which might be interpreted as a form of metabolic energy minimization. Since the non-negativity constraints are already in place, this sum of fluxes at all time points equals the so-called minimum L^1- or Manhattan- norm, which is defined as $\min\limits_{\substack{\mathbf{v}>0 \\ A\mathbf{v}=\mathbf{b}}} \|\mathbf{v}\|_1 = \min\limits_{\substack{\mathbf{v}>0 \\ A\mathbf{v}=\mathbf{b}}} \sum_{i=1}^{m} |v_i| = \min\limits_{\substack{\mathbf{v}>0 \\ A\mathbf{v}=\mathbf{b}}} \sum_{i=1}^{m} v_i$. The optimization problem leading to this result in terms of $\boldsymbol{\gamma}$ is shown in Eq. 14. The constraint $A\mathbf{v} = \mathbf{b}$ is already taken into account, since the representation in Eq. 11 only allows for fluxes that satisfy this constraint. Thus, the optimization simplifies to:

$$\begin{aligned} \min_{A^{+}\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} & \sum_{i=1}^{m} A^{+}\mathbf{b}(t) + null(A)\boldsymbol{\gamma}(t) \\ &= \min_{A^{+}\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} \sum_{i=1}^{m} null(A)\boldsymbol{\gamma}(t) \end{aligned} \tag{14}$$

Eq. 14 shows that the linear program can be translated into a simpler linear program in terms of $\boldsymbol{\gamma}(t)$, which can be solved using algorithms for linear programming, such as the simplex method. In practice, testing the corner points of the feasible polyhedron for identifying the corner with the minimum sum is a very well established way of arriving at the optimal solution [21].

An alternative optimization approach for minimizing the sum of squared flux norms for all time points, *i.e.* the L^2-norm of the flux vector at each point in time, is depicted in Equation 15. Again, it represents in some sense the minimum-energy flux set.

$$\min_{\substack{\mathbf{v}>0 \\ A\mathbf{v}=\mathbf{b}}} \|v\|_2^2 \tag{15}$$

The optimization problem in Equation 15 can be reformulated as the optimization problem of minimizing the L^2-norm of the vector $\boldsymbol{\gamma}(t)$. Equation 16 shows this reformulation.

$$\begin{aligned}
&\min_{A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} (A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t))^T(A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)) = \\
&\min_{A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} (A^+\mathbf{b}(t))^T A^+\mathbf{b}(t) + \boldsymbol{\gamma}(t)^T null(A)^T null(A)\boldsymbol{\gamma}(t) \\
&\qquad + \boldsymbol{\gamma}(t)^T null(A)^T A^+\mathbf{b}(t) + (A^+\mathbf{b}(t))^T null(A)^T null(A)\boldsymbol{\gamma}(t) = \\
&\min_{A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} \boldsymbol{\gamma}(t)^T \mathbb{I}_m \boldsymbol{\gamma}(t) \\
&\min_{A^+\mathbf{b}(t)+null(A)\boldsymbol{\gamma}(t)\geq 0} \|\boldsymbol{\gamma}(t)\|^2
\end{aligned} \tag{16}$$

Here, $null(A)^T null(A) = \mathbb{I}_m$ is the identity matrix of dimension m, because the columns of $null(A)$ are orthonormal base vectors of the null space. Furthermore, the pseudo-inverse solution $A^+\mathbf{b}(t)$ is orthogonal to the null space, and thus $null(A)^T A^+\mathbf{b}(t) = (A^+\mathbf{b}(t))^T null(A) = 0$. Additionally, $(A^+\mathbf{b}(t))^T A^+\mathbf{b}(t)$ does not change with $\boldsymbol{\gamma}(t)$, so that its removal from the optimization problem does not change the result. Thus, Eq. 16 is equivalent to the quadratic program of Equation 15.

Other optimization problems could be formulated, but the challenge is that it is not really known what optimality means for the fluxes in a biological system or organism. Optimal solutions, with respect to various criteria, could be suggested, but whether these solutions are compatible with additional information about the functional form or about effectors of fluxes needs to be tested for specific problems. Section 4.1.6.4 examines the minimum-energy solution for a realistic biological system

and indeed challenges the validity of this particular solution. This discussion shows that optimization, which at this stage does not assume any functional form for the fluxes, may lead to fluxes that can become questionable later. At the same time, these optimal solutions can be utilized for approaching solutions that appear to be biologically meaningful.

4.1.5 Generic Information Regarding Alleged Flux Characteristics Can Restrict the Feasible Space Further

After characterizing a feasible set of fluxes, as discussed in previous sections, optimizing the parameters for these fluxes yields a reasonable default solution. Nonetheless, accounting additionally for generally expected features of fluxes can lead to more biologically relevant flux sets. Such generic features may include knowing that a certain flux is a function of only one variable, *i.e.*, its substrate. Another piece of generic information could be that, when a substrate of a flux is zero, the flux has to equal zero as well. These types of constraints will be explained in more detail with an example.

In the following sections, the task of flux identification is performed with a realistic example from the literature that has the right degree of complexity for illustrating the methods described before. The example concerns the biosynthetic pathway of aspartate-derived amino acids in the plant *Arabidopsis thaliana*. In reference to the lead author of a model of this system, we will call it the Curien model. Since the complete model and the fluxes are known, the pathway system constitutes a good test case. The Gamma-trajectory for the Curien model will be plotted, the criterion of non-negativity and its implication in Gamma-space will be investigated and determined, and the result of optimization will be studied and compared to the original fluxes. Finally, auxiliary methods of flux improvement will be suggested.

4.1.6 Flux Identification for the Biosynthesis of Aspartate-derived Amino Acids in the Plant *Arabidopsis thaliana*

Curien and coworkers developed a regulated metabolic reaction network model of the biosynthesis of aspartate-derived amino acids for the plant *Arabidopsis thaliana* [18]. This pathway is responsible for the distribution of the carbon influx into the synthesis of threonine, lysine, methionine, and isoleucine (Figure 14). The kinetic model was constructed based on *in vitro* kinetic measurements, using functional forms of the fluxes in the tradition of Michaelis and Menten. The model contains seven dependent variables, namely, X_1 = [Aspartyl-phosphate], X_2 = [Aspartate semialdehyde], X_3 = [Lysine], X_4 = [Homoserine], X_5 = [Phosphohomoserine], X_6 = [Threonine], and X_7 = [Isoleucine] [18]. We additionally consider the output variable X_8= [Threonyl-tRNA].

This specific case of a metabolic reaction network model is selected for the illustration of the proposed techniques of flux identification, because it is representative and of moderate complexity, and because it is fully known, which facilitates method development and multiple diagnoses of problems that are likely to arise.

The equations for the model are directly taken from the original paper (Eq. 17). The functional forms of the fluxes are presented in Eq. 18.

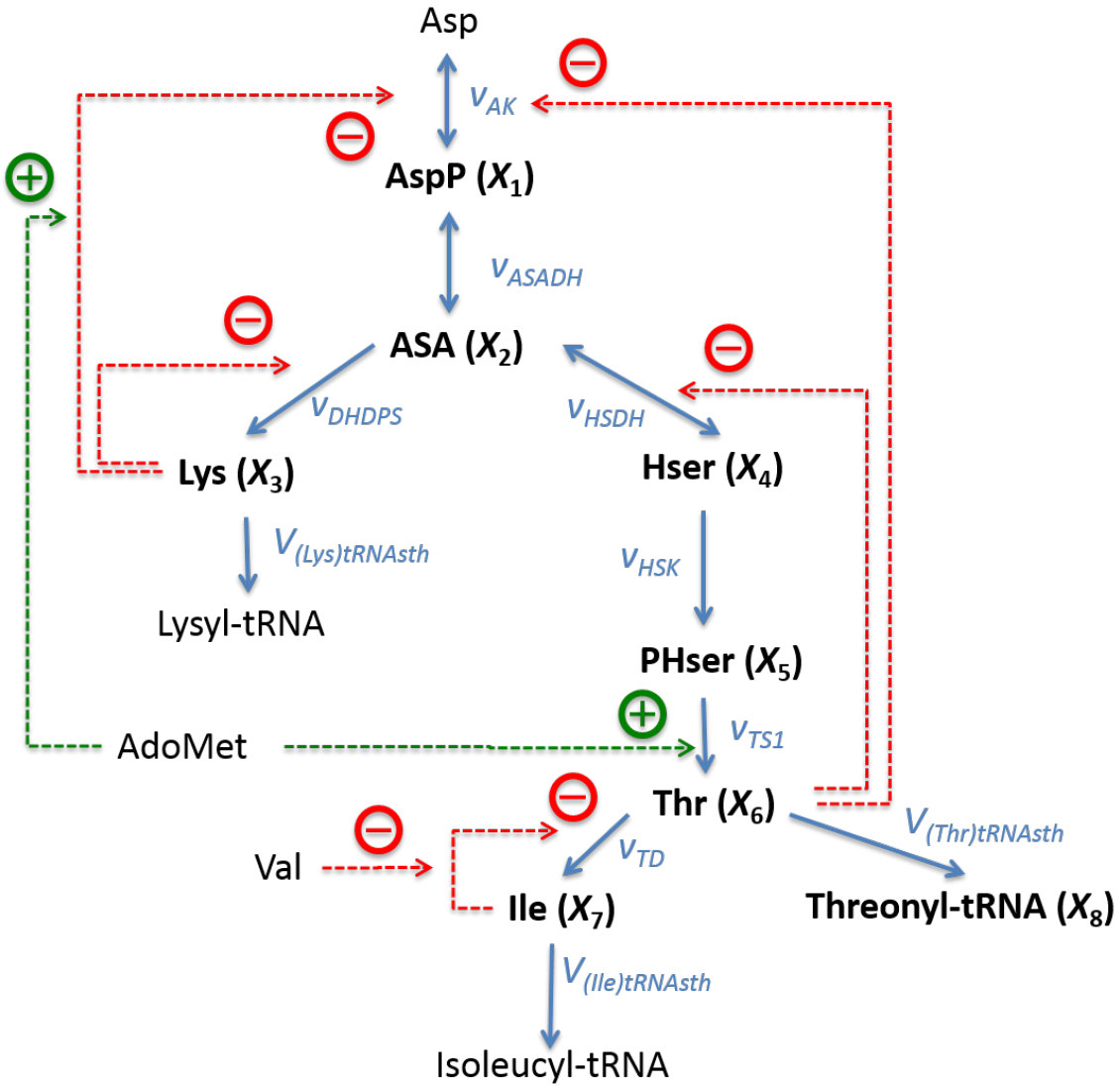

Figure 14: Metabolic reaction network of the biosynthesis of aspartate- derived amino acids in *Arabidopsis thaliana*. Abbreviations are: Asp: L-Aspartate, AspP: L-Aspartate-4-phosphate, ASA: L-Aspartate- semialdehyde, Lys: L- Lysine, Hser: Homoserine, PHser: O-Phospho-L-homoserine, AdoMet: S-Adenosylmethionine, Thr: L-Threonine, Ile: L-Isoleucine, Val: L-Valine. Lysyl-tRNA and Isoleucyl-tRNA are shown here as end products, but they are not explicitly included in the model. Adapted from [18].

$$\frac{dX_1}{dt} = v_{AK} - v_{ASADH} \tag{17a}$$
$$\frac{dX_2}{dt} = v_{ASADH} - v_{DHDPS} - v_{HSDH} \tag{17b}$$
$$\frac{dX_3}{dt} = v_{DHDPS} - v_{(Lys)tRNAsth} \tag{17c}$$
$$\frac{dX_4}{dt} = v_{HSDH} - v_{HSK} \tag{17d}$$
$$\frac{dX_5}{dt} = v_{HSK} - v_{TS1} \tag{17e}$$
$$\frac{dX_6}{dt} = v_{TS1} - v_{TD} - v_{(Thr)tRNAsth} \tag{17f}$$
$$\frac{dX_7}{dt} = v_{TD} - v_{(Ile)tRNAsth} \tag{17g}$$
$$\frac{dX_8}{dt} = v_{(Thr)tRNAsth} \tag{17h}$$

$$v_{AK1} = [AK1] \cdot \frac{5.56 - 1.6[AspP]}{1 + \left[[Lys] / \left(\frac{550}{1+[AdoMet]/3.5}\right)\right]^2} \tag{18a}$$

$$v_{AK2} = [AK2] \cdot \frac{3.15 - 0.86[AspP]}{1 + ([Lys]/22)^{1.1}} \tag{18b}$$

$$v_{AKI} = [AKI - HSDHI] \cdot \frac{0.36 - 0.15[AspP]}{1 + ([Thr]/109)^2} \tag{18c}$$

$$v_{AKII} = [AKII - HSDHII] \cdot \frac{1.35 - 0.22[AspP]}{1 + ([Thr]/109)^2} \tag{18d}$$

$$v_{AK1} = v_{AK1} + v_{AK2} + v_{AKI} + v_{AKII} \tag{18e}$$

$$v_{ASADH} = [ASADH] \cdot (0.9[AspP] - 0.23[ASA]) \tag{18f}$$

$$v_{HSDHI} = [AKI - HSDHI] \cdot 0.84 \cdot \left(0.14 + \frac{0.86}{1 + [Thr]/400}\right) \tag{18g}$$

$$v_{HSDHII} = [AKI - HSDHI] \cdot 0.64 \cdot \left(0.25 + \frac{0.75}{1 + [Thr]/8500}\right) \tag{18h}$$

$$v_{HSDH} = v_{HSDHI} + v_{HSDHII} \tag{18i}$$

$$v_{DHDPS1} = [DHDPS1] \cdot [ASA] \cdot \frac{1}{1 + ([Lys]/10)^2} \tag{18j}$$

$$v_{DHDPS2} = [DHDPS2] \cdot [ASA] \cdot \frac{1}{1 + ([Lys]/33)^2} \tag{18k}$$

$$v_{DHDPS} = v_{DHDPS1} + v_{DHDPS2} \tag{18l}$$

$$v_{(Lys)tRNAsth} = V^{AaRS} \cdot \frac{[Lys]}{25 + [Lys]} \tag{18m}$$

$$v_{HSK} = [HSK] \cdot \frac{2.8[Hser]}{14 + [Hser]} \tag{18n}$$

$$v_{TS1} = [TS1] \cdot \frac{\left(\frac{0.42+3.5[AdoMet]^2/73}{1+[AdoMet]^2/73}\right)[PHser]}{\left[\frac{250\left(\frac{1+[AdoMet]/0.5}{1+[AdoMet]/1.1}\right)}{1+\frac{[AdoMet]^2}{140}}\right]\left(1 + \frac{[P_i]}{1000}\right) + [PHser]} \tag{18o}$$

$$v_{(Thr)tRNAsth} = V^{AaRS} \cdot \frac{[Thr]}{100 + [Thr]} \tag{18p}$$

$$v_{TD} = [TD] \cdot \frac{0.0124[Thr]}{1 + \left[[Ile] / \left(30 + \frac{74[Val]}{610+[Val]}\right)\right]^3} \tag{18q}$$

$$v_{(Ile)tRNAsth} = V^{AaRS} \cdot \frac{[Ile]}{20 + [Ile]} \tag{18r}$$

Equations (17a-17h) can equivalently be written in vector form as shown in Eq. 19, where vector $\mathbf{v}$ and matrix A are the corresponding vector of reaction rates (*i.e.*, fluxes) and the stoichiometric matrix, respectively; they are shown in Eqs. 20 and 21.

$$\frac{d\mathbf{X}}{dt} = \dot{\mathbf{X}} = A \cdot \mathbf{v} \tag{19}$$

$$\begin{aligned}\mathbf{v} = [&v_{AK}, v_{ASADH}, v_{HSDH}, v_{DHDPS}, v_{(Lys)tRNAsth}, v_{HSK}, \\ &v_{TS1}, v_{(Thr)tRNAsth}, v_{TD}, v_{(Ile)tRNAsth}]^T\end{aligned} \tag{20}$$

$$N = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{21}$$

4.1.6.1 Gamma-trajectory of the Curien model

The fluxes and metabolite concentrations for this system are known, which allows us to plot the true Gamma-trajectory in the Gamma-space representation *vs.* time:

$$\mathbf{v}(t) = A^{+}\mathbf{b}(t) + null(A)\boldsymbol{\gamma}(t) \tag{22}$$

Here,

$null(A) = [\mathbf{vec}_1, \mathbf{vec}_2] =$

$$\begin{bmatrix} 0.5374 & 0.5374 & 0.1162 & 0.4212 & 0.4212 & 0.1162 & 0.1162 & 0 & 0.1162 & 0.1162 \\ 0.0534 & 0.0534 & 0.3914 & -0.3380 & -0.3380 & 0.3914 & 0.3914 & 0 & 0.3914 & 0.3914 \end{bmatrix}^T$$

spans the null space of A. This solution is easily found, as $null(A)$ is a simple MATLAB command that results in these two orthonormal vectors. $\boldsymbol{\gamma}(t) = [\gamma_1(t), \gamma_2(t)]^T$ is the vector of coefficients. With this information, the two-dimensional Gamma-space can be explored instead of the feasible subset of the 10-dimensional space of fluxes.

For each time point t, the gamma coefficients can be calculated by projecting $\mathbf{v_{null}}(t) = \mathbf{v}(t) - A^+\mathbf{b}(t)$ onto the vectors $\mathbf{vec}_1$ and $\mathbf{vec}_2$. The result is equivalent to the dot product of $null(A)$ and $\mathbf{v}(t)$, since $A^+\mathbf{b}(t)$ is orthogonal to the null space and the dot product would result in zero.

Figure 15 shows the trajectory starting at time zero and ending at steady state shown with a red dot.

