Analyzing the Fluxome of *P. chrysogenum* in an Industrial Environment

Workflows for ^{13}C Metabolic Flux Analysis in Complex Systems

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Sebastian Niedenführ

Berichter: Univ. Prof. Dr. rer. nat. Wolfgang Wiechert
Univ. Prof. Dr.-Ing. Ralf Takors

Tag der mündlichen Prüfung: 12.03.2014

Kurzfassung

Metabolic Engineering ist eine gerichtete und wissensbasierte Methode, um die Produktionseigenschaften von Mikroorganismen zu verbessern. Sie zielt darauf ab, die Produktbildung der Mikroorganismen zu erhöhen, um ökonomische Produktionsprozesse zu schaffen. Das *Fluxom,* die Gesamtheit metabolischer Reaktionsraten, des Mikroorganismus wird durch mathematische Modellierung abgebildet. Das leistungsfähigste Werkzeuge der Fluxomanalyse, die ^{13}C metabolische Stoffflussanalyse (^{13}C-MFA), verwendet isotopisch markierte Substrate zur Kultivierung von Mikroorganismen. Die nach der Aufnahme und metabolischen Umsetzung entstehende Markierungsmuster in den Metaboliten werden durch hoch präzise Messapparate (zum Beispiel Massenspektrometer) quantifiziert. Aus diesen Messungen können die intrazellulären Reaktionsraten abgeschätzt werden.

Die vorgelegte Arbeit befasst sich mit der ^{13}C-MFA in komplexen Systemen. Dazu wurde *P. chrysogenum* BCB1 verwendet und unter industriellen Bedingungen die Bildung des Antibiotikums von Penicillin V untersucht. "Komplex" bezeichnet hier nicht nur das Wachstumsverhalten von *P. chrysogenum,* sondern auch die komplexe Nebenproduktbildung und Kompartimentierung des Metabolismus. Zudem zielt "komplex" auf die Übertragung der ^{13}C-MFA von einer wissenschaftlichen Anwendung unter idealen Bedingungen zu einem industriellen Standard ab. Aus diesem Grund werden Voraussetzung und Annahmen für die Durchführung der Technik diskutiert und die notwendigen Anpassungen für den Einsatz im industriellen Umfeld beschrieben.

Um gesicherte Erkenntnis über den Organismus zu erlangen, wird der Arbeitsablauf zur Durchführung der ^{13}C-MFA vorgestellt. Durch die Untersuchungen werden die Defizite und Limitierungen der Technik aufgedeckt. Zuerst wurde die Technik in Chemostat Experimenten etabliert und auf *Fed-batch* Kultivierungen nahe am industriellen Prozess übertragen. Als Resultat liegen erstmals Stoffflusskarten unter diesen Prozessbedingungen vor.

Die beiden Prozesse wurden mittels kinetischer Modellierung abgebildet um extrazelluläre Raten für die ^{13}C-MFA abzuschätzen. Um die ^{13}C-MFA anwenden zu können, mussten die zeitaufgelösten Markierungsanreicherungen extrapoliert und um die natürliche Markierung korrigiert werden. In der Arbeit wurden Toolboxen zur nichtlinearen Regression und globalen Sensitivitätsanalyse entwickelt.

Großskalige metabolische Modelle wurden für *P. chrysogenum* aufgebaut, wobei experimentelle Daten und Datenbanken, sowie Literatur-Quellen, zum Einsatz kamen. Stammspezifische Messungen der Biomassezusammensetzung wurden in das Modell integriert. Um die Modelle auf ihre Anwendbarkeit zur Flussschätzung zu prüfen, wurde erstmals eine globale Sensitivitätsanalyse für die ^{13}C-MFA durchgeführt. Zudem wurde diese Technik auf die zuvor beschriebenen Bioprozesse angewendet um die intrazellulären Reaktionsraten zu berechnen. Die erhaltenen, statistisch hochaufgelösten, Stoffflusskarten wurden weiter verbessert durch den Einsatz gezielter Versuchsplanung.

Dazu wurden die konventionellen Versuchsplanungstechniken um die optimale Planung von sequentiellen Experimenten und Mehrzieloptimierung erweitert. Durch letzteres konnten optimale, und dennoch ökonomische, Versuche geplant und gleichzeitig die Nachteile konventioneller Techniken vermieden werden.

Die Interpretation der generierten Stoffflusskarten zeigte, dass neben der Penicillinproduktion das Wachstum einen größeren Einfluss auf den oxidativen Pentosephosphatweg besitzt als bisher in der Literatur diskutiert. Dies führt zu einer sorgfältigen Ausbalancierung von Wachstum im Produktionsprozess und der gleichzeitigen Stammoptimierung.

Summary

Metabolic engineering is a targeted and knowledge-based approach to improve production capabilities of microorganisms. It aims at increasing metabolic reaction rates towards product formation to obtain economic production processes. The fluxome, i.e. the computation of all metabolic reaction rates, is one cornerstone of metabolic engineering. For fluxomics, the study of intracellular reaction rates, several methods have been established. The most powerful one, ^{13}C-metabolic flux analysis (^{13}C-MFA), uses isotopically labeled substrates which are fed to cells. Emerging labeling patterns in the synthesized metabolites are measured by high-precision measurement devices like mass spectrometry. From the measured labeling pattern, the intracellular reaction rates can be estimated by mathematical modeling.

The present work faces ^{13}C-MFA of complex systems. The non-model organism *P. chrysogenum* strain BCB1 is investigated in industrial environment with special focus on penicillin V production. The term "complex" is not only referring to the growth behavior of *P. chrysogenum*, but also includes side-product formation and compartmentalization of metabolism. This thesis aims at the transfer of ^{13}C-MFA from a scientific application in a nearly ideal environment to an industrial standard. For this reason, the prerequisites needed and compromises made for ^{13}C-MFA work-flow are thoroughly discussed and adaption of the industrial process is highlighted.

To systematically gain knowledge about the organism, the state-of-the-art work-flow for ^{13}C-MFA is presented, pitfalls and limitations of the technique are revealed for a close-to-industrial context. The technology is established using chemostat experiments. In a second step, close-to-industrial fed-batch cultivations are investigated and the first quantitative flux map of *P. chrysogenum* for industrial process conditions is presented.

Therefore, a kinetic model was implemented for the processes aiming at an accurate extracellular rate estimation. To apply ^{13}C-MFA, stationary labeling patterns are derived from time resolved labeling data by extrapolation and correction for natural abundance.

Large scale metabolic models for *P. chrysogenum* were built based on experimental, literature and database knowledge. Strain specific measurements of biomass compounds were introduced into the models. Using the constructed model, the first global sensitivity analysis was performed for ^{13}C-MFA to evaluate its suitability for flux elucidation. Finally, ^{13}C-MFA was conducted to gain knowledge about intracellular fluxes within *P. chrysogenum* BCB1 and experimental design was used to increase the information content of isotope labeling experiments.

The conventional experimental design tools were extended by diversification-driven and multi-objective experimental design. The design space was explored and compared to single-objective applications. Thereby optimal, yet economic, experimental designs can be planed, fighting shortcomings of conventional techniques.

From the results of the deduced flux maps, hints for strain and process development are derived. One major finding was that the flux in oxidative pentose phosphate pathway is strongly influenced by the biomass formation, leading to carefully balanced growth in cultivations and strain optimization.

Publications and Further Work

Courses

The *1st* and *2nd Advanced Course on* ^{13}*C-based Metabolic Flux Analysis* (each 5 day long) were run in 2012 and 2013, respectively. Materials were prepared for oral presentation and exercises. Most lectures were given. The course was prepared with Katharina Nöh, Birgit Stute and co-workers.

Papers

- Beste, D.; Nöh, K.; Niedenführ, S.; Mendum, T.; Hawkins, N.; Ward, J.; Beale, M.; Wiechert, W.; McFadden J.
  Carbon fixation and a mixed diet for the intracellular tuberculosis bacillus.
  Cell: Chemistry and Biology (accepted)

- Weitzel, M.; Nöh, K.; Dalman, T.; Niedenführ, S.; Stute, B.; Wiechert, W.
  13CFLUX2-high-performance software suite for ^{13}*C-metabolic flux analysis.*
  Bioinformatics 29, 1 (2013), pp. 143–145.

- Droste, P.; Miebach, S.; Niedenführ, S.; Wiechert, W. and Nöh, K.
  Visualizing multi-omics data in metabolic networks with the software Omix: a case study.
  Biosystems 105, 2 (2011), pp. 154–161.

Talks

- Niedenführ, S.; Stute, B.; Weitzel, M.; Wiechert, W.; Nöh, K.
  Multiobjective experimental design for carbon labeling experiments: a case study.
  1st European Congress of Applied Biotechnology (Berlin, Germany, September 25, 2011).

- Niedenführ, S.; Meinert, S.; Schmitz, K.; Hönlinger, C.; Hardiman, T.; Kornfeld, G.; Mitterbauer, R.; Wiechert, W.; Noack, S.; Nöh, K.
  Industrial Production of Penicillin V with Penicillium chrysogenum: Towards efficiency Optimization by Systems Understanding.
  ECCE9/ECAB2 (Den Haag, Netherlands, April 22, 2013)

Posters

- Niedenführ, S.; Meinert, S.; Hardiman, T.; Kornfeld, G.; Wiechert, W.; Nöh, K.
  Fluxome analysis for Penicillium chrysogenum – lessons from an industrial collaboration project.
  30. DECHEMA-Jahrestagung der Biotechnologen und ProcessNet-Jahrestagung (Karlsruhe, Germany, September 10–13, 2012).

- Niedenführ, S.; Weitzel, M.; Wiechert, W.; Nöh, K.
  Multi-objective experimental design for carbon labeling experiments: a case study.
  8th ASIM Workshop 2011 (Garching, Germany, March 14, 2011).

- Niedenführ, S.; Schmitz, K.; Wiechert, W.; Noack, S.; Nöh, K.; Meinert, S.; Hönlinger, C.; Hardiman, T.; Mitterbauer, R.; Kornfeld, G.
  Industrial Fluxome Analysis for P. chrysogenum: Systems Understanding for Efficiency Optimization.
  Copenhagen Bioscience Conference, Cell Fatories and Biosustainability (Copenhagen, Denmark, May 5-8, 2013)

- Beste, D.J.V.; Nöh, K.; Niedenführ, S.; Mendum, T.; Hawkins, N.; Ward, J.L.; McFadden, J.
  ^{13}*C isotopologue profiling of intracellular Mycobacterium tuberculosis*
  Tuberculosis 2012 (Paris, France, September 11, 2013)

Developed Toolboxes

- Nonlinear Regression Toolbox (NoReTo)
- Global Sensitivity Analysis Toolbox (GloSA)

Further work was conducted:

- ^{13}C-MFA for *P. chrysogenum* BCB1 in cooperation with Katja Schmitz for bi-phasic growth in batch cultivations
- ^{13}C-MFA for *Mycobacterium tuberculosis* in cooperation with D. J. V. Beste and J. McFadden (University of Surrey)

This work was conducted at Forschungszentrum Jülich IBG-1:Biotechnologie in close collaboration with Sabine Meinert, Georg Kornfeld and Timo Hardiman at Sandoz in Kundl, where most wet-lab experiments were conducted. At the Kundl site penicillin V was discovered in 1952 [209].

Statement of Authorship

Herewith I certify that this thesis is my original work and I properly referenced all sources used. Except where a reference is made in the text, the document contains no material extracted as a whole or in part from a document previously published by me. This thesis was not previously submitted to another academic institution.

_____

Date, Place

Acknowledgments

Thanks to Katherina Nöh and Wolfgang Wiechert for the opportunity. My Mom and Dad for the support. Sabrina Witthoff, Stefan Helfrich for helping hands. Katja Schmitz, Sabine Meinert, Georg Kornfeld, Timo Hardiman, and Stephan Noack for being my co-workers in this great project. Samuel Leweke, Elisabeth Zelle and Timm Severin for thoroughly working together.

Contents

Nomenclature

Abbreviations

^{13}C-MFA	^{13}C metabolic flux analysis
23PG	Joined pool of 2- and 3-phosphoglycerate
3PSer	3-Phosphoserine
6APA	6-Amino penicillianic acid
6PG	6-Phosphogluconate
8HPA	8-Hydroxypenillic acid
Aa	Amino Acid
AAA	L-2-Aminoadipate
AcCoA	Acetyl-CoA
ACO	Aconitate
ACV	L-α-Aminoadipyl-L-cysteinyl-D-valine
ACVS	L-α-Aminoadipyl-L-cysteinyl-D-valine synthase
ADP	Adenosine diphosphate
ALA	Alanine
AMP	Adenosinemonophosphate
ANOVA	Analysis of Variance
Ant	Anthranilate
ARG	Arginine
ASN	Asparagine
ASP	Aspartate
AT	Acyltransferase
ATP	Adenosine triphosphate

BIOM	Biomass
BM	Biomass
C1	One carbon pool compunds
Chor	Chorismate
Cit	Citrate
CO2	Carbon dioxide
CoA	Coenzyme A
CYS	Cysteine
Cyst	Cystine
DHAP	Dihydroxyacetone phosphate
DHIV	Dihydroxyisovalerate
DNA	Deoxyribonucleic acid
ED	Experimental design
ES	Ergosterol
F6P	Fructose-6-phosphate
FBA	Flux balance analysis
FBP	Fructose-1,6-phosphate
FUM	Fumarate
FVA	Flux variability analysis
G6P	Glucose-6-Phosphate
GAP	Glyceraldehyde-3 phosphate
GC	Gas chromatography
GEM	Genome scale metabolic network reconstruction
Glc	Glucose
GLN	Glutamine
GLU	Glutamate
GLY	Glycine
GSH	Glutathione

GSSG	Glutathione disulfide
HIS	Histidine
HOPenV	*p*-Hydroxypenicillin V
HOPOA	*p*-Hydroxyphenoxyacetate
HSer	Homo-serine
ILE	Isoleucine
INST-^{13}C-MFA	Non-stationary ^{13}C metabolic flux analysis
IPNS	Isopenicillin N synthetase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIL	Keto-isoleucine
KIV	Keto-isovalerate
LC	Liquid chromatography
LEU	Leucine
LYS	Lysine
MAL	Malate
MET	Methionine
MIMO	Multi input multi output
MS	Mass spectrometry
NADH	Reduced form of nicotinamide adenine dinucleotide
NAD	Nicotinamide adenine dinucleotide
NADPH	Reduced form of nicotinamide adenine dinucleotide phosphate
NADP	Nicotinamide adenine dinucleotide phosphate
NMR	Nuclear magnetic resonance spectroscopy
OAA	Oxaloacetate
OPC	6-Oxo-piperidine-2-carboxylic acid
ORN	Ornithine
oxPPP	Oxidative pentose phosphate pathway
PAA	Phenylacetic acid

PA	Penicillin Amidase
PCL	Phenoxyacetic acid-CoA ligase
PeAl	Penilloaldehyde
PeA	Penaldic acid
PenG	Penicillin G
PeNH2	Penicillamine
PenV	Penicillin V
PEP	Phosphoenolpyruvate
PHE	Phenylalanine
PHP	3-Phospho-hydroxypyruvate
PIA	Penilloic acid
PIO	Penicilloic acid
Pi	Ortho-phosphate
PPP	Pentose phosphate pathway
PRO	Proline
PSer	3-Phosphoserine
PYR	Pyruvate
R5P	Ribose-5-phosphate
RNA	Ribonucleic acid
ROS	Reactive oxygen specy
Ru5P	Ribulose-5 phosphate
S-MFA	Stoichiometric metabolic flux analysis
S7P	Sedoheptulose-7 Phosphate
SER	Serine
SUC	Succinate
THR	Threonine
TRP	Tryptophane
TR	Thioredoxin-thioredoxin reductase system

TYR	Tyrosine
VAL	Valine
vvm	Volume gas per volume liquid and minute
X5P	Xylulose-5-phosphate

Mathematical Symbols

$\|\|\|_1$	L1-norm
$\|\|\|_{2,n}$	Scaled l^2-norm
$\|\|\|_\infty$	Infinity norm
aa'	Labeling state (m+0, m+1,...) in the daughter ion in all contained atoms
$aa'bb'$	Labeling state (m+0, m+1,...) in the mother ion in all contained atoms
ab	Labeling state (m+0, m+1,..) of skeletal atoms in mother ion
α	HDMR expansion coefficient for first order component function
α	Significance level
A	Number of skeletal atoms in detected fragment as vector
A'	Number of non skeleton atoms in detected fragment as vector
$\mathbf{B}$	Decision space
a	Number of labeled skeletal atoms in daughter ion
B	Number of skeleton atoms in non-detected fragment as vector
b	Vector reffering to fluxes over systems boundaries
β	HDMR expansion coefficient for secon order component function
B'	Number of non-skeleton atoms in non-detected fragment as vector
$\mathbf{C}$	Costs
c	Confidence interval of the standard deviation
C_I	Correction matrix for atom specy I
c_{init}	Natural labeling of measured ion
Cov	Covariance matrix

C_y	Length of vetor y_C
f	Number of occurences of the atom specy in the observed molecule
ϕ	Orthogonal polynomial
f_0	Zeroth order component function
f_i	First order component function
f_{ij}	Second order component function
f_{ijk}	Third order component function
Fish	Fisher Information matrix
f_{p_j}	Component function of arbitrary cooperative combination of inputs
$f(x)$	Function **or** HDMR model
G	Stoichiometric matrix including in- and effluxes
γ	HDMR expansion coefficient for third order component function
I	Information criterion
J	Jacobian
k	Maximum reaction rate
K_M	Michales constant
K^p	Unit hypercube in p dimensions
K_S	Affinity constant
λ	Eigenvalue
M	Molar mass
M	Measurement matrix for labeling measurements
y_M	Vector of all possible labelings in mother and daughter ion
m_s	Non-growth associated maintenance coefficient
μ	Specific growth rate
M_y	Length of vetor y_M
N	Number of a type $0, 1, 2, 3, \ldots$
N	Sample number
n_C	Number of carbon atoms in the measured ion

n_p	Number of component functions in HDMR model
n_v	Number of measured fluxes
n_y	Number of labeling measurements
p_{HDMR}	Number of parameters in HDMR model
φ	Feasible solution in decision space B
$p(I)$	Natural abundance of the isotope I
q	Reaction rate $[\mathrm{kg\,kg^{-1}\,h^{-1}}]$
r	Mass balance $[\mathrm{kg\,kg^{-1}\,h^{-1}}]$
RSS	Residual sum of squares
$\mathbf{S}$	Stoichiometric matrix
C_{corr}	Correction matrix for isotope correction of a MSMS measurement
S	Sensitivity index
s	Number of labeled skeletal atoms in mother ion
S^{cor}	Correlative global sensitivity index
s_{meas}	Measurement standard deviation
s'	Number of labeled skeletal atoms in mother ion
y_C	Labeling state of skeletal atoms in mother and daughter ion
S_T	Summed sensitivity indices of a HDMR model
S^{tot}	Total sensitivity index
S^{struc}	Structural sensitivity index
S_y	Selection matrix to choose measured entries from the possible of labeling states y_M
$\mathbf{t}$	Maximal possible labeling state in mother ion
t	time
θ	Model parameters for extrapolation
θ	Parameter
$\mathbf{t}'$	Maximal labeling state in daughter ion
V	Variance
v	Metabolic reaction rate also called flux

v_{free}	Vector of free fluxes
w	Propability density function w
WRSS	Weighted residual sum of squares
$\mathbf{X}$	Concentration
$\mathbf{x}$	Model input as a set of parameter values
x	Vector of simulated labeling pattern
x_{inp}	Vector of input substrate labeling pattern
x_{p_j}	Subset of inputs for a component functions
y_M	Labeling measurement
Y_{PS}	Product yield coefficient
Y_{XS}	Biomass yield coefficient
$\mathbf{z}$	Single objective values for multi-objective optimization
Z	Objective values for multi-objective optimization
z	Vector of experimentally determined measurements for a combination of mother and daughter ion

Subscripts

$_{A'}$	Detected fragment, without skeletal atoms
$_{meas}$	Measured value
$_{min}$	Minimum
$_{B'}$	Not detected part of the mother ion, without skeletal atoms
$_{sim}$	Simulated value

Superscripts

$^{(s)}$	Refering to the sample with index s

Part I

Introduction

Chapter 1

Penicillin

To date, more than 80 years after its discovery, penicillin is still one of the predominantly used antibacterial agents. Penicillin is produced in biotechnological processes by use of the fungus *Penicillium chrysogenum*. In the 1940ies penicillin was the first efficient pharmaceutical compound used for infection control and health care. Moreover, the first defined and successful chemotherapeutic product was isolated from a microbial source, founding the era of therapeutic biotechnology [118].

1.1 History of Penicillin Production

In 1929, Alexander Fleming detected the antibacterial effects of *Penicillium* towards the Gram-positive pathogenic bacterium *Staphylococcus aureus* [81]. This newly found mold was later identified as *Penicillium notatum* [118]. The agent obtained via filtration of the lysed mold was non-toxic for animals and was named penicillin. However, penicillin is in-stable and in 1932, Fleming stopped work on it, because its purification and stabilization was difficult [118]. In the 1930ies sulphonamides, another class of chemically synthesized antibacterial agents, possessed a small spectrum of activity, had severe side effects and bacterial resistances developed rapidly [118, 57].

Howard Florey, Ernst Chain, Norman Heatly, and Edward Abraham were the first recognizing Flemings work. They continued purifying penicillin using solvent/water mixtures, stabilized the extract, and used it as pharmaceutical for the first clinical trials [118, 83]. Later, in 1945, Alexander Fleming, Ernst Boris Chain, and Howard Walter Florey received the Noble price "for the discovery of penicillin and its curative effect in various infectious diseases" [274]. In 1940-1941 the enormous potential of penicillin as antibacterial agent became clear as it was tested on humans for the first time [83]. In the subsequent years, the Northern Regional Research Laboratory (NRRL), the US Department of Agriculture (USDA), and several companies (including Pfizer, Merck, and Squipp) were involved in the development of a production process for penicillin [118, 34, 83, 12].
Starting in 1943, pilot plants for penicillin production were developed by Pfizer, Merck, Squipp, and the Commercial Solvents Corporation and penicillin was used for the first time for those wounded in World War II [118, 31]. Later, penicillin helped curing gonorrhea and syphilis which was widespread at the end of World War II and lead to a fast post-war recovery [121, 118]. The US government subsidized entry into the penicillin production at that time,

broadening the spectrum of manufacturers. Hand in hand with the steep production increase by strain and process optimization, prices of penicillin decreased (see figure 1.2) [12, 196]. Between 1943 and 1945 the US production increased from about 0.64 kg to 1,180 kg penicillin per month.

Simultaneous to the development of production processes, the first "strain development" in history took place (cf. figure 1.1). The first isolate *P. chrysogenum* NRRL 1951 was taken from a moldy cantaloupe. The strain was capable of producing 60-150 $mg\,L^{-1}$ penicillin (see figure 1.2) which was 11-15 fold more than Flemings *Penicillium notatum* strain (production titer: 1 $mg\,L^{-1}$) [118, 110]. X-ray treatment of *P. chrysogenum* resulted in a mutant strain X-1612 which was capable of producing 161-300 $mg\,L^{-1}$ [118, 54, 196]. Finally, the strain Q-176 was obtained by ultraviolet induced mutation at the University of Wisconsin. It was capable of producing up to 550 $mg\,L^{-1}$ [54, 196, 66]. Based on this mutant, the WISCONSIN family of strains was generated. Titers up to 1.8 $g\,L^{-1}$ for WISCONSIN 54-1225 are reported [25, 66, 110]. In 1973, 2.8-7.3 $g\,L^{-1}$ penicillin were produced by the P2 strain emerged from WISCONSIN 54-1225 (cf. figure 1.1) [110, 174, 18].

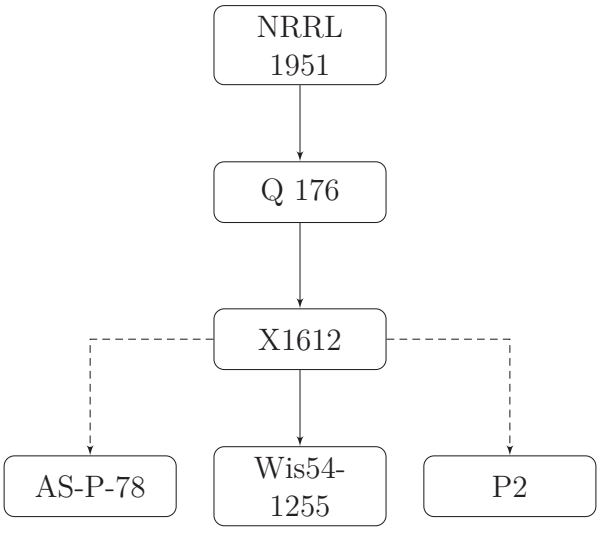


figure 1.1: Strain development for *P. chrysogenum*. Solid lines indicate one type of treatment (X-ray, UV, ...) used for mutagenesis, dashed lines indicate multiple treatments [18].

Penicillin is produced by *P. chrysogenum* after addition of a side-chain precursor. It was in the 1950s when it became evident that *P. chrysogenum* could use other acyl-group precursors than phenylacetic acid (PAA) unlocking production of further penicillins [219, 118]. One of them was penicillin V (phenoxymethylpenicillin, V refers to the German word *"vertraulich"* which translates to confidential) [118, 31].

In parallel to the strain optimization programs, the production process for penicillin was continuously developed. Early on, the production of penicillin was shown to be catabolite repressed by glucose [200]. Thus, the first media was optimized for growth of *P. chrysogenum* and penicillin production. It was found that continuous addition of glucose results in high penicillin production [223].

The first established processes for penicillin production used lactose as carbon source and were performed in uncontrolled batch mode. Today, glucose/sucrose or other crude sugars are used in a fed-batch process with additional feeds for precursor supply and pH control. Cultivations are performed in 100-400 m^3 bioreactors and process duration is 120 to 200 h, the growth form is pelleted [66]. A production of at least 66 $g\,L^{-1}$ penicillin V is needed for a modern fed-batch process to be amortized. The final biomass concentration is around 45 $g\,L^{-1}$ (data estimated by Monte-Carlo simulations by Biwer *et al.*[26]). The strong increase in production of penicillin since 1929, resulting in titers 10,000-fold higher than the ones produced by the initial isolate of Alexander Fleming (see figure 1.2), is a result of a simultaneous strain and process improvement. Current data from industrial processes were not available, estimates for them are present in literature and

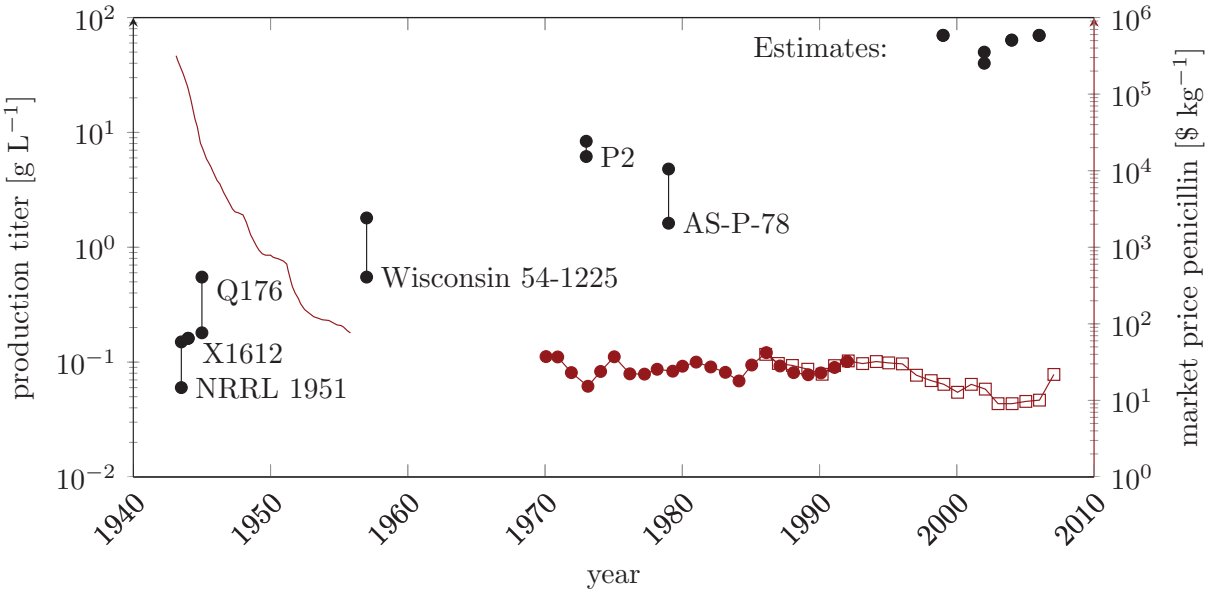


figure 1.2: Product titer and market prices for penicillin G. Measured and estimated product titers (—●—) and were taken from: [120, 144, 25, 66, 118, 142, 174, 54, 26, 11]. Market price for penicillin G taken from: Nielsen [174] (—●—), King [120] (——) and data taken from Scott and Oldenhof [215] (—□—).

shown in figure 1.2.

To date, penicillins and cephalosporins are the most widely used antibiotics in therapeutic medicine. In 2009, the market for antibiotics possessed a volume of US$ 42 billion which makes up 5 % of the world wide pharmaceutical market. Penicillins and cephalosporins possess similar sales data and penicillin alone accounted for 16 % of the sales in 2009 for antibiotics [92]. The annual production of penicillin G and penicillin V in 2006 was 16,900 and 7,500 tons, respectively.

In the last 30 years a high amount of pharmaceutical ingredients were produced in India and China, which led to a strong decrease of the price for penicillin (figure 1.2). The price of penicillin G crushed to its minimum of 9.28 US$ from 16.8 US$ in April 2003 [24]. In the following years, the price rose again until a strong increase in 2007 was observed, due to a high market share of China mainly influencing prices. China was in the last years the major producer of penicillin G accounting for up to 75 % of the world production [24, 215].

1.2 Growth Behavior

P. chrysogenum is a filamentous growing fungus (cf. figure 1.3). Growth occurs at the tip of the hypha in the apical part. Below, the sub-apical and hyphal part of the cell are located. The growth is, thus, polarized and strong activity is found only in specialized parts of the cells. In contrast to the apical part, the sub-apical region contains septa that are not yet fully developed. It is only the hyphal part of the cell where the septa lead to a full compartmentalization of nutrients. Concomitantly with further differentiation, highly vacuolized cell parts are formed in the hyphal region [121, 249, 281, 184, 89].

Due to these aging effects, penicillin producing cultures of *P. chrysogenum* undergo active

changes during cultivations and the cellular states are diverse [249]. Roughly 5 % of the overall cells are active in their apical parts, 45 % consist of sub-apical parts and 20 % are in an hyphal differentiation state and up to 40 % of the cells can be lysed [249]. The sub-apical parts of the cell and to some extent also parts of the hyphal segments are assumed to produce most of the penicillin [281, 249].

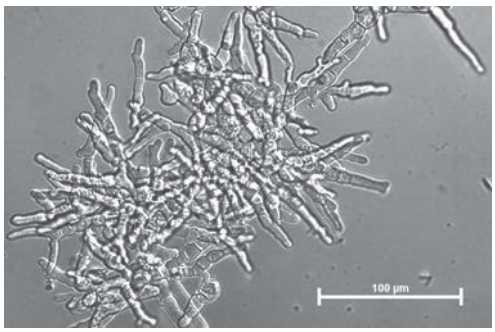


figure 1.3: Microscopic picture taken from *P. chrysogenum* strain BCB1 used in this thesis. Kindly provided by Alexander Grünberger (IBG-1, Forschungszentrum Jülich) .

Usually three forms of growth are distinguished: (I) dispersed growth, with the formation of branched and unbranched hyphae (free dispersed), (II) dispersed growth with the formation of clumps (aggregates), and (III) pellet growth. Due to the polarized growth of *P. chrysogenum*, the separate regions of the cell are in different aging and physiological states, leading to continuous changes in growth. One influencing factor on cell morphology and also on production are mechanical shear forces during bioreactor cultivations which cause breakage or aggregation of cells [114, 185]. Furthermore, the pH, growth rate, inoculation/germination, and nutrition may cause differences in the morphology of *P. chrysogenum* [174, 159, 114, 176, 177].

1.3 Penicillin Biosynthesis in *P. chrysogenum*

Penicillins are composed of a nucleus, formed from amino acids, and a side-chain precursor, which is specific for the type of penicillin. The side-chain determines its activity spectrum and is added to the cultivation medium. For penicillin G and penicillin V the side-chain precursors are phenylacetate and phenoxyacetate, respectively. Penicillin V outperforms penicillin G in terms of stability towards acids, making it more effective for the oral administration [103]. In figure 1.4 the chemical structure of penicillin G and V (PenG and PenV) are shown.

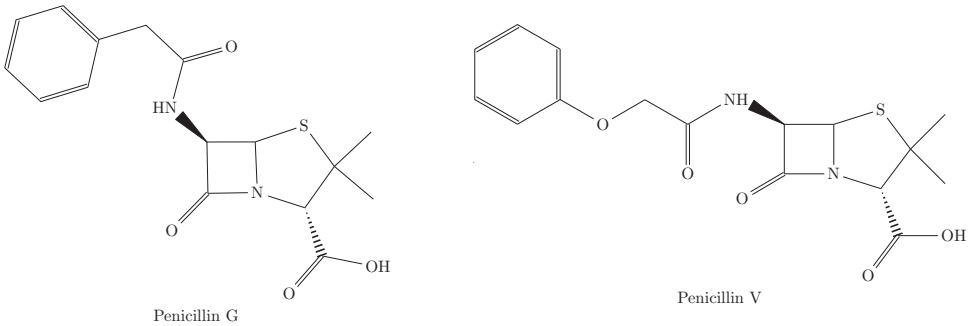


figure 1.4: Penicillin G and V chemical structures.

The main building blocks of penicillin were elucidated in the 1950s by inhibition of biosynthesis using structural analogs and tracer studies, it was deduced that L-cysteine (CYS) and

L-valine (VAL) form the nucleus of penicillin [53, 11, 227]. Penicillin V is formed in three subsequent steps. Figure 1.5 provides an overview of the penicillin pathway starting from the main precursors. Initially, the amino acids L-valine, L-cysteine and the diamino acid L-2-aminoadipate (AAA) form the tripeptide L-α-aminoadipyl-L-cysteinyl-D-valine (ACV) by a non-ribosomal peptide ACV synthase (ACVS). Following, the isopenicillin N synthetase (IPNS) catalyzes the β-lactam ring formation by closing the β-lactam and thiazolidin ring. This reaction requires an equi-molar amount of oxygen [14, 164]. Finally, L-2-aminoadipate is substituted with the acyl-side-chain precursor added to the medium by the acyltransferase (AT). Previously, the acyl-chain precursor is activated by phenoxyacetic acid-CoA ligase (PCL). The side-product L-2-aminoadipate is recycled or used for lysine synthesis.

The genes encoding for the enzymes ACVS, IPNS and AT are arranged in a gene cluster consisting of *pcAB*, *pcbC* and *penD*. The clustering of genes is a feature often found in secondary metabolite pathways, because they are subjected to a tight control. Several other open reading frames were detected in this region, but had no significant effect on the penicillin production [240]. The gene encoding for phenoxyacetic acid-CoA ligase (PCL) is not located in this cluster [179].

In *P. chrysogenum*, the formation of penicillin is taking place within two compartments. The reactions catalyzed by the ACVS and IPNS are assumed to occur in the cytoplasm [164, 239]. The final steps of penicillin synthesis, the activation of the side-chain precursor and side-chain substitution catalyzed by the acyltransferase, take place in the peroxisomes [165, 164, 241]. The peroxisomes are thus crucial for penicillin production in *P. chrysogenum* [154, 119].

Energy Consumption

The energy demand for penicillin production is high, for CYS biosynthesis 5-8 $\mathrm{mol\,mol}_{\mathrm{CYS}}^{-1}$ and for VAL $1\,\mathrm{mol\,mol}_{\mathrm{VAL}}^{-1}$ cytosolic NADPH is used [282]. The energy demand of penicillin synthesis will be later thoroughly discussed.

Side Products

In penicillin biosynthesis several side-products have been observed emerging from the precursors and penicillin itself (cf. figure 1.5). One of the penicillin precursors, L-2-aminoadipic acid, is found in significant amounts in the medium supernatant. By cyclization of L-2-aminoadipate the lactam 6-oxo-piperidine-2-carboxylic acid (OPC) is formed as a side product [188, 55]. ACV is formed from the three precursor metabolites L-valine, L-cysteine and L-2-aminoadipate. From ACV the formation of bis-ACV is favorable, it is formed by utilization of oxygen. The reverse reaction is catalyzed by the thioredoxin-thioredoxin reductase system (TR). Besides this, glutaredoxin system might also be involved in the reverse reaction [188, 55, 121]. Next, acyltransferase catalyzes the formation of 6-aminopenicillanic acid (6APA) from isopenicillin N, by release of L-2-aminoadipic acid. 6APA can re-enter the penicillin formation pathway by conversion to penicillin V by action of acyltransferase [188, 55]. Besides this, a side-product 8-hydroxy-penillic acid (8HPA) is formed by decarboxylation of 6APA [188, 55].

The final product penicillin V is subject to degradation by penicillin amidase (PA) to 6APA or it can be converted into several side-products which cannot re-enter the penicillin V pool. Penicillin V can either be converted to penicilloic acid (PIO), which can be further

decarboxylated to penilloic acid (PIA) or POA and penicillin V can be converted to p-hydroxypenicillin V (HOPenV) and p-hydroxyphenoxyacetate (HOPOA) [188, 55].

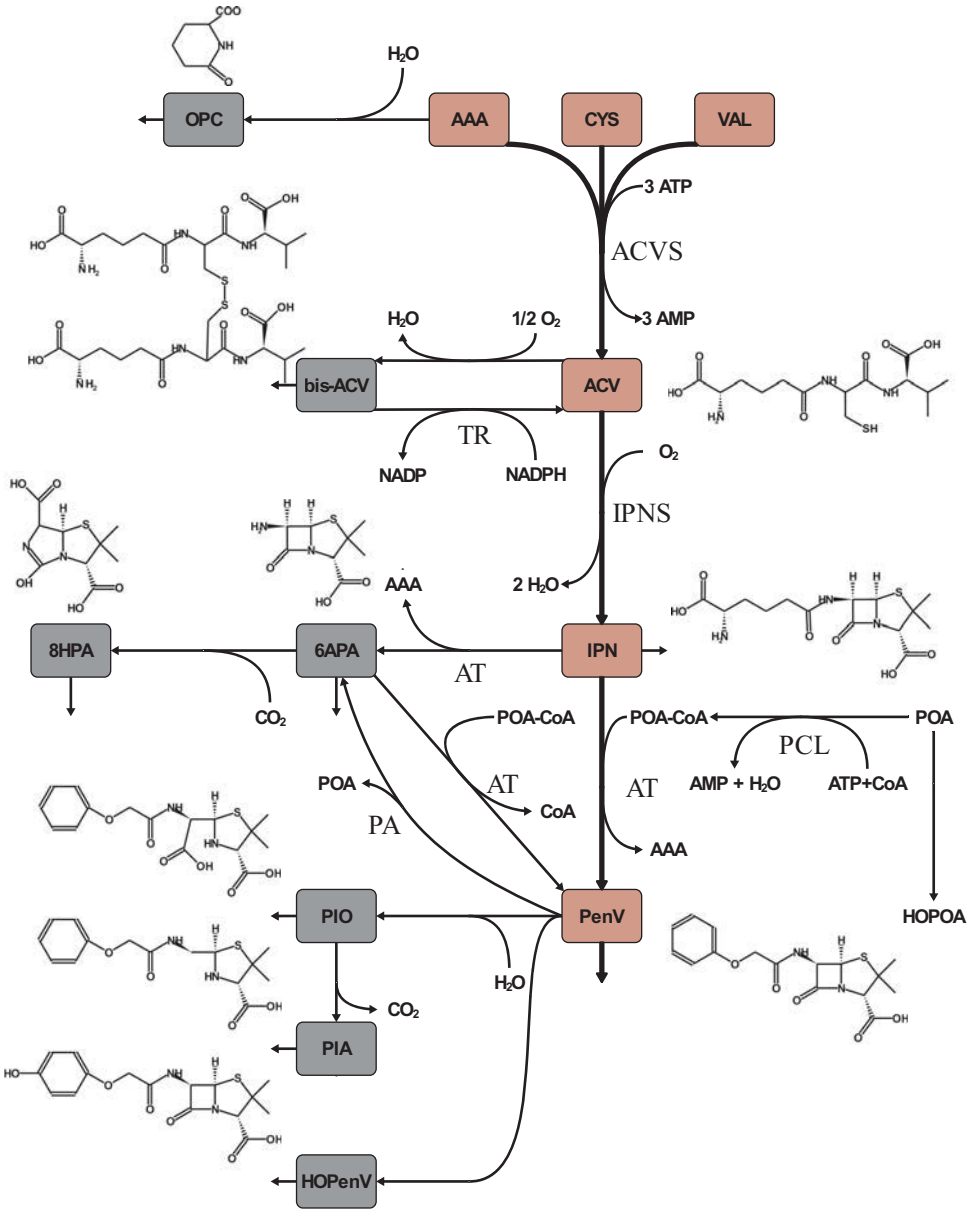


figure 1.5: Penicillin pathway and side-product formation in *P. chrysogenum.* Main penicillin pathway is colored in red.

1.4 Penicillin - A Global View

In summary, the production processes for penicillin V in industry have to face low market prices and, still, need to be profitable. This demands for economic production processes that are competitive with the global market. One of the central aims for process improvement is an increase in the production capabilities of the penicillin V producing organism *P. chrysogenum.*

To enhance the production of the organism, the complex formation of penicillin V has to be elucidated.
In the next chapter, tools will be introduced that elucidate the complex interaction of penicillin formation and reaction rates in the metabolism.

Chapter 2

Fluxomics Tools

Modern industrial biotechnology aims at producing valuable compounds by microorganisms. The ultimate goal is to develop economic processes by increasing production rate and yield of the desired product. To achieve this, targeted improvements of the strains are needed via means of metabolic engineering. Several (omics) disciplines are applied to identify bottlenecks in metabolism to increase production [113]. Besides, most omics techniques like genomics, transcriptomics, and proteomics provide only indirect and limited information about the intracellular reaction rates [104, 84, 106]. However, providing information about metabolic reaction rates of a cell, or synonymously the fluxome, is crucial to understand the function of the cell‘s metabolism and, ultimately, to provide hints for strain improvement. To quantify and describe the fluxome or parts of it, several fluxomics tools were developed for more than a decade and were established for prokaryotic and eukaryotic model organisms [268, 261, 183, 74, 75, 140, 277]. Fluxes describe the material transport through metabolic pathways. They are, usually, normed to cell dry weight (CDW) as reference system and are expressed in $[\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}]$. Important Fluxomics tools used in this work will be introduced in this chapter.

Stoichiometric Reactions

One of the prerequisites for applying fluxomics tools is the knowledge about the set of reactions in a cell, which is also called the metabolic reaction network. Nowadays, the complete sequence of an organism‘s genome can be generated in a high throughput manner [218]. This enables the reconstruction of metabolic reaction networks from known enzymatic functions by sequence analysis. However, these “genome scale metabolic network reconstructions” approaches can provide a metabolic blueprint, the network topology, rather than quantitative information [234]. Several fluxomics tools were established to quantify *in vivo* fluxes for a given metabolic network (flux balance analysis, ^{13}C-MFA, kinetic flux profiling,...) [74, 210, 269, 277]. Metabolic flux analysis is the core technique to understand metabolism and to implement metabolic engineering for targeted strain improvement for production of such valuable compounds as penicillin V.

In all of these techniques, the metabolic network is described by a set of mass balance equations [251]. In a reaction sequence of $\mathrm{A} \xrightarrow[v_1]{} \mathrm{B} \xrightarrow[v_2]{}$ with fluxes v_1 and v_2, the mass balance equation for metabolite B can be written for a time point t, with C_X denoting the concentration of a metabolite X:

$$\frac{d(C_B)}{dt} = v_1 - v_2 \tag{2.1}$$

Assembling mass balance equations for all intracellular metabolites can be re-formulated as an equation system with $\mathbf{X}$ being a vector of metabolite concentrations in the cell, $\mathbf{G}$ is a stoichiometric matrix (including in- and effluxes) that comprises the stoichiometric coefficients of the reactions, v is the flux vector, and b is a vector referring to fluxes over the system's boundaries:

$$\frac{d\mathbf{X}}{dt} = \mathbf{G} \cdot v + b \tag{2.2}$$

By suitable process realization a dynamical equilibrium, called steady-state, is accomplished for the bio-process under investigation. This is done by maintaining constant state variables in the process. If this macroscopic steady-state is accomplished, it is assumed that also a microscopic steady-state is found within the cell. Thus, the metabolite concentrations $\mathbf{X}$ and reaction rates v are constant. This state is also called *metabolic steady-state* of the system. At metabolic (pseudo-)steady-state no changes in the concentrations of metabolites are found ($\frac{dC_X}{dt} = 0$). The underlying mathematical system is simplified as it holds that $\mathbf{G} \cdot v + b = 0$. Thus, the influx into a pool equals its effluxes (for the example above: $v_1 = v_2$). Importantly, for real metabolic networks, the system of mass balance equations is under-determined. One possibility solve the system is flux balance analysis.