4.1.6.2 Feasible solutions

Similar to the introductory example, this model permits an infinite number of solutions, which may be quite different. These feasible solutions are generated by starting at some initial point in the Gamma-space and computing a phase-plane trajectory according to the linear state-space model of $\dot{\boldsymbol{\gamma}}(t) = B\boldsymbol{\gamma}(t)$. A Monte-Carlo approach is utilized, in which a stable 2×2 matrix B is randomly generated and where only those matrices are retained that result in non-negative fluxes for all time points, as described before. The resulting trajectories exhibit a variety of different dynamical characteristics for the fluxes. Panels 1 through 9 of Figure 16 show in multiple colors a selection of feasible solutions for fluxes v_1 through v_{10}, with the exception of the output flow v_8. Flux v_8 is not shown since it belongs to the only full rank subset of the system and is fully determined by differentiating X_8. The thin lines representing

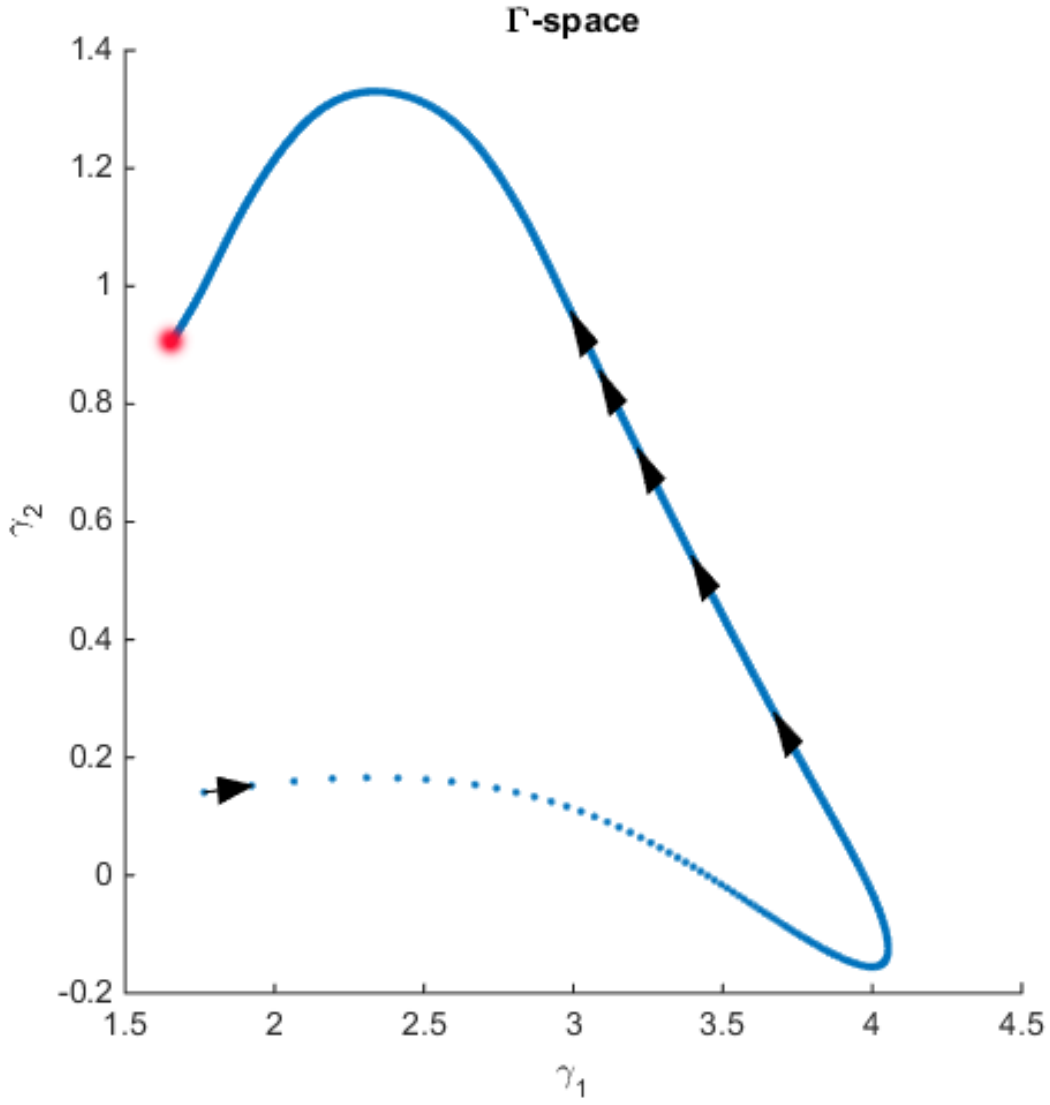

Figure 15: Gamma-trajectory for the Curien model. The spacing of arrows show the progression of time. The steady state is shown in red.

these solutions are superimposed on the actual flux (black) that is known from the model. It is evident that some of the inferred fluxes are similar to the actual fluxes, but that many are not even qualitatively of the same shape. In order to facilitate easier comparisons, the fluxes shown are shifted so that their initial value matches. Interestingly, the inferred fluxes show different behaviors ranging from monotonic to various oscillatory shapes.

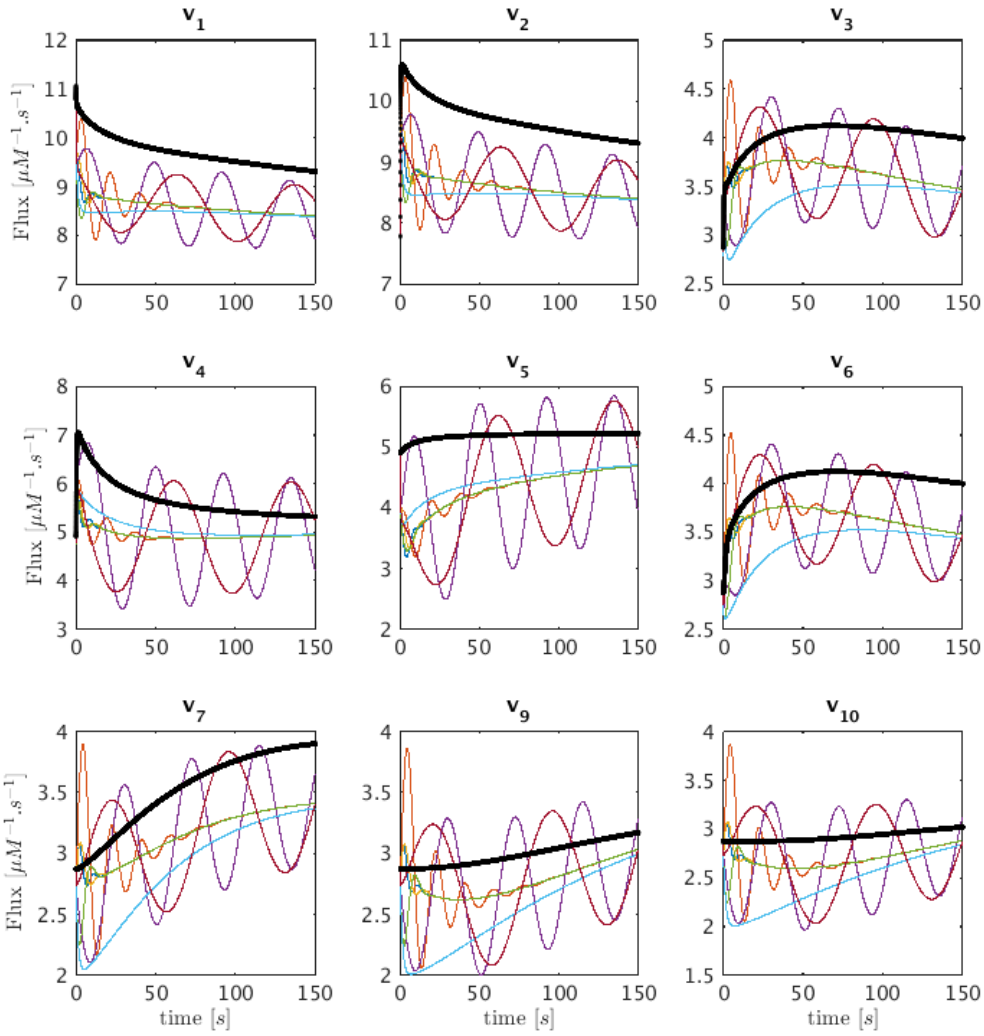

Figure 16: Sets of feasible solutions for each flux v_1 to v_7 and v_9 to v_{10} is shown in each panel. The actual flux from the model is superimposed as a thick black line for comparison.

One should note that these feasible solutions are representative examples if we assume a trajectory from a linear state-space solution and by no means represent all the possible trends.

An interesting observation is that one may add an equal amount to each flux in Set 1 = $\{v_1, v_2, v_4, v_5\}$ and/or Set 2 = $\{v_1, v_2, v_3, v_6, v_7, v_9, v_{10}\}$ without a change in the metabolite concentration profiles. The reason is that these shifts cancel out in the original differential equations and $\dot{\mathbf{X}}(t)$ therefore stays the same. Figure 17 indicates that the shape of the Gamma-trajectory in Figure 15 is shifted along the red line if one adds different positive constant amounts to Set 1 and along the cyan line if one adds different positive constant amounts to Set 2. One could also pick negative constant values as long as the fluxes stay positive. This way, the whole Gamma-space can be spanned by feasible solutions satisfying Eq. 19. This is an equivalent, and perhaps more comprehensible, explanation of the two degrees of freedom for this pathway. As an alternative to constant shifts, one could add the same function of time to all fluxes in the sets.

4.1.6.3 Admissible subset of the Gamma-space: the subspace of non-negative fluxes for the Curien model

For each time point t, we determine the set of s for which the corresponding v(t) consists entirely of non-negative fluxes. From Eq. 13, the feasible space is an intersection of 10 half spaces characterized by Eq. 23.

$$A^{+}(i,:)\mathbf{b}(t) + \gamma_1 vec_{1,i} + \gamma_2 vec_{2,i} \geq 0, \quad i = 1, 2, \cdots, 10 \tag{23}$$

Here, $A^{+}(i,:)$ denotes the i^{th} row of the 10×8 Moore-Penrose pseudo-inverse matrix.

In this example, only 2 out of the total of 10 inequalities happen to be active inequalities, which results in a feasible subspace in the shape of an open triangle. One should note, however, that $\mathbf{b}(t)$ changes with time, so that there is a new open

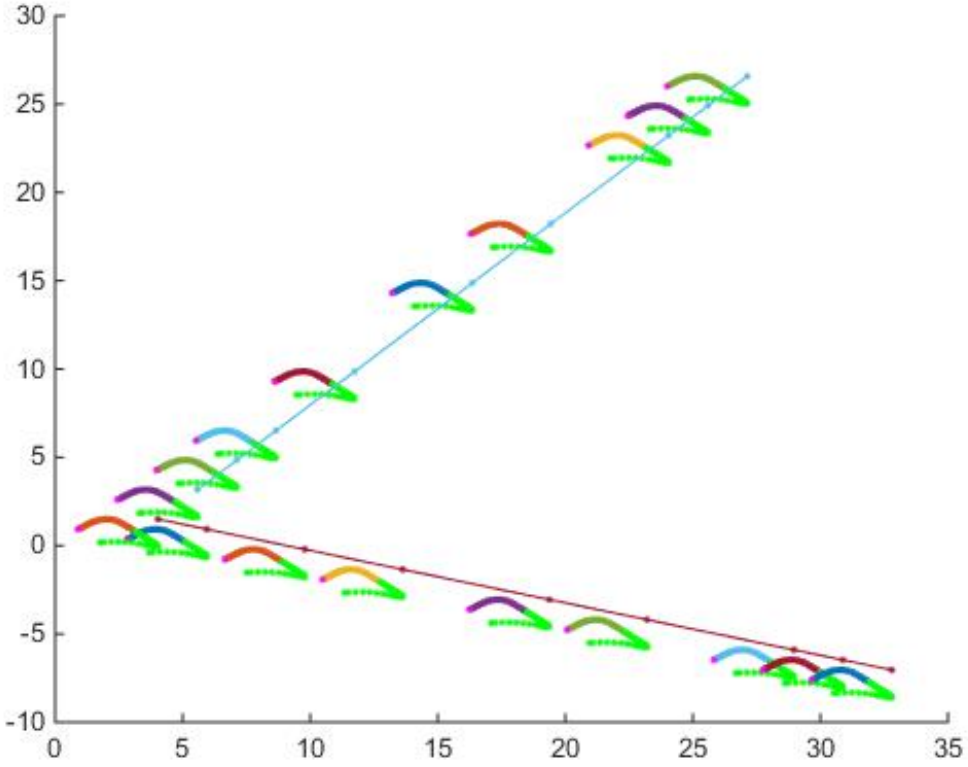

Figure 17: Adding a constant amount to the fluxes in Set 1 for all time points shifts the Gamma-trajectory along the dark red line without any change in the concentration profiles for all metabolites. Similarly, adding a constant amount to the fluxes in Set 2 for all time points shift the Gamma-trajectory along the cyan line without any change in the concentration profiles for all metabolites.

triangle for each time point. Expressed differently, the feasible region resulting in non-negative flux sets varies with each time point. Figure 18 exhibits 7 of these open triangles in different shades of red. There is one such triangle for each time point; most of these triangles are not shown for the following time points to avoid over-population of the plot.

The corners of these open triangles are shown as black dots, which lie on a curve. The blue curve shows the actual Gamma-trajectory of Figure 15. One interesting phenomenon is that, for the initial time points, the two curves (true and inferred) are overlapping. For later time points the blue curves lie inside the corresponding open triangle of non-negative solutions.

Any continuous trajectory whose points fall inside these non-negative open triangles for all time points result is a feasible flux profile satisfying Eq. 19.

4.1.6.4 The minimum-energy flux set

Searching the solutions for the set of flux profiles that minimize the sum of squared flux norms for all time points results in the minimum-energy flux. This procedure is equivalent to solving the quadratic programming of Equation 16, and results in the same flux profile as solving the linear programming of Equation 14. For the case of Curien model, both of these methods yield the same set of fluxes as the corner solution introduced in the previous section. This solution is also equivalent to the result of a nonnegative least-squares optimization problem performed in MATLAB.

The minimum energy flux profiles set are plotted *vs.* time (depicted in red) together with the actual fluxes of the Curien model (blue) in Figure 19. It is clear that the two solutions are different, although they match the metabolite data perfectly. Next we will introduce strategies to alleviate this discrepancy.

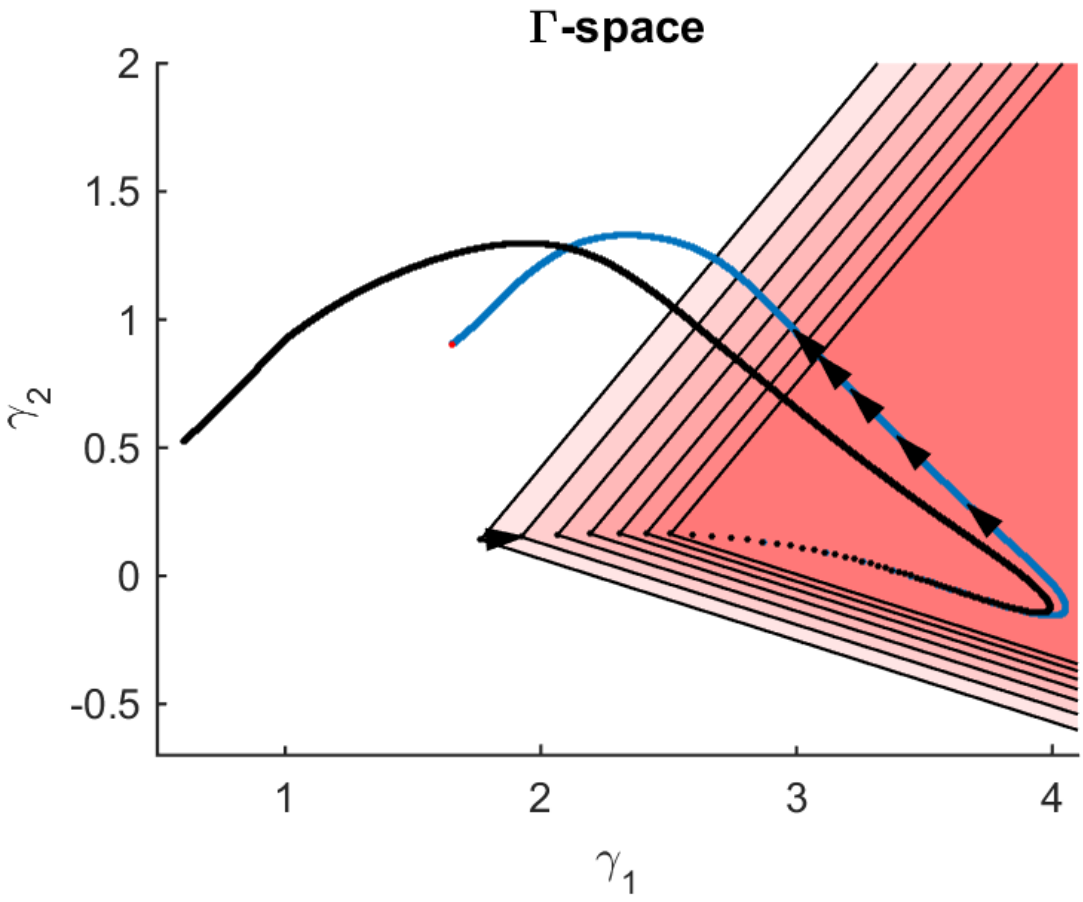

Figure 18: The Gamma-trajectory of the Curien model is depicted in blue color. The black arrowheads shown halfway through the blue curve are equally spaced in time. The open red triangles show the subset of the Gamma-space where the corresponding flux set is non-negative at each point in time. Only the first 7 triangles are shown for illustration purposes. The black doted curve shows the corners of these open triangles for different time points. We will later see that, for the Curien model example, this curve is the same as the minimum-energy curve as described in Section 4.1.6.4. Interestingly the blue and black curves are overlapping in the beginning but then diverge.

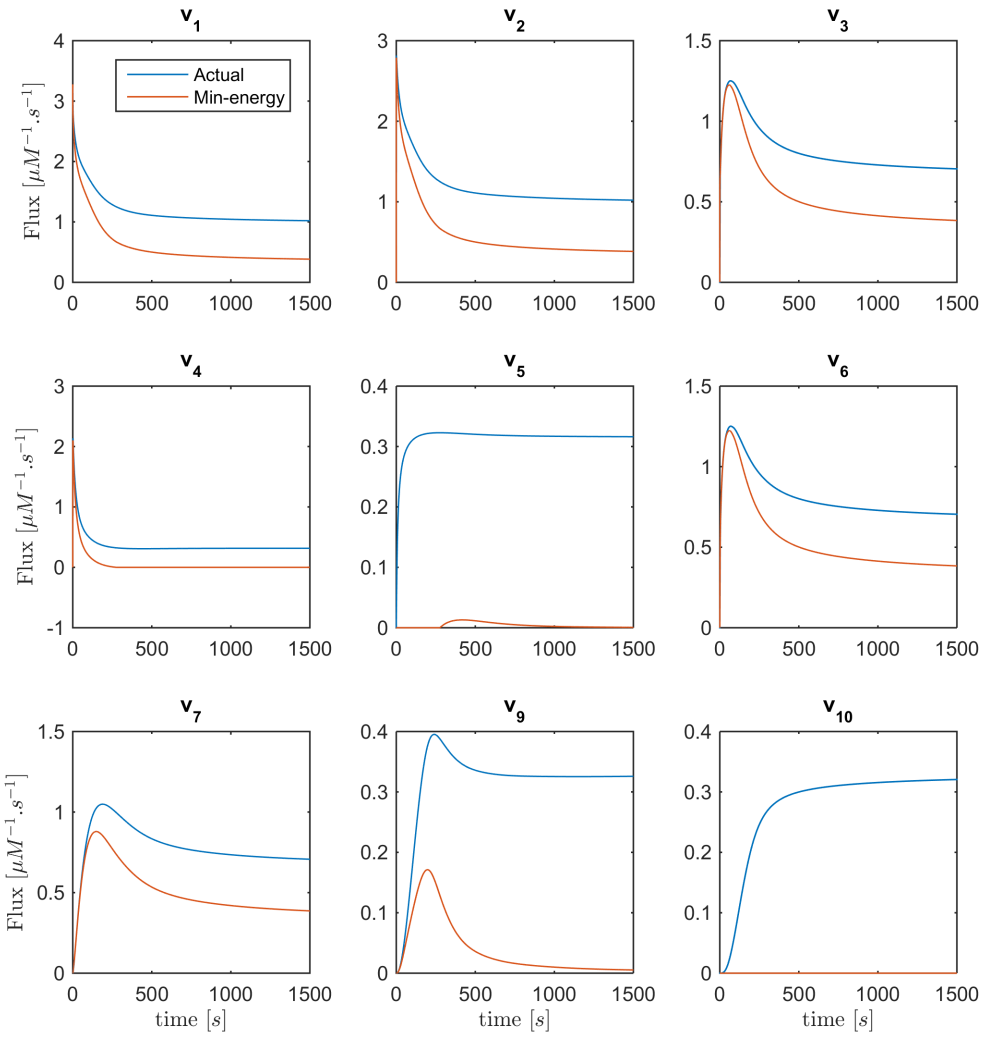

Figure 19: Fluxes v_1 to v_{10} with the exception of v_8 are plotted *vs.* time. Curves in red are the min-energy fluxes, while the blue curves show the actual fluxes of the Curien model. Flux v_8 is not shown because it belongs to the full-rank subset of the system and can be recovered exactly.

4.1.6.5 Generally Expected Features Regarding Fluxes Can Restrict the Feasible Space Further

In this section, general expectations regarding metabolic fluxes are examined to help further constrain the feasible flux profiles.

Some fluxes are functions of one variable only- Let us start from the min-energy solution and incorporate the additional piece of information that each of the fluxes v_5, v_6, v_7, and v_{10} is known to be a function of its substrate only. Using DFE as a diagnostic tool, the min-energy fluxes *vs.* their substrate concentrations are depicted in Figure 20. One notes that the plots of v_6 *vs.* X_4 and v_7 *vs.* X_5 show a behavior that is not allowable, namely a folding-over. For example, if the concentration of X_4 is 1.2 μM, flux v_6 may take two values, and therefore cannot be a function in the mathematical sense. Assuming that we know that no other variables affect this flux, this folding-over phenomenon is not acceptable.

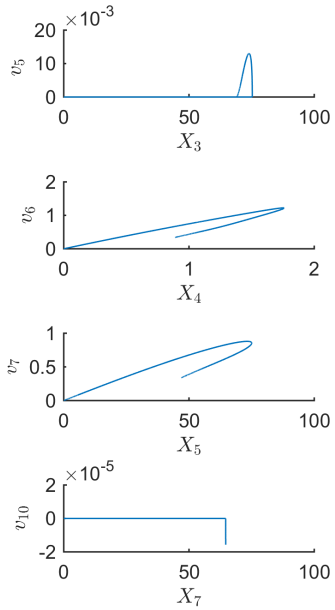

Figure 20: One-substrate fluxes of the system are plotted against their substrate concentrations. The fluxes v_6 and v_7 exhibit a folding-over phenomenon.

In an effort to ameliorate this problem, one may remove or cut the folded-over section. Specifically, for the time points corresponding to those folded-over values, let v_6 take values according to the top branch. Using this technique, $v_6(t)$ becomes uniquely determined and can be considered a known flux (see Section 4.1.6.4 for details). Subsequently, a new min-energy response can be computed with exactly the same methods as before.

For illustration purposes, Figure 21 shows all actual fluxes *vs.* their substrates and effectors in blue, super-imposed on the min-energy fluxes *vs.* their substrates and effectors in red. Fluxes v_2, v_3, v_4, and v_9 have two substrates/regulators, and v_1 has three. Figure 22 depicts the same plots after removing the folding-over phenomenon. Interestingly, all fluxes in Set 2 $= \{v_1, v_2, v_3, v_6, v_7, v_9, v_{10}\}$, as introduced in Section 4.1.6.2, are now fixed and almost equivalent to the actual fluxes. This means that the number of degrees of freedom has decreased to 1 after incorporating the information that one of the fluxes is a function of one variable only. The discrepancy between fluxes in Set 1 $= \{v_1, v_2, v_4, v_5\}$ remains unsolved, and there is no further folding-over case among the one-variable fluxes.

In order to recover the fluxes in Set 1, additional information is needed. As an example if it is known that v_5 assumes a Michaelis-Menten functional and the corresponding kinetic parameters km and Vmax can be extracted from the literature, one could find v_1, v_2, v_4, by the following simple procedure: Find $f_{shift}(t) = \frac{V_{max} X_3(t)}{k_m + X_3(t)} - v_{5.min}(t)$. $f_{shift}(t)$ is the shift function that needs to be added to the rest of fluxes in Set 1 to find the actual fluxes.

$$v_j(t) = v_{j.min}(t) + f_{shift}(t), \quad j \in \{1, 2, 4\} \tag{24}$$

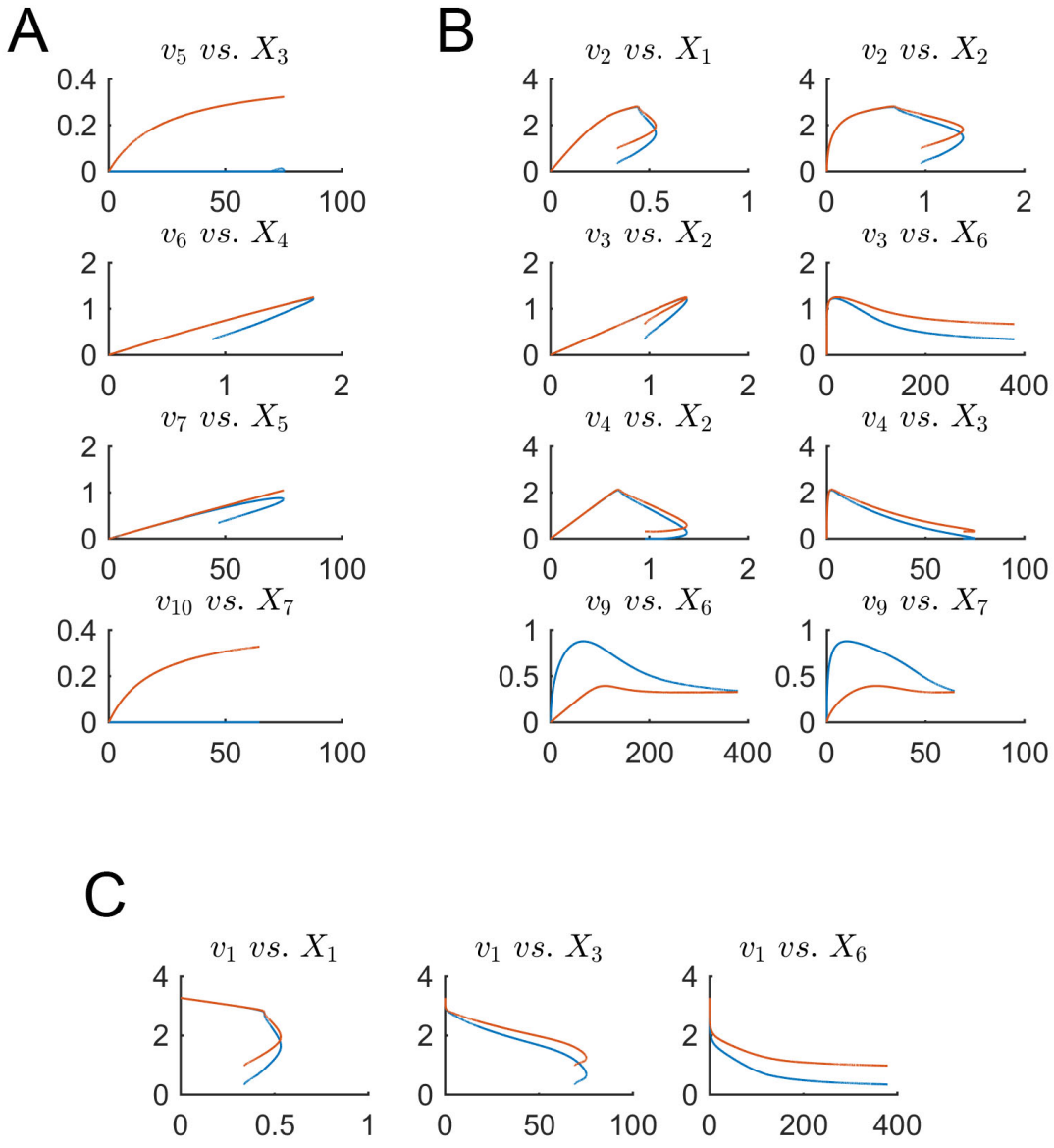

Figure 21: Panel (A) shows the one-variable fluxes *vs.* their substrates. Panel (B) depicts the plots of fluxes that have two substrates/effectors *vs.* each variable separately. Panel (C) shows flux v_1 *vs.* its participating variables. In all plots, the actual fluxes, as known from the original model, are plotted in red, while blue shows the min-energy fluxes.

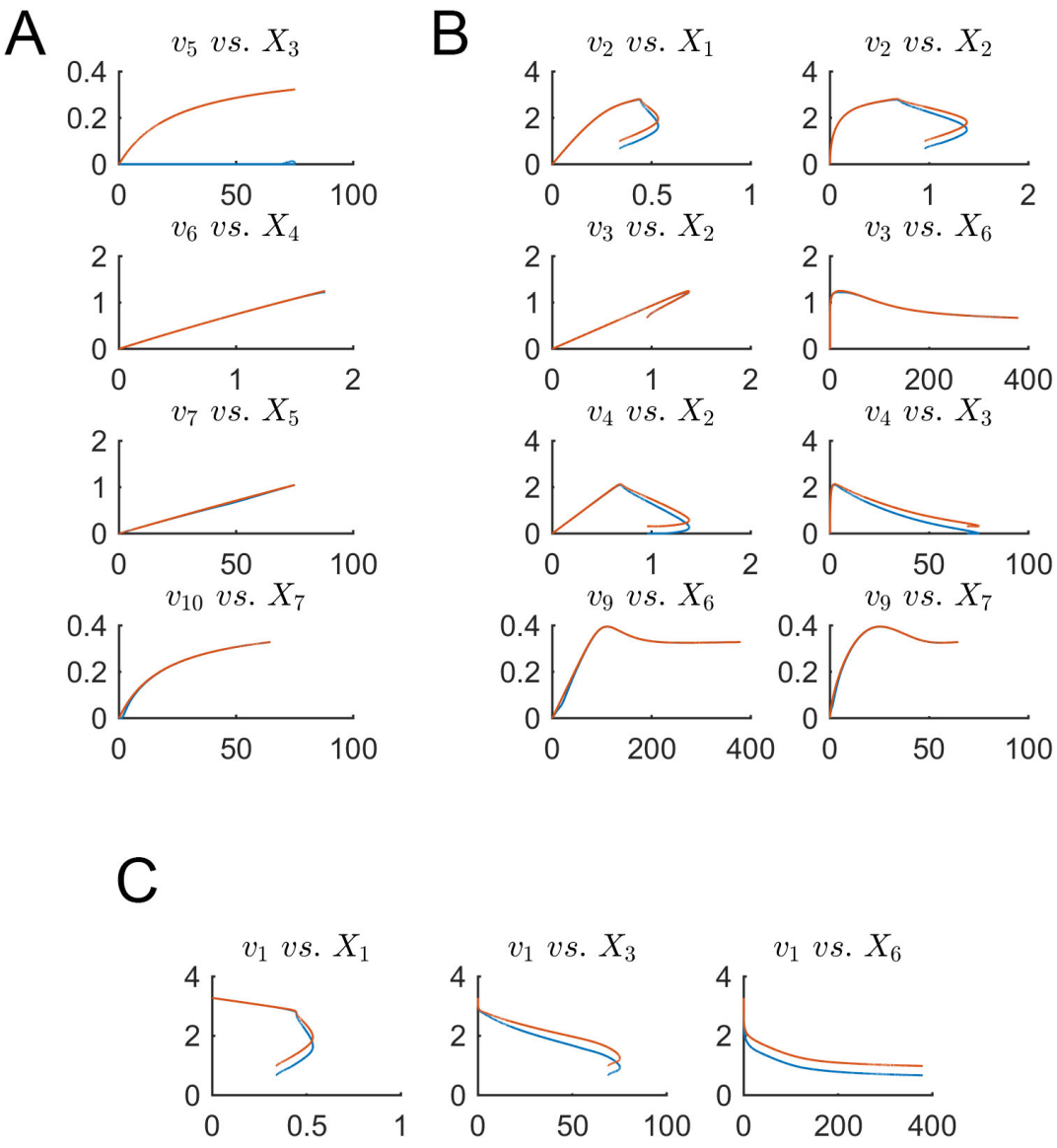

Figure 22: This figure shows the same plots as in Figure 21 with the difference that the plots in blue are the min-energy fluxes after fixing the folding-over problem. Panel (A) shows the one-variable fluxes *vs.* their substrates. Panel (B) depicts the plots of fluxes that have two substrates/effectors *vs.* each variable separately. Panel (C) shows flux v_1 *vs.* its participating variables. In all plots, the actual fluxes, as known from the original model, are plotted in red, while blue shows the min-energy fluxes after resolving the folding-over problem.

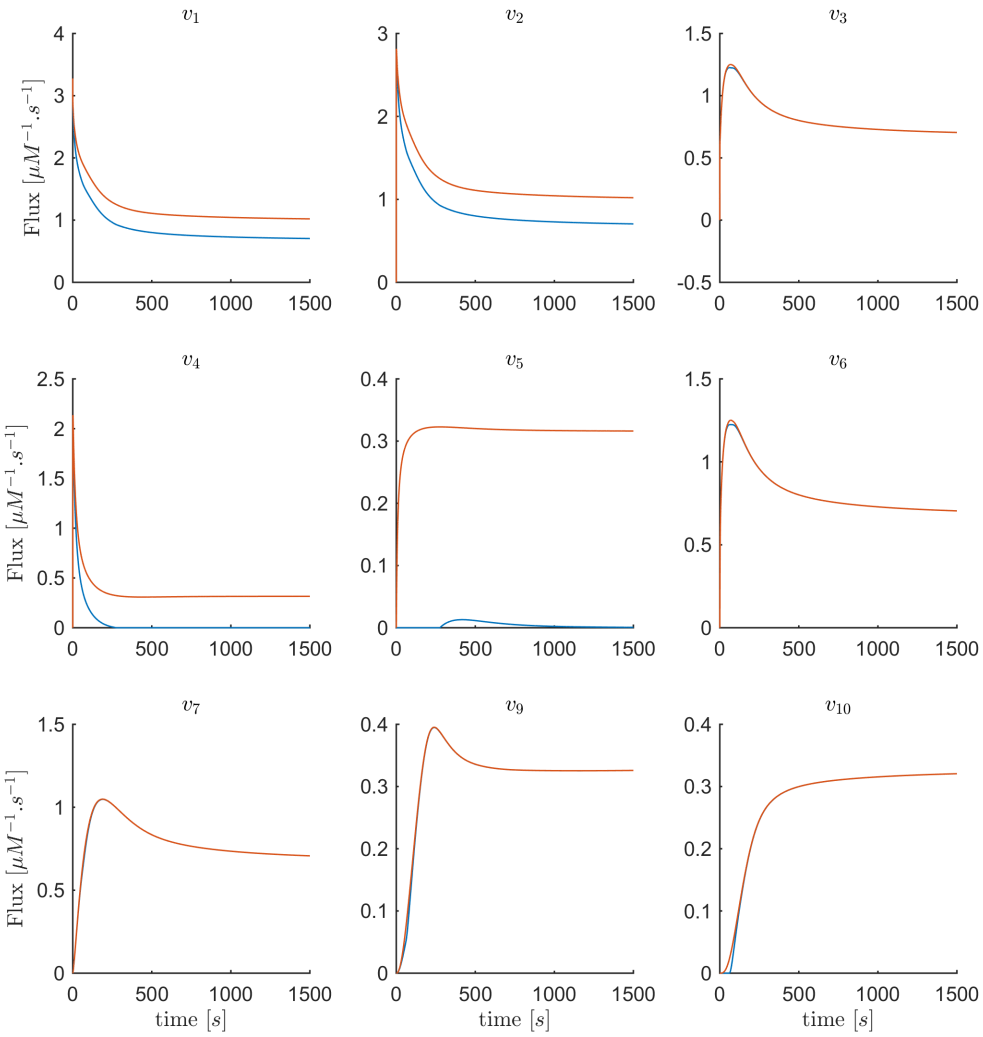

Figure 23: Fluxes v_1 to v_{10} with the exception of v_8 are plotted *vs.* time. Curves in red curves are the min-energy fluxes after solving the folding-over problem, while the blue curves show the actual fluxes. It is evident that the fluxes v_3, v_6, v_7, v_9, v_{10} are almost identical and overlapping and that our method has recovered these fluxes.

4.2 Extension of DFE and Parameter Estimation for the Lactococcus Model

In practical scenarios, some data are often missing, and a number of fluxes cannot be determined fully even after employment of the techniques described in Section 4.1. When this situation arises, there is the need for additional strategies that make maximal use of DFE's flux-by-flux parameter estimation capabilities and diagnostic features, while also using with random search and global optimization techniques.

This section introduces a multi-step strategy that takes advantage of the diagnostic and computational benefits that DFE offers as far as the data characteristics allow, and augments them with auxiliary methods and global optimization approaches to arrive at full-system parametrizations (Figure 24). These procedures are illustrated with the construction of the model of the glycolytic pathway of *Lactococcus lactis* from NMR data, as it was described in Chapters 2 and 3. Due to missing data and other features of the data, this estimation of parameters for the *Lactococcus* model is not straightforward.