2.1 Flux Balance Analysis

Flux balance analysis (FBA) is commonly applied in systems biotechnology to understand metabolism and to obtain hints for strain optimization. The technique is appealing as it can provide fast results and predictive information about potential states of the metabolism. The metabolic network model is typically under-determined, thus, the solution is not unique. To circumvent this a so called objective function (usually a reaction rate) is minimized or maximized to obtain a single solution. Commonly, maximizing product, biomass yield, or energy generation are used as objectives [183]. Thus, it is assumed that the biological system under investigation is behaving goal-oriented towards highest evolutionary pressure. Similar techniques can also be applied to estimate lower and upper boundaries for fluxes in a reaction network (called flux variability analysis, short: FVA) [90].
Metabolic networks used for FBA rely heavily on cofactor balancing. Thus, additional assumptions for energy consumptions are introduced into the network, although it is known that complete cofactor balancing is difficult to achieve [245].
In order to perform flux balance analysis, a stoichiometric reaction network is built, describing the metabolism of the organism under investigation. By the underlying steady-state assumption, mass balance equations can be formulated for the stoichiometric reaction network. It is expressed in the form of the stoichiometric matrix $\mathbf{S}$ with size $m \times n_S$ (m metabolites and n_S reactions). This matrix determines the stoichiometric relation between substrates and products of a reaction. It does not contain in- and effluxes of the system.

$$\mathbf{S} \cdot v = 0 \quad \text{with} \quad v_{i,lower} \leq v_i \leq v_{i,upper} \quad (i = 1, 2, 3, \ldots n_s) \tag{2.3}$$

Here, lower $v_{i,lower}$ and upper $v_{i,upper}$ bounds are defined for each flux in vector v based on thermodynamical, process, or physiological data (e.g. capacity of enzymatic activity). Thereby, the flux values are restricted to biologically feasible solutions. Importantly, measurements with standard deviations are not used, but lower and upper bounds for fluxes [183].
Finally, if degrees of freedom are left within the given system ($n_S > m$), the system is underdetermined. Usually a unique solution cannot be obtained from equation 2.3 by measuring extracellular rates. Thus, FBA searches either the minimum or maximum of the user-defined objective function consisting of a linear combination of fluxes weighted by user chosen factors in vector $\mathbf{c}$ (with length n_S). The objective value is $\mathbf{Z} = \mathbf{c}^\intercal v$.
From the infinite number of solutions in such a network only one is obtained in the end. However, the chosen flux distribution is not representing the actual state of the cellular metabolism but only a possible state.

To estimate *in vivo* fluxes (intracellular fluxes), further information is needed. Therefore, adding measurements containing information about intracellular fluxes is necessary. For this, labeled material is introduced into the cells by the fed substrate. How this can increase the information content of fluxes, will be described in the next section.

2.2 ^{13}C Metabolic Flux Analysis

In contrast to FBA, ^{13}C-metabolic flux analysis (^{13}C-MFA) explores the fluxome by relying on extracellular fluxes and intracellular labeling measurements. In this state-of-the-art technology metabolic fluxes are inferred from isotope labeling experiments using ^{13}C labeled substrates. The resulting labeling patterns in metabolites are quantified. Finally, fluxes are inferred based on a model describing the distribution of (^{13}C and ^{12}C) carbon atoms in the metabolism.
Next, the practices to obtain experimental data for ^{13}C-MFA are described.

2.2.1 Carbon Labeling Experiment

To conduct a ^{13}C-MFA, a carbon labeling experiment is performed (see figure 2.2) [280]. Importantly, this work deals with stationary ^{13}C-MFA and, thus, also the isotopic labeling in the metabolites needs to be stationary (isotopic steady-state).
In a metabolic steady-state all extracellular and, thus, intracellular reaction rates are constant. To establish an isotopic steady-state needed for ^{13}C-MFA, first a metabolic steady-state needs to be established. This is usually accomplished in a process called chemostat, it aims at preserving the steady-state conditions by a constant feed and withdrawal of bioreactor content throughout the cultivation. The organism under investigation is, thus, cultivated under controlled conditions to reach the metabolic steady-state. At this stage, natural labeled substrates are exchanged by isotopically labeled ones. In turn, driven by the metabolic activity of the cells, characteristic labeling patterns emerge in the metabolite pools; first in the vicinity of the carbon source, later in pools further downstream. This can take up to hours or even days in eukaryotic organisms [96, 284]. The speed of labeling distribution depends on the pools size (intracellular concentration) and the flux through the pool under observation, it is referred to as *turnover* (see figure 2.1). Metabolites in glycolysis possess small pool sizes and high fluxes (high turnover) and show fast equilibration of labeling. In

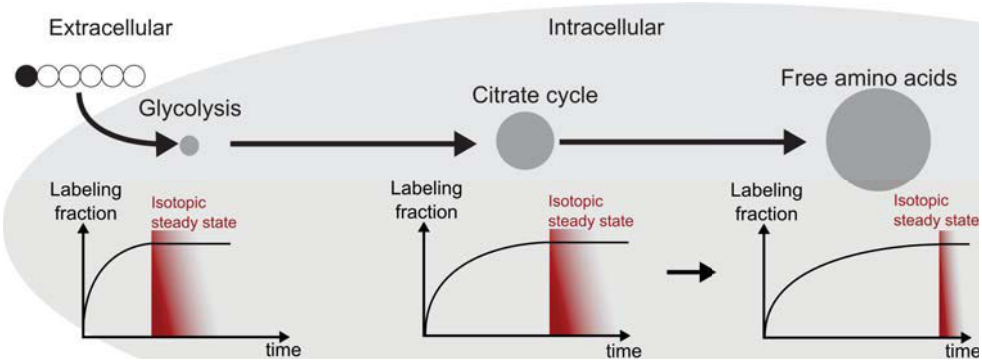


figure 2.1: Turnover of metabolite pools and the time resolved spread of labeling in metabolism.

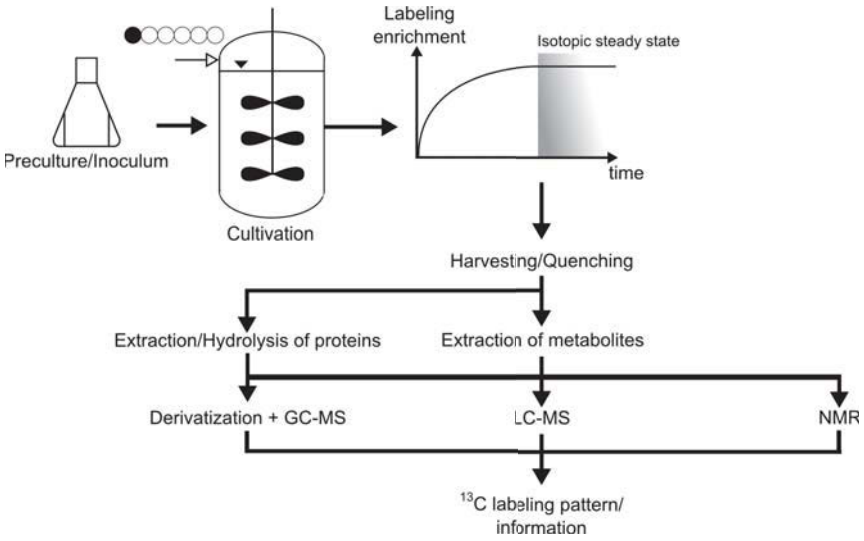


figure 2.2: Experimental work-flow for ^{13}C-MFA

this case, the so called isotopic steady-state is reached fast. Pools of amino acids are usually larger with low fluxes and exhibit, thus, slow labeling enrichment (low turnover) and reach the isotopic steady-state later in an experiment [96, 284].

In stationary ^{13}C-MFA, a sample is taken at isotopic steady-state. Cells are quenched to stop metabolism and preserve metabolites containing labeling information, the metabolites are extracted by specialized protocols [236]. Their respective labeling pattern is measured by state-of-the-art techniques, e.g. nuclear magnetic resonance spectroscopy (NMR) or mass spectrometry (MS). MS is one of the techniques which is frequently applied for ^{13}C-MFA. By MS, the measured ions can be separated by their masses, emerging from the labeling state in the metabolite [271]. Finally, the measured labeling pattern can be used for flux estimation in ^{13}C-MFA.

Summarizing for stationary ^{13}C-MFA, a metabolic steady-state has to be maintained for a sufficient time to also obtain isotopic stationary conditions (isotopic steady-state). Thus, prerequisites for the application of (stationary) fluxomics tools are optimally met in chemostat cultivations.

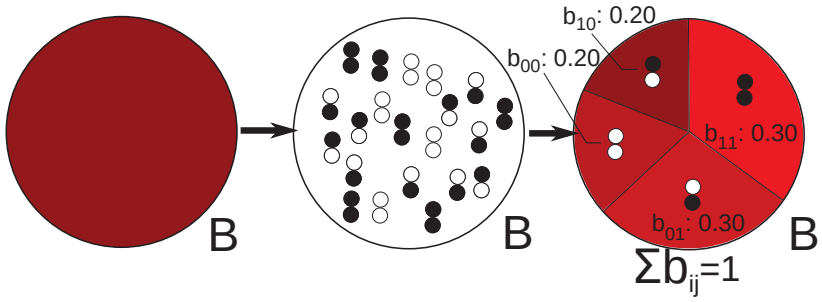


figure 2.3: Metabolite B contains two carbon atoms. The metabolite pool of B consists of a mixture of four isotopomers (B#00, B#10, B#01, B#11), the fraction of the separate isotopomers can be measured and sum up to one.

2.2.2 Modeling of Labeling Experiment

In this section the ingredients for modeling of carbon labeling experiments are presented:

- the labeling patterns of metabolites and substrates
- simulation of labeling patterns for given fluxes
- modeling of labeling measurements

The main steps of the ^{13}C-MFA work-flow are visited by usage of a toy example. At the end of the chapter a summary is given and the state-of-the-art work-flow is recapitulated.

Isotopomers

The stable carbon isotope ^{13}C is present in about 1.07 % of natural carbon. This labeling is referred to as natural labeling [22]. Labeled substances possesses ^{13}C atoms enriched at specified carbon atom positions. Labeling patterns are often described in binary notation: ^{12}C is denoted by a "0" and ^{13}C is denoted by a "1".
For a metabolite with n_C carbon atoms 2^{n_C} labeling patterns are possible. Metabolite B contains two carbon atoms and thus four different isotopomers are possible B#00, B#10, B#01, B#11 (cf. figure 2.3).
If isomer forms of a molecule share the same isotopic composition, but differ in their positions, they are called isotopomers. By measurements, the fractions of the isotopomers for a metabolite can be elucidated (see figure 2.3). The fractions of all isotopomers of a compound sum up to one. Thus, labeling patterns for ^{13}C metabolic flux analysis are using relative amounts of isotopomers.

Substrates

Labeled substrates are used in an carbon labeling experiment. Frequently used substrate for carbon labeling experiment are:

- glucose labeled at the first position, called 1-^{13}C glucose or Glc#100000,
- uniformly labeled glucose, called U-^{13}C glucose or Glc#111111,
- naturally labeled glucose Glc#000000.

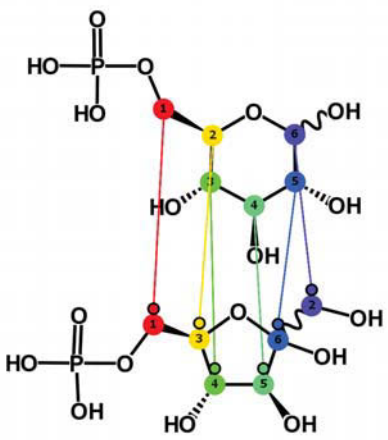


figure 2.4: Atom transition for glucose-6-phosphate isomerase. Numbering of carbon atoms by INCHI™ strings to identify uniquely the carbon atoms, see Mu *et al.* for details [163]. Figure generated using OMIX [60].

Further, more or less costly, substrates and labeling patterns are commercially available [1]. The labeled substrates are not only differing in their prices, but also in purity. Usually 99% atom purity is used. It refers to the purity of ^{13}C at the enriched positions in the substrates. In the experiments a mixture of them is used, called input substrate mixture. The labeling state of an input substrate is denoted x_{inp}.

Isotopomer Balancing

To balance isotopomers in a metabolic network, the mass balance equations for a metabolite pool can be extended. For the reaction sequence A $\underset{v_1}{\longrightarrow}$ B $\underset{v_2}{\longrightarrow}$, it is assumed that each metabolite contains two carbon atoms. In equation 2.1 isotopomer pools can be integrated by resolving the metabolite pool concentration into separate isotopomer pools for which the steady-state assumption can be applied as well.
The fate of the carbon atoms need to be specified for each reaction containing carbon atoms and are called atom transitions, an example can be seen in figure 2.4. For the reaction of glucose-6-phosphate isomerase the fate of carbon atoms in glucose-6-phosphate is assigned to fructose-6-phosphate.
Similar to metabolite pool balances, isotopomer pools are balanced, whereas the isotopomer fraction of metabolite A and B are denoted by a and b and their labeling pattern is given in subscripts:

$$\begin{aligned}
\frac{d(C_B \mathrm{b}_{00})}{dt} &= v_1 \mathrm{a}_{00} - v_2 \mathrm{b}_{00} \\
\frac{d(C_B \mathrm{b}_{10})}{dt} &= v_1 \mathrm{a}_{10} - v_2 \mathrm{b}_{10} \\
\frac{d(C_B \mathrm{b}_{01})}{dt} &= v_1 \mathrm{a}_{01} - v_2 \mathrm{b}_{01} \\
\frac{d(C_B \mathrm{b}_{11})}{dt} &= v_1 \mathrm{a}_{11} - v_2 \mathrm{b}_{11}
\end{aligned}$$

The input substrate, here A, is labeled. The labeling is distributed by the reactions (here v_1 and v_2) through the metabolic network, resulting in labeling patterns in the emerging metabolites (here: B).
A general formulation of the above described reactions can be given in compact notation. The function f describes the labeling change over time and depends on the fluxes v, the input substrate labeling state x_{inp} and the labeling state of the system x [181]:

$$\mathrm{diag}(\mathbf{X}) \cdot \frac{dx}{dt} = f(v, x_{inp}, x) \tag{2.4}$$

At metabolic and isotopic steady-state, the labeling fractions x are constant and the left hand side of equation 2.4 vanishes:

$$\mathrm{diag}(\mathbf{X}) \cdot \frac{dx}{dt} = 0 \tag{2.5}$$

In contrast to purely stoichiometry-based analysis like FBA, in modeling carbon labeling experiments the exchange of labeled material in reversible reaction steps has to be accounted for.

Flux Coordinate System

To model the labeling patterns in the metabolites, transport of labeled material needs to be fully described by isotopomer balance equations. This is important for reactions close to thermodynamic equilibrium. They are called reversible reactions [262]. In these reactions, transport of material occurs in both directions denoted by two arrows in the reaction equation. For example:

$$q : \mathrm{B} \rightleftharpoons \mathrm{E}$$

In reaction q, a fraction of the labeled material is transported in the direction of reaction equation (left to right) $\mathrm{B} \underset{v_{q,forward}}{\rightharpoonup} \mathrm{E}$ by a so called forward flux and another fraction is reacting in the opposite direction $\mathrm{B} \underset{v_{q,backward}}{\leftharpoondown} \mathrm{E}$, called backward flux. Both fluxes possess by definition positive values and are acting at a time, resulting in equilibration of the pools B and E. Forward and backward fluxes are hard to interpret in the context of a biological network, thus they are converted to net and exchange fluxes.

$$v_{net} = v_{forward} - v_{backward} \tag{2.6}$$

$$v_{xch} = \min(v_{forward}, v_{backward}) \tag{2.7}$$

Having defined the metabolic network stoichiometry, the rank of the stoichiometric matrix S determines the number p of free fluxes v_{free} [263]. They can be converted to all net and exchange fluxes v_{net} and v_{xch} by using the kernel matrix $\mathcal{N}$ and a transformation Φ^{-1} [263]:

$$v^{\rightleftharpoons} = \Phi^{-1} \begin{pmatrix} v_{net} \\ v_{xch} \end{pmatrix} = \Phi^{-1}(\mathcal{N} \cdot v_{free}) \tag{2.8}$$

Setting up all equations as described above, results in a large system of differential algebraic equations (cf. equation 2.4). By specifying input substrate and a feasible set of free fluxes, the system can be solved; the labeling state of the system x can now be calculated [267]:

$$f(x_{inp}, v^{\rightleftharpoons}, x) = 0 \tag{2.9}$$

In general, nonlinear terms arise in reactions using multiple labeled substrates. For small scale systems, isotopomer balance equations can be efficiently applied to describe the labeling state. For larger systems, the number of isotopomers increases significantly, because metabolites with up to 11 (or even more) carbon atoms are contained. In a genome scale network metabolites with an even larger carbon skeleton may exist. Such large system can no longer be handled efficiently by simulating all 2^{nC} isotopomers. Thus, mathematical transformations like EMUs, cumomers and topological network reduction by forward/backward tracing were developed to make these systems computationally feasible [259]. A multitude of techniques have been published so far, nevertheless they are not subject of discussion in this work [226, 225, 7, 267]. Throughout this work 13CFLUX2 is used, which uses both cumomer or EMU transformation. For illustrative purposes only isotopomers are discussed.

Measurements

Finally, from the labeling state x the labeling measurements need to be calculated. The labeling state can be converted to a simulated labeling measurement y_M by the help of a measurement matrix M [266].

$$y_M = M \cdot x(v_{free}, x_{inp}, ...) \tag{2.10}$$

Likewise a measurement matrix can be built for the simulated flux measurements y_v. From the given $n_y + n_v$ labeling and flux measurements, p free fluxes are estimated.

Assumptions on ^{13}C-MFA

Summarizing, the assumptions on which stationary ^{13}C metabolic flux analysis and (in parts) flux balance analysis rely on are discussed (taken and adapted from Wiechert [261], Wiechert and de Graaf [263]):

1. The biological system can be represented by a finite set of homogeneously distributed pools. Compartmentalization may lead to separate pools for a metabolite per compartment. Besides, it is assumed that the cultivated organism‘s population is homogeneous.

2. The observed system must be kept in a well-defined physiological stationary state. Thus, state variables like temperature, pressure, concentrations in liquid and gas phase need to be kept constant. The metabolic and isotopic steady-state has to be reached.

3. The set of relevant reactions and atom transitions are known and incorporated in the network model.

4. No isotopic mass effects can be observed in the system under study.

Several counter-arguments exist for the conduction of the analysis. In *P. chrysogenum* the hyphae structure, the spatially resolved metabolism, aging and changing product formation is known [249]. In this study, these effects are neglected for conduction of ^{13}C-MFA, because the technique is not capable of resolving them. Besides, fungi exhibit significant changes of the metabolism during the cell cycle. Thus, performing analysis of labeling patterns will give

averaged results over the cell cycle in these cells. To give an example, in *Saccharomyces cerevisiae* even synchronization of cell cycle was frequently observed in chemostat culture [190].

To introduce the application and the work-flow for ^{13}C-MFA, an illustrative example is subject of the next section.

2.2.3 Illustrative Toy Example: The Spiral Model

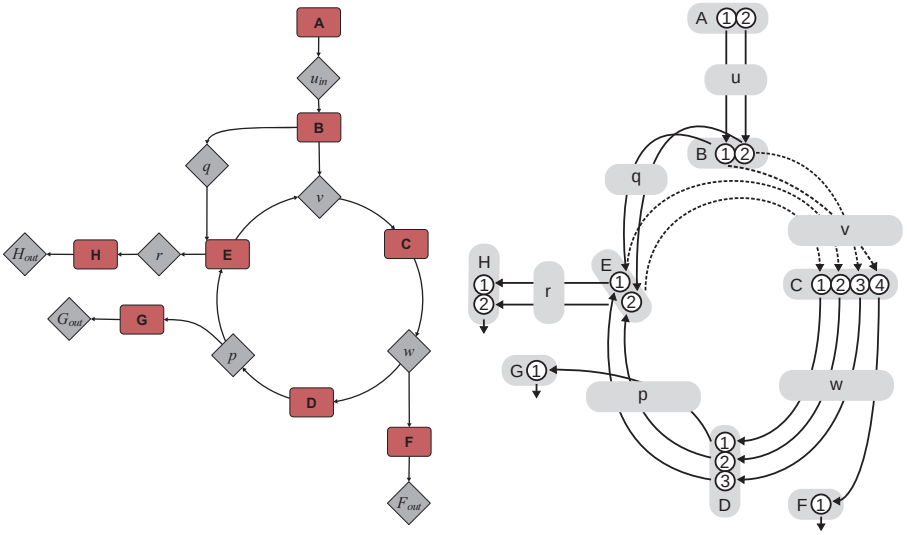
^{13}C-MFA is usually dealing with high dimensional metabolic models with 20-40 degrees of freedom. Most of the characteristics can be seen also in lower dimensions. In this section certain aspects of ^{13}C-MFA, becoming later important for large scale system, will be introduced with an artificial toy example.

This model, called Spiral model, was first introduced by Wiechert *et al.* [267]. The metabolic network model consists of nine reactions with an input pool A, see figure 2.5a. The influx u_{in} is fixed to 1.0. Two free fluxes exist q_{net} and q_{xch}. The exchange fluxes will be rescaled to match the interval $[0, 100\%]$ by applying the transformation $q_{xch01} = 100 \cdot q_{xch}/(1 + q_{xch})\%$, whereas 0 means no exchange flux and 100 % means an infinite exchange flux [263]. As q_{net} depends on the influx u_{in} which is constrained to a value of 1.0, it can only be varied within the interval $[0, 1]$.

The atom transition network is shown in figure 2.5b. It specifies the fate of the carbon atoms for each reaction. The atom transition network and the metabolic network can be formulated in the short notation including atom transitions:

$$
\begin{aligned}
u_{in} &: \mathrm{A}\#ab \longrightarrow \mathrm{B}\#ab \\
q &: \mathrm{B}\#ab \longrightarrow \mathrm{E}\#ab \\
v &: \mathrm{B}\#ab + \mathrm{E}\#cd \longrightarrow \mathrm{C}\#abcd \\
w &: \mathrm{C}\#abcd \longrightarrow \mathrm{F}\#a + \mathrm{D}\#dcb \\
p &: \mathrm{D}\#abc \longrightarrow \mathrm{E}\#cb + \mathrm{G}\#a \\
r &: \mathrm{E}\#ab \longrightarrow \mathrm{H}\#ab \\
F_{out} &: \mathrm{F}\#a \longrightarrow \\
G_{out} &: \mathrm{G}\#a \longrightarrow \\
H_{out} &: \mathrm{H}\#a \longrightarrow
\end{aligned}
$$

For this network the isotopomer balance equations can be written in short notations, with subscripts $i, j, k, l \in \{0, 1\}$ denoting labeling of single atoms. The lower case letters are referring to the respective isotopomer fractions of an metabolite:



(a) Metabolic network (b) Atom transition network

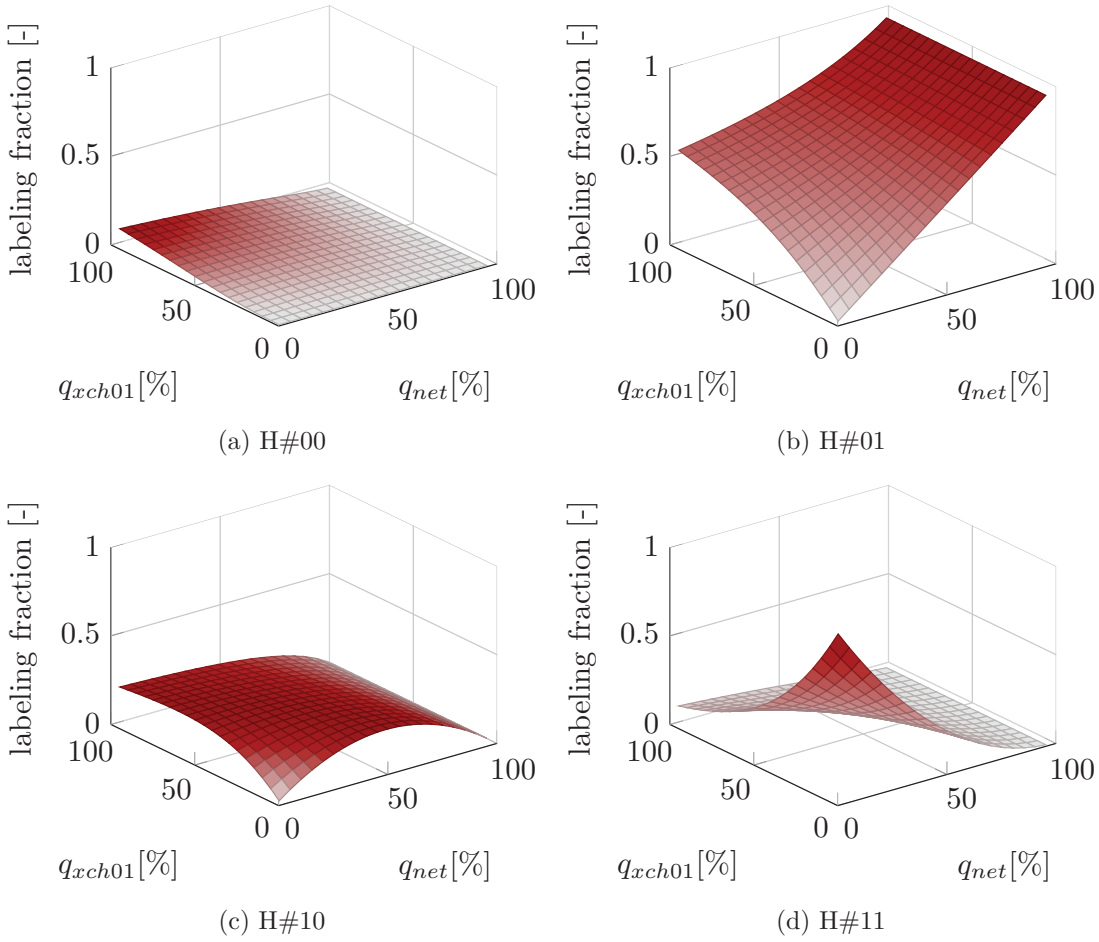
figure 2.5: Spiral network model posses one influx u_{in} and several effluxes F_{out}, G_{out} and H_{out}. (a) Metabolic network and (b) carbon atom transition network for the Spiral model.

$$
\begin{aligned}
\text{Pool Change} \quad &= \text{Arriving} && - \text{Leaving} \\
\text{B}: \frac{d([B]b_{ij})}{dt} \quad &= u_{in} \cdot a_{ij} && -(v+q) \cdot b_{ij} \\
\text{C}: \frac{d([C]c_{ijkl})}{dt} &= v \cdot b_{kl} \cdot e_{ij} && -w \cdot c_{ijkl} \\
\text{D}: \frac{d([D]d_{ijk})}{dt} &= w \cdot (c_{ijk0} + c_{jkl1}) && -p \cdot d_{ijk} \\
\text{E}: \frac{d([E]e_{ij})}{dt} \quad &= p \cdot (d_{0ij} + d_{1ij}) + q \cdot b_{ij} && -(r+v) \cdot e_{ij}
\end{aligned}
$$

In these isotopomer balancing equations, a special feature of the ^{13}C-MFA models can be seen: In the balance equation of metabolite C a nonlinear term $b_{kl} \cdot e_{ij}$ is introduced as two metabolites (B and E) react to a new metabolite C. Thus, it is a system of nonlinear differential algebraic equations.

Following, it is assumed that all isotopomers for metabolite H are measurable and the system is in metabolic and isotopic steady-state ($\frac{d([X]x)}{dt} = 0$). Now, the labeling pattern in H can be calculated by specifying the free fluxes q_{net} and q_{xch} and an input substrate. In this example, the input substrate A_{in} is labeled at the second position A#01. Equation 2.9 is solved and the isotopomers fractions for H are obtained. This procedure is called henceforth (forward) simulation.

Calculating the labeling pattern for all flux values of q_{net} and q_{xch} results in the surface plots seen in figure 2.6. The resulting isotopomer fractions are smooth, but several of them (H#00 and H#10) show only small changes in value. Additionally, some isotopomer fractions show strong changes in function values with respect to the net flux (cf. H#01). For higher exchange fluxes q_{xch01} within 75-100 %, small changes are observed in the labeling patterns, because pools H and B are equilibrated. At a high net flux $q_{net} = 1.0$ all flux is redirected



(a) H#00 (b) H#01

(c) H#10 (d) H#11

figure 2.6: Simulation results for isotopomers of metabolite H in Spiral model with varying free fluxes q_{net} and q_{xch}. Influx u_{in} was constrained to 1.0. Input substrate was chosen to be A_{in}#01.

from metabolite A via B to E, bypassing the route via C and D. Thus, the exchange flux between pool B and E becomes non-influential as both metabolites share already the same labeling pattern.
However, if two sets of flux distributions are chosen, we can see that they possess different isotopomers fractions in metabolite H. Thus, to determine the fluxes for this model, the measurements of the isotopomers of H need to be known. This is true, as long as the q_{net} flux is not equal to 1.0. In this case, no information content about this exchange flux is found in the labeling pattern, because two flux values for q_{xch} possess the same isotopomer measurements. This is called structural non-identifiability [199]. Thus, q_{xch} is only locally structural non-identifiable at high fluxes of q_{net}.

2.2.4 From Labeling Patterns to Fluxes

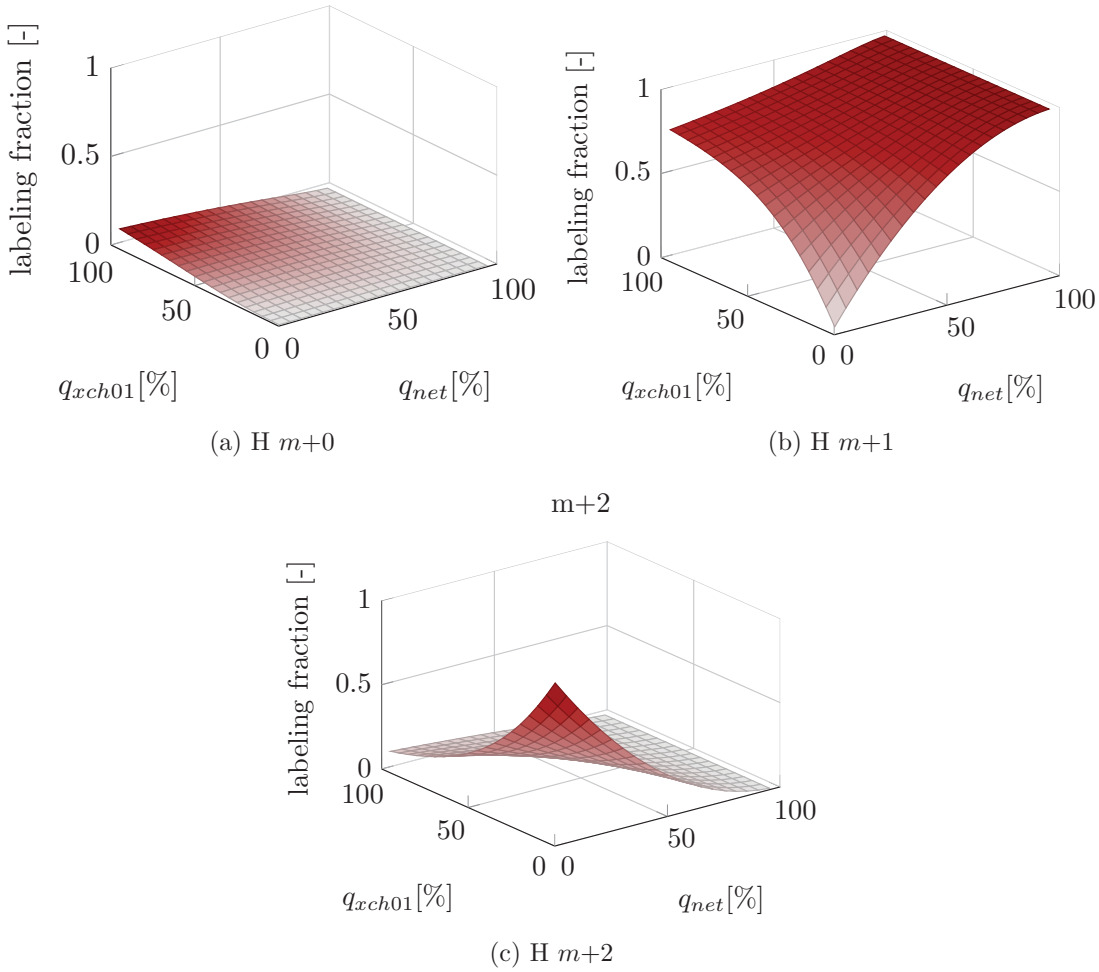
It was shown that measuring the labeling distribution of the metabolite H will determine the fluxes of the Spiral model. Nevertheless, there are some drawbacks emerging from this: (I) measuring the whole isotopomers of a metabolite is cumbersome, if not impossible for

large molecules, and (**II**) measurements are never exact.
Here we will see how mass spectrometic measurements of a metabolite are built from the simulated isotopomer distribution. Metabolite H is measured by mass spectrometry and possesses the isotopomers H#00, H#10, H#01 and H#11. The mass spectrometric devices can separate metabolite H by mass m and quantify the masses. The fraction of completely unlabeled H (m+0; H#00), labeled at one position (m+1; H#10 + H#01) and labeled at two positions (m+2; H#11) can be measured. The detected part of H contains n_C carbon atoms and thus $m + 0, 1 \ldots (n_C + 1)$ so called mass traces are measured (here: three were measured). Thus, a single MS measurement can only measure a subset of the isotopomers information content. By solving equation 2.9, the full labeling state of H is obtained, describing the fractions of all simulated isotopomers x. The measurements can be calculated from the labeling state using the measurement matrix M. For this case, equation 2.10 can be reformulated for metabolite H:

$$y_{M,H} = M_H \cdot x_H$$

$$\begin{bmatrix} y_{H,m+0} \\ y_{H,m+1} \\ y_{H,m+2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_{00} \\ h_{10} \\ h_{01} \\ h_{11} \end{bmatrix}$$

The results for the Spiral models metabolite H can be seen in figure 2.7. Clearly, at high flux values of the net flux q_{net} low changes are found compared to the full set of isotopomers. Here, the loss of information is reflected in the MS measurements‘ surface plots.



(a) H m+0

(b) H m+1

(c) H m+2

figure 2.7: Simulation results for MS measurement of metabolite H in Spiral model with varying free fluxes q_{net} and q_{xch}. Influx u_{in} was constrained to 1.0. Input substrate was chosen to be A#01.

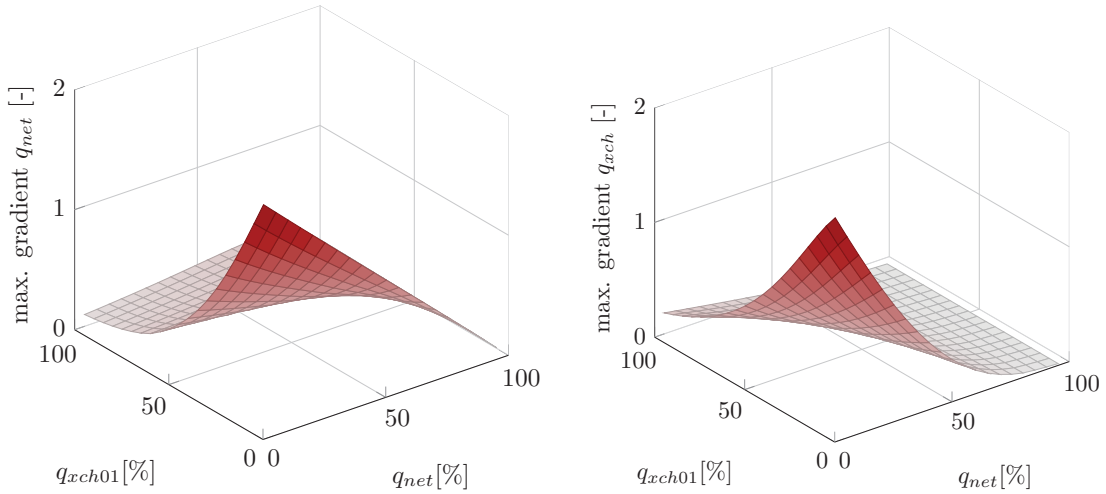
2.2.5 Sensitivities and Jacobian

Sensitivities describe the local effect of the input parameters on the output of the model. Usually, the partial derivative for one flux with respect to its output y is calculated at a specific flux distribution $\hat{v}$:

$$\left|\frac{\partial y}{\partial v_i}\right|_{\hat{v}}$$

The output of the model y can be a labeling y_M and/or flux measurement y_v.

This local sensitivity analysis indicates, at the point in flux space, the strength of the dependency between a measurement and a free flux. Calculating all local sensitivities for fluxes and measurement, one obtains the Jacobian with dimension $p \times (n_y + n_v)$:



(a) q_{net} (b) q_{xch}

figure 2.8: Sensitivities (absolute values) calculated for Spiral model with influx u_{in}=1.0 and input substrate 100 % A#01. The Jacobian was evaluated and the maximal gradient was extracted for (a) flux q_{net} and (b) flux q_{xch}. Only maximal absolute values are of all MS measurements of metabolite H.

$$J = \left| \begin{array}{ccc} \frac{\partial y_{M,1}}{\partial v_1} & \cdots & \frac{\partial y_{M,1}}{\partial v_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{M,ny}}{\partial v_1} & \cdots & \frac{\partial y_{M,ny}}{\partial v_p} \\ \frac{\partial y_{v,1}}{\partial v_1} & \cdots & \frac{\partial y_{v,1}}{\partial v_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{v,n_v}}{\partial v_1} & \cdots & \frac{\partial y_{v,n_v}}{\partial v_p} \end{array} \right|_{\hat{v}}$$

The Jacobian describes the local approximation of the change of all model outputs to a change in the free fluxes. The sensitivities were calculated for the Spiral model for the MS measurements of H for both fluxes q_{net} and q_{xch}. Only the maximal absolute values for the sensitivities for each parameter with respect to the labeling measurements are shown in figure 2.8. Both fluxes exhibit high gradients at low q_{net} and q_{xch}. Finally, both fluxes possess (close to) zero gradient in J when the q_{net} flux is high.

Until now, it was assumed that the measurements y_M possess no measurement error. However, measurements are subject to uncertainty and are in general assumed to possess normal distributed error. To quantify the uncertainty by error propagation from the measurements to the fluxes, statistics need to be evaluated. This can be achieved for the nonlinear model by linearization at the flux distribution $\hat{v}$. Therefore, the Jacobian $J(x_{inp}, \hat{v}_{free})$ is calculated and the measurement‘s covariance matrix $\Sigma = diag(s^2_{meas,1}, s^2_{meas,2}, \ldots, s^2_{meas,n_M+n_v})$ can be build from the measurement‘s standard deviation s_{meas} neglecting covariances/correlation between the measurements [197]. Using these matrices, the Fisher Information matrix is calculated:

$$Fish = J^{\intercal}\Sigma^{-1}J \tag{2.11}$$

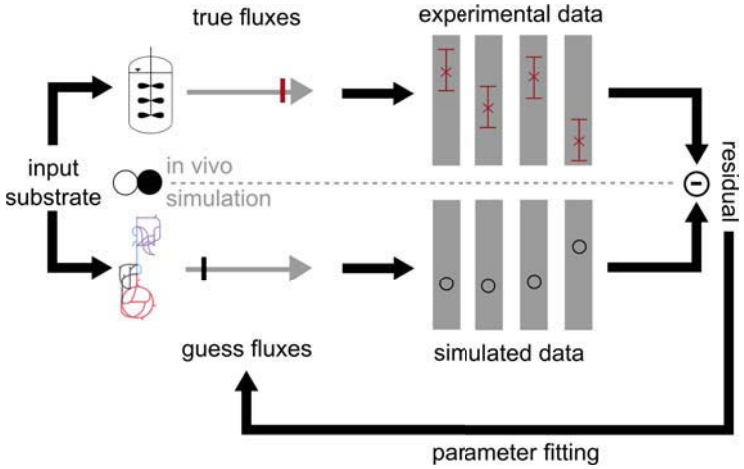


figure 2.9: Work-flow for flux estimation based on measurements. Adapted from Wiechert [261].

Inverting the Fisher Information will yield the covariance matrix of the free fluxes ($Cov = Fish^{-1}$). It describes the shape and size of the free fluxes‘ covariance ellipsoids and is a measure for the statistical identifiability of the fluxes [85].

Summing up, from a metabolic network and carbon atom transitions a unified description of the isotopomer balance equation was deduced. Labeling measurements were calculated by isotopomers and sensitivities were evaluated for the measurements with respect to the fluxes to perform a globalized sensitivity analysis in flux space. Finally, by assuming some error on the measurement it was shown that calculation of error propagation from fluxes to measurements is straightforward and closely related to sensitivity analysis.
Estimation of the free fluxes for a given set of measurements is the focus of the next chapter.

2.2.6 Flux Estimation

Up to now, a toy example was used and it was shown that there is a correspondence between isotopomers and fluxes. However, direct calculation of the fluxes from the labeling measurements is not possible. Besides ^{13}C-MFA deals usually with high dimensional systems ($p = 20 - 40$) and thus plotting of isotopomers (as in the case of the Spiral model) is not possible. The work-flow for estimating the fluxes from the experimental measurements is shown in figure 2.9.
First, an experiment is performed in the wet-lab with a chosen substrate mixture and the resulting stationary labeling patterns in the cell is measured. In the dry-lab the estimation of the unknown *in vivo* metabolic fluxes from the labeling patterns is started by choosing a set of free fluxes. With the same input substrate used in the model, the labeling measurements are simulated and compared to the real measurements. By minimizing the difference between simulated data and measurements, the simulated flux distribution is getting similar to the *in vivo* fluxes. If the difference is minimal, one assumes that the found flux distribution is the same as *in vivo.*
Each MS and flux measurement is prone to errors, thus they posses a standard deviation by the stochastic error introduced during experimental/measurement procedures. To measure how good a measurement fits the simulated data, an adequate measure is needed. The weighted residual sum of squares (WRSS) provides a measure for the goodness of fit [266].

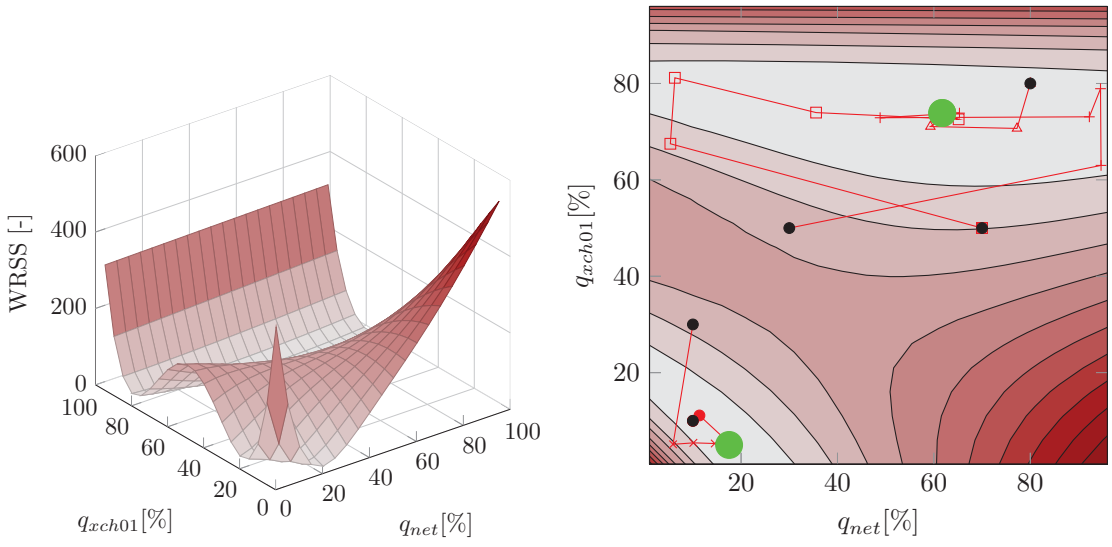
The WRSS weighs the squared deviate of measured and simulated measurement values by their experimentally determined standard deviations s_{meas}.
Measurements given in arbitrary unit or incomplete mass spectrometric measurements (e.g. m+0 measurement missing) are scaled to the simulated measurements by a scaling factor ω. If all mass traces were available and scaled to sum up to 1, ω is set to 1. Thus, the weighted deviate of simulated and measured labeling patterns and fluxes are minimized [259]:

$$\hat{v} = \arg\min_{v_{free}} \mathrm{WRSS} \tag{2.12}$$

$$= \arg\min_{v_{free}} \|y_{meas} - \omega \cdot y_M(v_{free}, x_{inp})\|_\Sigma + \|v_{meas} - y_v\|_\Sigma \tag{2.13}$$

$$= \arg\min_{v_{free}} \sum_i \frac{(y_{meas,i} - \omega \cdot y_{M,i}(v_{free}, x_{inp}, ...))^2}{s^2_{meas,i}} + \sum_j \frac{(v_{meas,j} - y_{v,j}(v_{free}))^2}{s^2_{meas,j}} \tag{2.14}$$

The optimization landscape given by the WRSS is visualized in figure 2.10a. For the Spiral example, a subset of the isotopomers of H were measured: H#00 0.02 ± 0.004; H#00 0.15 ± 0.01; H#10 0.17 ± 0.01. Minimizing the value of the WRSS results in a good fit of simulated and real measurements. It is assumed, that the flux distribution $\hat{v}$ at the minimal value of the WRSS resembles the true fluxes in the cell. In this example, two optima are obtained, one at q_{net}=0.18 and q_{xch}=0.05 with a WRSS of 37. The second optimum, is at q_{net}=0.62 and q_{xch}=0.74 with a value of WRSS=22. Due to the lower score, the latter is the global optimum and assumed to be the *in vivo* fluxes.