In order to use DFE, we first identify full rank subsets of fluxes within the system (see flux estimation module in Figure 24). For instance, supposing that the leakage terms v_7, v_8, v_{7r}, and v_{8r} are negligible, the first six differential equations in Eq. 1 of the Chapter 2 are supported in Dataset 1 with time series representing the concentrations of five of the six metabolites (X_1, X_2, X_3, X_4, X_6). We can directly use these to determine the shapes of the five fluxes v_1+v_P, v_2, v_3, v_4, v_5. For Experiments 2 and 3, G6P data are not available and additional strategies are needed and will be discussed later.

If data for one or more of the variables in a flux v_i are missing, the "missing metabolite estimation modules" in Figure 24 is used. The goal is to constrain the parameters for the following steps of randomized search and global full system optimization. This module involves a high-dimensional optimization task, which ideally yields

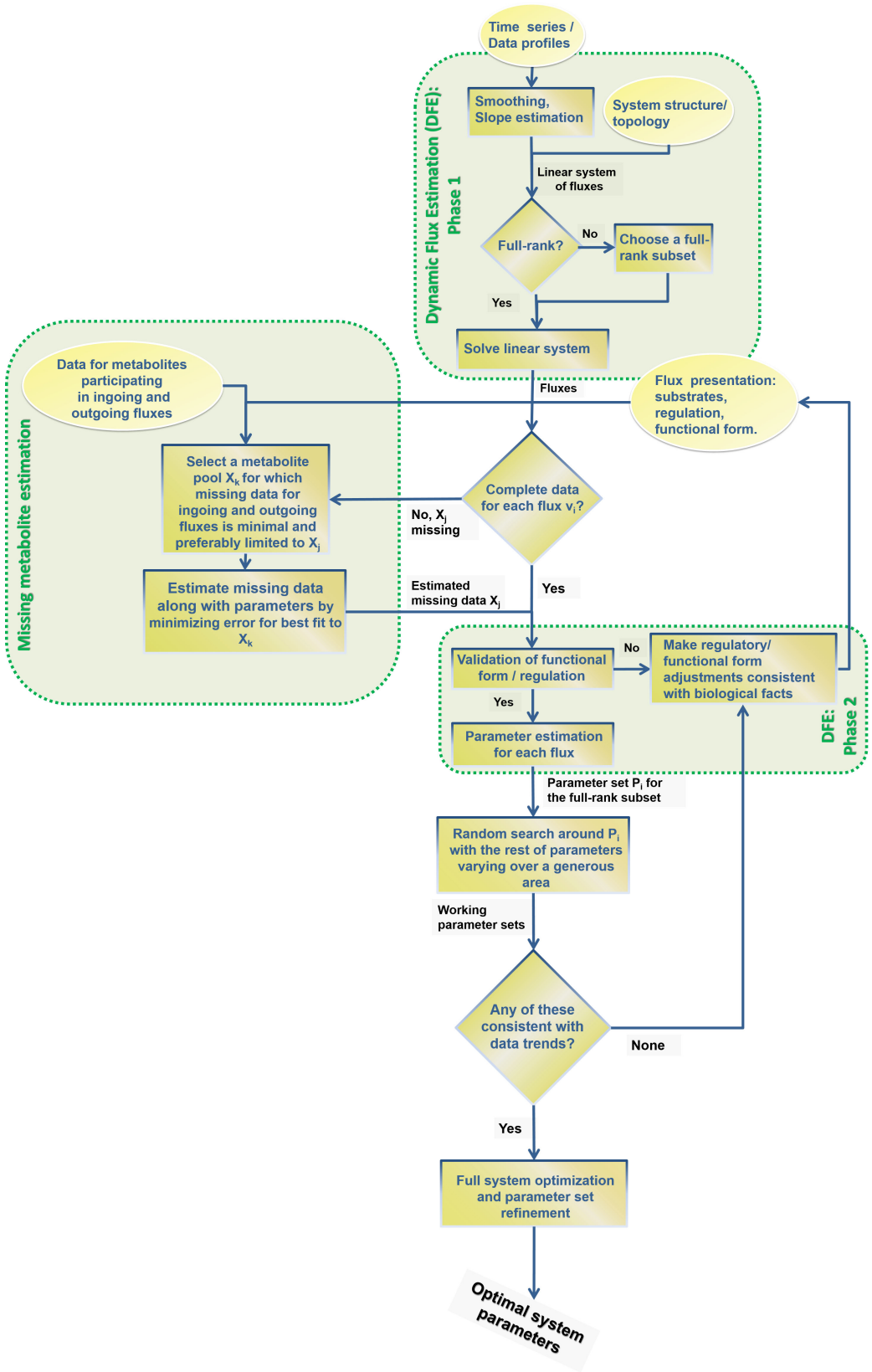

Figure 24: Step-by-step procedure for the proposed extension of dynamic flux estimation (DFE).

valuable information regarding the likely profile of the missing data. The first step in this module consists of selecting a metabolite pool for which the dataset includes a concentration profile and whose influxes and effluxes have the least number of missing data. For example, we select X_3 to estimate missing data for X_2 in Datasets 2 and 3, because we have all data regarding $v_2 = \alpha_2 X_2^{h_{2,2}} X_9^{h_{2,9}}$ and $v_3 = \alpha_3 X_3^{h_{3,3}} X_7^{h_{3,7}} X_{10}^{h_{3,10}}$ except for X_2. One could theoretically use the flux profiles for this estimation procedure. However, the errors in the estimated fluxes would propagate through the system and could result in an inaccurately inferred time profile for the missing metabolite concentration. Thus, we solve the optimization problem of minimizing the sum of squared errors between the ODE simulation of X_3 and the existing data for metabolite X_3, as well as other contributors to the fluxes v_2 and v_3, and simultaneously estimate X_2 for all time points, along with the parameters α_2, $h_{2,2}$, $h_{2,9}$, α_3, $h_{3,8}$, $h_{3,7}$, $h_{3,10}$.

The same procedure is performed for X_5, where we select X_6 as the target metabolite. In some cases, measurements fall below the detection limit, so that no numerical data are available, although the biology of the system mandates that the concentrations are not zero. The detection limit, mass conservation, and possibly other considerations can serve as useful constraints for the optimization algorithm. The outputs of this module thus consist of substitutes for some of the missing data profiles, along with their associated parameter values. In other parts of the workflow, these are treated like experimental data.

The "validation of functional form and regulations" step assesses the appropriateness of the functional formats for the flux representations. A first and obvious criterion is the quality of the fit, which is necessary, although not sufficient [68]. A second criterion is the detection or lack of "runs in residuals" [25]. If no appropriate format and parameterization can be found, it is probable that important components of the pathway are missing from the model. Beyond the quality of fit and run test, no true validation is possible, because the fluxes are unknown. Nevertheless, this step

ensures reasonableness and, for instance, flags fluxes that are computed as negative or exhibit unduly high magnitudes. Similarly, this step assesses the pathway structure in terms of regulatory signals. As an example, the estimation may suggest a formerly unknown inhibition signal, whose biological likelihood is to be evaluated in collaboration with subject area specialists.

As an illustration of the workflow in Figure 24, and more specifically the "validation of functional form and regulation" module, let us focus on one set of fluxes for one of the datasets, such as the 20 mM experiment of Dataset 1. For each flux in this set we check if time profiles for all participating substrates and regulating metabolites are available. If so, we select a functional format and directly estimate its parameters using a simple optimization algorithm. Two examples of this situation are given for v_2, which is a function of G6P and ATP, and for v_6, which is a function of ATP. If we choose power-laws, it is beneficial to take logarithms, which allows us to use linear regression techniques for estimating the coefficients.

In the first example, the dynamics of v_2 appears to be captured well by a power-law function of its substrates. By contrast, v_6, which collectively represents all processes that consume ATP outside glycolysis, cannot be modeled as a non-decreasing function of ATP only. In particular, neither a power-law nor a Michaelis-Menten or Hill functional form can possibly yield an appropriate fit. As a result, we need to explore possible regulators. When scanning through all options within the system, the availability of glucose, or more specifically the transport rate of glucose into the cell, seems to be both mathematically and biologically relevant. Figure 25 shows how the DFE-inferred fluxes (solid lines) are fitted as a power-law function of ATP and the transport rate of glucose (dashed lines). Since v_6 is an aggregate of the processes outside glycolysis, which we are not modelling, the assumption that ATP consumption is regulated by glucose availability seems to be reasonable. Other interesting examples of the application and effectiveness of DFE in elucidating missing regulations of this

nature are presented in Chapter 2.

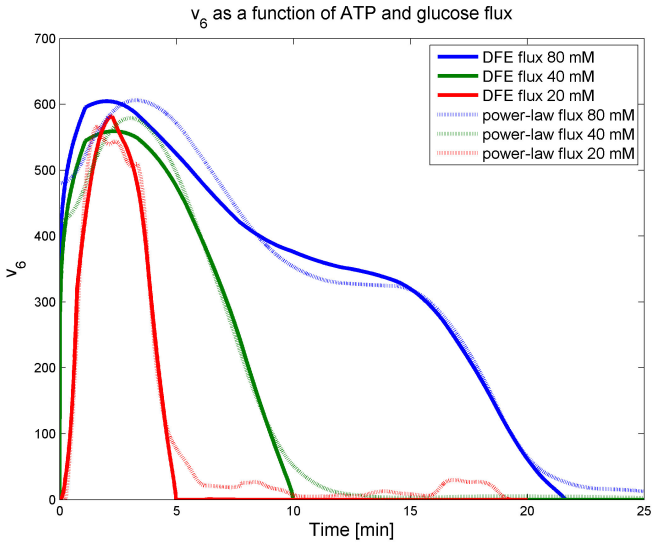

Figure 25: Flux v_6 *vs.* glucose concentration for Experiments 1 (blue), 2 (green), and 3 (red). The solid lines show the DFE-inferred fluxes v_6. Power-law functions of ATP and the glucose transport rate were fitted to these inferred fluxes (dots). These functions fit the inferred fluxes well.

Once the functional forms and regulations are considered satisfactory and the corresponding parameters are estimated, it is necessary to test whether the estimated parameter set is essentially unique or whether substantially different solutions exist. This step is particularly pertinent if the data are noisy and some of the data were not measured but inferred in earlier steps. This global analysis utilizes Monte Carlo simulations, in which a large-scale random search is anchored in the estimated, optimal parameter set $\{P_i\}$, which serves as the starting point for the global optimization. The differences in the sets of estimated parameter values for each flux and each experiment are collectively used to determine admissible ranges for the parameters of the system and starting values for global optimization. Specifically for our analysis, missing parameters were at first allowed to vary within generous intervals. For instance, kinetic orders in BST models were constrained within $[0, 2]$ for positive effectors, and within $[-2, 0]$ for inhibition signals, while the non-negative rate constants were allowed to reach 100 times the value of the biggest estimated rate constant in

the system; in fact they could also have been kept unbounded.

The next step entails a combination of different optimization techniques including evolutionary (genetic) and steepest descent algorithms. The objective function utilized here is the sum of squared errors over all time points, metabolites, and datasets; it also includes a relatively smaller penalty for metabolite concentrations that were inferred rather than directly measured. A smaller penalty for the sum of squared errors over all time points of fluxes, which were directly identified in the 1st phase of the DFE module, was also added. The optimized parameters were first tested with respect to biological reasonableness and, once attested as meaningful, used in model simulations.

4.3 Conclusion and Outlook

The goal of this chapter was to extend the usage of DFE to the more common scenarios when the algebraic system of fluxes are underdetermined (Section 4.1) or the time series data are missing or incomplete (Section 4.2).

The important task of flux identification consists of inferring the fluxes of a metabolic pathway systems using, as input, time series of metabolite concentrations. More specifically, the purpose of this study was to develop a flux identification procedure for metabolic pathway systems where the stoichiometric matrix is underdetermined. Initially, a lower-dimensional representation of a so-called Gamma-space and a Gamma-trajectory was introduced. This representation is especially useful when the degrees of freedom are low (between one and three) so that it can be used as a helpful visualization technique. Reasonable conditions like smoothness over time and non-negativity of fluxes were taken into account to constrain the feasible space even further. Optimization problems were formulated using biologically relevant conditions; in particular, a minimum-energy criterion was considered. The concepts were illustrated with a model of aspartate metabolism in the plant *Arabidopsis*. The

minimum-energy flux set did not match the actual flux profiles for this pathway. However, the addition of biologically reasonable constraints improved the situation considerably. Namely, knowing that a certain flux, v_6, is a function of only its substrate, helped in reshaping the min-energy flux, with the consequence that more than half of the resulting fluxes matched the original flux profiles. More knowledgeor assumptionsabout the fluxes translates into more constraints for the feasible space of solutions and could potentially recover the original flux set. For example knowing that a certain flux follows a specific functional form can potentially determine that flux and decrease the degrees of freedom by one.

It is not clear what the optimal criteria or constraints are to reduce the feasible set of solutions further, because all fit the concentration data exactly. But characterization and a closer look at the set of feasible flux sets may lead to a better understanding of the system and possibly the design of experiments to more efficiently and effectively fill the gap and recover the true fluxes.

Guidance to further experiments should be given: All (most, many) X's should span as much of the relevant substrate range as possible.

On a different but complementary trajectory, incomplete or missing data render the employment of DFE for the task of parameter estimation impossible. I introduced a mixed strategy to alleviate the problem and take advantage of the computational and diagnostic benefits of DFE maximally. This was explained in detail in Section 4.2.

CHAPTER V

DATA PREPROCESSING: A CONSTRAINED WAVELET SMOOTHER FOR PATHWAY IDENTIFICATION TASKS IN SYSTEMS BIOLOGY[1]

As a necessary prerequisite for the tasks of model identification for metabolic pathway systems is data preprocessing and smoothing. Smooth data are especially important when this preprocessing is used as a first step of DFE, which requires slopes of the time courses in the model-free phase. Furthermore, the use of slopes is very advantageous, because parameter values may be estimated without the integration of differential equations [61, 74, 73]. Indeed, the integration of a system of differential equations is computationally expensive and prone to a host of technical challenges, associated with complicated error surfaces that can contain numerous local minima [71]. Experience shows that the slopes are rather sensitive to noise in the time courses, which renders it necessary to smooth and balance the data. Smoothing reduces noise, while balancing assures that there is no gain or loss of mass over time in a closed system.

Numerous methods have been proposed for smoothing time course data. They include splines, moving average algorithms, finite difference approximations, and various types of nonlinear programming [26, 63, 78].These methods are time consuming and need to be performed interactively, or at least in a closely supervised manner. Furthermore, this type of smoothing process can lead to secondary issues. Especially important for the purposes of metabolic pathway analysis is the potential problem that the overall mass in a system may no longer be constant if the data are smoothed.

[1]THIS MATERIAL HAS BEEN PUBLISHED [23]

To address these issues, we propose here an automated smoothing technique that takes as input any given data set and estimates and removes noise while at the same time satisfying the required mass balance within the system. The proposed approach is iterative and called *Constrained Iterative Wavelet-based Smoother* (CIWS).

5.1 Background and Data

5.1.1 Multiresolution Analysis Using Wavelets

The proposed smoothing technique is built upon the notion of multiresolution analysis (MRA) from wavelet theory, which we will briefly explain here.

Wavelets are becoming a standard data analysis tool that is excellent for tasks of data compression as well as for denoising and smoothing. One of their advantages is that they are flexible as well as local, which means that they do not ignore desirable functional details. The reason is that the resolution in MRA can be adapted to the situation at hand.

Mathematically speaking, wavelets are orthogonal basis functions which span the space of all square-integrable functions ($L^2(R)$). Thus, any element in $L^2(R)$ may be represented as a possibly infinite linear combination of these basis functions. An important property of this linear representation is that it may be partitioned into orthogonal subspaces $W_j = span[\psi_{j,k}(x)]$, each of which captures a certain level of "detail" information. The key concept of orthogonal MRA is to partition a given function $f(x)$ into its components $f^{(j)}(x) \in W_j$. Here, the space W_j consists of functions with lower resolution than the ones in W_{j+1} which means that if some arbitrary function $g(x)$ is in W_j, then $g(2x)$ is in W_{j+1} [59].

For example, in the traditional wavelet representation $f(x) = \sum_{k\in Z} c_{J_0,k}\phi_{J_0,k}(x) + \sum_{j\geqslant J_0,k\in Z} d_{j,k}\psi_{j,k}(x)$, the second sum contains the terms which capture the higher levels of detail (*i.e.*, $\cup(j \geq J_0)W_j$, which is the union of all levels of detail greater than or equal to J_0). Choosing the appropriate coarsest resolution J_0 gives rise to

different transforms. We can also just approximate $f(x) = \sum_{k \in Z} c_{J_0,k}\phi_{J_0,k}(x)$. The choice of J_0 provides us with the flexibility of selecting the desired level of detail, which is traded against the desired level of smoothness. In the above representation of $f(x)$, the functions $\phi_{J_0,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ are scaling and wavelet functions, which correspond to commonly called "smooth" and "detail" coefcients, respectively; j is the dilation/scale index, and k indicates shift or position [62].

In wavelet decomposition, the wavelet coefficients represent details, and if these are small, they can actually be removed without affecting the general trend of the data. In fact, wavelet transformations are known to be parsimonious in that they can be well described by a relatively small number of "energetic" wavelet coefficients.

Wavelet thresholding is the process of removing the wavelet coefficients that are smaller in magnitude than some threshold λ. The resulting signal, after the inverse wavelet transformation, is expected to have its noise removed or at least reduced. The characteristics of the data determine the magnitude of noise, and it is therefore useful to specify the threshold value λ based on the variability of the data at hand. Different thresholding policies and threshold values are discussed in Section 5.2.2 in more detail.

All wavelet computations were performed in *WaveLab*, a MATLAB wavelet toolbox available from the website of Stanford University [3] . Sample MATLAB codes using the functions available in the *WaveLab* toolbox are available in the Appendix B.

5.1.2 Description of data

The proposed smoothing method was tested with *in vivo* NMR time series data of metabolite concentrations in the glycolytic pathway of the bacterium *Lactococcus lactis* (see Chapter 2). The dataset, which is published in [6], consists of time courses

of glucose, lactate, UDP-glucose (uridine diphosphate glucose), minor amounts of ethanol and 2,3-butanediol, as well as intermediate metabolites including FBP (fructose 1,6-bisphosphate) and 3PGA (3-phosphoglyceric acid). These time courses were obtained from non-growing cells of *L. lactis*, following a bolus of 40 mM of [1-^{13}C] glucose, at 30C. The experiments were executed under anaerobic conditions, and pH was controlled at 6.5 and monitored online.

5.2 Constrained Iterative Wavelet-based Smoother (CIWS)

5.2.1 Basic Concepts of CIWS

During the model-free phase of DFE, noisy time courses are to be smoothed for later slope estimation and balanced in such a fashion that no material is gained or lost [31]. The latter aspect is not a triviality because the model is based on the implicit assumption that all material is accounted for. Thus, if the (smoothed) data do not maintain mass conservation, the model structure is immediately at odds with the data, and the estimation process will introduce undesirable means of numerical compensation.

While several advanced algorithms for general smoothing have been developed (*e.g.*, [26, 63, 71, 78]), they have not ascertained the need to conserve mass. The task at hand here is therefore to smooth the time series $f_i(t), i = 1, \cdots, n$, conditioned on the constraint that their sum remains constant in time. We approach this task by constructing a wavelet transform of each $f_i(t)$ in the following form $f_i(t) = \sum_{k\in Z} c^{(i)}_{J_0,k}\phi_{J_0,k}(t) + \sum_{j\geq J_0,k\in Z} d^{(i)}_{j,k}\psi_{j,k}(t)$. Again, the functions $\phi_{J_0,k}$ and $\psi_{j,k}$ are scaling and wavelet functions respectively, and (j,k) is standard scale/shift wavelet indexing. The two sets of coefficients $c_{J_0,k}$,$d_{j,k}$ have a direct interpretation: the smooth and coarse coefcients $c_{J_0,k}$ are responsible for trends and global features, while the detail coefcients $d_{j,k}$ describe mostly the noise in the decomposed time series $f_i(t)$. By thresholding the detail coefcients, that is, by setting to zero those coefcients

small in magnitude, and inverting back to the domain of original data, the individual functions are smoothed. We will discuss later how small this magnitude is to be set. In the following, we will denote a smoothed version of f_i as f_i^*. In practical applications, wavelet decompositions of a given sampled function f are found by Mallat's algorithm [40] . This algorithm consists of data filtering by two filters h and g which are low pass and high pass wavelet filters. The form of filters is fully determined by the choice of the scaling and wavelet functions.

Unless further precautions in the process of smoothing are introduced, the mass balance among all metabolites will likely be violated, leading to $\sum_i f_i^*(t) = g(t) \neq C$. We propose to balance the sum by rescaling each $f_i^*(t)$ to $f_i^{**}(t) = C f_i^*(t)/g(t)$ and to repeat the process of wavelet-transforming, thresholding, back-inverting, and rescaling. While a rigorous proof of convergence is not possible in generality, simulation results with various representative examples (see later) suggest that this procedure generally leads to a set of smooth metabolic time trend functions, while ascertaining a constant mass balance over time. Indeed, as we will discuss later, the method leads to smooth time courses in our application. Figure 26 depicts the concepts of this procedure.

This Constrained Iterative Wavelet-based Smoother (CIWS) consists of an iterated 3-step procedure.

Step 1. Construct wavelet transforms of each time series in terms of the scaling and wavelet functions ϕ and ψ. For this task, we have a choice among an infinite set of basis functions. As a default, we use the Daubechies, Symmlet, and Coiflet families [22]:

$$\begin{cases} f_i(t), \quad i = 1, 2, \cdots, m \\ \sum_i f_i(t) = C, \quad \forall t \end{cases} \xrightarrow{Wavelet \quad Transform} \tag{25}$$

$$f_i(t) = \sum_{k \in Z} c^{(i)}_{J_0,k} \phi_{J_0,k}(t) + \sum_{j \geqslant J_0, k \in Z} d^{(i)}_{j,k} \psi_{j,k}(t)$$

Step 2. Threshold the detail wavelet coefficients and invert back to smooth time-domain functions. Detail coefficients describe the noise in the decomposed time series $f_i(t)$, while the smooth and coarse detail coefficients are responsible for trends and global features. Hard thresholding operates on the detail coefficients by setting $d^{*(i)}_{j,k} = \begin{cases} 0, & \text{if} \quad |d^{(i)}_{j,k}| < \lambda \\ d^{(i)}_{j,k}, & \text{if} \quad |d^{(i)}_{j,k}| \geq \lambda \end{cases}$ The threshold λ is estimated from the standard deviation of the time series data, as outlined before. Once step 2 is performed and the result inverted back to the time domain, one obtains time series $f^*_i(t), i = 1, 2, \cdots, m$, which however are not mass conserving throughout time (*i.e.*, $\sum_i f^*_i(t_l) = g(t) \neq C$)).

Step 3. Recover mass balance by appropriately rescaling each time series. If the rescaled functions $f^{**}_i(t) = C f^*_i(t)/g(t)$ are sufficiently smooth, then terminate the smoothing process. Otherwise return to Step 1. Here, sufficient smoothness is defined as $\vec{g} - C < \epsilon$, where $\epsilon > 0$ is an acceptable error for deviations from mass conservation and $\vec{g} = \big(g(t_1), g(t_2), \cdots, g(t_N)\big)$, $g(t_l) = \sum_i f^*_i(t_l) \neq C$.

5.2.2 Estimating the appropriate threshold and wavelet functions

We consider N noisy time points of a function $f_i(t)$ sampled at time points $t_j, j = 1, \cdots, N$:

$$f_i(t_j) = y_j + z_j, \quad j = 1, \cdots, N \tag{26}$$

Here the noise is assumed to be white $z_j \sim WN(0, \sigma^2)$; expressed differently the noise is independent and identically distributed with

$$E\{z_j\} = 0, Var\{z_j\} = \sigma^2$$

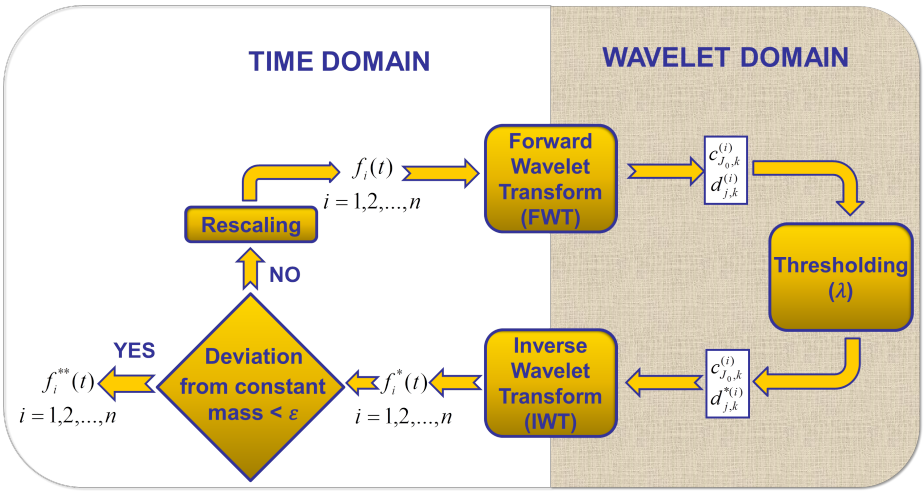

Figure 26: Diagram of the Constrained Iterative Wavelet Smoothing (CIWS) technique.

. Our goal is to estimate a smoothed vector of time points $f_i^{**} = [y_1, y_2, \cdots, y_N]$. As a first step it seems that we need to specify a class $\mathcal{F}$ of sampled functions to which $y_i, i = 1, \cdots, m$ are supposed to belong. However, we often have no a priori knowledge about such a class $\mathcal{F}$. Correspondingly, we assume no functional form for the manner with which the concentrations of metabolites change over time. Instead, determining the appropriate wavelet function, thresholding, and defining a threshold value are central to the performance and convergence of the proposed CIWS smoothing technique.

The process of thresholding wavelet coefficients may be divided into two steps. First, we need to choose an appropriate thresholding rule. In the step-by-step description of the CIWS algorithm, hard thresholding was introduced as the method of choice in principle, but other standard choices are available, including soft thresholding [24], the non-negative garrote shrinkage [2], and others [62]. The most prominent thresholding rules are represented in equations (1)-(3) below:

$$T^{Hard}(d,\lambda) = d\mathbf{1}(|d| > \lambda) \tag{27a}$$

$$T^{soft}(d,\lambda) = (d - sgn(d)\lambda)\mathbf{1}(|d| > \lambda) \tag{27b}$$

$$T^{garrote}(d,\lambda) = (d - \frac{\lambda^2}{d}\mathbf{1}(|d| > \lambda) \tag{27c}$$

, where $\mathbf{1}(|d| > \lambda) = \begin{cases} 0 & |d| \leq \lambda \\ 1 & |d| > \lambda \end{cases}$. Both hard and soft thresholding have advantages and disadvantages. Hard thresholding was discussed before as a viable option. Soft shrinkage tends to have a higher bias since it shrinks large coefficients. This tends to impair the convergence of our iterative method. The non-negative garrote shrinkage function was first introduced by Breiman [2] in a different context and tends to improve the performance of hard thresholding slightly while preserving the convergence properties. No matter what thresholding policy is chosen, a threshold value is to be determined. For each iteration of CIWS, the standard deviation of the remaining noise is estimated directly from the data. The threshold λ is defined as a linear function of the standard deviation of the sampled $f_i^*(t)$, because the noise is expected to decrease as the number of iterations increases. Thus, we set $\lambda = \gamma \cdot \sigma_i$ where σ_i is the standard deviation of the ith time series of concentrations. The coefficient γ still needs to be determined in order for the smoothed function to have the required and desired properties.

Because of implementation issues, the number of data points needs to be a power of two. When this is not the case in a given dataset, we artificially do a mirror-image extension of the last q data points so that the total length is a power of 2. This strategy facilitates smoothness at the boundaries of the observed dataset and avoids the so-called Gibbs effect associated with very short time series; it also yields more robust standard deviation estimates. Assuming that noise in the data is white with variance σ_i^2, the universal threshold of $\lambda_i = \sqrt{2\log n}\sigma_i$ was shown to remove noise

with high probability, thus contributing to the visual quality of the reconstructed signals [24].

5.2.3 Selecting an Appropriate Wavelet Function

For any wavelet generating multiresolution analysis (MRA), there is a trade-off between the smoothness of the wavelet and scaling functions on the one hand, and the locality (that is, the accuracy of local representation) of the wavelets and wavelet representations on the other. Since a wavelet representation of discrete data must interpolate between the points with a shift of the scaling function, it is important that the scaling function generating MRA is smooth, because the wavelet decomposition represents a smooth physical process. However, a wavelet that is too smooth, due to a long wavelet filter, is not sufficiently local and modifies possibly important fluctuations far from the location that it wants to address. In the CIWS algorithm, the functions are iteratively rescaled to ensure a constant sum after each iterative smoothing step. This rescaling is driven point-wise, and for this reason it is important that wavelets retain sufficient locality. Indeed, wavelets that are too smooth and not sufficiently local may cause the CIWS algorithm to diverge. Details regarding the selection of appropriate wavelet functions among the standard families (Daubechies, Coiflet, Symmlet, and Pollen) are presented in Appendix A.

A traditional measure of smoothness of a function f is the Holder exponent α, which is defined through the following inequality: $(\forall x, y)\ \exists C \geq 0, \alpha \geq 0 : |f(x) - f(y)| \leq C|x-y|^{\alpha}$. It was shown that the Holder exponent may be expressed in terms of a large enough vanishing moment M, namely as $\alpha_M = 0.2075M$ [62]. Here, the k^{th} moment of a wavelet function ψ is defined as $\int_{-\infty}^{\infty} x^k \psi(x) dx$, and the terminology that "a wavelet function has M vanishing moments" means that this integral is zero for $k = 0, 1, \cdots, M-1$. For the Daubechies family, for example, M vanishing moments are achieved by discrete wavelet filters that are of length $2M$.

One useful criterion for the choice of an adequate wavelet basis is entropy [12]: Among different candidate wavelet functions, the most appropriate is chosen by minimizing the maximum entropy among different time series of concentrations. Appendix A describes how this criterion facilitates the identification of the best wavelet functions for the purpose of smoothing. For our case of metabolic time series data, Coiflet 1 results in the lowest maximum entropy. This wavelet also well balances smoothness and locality and is differentiable. Appendix Table 3 provides entropy values for our case study.

5.2.4 Avoidance of Negative Concentrations During Back-conversion to the Time Domain

A typical problem with any wavelet-based estimation of non-negative functions and densities is the fact that the estimators may not be fully non-negative. Only the so-called Haar wavelet has a non-negative scaling function and induced kernel, and always results in a non-negative estimator; however, it is known that the convergence rates for strictly positive kernels are inferior to general kernel-based methods, where the kernels can be locally negative [77]. Wavelet-based kernels (except for the Haar) are necessarily locally negative.

A common strategy to circumvent this issue is to fit and smooth log functions or square root functions and, once fitting and smoothing are accomplished, use an exponential transformation or square the results. For our case, this strategy poses a problem with controlling the mass balance. Thus, instead of a transformation, our solution uses a smooth wavelet and always curtails the negative components when iterating. Beyond satisfying our needs, this strategy could be tailored toward constraints other than mass conservation.

5.3 RESULTS

5.3.1 Convergence of the Constrained Smoother

For the metabolic time series in our case study, CIWS converges in fewer than 40 steps to smooth, mass-conserving time courses, and the degree of smoothness is adjustable to our specifications.

A formal, general proof of convergence of CIWS seems not possible, as it would require knowledge of the nature of the functions $f_i(t)$, constrained to some smoothness spaces, and the interplay between thresholding strategies and types of wavelets. The task might be feasible if the goal function, consisting of the residual error to be minimized, were convex.

However due to the arbitrary nature of the functions to be smoothed and the wavelets, the task of analytically optimizing the goal function is intractable. Instead we assess convergence with Monte Carlo simulations and representative functional shapes. Specifically, we use synthetic data from a battery of standard test functions proposed by Donoho & Johnstone [24], known as *Blocks*, *Bumps*, *Doppler* and *HeaviSine*. These functions had been selected because they portray significant spatial inhomogeneity and mimic functions arising in signal processing tasks, including imaging and NMR spectroscopy. An additional test function, *Mishmash*, is defined as: *Mishmash* = C - (*Blocks* + *Bumps* + *Doppler* + *HeaviSine*).

Here C is the constant sum of all five functions. In our case, *Mishmash* reflects the need that the total mass remains constant over time. For simulation purposes, the test functions are sampled at 2048 equally-spaced time points within the interval $[0, 1]$. Choosing the Coiflet 1 wavelet function, CIWS applied to this set converges after only two iterations for a smoothness parameter $\epsilon = 0.1$, if the signal-to-noise ratio (SNR) is set to either 5 or $15dB$. Figures 27-29, show four of these functions without noise, with added white (Gaussian) noise of $SNR = 15dB$, and the output of the CIWS algorithm.

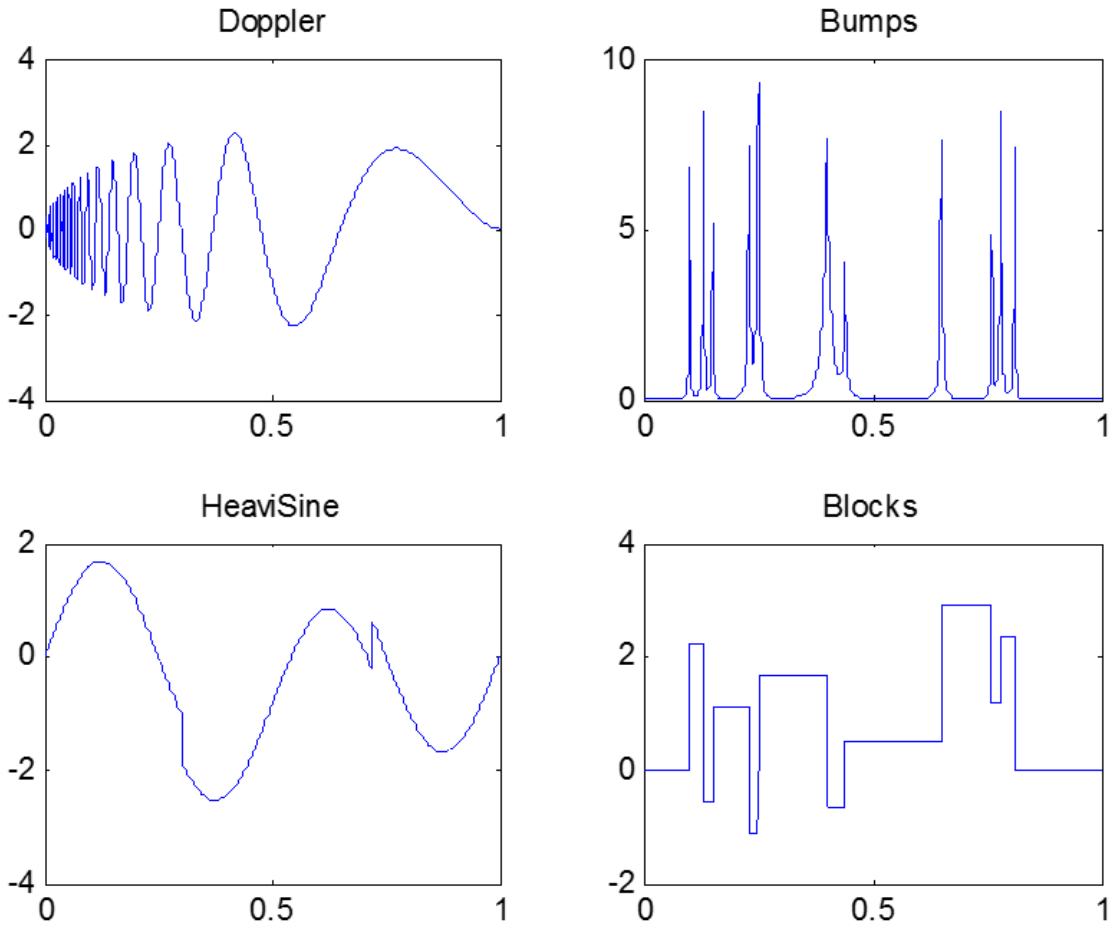

Figure 27: Four of the set of five test functions, called *Doppler*, *Bumps*, *HeaviSine*, and *Blocks* without noise.