(a) WRSS (b) Contour plot of WRSS

figure 2.10: Plot of WRSS for the Spiral model for varied free fluxes q_{net} and q_{xch}. Influx u_{in} was constrained to 1.00. Measured were isotopomers of metabolite h: H#00 0.02 ± 0.004, H#00 0.15 ± 0.01 H#10 0.17 ± 0.01. With global optimum of at with q_{net}=0.62 and q_{xch}=0.74 and WRSS=22. And local minimum at q_{net}=0.18 and q_{xch}=0.05 with a WRSS of 37. (a) WRSS values. (b) Contour plot of WRSS and paths of MATLAB® *fmincon* optimizer. Black dots are starting points, green dots mark the optima. Red lines indicate search path.

The problem is a constrained nonlinear optimization problem. In figure 2.10b the trail of an optimizer started at separate points in flux space is shown. Depending on the starting point the optimizer may be stuck in a local minimum. Clearly, the constrained nonlinear optimization problem can be solved by heuristic or global optimization strategies. This circumvents local but sub-optimal solutions. In the figure 2.10b a globalized multi-start strategy was applied. For this, several starting points were chosen to identify separate optima. Two out of five optimization runs found the local optima and stopped. Three found the global optimum.

2.2.7 Experimental Design

Experimental design aims at improving quality of the information about the estimated quantities of the system under investigation. Besides other potentially possible measures for quality (like costs, duration of experiment,...), it usually aims at improving the statistical identifiability of the estimated model parameters. In ^{13}C-MFA it is aimed at improving the statistical identifiability of the estimated fluxes $\hat{v}$. This can be achieved by changing the set of design variables directly influencing the information content of an experiment: (I) standard deviation, (II) measurement configurations, and/or (III) the input substrate.
The quality of the experiments is intended to be increased by shrinking the volume of the covariance ellipsoids leading to better identifiable fluxes. However, the calculation of the covariance matrix depends on the input substrate mixture x_{inp}. Choosing the substrate mixture is crucial for the statistical quality of the estimated fluxes. For example, using only uniformly labeled substrate (^{12}C or ^{13}C) results in no information at all.
In classical experimental design the covariance matrix of the parameters (here: free fluxes) is converted to a single value which describes the overall statistical estimation quality of the system. The criteria are called alphabetical design criteria as they are denoted by capital letters. Examples for frequently applied criteria are in figure 2.11, an overview can be found in Franceschini and Macchietto [85]. Maximizing the information criteria for the Fisher Information or minimizing them for the covariance matrix results in the same outcome. The inversion step in the calculation of the covariance matrix may introduce numerical errors. Therefore, the Fisher Information is used.
If frequently used substrates in a experimental design study are applied, an optimal substrate mixture can be found. Frequently applied substrates are uniformly labeled glucose, 1-labeled glucose and naturally labeled glucose.

Switching back to the Spiral model: it is used to perform an experimental design study at separate points in a ternary mixture triangle of the substrates A#00, A#01 and A#11. At each point in mixture triangle information criterion calculation is performed, using the respective mixture (see figure 2.11).
A distinct optimum can be found for A#01. All other mixtures show a lower potential to increase statistical identifiability for the chosen fluxes indicated by lower D-criteria values. For performing an experimental design study, like shown above, preliminary information about the flux values are needed (*a posteriori* experimental design), because the Jacobian and thus also the Fischer Information and covariance matrix provide local measures for a given flux distribution $\hat{v}$. To summarize: for another flux distribution the mixture traingle may look different. Next light is shed on this issue. For the mixture of 100% A#01, the statistics for the overall flux space is calculated (see figure 2.12). At each point the separate

Criterion	Criterion value
Modified A	$\max trace(Fish)$
C	$\max min(diag(Fish))$
D	$\max det(Fish)$
E	$\max \lambda_{min}(Fish)$

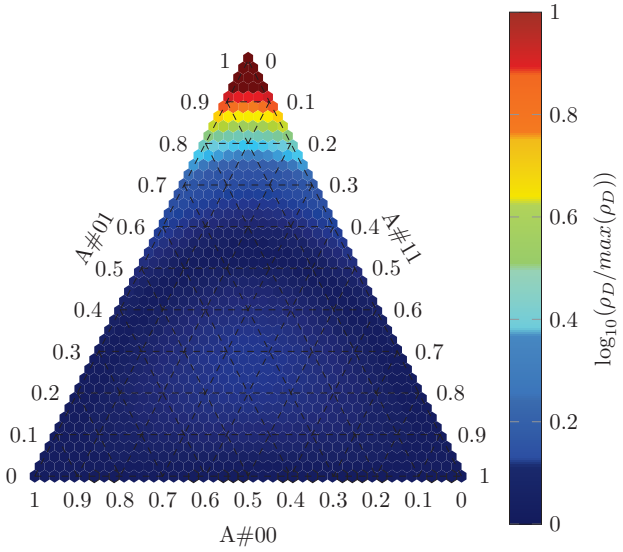


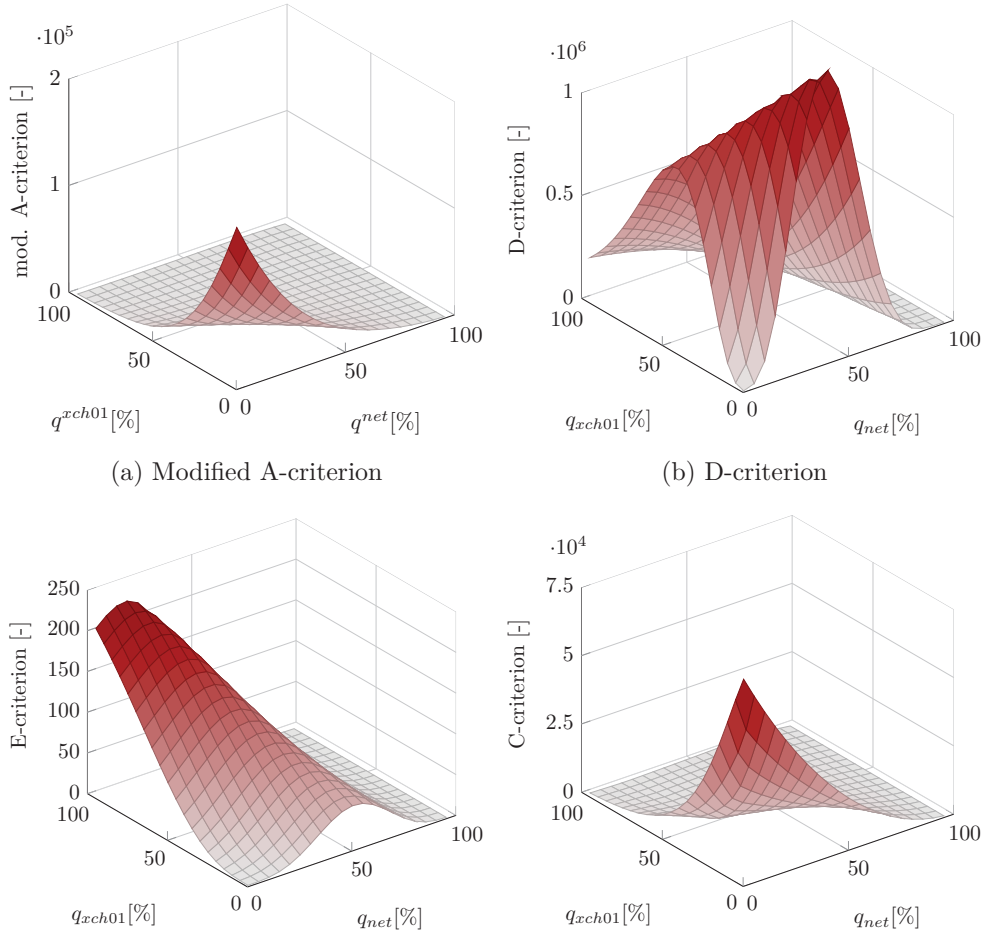
figure 2.11: Experimental design study for spiral model with fluxes $q_{net} = 0.2$ $q_{xch} = 0.01$ based on Fisher information matrix. A full MS measurement (m+0, m+1, m+2) of metabolite H was used. Logarithmic D-criterion values were scaled to interval $[0, 1]$. A value of 1 corresponds maximal statistical identifiability. In the table the optimal experimental design criteria are given, where $\lambda_{min}$is the minimal eigenvalue of the Fisher information [192, 250].

criteria are evaluated. The criterion values are strongly dependent on the fluxes and showing deviating behavior as their optima are located at different positions. This poses a problem on the conduction of the experimental design: Preliminary knowledge about the fluxes from literature or an reference experiment is needed. In the event that some fluxes are weakly statistical identifiable, obtaining an optimal substrate mixtures becomes harder, because the criterion value of these fluxes may vary strongly within their range of uncertainty.
In contrast to the classical experimental design, where an estimate for the parameter is already available, an experimental design without preliminary information about parameters is called *a priori* experimental design henceforth.

The used statistics in this section were founded on local approximation of the change of model outputs to a change in fluxes (Jacobian). Thus, this type of statistic is a so-called linearized statistics. In the next section, the more general nonlinear statistics will be applied to the Spiral model.

2.2.8 Nonlinear Statistics

The Spiral model is nonlinear, thus, evaluation of introduced bias by the linearized statistics (cf. equation 2.11) needs to be performed. Nonlinear statistics can be based on a nonlinear measure or a bootstrap method, both types are explained in section 4.2.2 and 4.2.4. As already introduced, the linearized statistics are a local approximation specific for given flux values. In contrast to that, nonlinear statistics are not a local measure as they are not calculated for a single point in flux space.
Based on a given set of measurements linearized, bootstrap, and nonlinear statistics (based on



(a) Modified A-criterion

(b) D-criterion

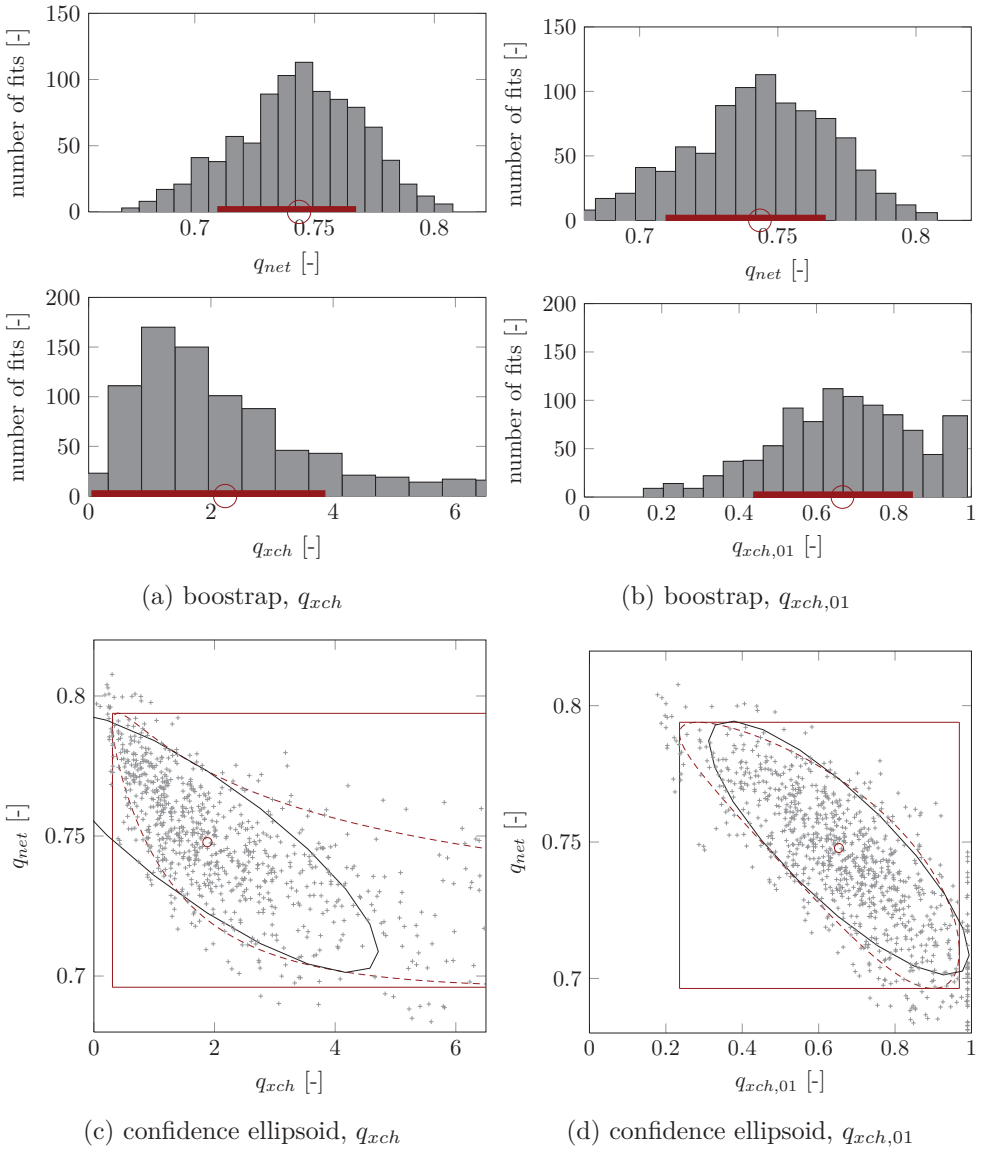
(c) E-criterion

(d) C-criterion

figure 2.12: Changes in optimal experimental design criteria by altering flux distribution for the spiral model with a fixed mixture of A#01 as input substrate. A full MS measurement (m+0, m+1, m+2) of metabolite H was used.

equation 4.30) were calculated for the free fluxes q_{net} and q_{xch} (see figure 2.13). The linearized statistics results in standard deviations of the free fluxes: $q_{net} = 0.75 \pm 0.03$ and $q_{xch} = 0.65 \pm 0.20$ mmol $g_{CDW}^{-1}\, h^{-1}$. The bootstrap method predicts a confidence interval of q_{net} of $[0.71, 0.77]$ and for q_{xch} of $[0.05, 3.87]$. For the nonlinear statistics (based on equation 4.30), the confidence interval is greater compared to the bootstrap and linearized statistics: $q_{net} = [0.70, 0.80]$ and $q_{xch} = [0.31, \infty)$ (upper boundary not determinable). The confidence intervals of the net flux corresponds well to the two methods of nonlinear statistics. However, the exchange flux shows strong deviation.
After transformation of exchange flux from q_{xch} to q_{xch01} by using $q_{xch01} = q_{xch}/(1 + q_{xch})$, the exchange flux in the Spiral model is close-to-linear in its behavior, see figure 2.13d. By this transformation of the exchange fluxes to the interval $[0, 1]$, linearized statistics provides a good approximation. The effect of transformation to obtain close-to-linear behavior for confidence intervals was also extensively described by Ratkowsky [197].
To sum up, linearized statistics applied on these nonlinear models can result in false approximations of the confidence intervals. Thus, care has to be taken in interpretation of the results.

By using the Spiral model as a toy example, basic steps of the work-flow for ^{13}C-MFA were introduced. Next, the major aim of this study is shown and the big picture of the complex work-flow for applying ^{13}C-MFA is demonstrated for real-world examples.



(a) boostrap, q_{xch}

(b) boostrap, $q_{xch,01}$

(c) confidence ellipsoid, q_{xch}

(d) confidence ellipsoid, $q_{xch,01}$

figure 2.13: Nonlinear and linearized statistics for Spiral model using NoReTo (cf. section 4.2.4). (a-b) Result of Monte Carlo bootstrap to calculate error propagation to the flux q_{net} and q_{xch}, exchange flux q_{xch} was rescaled to "01" coordinates in (b). Calculations were performed according to Joshi *et al.* [112]. (c-d) Comparison of nonlinear error propagation by bootstrap (+), linearized (——) and nonlinear (- - - -) confidence intervals both with significance level of 0.32, exchange flux q_{xch} was rescaled to "01" coordinates in (d). Measurements were: H#01 0.15, H#01 0.17, H#00 0.02, H#01 0.14, H#01 0.16, H#00 0.02. For bootstrap the same measurement values and standard deviation of 0.012 was used, it was estimated using the best fit's residual to calculate the variance of unit weight $\sigma_0^2 = \frac{RSS_{min}}{n-p}$ applying the NoReTo toolbox [5].

Chapter 3

Transferring ^{13}C-MFA to an Industrial Process

The ultimate aim of this study is to elucidate the fluxome of *P. chrysogenum* under industrially relevant conditions. As discussed in the beginning, *P. chrysogenum* is cultivated on glucose/sucrose or other crude sugars in controlled fed-batch processes in industry. The production process takes place in 100-400 m^3 bioreactors and process duration is 120 to 200 h, the growth form is pelleted [66]. A production of at least 66 $g\,L^{-1}$ penicillin V is needed for a modern fed-batch process to be amortized, while the final biomass concentration is in the order of 40-50 $g\,L^{-1}$ (data estimated by Monte-Carlo simulations by Biwer *et al.*)[26]. Thus, high cell densities are found within such a process. The increased biomass density in a pellet can result in substrate limitation in the center [89].
In order to perform ^{13}C-MFA, certain prerequisites are to be met. Ideally, ^{13}C-MFA is performed in well-controlled chemostats under metabolic and isotopic steady-state. As valuable labeled compounds are used, the investigated process is scaled down to the liter scale. Importantly, this step is also done to obtain homogeneous conditions in the bioreactor and to prevent effects by slow mixing, because this can result in heterogeneous conditions within the cultivation. Finally the varying conditions result in a multitude of metabolic states that are not resolvable by ^{13}C-MFA. Besides it is known that different metabolic states can be found within the pellets formed by *P. chrysogenum* [130, 89]. The sum of these different states will ultimately result in the macroscopic, averaged, characteristics of the cultivation process yielding the final product (penicillin) of interest. Thus, the inhomogeneities are important for the understanding of the process.
Aiming at characterization of the industrial process ^{13}C-MFA needs to be conducted as close to the industrial process as possible, without severely compromising the basic modeling assumption beforehand mentioned. Nevertheless, this requires simplifications: The industrial process was scaled down and the growth form was changed to dispersed growth with reduced biomass to minimize inhomogeneities emerging from mixing and transport processes within the pellets to optimally apply ^{13}C-MFA. Thus it was a balancing act between the shortcomings of the technique and the industrial application. By these changes, the adapted process differs from the original process and is, henceforth, called close-to-industrial.
To mimic the large scale processes, the sensor bioreactor concept was applied (see figure 3.1) [61, 62, 150]. A large scale bioreactor with 150 liter working volume, representative for the industrial process, and a small "sensor" bioreactor with 1 liter working volume were used. The sensor bioreactor is controlled in master/slave mode with respect to the large fermenter,

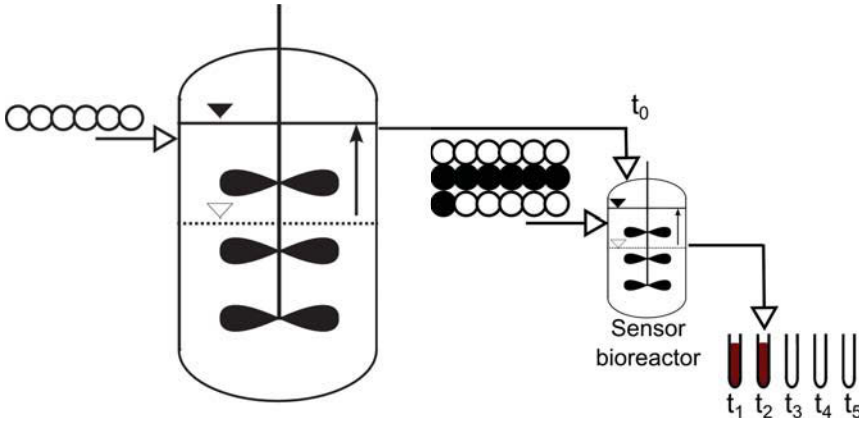


figure 3.1: Experimental setup of the sensor bioreactor for ^{13}C-MFA

resulting in close-to-industrial conditions in the small sensor bioreactor. This approach has the advantages that, due to the sensor bioreactor‘s considerable smaller volume, the usage of costly labeled material is significantly reduced. Additionally, several labeling experiments can be conducted sequentially within one cultivation to enable time-resolved monitoring of the process. Moreover, it provides less heterogeneous areas although it is comparable in process conditions to the process in the large scale fermenter.

Furthermore, implications caused by the non-model organism *P. chrysogenum* have to be solved when establishing ^{13}C-MFA. First, aging of culture and potentially decreasing product titer (degeneration) were often observed [59]. Second, polarized growth results in subpopulation of cells have been found, whereas ^{13}C-MFA will only represent averaged values for the overall cultivation.

As the culture undergoes subsequent changes and *P. chrysogenum* is a slow growing organisms, the time frame for experimental observation was shifted from estimated process duration of about 120-200 h to a time frame of 95 h for the fed-batch process. By this time saving, experimental cycles with subsequent experimental design could be conducted.
Additionally, further constraints were present for evaluation of the experimental data used in this study. Time constants of labeling are heterogeneous leading to the observed labeling data in transient as well as stationary labeling regimes. In order to apply ^{13}C-MFA to estimate intracellular fluxes, stationary values have to be estimated from the measured enrichment.

Unfortunately, quantitative pool size measurements protocols and highly time resolved data for central carbon metabolism were missing, preventing conduction of non-stationary ^{13}C-MFA for the data-sets [181].

Having these measurements at hand, stationary ^{13}C-MFA was applied to several process conditions. First, chemostat experiments were conducted to establish the ^{13}C-MFA for the organism and to circumvent the shortcomings of a fed-batch process. Afterward, the fed-batch process was carried out using the sensor bioreactor concept for close-to-industrial conditions. Additionally, experimental design was applied to increase statistical significance of the results.

As motivated with the simple Spiral example, the work-flow for ^{13}C-MFA was introduced. It is revisited in the next section, to provide the outline of this work.

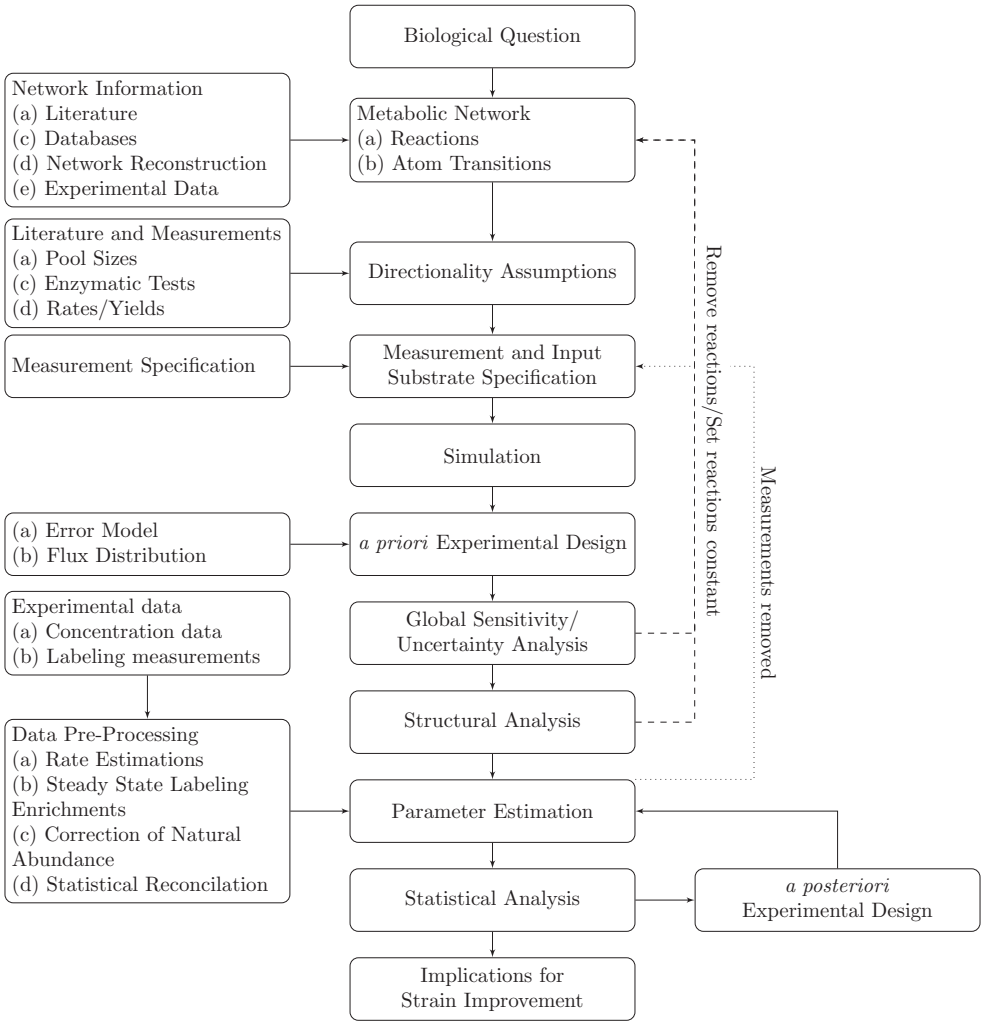


figure 3.2: Work-flow of ^{13}C-MFA.

3.1 State-of-the-Art Work-Flow for ^{13}C-MFA

^{13}C-MFA is applied in the industrial environment to gain knowledge about the organism and get hints for strain improvements to increase its production capabilities. Here, the optimal work-flow for conduction of this analysis is presented that generally applies to all ^{13}C-MFA studies (see figure 3.2). It is important to note, that the work-flow considers all steps related to model based evaluation and analysis.

^{13}C-MFA Model

Construction of a metabolic network with reactions and atoms transitions is the first step in every ^{13}C-MFA study. Necessary organism-specific information can be taken from databases or literature. Besides, it is also possible to build or use a genome scale metabolic network and to reduce the number of reactions for the ^{13}C-MFA. The generation process of these networks cannot clearly be classified as *bottom-up* or *top-down*, since it depends on the actual state of knowledge about the organism [175]. This also includes atom transitions which can be specified by literature and text-book knowledge [265].

Assumptions about the directionality of reactions are introduced in the model. They can be based on pool size measurements for thermodynamic analysis [265]. Enzymatic tests can be used to exclude "inactive" reactions from the network or characterize the enzymes preferred directionality like shown by Stols and Donnelly [228].
Next, measurements' specifications are introduced in the model for the potentially measurable metabolites; experimental data are not yet introduced. Up to this point only the raw model without measurement data is built. Nevertheless, at this point the first (forward) simulation can be performed, simulating the measurement values for an arbitrarily chosen flux distribution as previously shown for the Spiral model. This unlocks the basis for investigation of the model. Based on the simulation model *a priori* experimental design can be performed. This may be done by, for example, choosing several flux distributions from literature and assuming an error model for the measurements to estimate the optimal substrate mixture. The determined mixture is used henceforth to conduct subsequent steps.

The models build for P. chrysogenum and their generation process are described in chapter 5. From a large scale network, including cofactors, a focused network was built and assumptions were stated. The effluxes to biomass compounds are calculated and introduced into the model. A substrate mixture was chosen based on literature data and measurements are specified. Finally, FBA is used to perform yield analysis.

Exploration of ^{13}C-MFA

To determine sensitivity of the model parameters, i.e. fluxes, a global sensitivity analysis is performed. This is important in order to prevent over-parametrization. Usually, the sensitivity analysis is accompanied by uncertainty analysis taking also into account the measurements uncertainties. Next, a structural analysis is accomplished to check which fluxes are structurally non-identifiable, i.e. cannot be resolved even assuming error-free measurements [45]. By the results of these analyses, either fluxes may need to be removed or set constant leading to cycles of model adaption. Besides, if important fluxes for the to-be-answered biological question are not sensitive or identifiable, cycles of experimental design need to be conducted to find measurements that increase their sensitivity.

In this work, the first global sensitivity analysis for ^{13}C-MFA models is performed. Adequate visualization and influences of the fluxes on the labeling pattern are shown in chapter 6. Global uncertainty and structural identifiability analysis are not subject of this work.

Data Pre-Processing and Conduction of ^{13}C-MFA

After construction and theoretical investigation of the model, the measured experimental data can be introduced to perform parameter estimation. For this, the extracellular rates and the steady-state labeling patterns are needed.

In this work, extracellular rates are estimated by kinetic modeling of the complex close-to-industrial fed-batch and chemostat cultivations. The steady-state labeling data are derived by model-driven data extrapolation, followed by their correction for natural abundance and, finally, a statistical reconciliation step (see chapter 4).

Proceeding in the work-flow, parameter estimation has to be performed by globalized (heuristic) strategies to circumvent local minima. At this point it might happen that flawed measurement data are found which need to be removed from the model. Likewise, the model (including stoichiometry and flux directionality) may be subject to a thorough review procedure. This results in cycles of adaption by re-specification of the model or measurements in the work-flow and re-performing of some steps. Finally, the results of parameter estimation need to be re-examined, to detect whether alternative local or multiple, equally good, optima are present. If several optima are present, these need to be carefully analyzed. Infeasible solutions (identified by expert knowledge about thermodynamics, biology, ...) may be discarded. If the resulting flux maps represent distinct but different possible solutions, further experimental data are needed to exclude the one or the other hypothetical flux map.

Results of parameter estimation, statistical analysis of results for chemostat and fed-batch cultivations are shown in chapter 7.

Experimental Design

If the resulting flux maps is weakly statistical identifiable, flux maps can be improved, in a statistical sense, by a second *a posteriori* experimental design and a thereafter performed carbon labeling experiment. Finally, after calculation of the statistics, the flux map is interpreted and hints for strain improvement can be obtained.

Experimental design is conducted and further experiments are performed, see chapter 7.1.2. The results of the outcome of the studies are discussed chapter 8. Conventional experimental design is extended by diversity-driven experimental design chapter 9. Finally, planing of optimal, yet economic, experimental design facing several optimality criteria is described in chapter 10.

Part II

^{13}C Metabolic Flux Analysis for *P. chrysogenum*

Chapter 4

Data Pre-Processing

For ^{13}C-MFA, two types of measurements are required: (I) labeling patterns of intracellular intermediates and (II) extracellular rate measurements, i.e. substrate uptake and product secretion rates, need to be provided. From the design of the study and the constraints imposed by the biological system in combination with the close-to-industrial process, limitations arise which need to be accounted for in the pre-processing pipeline for ^{13}C-MFA (see also figure 4.1):

1. Sample withdrawal from the sensor bioreactor is only possible at certain time intervals to preserve the metabolic steady-state. Overall ~40 % of the reactor volume is removed by sampling in a fed-batch cultivation. This hampers direct calculation of extracellular rates. Additionally, side-product formation is complex in *P. chrysogenum*. Therefore, a kinetic model is used to estimate rates based on experimental concentration measurements, weight measurements of withdrawn samples, and feeds. The model is fit to the concentration data and rates are estimated. The same procedure was likewise applied for the chemostats (in section 4.1).

2. Beside the concentration data, the labeling pattern of the metabolites were measured time resolved. Several metabolites possessed labeling far from isotopic steady-state (in tricarboxylic acid (TCA) cycle and amino acid synthesis pathways). This forces extrapolation to steady-state labeling enrichment (see section 4.2).

3. The extrapolated MS measurements need to be corrected for natural abundance (see section 4.3).

These three steps will be conducted in this chapter.

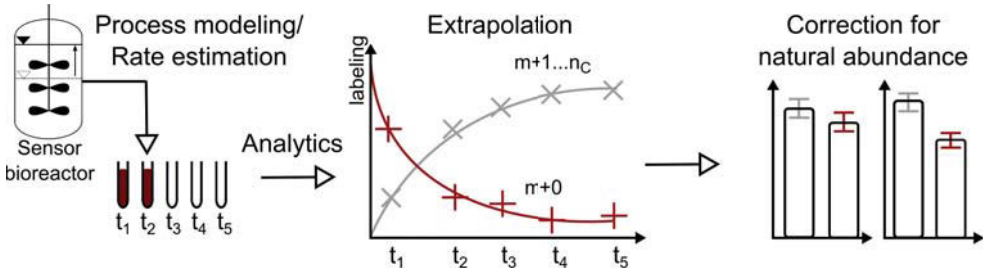


figure 4.1: Data pre-processing pipeline for ^{13}C-MFA established in this work.

4.1 Model-Based Extracellular Rate Estimation for *P. chrysogenum* Cultivations

In this section, the cultivations in chemostat and fed-batch processes are described by kinetic models to obtain consistent specific rate estimates for ^{13}C-MFA with closed mass balances. The cultivations of *P. chrysogenum* were described by an unstructured kinetic model for cellular growth [178]. The kinetic model describes all reactions participating in the formation of all known side- and main products measured in the chemostat and fed-batch cultivations. Extracellular concentration measurements were supplied for biomass, glucose, penicillin V, IPN, HOPenV, HOPOA, penicillamine (PeNH2), PIA, PIO, 6APA, 8HPA, and OPC. Additionally, off-gas analysis (of carbon dioxide) was used. bis-ACV was not detected during cultivations. Additionally, the uncertainty of the obtained rates is assessed by calculating standard deviations. In kinetic models, the mathematically expressed correlations between reaction rates q and the concentration C of a reactant are inserted into mass balances r. This allows description of whole bio-processes (e.g. cultivation in chemostat) [178]. In this chapter kinetic models are used to describe the complex side-product formations and growth of the *P. chrysogenum* BCB1 strain. In the end, estimates for the specific rates of the chemostat and fed-batch processes are derived from the primary cultivation data.
Next, the general nomenclature will be introduced in this section by starting with simple examples. For a product P the change in concentration C_P can be expressed by a equation r_P. It depends on the reactions q which consume or produce P. In this work only simple kinetics are used. To give an example:

$$q_{P,1} = k \cdot C_S$$

In $q_{P,1}$ the formation of P is linearly dependent on the concentration of the substrate S, with the reaction rate constant k. Furthermore Michaelis Menten kinetic are used describing the reaction catalyzed by a single enzyme [178, 262]. Using these kinetics a second reaction rate $q_{P,2}$ with the maximal possible reaction rate k_{max} and a Michaelis constant K_M can be formulated:

$$q_{P,2} = k_{max} \frac{C_S}{K_M + C_S}$$

A similar type of kinetics are called Monod kinetics. They are formally identical to Michaelis Menten kinetics and describe the growth behavior of an organism and not the reaction of a single enzyme as it is the case for Michaelis Menten kinetics [178]. In Monod kinetics the Michaelis Menten constant is changed to an affinity constant K_S.
The parameters (reaction rate constants) in the reactions rates q need to be estimated by parameter fitting to experimentally derived concentration data. Usually, estimates for these constants are available in literature and can be used to specify lower/upper boundaries.

In this work, an unstructured general kinetic model is built to consistently describe, both, chemostat and fed-batch process.

4.1.1 Penicillin and Side-Product Formation Revisited

The side product formation is important because it interferes with the penicillin production. A multitude of side-products are formed and can reduce the yield of the penicillin V. The

penicillin degradation pathway modeled in this study is sketched in figure 4.2. It is comprised of the main penicillin formation pathway (red) and the side-product formations (gray). Most side-products were detected and quantified in chemostat and fed-batch of the BCB1 strain. In addition to the degradation pathways which were already described in chapter 1.3, further degradation of PIO can occur. It is degraded to penaldic acid (PeA) and penicilllamine (PeNH2). PeA can be further degraded to penilloaldehyde (PeAl).
It was already described that degradation of penicillin V to HOPenV can occur. Additionally, POA is known to be degraded to p-hydroxy-phenoxyacetate (HOPOA), which can serve as a substrate for penicillin production to produce HOPenV [55].

4.1.2 Kinetic Model

An unified kinetic model was constructed for both, chemostat and fed-batch cultivations. In total five experiments, three chemostat and two fed-batch cultivations were performed, whereby the fed-batch processes were conducted in the sensor-bioreactor.
The reaction pathway of penicillin was based on the work of Pissara *et al.* [188]. In contrast to his work, intracellular concentration data were not available for the process under study and, thus, modeling was conducted purely based on extracellular measured concentrations. The underlying model of Pissara *et al.* was used as a reference and extended based on recent publications on side-product formation in *P. chrysogenum*.
All parameters, their limits and units used for parameter estimations are given in the appendix B.3.

Growth and Penicillin Formation

Growth of *P. chrysogenum* was modeled with standard Monod growth kinetics with the growth rate μ, the maximal growth rate μ_{max}, glucose concentration C_{Glc} and an affinity constant $K_{S,Glc}$ [178]:

$$\mu = \mu_{max} \cdot \frac{C_{Glc}}{(C_{Glc} + K_{S,Glc})} \tag{4.1}$$

The substrate uptake rate can be calculated based on the amount of substrate used for growth, penicillin production and maintenance. The maintenance coefficient m_s describes the amount of substrate used for processes not directly linked to biomass or product formation. To give some examples: it describes the amount of substrate/energy used for osmoregulation, turnover of macromolecules and cellular organization [187, 201]. To describe the substrate uptake rate the specific growth rate μ, penicillin production, and side-product formation rates (r_{PenV}, $r_{HOPenV,2}$, r_{IPN}, r_{OPC}) are taken into account. Moreover, biomass Y'_{XS} and product yield Y'_{PS}, and maintenance coefficient m_s are incorporated[178]:

$$q_{Glc} = \frac{\mu}{Y'_{XS}} + \frac{(q_{PenV} + q_{HOPenV,2} + q_{IPN} + q_{OPC})}{Y'_{PS}} + m_s \tag{4.2}$$

Here, only one yield coefficient was used for all formed side products, because determining separate yield coefficients was not possible based on the provided data-sets.
To prevent modeling of intracellular pools, product formation was represented by formation of the product “Base” (the penicillin nucleus, see figure 4.3). From this, the formation of

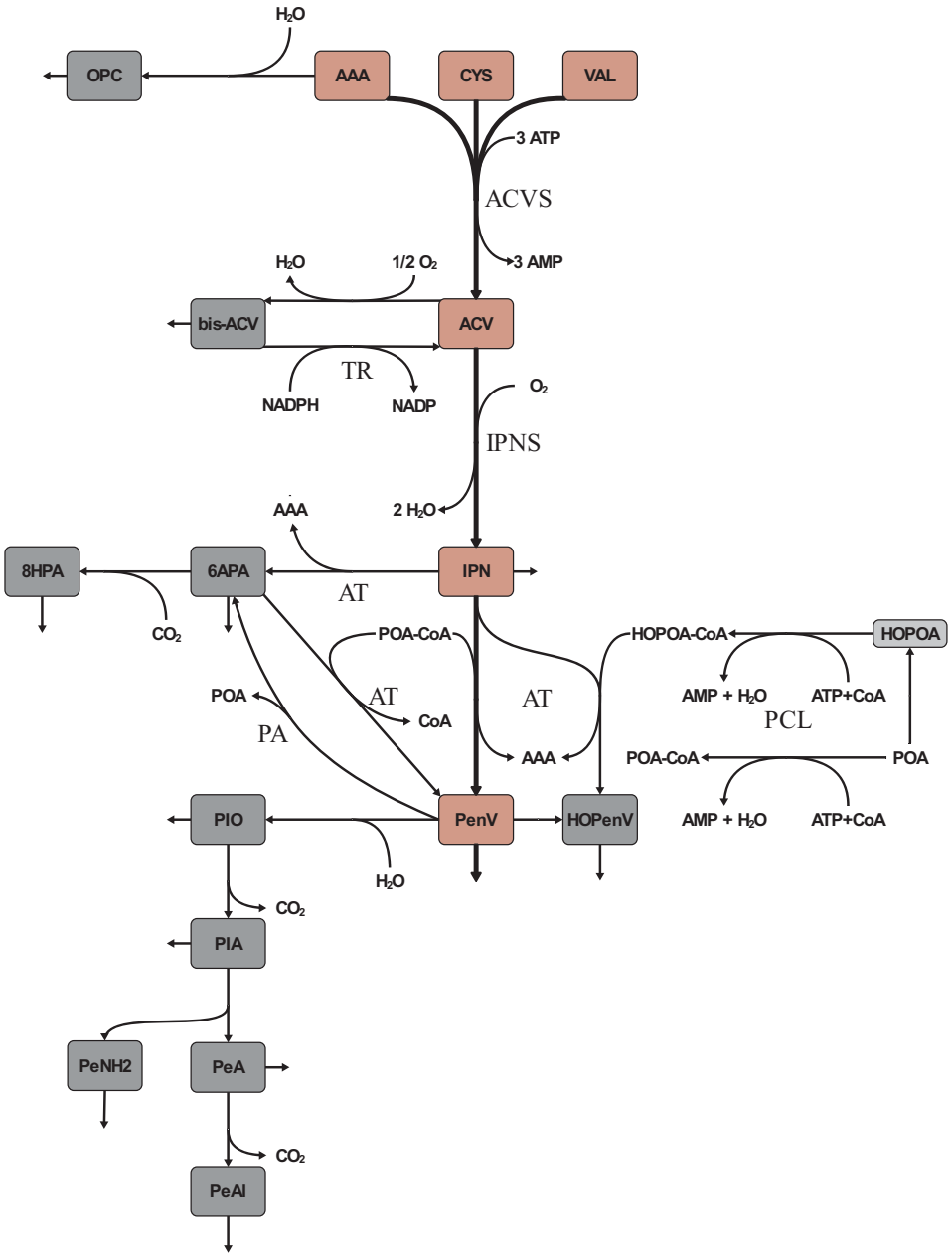


figure 4.2: Penicillin degradation pathways involved in fed-batch and chemostat cultivations. In comparison to figure 1.5 side-product formation was extended by the degradation of PIA and formation of HOPOA and HOPenV.

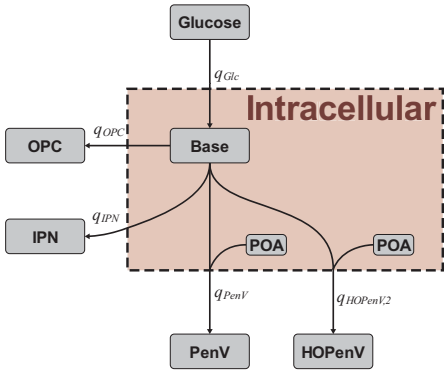


figure 4.3: Modeling of penicillin V, OPC, HOPenV and IPN formation. From the precurso glucose, a base

penicillin V, HOPenV, 6APA and OPC were modeled. In the conducted simulations significant product inhibition by glucose was not found to be important and was removed from the model. This was also described by Tiller *et al.* [237]:

$$q_{Base} = C_{BM} \cdot \frac{C_{Glc}}{(C_{Glc} + K_{S,Base})} \tag{4.3}$$

From this product formation rate, the separate substances are formed: penicillin V, OPC, HOPenV and IPN. The formation of penicillin V is assumed to be dependent on the extracellular POA concentration [102]:

$$q_{PenV} = k_{PenV,max} \cdot q_{Base} \cdot \frac{C_{POA}}{(C_{POA} + K_{S,POA})} \tag{4.4}$$

Side-Product Formation

A multitude of side-products can be accumulated in cultivations of *P. chrysogenum*. They are generated by enzymatic and non-enzymatic catalyzed reactions. The reactions and their kinetics are stated below.
The cyclic lactam OPC is formed from AAA by cyclization. It was indicated by Pissara *et al.* [188] that the formation is dependent on the isopenicillin N amidohydrolase activity. Additionally, it was mentioned that the reaction is not spontaneous, because formation rates were high. The current status in literature suggests a formation path either from aminoadipate or as a by-product of 6APA formation (for a discussion see Henriksen *et al.*)[101]. Formation of OPC is modeled as fraction of the substrate flux into the products:

$$q_{OPC=k_{OPC}} \cdot q_{Base} \tag{4.5}$$

Another side-product IPN is one product of ACV besides penicillin V and obeys first order kinetics with respect to oxygen [14]. Throughout the performed cultivations, dissolved oxygen concentration was kept constant high. Thus, the kinetic is simplified to:

$$q_{IPN} = k_{IPN} \cdot q_{Base} \cdot C_{O_2} \approx k`_{IPN} \cdot q_{Base} \tag{4.6}$$

The penicillin nucleus 6APA is formed by action of AT by releasing AAA. This reaction was simplified compared to Pissara *et al.* (who assumed Michaelis Menten kinetics) and approximated by first order kinetics:

$$q_{IPN\rightarrow 6APA} = k_{IPN\rightarrow 6APA} \cdot C_{BM} \cdot C_{IPN}$$

Furthermore, 6APA can also be formed from penicillin V by penicillin amidase [188]:

$$q_{PenV\rightarrow 6APA} = k_{PenV\rightarrow 6APA,max} \cdot C_{BM} \cdot \frac{C_{PenV}}{C_{PenV} + K_{M,6APA}} \tag{4.7}$$

The side-product 8HPA reacts spontaneously to 6APA: The reaction is first order with respect to the concentration of carbon dioxide and 6APA concentration [99]. Besides, POA is released by this reaction. As the concentration of carbon dioxide is constant throughout cultivations, the equation is simplified to:

$$q_{8HPA} = k_{8HPA} \cdot C_{CO_2} \cdot C_{6APA} \approx k'_{8HPA} \cdot C_{6APA} \tag{4.8}$$

Degradation of penicillin to PIO is spontaneous and depends on the phosphate concentrations [43]. Phosphate was added at the beginning of the cultivation, the equation was simplified:

$$q_{PIO} = k_{PIO} \cdot C_{PO_4} \cdot C_{PenV} \approx k'_{PIO} \cdot C_{PenV} \tag{4.9}$$

PIO is further degraded to PIA by decarboxylation [42]:

$$q_{PIA} = k_{PIA} \cdot C_{PIO} \tag{4.10}$$

Note that the last three equations do not contain dependency on the biomass (C_{BM}), because they are not catalyzed by the action of any enzyme (biomass component). It is a speciality of the penicillin pathway that a number of sideproducts are formed by nonenzymatic catalyzed reactions.
Furthermore, according to Desphande *et al.* PIA can be degraded to PeA and PeNH2 in an acidic environment. PeA can be further degraded to PeAl [55]. Of these three side-products (PeNH2, PeA and PIA), only PeNH2 was measurable but it was not detected. Additionally, sinks for PIA and PIO were included to model not-accounted side-products: $q_{PIA,sink} = k_{sink} \cdot C_{PIA}$ and $q_{PIO,sink} = k_{sink} \cdot C_{PIO}$, see Deshpande *et al.* for further description [55].