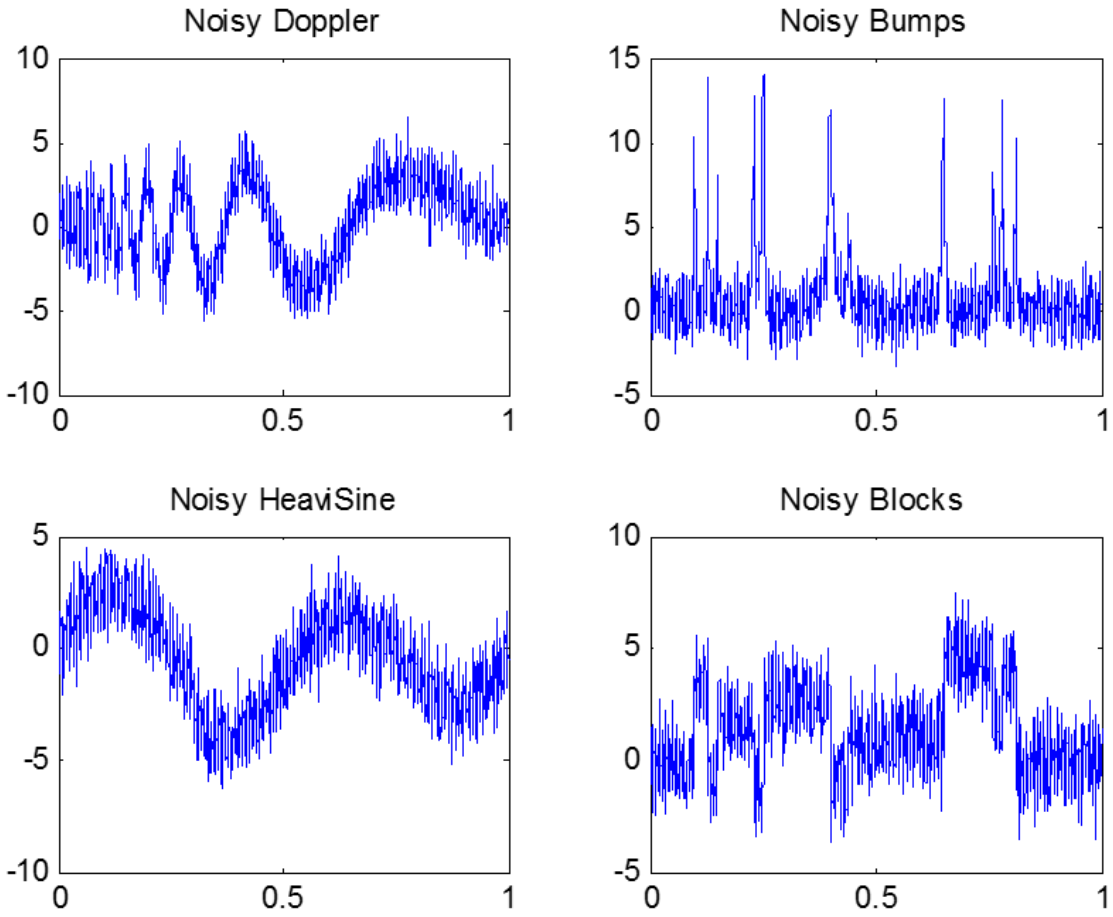

Figure 28: Four of the set of five test functions called *Doppler*, *Bumps*, *HeaviSine*, and *Blocks* with additive white (Gaussian) noise of SNR = 15dB.

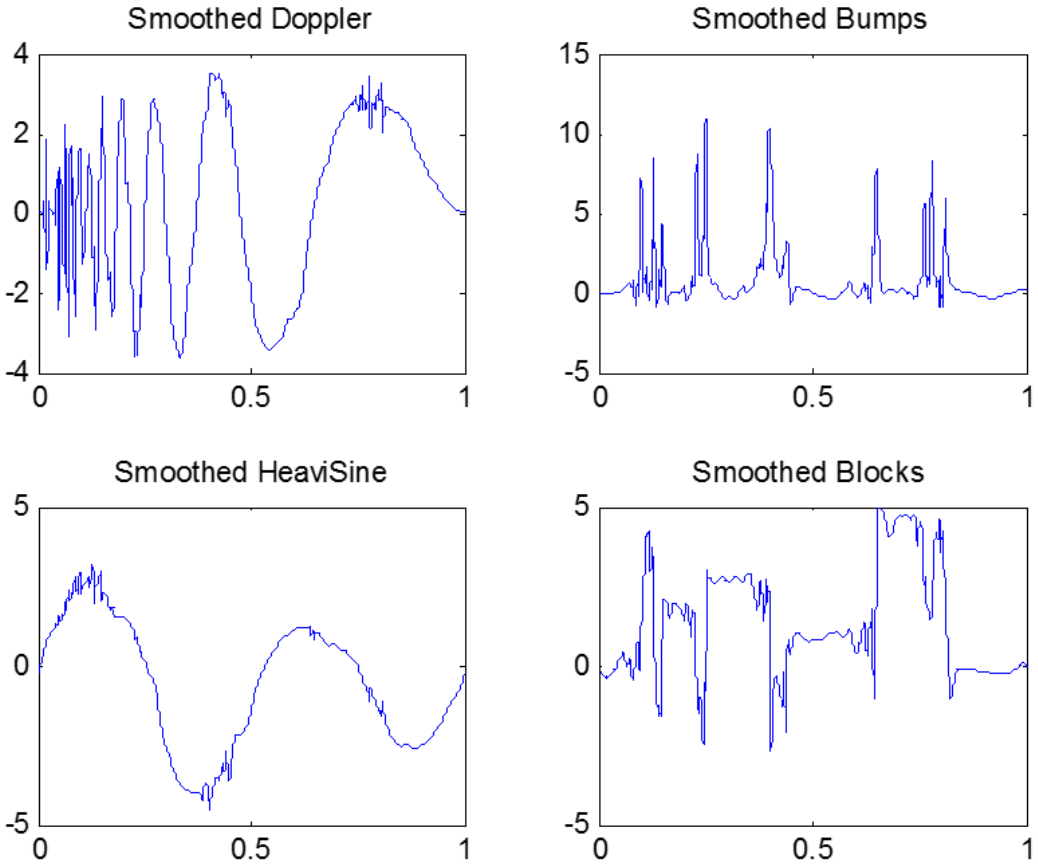

Figure 29: Estimated test functions as output of the CIWS algorithm (compare to Figure 28).

The estimated functions in Figure 29 together with the smoothed *Mishmash* function preserve the balance of energy C, and at the same time have noise mostly eliminated. Locally, the presence of some remaining noise is visible and it could be eliminated at the expense of losing resolution at high frequency features (as in *Doppler*).

5.3.2 Data Analysis

CIWS was tested with different time series data of metabolites in the glycolytic pathway of the bacterium *Lactococcus lactis*, which had been measured under anaerobic conditions following a glucose pulse of 40 mM. The data were obtained with *in vivo* Nuclear Magnetic Resonance (^{13}C-NMR) techniques. The algorithm converged to mass-conserving time courses, whose smoothness was adjustable to our specifications.

Extensive simulations with different datasets demonstrated that CIWS is very time efficient and converges in quite a small number of iterations (between 2-60) if a reasonable value for the mass conservation error constant is chosen, such as $\epsilon \sim 0.01$. As an illustration, Figure 30 depicts the CIWS results for the glycolytic time series data from *L. lactis*. The wavelet function used in this case is *Coiflet* 1. It converges in 28 iterations for $\epsilon = 0.1$. Decreasing ϵ to 0.001 does not result in a significant visual difference but increases the iterations until CIWS converges to 86. Note that CIWS retains the observed drop in lactate around time 14, while smoothing the ascent in this metabolite $t \in (0, 10)$, as well as its more or less stationary phase $t \in (16, 47)$.

5.4 Discussion

In this chapter, we propose a novel Constrained Iterative Wavelet-based Smoothing method that permits noise reduction and smoothing, while assuring a mass conservation constraint. Unlike curve fitting, where the main emphasis is on matching the data as closely as possible, smoothing contains the somewhat vague concept of time course values that are expected to change relatively slowly from one time point to the next. This concept renders the performance criterion for smoothing techniques

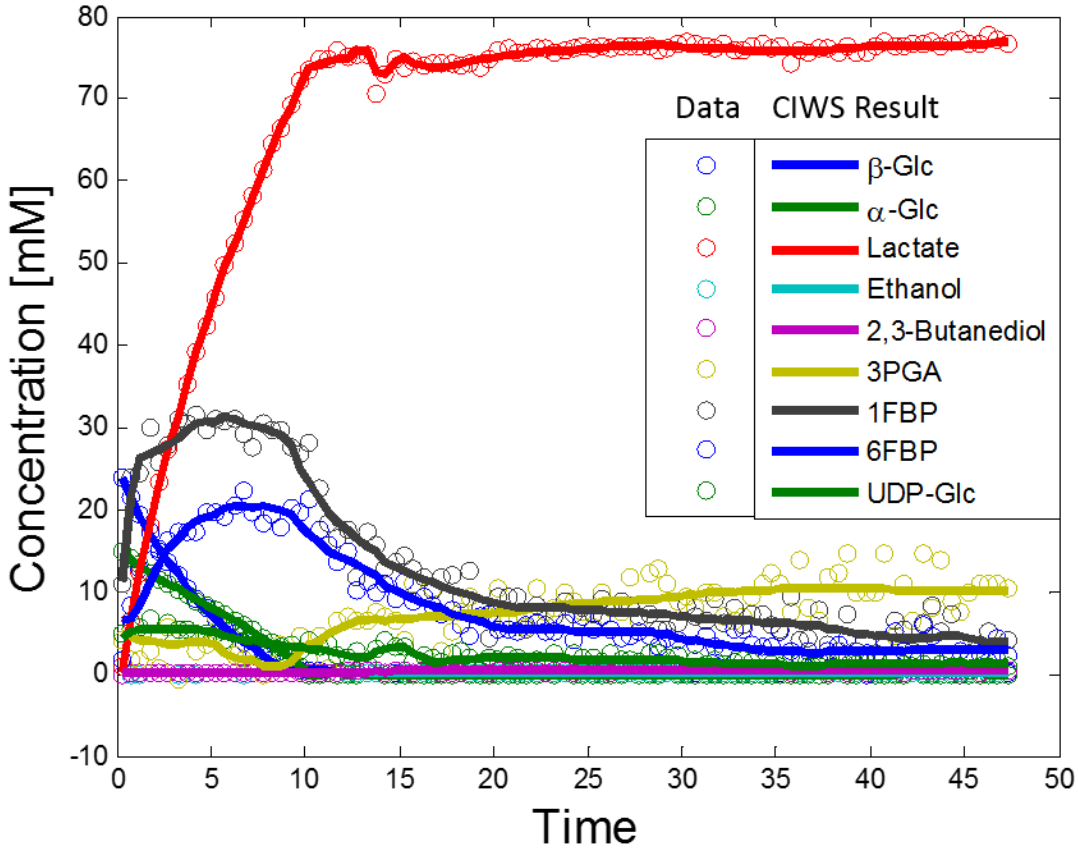

Figure 30: Results of CIWS applied to one sample set of time series data characterizing the dynamics of the glycolytic pathway in *Lactococcus lactis* under anaerobic conditions and with an input glucose pulse of 40 Mm. Circles represent the measured time series data, while CIWS results are represented with lines of the corresponding color.

application dependent. For example, what might be considered as a good smoothing technique in one application where removing of outliers is of interest, might not be considered satisfactory in another application where one might be interested in local artifacts and features of a signal. This data-adaptive interplay between smoothness and local representation renders wavelets suitable tools for flexible smoothing tasks. For the dataset of our case study, the primary purpose is to smooth the data and prepare them for slope estimation while avoiding spurious spikes in the dynamics of the data.

The main novelty of the proposed CIWS smoothing technique is the property of mass conservation, which is practically important since it directly affects the consistency between experimental data and the fitted model.

A secondary advantage with respect to computation and implementation is the fact that CIWS does not require operator interaction or supervision and thus can be automated. Finally, the fast convergence properties of the wavelet transform techniques render this algorithm computationally very efficient. Mallats cascade algorithm can be used in the implementation of each iteration, which results in computational complexity of order $O(N)$. This fact is particularly beneficial when have larger datasets with thousands of data points are to be smoothed. Such larger datasets, which are increasingly more prevalent due to modern high-throughput experimental techniques, will render the proposed CIWS smoothing method even more effective, because short signals decompose in only a few multiresolution spaces and render the thresholding ineffective due to the availability of limited number of scales.

CHAPTER VI

CONCLUSIONS AND OUTLOOK

New types of data, obtained with methods of high-throughput experimental techniques, have presented us with enormously exciting opportunities and challenges. Of specific interest are omics time series data, which contain very valuable information about the structure, dynamics and regulatory mechanisms governing biological systems and are as close to the organisms natural conditions as is presently achievable in the lab. They offer a unique glimpse into the multi-level control of microorganisms. While extraordinarily promising, these types of data pose the substantial challenge of extracting novel, quantitative information and translating this information into integrative dynamic models. At the core of this task is the identification of suitable functions and the estimation of optimal (or at least suitable) parameter values.

In this thesis, I developed or advanced methods for these purposes and applied them to the design of a comprehensive kinetic-dynamic model of the control of carbohydrate metabolism in *Lactococcus lactis*. This model is an expansion of an existing model of glycolysis in *L. lactis* under aerobic conditions [72], but is considerably more complicated because it addresses the organisms preferred anaerobic conditions. This switch between conditions mandated the consideration of the dynamics of NAD^+ and NADH and of ATP and ADP. The use of an advanced modeling methodology, namely DFE and its extensions, as well as the simultaneous account of three datasets within one model instantiation, offered several novel insights into the complex control strategies with which this organism controls glycolysis.

The model proposed in this work represents the available datasets well, but still needs to be tested further against new data. In addition to bolus experiments, as

they were analyzed here, one could imagine testing the model for fermenters under continuous flow conditions, where glucose is constantly supplied. Such fermenters are often preferred in industrial settings, as they yield product on an ongoing basis. An entirely different potential use of the model could be the following scenario, which is of great academic and practical interest, namely the exploration of the effect of pH on glycolysis. If it became possible to manipulate *L. lactis* into quasi-normal operation under low pH conditions, the organism could potentially survive the acidic environment of the human stomach and be used as a non-invasive vehicle for the delivery of proteins or drugs to the intestine [65]. Such a mechanism would be a very welcome pharmaceutical tool for the treatment of diseases such as Crohns [1]. This use of the model would require the explicit inclusion of the driving biological factors affecting the pH inside the cells. It would also require biological experiments assessing the viability of altered cells under such hostile conditions. So far, methods of molecular biology and genetics have resulted in *L. lactis* strains that can tolerate a pH of about 4.8 for some time [6], but this tolerance is not sufficient for the purpose outlined above. A major milestone toward these goals has been reached here, as the model offers a rich array of explanations regarding the strategies that *L. lactis* uses to survive under different conditions.

Another objective of this thesis was the refinement of methods of information retrieval from metabolic time series data, including parameter estimation and structure identification beyond the applications of *L. lactis*. Dynamic flux estimation methods were adapted to become applicable to incomplete and less than ideal data. Finally, a roster of methods was developed for addressing and characterizing underdetermined flux systems. Future studies of systems with higher degrees of freedom and characteristics should be assessed as they could lead to further insights into better characterization strategies for metabolic flux systems. Another future direction could be the exploration of additional, biologically relevant optimization problems

and their constraints, as well as suggestions for experimental designs that would add genuinely new information content and render the task of flux identification more reliable, comprehensive, and possibly exact.

Methods for noise reduction and smoothing, constrained on mass conservation, were developed. As part of the model-free phase of DFE, the estimation of slopes from noisy time series data is crucial. An interesting possible future direction would be the utilization of wavelet multi-resolution analysis for the direct estimation of slopes of the time series data while maintaining mass balance. As a starting point for this research, the chapters on wavelet calculus and connection coefficients by Reskinkoff and Wells [50] and by Mallat [41] are recommended.

Finally, an interesting topic for future investigation could be the study of the complex control mechanisms of *L. lactis* from a perspective of control theory and synthetic biology. More specifically, one could examine if the nonlinear system under study could be representable by an essentially equivalent linear system yielding similar concentration data. This investigation could constitute a compelling system identification task, which could potentially reveal further systems level details of this particular pathway, and could also have intriguing implications for the manipulation or synthesis of new organisms. A thought-provoking question could be if certain behaviors of *L. lactis*, including its dynamics around the ready-to-respond state as explained in Section 3.4.3, could emerge from linear system components and be used to construct a biobrick for similar tasks.

APPENDIX A

SELECTING APPROPRIATE WAVELET FUNCTIONS

Many different wavelets could be utilized for the smoothing task at hand. They possess different smoothness, convergence, and locality properties. Smoothness is a qualitative and somewhat subjective property. Nevertheless, different objective measures, including Sobolev and Holder regularity exponents, have been proposed to quantify the smoothness of scaling functions. The first step of the smoothing process is the choice of an appropriate wavelet function; typical choices include wavelet functions from the *Coiflet*, *Daubechies*, and *Symmlet* families. Among these, we need to eliminate those that are not differentiable, since the initial motivation behind smoothing the time series data is to calculate derivatives. For example, Daubechies 4 and Haar wavelets are not differentiable. As a consequence, functions that are approximated with these wavelets are not truly representative of the physical and biological processes under study.

There is a close relationship and tradeoff between the length of support and the regularity index of the scaling functions [62]. The larger the support interval for a wavelet function, the smoother it can be; however it tends to become less local. The inherent locality-smoothness tradeoff becomes especially important due to the observation that the CIWS method tends to diverge for wavelet functions that are smoother than some value and have a large number of taps. For instance, using Daubechies 10 (or higher) as the wavelet of choice for CIWS will lead to divergence for our test dataset. This may be explained by the fact that the error resulting from the smoothing process is repeatedly rescaled in a non-local manner to keep the total mass constant for all time points. The error therefore propagates and causes

Table 3: Members of the wavelet families *Coiflet*, Daubechies, *Symmlet*, and *Haar*, for which CIWS converges. One can see that their maximum entropies in the wavelet domain, as well as their means and variances, are quite similar.

Wavelet	Max entropy	Mean entropy	Entropy variance
Coiflet 1	2.9707	2.3599	0.1590
Daubechies 4	3.0463	2.4323	0.1372
Daubechies 6	3.0142	2.5804	0.0918
Daubechies 8	2.8991	2.4430	0.0793
Symmlet 4	3.0343	2.4620	0.1326

the iterative method to diverge and deviate from the actual values at different time points.

For the second step in choosing the most suitable wavelet function, considerations of convergence necessitate the removal of those for which the iterative method diverges on the other extreme of the spectrum of smoothness. This restriction leaves us with a pool of candidates from which we are free to select the most appropriate functions for our purposes. If we only consider the three aforementioned families, the remaining pool of plausible candidates consists of *Coiflet* 1, *Symmlet* 4, and *Daubechies* 6 and 8. The smoothed datasets using all of these wavelet functions are depicted in Figure 10Figure 13. Other families of wavelets may also be considered. Again, it is then necessary to eliminate family members that are not differentiable as well as the ones for which CIWS converges.