Products Formed by Hydroxylation

The hydroxylation products are formed by enzymatic activity. The activity of a monooxygenase was proposed by Emri *et al.* [71]. For the side-chain precursor phenoxyacetic acid, which was used in this work, elucidation of the complete degradation pathways is missing. Therefore, the degradation was assumed to be driven by enzymatic reaction steps. Thus, the following Michaelis Menten kinetics is assumed:

$$q_{HOPOA} = k_{HOPOA,max} \cdot C_{BM} \cdot \frac{C_{POA}}{C_{POA} + K_{M,HOPOA}} \tag{4.11}$$

Further degradation of POA and usage as carbon source was not incorporated into the model, because no evidence was found for this hypothesis in the data-sets at hand.

The formation of HOPenV can occur either by hydroxylation or by incorporation of hydroxylated phenoxyacetate into penicillin. Emri *et al.* and Chang *et al.* demonstrated that experimentally [70, 35]. The formation of HOPOA was modeled by

$$q_{HOPenV,1} = k_{HOPenV,1,max} \cdot C_{BM} \cdot \frac{C_{PenV}}{C_{PenV} + K_{M,HOPOA}} \tag{4.12}$$

$$q_{HOPenV,2} = k_{HOPenV,2,max} \cdot q_{Base} \cdot C_{HOPOA} \tag{4.13}$$

In the model it is assumed that penicillin V possesses the same affinity constant as POA. The resulting HOPenV is potentially subject of degradation via PIO to the other compounds (PIA, PeAl, ...). An efflux from HOPenV was used to model this sink:

$$q_{HOPenV,sink} = k_{HOPenV,sink,max} \cdot \frac{C_{HOPenV}}{(C_{HOPenV} + K_{M,sink})} \tag{4.14}$$

Mass Balances

To model the whole bio-process, mass balances are required for each compound. All reactions stated above were calculated based on concentrations in gram per kilogram of reactor broth. Thus conversion from mass to molar concentrations need to be applied for mass balances. This is done by normalizing the reactions by the molar masses $\mathbf{M}$ of the separate compounds:

$$\frac{dC_{BM}}{dt} = r_{BM} = \mu C_{BM} \tag{4.15}$$

$$\frac{dC_{Glc}}{dt} = r_S = -q_{Glc} \tag{4.16}$$

$$\frac{dC_{PenV}}{dt} = r_{PenV} = q_{PenV} - q_{HOPenV,1} \cdot \frac{\mathbf{M}_{PenV}}{\mathbf{M}_{HOPenV}} - q_{PenV \to 6APA} \cdot \frac{\mathbf{M}_{PenV}}{\mathbf{M}_{6APA}} - q_{PIO} \cdot \frac{\mathbf{M}_{PenV}}{\mathbf{M}_{PIO}} \tag{4.17}$$

$$\frac{dC_{OPC}}{dt} = r_{OPC} = q_{OPC} \tag{4.18}$$

$$\frac{dC_{6APA}}{dt} = r_{6APA} = q_{IPN \to 6APA} + q_{PenV \to 6APA} - q_{8HPA} \cdot \frac{\mathbf{M}_{6APA}}{\mathbf{M}_{8HPA}} \tag{4.19}$$

$$\frac{dC_{8HPA}}{dt} = r_{8HPA} = q_{8HPA} \tag{4.20}$$

$$\frac{dC_{POA}}{dt} = r_{POA} = q_{PenV \to 6APA} \cdot \frac{\mathbf{M}_{POA}}{\mathbf{M}_{6APA}} - q_{PenV} \cdot \frac{\mathbf{M}_{POA}}{\mathbf{M}_{PenV}} - q_{HOPOA} \cdot \frac{\mathbf{M}_{POA}}{\mathbf{M}_{HOPOA}} \tag{4.21}$$

$$\frac{dC_{HOPOA}}{dt} = r_{HOPOA} = q_{HOPOA} - q_{HOPenV,2} \cdot \frac{\mathbf{M}_{HOPOA}}{\mathbf{M}_{HOPenV}} \tag{4.22}$$

$$\frac{dC_{PIO}}{dt} = r_{PIO} = q_{PIO} - q_{PIA} \cdot \frac{\mathbf{M}_{PIO}}{\mathbf{M}_{PIA}} - q_{PIO,sink} \tag{4.23}$$

$$\frac{dC_{PIA}}{dt} = r_{PIA} = q_{PIA} - q_{PenNH2} \cdot \frac{\mathbf{M}_{PIA}}{\mathbf{M}_{PenNH2}} \tag{4.24}$$

$$\frac{dC_{HOPenV}}{dt} = r_{HOPenV} = q_{HOPenV,1} + q_{HOPenV,2} - q_{HOPenV,sink} \tag{4.25}$$

$$\frac{dC_{IPN}}{dt} = r_{IPN} = q_{IPN} - q_{IPN \to 6APA} \cdot \frac{\mathbf{M}_{IPN}}{\mathbf{M}_{6APA}} \tag{4.26}$$

Carbon dioxide in off-gas was modeled by balancing carbon flux for all substances. Here, the i separate rates were normalized by a factor ν_i describing the carbon fraction in the metabolite of the respective reaction rate q_i.

$$r_{CO_2} = \sum_i \nu_{i,educt} \cdot q_{i,educt} - \sum_i \nu_{i,product} \cdot q_{i,product} \tag{4.27}$$

These rate expressions can be inserted into bioreactor‘s mass balance equations to calculate concentration data for each compound.
Materials and methods of the conducted experiments and simulation studies can be found in the appendix B.1.

4.1.3 Results of Rate Estimation

In this section, the results for (kinetic) model-based rate estimation for chemostat and fed-batch are shown. Concentration data for both cultivations in chemostat and fed-batch were available. Two chemostats were conducted with a dilution rate of $0.03\,\mathrm{h}^{-1}$ (chemostat 1 and 2) and one with $0.05\,\mathrm{h}^{-1}$ (chemostat 3). The fed-batch process was executed in duplicate using the sensor bioreactor concept (fed-batch 1 and 2).
The model described beforehand was constructed in Dymola. Concentration measurements were introduced and SSmGO toolbox was used for global optimization [65]. The parameters (rate constants, affinity constants, ...) were constrained to biological meaningful limits, see appendix B.2. The simulated concentration data were fitted to the concentrations measurements of the cultivations. From the best fit model, estimates for reaction rate r were obtained (see appendix B.3). Here only the specific substrate uptake and product formation rates are shown and discussed, because they will be used later for the ^{13}C-MFA.

Chemostat

For chemostat experiments a batch phase of 20 h was conducted. Afterward the measurements were used to fit the constructed kinetic model. Measured concentration of PIA and PIO acid were added to penicillin V as they could not be described by the model.
The best obtained fit of the model to the experimental data for chemostat 2 at a dilution rate of $0.03\,\mathrm{h}^{-1}$ can be seen in figure 4.4. The measured cell dry weight was changing between 3.2 and $4.0\,\mathrm{g\,kg}^{-1}$ during chemostat phase. Phenoxyacetate concentration were at ca. $1.3\,\mathrm{g\,kg}^{-1}$ throughout the cultivation process. Glucose concentrations were below detection limit. The measured penicillin V concentrations were ranging from 0.2 to $0.3\,\mathrm{g\,kg}^{-1}$ and showed fluctuations. The simulated data could not account for these fluctuations in cell dry weight and penicillin V concentration. Potentially, they were caused by effects of changing from batch to chemostat process.

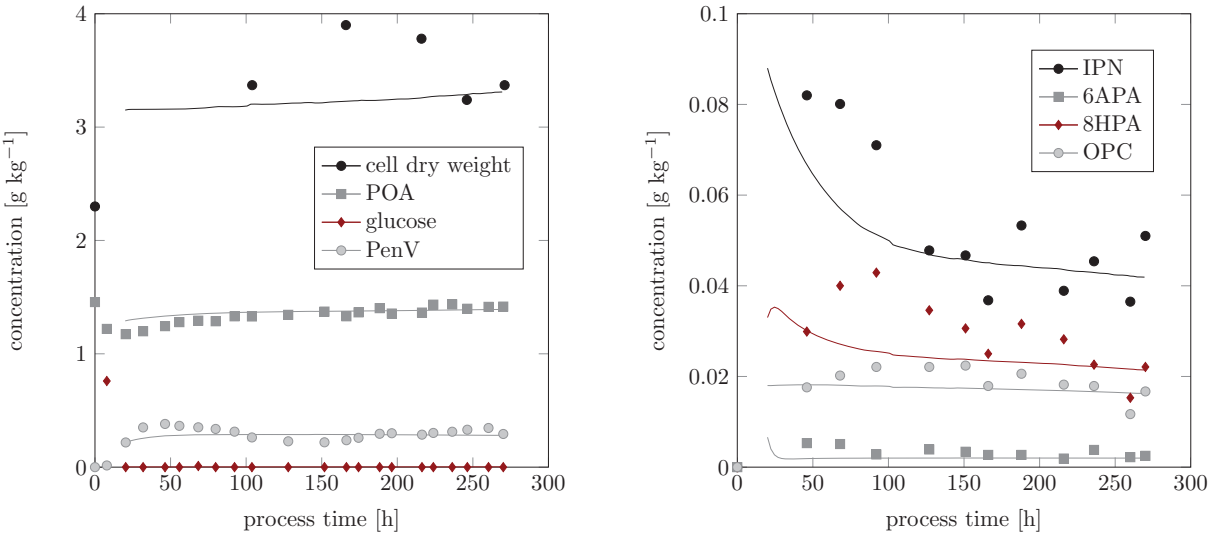


figure 4.4: Results of process simulation for chemostat 2 with a growth rate of 0.03 h^{-1}. The marker indicate measured concentrations and lines represent simulation results. Feed was started at 20 h cultivation time.

Concentration measurements of OPC, 6APA, IPN and 8HPA are in good agreement with the simulated data. For those side-products experimental and simulated data showed a decrease from high concentrations at the beginning of the chemostat (at 20 hours) to low concentrations till the end of the process.
The yield coefficient for biomass was calculated to 0.41 g_{CDW} g^{-1}. The specific growth rate of the strain was 0.031 h^{-1}.
The resulting substrate consumption and product formation rates can be found in table 4.1. Chemostat 1 and 2 were conducted with a growth rates of 0.03 h^{-1} and their biomass and glucose uptake rates were comparable. Penicillin V production was 35 % lower in chemostat 1 compared to chemostat 2. The simulation results indicates for chemostat 1 five times higher net production of IPN, although it was not measured for this experiment. OPC and 8HPA were formed as major side-products. In the third chemostat with higher growth rate of 0.05 h^{-1}, the specific penicillin production was strongly reduced. Side-product formation was similar compared to the chemostat with a growth rate of 0.03 h^{-1}, although higher production of 8HPA was found.

Fed-Batch

The fed-batch data (see figure 4.5) showed flat curve progression in dry cell weight with 12-14 g kg^{-1}, and 1.5-2.3 g kg^{-1} POA. Penicillin production was high and resulted in 4 g kg^{-1} penicillin V. This is surprising, because low cell dry weight was present for the production of penicillin. The simulated products showed qualitatively a good fit with the measured concentrations. The side product PIO and PIA indicate a decrease in formation towards the end of cultivation. In contrast, the estimated concentration profiles, based on first order kinetics for these products, were further increasing. Simulated IPN data show an increase till the end of cultivation, where a sharp drop of concentration was detected experimentally. Introduction of backward reaction from 6APA to penicillin V by AT showed no improvement. OPC and HOPOA were fitted with good agreement to the data.

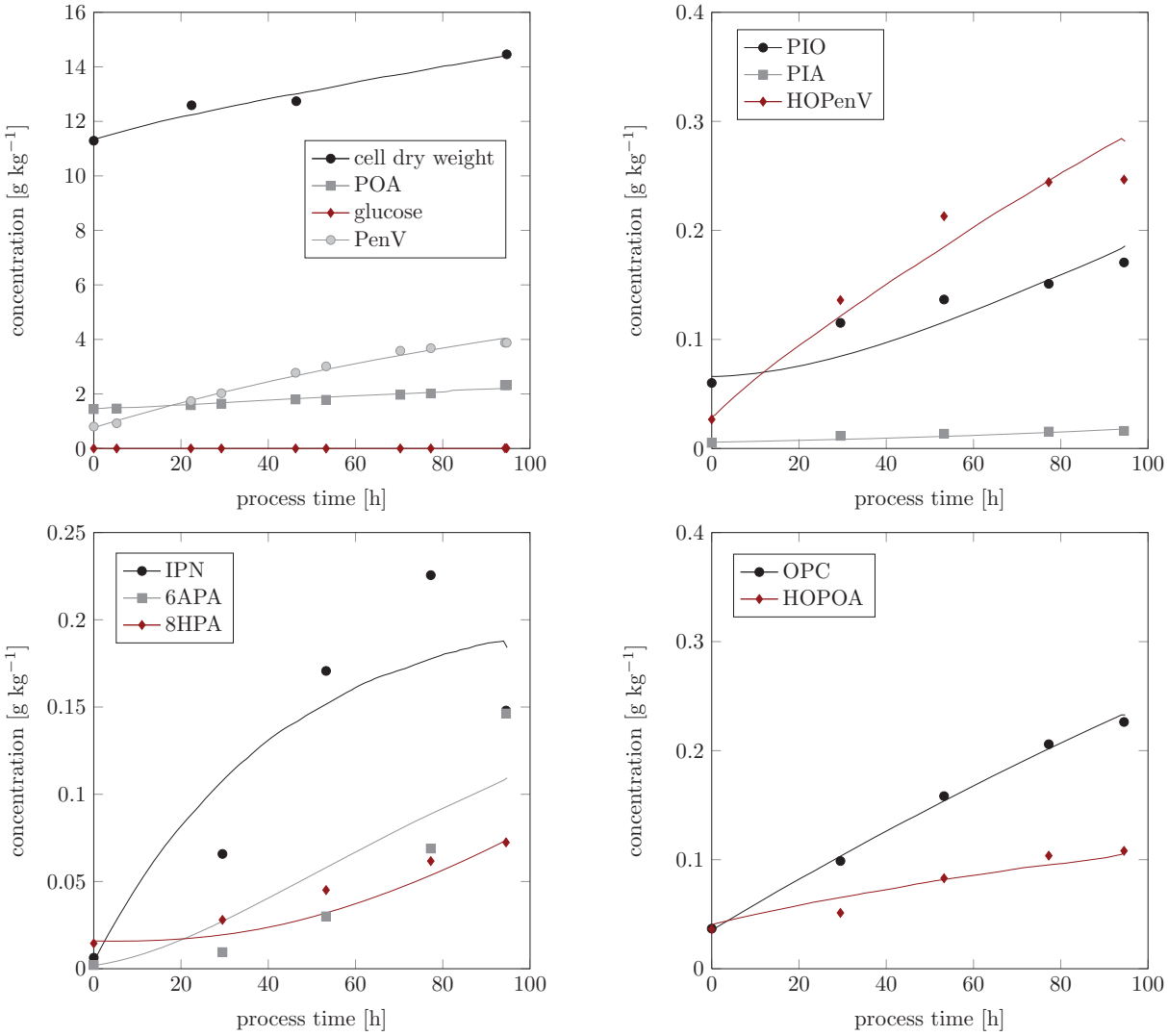


figure 4.5: Results of process simulation for fed-batch 1 with growth rate of 0.008 h^{-1}. The marker indicate measured concentrations and the lines represent simulation results.

table 4.1: Specific substrate uptake/product formation rates for the chemostat process in [mmol $g_{CDW}^{-1} h^{-1}$]. Contribution of HOPenV and HOPOA formation to the POA consumption rate was corrected, because these metabolites were not detected in the cultivation. IPN production rate of chemostat 1 was neglected for the conduction of ^{13}C-MFA. n.d. = not detected.

	Chemostat 1	Chemostat 2	Chemostat 3
Labeled substrate [%] 1-^{13}C/U-^{13}C/^{12}C	60/20/20	70/30/0	70/30/0
Total Process duration	262 h	270 h	264 h
v_{Glc}	0.389 ±0.015	0.420 ±0.018	0.517 ±0.028
v_{BM}	0.030 ±0.002	0.031 ±0.002	0.045 ±0.003
v_{PenV}	5.362 ±0.061·10^{-3}	8.211 ±0.144·10^{-3}	2.966 ±0.184·10^{-3}
v_{POA}	5.362 ±0.061·10^{-3}	8.400 ±0.146·10^{-3}	2.966 ±0.184·10^{-3}
v_{IPN}	5.108 ±0.003·10^{-3}(n.d.)	1.178 ±0.000·10^{-3}	1.502 ±0.078·10^{-3}
v_{6OPC}	1.189 ±0.001·10^{-3}	1.148 ±0.000·10^{-3}	1.186 ±0.061·10^{-3}
v_{6APA}	0.014 ±0.001·10^{-3}	0.090 ±0.002·10^{-3}	0.127 ±0.007·10^{-3}
v_{8HPA}	0.844 ±0.107·10^{-3}	0.846 ±0.009·10^{-3}	1.116 ±0.030·10^{-3}

The estimated specific production and consumption rates are shown in table 4.2. The specific substrate uptake rates were reduced in fed-batch process compared to the chemostat process (0.251-0.263 mmol $g_{CDW}^{-1} h^{-1}$ versus 0.389-0.517 mmol $g_{CDW}^{-1} h^{-1}$). Biomass formation was reduced to low specific growth rates (0.007-0.008 h^{-1}), although specific production rates of penicillin were increased in fed-batch (9.8-10.9·10^{-3} mmol $g_{CDW}^{-1} h^{-1}$) compared to the chemostat processes (3.0-8.2·10^{-3} mmol $g_{CDW}^{-1} h^{-1}$).
Besides, side product formation in fed-batch was revealed to be more complex compared to chemostats. IPN was not formed as major side-product, but OPC and HOPOA. Even PIA and PIO were detected in significant amounts and were well described by the model. Whereat PIA showed lowest specific production rate of all side-products.
The biomass yield was significantly reduced compared to chemostat data with 0.17 $g_{CDW} g^{-1}$.

4.1.4 Discussion

The concentration data in chemostat and fed-batch process were fitted with an unified kinetic model. Although fluctuations in the concentrations were measured, most data could be represented by the constructed model. Some simulated side-products (e.g. PIA, PIO, PeNH2, and IPN) were deviating with respect to the measurements. Potentially, these deviations are caused by degradation effects in analytics or after sample withdrawal. Specific product formation and substrate uptake rates were computed. Their standard deviations were calculated based on the simulated rates and are shown in table 4.1 and 4.2. Bootstrap methods for error propagation were not accessible because parameter estimation took approximately 45 minutes for the model and experimental uncertainties of the measurements were not known. It was shown that the side-product formation was increased in fed-batch cultivations compared to chemostat. Importantly, *P. chrysogenum* strain BCB1 can produce high amounts of penicillin V in fed-batch processes with a reduced growth rate. The penicillin formation was reduced at elevated dilution rate in the chemostats; this is an known effect from literature [242]. The estimated rates for the ^{13}C-MFA were calculated and the standard de-

table 4.2: Specific substrate uptake and product formation rates for the fed-batch process in [mmol g_{CDW}^{-1} h^{-1}].

	Fed Batch 1	Fed Batch 2
Labeled substrate [%] 1-^{13}C/U-^{13}C/^{12}C	60/20/20	60/20/20
Total Process duration	94.5 h	94.5 h
v_{Glc}	0.251 ±0.014	0.263 ±0.021
v_{BM}	7.611 ±0.701·10^{-3}	6.738 ±0.824·10^{-3}
v_{PenV}	10.925 ±1.021·10^{-3}	9.823 ±0.793·10^{-3}
v_{POA}	12.616 ±0.971·10^{-3}	11.950 ± 0.780·10^{-3}
v_{IPN}	0.563 ±0.209·10^{-3}	0.711 ±0.226·10^{-3}
v_{PIO}	0.366 ±0.105·10^{-3}	0.569 ±0.183·10^{-3}
v_{PIA}	0.040 ±0.010·10^{-3}	0.000 ±0.000·10^{-3}
v_{6OPC}	1.526 ±0.067·10^{-3}	1.415 ±0.100·10^{-3}
v_{6APA}	0.509 ±0.090·10^{-3}	0.121 ±0.0018·10^{-3}
v_{8HPA}	0.225 ±0.125·10^{-3}	0.588 ±0.269·10^{-3}
v_{HOPenV}	0.753 ±0.054·10^{-3}	0.484 ±0.489·10^{-3}
v_{HOPOA}	0.532 ±0.050·10^{-3}	1.034 ±0.562·10^{-3}

viations obtained for the rate measurements in fed-batch process were higher compared to the chemostats.

Next, the labeling measurements are preprocessed for ^{13}C-MFA.

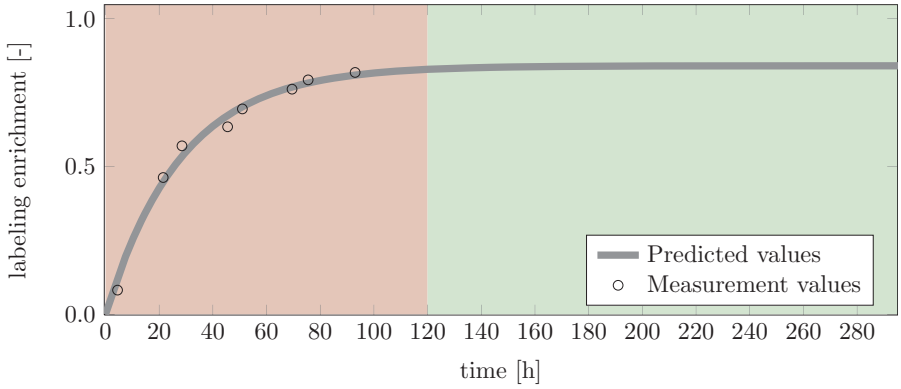


figure 4.6: Transient labeling enrichment data and estimation of steady-state value. Only transient data ( ) have been measured. Steady-state data ( ) were not measured.

4.2 Extrapolation of Measurement Data

In chemostat and fed-bath cultivations samples were drawn throughout cultivations and analyzed by LC-MS/MS (a mass spectrometric measurement device). By this, labeling measurements are provided for central carbon metabolites of glycolysis, citrate cycle, pentose phosphate pathway, amino acids, and their precursors. The measurement data show time resolved labeling enrichment. Additionally, the number of samples and time points were varying throughout the cultivations (see table B.1).
For stationary ^{13}C-MFA only steady-state labeling measurements are used. In this chapter the extrapolation from transient to stationary labeling data is explained and the results are discussed.

4.2.1 Extrapolation

A set of measurements is given by isotopic labeling enrichment data derived from isotopic labeling experiments in chemostat and fed-batch. The measurement values were determined by a mass-spectrometric device at time points distributed evenly in the experimental time frame. To obtain measurement values at isotopic-steady-state a labeling experiment has to be performed until all the labeling patterns are in steady-state within the cell. For *P. chrysogenum* this may take in the order minutes or up to several hours [96, 284]. In figure 4.6 the transient labeling enrichment of a (hypothetical) mass trace of an MS measurement is shown.
It is assumed, based on previously conducted transient labeling experiments, that enrichment follows an exponential growth/decay law. This provides the opportunity for model-based extrapolation. Thus, nonlinear regression is used to extrapolate the data, as the kinetics of labeling enrichment are described by nonlinear functions. Confidence intervals for the extrapolated data are calculated to evaluate the error introduced by the extrapolation method and the measurements‘ error.
To illustrate the effect of sample points on the confidence region of labeling data, three artificial example data sets were generated. In the examples the turnover of the metabolites are changed from low to high values. In figure 4.7 confidence regions are given for the fitted curves, depending on the distribution of the measurements points. In the first case the quasi-steady-state is not reached (it is defined as $> 99\,\%$ of the final enrichment value), i.e. only transient labeling data points are available. The resulting confidence (red dotted line) regions

are large. Not enough data were provided to extrapolate with the given transient data to a well defined steady-state labeling measurement. In the other examples further data are provided in the steady-state region, thus the calculated confidence region provided covers smaller region and the extrapolated data are statistically well determined.

It was already mentioned that labeling enrichment data show the behavior of exponential (growth/decay) functions. It originates from the similarity to a dilution/wash-out kinetic. For each intracellular pool reached by labeling the abundance of labeled species will increase and the abundance of unlabeled species will decrease. This is explained by simultaneous growth and uptake of substrate from the environment, which will dilute each of the pools consecutively.

4.2.2 Regression Model

Next, the regression model is introduced which is used to fit each mass trace of the MS measurement separately. For simplicity, the process will be explained for MS-measurements but can be easily extended for MS/MS data-sets. The observed fragment of a metabolite possesses n_C carbon atoms. In the optimal case, $n_C + 1$ mass isotopomers can be measured per metabolite ($l = 0, 1, 2, 3 \dots n_C$). The first equation (shown below) can be used to fit the behavior of $m + 0$ measurements as they are decreasing. The second equation can be used to fit the enrichment of higher masses ($m + l$, $l > 0$) as they are increasing during an labeling experiment. Besides these assumptions about the system, the parameters θ_1 and θ_2 in the equations have to be positive. The model equation is shown in the next equation, where t denotes the time and y is the mass trace of an MS measurement:

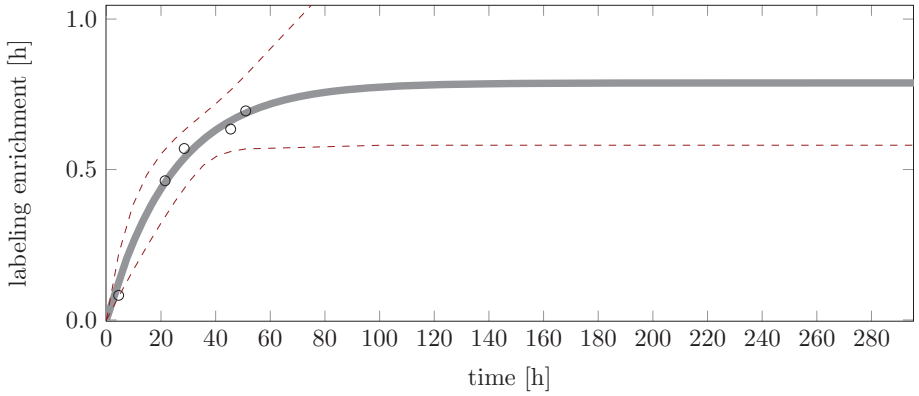
$$y_{m+l}(t, \theta_{1,l}, \theta_{2,l}) = \begin{cases} \theta_{1,l} + (1 - \theta_{1,l}) exp^{-\theta_{2,l} \cdot t} + c_{init,m+l} & \text{for} \quad l = 0 \\ \theta_{1,l} \cdot (1 - exp^{-\theta_{2,l} \cdot t}) + c_{init,m+l} & \text{for} \quad l = 1, 2, \dots n_C \end{cases} \tag{4.28}$$

The parameters (θ_1 and θ_2) are estimated and, in turn, the simulated measurement values $y_{m+l}(t)$ will provide a good estimate of the measurement data‘s course. The parameter θ_1 gives information about the steady-state labeling enrichment and θ_2 about the turnover of the metabolite pool. In contrast to these parameters, the constant $c_{init,m+l}$ accounts for natural labeling and is calculated for each mass trace of a metabolite. Calculation of natural labeling for MS and MS/MS mass traces will be explained in the next chapter 4.3.

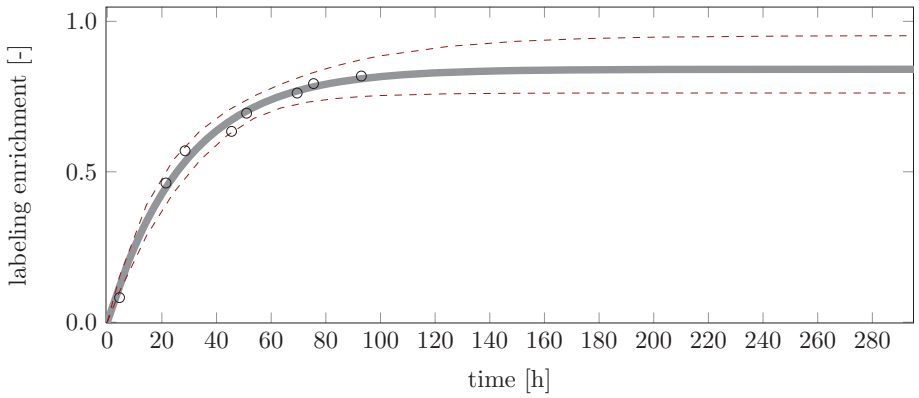
The regression model f is build using equation 4.28. The unknown model parameters (θ_1 and θ_2) are to be estimated from the values $y_{m+l}(t)$ at n measurement values at specific time points $t_1, t_2, \dots t_n$. The errors of the measurements $\epsilon(t)$ are assumed to be identically distributed normal random variables with zero mean and a finite variance σ^2, which is unknown and has to be estimated from the data. The regression model can be defined, and the general form is $f_{m+l}(t, \theta_{1,l}, \theta_{2,l}) = y_{m+l}(t, \theta_{1,l}, \theta_{2,l}) + \epsilon(t)$.

Residual Sum of Squares

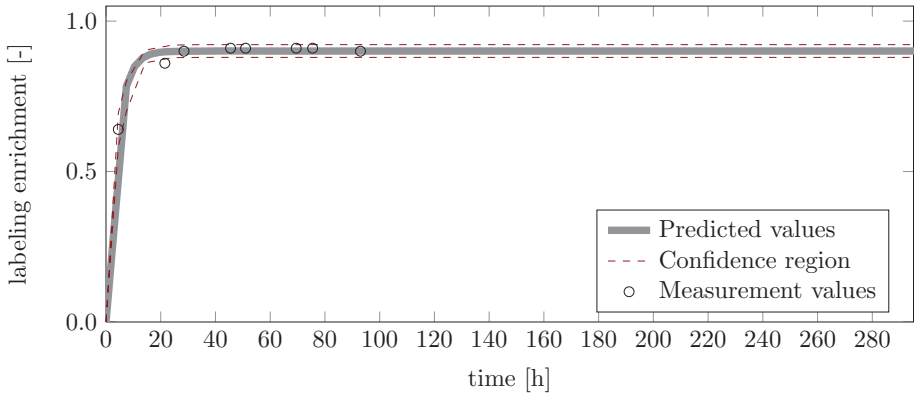
From the definition of the model, the residual sum of squares is defined (RSS). The RSS is used as a measure for discrepancy between model and data. A smaller residual corresponds



(a) Low turnover, measurement data far from steady-state



(b) Low turnover, measurement data close to steady-state



(c) High turnover, measurement data at steady-state

figure 4.7: Regression model used for estimating steady-state data point. Three examples are shown with labeling measurement (a) far away, (b) close and (c) at steady-state. Metabolite pool in (a) and (b) show low turnover and in (c) a high turnover.

to a better fit, thus the RSS is minimized in order to estimate the parameters $\hat{\theta}$:

$$\hat{\theta} = \arg\min_{\theta} \mathrm{RSS} = \sum_{i=1}^{n} (y_{t_i} - f_{m+l}(t_i, \theta_{1,l}, \theta_{2,l}))^2 \quad \text{subject to} \quad \theta_{s,lower} \leq \theta_s \leq \theta_{s,upper} \, (s = 1, 2) \tag{4.29}$$

The lower and upper limits $\theta_{s,lower}$ and $\theta_{s,upper}$ were set to the interval $[0, 1]$. The best fit results in optimal parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ with the RSS_{min}.

Estimation of Confidence Intervals

Here, the boundaries of a $100 \cdot (1 - \alpha)\,\%$ confidence interval and region are defined for θ and y, respectively. F is the F-distribution with a significance level of α. The F-distribution gives the ratio of χ^2 values of two independent χ^2-distributed random variables with p and $n - p$ degrees of freedom. The above given model y_{m+l} was used with $p = 2$ parameters and n measurements. Based on the principle of maximizing likelihood a maximal value for the RSS is defined [197]:

$$RSS_{max}(\hat{\theta}) = RSS_{min}(\hat{\theta}) \left[1 + \frac{p}{n - p} F(\alpha; p, n - p) \right] \tag{4.30}$$

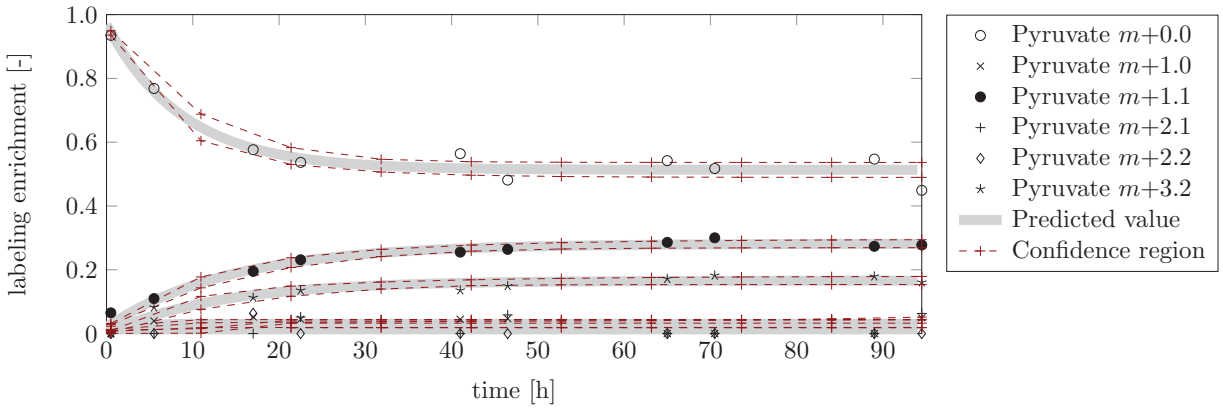
Minimizing/Maximizing the parameters θ within the boundary of RSS_{max} results in lower $\theta_{conf,lower}$ and upper $\theta_{conf,upper}$ confidence bounds. The intervals of θ_1 can be used to calculate the confidence band on the estimated steady-state confidence intervals (by using θ_1); they can be asymmetric. However, symmetric intervals are required for 13CFLUX2, these are calculated according to the following protocol for θ_1:

$$s \;=\; max(\theta_i - \theta_{conf,lower,i}, \theta_{conf,upper,i} - \theta_i) \tag{4.31}$$

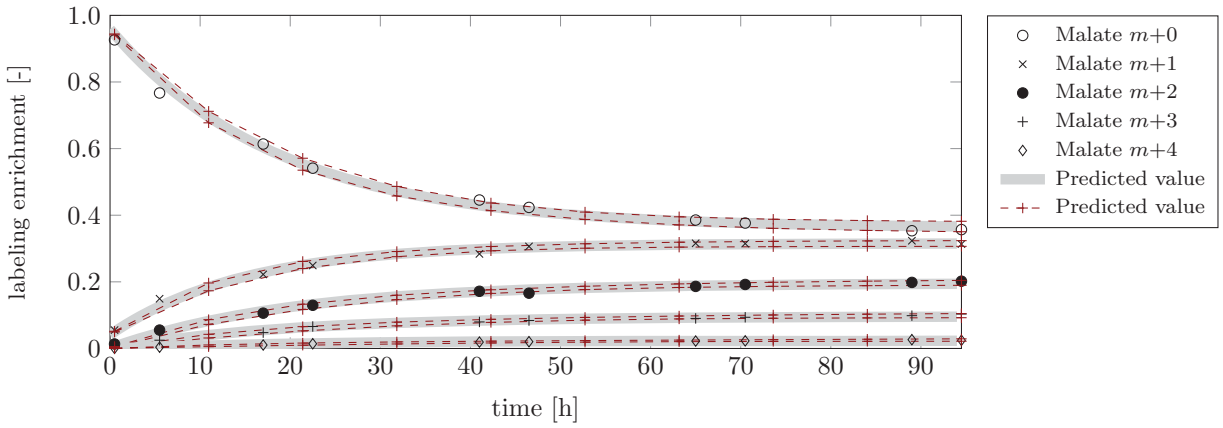
Model fits with bad fit quality were sorted out manually and are excluded from further analysis.

4.2.3 Results of Extrapolation

Samples are drawn time resolved from the bioreactor, the measurements for the fed-batch process are visualized in figure 4.8. The estimation of a confidence interval is dependent on the deviation of the sample points to the regression model. The more the data points deviate from the regression curve in the optimal fit, the more the confidence interval of the parameters will increase. This is also indicated in figure 4.8a by time resolved labeling data derived from fed-batch 2. Pyruvate mass traces are more variable. Thus, also the final confidence interval is higher for the m+0.0 trace compared to the m+0 trace of malate. As shown in the example, uncertainties in the measurements are directly propagated to the steady-state MS measurement values.
Nevertheless, beside using parameter $\hat{\theta}_1$ for the estimation of steady-state values of labeling data, $\hat{\theta}_2$ can be interpreted. It describes the speed of the labeling enrichment in the respective pool. This parameter is defined as the ratio of influx into this pool in relation to the pool-size at steady-state, this ratio is known as turnover [264, 180]. Using this parameter, direct estimation of fluxes was not possible because pool size measurements were not available.



(a) Pyruvate



(b) Malate

figure 4.8: Examples for extrapolation of pyruvate and malate with confidence region (- + -) for measurement data of 68.2 %. Each mass trace was fitted separately.

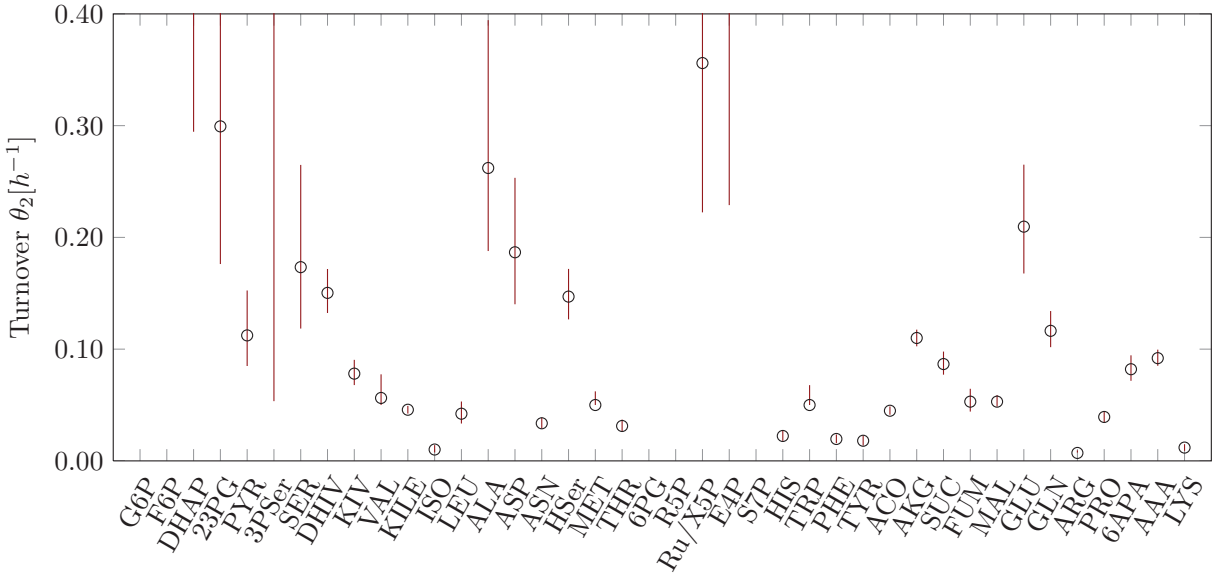


figure 4.9: Turnover estimate for $\hat{\theta}_2$ with lower and upper confidence interval (confidence of 68.2 %) determined for fed-batch 2 experiment. θ_2 was taken from fitted curves of $m{+}0$ mass traces. Estimate for $\hat{\theta}_2$ (○) and its confidence interval (——) computed by nonlinear regression.

Nevertheless, it is possible to evaluate the spread of the labeling through the reaction network by comparing the turnover rates (in figure 4.9). For this, sequences of reactions are compared with respect to their turnover to test consistency of the data, with respect to published reaction networks.

The spread of the labeling through pools of upper glycolysis and pentose phosphate pathway was too fast to be observable with the drawn samples. After dihydroxyacetonephosphate (DHAP), 2- and 3-phosphoglycerate (2/3PG) and finally pyruvate (PYR) were labeled. Serine is produced from its precursor 3-phosphoglycerate. Although the turnover rate of intermediate 3-phosphoserine for serine synthesis is weakly determined, we can see that the estimates for turnover show the same sequence of reactions. From pyruvate originates the pathway via dihydroxyisovalerate (DHIV) and ketoisovalerate (KIV) to valine, which shows perfect fit of the reaction sequence proposed. From ketoisovalerate, one step before valine synthesis, isoleucine is formed via ketoisoleucine (KILE) and leucine is formed from pyruvate in the mitochondria. Interestingly, aspartate is labeled faster compared to pyruvate. This can be explained by the compartmentalization. The cytosolic pool of pyruvate is fueling anaplerotic reactions, probably, via pyruvate carboxylase resulting in the formation of oxaloacetate. These pools are faster labeled in the cytosol. However, aspartate can be formed in the cytosol and in the mitochondria. As proposed by Kleijn, it is predominantly formed in the cytosol [121]. The data obtained underpin this finding as the pools assumed to be predominantly in the mitochondria are considerably later labeled, compared to the aspartate pool.

The labeling enrichment in glutamate and glutamine is faster than in their respective precursor α-ketoglutarate. This finding is at first contradictory. An explanation can be differing pool-sizes in the separate compartments. Besides, α-ketoglutarate can also be formed in the cytosol by NADPH dependent cytosolic isocitrate dehydrogenase and could, thus, be

faster labeled compared to mitochondrial α-ketoglutarate. The amino acids glutamate and glutamine are expected to be formed in the cytosol, according to Kleijn [121].
The main flow direction in the cell of the TCA cycle intermediates can also be seen in the data from α-ketoglutarate to the pool malate, although cis-aconitate (ACO) shows lower turnover rate.
To sum up, the turnover of metabolite pools estimated by nonlinear statistics correspond in general well with literature data. However, deviating examples were also found (i.e. ACO).

4.2.4 Nonlinear Regression Toolbox: Work-flow of Nonlinear Regression

All extrapolated results were computed by nonlinear statistics. Here, a short introduction to work-flows for nonlinear regression problems is given. There are basically two methods to perform a nonlinear regression: the first is founded on least squares method and the second is a bootstrap method.
The method using least squares can be applied in two ways. First, checks for non-linearity are performed according to local non-linearity measures of Bates and Watts [145, 146, 19, 197]. Thus the combination of model/data-set is tested for parameter effects non-linearity and intrinsic non-linearity. If the model behavior is close to linear (**I**), most of the quantities (covariance matrix, confidence intervals, etc.) can be derived directly by using linearized statistics [197]. These derivatives can be calculated based on finite difference or complex-step approximation [148].
If the model behavior is nonlinear (**II**), the confidence regions have to be derived by a iterative process using solvers for nonlinear problems. Confidence regions are estimated for the regression model according to Dolan *et al.* [58]. The bootstrap method can also provide the user with approximations for confidence regions of the parameters. Additionally it can be used to verify the estimates produced by least square method, because tests for non-linearity of the model are based on local measures [197, 216, 162, 5, 112].
The Nonlinear Regression Toolbox (NoReTo) was developed to extrapolate measurement data and implemented in MATLAB 2013a. It can be used to compare linearized and non-linear statistics. Both described work-flows (bootstrap and least square) were implemented including all techniques described above.

4.3 Correction for Natural Abundance

The measurement of isotopomers by mass spectrometry is of crucial importance to the outcome of a ^{13}C-MFA study. In ^{13}C-MFA the measured labeling state of the metabolite‘s carbon atoms are used. However, several of the elements found in metabolites contain a considerable fraction of natural isotopic labeling, which can alter the labeling state of the metabolite [22]. For ^{13}C-MFA the measurement need to be corrected for this natural labeling to obtain exclusively the carbon atom‘s labeling state.
Tools and algorithms for the correction of tandem MS measurements are still lacking, while several tools can be found for the correction of simple MS measurements [255]. Besides, the description of correction procedures are spread over wide parts of literature. Here, the complete correction procedure is described in formal notation.
In this section the method for correction is described and applied to the extrapolated MS/MS measurements.