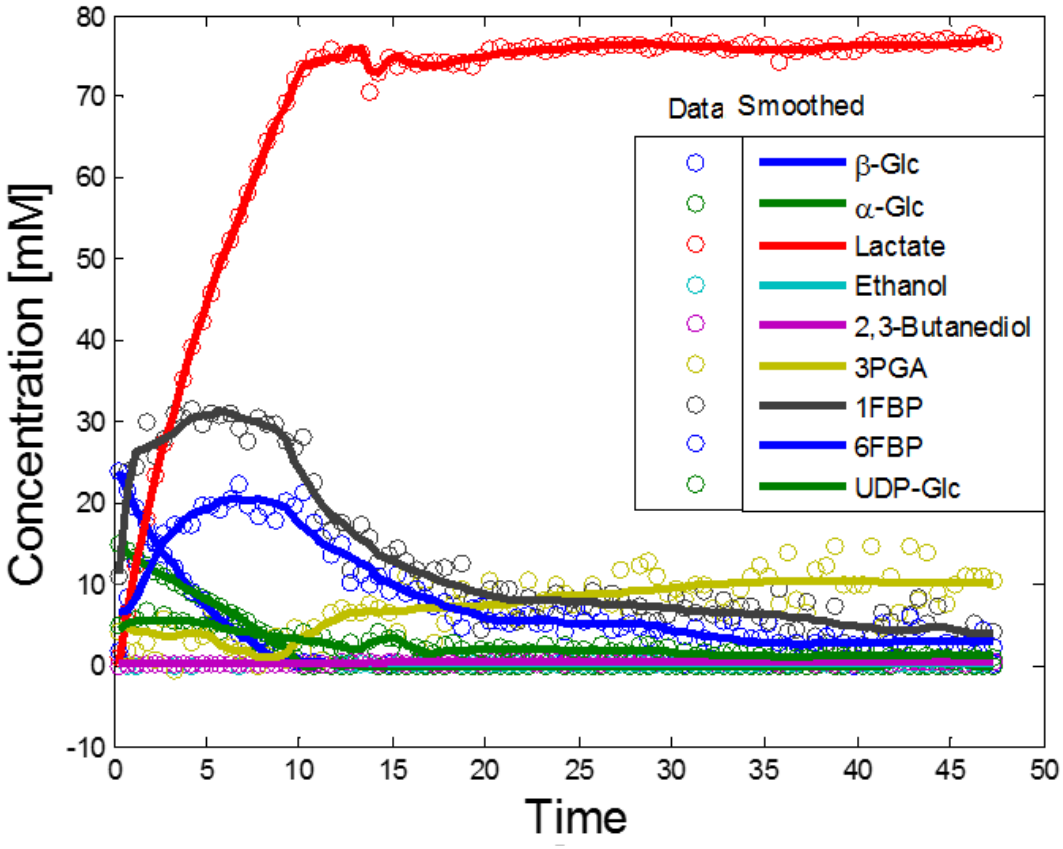

Figure 31: The result of smoothing using *Coiflet* 1, which converges in 28 iterations for $\epsilon = 0.1$.

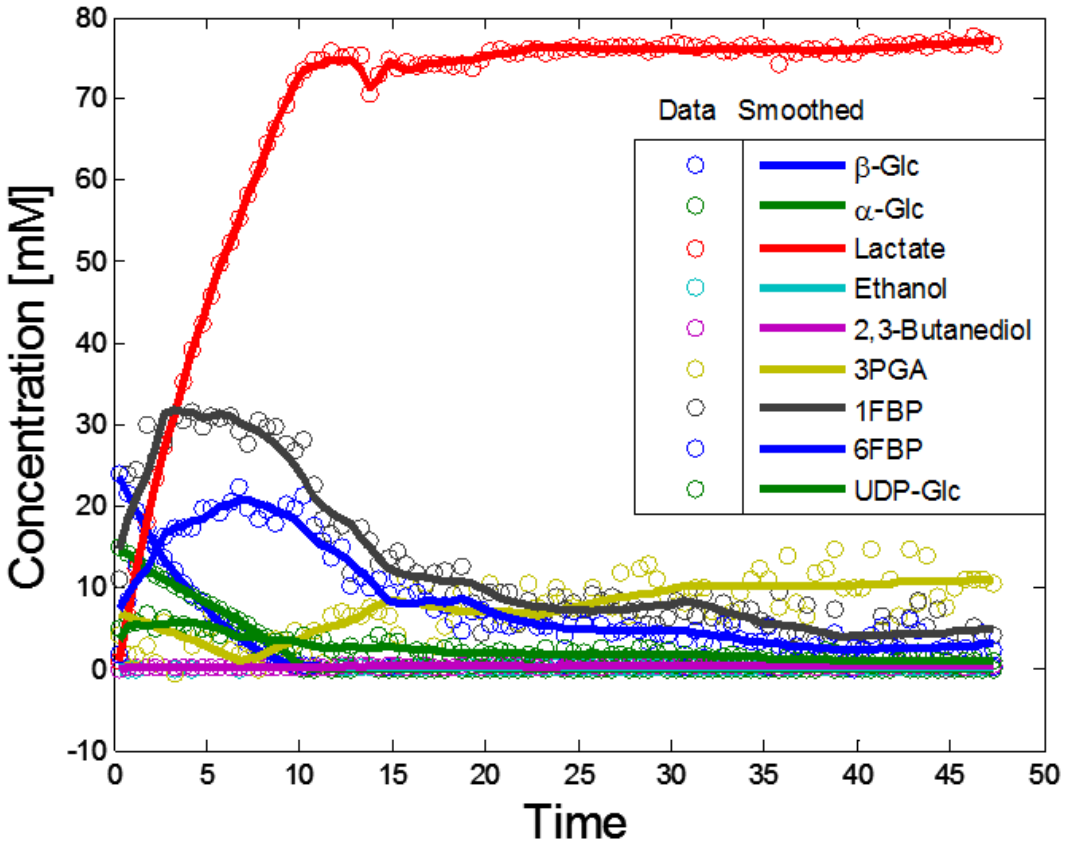

Figure 32: The result of smoothing using *Daubechies 6*, which converges in 6 iterations for $\epsilon = 0.1$.

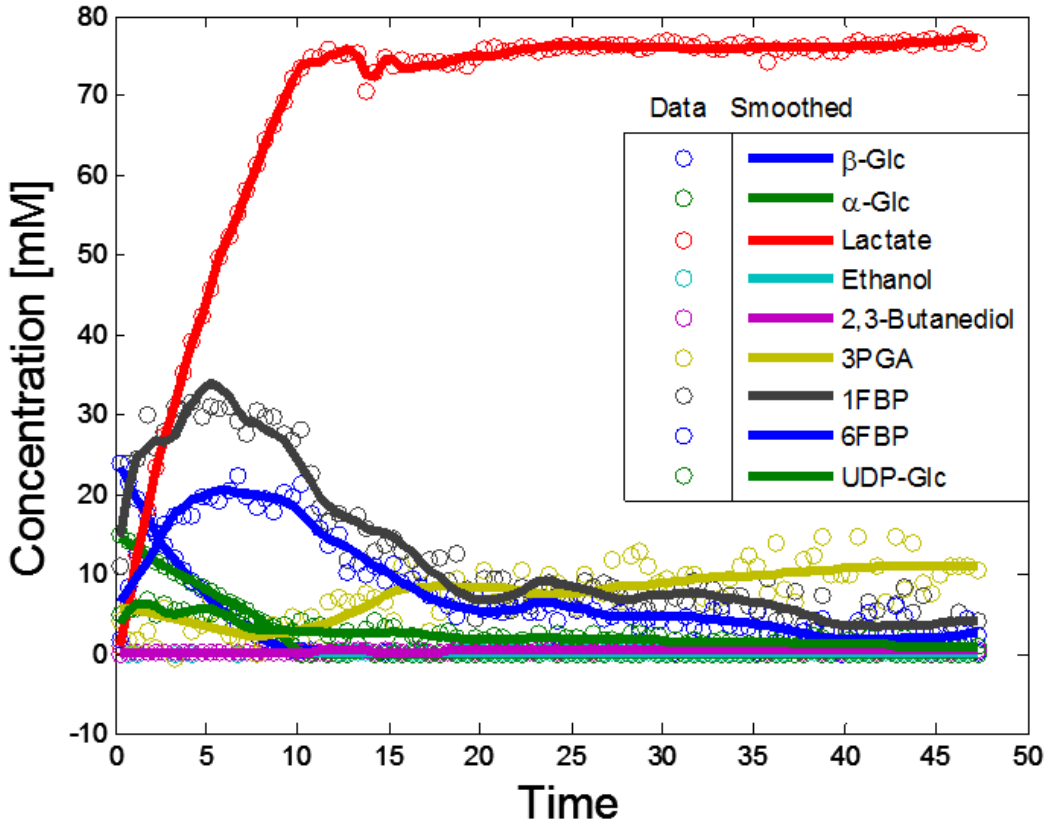

Figure 33: The result of smoothing using *Daubechies 8*, which converges in 6 iterations for $\epsilon = 0.1$.

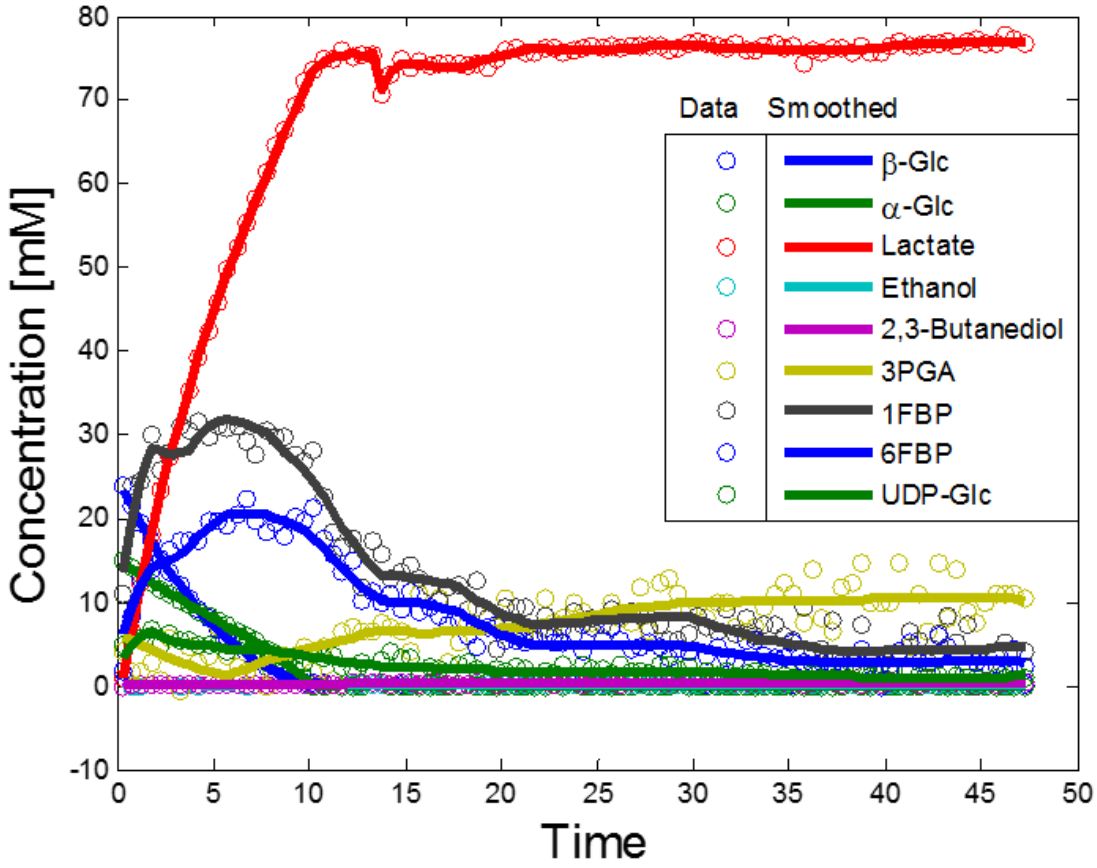

Figure 34: The result of smoothing using *Symmlet* 4 wavelet, which converges in 7 iterations for $\epsilon = 0.1$. Entropy was utilized as the criterion for choosing between the remaining functions in Section 5.2.3. Table 3 shows the admissible wavelets along with the maximum entropy in the wavelet domain among different time series of metabolite concentrations for each wavelet. Average entropy and the corresponding variance is also included. Among the admissible wavelets *Coiflet* 1 exhibits the lowest average and maximum entropy and was thus chosen as the wavelet of choice.

APPENDIX B

MATLAB CODES

The MATLAB code for the CIWS technique is provided below. The test data "wt104_ana40_1.mat" is also provided.

```
function y = CIWS(x,Type,Par,epsilon,Stoi_Vol_Vect)
% First we need to install the Wavelab toolbox from
%http://statweb.stanford.edu/~wavelab/Wavelab_850/download.html
% Then Run  WavePath only for the first time you run this code in a
%MATLAB session

%  USAGE
%    y = CIWS(x,Type,Par,epsilon,exvol,invol)
%  INPUTS
%    x       The time series data with the first column representing time
%            and each of the following columns representing another time
%            series of data.
%    Type    string, 'Haar', 'Coiflet', 'Daubechies',
%            'Symmlet', 'Vaidyanathan','Battle', 'Lemarie','Pollen'
%    Par     integer or angle(s), it is a parameter related to either
%            the support and vanishing moments of the wavelets or
%            the angle, explained below for each wavelet.

%    epsilon   low-pass filter corresponding to orthogonal WT

%    Stoi_Vol_Vect The vector which includes the Stoichiometric as
%         well as volume information for compartmental models
```

```
%  OUTPUT
%  y      The smoothed data.
%

%  DESCRIPTION
%   The algorithm was tested on the data matrix which had the following
%   columns:
%   1)Time(min) Glucose: <-- 2)betta-glc    3)alpha-glc 4)Lactate ...
%   5)Ethanol   6)2,3-But 7)3PGA 8)1FBP + 9)6FBP -->FBP 10)UDP-Glc
%   The following lines were added before calling the algorithm:
%       load 'wt104_ana40_1.mat'
%       x=wt104_ana40_1;
%       exvol = 50;
%       invol = 2.15;
%       Stoi_Vol_Vect = [2*exvol 2*exvol exvol exvol exvol invol
%                                    2*invol 2*invol 2*invol];
%       epsilon = 0.1;
%       Type = 'Coiflet';
%       Par = 1;
%       x_smoothed = CIWS(x,Type,Par,epsilon,Stoi_Vol_Vect);

% Extract number of data points from the input data matrix x
N = 2^(ceil(log2(size(x,1))));
stoi_vol = repmat(Stoi_Vol_Vect,N,1);

% Make the data matrix of size 2^N: Filling by mirroring
x_mirrored = zeros(N,size(x,2)-1); %minus 1: time
x_mirrored(1:size(x,1),:) = x(:,2:end);
x_mirrored(size(x,1)+1:end,:) = x(end:-1:2*size(x,1)-N+1, 2:end);

x_data = x_mirrored;
x_mirrored = stoi_vol .* x_mirrored;
```

```
%Introducing the variable z to account for the lost data(keeps the mass
%constant)
tot_mass = max(sum(x_mirrored,2)); %total mass equals to the
% maximum of the sum
z = tot_mass*ones(N,1)-sum(x_mirrored,2);
exvol = 50;
x_data = [x_data z/exvol]; %z is considered as an internal(inside the
% cell) metablite
stoi_vol = [stoi_vol exvol*ones(N,1)];

% Choose the family of wavelet and make the corresponding quadrature
% mirror filter
if strncmpi(Type,'Haar',2),
    wf = MakeONFilter('Haar',Par);
end

if strncmpi(Type,'Coiflet',2),
    wf = MakeONFilter('Coiflet',Par);
end

if strncmpi(Type,'Daubechies',3),
    wf = MakeONFilter('Daubechies',Par);
end

if strncmpi(Type,'Symmlet',3),
    wf = MakeONFilter('Symmlet',Par);
end
if strncmpi(Type,'Vaidyanathan',2),
    wf = MakeONFilter('Vaidyanathan',Par);
end
```

```
if strncmpi(Type,'Battle',2),
    wf = MakeONFilter('Battle',Par);
end

if strncmpi(Type,'Lemarie',2),
    wf = MakeONFilter('Lemarie',Par);
end

if strncmpi(Type,'Pollen',2),
    wf = MakeONFilter('Pollen',Par);
end

% Transformation Matrix of FWT_PO: W = WavMat(h, N, k0, shift).
% Here, N is the size of matrix/length of data, which should be
% a power of 2.
% K0 is the depth of transformation. Ranges from 1 to J=log2(N).
%  Default is J. We use J = log2(N)-2.
% shift:  the matrix is not unique an any integer shift gives. We use 0.
W = Wavmat(wf,N,log2(N)-2,0); %N x N transformation matrix


%% Iterative method:
g1 = zeros(size(x_data,1),1); %number of data points in time
g0 = ones(size(x_data,1),1) * tot_mass;

j = 0;
while norm(g1-g0)> epsilon && j<=1500
    j = j+1;
    D =  W * x_data; %This is the vector containing all coefficients

    % Standard deviation estimation:
    % sd = std(D)
```

```
    sd = mad(D,1)/0.6745; % This is more robust and less prone to
    %  outliers influence than the MATLAB std function.

    %Threshold: according to universal thresholding
    lam = repmat(sqrt(2*log(N)).* sd,N,1);

    % Choice of thresholding rule:
    D_th = D .*(abs(D)> lam); %Hard thresholding
    %D_th = (abs(D)-lam.*ones(size(D))).*(abs(D)> lam); %Soft threshold
    %D_th = (D-lam.^2.*D.^(-1)).*(abs(D)> lam); %Non-negative garrotte
    % thresholding
    x_s = W'*D_th ;
    x_smoothed = x_s .* (x_s >= 0);
    g1 = sum(stoi_vol .* x_smoothed,2);
    g = repmat(g1,1,size(x_data,2)); %Mass balance matrix
    x_data = tot_mass * x_smoothed./g;
end

if j>=1500
    disp('CIWS does not converge!')
else
    A = Type;
    A1 = [Par;j];
    formatSpec = 'CIWS algorithm using the %s %u converged in %u
    iterations:')
    fprintf(formatSpec,A,A1)
end


y = x_smoothed;

figure(1)
```

```
tmph = plot(x(:,1),[x(:,2:end)],'--o');
set(tmph,'linewidth',1.4);
legend('\beta-Glc','\alpha-Glc','Lactate','Ethanol','2,3-Butanediol', ...
    '3PGA','1FBP', '6FBP','UDP-Glc');
hold on
tmph = plot(x(:,1),x_smoothed(1:size(x,1),1:end-1));
set(tmph,'linewidth',3);
xlabel('Time', 'fontsize',16);
ylabel('Concentration [mM]','fontsize',16);
set(gca, 'fontsize', 10);
hold off


save smoothed x x_smoothed


end
```

REFERENCES

[1] Braat, H., Rottiers, P., Hommes, D. W., Huyghebaert, N., Remaut, E., Remon, J., van Deventer, S. J., Neirynck, S., Peppelenbosch, M. P., and Steidler, L., "A phase i trial with transgenic bacteria expressing interleukin-10 in crohns disease," *Clinical Gastroenterology and Hepatology*, vol. 4, no. 6, pp. 754 – 759, 2006.