4.3.1 Introduction

Frequently found elements in metabolites possessing naturally occurring isotopes (e.g. ^{2}H, ^{18}O, ^{13}C...) are given in table 4.3. High abundances of natural occurring isotopes are found in carbon and sulfur. Some sample preparation procedures for measurement techniques (e.g. GC-MS or GC-MS/MS) introduce additional natural labeling in the measured analyte by derivatization. Here large side-groups are attached by sylilation, e.g. reaction with N-(tert-butyldimethylsilyl)-N-methyl-trifluoroacetamide (MTBSTFA) or tert-butyldimethylsylil (tBDMS). These side groups increase volatility of the substances in GC separation [49, 189]. In particular introduced silicon (besides introduced carbon) contains high amounts of heavy isotopes (92.223 % ^{28}Si, 4.685 % ^{29}Si and 3.092 % ^{29}Si) and will thus alter the labeling pattern significantly [22].

table 4.3: Isotopic compositions of elements found in biologically relevant substances [22].

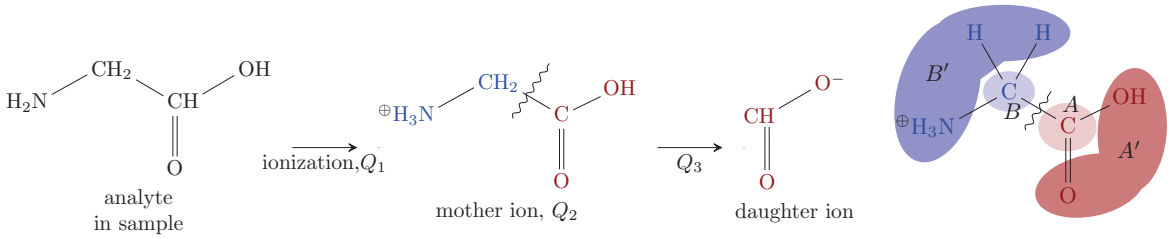
Element	Mass number	Mole fraction [%]
H	1	99.9885
	2	0.0115
C	12	98.93
	13	1.07
N	14	99.757
	15	0.364
O	16	99.757
	17	0.038
	18	0.205
P	31	100.0
S	32	94.99
	33	0.75
	34	4.25
	36	0.01

The measurements of metabolites containing non-carbon atoms are biased as the naturally occurring isotopes of these atoms will alter the mass of the measured metabolite. Thus, the natural abundance of the non-carbon atoms need to be corrected in order to receive the measurements representing exclusively the skeletal carbon atom‘s labeling state. The corrected measurements for the carbon atom skeleton can be used to perform ^{13}C-MFA. The described method can be applied for correction of MS/MS measurements, correction of MS measurement was described by Wahl *et al.* and several tools are available [255, 160, 279]. In this chapter, first, a work-flow is given for the correction of MS/MS data. The complete work-flow is performed, examples are given, and an algorithm is developed to correct MS/MS data. The work-flow is, finally, applied to the extrapolated MS/MS measurements of chemostat

and fed-batch.

4.3.2 Principle of MS/MS Measurement

In the mass spectrometric device, the analyte molecule is first ionized. The emerging positively/negatively charged ions are then separated by their mass to charge ratio (m/z). The first quadrupole (Q_1) selects for a specific ion by selecting for a m/z ratio, like shown in figure 4.10a. This ion, called *mother ion*, does contain a fraction or all atoms of the original molecule. In a collision cell (Q_2) fragment ions of the mother ion are produced, which are separated by a second quadrupole (Q_3) to yield one specific daughter ion, which is detected. Upon detection the masses of mother and daughter ion can be different, depending on which m/z ratio was selected for in the two quadrupoles [153].



(a) Fragmentation in MS/MS (b) Notation of atoms

figure 4.10: (a) Hypothetical mother and daughter ion of L-glycine derived by MS/MS technique. Atoms contained in the final fragment are marked in red. (b) Notation of A, A', B, B' for glycine mother ion, adapted from Miebach *et al.* [158] .

An example for the measurements in an MS/MS (tandem MS) device is given in figure 4.11. A metabolite X with three carbon atoms is analyzed. The first quadrupole Q_1 selects for the mass of the mother ion. Depending on the selected mass for the mother ion (m+0, m+1, ...), different isotopomers are chosen. For example: if it is selected for the m+1 of the mother ion, three possible labeling states of the metabolite are chosen (X#100+X#010+X#001). After fragmentation in Q_2 it is selected for the mass of the daughter ion (which contains here one carbon atom). Depending on the selected mass in Q_3, a subset of the isotopomers of the metabolite are measured. For example m+1 of the mother ion is further separated into the fractions m+1.0 and m+1.1, where the second number indicates the labeling in the daughter ion. The fractions m+1.0 and m+1.1 correspond to the labeling pattern X#100+X#010 and X#001, respectively.

4.3.3 Work-flow of Correction

Before starting with the correction it needs to be clarified for which atoms a correction has to be done. The analyte, shown in figure 4.10b, is composed of the detected part of the mother ion (red) and the non-detected part (blue). Each contains skeletal atoms (light blue and red) and non-skeletal atoms (dark blue and red). The term *skeletal atoms* is referring to the atoms of interest, in case of this work: the carbon backbone of the metabolite. Henceforth, these atoms are only called skeletal atoms. Note, that eventually introduced side groups (sylilation) for separation in GC or LC are counted as *non-skeletal atoms*.

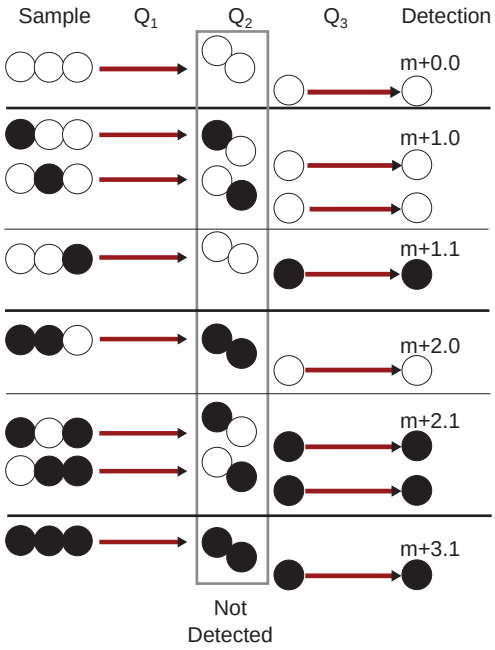


figure 4.11: Tandem MS measurement principle for a metabolite X containing three carbon atoms. The mother ion is selected in the first quadrupole (Q_1). In the collision cell (Q_2) the third carbon atom is split off and in the second quadrupole (Q_3) it selected for it.

The work-flow described is needed for MS/MS measurement in which, upon fragmentation, a different subset of skeletal atoms is found in mother and daughter ion.

The work-flow described below is *not* needed, if (I) no skeletal atoms are found within the daughter ion (this is usually the case for measurements of sugar-phosphates in this thesis). As the labeling of the daughter is only influenced by natural labeling, the underlying measurement simplifies to an MS measurement correction. The measurement values need to be corrected exclusively for all non skeletal atoms in the mother ion. Thus, the mother ions chemical formula needs to be known. (II) If the daughter ion contains all skeletal atoms, the correction is done exclusively for the daughter ion. Thus, only the chemical formula of the daughter ion is needed for correction. In both cases a standard MS correction can be performed.

The steps of the work-flow are shown in figure 4.12. To remove the natural labeling of an MS/MS measurement with different subset of skeletal atoms in mother and daughter ion, the elemental composition of mother and daughter ions needs to be known. So-called mass shift matrices are built to describe the natural labeling pattern in the daughter ion and the non-detected part of the mother ion. The correction is calculated in a constrained optimization problem, to prevent negative mass traces after correction (caused by stochastic errors in the measurements): initially, a labeling pattern of the metabolites skeletal atoms is guessed. The natural labeling of all non skeletal atoms is added to the estimated skeletal labeling pattern to obtain the labeling state of the metabolite. The obtained labeling pattern is compared to the measured labeling pattern and the difference is minimized using an optimization procedure. Finally, a statistical reconciliation is performed to obtain error transduction from the mass traces to the obtained skeletal atoms labeling measurements, which are the labeling state of carbon atoms. This labeling state of the carbon atoms is of interest for the study.

4.3.4 How to correct natural abundance ?

The correction of the natural labeling can be accomplished by building the mass shift matrix, called C_{corr}. This matrix contains probabilities for the detected and non-detected part of the mother ion at each single entry. To be more precise, the entries corresponds to probability for a specific labeling pattern of the mother and daughter ion non skeletal atoms to appear. This matrix is used to correct the labeling state of the skeletal carbon atoms y_C to obtain the full measurable labeling state y_M of the metabolite.

$$y_M = C_{corr} \cdot y_C \tag{4.32}$$

It is obvious that this is a linear problem. In the following section, the first formal notation is established to caclulate these mass shift matrices. For this, several literature sources are used and adapted to calculate correction for tandem MS-measurements. As no publicly available tool is known to the author, a pseudocode is presented for the generation of the mass shift matrix.

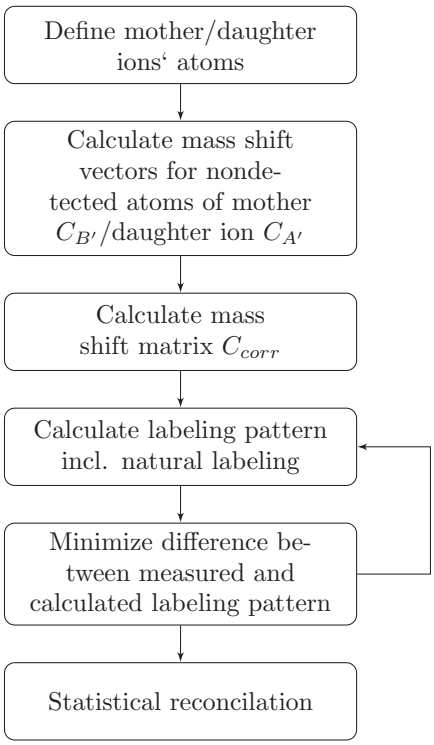


figure 4.12: Work-flow for natural labeling correction.

4.3.5 Formal Notation

To indicate the difference in the masses of the ions, a formal notation is established. The labeling of the mother and daughter ion is indicated by two numbers, separated by a colon $ab : a$, the first variable ab indicates the number of labeled skeletal atoms in the mother ion, the second variable a the number of labeled atoms in the daughter ion $ab = 0, 1, 2...s$ and the number of daughter ion skeletal atoms s' with $s \leq s'$, $with\, s, s' \epsilon\, \mathbb{N}^{\star}$ as the daughter ion is smaller or equal in size compared to the mother ion. We can define now a labeling $y_C^{ab:a}$. All labeling states are not allowed for a as the daughter ion can contain a subset of the mother ions‘ skeletal atoms. Thus

$$a = max(0, ab - (s - s')),\, max(0, ab - (s - s')) + 1,\, \ldots min(ab, s'). \tag{4.33}$$

$y_C^{2:1}$ indicates that the mother ion is labeled twice at any skeletal atom and the one indicates that the daughter ion is labeled only at one position (because not all skeletal carbon atom of the mother ion are contained in the daughter ion).
A measurement vector

$$y_C = [y_C^{0:0}, y_C^{1:0}, y_C^{1:1}, y_C^{2:0}, \ldots, y_C^{s:s'}]^{\intercal} \tag{4.34}$$

of the skeleton atoms (in our case skeletal carbon atoms) with length C_y is defined. It is composed of the labeling pattern for a given pair of mother and daughter ion and describes only the skeletal atoms of interest. Introduced carbon atoms by modifications (e.g. sylilation in GC-MS/MS) are not accounted for here.

Example 1 *For the above given example of L-glycine (see figure 4.10), the vector of C_y can be build using equation 4.33:*

$$y_C = [y_C^{0:0}, y_C^{1:0}, y_C^{1:1}, y_C^{2:1}]^\intercal$$

Thus, only four different labeling patterns are possible for the MS/MS measurement of skeletal carbon atoms of glycine, because only two labeling states (^{12}C or ^{13}C) exist for carbon.

Nevertheless, in the measured metabolites, more labeling patterns than the pure skeletal carbon atoms of interest are possible; there are also the non skeletal atoms contributing to the labeling state (like H, O, N, P, ...). Thus, there is a second "true" mass vector for the measurement incorporating all possible labeling state*s*

$$y_M = [y_M^{0:0}, y_M^{1:0}, y_M^{1:1}, y_M^{2:0}, \dots y_M^{\mathbf{t}:\mathbf{t}'}]^\intercal \tag{4.35}$$

and $y_M^{aa'bb':aa'}$ with $aa'bb' = 0, 1, 2...\mathbf{t}$ and $aa' = 0, 1, 2, ...\mathbf{t}'$ and as more than one labeling state can be found $\mathbf{t} \geq s$, $\mathbf{t}' \geq s'$ with $\mathbf{t}, \mathbf{t}' \epsilon \mathbb{N}^\star$. The vector has the length M_y. The maximal possible labeling states of skeletal carbon atoms of interest (s, s') and of the full set of all atoms in the fragments $(\mathbf{t}, \mathbf{t}')$ can be calculated by setting up an vector d of the maximal number of isotopes $N_{I,isotope}$ known for element $I \epsilon \{C, N, H, O, P, \dots\}$ present in the metabolite

$$d = [\mathbf{N}_{C,isotope} - 1, \mathbf{N}_{N,isotope} - 1, \mathbf{N}_{H,isotope} - 1, \mathbf{N}_{O,isotope} - 1, ...]^\intercal. \tag{4.36}$$

The values for $\mathbf{N}_{I,isotope}$ can be deduced from table 4.3 or taken from Berglund and Wieser [22], e.g. for oxygen $\mathbf{N}_{O,isotope} = 3$ and for carbon $\mathbf{N}_{C,isotope} = 2$. Multiplication of d with the a vector containing the number of atoms $\mathbf{N}_{atoms}$ of each element in the observed fragment gives the maximal labeling state of the ion. Thus the detected fragment (atoms shown in red in figure 4.10) is composed of

$$A = [\mathbf{N}_{C,atoms}, \mathbf{N}_{N,atoms}, \mathbf{N}_{H,atoms}, \mathbf{N}_{O,atoms}, ...] \tag{4.37}$$

skeletal atoms and

$$A' = [\mathbf{N}'_{C,atoms}, \mathbf{N}'_{N,atoms}, \mathbf{N}'_{H,atoms}, \mathbf{N}'_{O,atoms}, ...] \tag{4.38}$$

non skeletal atoms. The same holds true for the non-detected fragment (atoms shown in blue), which is composed of B skeleton atoms and B' non skeleton atoms. Finally, the absolute size of the measurement vectors can be calculated:

$$\mathbf{t} = A \cdot d + B \cdot d + A' \cdot d + B' \cdot d, \quad \mathbf{t}' = A \cdot d + A' \cdot d, \quad s = A \cdot d + B \cdot d, \quad s' = A \cdot d.$$

y_C and y_M are molar fractions of sums of isotopomers, it holds $\|y_M\|_1 = 1$ and $\|y_C\|_1 = 1$.

Example 2 *Carrying on with the example of L-glycine (see figure 4.10), the full labeling state can be calculated.*

$$y_M = \left[\; y_M^{0:0} \;\; y_M^{1:0} \;\; y_M^{1:1} \;\; y_M^{2:0} \;\; y_M^{2:1} \;\; y_M^{2:2} \;\; y_M^{3:0} \;\; y_M^{3:1} \;\; y_M^{3:2} \;\; y_M^{3:3} \;\; y_M^{4:0} \;\; y_M^{4:1} \;\; y_M^{4:2} \;\; \dots \;\; y_M^{13:6} \; \right]^\intercal$$

For the chosen vector of d, we can calculate for $I \epsilon \{C, N, H, O\}$.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$
$$A' = \begin{bmatrix} 0 & 0 & 1 & 2 \end{bmatrix}$$
$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$
$$B' = \begin{bmatrix} 0 & 1 & 5 & 0 \end{bmatrix}$$

The vector d can be easily generated:

$$d = \begin{bmatrix} 1 & 1 & 1 & 2 \end{bmatrix}^{\mathsf{T}}$$

So for the above example, $\mathbf{t} = 13$ *and* $\mathbf{t}' = 6$ *for the overall fragment. For the carbon backbone, we obtain* $s = 2$ *and* $s' = 1$.

Usually, just a fraction of the overall vector is measured, these measurements are denoted z with length C_y. A selection matrix S_y (with dimension $M_y \times C_y$) is defined choosing the measured entries z from y_M. Now, the simulated measurement

$$\tilde{y}_{sim} = S_y \cdot y_M \tag{4.39}$$

can be calculated. After correction, the obtained measurement vector is rescaled, so that $\|\tilde{y}_{sim}(y_C)\|_1 = 1$ holds.
Most importantly, at this point the main fraction of the potentially possible measurements are removed, reducing from M_y potentially possible measurements corresponding to the full labeling state of the metabolite to C_y measurements. Thereby, only the mass traces corresponding to the skeletal atoms labeling state are measured.

Example 3 *For the example of glycine, the full labeling state* y_M *of this MS/MS measurement with* $M_y = 57$ *is obtained. However, only the fragments corresponding to labeled carbon atoms are measured, in this case: m+0.0, m+1.0, m+1.1, and m+2.1. Thus z has the length* $C_y = 4$. *The other 57-4 unmeasured mass traces can be calculated, if the natural labeling is known (which is assumed throughout this chapter).*

For the detected and non detected non skeletal atoms emerging from the mother ion, we need to generate the mass shift matrices $C_{A'}$ and $C_{B'}$. They possess only one column, each entry is referencing to a specific weight. The specific meant weight is henceforth indicated by a superscript (e.g. weight 0 "$C^0_{A'}$" corresponds to the first entry, weight 1 "$C^1_{A'}$" to the second entry and so on). Using these vectors, the beforehand mentioned general matrix for the correction is built, called C_{corr}. Each entry in this matrix corresponds to a product of $C_{A'}$ and $C_{B'}$.
First it is assumed that the labeling vector y_C of the skeletal atoms is known with length C_y. The complete unknown vector y_M of all possible labeling states is then calculated as:

$$y_M = C_{corr} \cdot y_C \tag{4.40}$$

As already mentioned, only a subset of the vector y_M is measured, which is denoted z. In the next section the formation of the correction matrix C_{corr} is described.

4.3.6 Generation of Mass Shift Matrices

In the analyte $\mathbf{N}_{element}$ elements I (i.e. N, H, O, P, ...) are found to correct for. For each element $\xi = 1, 2, \ldots, N_{element}$ a correction matrix $C_{I\xi}$ is needed. After multiplication they yield the correction vector $C_{A'}$ and $C_{B'}$ for the detected and non detected non skeletal atoms. Below the calculation of $C_{A'}$ is shown:

$$C_{A'} = \prod_{\xi=1}^{N_{element}} C_{I\xi} \tag{4.41}$$

This calculation can be performed likewise for vector $C_{B'}$. Each of these correction matrices $C_{I\xi}$ is built separately. The procedure is shown in Example 4.

Example 4 *For glycine in figure 4.10 the mass shift matrix is to be generated, as described by van Winden et al. [246]. For this purpose both vectors $C_{A'}$ and $C_{B'}$ are to be build, starting with $C_{A'}$ for the measured fragment. The fragment is composed of two oxygen and one hydrogen.*
The matrices $C_{I\xi}$ are built in a consecutive manner, the order of building them needs to be fixed before starting. First, the correction matrix C_H is built, where each entry corresponds to the natural abundance of an isotopologue indicated by its chemical formula. C_H possesses one column as it is the first matrix build and consist of two rows as one hydrogen in the non skeletal atoms ($\mathbf{N}_{H,atoms,A'} = 1$) results in two possible labeling states of hydrogen: $n_{H,rows,A'} = (\mathbf{N}_{H,isotope} - 1) \cdot \mathbf{N}_{H,atoms,A'} + 1$.

$$C_H = \begin{bmatrix} {}^{1}H_1 \\ {}^{2}H_1 \end{bmatrix}$$

The number of rows needed for matrix C_H is determining the column number of the correction matrix for C_O. The number of potential possible labeling states of the respective atom species, here oxygen, and the number of rows of the last matrix build (here: C_H) determine the number of rows $n_{O,rows,A'} = (\mathbf{N}_{O,isotope} - 1) \cdot \mathbf{N}_{O,atoms,A'} + n_{H,rows,A'}$. Two oxygen atoms can result in $\mathbf{N}_{O,isotope} = 3$ different labeling patterns as ${}^{16}O$, ${}^{17}O$ and ${}^{18}O$ are naturally occurring. This results in 5 different labeling states:

$$C_O = \begin{bmatrix} {}^{16}O_2 & 0 \\ {}^{16}O_1 \cdot {}^{17}O_1 & {}^{16}O_2 \\ {}^{18}O_1 \cdot {}^{16}O_1 + {}^{17}O_2 & {}^{16}O_1 \cdot {}^{17}O_1 \\ {}^{18}O_1 \cdot {}^{17}O_1 & {}^{18}O_1 \cdot {}^{16}O_1 + {}^{17}O_2 \\ {}^{18}O_2 & {}^{18}O_1 \cdot {}^{17}O_1 \\ 0 & {}^{18}O_2 \end{bmatrix}$$

The first column of the matrix is describing a labeling state, where other atoms of the measured fragment (here: hydrogen) are unlabeled. The first row is describing the unlabeled state of the measured fragment (m+0): all oxygen is unlabeled ${}^{16}O_2$. The second row is describing the m+1 labeling of the measured fragment by the labeling state of oxygen. The occurrence of ${}^{16}O_1{}^{17}O_1$ results in m+1 labeling here. In the second column, the other atom labeling state in the measured fragment is increased to m+1 (here: hydrogen is ${}^{2}H_1$) and so forth.

The vector $C_{A'} = C_N \cdot C_H$ can be built.

$$C_{A'} = \begin{bmatrix} {}^{1}H_1 \cdot {}^{16}O_2 \\ {}^{1}H_1 \cdot {}^{16}O_1 \cdot {}^{17}O_1 + {}^{2}H_1 \cdot {}^{16}O_2 \\ {}^{1}H_1 \cdot ({}^{18}O_1 \cdot {}^{16}O_1 + {}^{17}O_2) + {}^{2}H_1 \cdot {}^{16}O_1 \cdot {}^{17}O_1 \\ {}^{1}H_1 \cdot {}^{18}O_1 \cdot {}^{17}O_1 + {}^{2}H_1 \cdot ({}^{18}O_1 \cdot {}^{16}O_1 + {}^{17}O_2) \\ {}^{1}H_1 \cdot {}^{18}O_2 + {}^{2}H_1 \cdot {}^{18}O_1 \cdot {}^{17}O_1 \\ {}^{2}H_1 \cdot {}^{18}O_2 \end{bmatrix}$$

Note, the first correction matrix C_H and each following correction matrices' columns determine the rows of the subsequent ones. A similar matrix needs to be built also for the non detected non skeletal atoms of the mother ion, whereas here five hydrogen atoms ($\mathbf{N}_{H,isotope} = 2$), one carbon and one nitrogen ($\mathbf{N}_{N,isotope} = 2$) are present. First, again, the hydrogen correction is built $n_{H,rows,B'} = (\mathbf{N}_{H,isotope} - 1) \cdot \mathbf{N}_{H,atoms,B'} + 1$:

$$C_H = \begin{bmatrix} {}^{1}H_5 \\ {}^{1}H_4 \cdot {}^{2}H_1 \\ {}^{1}H_3 \cdot {}^{2}H_2 \\ {}^{1}H_2 \cdot {}^{2}H_3 \\ {}^{1}H_1 \cdot {}^{2}H_4 \\ {}^{2}H_5 \end{bmatrix}$$

Next the correction matrix for nitrogen is built $n_{N,rows,B'} = (\mathbf{N}_{N,isotope} - 1) \cdot \mathbf{N}_{N,atoms,B'} + n_{H,rows,B'}$:

$$C_N = \begin{bmatrix} {}^{14}N_1 & 0 & 0 & 0 & 0 & 0 \\ {}^{15}N_1 & {}^{14}N_1 & 0 & 0 & 0 & 0 \\ 0 & {}^{15}N_1 & {}^{14}N_1 & 0 & 0 & 0 \\ 0 & 0 & {}^{15}N_1 & {}^{14}N_1 & 0 & 0 \\ 0 & 0 & 0 & {}^{15}N_1 & {}^{14}N_1 & 0 \\ 0 & 0 & 0 & 0 & {}^{15}N_1 & {}^{14}N_1 \\ 0 & 0 & 0 & 0 & 0 & {}^{15}N_1 \end{bmatrix}$$

The correction matrix can be calculated:

$$C_{B'} = \begin{bmatrix} {}^{14}N_1 \cdot {}^{1}H_5 \\ {}^{15}N_1 \cdot {}^{1}H_5 + {}^{14}N_1 \cdot {}^{1}H_4 \cdot {}^{2}H_1 \\ {}^{15}N_1 \cdot {}^{1}H_4 \cdot {}^{2}H_1 + {}^{14}N_1 \cdot {}^{1}H_3 \cdot {}^{2}H_2 \\ {}^{15}N_1 \cdot {}^{1}H_3 \cdot {}^{2}H_2 + {}^{14}N_1 \cdot {}^{1}H_2 \cdot {}^{2}H_3 \\ {}^{15}N_1 \cdot {}^{1}H_2 \cdot {}^{2}H_3 + {}^{14}N_1 \cdot {}^{1}H_1 \cdot {}^{2}H_4 \\ {}^{15}N_1 \cdot {}^{1}H_1 \cdot {}^{2}H_4 + {}^{14}N_1 \cdot {}^{2}H_5 \\ {}^{15}N_1 \cdot {}^{2}H_5 \end{bmatrix}$$

The entries given in the correction matrices (in example 4) are abundances of the separate atom species' isotopes. Each element I^{ξ} has $\mathbf{N}_{I^{\xi},isotope}$ naturally occurring isotopes $I^{\xi}_1, ..., I^{\xi}_{N_{isotope}}$ with the natural abundance $p(I)$ given in table 4.3. For hydrogen $I^{H}_1 = {}^{1}\mathrm{H}$ and $I^{H}_2 = {}^{2}\mathrm{H}$ are found ($\mathbf{N}_{H,isotope} = 2$). The frequency $f(I^{\xi}_1), ..., f(I^{\xi}_{\mathbf{N}_{isotope}})$ is the number of

occurrences of the respective isotope in the isotopologue observed. Thus, the abundance in the correction matrices C_{I^ξ} can be calculated (taken from [246]):

$$Abund(I^\xi) = \left(\left(\sum_{i=1}^{\mathbf{N}_{I,isotope}} f(I_i^\xi)\right)!\right) \cdot \prod_{i=1}^{\mathbf{N}_{I,isotope}} \frac{p(I_i^\xi)^{f(I_i^\xi)}}{f(I_i^\xi)!} \tag{4.42}$$

Using the correction matrices for the detected and non-detected non skeletal atoms, the final correction matrix C_{corr} is calculable. For example, the abundance of $^{17}O_1 \cdot {}^{18}O_1$ in the mass shift vector can be calculated. The frequency is given by $f(I_1^O) = 0$, $f(I_2^O) = 1$, $f(I_3^O) = 1$ and natural abundance $p(I_1^O)$ =0.99757, $p(I_2^O)$ =0.00038, $p(I_3^O)$ =0.00205 thus the abundance of I^O is calculated to be $7.97{\cdot}10^{-7}$. It has to be noted, that for sums of isotopomer abundances each single term needs to be calculated separately with equation 4.42 and afterwards summed. The matrix C_{corr} contains in the columns the labeling states of the skeleton whose masses are shifted to higher masses by the natural labeling. Mass shifts leading to a specific $aa'bb' : aa'$ value are found in the rows (see example 5). To build this matrix each possible labeling state of the mother and daughter ion's non skeletal atoms $C_{A'}^{a'}$ and $C_{B'}^{b'}$ are calculated, the superscripts indicate the labeling state of the respective non skeletal atoms and referring to an entry in the vector. The first element in the vector is referring to unlabeled non skeletal atoms (m+0), the second m+1 labeled and so on. Their product is equal to the overall labeling state of daughter and mother ion's non skeletal atoms.
These entries are subject to the following constraints:

- The labeling state ($aa'bb'$) of the mother ion has to be larger or equal to the number of labeling states in the detected fragment (aa').
- The sum of all skeletal atoms labeling state (ab) has to be larger than the labeling state of detected skeletal atoms.
- And finally, the non skeletal atoms labeling states ($aa' - a$) in the daughter ion needs to be smaller than maximal labeling state in the detected fragment ($A'd$) and the same holds for the not-detected ion fragment.

Going through each potential labeling state of the skeletal atoms (iterating through a and ab) and all atoms in mother ion (iterating through $aa'bb'$ and aa'), the separate constraints can be checked (adapted and extended by constraints from [158]):

$$C_{corr}^{aa'bb':aa'} = \begin{cases} C_{A'}^{a'} \cdot C_{B'}^{b'}, & aa'bb' \geq aa' \wedge ab \geq a \wedge A'd \geq (aa' - a) \wedge \\ & B'd \geq (aa'bb' - aa' - (ab - a)) \\ 0, & otherwise \end{cases} \tag{4.43}$$

If these constrains are fulfilled, the product of the two matrices $C_{A'}^{a'}$ and $C_{B'}^{b'}$ entries are written to the specified position in matrix C_{corr}.

Example 5 *The correction procedure from the assumed to be known skeletal labeling vector to the overall vector is conducted.*

The maximal labeling for glycine is m+13 for the mother ion and m+6 for the daughter ion. The maximal labeling state of the detected fragment non-skeletal atoms is 5 and the not detected non skeletal atoms‘ is 6. The matrix can be built for glycine:

$$\begin{bmatrix} y_M^{0:0} \\ y_M^{1:0} \\ y_M^{1:1} \\ y_M^{2:0} \\ y_M^{2:1} \\ y_M^{2:2} \\ y_M^{3:0} \\ y_M^{3:1} \\ y_M^{3:2} \\ y_M^{3:3} \\ y_M^{4:0} \\ y_M^{4:1} \\ y_M^{4:2} \\ y_M^{4:3} \\ y_M^{4:4} \\ y_M^{5:0} \\ \vdots \\ y_M^{13:6} \end{bmatrix} = \begin{bmatrix} C_{A'}^0 \cdot C_{B'}^0 & & & \\ C_{A'}^0 \cdot C_{B'}^1 & C_{A'}^0 \cdot C_{B'}^0 & & \\ C_{A'}^1 \cdot C_{B'}^0 & & C_{A'}^0 \cdot C_{B'}^0 & \\ C_{A'}^0 \cdot C_{B'}^2 & C_{A'}^0 \cdot C_{B'}^1 & & \\ C_{A'}^1 \cdot C_{B'}^1 & C_{A'}^1 \cdot C_{B'}^0 & C_{A'}^0 \cdot C_{B'}^1 & C_{A'}^0 \cdot C_{B'}^0 \\ C_{A'}^2 \cdot C_{B'}^0 & & C_{A'}^1 \cdot C_{B'}^0 & \\ C_{A'}^0 \cdot C_{B'}^3 & C_{A'}^0 \cdot C_{B'}^2 & & \\ C_{A'}^1 \cdot C_{B'}^2 & C_{A'}^1 \cdot C_{B'}^1 & C_{A'}^0 \cdot C_{B'}^2 & C_{A'}^0 \cdot C_{B'}^1 \\ C_{A'}^2 \cdot C_{B'}^1 & C_{A'}^2 \cdot C_{B'}^0 & C_{A'}^1 \cdot C_{B'}^1 & C_{A'}^1 \cdot C_{B'}^0 \\ C_{A'}^3 \cdot C_{B'}^0 & & C_{A'}^2 \cdot C_{B'}^0 & \\ C_{A'}^0 \cdot C_{B'}^4 & C_{A'}^0 \cdot C_{B'}^3 & & \\ C_{A'}^1 \cdot C_{B'}^3 & C_{A'}^1 \cdot C_{B'}^2 & C_{A'}^0 \cdot C_{B'}^3 & C_{A'}^0 \cdot C_{B'}^2 \\ C_{A'}^2 \cdot C_{B'}^2 & C_{A'}^2 \cdot C_{B'}^1 & C_{A'}^1 \cdot C_{B'}^2 & C_{A'}^1 \cdot C_{B'}^1 \\ C_{A'}^3 \cdot C_{B'}^1 & C_{A'}^3 \cdot C_{B'}^0 & C_{A'}^2 \cdot C_{B'}^1 & C_{A'}^2 \cdot C_{B'}^0 \\ C_{A'}^4 \cdot C_{B'}^0 & & C_{A'}^3 \cdot C_{B'}^0 & \\ C_{A'}^0 \cdot C_{B'}^5 & C_{A'}^0 \cdot C_{B'}^4 & & \\ \vdots & \vdots & \vdots & \vdots \\ & & & C_{A'}^5 \cdot C_{B'}^6 \end{bmatrix} \cdot \begin{bmatrix} y_C^{0:0} \\ y_C^{1:0} \\ y_C^{1:1} \\ y_C^{2:1} \end{bmatrix}$$

The final measurement vector is just containing the equivalents to the possible skeletal labeling $z = [y_M^{0:0} y_M^{1:0} y_M^{1:1} y_M^{2:1}]^\intercal$. However, the entries in this matrix are readily explainable. The first row is addressed to the full labeling pattern y_M of the mother and daughter ion and is depending on the unlabeled skeletal atoms fraction $y_C^{0:0}$. The probability for obtaining unlabeled compound $y_M^{0:0}$ is given by the probabilities for the detected and non-detected non skeletal atoms to be unlabeled $C_{A'}^0 C_{B'}^0$. For the mass trace $y_M^{1:0}$, the labeling is only present in the non-detected fragment, thus the labeling is influenced by skeletal atoms of the non-detected skeletal atoms $y_C^{1:0}$ if all non skeletal atoms are unlabeled $C_{A'}^0 C_{B'}^0$. Additionally, the probability for natural labeling of the non-detected non skeletal atoms $C_{A'}^0 C_{B'}^1$ will also influence the labeling state $y_M^{1:0}$ as it is multiplied with the labeling state for all skeletal atoms $y_C^{0:0}$.

Entering the abundances from equation 4.42, the final correction matrix can be calculated:

$$\begin{bmatrix} y_M^{0:0} \\ y_M^{1:0} \\ y_M^{1:1} \\ y_M^{2:0} \\ y_M^{2:1} \\ y_M^{2:2} \\ y_M^{3:0} \\ y_M^{3:1} \\ y_M^{3:2} \\ y_M^{3:3} \\ y_M^{4:0} \\ y_M^{4:1} \\ y_M^{4:2} \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.991 & & & \\ 4.20 \cdot 10^{-3} & 0.991 & & \\ 8.68 \cdot 10^{-4} & & 0.991 & \\ 2.24 \cdot 10^{-6} & 4.20 \cdot 10^{-3} & & \\ 3.71 \cdot 10^{-6} & 8.68 \cdot 10^{-4} & 4.20 \cdot 10^{-3} & 0.991 \\ 4.07 \cdot 10^{-3} & & 8.68 \cdot 10^{-4} & \\ 4.99 \cdot 10^{-10} & 2.24 \cdot 10^{-6} & & \\ 1.96 \cdot 10^{-9} & 3.71 \cdot 10^{-6} & 2.24 \cdot 10^{-6} & 4.20 \cdot 10^{-3} \\ 1.73 \cdot 10^{-5} & 4.07 \cdot 10^{-3} & 3.71 \cdot 10^{-6} & 8.68 \cdot 10^{-4} \\ 2.02 \cdot 10^{-6} & & 4.07 \cdot 10^{-3} & \\ 5.65 \cdot 10^{-14} & 4.99 \cdot 10^{-10} & & \\ 4.38 \cdot 10^{-13} & 1.96 \cdot 10^{-9} & 4.99 \cdot 10^{-10} & 2.24 \cdot 10^{-6} \\ 9.19 \cdot 10^{-9} & 1.73 \cdot 10^{-5} & 1.96 \cdot 10^{-9} & 3.71 \cdot 10^{-6} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \cdot \begin{bmatrix} y_C^{0:0} \\ y_C^{1:0} \\ y_C^{1:1} \\ y_C^{2:1} \end{bmatrix}$$

From the nominal entries in the correction matrix, it is obvious that the correction will not

significantly alter the measured labeling pattern. From each measured carbon labeling state, roughly 1 % of the carbon labeling state is addressed to natural labeling.

To summarize, the calculation of the matrix C_{corr} is formulated as pseudo-code in algorithm 4.1. The required input for the program are the number atoms per element of mother and daughter ion (*AtomsMoth*, *AtomsDau*), as well as the number of skeletal atoms in mother and daughter ion (*CMoth*, *CDau*). Besides, also a list with the maximal number of labeled states per atom specy ($N_{isotope}$) and their natural abundances ($p_{isotope}$) as shown in table 4.3 need to be provided.

4.3.7 Optimization Procedure

Finally, coming from the real carbon measurement vector, the overall labeling state can be calculated. In a real setup, the carbon vector is of interest and needs to be estimated. By the mass spectrometric device we get measurements $\tilde{z}$ with standard deviations corresponding to the simulated vector $z(y_C)$. The problem can be reformulated as a least square problem [255]. Nevertheless, this might cause problems as several mass traces can become negative during this procedure (by stochastic measurement errors) and need to be corrected manually afterwards. To prevent this, the problem is formulated as an constrained least square optimization problem, using the covariance matrix of the measurement error Σ:

$$\widetilde{y}_C = \arg\min_{y_C} \parallel z(y_C) - \tilde{z} \parallel_{\Sigma}^2 \; subject\, to \begin{cases} 0 \leq y_C \leq 1 \\ \|y_C\|_1 = 1 \end{cases}, \tag{4.44}$$

Finally, $\widetilde{y}_C$ is estimated.

Statistical Analysis and Consistency Check

The statistics of the system can be calculated by using the Jacobian J and the covariance matrix of the measurement error Σ. It is assumed that the measurements show no dependency and thus the matrix contains only the measurements standard deviations as diagonal elements $\Sigma_{i,j} = \sigma^2,\ i = j, \Sigma_{i,j} = 0,\ i \neq j$.

$$Cov = J(\widetilde{y}_C) \cdot \Sigma^{-1} \cdot J(\widetilde{y}_C)^{\intercal} \tag{4.45}$$

The Jacobian is calculated by NoReTO toolbox using complex differentiation, see Martins et al. for details about this method [148].

4.3.8 Results of Correction

Application to Test Cases

In figure 4.13 examples are given for the correction. The first example is an alanine LC-MS/MS measurement. Here the correction does not lead to a high change in the labeling pattern. For methionine, the effect of the correction is in the order of 0.01 of the measurement. For GC-MS measurements the effect is more pronounced as alanine is derivatized by sylilation reagents. Introducing silicon results in a significant amount of natural labeling and correction

Algorithmus 4.1 Calculation of C_{corr}

Algorithm 4.3.1: CALC_CCORR($AtomsMoth, AtomsDau, CMoth, CDau, N_{isotope}, p_{isotope}$)

$A' \leftarrow AtomsDau$
$B' \leftarrow AtomsMoth - AtomsDau$
$A \leftarrow CDau$
$B \leftarrow CMoth - CDau$
$C\_A' \leftarrow$ CALC_MASSSHIFTMATRIX($A', N_{isotope}, p_{isotope}$)
$C\_B' \leftarrow$ CALC_MASSSHIFTMATRIX($B', N_{isotope}, p_{isotope}$)
$d \leftarrow$ CALC_MAXLABEL($N_{isotope}$)
$s \leftarrow A * d + B * d$
$s' \leftarrow A * d$
$t \leftarrow A * d + B * d + A' * d + B' * d$
$t' \leftarrow A * d + A' * d$
$Column \leftarrow 1$
for $ab \leftarrow 0$ **to** $CMoth$
  for $a \leftarrow max(0, ab - (s - s'))$ **to** $min(ab, s')$
    $Row \leftarrow 1$
    for $aa'bb' \leftarrow 0$ **to** t
      for $aa' \leftarrow max(0, aa'bb' - (t - t'))$ **to** $min(aa'bb', t')$
        $a' \leftarrow aa' - a$
        $b' \leftarrow aa'bb' - aa' - ab + a$
        $BB't \leftarrow Mother - Daughter$
        if $aa'bb' \geq aa'$
        $ab \geq a$,
        $A' * d \geq (aa' - a)$,
        $B' * d \geq (aa'bb' - aa' - (ab - a))$
          do $Ccorr[Row, Column] \leftarrow C\_A'(A') * C\_B'(B')$
          else
          do $Ccorr[Row, Column] \leftarrow 0$
        $Row \leftarrow Row + 1$
    $Column \leftarrow Column + 1$
return $(Ccorr)$

is mandatory. This is especially important if GC-MS is used as here large side groups are attached. By introduction of sylilation reagents, the labeling pattern can be altered by ~5-10 %.
Generally, in LC-MS(/MS) the introduced error by natural labeling is in the order of 0.1-1.0 %, but is significant.

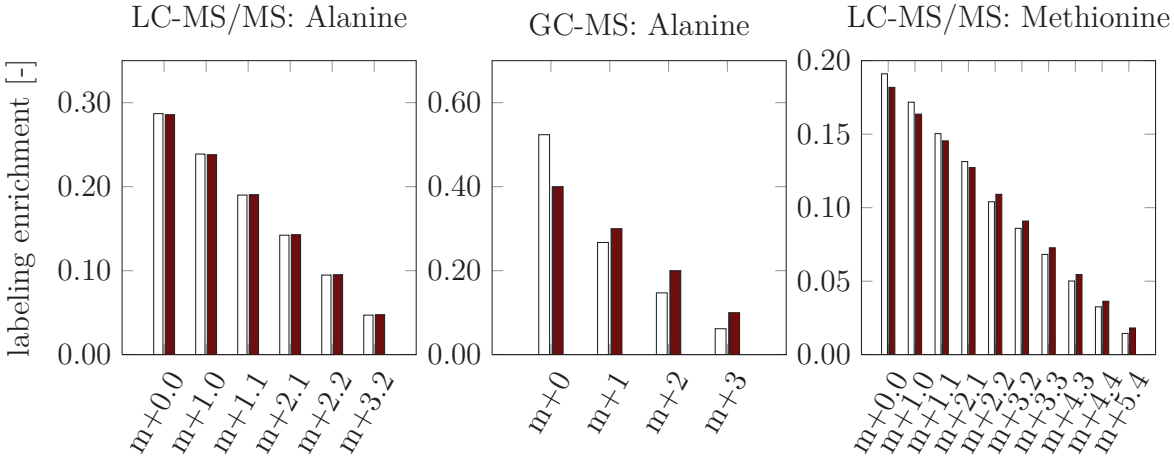
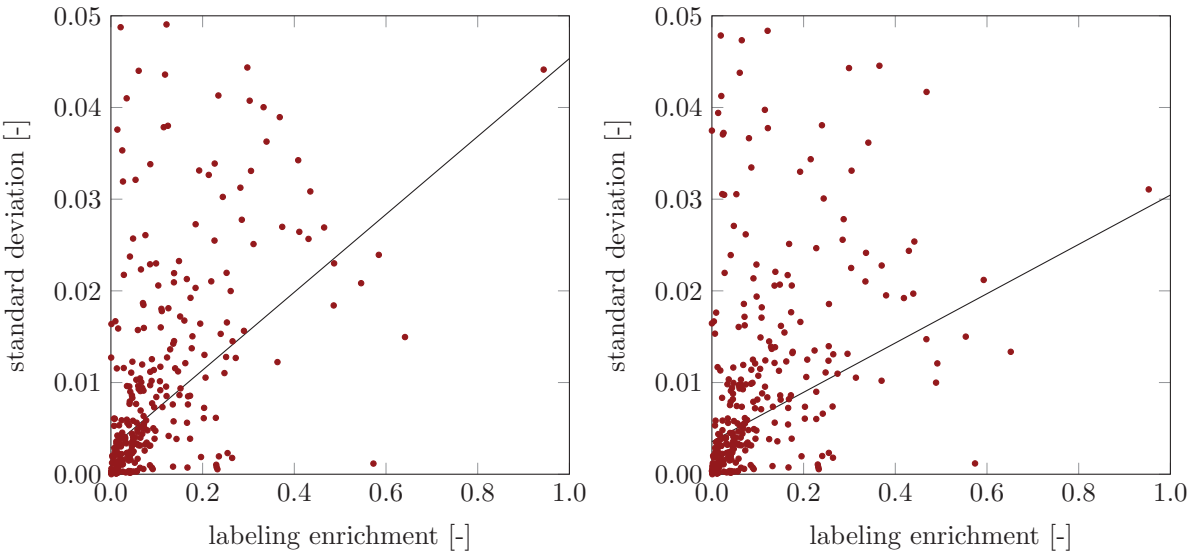


figure 4.13: Correction results of measurements using LC-MS/MS and GC-MS for the metabolites alanine and methionine. Alanine measurement for GC-MS is derivatized with two tert-butyldimethylsilyl groups at carboxy and amine group, the $[M-0]^+$ ion was used. Uncorrected measurements (▬) and corrected measurement (▭) are shown. The correction was performed for non skeletal carbon atoms. Artificial measurement values shown.

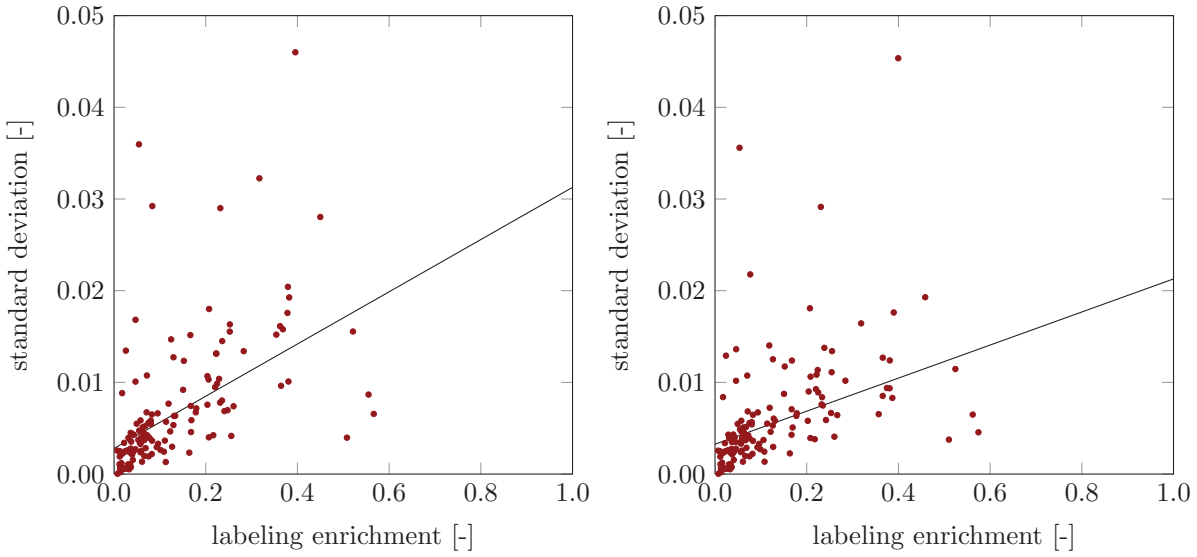
Application to Chemostat and Fed-Batch Data

The correction for natural abundance was performed for all extrapolated measurements. Additionally to the change in the absolute value, the uncertainties of the labeling measurements are also effected. The extrapolation was performed for each mass trace separately. During error propagation calculations, the peaks are evaluated as dependent of each other (their sum equals "one"). Thus, their respective standard deviation is changing during correction. This statistical reconciliation can be important if some peaks show large standard deviation. In figure 4.14 the standard deviations for a fed-batch and a chemostat are shown, before (left plot) and after (right plot) correction for natural abundance and statistical reconciliation.
To visualize the difference in the standard deviations before and after correction and statistical reconcilation, a linear regression model ($y = b + m \cdot x$) was fit to the data set. The linear regression model for the fed-batch data was $y = 0.003 + 0.042 \cdot x$ for the extrapolated mass traces and after statistical reconciliation $0.004 + 0.036 \cdot x$. Thus, lower measurement values (i.e. values close to zero) became less statistically identified, whereas large measurement values‘ (i.e. values close to one) standard deviation were decreasing. For the chemostat the estimated standard deviations for the extrapolated mass traces were lower, the regression model yields $y = 0.003 + 0.028 \cdot x$. After correction $y = 0.003 + 0.018 \cdot x$. The statistical reconciliation resulted in better statistics.
It is important to note that several labeling measurements, especially in chemostat experiments, showed low measurement standard deviations after extrapolation (see figure 4.14, left

plots). Low standard deviations result if the measurement of the mass traces were at steady state or showed no change in their labeling state. Besides, mass traces with low signal can be subject of peak integration error (peak is too low to be integrated). Thus, also a systematic error can be introduced at this stage of the analysis.
In case a lower standard deviation was found than the threshold of 0.005, the measurements' standard deviation were set to 0.005 (before and after statistical reconciliation). This is done to prevent false prediction in statistics later in ^{13}C-MFA based on falsely under-estimated standard deviations.



(a) Fed-batch process



(b) Chemostat

figure 4.14: Effects of statistical reconciliation for (a) fed-batch 1 and (b) chemostat 1 cultivation. Measurement with standard deviation larger than 0.05 were excluded.

4.4 Results

In this chapter, three important steps are addressed that concern measurement data to be incorporated in ^{13}C-MFA models and, thus, directly impact model results:

- A bio-process model for rate estimation:
  In total the kinetic model contains 24 parameters (mainly first order and Michaelis Menten kinetics). Model parameters were estimated for three chemostat and two fed-batch cultivations. The calculated concentration data possessed, for most substances, good reproducibility, although fluctuations in concentration profile were not described by the model. It was shown that *P. chrysogenum* BCB1 produces high amounts of penicillin V, especially in the fed-batch process. However, side-product formation was also increased in this process compared to chemostats at a growth rate of 0.03 and $0.05\,\mathrm{h}^{-1}$.

- Extrapolation of labeling time series:
  Time dependent labeling enrichments from MS are, by use of a nonlinear regression approach, extrapolated to the corresponding steady-state values. Here, the underlying model is based on exponential growth ($m+1, 2, \dots$) and decay (m+0) functions. Estimates for the turnover of metabolite pools were obtained and discussed in the context of published literature. Besides, a nonlinear regression toolbox was implemented.

- Correction for natural abundance for tandem MS:
  While plenty of tools are available for the correction of MS data, none of these is able to correct for tandem MS. An algorithm was developed to build the mass shift matrices. The correction for natural abundance was applied to the obtained extrapolated mass traces and statistical reconciliation was performed.

In the next chapter, the ^{13}C-MFA model is explained in detail.

Chapter 5

^{13}C-MFA Model for *P. chrysogenum* BCB1

The conduction of experiments and data pre-processing described in the last chapter are only two major steps needed for ^{13}C metabolic flux analysis. The model generation process is of equal importance. Here, this process is described for the strain BCB1 of *P. chrysogenum*. The metabolic pathways included in the ^{13}C-MFA model are described and critical design decisions and their importance for the outcome of the study are discussed. Initially, a model composed of 500 reactions containing cofactors and oxidative phosphorylation was built for *P. chrysogenum*. Then, for ^{13}C-MFA purposes, a focused ^{13}C metabolic network with 140 reactions was constructed, by simplifying the large network. It covers all pathways in *P. chrysogenum* that lead to the formation of measurable metabolites.
For both networks a strain specific biomass equation is derived for the strain BCB1 of *P. chrysogenum*. Finally, with the large scale metabolic network flux balance analysis is performed and the strain‘s performance is characterized by comparison to literature values.

5.1 Metabolic Network

Metabolic networks used for ^{13}C-MFA are generated by genome scale metabolic network reconstruction (GEM) or by best knowledge on the organism if a GEM is not available. Usually, these GEMs are assigning functions to potential genes (open reading frames) if their amino acid sequence possesses a high homology to proteins with known function. At the starting point of this work, no GEM was available for *P. chrysogenum*. In 2013 the first GEM for *P. chrysogenum* was published and it was used to proof read the built networks. Beforehand, a draft of this GEM was provided by Sandoz [2].

5.1.1 Large Scale ^{13}C-MFA Network

A metabolic network was created following the work-flow shown in figure 5.1. Initially, the reaction network was built with the help of the Kyoto Encyclopedia of Genes and Genomes (KEGG). For this process, the reactions needed to form the known biomass components and products were identified and their formation pathways included in the model (see chapter 5.1.4). KEGG is a pathway database that maps genomic information to chemical and systemic function, e.g. it maps a gene to its specific chemical reaction or regulatory function

[117]. The assignment of gene functions in KEGG is performed by the so called "KEGG orthology" which contains information of ortholog groups and links them to a function in the metabolic network. In addition, KEGG uses information about so called "ortholog clusters" containing similar genes based on their sequence. KEGG updates its contents every 3 month and contains by now (in 2013) 2112 complete genomes, including 153 genomes of eukaryotes [167]. However, KEGG can also be used to model the fate of atoms in the reactions [116, 115].

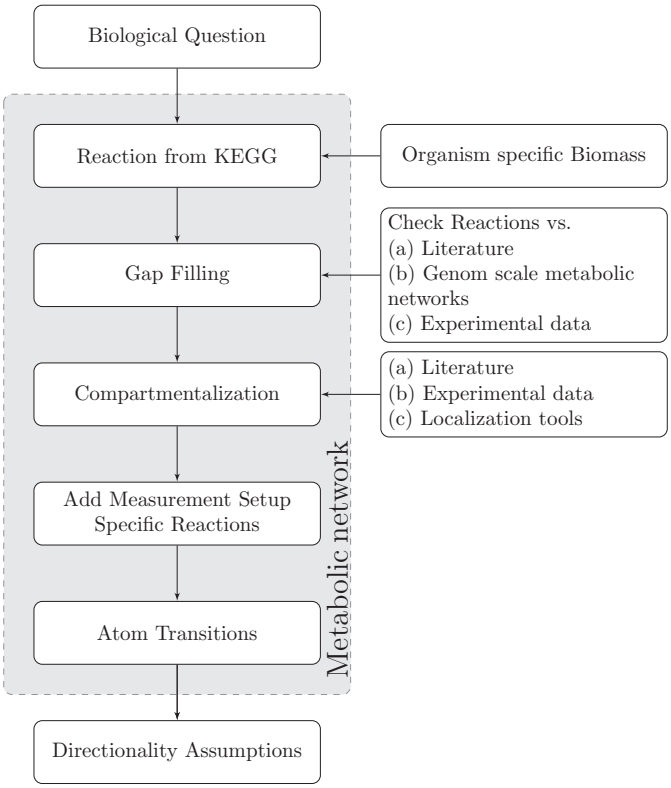


figure 5.1: Work-flow of ^{13}C-MFA model construction.

During generation of the metabolic reaction network, several enzymes could not be identified in KEGG but were known to be necessary for *de novo* synthesis of cellular components. This leads to the gap filling process described by Ogata et al. [182]. In the large network missing reactions were filled manually by literature knowledge or by similarity to other organisms. Based on the known components of the cell, which are crucial or typically found in this organism, a large scale network was constructed applying the data derived from the KEGG database. The compartmentalization of *P. chrysogenum* and finally the atom transitions were specified and the resulting network was compared to literature data. Finally, directionality assumptions were made.

Compartmentalization

Penicillin production in *P. chrysogenum* occurs throughout several compartments, especially involved are mitochondria and peroxisomes [72, 239]. The localization of several enzymes involved in the penicillin biosynthesis pathway can be found in literature. An overview is given in table 5.1. As the experimental knowledge about localization in *P. chrysogenum* is limited, tools for the prediction of the sub-cellular localization of proteins were applied.
MultiLoc2 was used for sequence based prediction of the sub-cellular localization of proteins and the results obtained by this program are visualized in figure 5.2 [27]. MultiLoc2 provides in most cases consistent estimates of the protein localization compared to literature references (e.g. particularly localization of amino acid biosynthesis for penicillin production) [72, 239]. In reaction sequences for the formation of biomass components, often predictions for other deviating compartments were received. However, the pathway was placed in the compartment where most reactions were predicted for. Finally compartmentalization is consistent in the model with the compartmentalization for amino acid formation proposed by Kleijn based on

table 5.1: Localization of single enzymes of the penicillin formation pathway. cyt.: cytosolic, vac.: vacuolar.

Protein	Localization	Exp. Method	Literature
α-Isopropylmalate synthase	cyt.	Product labeling pattern	[40]
Homocitrate synthase	cyt.	Fusion to GFP; labeling pattern	[40, 13]
ACVS	cyt./vac. membrane bound	immunogold	[164]
IPNS	cyt.	immunogold	[164]
AT	microbodies	immunogold	[164, 165]
PCL	microbodies	sequence analysis	[129]

comparison of labeling patterns, except for formation pathways for aspartate and glutamate which might be cytosolic or mitochondrial based on the prediction of the localization [121]. Correction of compartmentalization by literature and prediction software was performed manually. Proteins localized within the endoplasmatic reticulum, nucleus and cytosol were all assigned to the cytosolic compartment.

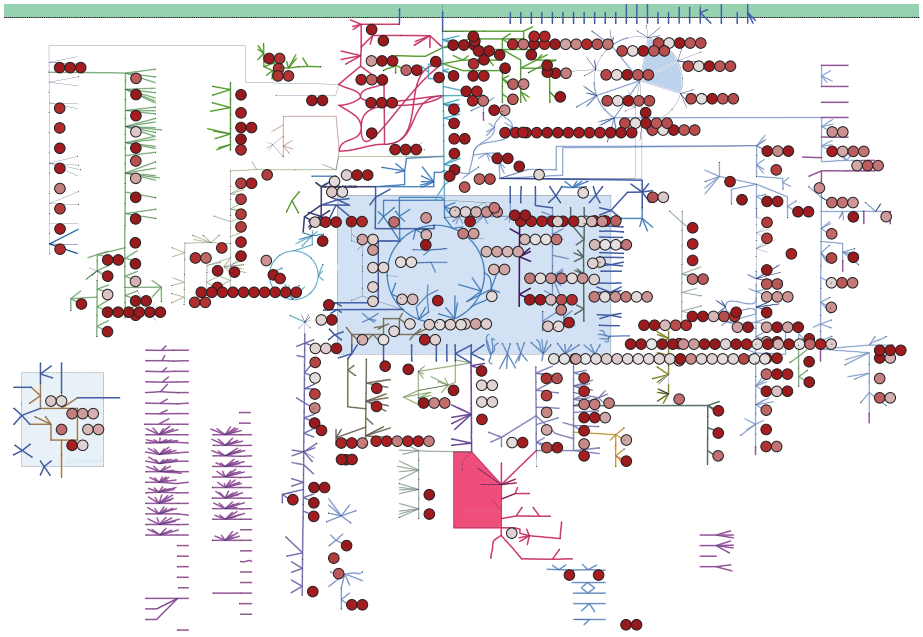
Proofreading and Comparison

Proofreading of the final model was performed by comparing the reaction stoichiometry to the recent yeast model version 6.05 and the recently published genome scale network of *P. chrysogenum* [2, 95]. The *P. chrysogenum* model was used to verify the compartmentalization and the enzyme‘s assigned gene identifier.

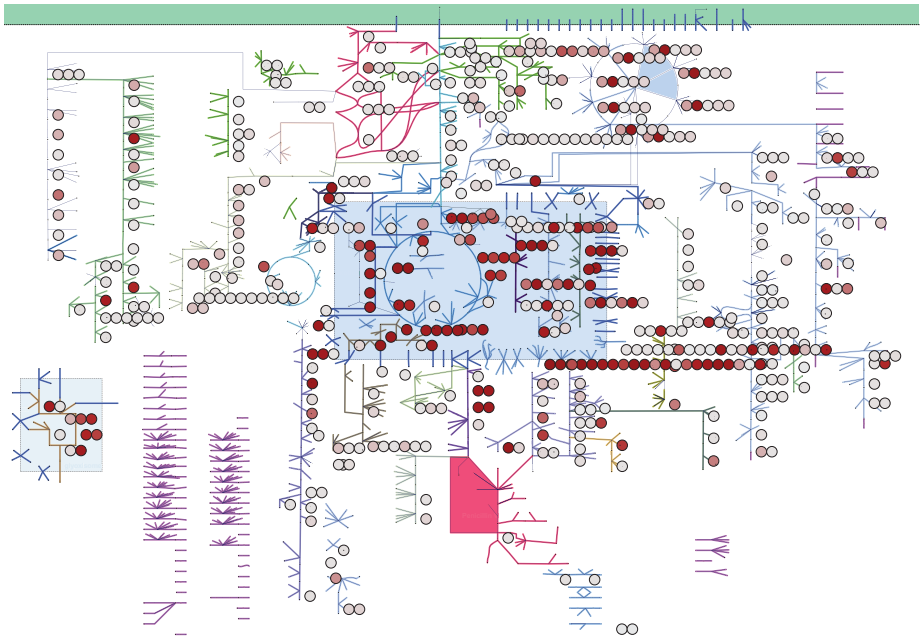
5.1.2 Pathways in the Large Scale Model

In this section, the importance of selected pathways for the ^{13}C-MFA in *P. chrysogenum* is discussed and uncertainties regarding their compartmentalization and regulation are exposed.

The network contains pathways for the uptake and metabolism of glucose, gluconate, ethanol, acetate, ammonia, phosphate, sulfate, and oxygen. Although in the performed chemostat and fed-batch experiments only glucose as sole carbon source is used. Additionally, it contains pathways for the formation of the most frequently found products like carbon dioxide, pyruvate, succinate, oxalate, penicillin, and its side products. Moreover, pathways for *de novo* synthesis of nucleotides, amino acids, polyols, sugars, polysaccharides, acylglycerols, phospholipids, phosphatidates, and ergosterolester were integrated.
The glycolysis, PPP and citrate cycle were modeled according to textbook knowledge [253]. Amino acid formation pathways were taken from KEGG and proofread using the recent GEM for *P. chrysogenum* [2]. The formation of glycine by the action of enzymes threonine aldolase and serine hydroxymethyltransferase were also incorporated in the model [40].
Fatty acids containing metabolites can possess more than 50 carbon atoms, they were modeled stoichiometrically (i.e. no atom transition were used), and connected by constraints to the rest of the network. Also cofactors like NADH, NADPH and ATP were modeled stoichiometrically.



(a) Cytosolic localization predicted



(b) Mitochondrial localization predicted

figure 5.2: Compartmentalization of metabolic pathways predicted by MultiLoc2. HighRes setting was used. (a) Protein‘s summed prediction value for cytosol, endoplasmatic reticulum and nucleus. (b) Prediction for mitochondrial compartmentalization. For each protein linked to the respective enzymatic function shown, a circle is drawn. The fill color indicates the prediction. All prediction values for the separate compartments (nuclear, cytoplasmic, mitochondrial, chloroplast, extracellular, plasma membrane, peroxisomal, endoplasmic reticulum, Golgi apparatus, lysosomal and vacuolar) sum up to 1, whereas in the shown network (▭) equals a prediction of 0 and (▬) equals 1. Colors in between were linearly interpolated.

table 5.2: Localization of anaplerotic enzymes predicted by YLoc [29, 30]. cyt.: cytosolic, mit.: mitochondrial, nu.:nucleus,

		Network		YLoc			
	Gene	cyt.	mit.	cyt.	mit.	nu.	Propability
PEP Carboxykinase	Pc12g09360	✓		99.4	0.0	0.6	0.86
PYR Carboxylase	Pc13g07230	✓		85.9	0.0	14.0	0.56
Malic enzyme (NAD)	Pc21g20250		✓	0.0	100	0.0	0.99
Malic enzyme (NADP)	Pc13g04510		✓	0.0	99.9	0.0	0.99

Anaplerosis and Gluconeogenesis

Anaplerotic reactions replenish citrate cycle intermediates which were removed for the formation of biomass components. In *P. chrysogenum* pyruvate carboxylase, phosphoenolpyruvate carboxykinase and malic enzyme are found, whereas the fructose-1,6-bisphosphatase, responsible for the gluconeogenetic flux in upper glycolysis, is expected to be in-active, because it is known to be repressed in yeasts by glucose [217]. The reactions catalyzed by these enzymes form additional cycles in the metabolic network and are subject to compartmentalization. As they are crucial for the conduction and interpretation of the final results, the localization was determined using YLoc and are listed in table 5.2 [30, 29]. YLoc predicts a mitochondrial location of the NAD- and NADP- dependent malic enzyme, which is consistent with the network for ^{13}C-MFA used by Kleijn for *P. chrysogenum*. Activity of the enzyme in *P. chrysogenum* was shown by Harris et al. [94]. In yeasts, the malic enzyme is also located in the mitochondria (and potentially in the peroxisomes) [28], which supports the results obtained via YLoc for the localization of it in *P. chrysogenum*. In contrast, Zheng removed the malic enzyme from the *P. chrysogenum* network, because malic enzyme activity was not detected [284, 121].
For the pyruvate carboxykinase and pyruvate carboxylase a cytosolic localization was predicted by YLoc and, thus, both enzymes were incorporated in the model. The same localization can be found in networks of Kleijn and Zheng for *P. chrysogenum* [121, 284].

Glyoxylate Shunt

The glyoxylate shunt is mainly used in anaplerotic reactions. Elevated expression levels of the genes encoding for isocitrate lyase and malate synthase are found in several fungi growing on ethanol [151]. Usually, the glyoxylate shunt is not integrated in metabolic networks, because it is assumed to be of minor importance during growth on glucose as sole carbon source. [121, 111, 157]. However, glyoxylate was detected in *P. chrysogenum* for growth on glucose and ethanol as sole carbon sources [170]. Zheng incorporated this pathway for in-stationary ^{13}C-MFA during growth on ethanol/glucose mixtures, but the respective reactions (isocitrate lyase and malate synthetase) showed zero flux [284]. In this work, the glyoxylate shunt is introduced into the large scale model to test if it exhibits low activity in the presence of glucose as sole carbon source.

Polyol Synthesis

Polyols were frequently found in fungi. They are part of their biomass and can be secreted into the media [56, 141]. Polyols like erythritol, arabitol, and mannitol were measured in the

biomass of the used *P. chrysogenum* BCB1 strain (cf. chapter 5.1.4) and were also detected for other *P. chrysogenum* strains; e.g. mannitol was found to account for 3.2 % of the cell dry weight [141, 15, 284]. Thus, the formation of these compounds has to be considered in the model.
Mannitol can be synthesized as well as degraded by cells of *P. chrysogenum*, which leads to a cycle [284]. However, in *Aspergillus niger*, an organism closely related to *P. chrysogenum*, the mannitol synthesizing enzymes were exclusively found in vegetative cells, while degrading enzymes are active in conidiospores. Thus, co-expression of all enzymes needed for the complete mannitol cycle was not detected [3, 222]. Based on this, it is assumed that this cycle is also not present in *P. chrysogenum*. This is important, as mannitol is a symmetric molecule. The presence of the mannitol cycle could, thus, alter labeling patterns significantly and contribute to NADPH regeneration.
Although little is known about the arabitol and erythritol formation in *P. chrysogenum*, it was assumed to be consistent with *A. niger*. In this fungi, the erythritol formation is catalyzed by a NAD(P)- dependent erythrose reductase. Arabitol is synthesized either via conversion of ribose 5-phosphate via a ribulose phosphatase and a NADP-dependent ribulose reductase [56]. Another route is via conversion of xylulose 5-phosphate via NAD- or NADP-dependent xylulose reductase [56]. However, both routes use equal amounts of cofactors [56, 2]. In the genome scale reconstruction of *P. chrysogenum*, the route via xylulose-5-phosphate is used and was also integrated into the large scale ^{13}C-MFA network [2].

Oxidative Phosphorylation

Oxidative phosphorylation reactions were incorporated into the model as they are capable of specifying constraints on metabolic fluxes by limiting the amount of available energy in form of ATP. Usually, cofactor balancing is not accounted for in ^{13}C-MFA. Importantly, complete balancing of cofactors is in general not possible; thus, sinks for NADPH and ATP were incorporated into the large model to account for yet unknown processes [108].
Oxidative phosphorylation is driven by the gradient of protons at the inner mitochondrial membrane maintained by activity of the succinate dehydrogenase and NAD(P)H dehydrogenases. In addition, other reactions and proton driven transport processes result in formation or consumption of protons in the cell.
In higher eukaryotes, transport processes are important for the modeling of proton gradients across membrane, e.g. for each generated ATP in the mitochondria (which costs usually 2.67-3.3 protons) one additional proton is needed to transport ADP and P_i into the mitochondrial matrix [76].
Oxidative phosphorylation complexes were modeled according to data for yeast by Fitton et al. (see figure 4 within their work) [80]. However, the mechanistic ratio of ATP production per proton for the ATPase is not exactly known for *P. chrysogenum*. It is expected to range between 2.7 H^+/ATP for bovine heart mitochondria and 3.3 H^+/ATP for yeast [80, 76, 258]. Nevertheless, other proton dependent transport processes of (for example) P_i, pyruvate and oxaloacetate into the mitochondria will add up to these costs. In the recent GEM for *P. chrysogenum* a value of 3.87 H^+/ATP was assumed [2]. However, the model of Agren *et al.* does not, unlike our model, account for complete proton balancing and the ATP ratio therein used has to account also for further transport processes. In the network built here, it is assumed that 3-4 H^+/ATP are spent. Proton dependent transport processes are incorporated based on the recent yeast GEM [95].

Often, the P/O ratio (phosphate/oxygen ratio) is defined to measure the effectiveness of energy generation. It refers to the number of ATP molecules synthesized (by oxidative phosphorylation) for each pair of electrons. The electrons are passed from NAD(P)H or succinate to oxygen (O_2). The P/O ratio can also be described by the ratio of protons generated per NAD(P)H in relation to the protons used for the generation of one ATP including transport processes [76]:

$$\text{P/O ratio} = \frac{H^+/NAD(P)H}{H^+/ATP + Transport}$$

Clearly, the P/O ratio can be altered if further uncoupling systems, shortcuts through the electron transport chains or external NAD(P)H dehydrogenases are present (for a in-depth discussion see Guerrero-Castillo et al. [91]). Recently, the presence of external NADPH dehydrogenase in *P. chrysogenum* was indicated and incorporated in the model [94].

In the built network a P/O ratio for NADH of 2-2.5 is computed, ranging in the experimentally determined values of 2.3 for yeast (in yeast it is assumed that H^+/ATP=3.3; for this it is assumed that 10 protons per NADH were generated and one proton was used for transport of ADP and P_i into the mitochondria).

NADPH Production and Consumption in the Penicillin Pathway

The cytosolic NADH pool is assumed to be unconnected to the mitochondrial pool as the inner mitochondrial membrane is impermeable for reduced pyrimidine nucleotides [254]. The same is assumed for NADPH. NADPH is used mainly for amino acid synthesis in the cytosol, thus also for penicillin synthesis. In the cytosol, NADPH is mainly formed by pentose phosphate pathway (glucose-6-phosphat dehydrogenase and 5-phosphogluconate dehydrogenase) and during the reaction catalyzed by the cytosolic isocitrate dehydrogenase. NADP dependent isocitrate dehydrogenase was detected in cytosol and mitochondria in *P. chrysogenum* [109]. NADPH is not only produced in the cytosol, but is also consumed by a multitude of reactions. The main factor is the generation of glutamate (1 $\text{mol}_{\text{NADPH}}\ \text{mol}^{-1}_{\text{GLU}}$), which is used in transamination of most amino acids. In several other amino acid and fatty acid biosynthesis pathways, NADPH consuming reactions occur (formation of chorismate, valine, cysteine and lysine). In fatty acid biosynthesis 2 $\text{mol}_{\text{NADPH}}$ per mol of attached malonyl-CoA is consumed. Furthermore, the formation of ergosterol is costly; 19-23 NADPH are used for production of an ergosterol molecule. And synthesis of desoxy-nucleotides and polyols are consuming NADPH as well.

Of major interest for penicillin production are the formation of cysteine and valine as they are the main building blocks of penicillin. Cysteine can be formed via two pathway in *P. chrysogenum* that use serine as precursor. The sulfhydrylation pathway incorporates reduced sulfur directly into serine. Including the formation of serine and reduced sulfur it requires 5 NADPH. The second pathway (transsulfuration pathway) incorporates reduced sulfur into O-acetylhomoserine and uses 7-8 NADPH. Homo-cysteine, the product of this reaction, is used together with serine to form cystathione which is converted to cysteine. In the course of this study, both reaction pathways cannot be elucidated by stationary metabolic flux analysis as they possess the same carbon atom transitions.

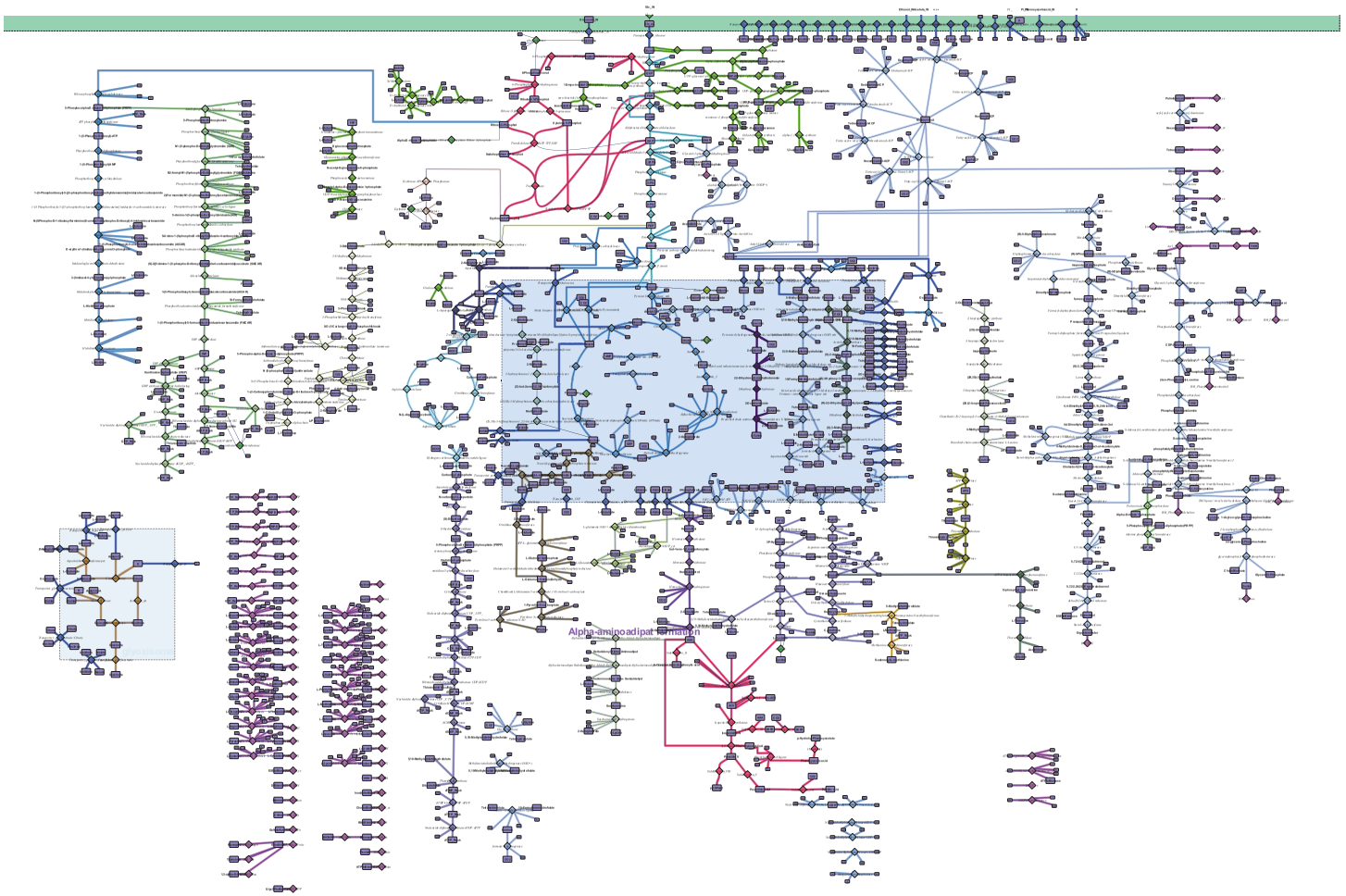


figure 5.3: Large scale metabolic network for *P. chrysogenum* including carbon atom transitions. The network contains 500 reactions and 352 metabolites, 31 pathways, four compartments. Additional reactions are added by constraint like biomass equation and fatty acid synthesis. The network is adapted for ^{13}C-MFA by removing influxes and corresponding pathways of unused substrates, thus reducing the number of reactions.

5.1.3 Focused ^{13}C-MFA Network

Focused ^{13}C-MFA networks are built for specific applications and do not rely on assumptions for cofactor balancing. Here, it was constructed based on the large metabolic network. The networks are typically focused on central metabolite pathways and product pathways resulting in a moderate size (50-150 reactions) of the final network compared to genome scale reconstructions. Linear reaction sequences and pools with rapid equilibrium are lumped.
The focused ^{13}C-MFA model accounts for 140 reactions and 76 metabolites in comparison to the large metabolic network with its 500 reactions and 352 metabolites. However, the same set of formed biomass components is modeled by both networks.
In the focused network most formation pathways of biomass effluxes were lumped. The number of reactions was significantly reduced, leaving only measured amino acids or precursors thereof and central carbon metabolism (glycolysis, citrate cycle, pentose phosphate pathway) metabolites in the network.
Nucleotides, ergosterol, and fatty acids were modeled as stoichiometric effluxes from their respective precursors. The penicillin formation from its precursors valine and cysteine is introduced as partially stoichiometric reaction system and is connected to the system by constraints.
Clearly, besides pool lumping in the biomass components, further simplification were made as stated below:

1. By-product formations were neglected in case of

    - Tryptophane synthesis: Conversion of serine to glyceraldehyde-3-phosphate and conversion of phosphoenolpyruvate to pyruvate per produced tryptophane molecule is neglected.
    - Cysteine synthesis via transsulfhydrylation pathway: the formation of by-product α-ketobutyrate and its degradation is neglected.
    - Purines and arginine synthesis: Conversion of aspartate to fumarate is neglected.

2. Simplification of pathways in case of:

    - Penicillin synthesis: Synthesis of penicillin is basically a linear pathway and the action of penicillin amidase and acyltransferase are leading to an exchange flux between 6APA and penicillin V, but are not resolvable by stationary ^{13}C-MFA as the reactions share the same atom transitions. Thus, it was modeled as a linear reaction sequence with measured effluxes to the side products of penicillin. Effluxes from valine and cysteine were constrained to this pathway.

3. Pathways were neglected in case of:

    - Glyoxylate shunt.
    - Degradation pathways, except for glycine and threonine.

4. Simplification of compartmentalization in case of:

    - Formation of glutamate: the glutamate formation is only possible from mitochondrial α-ketoglutarate, which is consistent with the result obtained by Kleijn [121].

- TCA intermediates: TCA intermediates are found in glyoxisomes, in mitochondria and in the cytosol, but were modeled to be present only in mitochondria (except oxaloacetate and pyruvate).

5. Simplification of reactions in case of:

   - Lumping the pools of bis-phosphoglycerate, 2-glycerate and 3-glycerate to the 23PG pool.
   - Cysteine pathways: both pathways were modeled as one pathway, because they share the same atom transitions.
   - Cofactor and energy balances: no cofactor and energy balances were incorporated.

The resulting network including names of reactions, atom transitions and directionality assumptions are given in the appendix C.2. During the optimization procedure, several additional assumption on directionality had to be made to prevent unfeasible flux distributions:

- Acetyl-CoA can be formed in the mitochondria and also within the cytosol. These reactions may lead to cycles. Thus, the flux of acetyl-coA from the cytosol to the mitochondria was prevented by constraining the transporter reaction to positive values ($0 \leq Transporter\_AcCoA$).

- Malic enzyme is assumed to catalyze exclusively the reaction from malate to pyruvate, as the malic enzymes possesses higher affinity towards malic acid compared to pyruvate. However, high CO_2 concentrations can reverse the reaction [228].

- Glycine can be formed by degradation of threonine catalyzed by the threonine aldolase and by the glycine hydroxymethyltransferase; both fluxes were assumed to possess an exchange flux ($0 \leq Gly1$, $Thr2 \leq 0$).

- Fluxes forming symmetric molecules lead to introduction of alternative atom transitions. The reaction of succinyl-CoA to fumarate catalyzed by succinyl-CoA synthetase is forming the symmetric fumarate molecule. This leads to the atom transition for the two scrambling reactions:

  $$\text{succinyl}-\text{CoA}\#\text{abcd} \longrightarrow \text{fumarate}\#\text{abcd}$$

  $$\text{succinyl}-\text{CoA}\#\text{abcd} \longrightarrow \text{fumarate}\#\text{dcba}$$

  Scrambling reactions are assumed to possess the same net and exchange reaction as both reaction occur with the same probability. They are found in the citrate cycle ($Tca5a - Tca5b = 0$, $Tca6a - Tca6b = 0$, $Tca7a - Tca7b = 0$).

- Anaplerotic reactions were assumed to run in the direction of ATP consumption ($Ana1 \leq gly6$, $Ana3 \leq 0$), which is consistent with textbook knowledge [173]. Nasution measured adenylate energy charge (ATP+ADP/2)/(ATP+ADP+AMP) in *P. chrysogenum* cultures, which showed high values (0.85-0.91) supporting the directionality assumption for the anaplerotic reactions [170, 168, 36].

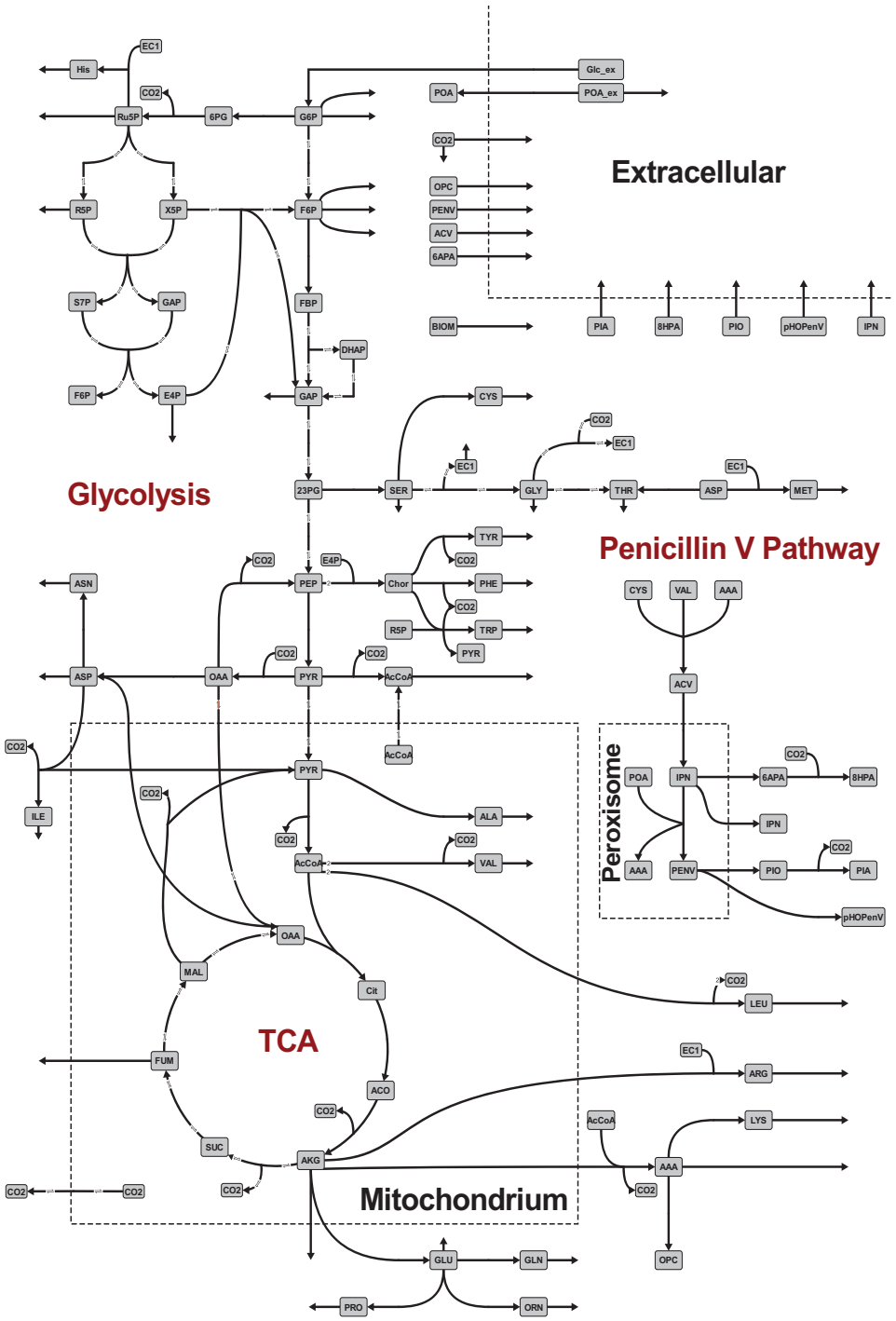


figure 5.4: Visualization of main reactions in the focused ^{13}C metabolic network. Effluxes to biomass and production of (side-)products are indicated by arrows without a product pool shown. Influxes are Glc_{ext} and POA_{ext}.

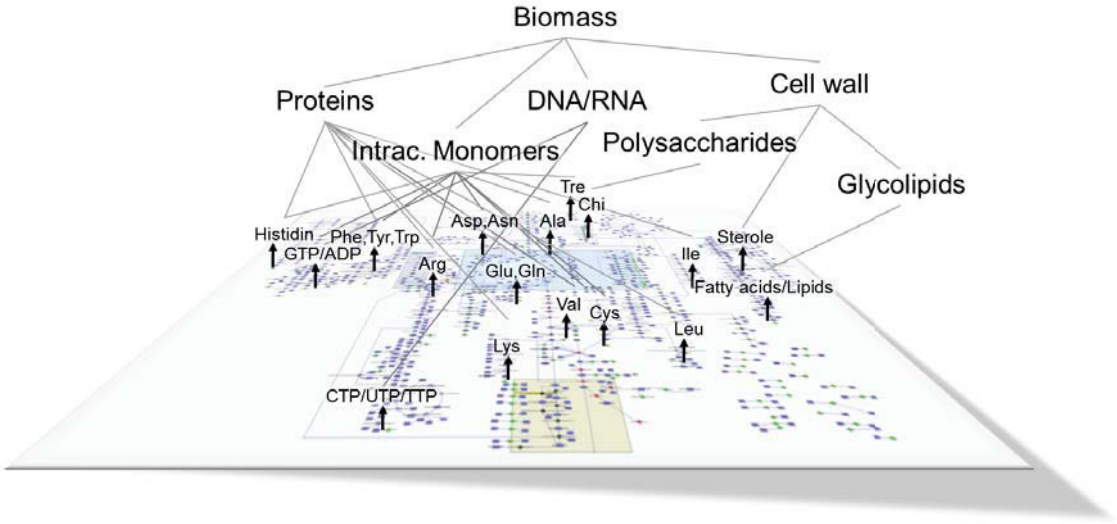


figure 5.5: Composition of biomass is shown schematically. Cellular compounds are produced from their respective precursors.

- Flux maps were found with considerably higher fluxes for phosphoenolpyruvate carboxykinase than the substrate uptake rate. A cycle was formed together with the fluxes of pyruvate kinase and pyruvate carboxylase. Reactions catalyzed by these enzymes consume ATP and, thus, these high fluxes were biologically infeasible. Parameter estimation of these fluxes for $gly7 \leq 2 \cdot gly6$ resulted in broad distribution of fluxes without clear optimum for phosphoenolpyruvate carboxykinase. These high fluxes were prevented by assuming flux for pyruvate kinase to be lower than enolase $gly7 \leq gly6$. Nevertheless, this is clearly an assumption, it reduces flux over phosphoenolpyruvate carboxykinase significantly.
  Thus, the phosphoenolpyruvate carboxykinase flux was not identifiable in the metabolism of *P. chrysogenum* under the conditions tested in this study. Besides, in yeasts it was already shown that the activity of PEP was reduced by 50-100 % when glucose was used solely as carbon source compared to growth on ethanol [87]. A similar regulation was also found in *Aspergillus nidulans* [105].

By now only the metabolic network including directionality assumptions was described. In the next section, the measurement data of the biomass components were used to specify intracellular effluxes for the formation of biomass.

5.1.4 Modeling of Biomass Effluxes

Each cell consists of highly diverse substances, ranging from small molecules (e.g. ions, precursors for amino acids, nucleotides, ...) to macromolecular constituents (e.g. proteins, cell wall components, carbohydrates). The sum of these compounds is referred to as biomass (see figure 5.5). All substances containing carbon have to be taken into account for ^{13}C-MFA to ascertain a closed carbon balance and to enable the modeling of the fate of the carbon atoms from substrate to product without bias. In *P. chrysogenum*, up to ~66 % of the glucose carbon can be found in the biomass ($Y_{SX}^{max} = 0.66\ \mathrm{mol_C\,mol_C^{-1}}$ taken from van Gulik *et al.*) [243].

To achieve correct modeling of the flux distribution, the strain-specific biomass composition of *P. chrysogenum* BCB1 was measured for a growth rate of $0.03\,h^{-1}$. Its biomass was hydrolyzed and the fractions of the separate hydrolysis products were analyzed, which resulted in a set of measurements describing the main building blocks of the biomass. To calculate the effluxes towards the biomass, the building blocks measured after hydrolysation (glucose, galactosamine, ...) need to be assigned to their macromolecular compounds (glucan, chitin, galactofuranan, ...) from which they were formed by hydrolysis. A number of compounds were not measured/measurable and were corrected by literature values (see below).

The elemental composition of the strain BCB1 ($\mu = 0.03\,h^{-1}$) was measured to be

$$CH_{1.734}O_{0.590}N_{0.162}S_{0.005} \text{ with an ash content of } 2.72\,\%.$$

Thus, 45.6 (w/w) % of the measured biomass is composed of carbon, whereas about 6.5 % of the biomass consists of pure water. Summing up the biomass constituents measured from the hydrolysis experiment and added data from literature results in 92.8 (w/w) % closed mass balance. The main biomass constituents are listed in table 5.3.