[2] Breiman, L., "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, pp. 373–384, Nov. 1995. Technometrics.

[3] Buckheit, J. and Donoho, D., "WaveLab and reproducible research," in *Wavelets and Statistics* (Antoniadis, A. and Oppenheim, G., eds.), vol. 103 of *Lecture Notes in Statistics*, pp. 55–81, Springer New York, 1995.

[4] Budin-Verneuil, A., Pichereau, V., Auffray, Y., Ehrlich, D., and Maguin, E., "Proteome phenotyping of acid stress-resistant mutants of *Lactococcus lactis* mg1363," *PROTEOMICS*, vol. 7, no. 12, pp. 2038–2046, 2007.

[5] Budin-Verneuil, A., Pichereau, V., Auffray, Y., Ehrlich, D., and Maguin, E., "Proteomic characterization of the acid tolerance response in *Lactococcus lactis* mg1363," *PROTEOMICS*, vol. 5, no. 18, pp. 4794–4807, 2005.

[6] Carvalho, A. L., Turner, D. L., Fonseca, L. L., Solopova, A., Catarino, T., Kuipers, O. P., Voit, E. O., Neves, A. R., and Santos, H., "Metabolic and transcriptional analysis of acid stress in *Lactococcus lactis*, with a focus on the kinetics of lactic acid pools," *Plos One*, vol. 8, p. 14, July 2013.

[7] Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D., "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. D459–D471, 2014.

[8] Castro, R., Neves, A. R., Fonseca, L. L., Pool, W. A., Kok, J., Kuipers, O. P., and Santos, H., "Characterization of the individual glucose uptake systems of *Lactococcus lactis*: mannose-PTS, cellobiose-PTS and the novel GlcU permease," *Molecular Microbiology*, vol. 71, pp. 795–806, Feb. 2009. Mol Microbiol.

[9] Chou, I.-C. and Voit, E. O., "Recent developments in parameter estimation and structure identification of biochemical and genomic systems," *Math Biosci*, vol. 219, pp. 57–83, June 2009. Mathematical biosciences.

[10] Chou, I. C. and Voit, E. O., "Estimation of dynamic flux profiles from metabolic time series data," *BMC Syst Biol*, vol. 6, p. 84, 2012. BMC systems biology.

[11] Clarke, J. B., Birch, M., and Britton, H. G., "The equilibrium constant of the phosphoglyceromutase reaction," *Biochemical Journal*, vol. 139, no. 3, pp. 491–497, 1974.

[12] Coifman, R. R. and Wickerhauser, M. V., "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, pp. 713–718, Mar. 1992. Ieee T Inform Theory.

[13] Collins, L. B. and Thomas, T. D., "Pyruvate kinase of streptococcus lactis," *J Bacteriol*, vol. 120, pp. 52–8, Oct. 1974. Journal of bacteriology.

[14] Crow, V. L. and Pritchard, G. G., "Purification and properties of pyruvate kinase from Streptococcus lactis," *Biochim. Biophys. Acta*, vol. 438, pp. 90–101, Jun 1976.

[15] Crow, V. L. and Pritchard, G. G., "Fructose 1,6-diphosphate-activated l-lactate dehydrogenase from *Streptococcus lactis*: kinetic properties and factors affecting activation.," *Journal of Bacteriology*, vol. 131, no. 1, pp. 82–91, 1977.

[16] Crow, V. L. and Pritchard, G. G., "Pyruvate kinase from *Streptococcus lactis*," *Methods Enzymol*, vol. 90 Pt E, pp. 165–70, 1982. Crow, V L Pritchard, G G Journal Article United states Methods Enzymol. 1982;90 Pt E:165-70.

[17] Crow, V. L. and Pritchard, G. G., "The effect of monovalent and divalent cations on the activity of streptococcus lactis {C10} pyruvate kinase," *Biochimica et Biophysica Acta (BBA) - Enzymology*, vol. 481, no. 1, pp. 105 – 114, 1977.

[18] Curien, G., Bastien, O., Robert-Genthon, M., Cornish-Bowden, A., Cárdenas, M. L., and Dumas, R., "Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters," *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[19] Dannelly, H. K. and Roseman, S., "Nad^+ and nadh regulate an atp-dependent kinase that phosphorylates enzyme i of the *Escherichia coli* phosphotransferase system," *Proc Natl Acad Sci U S A*, vol. 89, no. 23, pp. 11274–6, 1992. Dannelly, H K Roseman, S GM38759/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states Proc Natl Acad Sci U S A. 1992 Dec 1;89(23):11274-6.

[20] Dannelly, H. K. and Roseman, S., "Active site phosphorylation of enzyme i of the bacterial phosphotransferase system by an atp-dependent kinase," *Journal of Biological Chemistry*, vol. 271, no. 25, pp. 15285–15291, 1996.

[21] Dantzig, G. B., "Reminiscences about the origins of linear programming," *Operations Research Letters*, vol. 1, no. 2, pp. 43 – 48, 1982.

[22] Daubechies, I., *Ten lectures on wavelets*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, 1992.

[23] Dolatshahi, S., Vidakovic, B., and Voit, E. O., "A constrained wavelet smoother for pathway identification tasks in systems biology," *Computers & Chemical Engineering*, vol. 71, no. 0, pp. 728 – 733, 2014.

[24] Donoho, D. L. and Johnstone, I. M., "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, Sept. 1994. Biometrika.

[25] Draper, N. R. and Smith, H., *Serial Correlation in the Residuals and the DurbinWatson Test*, pp. 179–203. John Wiley & Sons, Inc., 1998.

[26] Eilers, P. H. C., "A perfect smoother," *Analytical Chemistry*, vol. 75, pp. 3631–3636, July 2003. Anal Chem.

[27] Galazzo, J. L. and Bailey, J. E., "Fermentation pathway kinetics and metabolic flux control in suspended and immobilized saccharomyces-cerevisiae," *Enzyme and Microbial Technology*, vol. 12, pp. 162–172, Mar. 1990. Enzyme Microb Tech.

[28] Garrigues, C., Loubiere, P., Lindley, N. D., Cocaign-bousquet, M., Bacteriol, J., Garrigues, C., Loubiere, P., Lindley, N. D., and Cocaign-bousquet, M., "Control of the shift from homolactic acid to mixed-acid fermentation in *Lactococcus lactis*: predominant role of the nadh/nad$^+$ ratio," *J*, pp. 179–5282, 1997.

[29] Gaspar, P., Carvalho, A. L., Vinga, S., Santos, H., and Neves, A. R., "From physiology to systems metabolic engineering for the production of biochemicals by lactic acid bacteria," *Biotechnology Advances*, vol. 31, no. 6, pp. 764 – 788, 2013. Bioenergy and Biorefinery from Biomass through innovative technology development.

[30] GENNEMARK, P. and WEDELIN, D., "Odeion a software module for structural identification of ordinary differential equations," *Journal of Bioinformatics and Computational Biology*, vol. 12, no. 01, p. 1350015, 2014. PMID: 24467754.

[31] Goel, G., Chou, I. C., and Voit, E. O., "System estimation from metabolic time-series data," *Bioinformatics*, vol. 24, pp. 2505–11, Nov. 2008. Bioinformatics.

[32] Goel, G., *Dynamic flux estimation - a novel framework for metabolic pathway analysis*. Ph.d. dissertation, Georgia Institute of Technology, Aug. 2009.

[33] Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., and Sethna, J. P., "Extracting falsifiable predictions from sloppy models," *Ann N Y Acad Sci*, vol. 1115, pp. 203–11, Dec. 2007. Annals of the New York Academy of Sciences.

[34] Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P., "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput Biol*, vol. 3, pp. 1871–78, Oct. 2007. PLoS computational biology.

[35] Hoefnagel, M. H., van der Burgt, A., Martens, D. E., Hugenholtz, J., and Snoep, J. L., "Time dependent responses of glycolytic intermediates in a detailed glycolytic model of *Lactococcus lactis* during glucose run-out experiments," *Mol Biol Rep*, vol. 29, no. 1-2, pp. 157–61, 2002. Molecular biology reports.

[36] Iwata, M., Shiraishi, F., and Voit, E. O., "Course but efficient identification of metabolic pathway systems," *International Journal of Systems Biology*, vol. 4, no. 1, pp. 57–72, 2013.

[37] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M., "Data, information, knowledge and principle: back to metabolism in kegg," *Nucleic Acids Research*, vol. 42, no. D1, pp. D199–D205, 2014.

[38] Kilstrup, M., *Proteomics of Lactococcus lactis: Phenotypes for a Domestic Bacterium*, pp. 149–178. John Wiley & Sons, Inc., 2005.

[39] Kuipers, O., de Jong, A., Baerends, R., van Hijum, S., Zomer, A., Karsens, H., den Hengst, C., Kramer, N., Buist, G., and Kok, J., "Transcriptome analysis and related databases of *Lactococcus lactis*," *Antonie van Leeuwenhoek*, vol. 82, no. 1-4, pp. 113–122, 2002.

[40] Mallat, S. G., "A theory for multiresolution signal decomposition - the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, July 1989. Ieee T Pattern Anal.

[41] Mallat, S. G., *A Wavelet Tour of Signal Processing (Third Edition)*. Boston: Academic Press, third edition ed., 2009.

[42] Malmstrom, B. G., *The Enzymes*, vol. V, book section 471. New York and London: Academic Press, 2nd ed., 1961.

[43] Marino, S. and Voit, E. O., "An automated procedure for the extraction of metabolic network information from time series data," *J Bioinform Comput Biol*, vol. 4, pp. 665–91, June 2006. Journal of bioinformatics and computational biology.

[44] Muiño, J. M., Voit, E. O., and Sorribas, A., "GS-distributions: A new family of distributions for continuous unimodal variables," *Computational Statistics & Data Analysis*, vol. 50, pp. 2769–2798, June 2006. Comput Stat Data An.

[45] Neves, A. R., *Metabolic Strategies to Reroute Carbon Fluxes in Lactococcus lactis: Kinetics of Intracellular Metabolite Pools by in vivo NMR*. Ph.d. dissertation, Universidade Nova de Lisboa, Sept. 2001.

[46] Neves, A. R., Ramos, A., Costa, H., van, S., Hugenholtz, J., Kleerebezem, M., de Vos, W., and Santos, H., "Effect of different NADH oxidase levels on glucose metabolism by *Lactococcus lactis*: kinetics of intracellular metabolite pools determined by *in vivo* nuclear magnetic resonance," *Appl Environ Microbiol*, vol. 68, pp. 6332–42, Dec. 2002. Applied and environmental microbiology.

[47] Neves, A. R., Ramos, A., Nunes, M. C., Kleerebezem, M., Hugenholtz, J., de Vos, W. M., Almeida, J., and Santos, H., "*in vivo* nuclear magnetic resonance studies of glycolytic kinetics in *Lactococcus lactis*," *Biotechnol Bioeng*, vol. 64, pp. 200–12, July 1999. Biotechnology and bioengineering.

[48] Neves, A. R., Ventura, R., Mansour, N., Shearman, C., Gasson, M. J., Maycock, C., Ramos, A., and Santos, H., "Is the glycolytic flux in *Lactococcus lactis* primarily controlled by the redox charge?: Kinetics of nad^+ and nadh pools determined *IN VIVO* by ^{13}c nmr," *Journal of Biological Chemistry*, vol. 277, no. 31, pp. 28088–28098, 2002.

[49] Poolman, B., Hellingwerf, K. J., and Konings, W. N., "Regulation of the glutamate-glutamine transport system by intracellular pH in streptococcus lactis," *J Bacteriol*, vol. 169, pp. 2272–6, May 1987. Journal of bacteriology.

[50] Resnikoff, H. L. and Wells, R. O., *Wavelet Analysis: The Scalable Structure of Information*. Springer, corrected ed., 2002.

[51] ROSE, I. A. and ROSE, Z. B., "Chapter {III} - glycolysis: Regulation and mechanisms of the enzymes*," in *Carbohydrate Metabolism* (FLORKIN, M. and STOTZ, E. H., eds.), vol. 17 of *Comprehensive Biochemistry*, pp. 93 – 161, Elsevier, 1969.

[52] Saier, M. H., Schmidt, M. R., and Lin, P., "Phosphoryl exchange reaction catalyzed by enzyme i of the bacterial phosphoenolpyruvate: sugar phosphotransferase system. kinetic characterization," *J Biol Chem*, vol. 255, pp. 8579–84, Sept. 1980. The Journal of biological chemistry.

[53] Sands, P. and Voit, E., "Flux-based estimation of parameters in s-systems," *Ecological Modelling*, vol. 93, no. 13, pp. 75 – 88, 1996.

[54] Savageau, M. A., "Biochemical systems analysis. i. some mathematical properties of the rate law for the component enzymatic reactions," *J Theor Biol*, vol. 25, pp. 365–9, Dec. 1969. Journal of theoretical biology.

[55] Savageau, M. A., *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Reading, Mass: Addison-Wesley, 1976.

[56] Savageau, M. A. and Voit, E. O., "Recasting nonlinear differential equations as s-systems: a canonical nonlinear form," *Mathematical Biosciences*, vol. 87, no. 1, pp. 83 – 115, 1987.

[57] Solem, C., Dehli, T., and Jensen, P. R., "Rewiring *Lactococcus lactis* for ethanol production," *Applied and Environmental Microbiology*, vol. 79, no. 8, pp. 2512–2518, 2013.

[58] Srinath, S. and Gunawan, R., "Parameter identifiability of power-law biochemical system models," *J Biotechnol*, vol. 149, pp. 132–40, Sept. 2010. Journal of biotechnology.

[59] Strang, G., "Wavelets and dilation equations - a brief introduction," *Siam Review*, vol. 31, pp. 614–627, Dec. 1989. Siam Rev.

[60] Thomas, T. D., "Activator specificity of pyruvate kinase from lactic *Streptococci*," *Journal of Bacteriology*, vol. 125, no. 3, pp. 1240–1242, 1976.

[61] Varah, J. M., "A spline least squares method for numerical parameter estimation in differential equations," *SIAM J. Sci. Stat. Comput.*, vol. 3, no. 2, pp. 28–46, 1982.

[62] Vidakovic, B., *Statistical modeling by wavelets*. New York: Wiley, 1999.

[63] Vilela, M., Borges, C. C., Vinga, S., Vasconcelos, A. T., Santos, H., Voit, E. O., and Almeida, J. S., "Automated smoother for the numerical decoupling of dynamics models," *BMC Bioinformatics*, vol. 8, p. 305, 2007.

[64] Vilela, M., Vinga, S., Maia, M. A., Voit, E. O., and Almeida, J. S., "Identification of neutral biochemical network models from time series data," *BMC Syst Biol*, vol. 3, p. 47, 2009. BMC systems biology.

[65] Villena, J., Medina, M., Raya, R., and Alvarez, S., "Oral immunization with recombinant *Lactococcus lactis* confers protection against respiratory pneumococcal infection," *Can J Microbiol*, vol. 54, pp. 845–53, Oct. 2008. Canadian journal of microbiology.

[66] Voit, E., Neves, A. R., and Santos, H., "The intricate side of systems biology," *Proc Natl Acad Sci U S A*, vol. 103, pp. 9452–7, June 2006. Proceedings of the National Academy of Sciences of the United States of America.

[67] Voit, E. O., *Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity*. New York: Van Nostrand Reinhold, 1991.

[68] Voit, E. O., "What if the fit is unfit? criteria for biological systems estimation beyond residual errors," in *Applied Statistics for Biological Networks* (Dehmer, M., Emmert-Streib, F., and Salvador, A., eds.), pp. 183–200, New York: J. Wiley and Sons, 2011.

[69] Voit, E. O., "Biochemical systems theory: A review," *ISRN Biomathematics*, 2013.

[70] Voit, E. O., "Characterizability of metabolic pathway systems from time series data," *Mathematical Biosciences*, vol. 246, pp. 315–325, Dec. 2013. Math Biosci.

[71] Voit, E. O. and Almeida, J., "Decoupling dynamical systems for pathway identification from metabolic profiles," *Bioinformatics*, vol. 20, pp. 1670–1681, July 2004.

[72] Voit, E. O., Almeida, J., Marino, S., Lall, R., Goel, G., Neves, A. R., and Santos, H., "Regulation of glycolysis in *Lactococcus lactis*: an unfinished systems biological case study," *Syst Biol (Stevenage)*, vol. 153, pp. 286–98, July 2006. Systems biology.

[73] Voit, E. O. and Savageau, M. A., "Power-law approach to modeling biological systems; II. application to ethanol production," *J Ferment Technol*, vol. 60, no. 3, pp. 229–232, 1982.

[74] Voit, E. O. and Savageau, M. A., "Power-law approach to modeling biological systems; III. methods of analysis," *J Ferment Technol*, vol. 60, no. 3, pp. 223–241, 1982.

[75] Voit, E. O. and Ferreira, A. E. N., *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*. Cambridge ; New York: Cambridge University Press, 2000.

[76] Voit, E., Goel, G., Chou, L.-C., and Fonseca, L., "Estimation of metabolic pathway systems from different data sources," *Systems Biology, IET*, vol. 3, pp. 513–522, Nov 2009.

[77] Walter, G. G., Shen, X., and Walter, G. G., *Wavelets and other orthogonal systems*. Boca Raton: Chapman & Hall/CRC, 2nd ed., 2001.

[78] Whittaker, E. T., "On a new method of graduation," *Proceedings of the Edinburgh Mathematical Society*, vol. 41, pp. 63–75, 2 1922.

VITA

SEPIDEH DOLATSHAHI

Sepideh Dolatshahi was born in Tonekabon, Mazandaran, Iran. She received her undergraduate degree in electrical and computer engineering from University of Tehran, Iran in 2007. She moved to the United States to pursue her master's degree at University of Massachusetts, Amherst. After receiving her M.S. in electrical and computer engineering, telecommunications and signal processing, she moved to Atlanta for her doctorate studies at Georgia Institute of Technology. She became interested in the field of systems and computational biology and started her interdisciplinary research in Dr. Eberhard Voit's Laboratory for Biological Systems Analysis in fall 2010. She earned a second master's degree in bioengineering from Georgia Tech in 2013 and finally her doctorate degree in August 2015. When she is not working on her research, Sepideh enjoys hiking, swimming, reading, traveling, and spending time with her family and friends.