In literature, further biomass constituents are described for *P. chrysogenum*. To be comparable, these compounds are added, as they are important constituents of the cell wall. Choline (0.0224 $\text{mmol}\,g_{CDW}^{-1}$), and inositol (0.0157 $\text{mmol}\,g_{CDW}^{-1}$) were added according to values by Nielsen [174]. The ergosterol content is based on van den Berg *et al.* [240] 0.0038 $g_{CDW}\,g^{-1}$ for strain Wisconsin54-1255. These compounds account for 0.9 (w/w) % of the biomass. The measured glutarate and malonate (assigned to 3.4 % "Rest" in table 5.3) could not be assigned to any biological constituent or biosynthetic pathway (they comprise 0.3 % of the biomass and were removed for the final biomass equation). Chitin was not directly determined, but instead the overall content of glucosamine, which was assumed to be derived from chitin. Acetyl residues of chitin were removed by hydrolysis of the biomass and were also not measured. However, values for acetyl-CoA and glucosamine were added to the biomass.

In table 5.3 the biomass composition of *P. chrysogenum* obtained including correction by literature values in this study is compared to Nielsen's and Jørgensen's data [174, 111]. Nielsen conducted experiments in chemostats, whereas Jørgensen in fed-batch. DNA and chitin content were strongly reduced and the carbohydrate content elevated in *P. chrysogenum* BCB1 compared to literature data. Thus, strong deviates in macromolecular constituents of the biomass composition are detected compared to literature data.

Comparison of the relative molar fractions of fatty acids found in *P. chrysogenum* BCB1 during this study to the results of Mumma *et al.* showed that the data-sets are consistent, except for linolenic acid and linoleic acid [166]. These fatty acids show a lower and accordingly higher amount in the data measured by Mumma. The fractions of the fatty acids for BCB1 strain (values by Mumma *et al.* given in brackets): 0.2 % myristic acid (0 %), 15.5 % (12.2 %) palmitic acid, 0.4 % (0 %) palmitoleic acid, 2.3 % (5.5 %) stearic acid, 6.4 % (10.9 %) oleic acid, 73.8 % (65.4 %) linoleic acid and 1.4 % (6.0 %) linolenic acid.

Amino acid composition of the *P. chrysogenum* strain BCB1 used in this study was also measured. Tryptophane and alanine could not be measured and were assigned based on Nielsen's data [174]. The separate fractions of amino acids contributing to the biomass are shown in table 5.4.

table 5.3: Biomass composition obtained for *P. chrysogenum* during this study in comparison to the biomass composition for *P. chrysogenum* determined by Nielsen and Jørgensen. The biomass was corrected for its water content of 6.5 %. Protein content was assumed to be 44.5 %, based on previous accomplished protein measurements with the strain and taking into account literature data [174]. Glucosamine measurement was evaluated as chitin. Nielsen obtained the data from *P. chrysogenum* cultivations showing growth rates of 0.022-0.066 h^{-1}. Jørgensen data are generated from three phases of a fed batch process.

	This study [(w/w)%]	Nielsen [174] [(w/w)%]	Jørgensen [111] [(w/w)%]
RNA	5.0	4.3-6.1	3.0-7.5
DNA	0.2	0.9-1.1	0.5
Proteins	44.5	39-45.3[1]	48-50.0[1]
Carbohydrates	34.6	18.2-27	23-30
Chitin	1.9	4.7-7.5	2-4.0
Lipids	2.7	3.2-5.5	4-5
Rest	3.4	14.9-20.4[1]	9.5-12[1]
$\sum$	93.2	100	100

[1] In case of Nielsen and Jørgensen only proteinogenic amino acids are determined, whereas the free non-proteinogenic amino acids are described as "Rest" in their data. In contrast, all amino acids are considered in this study. For BCB1 Rest accounts for organic acids, inositol, ergosterol, anions/ash.

Focused ^{13}C-MFA Network

Finally, the biomass composition was used to specify effluxes towards the biomass components listed in table 5.4. The portion of aspartate and asparagine as well as of glutamate and glutamine in the biomass were measured as summed peaks. To obtain the amount of every single amino acid it was split, e.g. 50 % accounted for glutamine and 50 % for glutamate. (Desoxy-)Nucleosides of RNA and DNA were incorporated in the focused ^{13}C-MFA model as effluxes from their precursors. (Desoxy-)Nucleosides are formed from ribose-5-phosphate, C1 compounds are delivered by folate, aspartate, glycine, and carbon dioxide according to reaction in the large scale metabolic network. Glycerol was formed from glyceraldehyde-3-phosphate.
The largest efflux towards biomass components (see table 5.4) is withdrawn from the glucose pool, followed by flux into acetyl-CoA. Acetyl-CoA efflux accounts for fatty acid, ergosterol and chitin synthesis. The overall sum of carbon per gram cell dry weight was calculated to be 35.36 $\mathrm{mmol_C\,g_{CDW}^{-1}}$, resulting in a fraction of 42.47 (w/w) % carbon per gram cell dry weight. Elemental measurements resulted in 45.6 (w/w) %, thus a 93.1 % closed carbon balance is obtained. For flux estimation in the ^{13}C-MFA metabolic network, the effluxes were scaled to obtain 100 % closed carbon balance. The biomass composition was used as an equation to determine effluxes to biomass, the fractions in table 5.4 are used as stoichiometric coefficients to yield 1 g of biomass. The measured growth rate was introduced into the ^{13}C-MFA model as measurement.

Large Scale ^{13}C-MFA Network

For the large scale metabolic network, the biomass composition is listed in table C.1 in the appendix. The fractions of glycogen, trehalose or glucan formed from glucose were not determined. Calculating back to the specific compounds was, thus, not possible. In this case

table 5.4: Biomass equation for the focused ^{13}C-MFA model of *P. chrysogenum*

Compound	Flux		Compound	Flux	
	[mmol g_{CDW}^{-1}]	[$mmol_C$ g_{CDW}^{-1}]		[mmol g_{CDW}^{-1}]	[$mmol_C$ g_{CDW}^{-1}]
Aspartate	0.269	1.08	Proline	0.118	0.59
Asparagine	0.182	0.73	Tryptophane	0.111	1.22
Glutamine	0.262	1.31	Alanine	0.279	0.84
Glutamate	0.262	1.31	Acetyl-CoA	1.047	2.09
α-Aminoadipate	0.028	0.17	Erythritol	0.016	0.06
Serine	0.286	0.86	Arabitol	0.010	0.05
Histidine	0.074	0.44	Mannitol	0.358	2.15
Glycine	0.330	0.79	Glucosamine	0.095	0.57
Threonine	0.218	1.09	Galactose	0.297	1.78
Arginine	0.198	1.19	Glucose	1.087	6.52
Tyrosine	0.116	1.04	Mannose	0.196	1.17
Cysteine	0.004	0.02	Fumarate	0.001	0.00
Valine	0.288	1.44	Isocitrate	0.001	0.00
Methionine	0.062	0.31	EC1	0.202	0.20
Phenylalanine	0.150	1.35	R5P	0.155	0.78
Isoleucine	0.195	1.17	CO_2	0.155	0.16
Ornithine	0.004	0.02	GAP	0.143	0.3
Leucine	0.292	1.75	$\sum$	-	35.36
Lysine	0.177	1.06			

the glucose units needed to build the separate compounds were summed to yield the biomass efflux to glycogen, trehalose and glucan. So each compound can be formed and the correct amount of glucose is used for them, nevertheless, the separate fractions (of glycogen, trehalose and glucan) are set free to be chosen within flux estimation procedure. The same was done for the composition of cell membrane constituents (only fatty acids, inositol, choline, ergosterol and glycerol production is fixed by the biomass). By adding these compounds, the carbon balance for this biomass equation is closed with 91.9 % closed carbon balance.

table 5.5: Maximal theoretical yield per gram substrate calculated by FBA model

Substrate	$Y_{XS,max}$[g g^{-1}]	$Y_{PS,max}$[mol$_C$ mol^{-1}]	$Y_{PS,max}$[mol mol^{-1}]
Glucose	0.48	0.071	0.45
Ethanol	0.72	0.077	0.16
Acetate	0.32	0.066	0.10
Gluconate	0.38	0.046	0.39

5.2 Flux Balance Analysis

Before conducting ^{13}C-MFA, *in silico* results for the metabolic capabilities of the organism are generated based on the large metabolic network. Additionally, measurements are available for the biomass composition and metabolic rates. Using these experimental data, estimates for productivity of the high-producing *P. chrysogenum* strain BCB1 can be computed.

Still, estimates for growth associated and non-growth associated ATP requirements are needed to obtain a reasonable model describing the metabolism of *P. chrysogenum*. Because limited data are available about strain BCB1, growth associated ATP requirements are taken from literature data. A growth associated ATP demand of of 108 $\mathrm{mmol_{ATP}\,g_{CDW}^{-1}}$ was published in the genome scale model of *P. chrysogenum* [2]. This is twice as high as in the recent yeast genome scale reconstruction (version 6.05; 59.3 $\mathrm{mmol_{ATP}\,g_{CDW}^{-1}}$) [95]. The non-growth associated maintenance coefficient was taken from literature to be 1.2-3.4 $\mathrm{mmol_{ATP}\,g_{CDW}^{-1}\,h^{-1}}$ as lower and upper boundaries [243].

To compare the model to literature data, uptake and metabolism of the mainly used substrates (acetate, ethanol, glucose and gluconate) were incorporated. First, the maximal theoretical biomass yields on these substrates are calculated by flux balance analysis(see table 5.5). The maximal yield on glucose was calculated to be 0.48 $\mathrm{g_{CDW}\,g_{glc}^{-1}}$. In literature, diverging values were found ranging from 0.35-0.61 $\mathrm{g_{CDW}\,g_{glc}^{-1}}$ but the maximal theoretical yield was calculated by van Gulik *et al.* to 0.61 $\mathrm{g_{CDW}\,g_{glc}^{-1}}$ (see table 5.6). For ethanol a theoretical maximum biomass yield of 0.72 $\mathrm{g_{CDW}\,g_{eth}^{-1}}$ was computed compared to 0.66 $\mathrm{g_{CDW}\,g_{eth}^{-1}}$ found by van Gulik *et al.* [243]. Acetate yield was 0.32 $\mathrm{g_{CDW}\,g_{ac}^{-1}}$, whereas an experimentally reported value from literature was 0.39 $\mathrm{g_{CDW}\,g_{ac}^{-1}}$. Gluconate showed yields below glucose but comparably higher than acetate of 0.38 $\mathrm{g_{CDW}\,g_{glt}^{-1}}$. However, as the model possesses a strain specific biomass composition, the data are in good agreement with literature values.

The mentioned substrates will also influence penicillin production. Ethanol is again performing best with 3.4 $\mathrm{mmol_{pen}\,g_{eth}^{-1}}$. Besides, it provides also highest C-atom efficiency, as most of the molecules carbon atoms are incorporated in the product. Glucose resulted in 0.45 $\mathrm{mmol_{pen}\,g_{glc}^{-1}}$ and gluconate is forming only 0.39 mol product per mol substrate. The experimentally obtained maximal theoretical production of penicillin by van Gulik *et al.* is reduced to ~40 % compared to *in silico* value. Gluconate results in a loss of productivity of roughly 20 % compared to glucose.

Based on the metabolic network a yield analysis was performed according to Song and Ramkrishna [224]. Here, the possible yield of the main products biomass and penicillin are evaluated based on the large FBA model, the result is shown in figure 5.6. In the figure a yield analysis for each substrate (glucose, ethanol, acetate and gluconate) is plotted, as well as the results for an optimal mixture of all substrates. In literature, several data-sets have been published for production of penicillin under controlled conditions and rate estimates are given. In

table 5.6: Experimentally determined literature values taken from literature for cultivations of *P. chrysogenum*.

Conditions	μ [h^{-1}]	Y_{XS} [$g_{CDW}\,g^{-1}$]	Y_{PS} [$mol\,mol^{-1}$]	m_s [$g\,g_{CDW}^{-1}\,h^{-1}$]	Comment	Citation
Chemostat	0.06	0.56	-	-	glucose	[41]
Chemostat	0.06	0.38	-	-	glucose	[41]
Chemostat	0.07	0.43	-	-	glucose	[41]
Chemostat	0.08	~ 0.35	-	-	glucose	[40]
Chemostat	~ 0.01-0.125	~ 0.61[2]	0.174[2]	~ 0.096[1]	glucose	[242, 243]
Chemostat	~ 0.03-0.09	~ 0.66[2]	0.068[2]	~ 0.006[1]	ethanol	[242, 243]
Chemostat	~ 0.03-0.09	~ 0.39[2]	0.040[2]	~ 0.0125[1]	acetate	[242, 243]
Chemostat	0.025-0.18	0.51	-	0.028	glucose	[44]
Chemostat	$\leq$ 0.025	0.45	-	0.024	lactose, pharmamedium	[204]
Chemostat	$\leq$ 0.075	0.45	-	0.022	glucose	[201]
Chemostat	-	~ 0.59 [1]	-	~ 0.002-0.007 [1]	glucose	[100]
Fed-Batch	$\leq$ 0.1	-	0.052-0.085	0.025	corn steep liquor, sucrose	[111]

[1]0.435 g_C per g_{CDW} assumed
[2]Here maximal yield coefficients are given: Y_{XS}^{max} and Y_{PS}^{max}

these datasets several different species of penicillins and side-products were formed, because penicillin is also degraded constantly to some of these side-products the amount of formed penicillin nucleus (i.e. 6APA) is used to compare the data-sets. First of all, the product yield of nearly all production processes is weak and a high biomass production regime is favored. For most of the data-sets, the major amount of carbon is not used for product or biomass, being far away from the diagonal lines, which would be optimal. The reason for this is not clear, yet. For the production process in this work based on glucose as substrate, the experimental yield is roughly 25 % of the maximum predicted yield and the biomass production is considerably lower compared to other literature data.

Finally, the penicillin yield obtained with strain BCB1 is high compared to literature data, providing evidence for a high producing strain with comparably high carbon efficiency towards the product with a low biomass yield. Another observation can be made: Most cultivations in literature are using glucose with addition of other substrates indicated by markers in figure 5.6.

Influence of Pentose-Phosphate Pathway on Penicillin Yield

For penicillin production, NADPH is needed in the cytosol for synthesis of cysteine and valine. The oxidative pentose phosphate pathway (oxPPP) and isocitrate dehydrogenase are main NADPH producing reactions. Based on the large scale metabolic network, the theoretical penicillin production can be calculated from the provided extracellular rates for glucose uptake and taking also into account biomass production. The substrate uptake rates and biomass formation of the fed-batch process are used. For this, the penicillin production is maximized by changing flux through oxPPP. The split ratio of oxPPP and glycolysis is computed by the flux through oxPPP divided by the uptake rate of glucose. The results

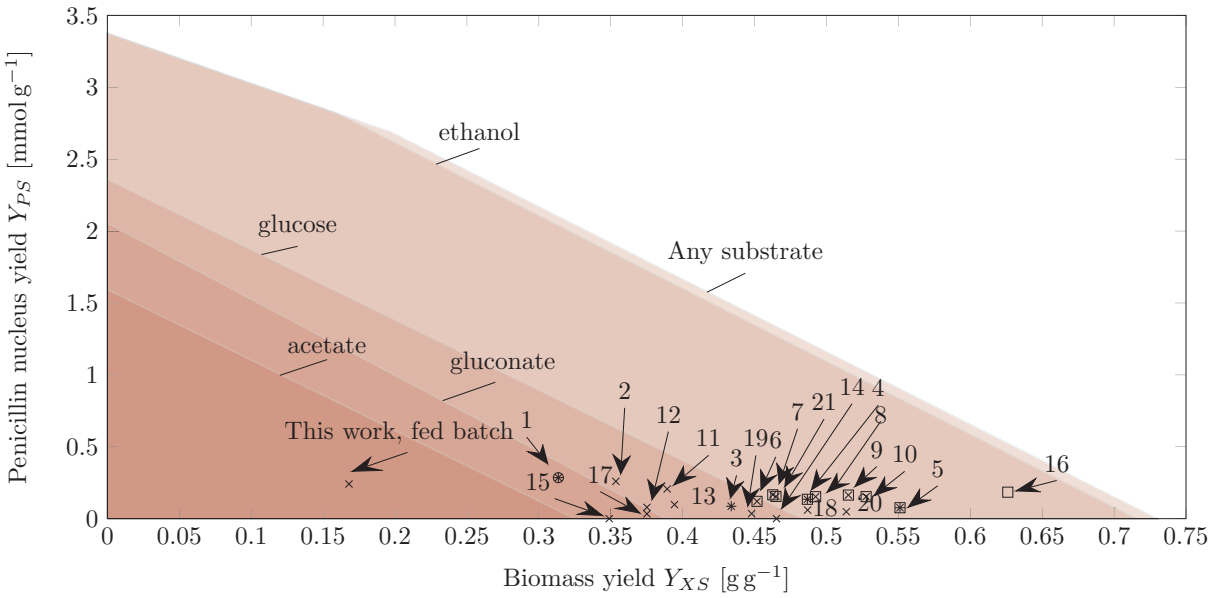


figure 5.6: Results of yield analysis for the large metabolic network for *P. chrysogenum* and published data in literature. Frequently applied substrates in literature: glucose (×), ethanol (□), gluconate (+), and acetate (○). Numbers in the figure indicate data-sets takenn from literature. 1: Kleijn [121], 2: van Gulik et al. [242], 3-10,21: Zheng [284], 11-12: van Winden et al. [247], 13-16: Nasution et al. [170], 17-20: Christensen et al. [44]. Side-products containing formed 6APA were added to penicillin nucleus production.

can be found in figure 5.7. A maximal production of 1.5 $\mathrm{mmol_{pen}\,g^{-1}}$ ($0.27\mathrm{mol_{pen}\,mol_{glc}^{-1}}$) was found for the fed-batch process. The action of cytosolic NADPH producing isocitrate dehydrogenase is beneficial if low fluxes ($< 90\,\%$) through oxPPP are observed. The isocitrate dehydrogenase is usually assumed to produce the precursor α-ketoglutarate for glutamate, arginine, lysine, glutamine and proline. Additionally, it produces NADPH needed for these mainly formed amino acids [283].

Higher fluxes/split ratios through oxPPP than 100 % are possible, as glucose-6-phosphate isomerase flux can be reversed. This results in split ratios up to 450 %. The production of penicillin is optimal in the regime of 60-120 % split ratio. If cytosolic isocitrate oxidoreductase is assumed to play no important role, roughly 90-110 % split ratio are optimal.

To summarize, high flux through oxPPP is optimal for penicillin production. The action of oxPPP can be positively accompanied by cytosolic NADPH producing cytosolic isocitrate dehydrogenase. The role of isocitrate dehydrogenase cannot be elucidated by later conducted stationary ^{13}C metabolic flux analysis as it shares the same atom transitions with the reaction in citrate cycle. Further experimental data need to be gathered to elucidate the role of cytosolic isocitrate dehydrogenase for the formation of penicillin. A direct link between the formation of penicillin and the activity of the isocitrate dehydrogenase has not been given in literature, yet.

As the conclusion, FBA predicts significantly higher production of penicillin V compared to experimentally found values, although biomass yields were adequately predicted. The formation of penicillin V is positively effected by flux through the oxPPP. Besides, the fed-batch

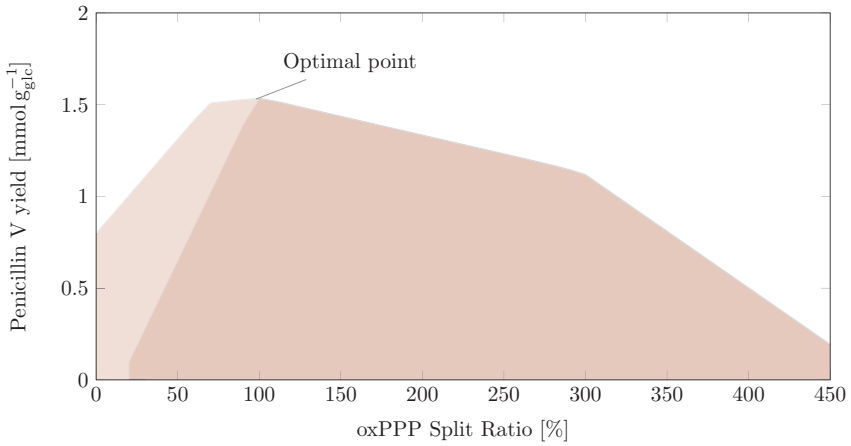


figure 5.7: Flux balance analysis conducted to estimate the production of penicillin in the fed-batch 1 in dependence of the flux through oxPPP. Growth on glucose with $0.251\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$ and growth rate of $0.008\,\mathrm{h^{-1}}$. Because of biomass effluxes, glucose-6-phosphate isomerase reaction is reversed at a split ratio of ~95 %. Including isocitrate dehydrogenase ( ) and without isocitrate dehydrogenase ( ), flux was set to zero.

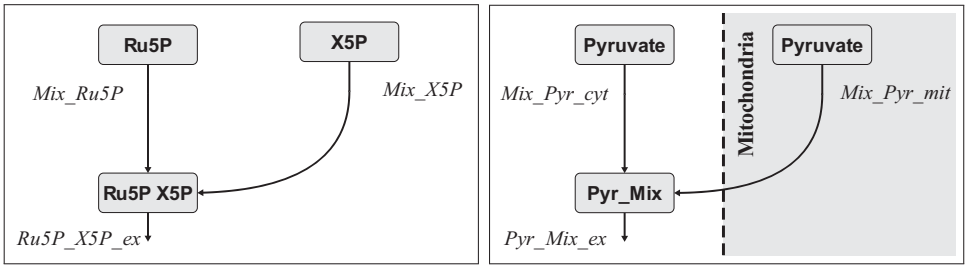
process for penicillin production shows significantly reduced biomass growth, but high penicillin and side-product formation compared to published literature data. Why FBA cannot adequately predict penicillin production, will be discussed in chapter 8.

5.3 Measurement and Input Substrate Specification for ^{13}C-MFA

For the first carbon labeling experiment later conducted, a 60 % 1-, 20 % uniformly, and 20 % naturally labeled glucose mixture was chosen, according to Kleijn's work [121]. For the generated networks 429 MS measurements are possible by LC-MS/MS. These measurements can be divided into 50 measurement groups, each an LC-MS/MS measurement of a specific metabolite.

Modeling of Labeling Measurements

Most labeling measurement can be directly incorporated into the ^{13}C-MFA model, because it contains the respective measured metabolites. Several metabolites' labeling measurements could not be separated upon detection because they show the same retention time and the emerging mother and daughter ions share equal masses. This was found for ribulose-5-phosphate and xylulose-5-phosphate, they were measured as summed peaks. To introduce them into the model, it was assumed that both metabolites can contribute to different extends to the measurements. An additional pool in the model and mixing of both metabolites using unidirectional fluxes in this pool, see figure 5.8a. An efflux out of the pool was set to a low value in order to prevent perturbations of the model ($10^{-5}\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$). Thus, the measurements were accurately modeled.
The *P. chrysogenum* model does also contain compartments. A number of metabolites can be found in several compartments (e.g. pyruvate, asparagine, glutamate, citrate, cis-aconitate,



(a) Summed Peak Measurements (b) Compartalization and Measurement

figure 5.8: Modeling of labeling measurements for (a) summed peak measurement and (b) pools subject to compartmentalization. The effluxes of *Pyr_Mix_ex* and *Ru5P_X5P_ex* were constrained to very low flux values. The labeling measurement is assigned to the *Pyr_Mix* and *Ru5P/X5P* pool.

oxaloacetate,...). Thus, their measurements are also sums of pools which (potentially) deviate in labeling enrichment. If measurements are specified for those metabolites, the above described method was likewise applied (see figure 5.8b). By applying this method, compartment specific contribution to labeling measurements by different pool sizes is possible.

5.4 Results

In this chapter the model construction process assisted by usage of literature, data bases (KEGG), genome scale metabolic network reconstructions, and localization prediction tools was described. The work-flow was carried out manually in OMIX [60]. From the large scale metabolic network comprising 500 reactions, a focused network was constructed. Assumptions for this were discussed. Additionally, a biomass equation describing effluxes to biomass was calculated from experimentally determined values. Finally, input substrate and possible measurement specifications were introduced, according to known analytical standard-operation-procedures.
Up to this point, a complete ^{13}C metabolic model was constructed. For a chosen set of free fluxes, forward simulation can be performed to obtain the labeling measurements. This unlocks theoretical investigation based on the model without introduced wet-lab data for labeling measurements. In the next chapter, the model is used to perform global sensitivity analysis.

Chapter 6

Global Sensitivity Analysis

Sensitivity analysis reveals insights into the interplay of measurable quantities (flux and labeling measurement) and the variables of the system (free fluxes and input substrates). This chapter aims at identifying and quantifying important fluxes influencing the model output, thereby preventing over-interpretation of the results of parameter estimation.
For this, the suitable method (high dimensional model representation) was selected and established. Here, for the first time, global sensitivity analysis is performed for ^{13}C-MFA and developed visualization of sensitivity measures assists its interpretability. Finally, by the chosen method, direct assessment and visualization of the interaction between fluxes and measurements is possible and can be used for model-based interpretation of labeling data.

Note that there is a change in nomenclature, x is no longer denoting the labeling pattern in this chapter, but the model input parameters.

6.1 Introduction

Traditional sensitivity analysis is a frequently applied method for evaluating the input/output behavior of models. In this method, changes in function values $y(\mathbf{x}^*)$ are observed by calculating the derivative of the model output with respect to the input variables $\mathbf{x}$ (in the following called "input") at a specified point $\mathbf{x}^*$:

$$\left|\frac{\partial y}{\partial x_i}\right|_{\mathbf{x}*}$$

Because the derivative is calculated at a single point in input space, the analysis is called *local*. This technique is applied in most studies, because only a small number of model evaluations is needed to obtain dependencies of the input and output of a model. This results in a sensitivity matrix for one point in input space. The method is applicable to (almost) linear systems; however, the subject of this study are nonlinear models and the resulting local sensitivity matrices may change if the same analysis is performed for multiple points in input space [221].
Global sensitivity analysis too aims at quantification of inputs‘ effect on the output of the model [207]. It is examining the global response by varying all inputs in a large range, not one at a time (as in local sensitivities). The general idea of most methods is that a number of data points in space contain more information than a local derivative.

This method is important in order to prevent errors in the model building process. Based on insensitive parameters, later, wrong conclusions can be drawn, because they are not supported by the used data. Additionally, sensitivity analysis is conducted together with uncertainty analysis. Uncertainty analysis is also taking into account the uncertainty of the measurements used for parameter estimation [207]. Uncertainty analysis is not subject of this chapter.

^{13}C-MFA - Requirements for Sensitivity Analysis

In ^{13}C-MFA high dimensional metabolic networks are built, whose fluxes are inferred by measuring direct information (extracellular fluxes) and indirect information (labeling measurements). The model outputs (the labeling information) show non-linear behavior with respect to the model inputs (free fluxes). Thus, sensitivities between the measurement and the flux values cannot be calculated globally *ad hoc*.
Assumptions for ^{13}C-MFA models are (cf. Spiral model chapter 2.2.3): (**I**) we are dealing with smooth functions for labeling measurements and (**II**) no rapid changes in the function values are observed within a ^{13}C-MFA model.
The characteristics of the input-output relation in the context of ^{13}C-MFA are summarized:

1. The models are high dimensional with 20-40 inputs (free fluxes) and possess several model outputs (measurements). Thus, it is a Multi-Input-Multi-Output (MIMO) system comprising 100-600 outputs $y^i(\mathbf{x})$, $i = 1, 2, ..., n$.

2. The flux values are constrained by lower/upper bounds and the stoichiometric matrix S. Due to the concept of free fluxes, the basis of the null space is not orthogonal.

For these models the computational cost of a simulation is low by employing the high performance simulator 13CLFUX2 [260]. Next, a short overview will be given about methods used for global sensitivity analysis in literature. Afterwards, an suitable method is chosen for application to ^{13}C-MFA.

6.1.1 Methods for Global Sensitivity Analysis

A multitude of methods has been published for sensitivity analysis. A selection of frequently applied methods is shown in table 6.1.

Visual Methods First, scatter plots are a graphical sensitivity analysis method. Here, pairs of model outputs and inputs are plotted. By examining the shape of the point clouds, dependencies of input and output can be detected. Naturally, this method is not applicable to a high number of inputs as it becomes tedious to examine all plots and their interpretation is only qualitative. Thus, quantifying the sensitivity of a single parameter is not possible [86].

Screening Methods are low cost methods for models with a large number of input parameters. They rank the parameters according to their importance. However, usually they do not give absolute quantities for the influence of an input.
For example, fractional factorial designs are known from experimental design studies by response surface techniques. Here, sample points are taken from the corners of a hypercube in input space [207]. A second screening technique called *elementary effect* is based on a

table 6.1: Sensitivity analysis methods (adapted and extended from [207]).

	Samples from	# Input	Advantages	Disadvantages	Literature
Scatter plot	Distribution	<10	• Fast • Visual	• Qualitative • Tedious	[207, 86]
Fractional factorial	Corners of hypercube	>100	• Moderate sample size	• Sampling on hypercube corners • Sampling in unit cube needed	[207]
Elementary effects	Trajectory on levels	20-100	• Moderate sample size	• Not resolving higher order interactions • Orthogonal inputs needed	[205, 33, 127]
"Classical" variance-based	Distribution	<20	• Quantification of higher order terms (total sensitivities) • Several techniques • Assigning of variance to inputs	• High sample number for purely variance based techniques	[208, 220, 194, 206]

similar sampling method. The inputs are changed at specified trajectories by a fixed level, where the level step size is in the order of the uncertainty of the system [127]. The sensitivity of the input is measured as the change per step from one to the other level. However, as these techniques change one factor at a time, they cannot distinguish between low and high order interactions of inputs [127]. Additionally, the elementary effects method becomes inefficient in higher dimensions (curse of dimensionality). Grid-based methods are also not reliable if the function under investigation is only changing in-between the grid points [205, 33]. Similar techniques rely on sampling radially from a central point or using star shaped designs, but rely only on estimating the changes produced by one factor at a time [33, 191].

Variance-Based Methods measure the variance of the model output. This offers advantages compared to the afore mentioned techniques. First, the variance is a model independent measure and it can quantify the variation of each input factor. Additionally, interaction effects can be described by their exhibited variance. In these methods it is possible to assign the variance of the output to the influence of the separate input parameters. These methods are usually applied to models with a medium number of inputs, because with rising dimensionality the number of necessary samples increases. Using 10-20 input factors results in several thousand samples needed, thus variance-based methods are more costly than elementary effects [207].

A growing number of techniques relying on variance-based methods have been proposed, e.g. eFAST (extended Fourier Amplitude Sensitivity Test) and Sobol's method [233, 220]. They are usually more efficient (need less samples) than purely variance-based techniques mentioned before.

Meta-Modeling Techniques are employed for high number of inputs (20-100) [207]. These techniques (usually) employ smoothing operations to obtain measures of sensitivity [194, 207, 198]. They have moderate requirements of sample size compared to variance-based techniques. However, higher order interactions of inputs are often not computable because they are computationally demanding. Two methods were found to be suitable for conduction of sensitivity analysis for models with non-orthogonal inputs: (**I**) state dependent parameter (SDP) meta-modeling and (**II**) random-sampling high dimensional model representation (RS-HDMR) [198]. However, these meta-modeling techniques can also rely on assigning the variance of the output to the input parameters and provide, thus, adequate measures for sensitivity.

6.1.2 Method-Selection

For ^{13}C-MFA models the method has to cope with a high number of inputs, and a non-orthogonal input space. Additionally, computation has to be fast and efficient. Thus, only meta-modeling techniques RS-HDMR and SDP modeling were left. The SDP implementation of Ratto was tested (SS-ANOVA-R toolbox), but was slow and memory consuming for increased sample size and input numbers (beyond 10-20 inputs and 5000 samples) [198]. Finally, RS-HDMR was chosen and implemented. The basic principle of RS-HDMR will be introduced next.

6.1.3 Random-Sampling High Dimensional Model Representation (RS-HDMR)

The RS-HDMR provides measures for sensitivity based on variance decomposition. It builds a surrogate model (i.e. it is a meta-modeling technique), which can substitute the original model (a so called meta-model). RS-HDMR decomposes model outputs into separate low dimensional functions. These functions are called component functions. They reflect the effect of a subset of inputs contributing to the output. Single and cooperative influence of two or more inputs on the output of the model can be reflected by these component functions. The single functions can be determined by parametric or non-parametric regression techniques. Random samples in input space are used for this procedure. By utilizing the separate component functions a meta-model is built. Finally, information about the sensitivity can be extracted from the meta-model. These measures are called sensitivity indices. This is done by extracting the component functions‘ variance and scaling it with the overall model outputs‘ variance.
In figure 6.1 the construction process of an RS-HDMR meta-model is shown. First, using random samples, the output of the function is evaluated. Second, the zeroth order component function f_0 (a scalar) is calculated (basically the mean of the sample points). Next, the separate component functions are built from the distribution of sample points, here only component functions of first order for each parameter are shown; for parameter 1 it is given by $f_1(x_1)$ and for parameter 2 it is given by $f_2(x_2)$. They describe the effect of one input parameter acting on the output of the model. Finally, the meta-model $f(\mathbf{x})$ is built from the separate component functions $f(\mathbf{x}) = f_0 + f_1(x_1) + f_2(x_2)$. From these component functions the contribution of each parameter to the output can be extracted by estimating their variance.

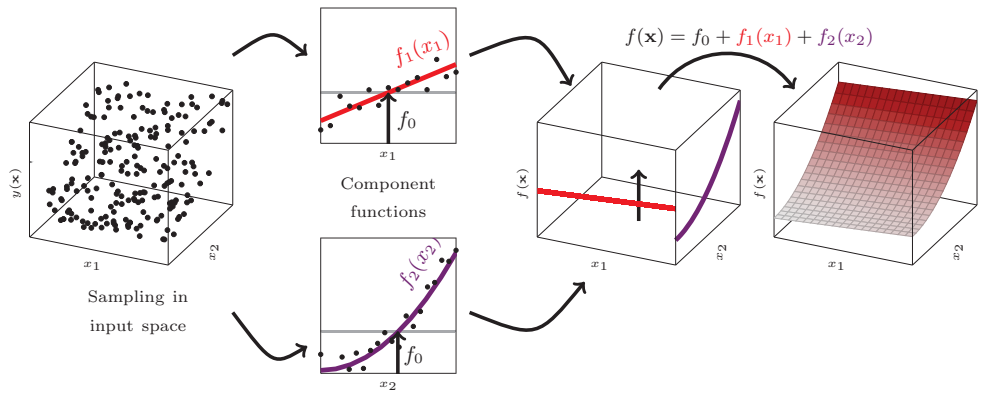


figure 6.1: Basic idea of RS-HDMR, picture provided by Samuel Leweke.

The RS-HDMR method provides several advantages compared to other methods:

- A moderate sample size is required (5,000-10,000 samples are used) [257].
- Independent and/or correlated inputs are usable [137].
- Estimation of second-order and higher order correlation terms is possible.

- Low computational cost (only vector dot products need to be evaluated for most techniques) and parallelizable.

The method is not capable of computing global sensitivities for all model outputs at once. For each output a separate meta-model has to be constructed. Out of all meta-models important information needs to be extracted and visualized to give the modeler a good overview.

Next, the basics of HDMR are introduced and later applied to ^{13}C-MFA.

6.2 Theory

The principle of HDMR is explained next. The RS-HDMR method is derived from it. Afterward, the measures for sensitivity are introduced.

6.2.1 HDMR in General

High dimensional model representation (HDMR) is based on Sobol's ANOVA (Analysis Of Variances) decomposition [220, 221]. It decomposes the output of a model $y(x)$ for a given number of p inputs $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ into a hierarchical function expansion:

$$y(\mathbf{x}) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{i,j}(x_i, x_j) + \ldots + f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p) \tag{6.1}$$

Where f_0 is a constant term. $f_i(x_i)$ are the first order component functions. They describe the effect of one input i to the output of the model. The second order component functions $f_{ij}(x_i, x_j)$ reflect cooperative effects of two inputs at a time. The last term $f_{(1,2,\ldots,p)}(x_1, x_2, \ldots, x_p)$ is the highest order cooperative contribution to the model output. The input and component functions are defined in the interval $0 \le x_i \le 1$, thus $K^n = \{x_1, x_2, ...x_n\} | 0 \le x_i \le 1$. Each component function possesses the property that its integral with respect to its own variables x is zero:

$$\int_0^1 f_{i_1 i_2 \ldots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) dx_s = 0 \quad \text{for each } s \,\epsilon\, \{i_1, i_2, \ldots, i_k\} \tag{6.2}$$

Furthermore, the orthogonality of two functions is assumed to be preserved:

$$\int_{K^n} f_{i_1 i_2 \ldots i_l}(x_{i_1}, x_{i_2}, \ldots, x_{i_l}) f_{j_1 j_2 \ldots j_k}(x_{j_1}, x_{j_2}, \ldots, x_{j_k}) d\mathbf{x} = 0 \ (\{i_1, i_2, \ldots, i_l\} \neq \{j_1, j_2, \ldots, j_k\}) \tag{6.3}$$

In HDMR, the above mentioned equation 6.1 is simplified. Several empirical studies have showed that some high dimensional models can be described by low order cooperative functions [135, 137]. Thus, HDMR uses only functions up to second or third order to describe $y(x)$. The HDMR contains only a truncated expansion:

$$y(\mathbf{x}) \approx f(\mathbf{x}) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{ij}(x_i, x_j) + \sum_{1 \le i < j \le k \le p} f_{ijk}(x_i, x_j, x_k)$$

Third order and higher order correlation are scarcely used, as they will introduce a high computational demand. The component functions in RS-HDMR are usually built by usage of Monte Carlo integration. For this, several techniques are applicable for regression/smoothing, e.g. wavelets, splines, polynomials, orthonormal polynomials, and fixed interval smoothing [134, 207]. For RS-HDMR, Li *et al.* compared orthonormal polynomials to cubic B splines, polynomials, and Monte Carlo integration. They showed that orthonormal polynomials decrease the needed sample size substantially while maintaining low approximation error [134], thus they will be used in this study.
The component functions consist of orthonormal polynomials ϕ and the model coefficients α_r^i, β_{rp}^{ij} and γ_{rpq}^{ijk}. For each parameter x_i a number orthonormal polynomials up to degree k, l, $l^{'}$, m, m', m'' are build.
The coefficients are determined by Monte Carlo integration (for more information see appendix).

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \phi_r^i(x_i) \tag{6.4}$$

$$f_{ij}(x_i, x_j) \approx \sum_{r=1}^{l} \sum_{p'=1}^{l^{`}} \beta_{rp'}^{ij} \phi_r^i(x_i) \phi_{p'}^j(x_j) \tag{6.5}$$

$$f_{ijk}(x_i, x_j, x_k) \approx \sum_{r=1}^{m} \sum_{p'=1}^{m^{`}} \sum_{q=1}^{m''} \gamma_{rpq}^{ij} \phi_r^i(x_i) \phi_{p'}^j(x_j) \phi_q^k(x_k) \tag{6.6}$$

The degree of the polynomials (k,l,$l^{'}$, m, m', m'') can be chosen individually for each component function. To build the first order component functions f_i of one input, $r = 1, 2, ...k$ expansion coefficients α_r^i are used for the separate orthonormal polynomials ϕ_r^i. For a second order function $l \cdot l'$ and for a third order function $m \cdot m' \cdot m''$ coefficients are used. Usually, the polynomial‘s degree (k,l,$l^{'}$, m, m', m'') is not increased beyond three or four. Higher order polynomials tend to oscillate. The number of needed coefficients p_{HDMR} will be increased also. For a model with 30 inputs with zeroth to second order correlations 4006 expansion coefficients need to be determined for a RS-HDMR meta-model (using third order polynomials).

The ingredients of a RS-HDMR meta-model was explained in this section. After construction of the meta-model, the sensitivity indices need to be extracted from the build model. This is described in the next section.

6.2.2 Sensitivity Indices

At the end of the sensitivity analysis, the importance of the input parameters has to be quantified. This is done by computing global sensitivity indices. In RS-HDMR this can be achieved by partitioning the variance of the model output with respect to the contribution of model inputs. Finally, the separate fractions of inputs‘ variance contribution are divided by the overall variance of the model output $y(x)$ and are called global sensitivity indices.
By knowing the sensitivity indices, the modeler can judge the importance of an input (combination) by its contribution to the output‘s overall variance. Non-influential model inputs can be identified and the parameters‘ importance is quantifiable.
These sensitivity indices are calculated from the conditional variance $V(y|x_i)$ for one input parameter x_i, which is always smaller than the overall variance of the given function $V(y(\mathbf{x}))$.

From these variances, one can calculate the first order global sensitivity index of x_i:

$$S_i = \frac{V_x(E(y(\mathbf{x})|x_i))}{V(y(\mathbf{x}))} \tag{6.7}$$

This principle can also be applied to second and third order interactions. From the ANOVA-HDMR decomposition, the overall variance of the model can be partitioned for separate contributions of component functions [220]:

$$V(y(\mathbf{x})) = \sum V_i + \sum_i \sum_{j>i} V_{ij} + \sum_i \sum_{j>i} \sum_{k>j} V_{ijk} + \ldots + V_{123\ldots p} \tag{6.8}$$

Dividing by $V(y(\mathbf{x}))$, the global sensitivity indices are obtained for all combinatorial possible interactions of inputs [221]:

$$\sum S_i + \sum_i \sum_{j>i} S_{ij} + \sum_i \sum_{j>i} \sum_{k>j} S_{ijk} + \ldots + S_{123\ldots p} = 1 \tag{6.9}$$

For RS-HDMR meta-models, the variance of the model $V(y(\mathbf{x}))$ can only be approximated by the calculated model outputs. If uniform distributed (i.e., uncorrelated) inputs sampled in a hypercube are used, sensitivity indices can be calculated by given N samples with input parameters $\mathbf{x}^{(s)}$ (where s gives the index a sample) [73]:

$$\begin{aligned}
S_i &= \frac{V(f_i)}{V(y(\mathbf{x}))} = \frac{V_i}{V(y(\mathbf{x}))} = \frac{\sum_{r=1}^{k}[\alpha_r^i]^2}{\frac{1}{N}\sum_{s=1}^{N}(f_0 - y(\mathbf{x}^{(s)}))^2} \\
S_{ij} &= \frac{V(f_{i,j})}{V(y(\mathbf{x}))} = \frac{V_{ij}}{V(y(\mathbf{x}))} = \frac{\sum_{r=1}^{l}\sum_{p=1}^{l'}[\beta_{r,p}^{ij}]^2}{\frac{1}{N}\sum_{s=1}^{N}(f_0 - y(\mathbf{x}^{(s)}))^2} \\
S_{ijk} &= \frac{V(f_{i,j,k})}{V(y(\mathbf{x}))} = \frac{V_{ijk}}{V(y(\mathbf{x}))} = \frac{\sum_{r=1}^{m}\sum_{p=1}^{m'}\sum_{q=1}^{m''}[\gamma_{r,p,q}^{ijk}]^2}{\frac{1}{N}\sum_{s=1}^{N}(f_0 - y(\mathbf{x}^{(s)}))^2}
\end{aligned} \tag{6.10}$$

If the inputs are non-orthogonal (independent and/or correlated, as in our case), these formulas can no longer be applied, because the uniqueness of the decomposition of the variance is lost. The calculated sensitivity indices will also depend on correlations between parameters [276, 137]. The calculation of sensitivity indices can then be performed according to Li *et al.* [276, 137]: An RS-HDMR function expansion can be seen as an additive model, where f_{p_j} are the separate component functions describing the contribution of single or subsets of inputs x_{p_j}:

$$f(\mathbf{x}) \approx f_0 + \sum_{j=1}^{n_p} f_{p_j}(x_{p_j}^{(s)}), \tag{6.11}$$

where n_p is the number of component functions f_{p_j}. The sensitivity indices can be decomposed into total ($S_{p_j}^{tot}$), structural ($S_{p_j}^{struc}$) and correlative ($S_{p_j}^{cor}$) contributions (where $\overline{y(\mathbf{x})}$ is the average value of y for all samples):

$$S_{p_j}^{tot} = Cov(f_{p_j}, y(\mathbf{x}))/V(y(\mathbf{x})) \approx \frac{\sum_{s=1}^{N} f_{p_j}(x_{p_j}^{(s)})(y(\mathbf{x}^{(s)}) - \overline{y(\mathbf{x})})}{\sum_{s=1}^{N}(y(\mathbf{x}^{(s)}) - \overline{y(\mathbf{x})})^2} \tag{6.12}$$

$$S_{p_j}^{struc} = Var(f_{p_j})/V(y(\mathbf{x})) \approx \frac{\sum_{s=1}^{N}(f_{p_j}(x_{p_j}^{(s)}))^2}{\sum_{s=1}^{N}(y(\mathbf{x}^{(s)}) - \overline{y(\mathbf{x})})^2} \tag{6.13}$$

$$S_{p_j}^{cor} = S^{tot}{}_{p_j} - S_{p_j}^{struc} \tag{6.14}$$

The coefficient $S_{p_j}^{tot}$ is called total sensitivity index, it can be positive or negative, depending on the effect on the overall variance. It is based on the covariance of the component function $Cov(f_{p_j}, y(\mathbf{x}))$. When $\sum_{j=1}^{n_p} S_{p_j}^{tot} \approx 1$, the results of RS-HDMR can be assumed a reliable estimate for the model behavior. The structural sensitivity index $S_{p_j}^{struc}$ is always positive and is based on the variance of the component function $Var(f_{p_j})$. $S_{p_j}^{cor}$ is the correlative sensitivity index. It can be positive or negative depending on the influence of the component function f_{p_j} and its correlation with other inputs. The correlative sensitivity index indicates correlation of the inputs ($S_{p_j}^{cor} \neq 0$), in this case the structural contribution is to be taken into account. Otherwise, if there is no correlation, the respective f_{p_j} are (approximately) orthogonal and the structural and total sensitivity indices are equal $S_{pj}^{tot} = S_{p_j}^{struc}$.

6.2.3 Measures for Goodness of Fit in RS-HDMR Techniques

Later, RS-HDMR techniques will be evaluated. To compare their results, three measures for goodness-of-fit are defined here to describe the approximation quality of a RS-HDMR model. In order to obtain comparability, the measures are independent of the sample size. First, the sum of all total sensitivity indices for an RS-HDMR meta-model is calculated.

$$S_T^{tot} = \sum S_i^{tot} + \sum_i \sum_{j>i} S_{ij}^{tot} + \sum_i \sum_{j>i} \sum_{k>j} S_{ijk}^{tot} + \dots \tag{6.15}$$

For separate outputs, the average values for them are given by $\overline{S}_T^{tot}$. To measure the deviation of sampling points and approximated function values, a scaled l^2-norm is used. It is normalized by the sample number N :

$$\|f(\mathbf{x}) - y\|_{2,n} = \sqrt{\frac{1}{N}\sum_{s=1}^{N}(f(\mathbf{x}^{(s)}) - y(\mathbf{x}^{(s)}))^2} \tag{6.16}$$

Additionally, also the maximum norm is used to measure fit quality:

$$\|f(\mathbf{x}) - y\|_{\infty} = \max_s \left\|f(\mathbf{x}^{(s)}) - y(\mathbf{x}^{(s)})\right\| \tag{6.17}$$

An RS-HDMR model fits the data optimal if $\|f(\mathbf{x}) - y\|_{\infty}$ and $\|f(\mathbf{x}) - y\|_{2,n}$ approach "zero", while it shows a similar overall variance as the sampling points ($S_T^{tot} \approx 1$).

6.2.4 Published RS-HDMR Techniques

In recent years, several new techniques for RS-HDMR were published. Without going into detail, only basic concepts of the techniques are described, which will be used later. The basic RS-HDMR was first published by Rabitz *et al.* [195]. This technique is applicable to orthogonal model inputs.

The calculation of model coefficients in RS-HDMR (by Monte Carlo integration, see appendix for formula) is an approximation, thus their application will result in weak fitting quality of

the model. The deviation between function value $y(\mathbf{x})$ and approximation $f(\mathbf{x})$ can be used to refine the RS-HDMR coefficient. One technique for this is the correlation technique for orthogonal inputs. For non-orthogonal inputs the backfitting algorithm can be applied [32, 137].
In some models important higher order component functions are found. Whereas the computational effort for calculation of the model coefficients increases. Li *et al.* showed that higher order component functions (order $\geq$2) can be efficiently approximated by the lower order product technique. Here, lower order component functions (usually first or second order) are used to approximate higher order component functions [135].
Usually, 5,000-10,000 samples are used for RS-HDMR model building. For low number of samples (less data points than expansion coefficients) D-MORPH (Diffeomorphic Modulation under Observable Response Preserving Homotopy) regression was shown to obtain a unique solution by applying a cost function. The cost function can be, for example, reduction of model variance or preservation of orthogonality between the hierarchical component functions [138, 133]. In this study, only the technique for preserving orthogonality with D-MORPH regression according to Li *et al.* is applied [133]. In this technique, the orthogonality of component functions in one hierarchical level like $f_i(x_i)$, $f_j(x_j)$, $f_k(x_k)$ is preserved to be orthogonal to the respective component functions on higher hierarchical levels like $f_{ij}(x_i, x_j)$, $f_{jk}(x_j, x_k)$, $f_{ik}(x_i, x_k)$. The same can be obtained by using extended bases for more samples than expansion coefficients [135].

6.3 Work-flow for Global Sensitivity Analysis

The work-flow for conducting RS-HDMR can be seen in figure 6.2. Before using the work-flow, the model input (samples) and the model output need to be computed. Samples can be generated by using a (quasi-)random number generator or, for example, a Gibbs or Hit-and-Run sampler [37]. Depending on the properties of the provided samples, the technique for global sensitivity analysis is chosen:

1. In case of detected uniformly distributed (assumed to be orthogonal) inputs, Legendre polynomials can be used for the RS-HDMR function expansion. Depending on the number of samples and expansion coefficients in RS-HDMR (p_{HDMR}) provided, the technique is chosen: For more samples than expansion coefficients, extended bases and non-extended bases can be applied. If less samples than expansion coefficients in the model are available, DMORPH regression has to be used (underdetermined system). This technique can also be beneficial if the parameters are not uniform distributed.

2. If non-uniformly distributed samples are used, two procedures are available for building orthonormal polynomials. The orthonormal polynomials are constructed by (**I**) Gram Schmidt procedure or (**II**) by forcing the discrete weighted orthonormality property of the polynomials to be preserved (see Li et al. for details) [135]. The first approximates the input parameter‘s density function by polynomials with automatically chosen degree and the latter by using the provided samples.
   The same techniques can be applied as in the uniform case, except that the backfitting procedure is used solely for non-extended bases [133].

For both methods it is possible to approximate higher order component functions by lower order product technique. Finally, it has to be decided if the inputs are correlated or uncorre-

lated. Depending on this choice, variance decomposition can be applied. If the model shall be used for calculations, non-informative component functions can be removed from it by reducing the coefficient number in the model based on the F-statistic [137].
A toolbox for global sensitivity analysis (GloSA toolbox) incorporating this work-flow using random sampling RS-HDMR was implemented in MATLAB 2012a.

6.4 Comparison of RS-HDMR Techniques

To establish RS-HDMR techniques for global sensitivity analysis in ^{13}C-MFA, preliminary tests of HDMR techniques were performed. The techniques were compared based on an analytically known test function (Sobol's g-function) and a ^{13}C-MFA model output (data in appendix). The performance of a technique was evaluated by comparing the measures of maximum norm $\|f(\mathbf{x}) - y\|_\infty$, l^2-norm $\|f(\mathbf{x}) - y\|_{2,n}$ and total sensitivity $\overline{S}^{tot}$. This was performed for extended bases and non-extended bases, lower order product techniques, and DMORPH regression. Additionally, the prediction quality of the sensitivity indices was examined. It was shown that by increasing sample size substantially above 5,000-10,000 samples, the approximation of sensitivity indices becomes better.
Among all tested algorithms non-extended bases with backfitting, DMORPH, and extended bases technique performed best. DMORPH regression showed, as expected, better performance at low sample number. DMORPH regression and extended bases are computationally demanding, but lead to (slightly) better measures for fit quality. In the end the backfitting procedure was used to build the RS-HDMR for ^{13}C-MFA models, because it is fast and produces reliable results.

In the next section RS-HDMR models are built for a complete network with 429 labeling measurements. From this a suitable visualization is chosen to condense the more than 200.000 sensitivity indices into a single matrix. From this matrix sensitive parameters can be identified. Finally, RS-HDMR is used as a meta-modeling technique to gain direct access to the complex interaction of fluxes and measurements.

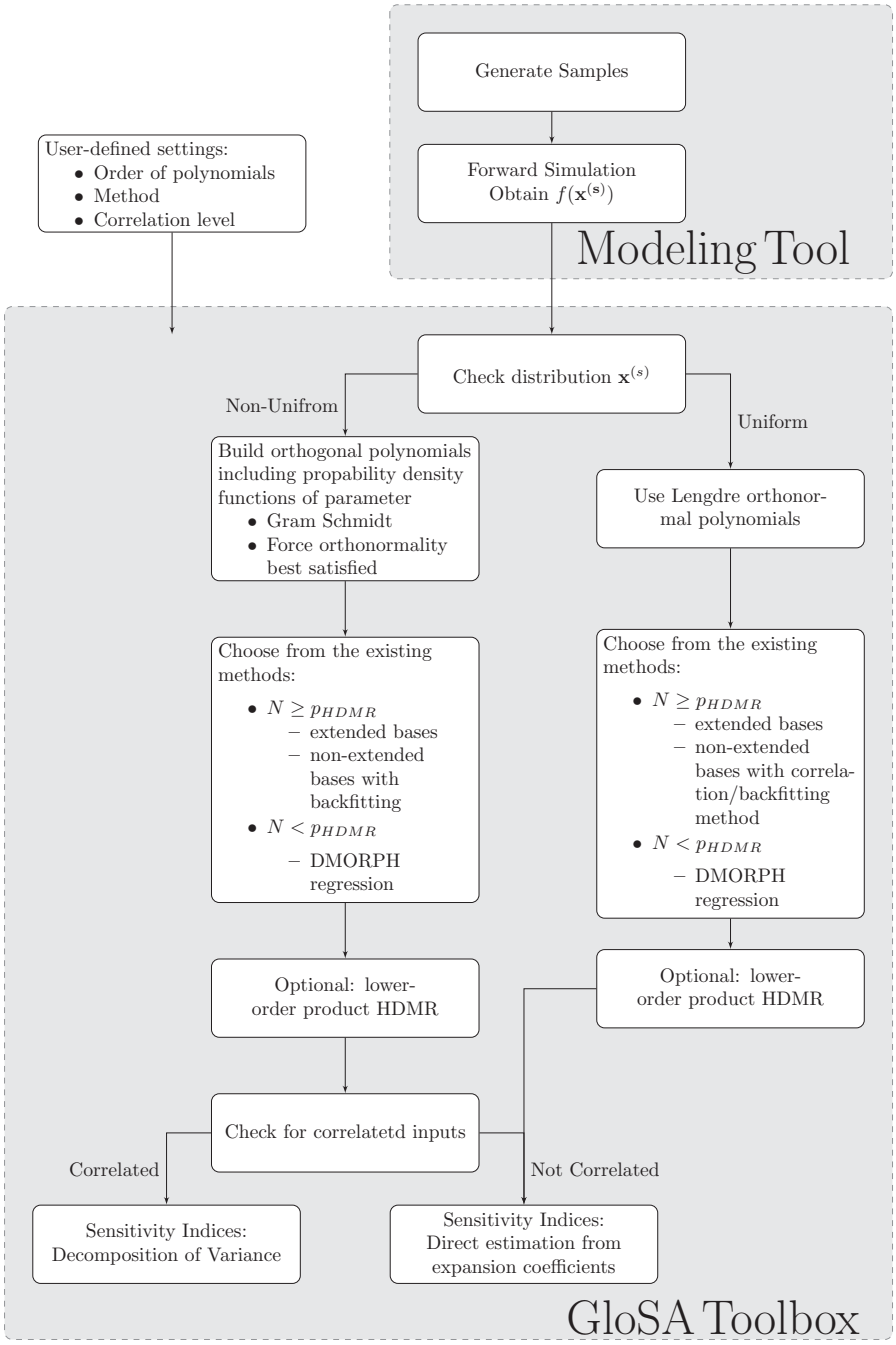


figure 6.2: GloSA toolbox and work-flow for optimal application of RS-HDMR techniques

6.5 Global Sensitivity Analysis for ^{13}C Metabolic Flux Analysis

In this section, global sensitivity analysis for a model of *P. chrysogenum* is performed. Here, the ultimate aim is the quantification of influence for the inputs, to identify the insenstive parameters. Visualization of the results is provided. Finally it is shown, that the generated RS-HDMR can be used to visualize and quantify the impact of fluxes on the labeling pattern.

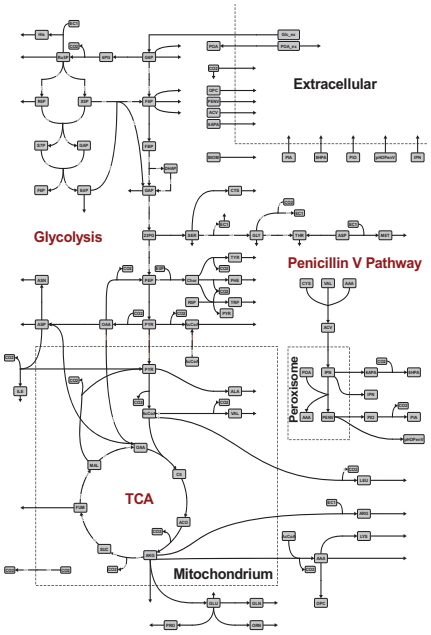


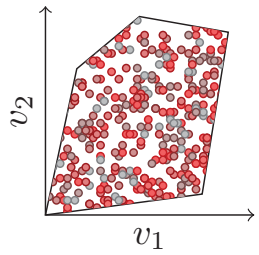
figure 6.3: *P. chrysogenum* model.

The model contains 50 LC-MS/MS measurement groups with overall 429 single measurement values (each measurement group refers to a specific metabolite measurement, consisting of single measurement values). The constructed model consists of 127 reactions with 76 pools for *P. chrysogenum* with a substrate mixture of 60 % Glc#100000, 20 % Glc#111111, and 20 % Glc#000000. Adaptions in constraints and flux directionality as well as penicillin pathway were made compared to the previous chapter. The uptake of glucose was set to $gly1.n$=1.00 and the biomass efflux to $BIOM\_ex.n$=0.0032 (according to experimental data). The measurements are situated in central carbon metabolism (glycolysis, pentose phosphate pathway, and citrate cycle) as well as (precursors of) amino acids. The fluxes were sampled in intervals given in table A.4 in the appendix. The exchange fluxes were sampled in the interval $[0, 6]$. Non-extended bases with backfitting procedure was used in RS-HDMR with 200 iterations and orthonormal polynomials of third order. The function was adaptively damped (see appendix for description).

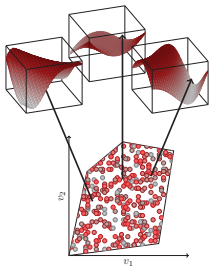
In the next section, the generation of a visualization for such a MIMO system is explained and applied for a model from literature for *E. coli* and later on for the previously build *P. chrysogenum* model. The *E. coli* model was chosen as a bad example.

6.5.1 Building Global Sensitivity Matrices for MIMO Systems

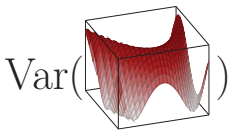
Sensitivity analysis is usually accompanied by a suitable visualization of its results to support strategic decisions for the model building process. Here, the work-flow for generating a sensitivity matrix for a MIMO system is described. It condenses the results of the sensitivity analysis into a readily accessible visualization.



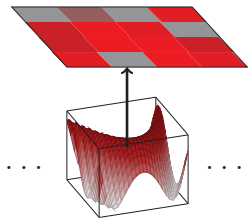
1. *Samples are generated in input space.* The flux space is a convex polytope constrained by the the stoichiometric matrix and given lower and upper bounds for the inputs. Sampling can be performed by Gibbs or Hit-and-Run sampler. In this study, the Gibbs sampler was used, only each p^3-th sample was used to guarantee independence of the sample points. For calculating the RS-HDMR 50,000 samples were generated.
From one given set of free fluxes, the model outputs are generated. Overall, for 429 model outputs $y_i(x)$ RS-HDMR meta-models were built.



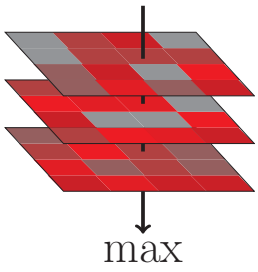
2. *Construct the component functions* (equation 6.6) *using RS-HDMR techniques.* The construction of the orthonormal polynomials $\phi(x_i)$ is performed only once for a set of samples, but for each model output a separate RS-HDMR has to be calculated. The chosen technique and settings depend on the system under investigation and desired accuracy. Naturally, the technique of choice has to be fast, and can be parallelized but provides sufficient accuracy. Here, the non-extended basis with backfitting procedure was chosen.



3. *Calculate sensitivity measures.* For each model output $y^i(x)$ the variance and covariance of the separate component functions $Var(f_{p_j})$, $Cov(f_{p_j}, f(\mathbf{x}))$ and $S^{cor}_{p_j}$ need to be calculated (equation 6.12-6.14). The first two quantities are not normalized to the model's output variance to weigh them with respect to the other model outputs. This is important as the variance of the model outputs of the ^{13}C-MFA model varies across five orders of magnitude (10^{-7}-10^{-2}). Thus, an input contributing weakly to a model output with high variance can (in absolute numbers) be overall more important than an input with high influence on a model output exhibiting low variance. By this the most influential inputs can be identified and lead to highest variance in model output.



4. *Build matrices sensitivity measures.* The extracted structural variance of each component function is written to a sensitivity matrix for each measurement $y^i(x)$ separately. Thus, 429 sensitivity matrices are built for one ^{13}C-MFA model.



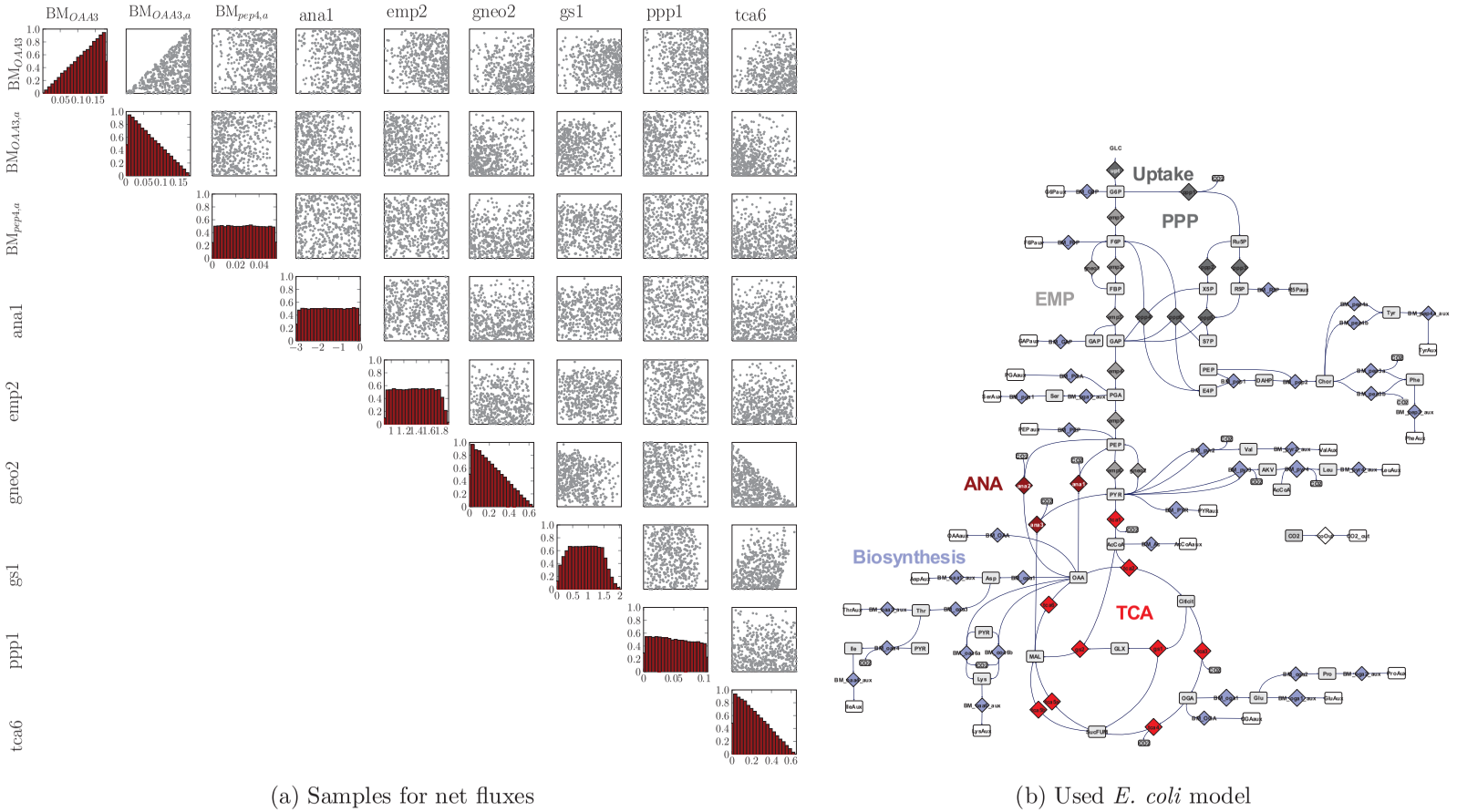
5. *Build MIMO sensitivity matrices.* Finally, the maximal absolute values for a specific component function f_i or f_{ij} in all constructed matrices $y^i(x)$ of $Var(f_{p_j})$ is searched for. The value is extracted and written to a new matrix. This matrix is the so-called structural sensitivity MIMO matrix.
Caution has to be taken, as the separate entries in the matrix might be intermixed if inputs are correlated. Also the entries corresponding to the extracted values of the structural sensitvity measures ($S_{p_j}^{struc}$) need to be visualized. The respective correlative sensitivity index (S_{pj}^{cor}) and the total sensitivity indices (represented by $Cov(f_{p_j}, f(\mathbf{x}))$) of the component function are visualized also in a matrix.

Next, the described workflow is applied to an example for a ^{13}C metabolic model from literature.

6.5.2 Example from Literature: *E. coli* model

In literature, several examples for stationary ^{13}C-MFA are found. A model of *E. coli* including measurements was published by Zamboni et al. [280]. In figure 6.5a this *E. coli* model is shown. The model contains 68 reaction and 37 pools. It is composed of 26 inputs and 194 output functions (35 MS measurement groups, only amino acids were measured).
For the generation of an RS-HDMR, random samples are needed as previously described. The flux values are constrained by lower/upper bounds and the stoichiometric matrix S. Due to the concept of free fluxes, the flux parameter space is correlated. Additionally, linear inequality constraints cut the space resulting in a convex polytope. Samples for the *E. coli* model can be seen in figure 6.4. The sample densities for the net fluxes are non-uniform and the sample points projected into two dimensions show areas without any samples.
The result of a sensitivity analysis described in section 6.5.1 is shown in figure 6.5b. One of the three measures (S^{struc}) for sensitivity of the separate fluxes is visualized in the structural sensitivity MIMO matrix. High values (▬) correspond to sensitive model inputs or combinations of them in respect to the measurements. Without further discussion, it is obvious that the sensitivity indices for all effluxes to biomass (prefix "*BM*") show only weak contributions (▬) to the output. Thus, several model inputs are not sensitive towards the MS measurements in the model. Later, interpretation of simulation outcome of this model may be severely hampered by these insensitive parameters because they are taken for granted. Besides, the model is over-parametrized. Potentially, even wrong conclusion may be drawn from the fluxes, if no statistics (confidence intervals of the fluxes) is calculated. In order to prevent this, the respective inputs can be set constant or removed from the model. In an optimal case, measurements can be added by experimental design to the model to increase the sensitivity of these fluxes. A possible way out in the case of the *E. coli* model is to measure the effluxes to the biomass and, thus, also these fluxes will become sensitive.

After this example from literature, the *P. chrysogenum* model is evaluated using the global sensitivity analysis.



(a) Samples for net fluxes

(b) Used *E. coli* model

figure 6.4: (a) Shown are 50,000 uniformly sampled points in net flux space for (b) the *E. coli* model (model taken from Zamboni et al. [280]). Density plots (▬) for all samples projected onto input space x_i for net fluxes are shown in diagonal plots. 500 Sampling points (•) are shown in two dimensional projections for each flux combination. The program *ssampler* of 13CFLUX2 using a Gibbs sampler was used for sample generation.

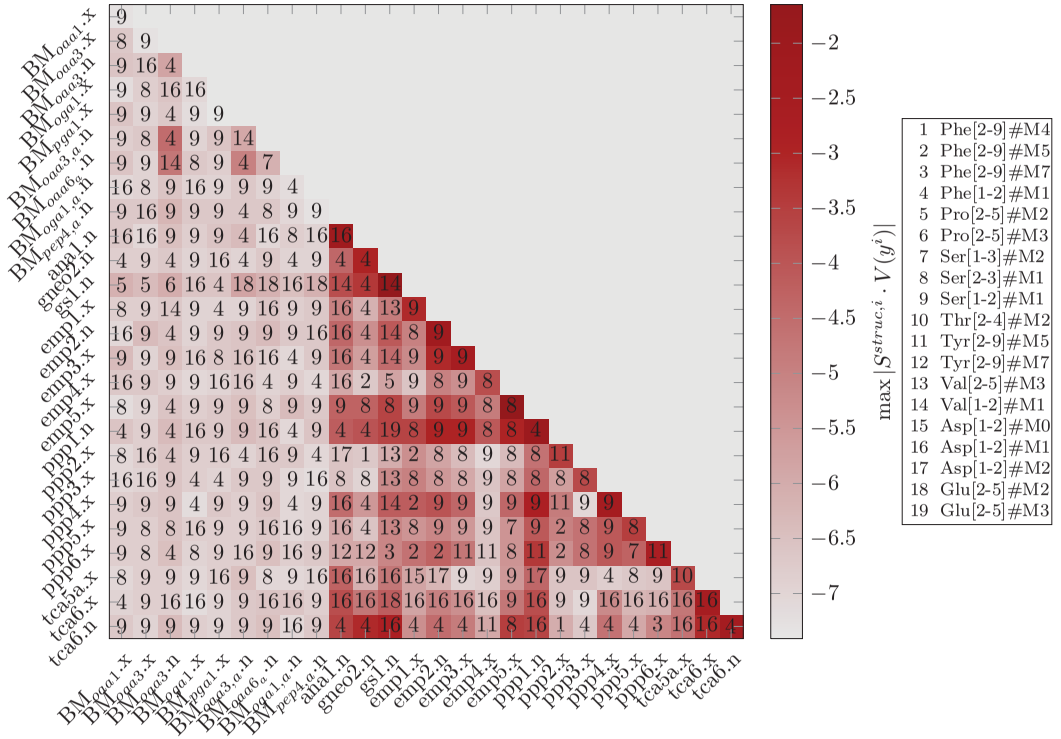


figure 6.5: Sensitivity analysis performed for the *E. coli* model (model taken from [280]). Sensitivity matrix (based on structural sensitivities scaled by model output) for the free fluxes. $\|f(\mathbf{x}) - y\|_{2,n} = 0.0015$, $\|f(\mathbf{x}) - y\|_{\infty} = 0.102$, $\overline{S}_T^{tot} = 0.953$.
The influx $upt.n$ was set to 1.00. The exchange fluxes were sampled in the interval $[0, 3]$: $BMooa3.n$: $[0.00, 0.18]$, $BMooa3aux.n$: $[0.00, 0.18]$, $BMoo6aux.n$: $[0.00, 0.18]$, $BMoga1aux.n$: $[0.00, 0.19]$, $BMpep4.n$: $[0.00, 0.05]$, $ana1.n$: $[0.00, 0.3.00]$, $emp2.n$: $[0.00, 3.00]$, $gneo2.n$: $[0.00, 3.00]$, $gs1.n$: $[0.02, 3.00]$, $ppp1.n$: $[0.00, 3.00]$, $tca6.n$: $[0.00, 3.00]$. The RS-HDMR was built by backfitting procedure with 50 iterations using orthonormal polynomials of third order. The function was adaptively damped. Calculations were performed by Samuel Leweke in a Python implementation of RS-HDMR. Suffix ".x" is referring to an exchange flux and ".n" to an net flux.

6.5.3 Results for *P. chrysogenum* model

RS-HDMR models were built for the separate measurements in the ^{13}C-MFA model and the above work-flow was applied for visualization. Next, the results will be discussed.

Sensitive Fluxes

The resulting total sensitivity MIMO matrix for 429 labeling measurements can be seen in figure 6.6. The terms $Var(f_{pj})$ and S_{pj}^{cor} corresponding to entries in the total covariance MIMO matrix are visualized in figure 6.7. In the model first order component functions were most influential, shown by the strong contribution of high values in the diagonal elements in figure 6.6. Four fluxes show only weak first order variances: Net flux through phosphoenolpyruvate carboxykinase and exchange flux of threonine aldolase and acetyl-CoA transporter (*Ana1.n*, *Thr2.x*, *Trans_AcCoA.x*) exhibit low effect on output variance of the model indicating that they can be removed from the model. Furthermore, there are two distinct clusters visible were high sensitivities were found in the fluxes of the TCA, glycolysis and transporter to mitochondria (*Tca1.n*, *Tca8.n*, *Trans_CO2.n*, *Trans_Pyr.n*, *gly2.n*, *gly7.n*). Another distinct cluster consists of the pentose phosphate pathway and fluxes in glycine formation for exchange fluxes (*PPP3.x*, *PPP3.x*, *PPP3.x*, *PPP3.x*, *PPP3.x*, *Gly1.x*, *Gly2.x*). They also showed strong effects on output variances by second order cooperative effects for fluxes of glycolysis (*gly2.n* and *gly7.n*).
The first order effects contribute 80 % of the variance and the second order effects 16 %. On average 96 % of the overall variance of model output was accounted for by the RS-HDMR meta-model. Total variance was usually well reproduced, a few model outputs were fitted only with low quality (lowest values $S_T^{tot} = 0.73$). Potentially, higher order interactions or polynomials need to be chosen to obtain better estimates for these measurements.

Correlations in the Network

Strong correlative effects can be seen in the plot of the correlative MIMO sensitivity matrix (cf. figure 6.7a). Especially the main net fluxes (*gly2.n*, *gly7.n* and *Tca1.n*) show strong correlative contributions to the total variance, and, thus, the structural MIMO sensitivity matrix has to be taken into account. In this matrix, main tendencies are not changed by the correlation. Caution is advised when interpreting the total sensitivity MIMO matrix, as correlative effects may hamper interpretability of the results as first and second order effect can contribute to each other by the effect of correlated inputs. Besides, orthogonality of the component function can be impaired in non-orthogonal inputs, resulting in the same effect and can also be indicated by the correlative MIMO sensitivity matrix.

Measurements

Besides identification of most influential inputs, the most influential measurement can be extracted from figure 6.6. As expected, the aspartate formation in the cytosol or mitochondrium shows strong correlation with the measurement of aspartate. Interestingly, most net fluxes in glycolysis are sensitive on the xylulose/ribulose-5-phosphate measurement. Anaplerotic fluxes (*Ana1.x* and *Ana2.x*) show also high sensitivity indices for phosphoenolpyruvate measurement. Exchange flux in upper glycolysis of glucose-6-phosphate isomerase show high sensitivity indices for the 6-phosphogluconate measurements.

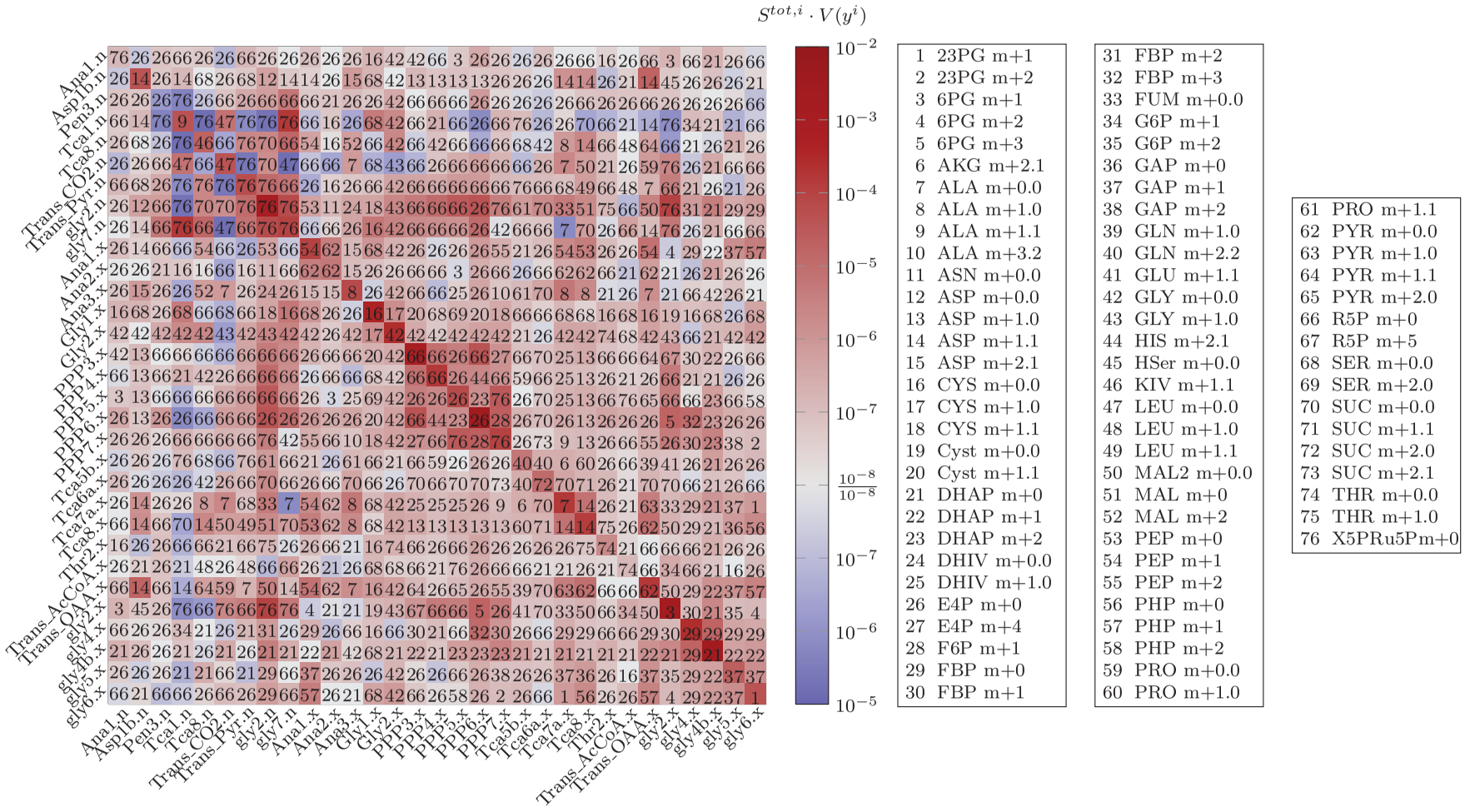


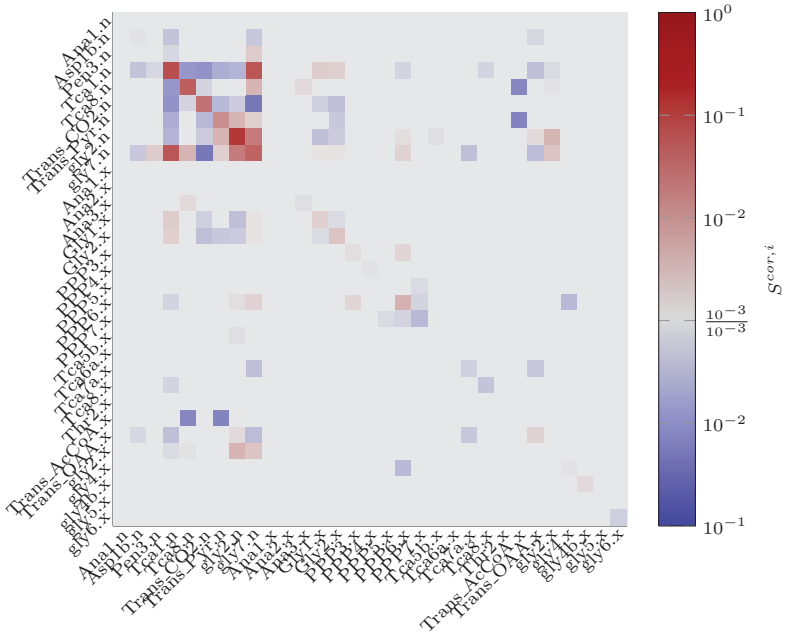
figure 6.6: Total sensitivity MIMO matrix for *P. chrysogenum* network. Maximal absolute value of total variance of the respective component functions were extracted from all model outputs. Red patches indicate positive and blue negative values, respectively. The numbers given in each patch are identifiers of the respective measurement that was used (see legend on right hand side). RS-HDMR built for all MS/MS measurements $\|f(\mathbf{x}) - y\|_{2,n} = 0.0028$,$\|f(\mathbf{x}) - y\|_{\infty}$ =0.183, $\overline{S}_T^{tot} = 0.955 \pm 0.036$. Suffix ".x" is referring to an exchange flux and ".n" to an net flux.

Finally, the fit quality and run-time of the program was evaluated for the shown 429 meta-models.

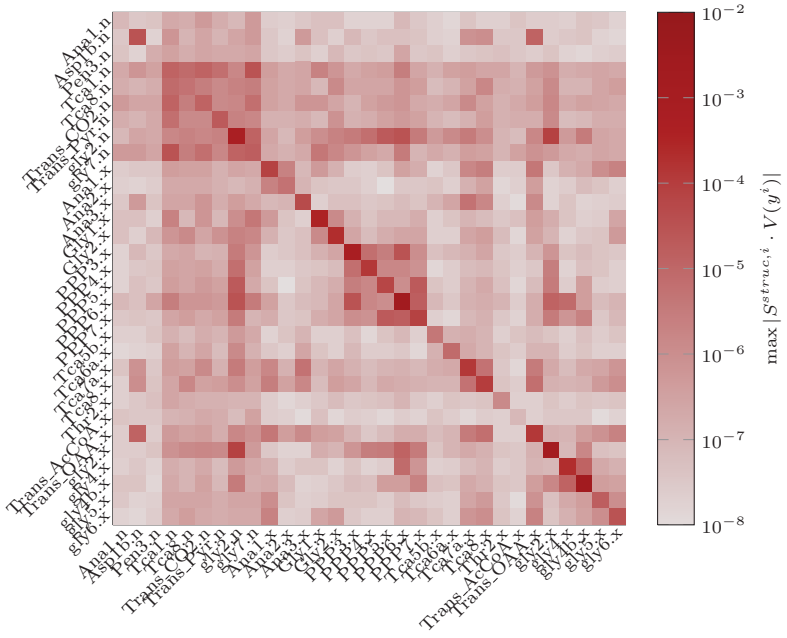
Fit quality and Run-time of RS-HDMR

RS-HDMR was efficiently applied to gain knowledge about global sensitivities of fluxes. Here, only polynomials up to third order were used. Increasing the order further will increase fitting quality and results in better estimation of the overall variance of model output $\overline{S}_T^{tot}$, which was 4 % below of the variance of the MIMO model.
The overall procedure took 20-24 hours for the generation of 50,000 samples (including for each generated sample p^3 discarded samples, where$p$ equals the number of free fluxes; one core was used) and forward simulation of all flux samples by 13CFLUX2. Additionally, construction of all RS-HDMR meta-models lasted between 1-10 days (1 MATLAB process) depending on the desired accuracy (10 days were needed for this study).

Until now, the built meta-models were used to extract single-value measures (sensitivity indices). At the same time, more information is contained in the 429 built meta-models. In the next section we focus on the built first and second order component functions for specific measurements. The component functions are visual representations of changes in measurements caused by flux values and will be discussed in the context of the metabolic network. The mass spectrometric measurement of fructose-6-phosphate is used, which is highly sensitive for fluxes in upper glycolysis.



(a) Correlative MIMO sensitivity matrix



(b) Structural MIMO sensitivity matrix

figure 6.7: Structural and correlative MIMO matrix for *P. chrysogenum* network. (a) Maximal absolute correlative variance of the respective component functions. (b) Maximal absolute structural variance of the respective component functions were extracted. Red and blue patches indicate positive and negative value, respectively. Both matrices are carrying the same entries as the total MIMO sensitivity matrix.

6.6 Assessing Key Influential Fluxes from Changes in Labeling Patterns

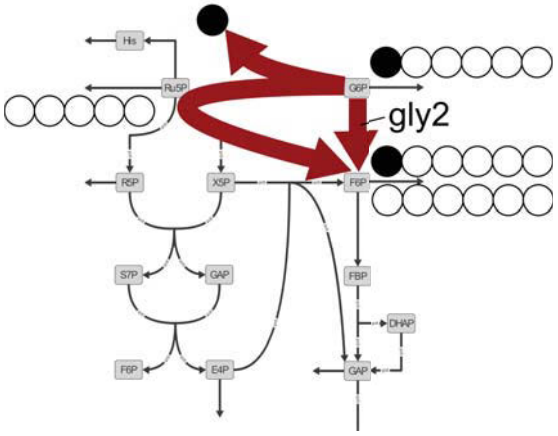
In literature the dependency of labeling patterns on fluxes has been extensively discussed [156]. Quantification of the effects of a flux on the labeling states requires interpretation in the context of the constrained flux space. Besides, forward simulation can be used to shed light onto the dependency of a labeling pattern of a metabolite on a given set of fluxes. In contrast to this, the changes in the MS labeling patterns are visualized in this section aiming at the reverse assessment of fluxes by the measured mass spectra.
The tracking of labeling through a reaction network is often used in literature to interpret labeling patterns. However, by the presented method, this can be simplified. Here, we aim at elucidating the question: *Can we use component functions to shed light on the interplay of MS measurements and fluxes in a ^{13}C-MFA model?*

P. chrysogenum Model

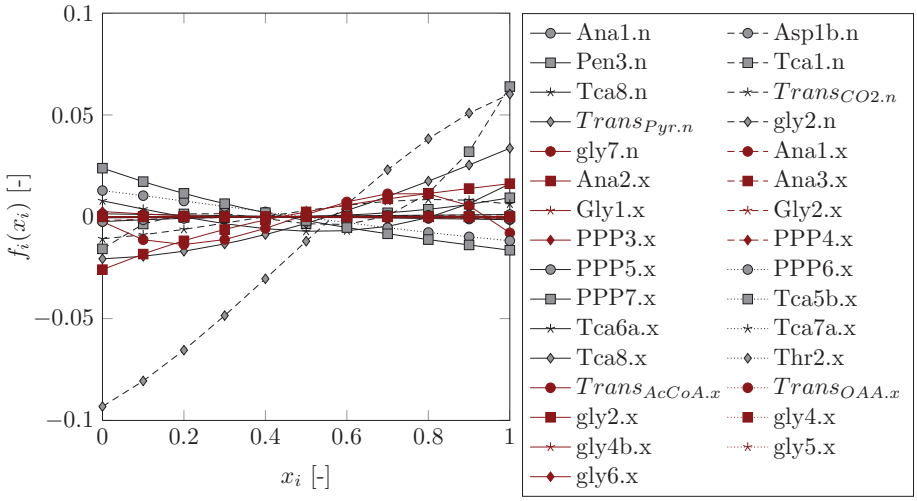
For this, the previously used *P. chrysogenum* network is employed. Deviating from the previous procedure, exchange fluxes were sampled in the interval $[0, 0.5]$. Net fluxes were sampled in intervals as for generation of MIMO matrices. In the following two examples, we will focus on the m+1 and m+2 mass trace of fructose-6-phosphate.

Fructose-6-Phosphate m+1 Mass Trace

An RS-HDMR meta-model was built for a MS measurement of the fructose-6-phosphate m+1 mass trace. The most influencing fluxes on the m+1 labeling pattern are the fluxes of glycolysis by glucose-6-phosphate isomerase (*gly2.n* and *gly2.x*). Here, a high net flux through glycolysis (by *gly2.n*) is resulting in a low flux through oxidative pentose phosphate pathway. For this model 60 % 1-, 20 % uniform, and 20 % unlabeled glucose was used. Thus, the higher the flux through glycolysis, the higher the m+1 trace as the first carbon atom of glucose-6-phosphate is removed in the pentose phosphate pathway by 6-phosphogluconate dehydrogenase (see figure 6.8a).
The resulting first order component for the m+1 mass trace can be seen in figure 6.8b. It decreases by 0.15 from lowest to highest value for *gly2.n*. Also *gly2.x* shows a similar effect, but less pronounced. As expected for this mass trace, *gly2.n* has a $S^{tot}_{gly2.n}$ of 0.61 and is the most influential flux acting on the m+1 labeling. Besides the exchange flux was rather low $[0, 0.5]$ and exhibits only a small change in the labeling pattern $S^{tot}_{gly2.x} = 0.03$. The other inputs of the model were less influential on the output.
Additionally, a high second order interaction is observed for *gly2.n* and *gly2.x*. The second order component function $f_{gly2.n,gly2.xch}$ of m+1 is visualized in figure 6.9 and results in a total sensitivity of $S^{tot}_{gly2.n,gly2.xch} = 0.02$. Here an additional effect of both fluxes acting together on the model output is shown. Another change of ~0.1 in labeling pattern is accounted for in the second order interaction term.
In the second order component function, it can be seen that all 1-labeling is removed by pentose phosphate pathway if flux through glycolysis is zero ($gly2.n = 0$ and $gly2.x = 0$). If the exchange flux and/or net flux to glycolysis is increasing, fructose-6-phosphate pool's m+1 labeling fraction elevates.



(a) Pentose phosphate pathway



(b) F6P m+1

figure 6.8: RS-HDMR results for fluxes effecting 1-labeling in fructose-6-phosphate mass spectrometric measurement. (a) Effect of flux over pentose phosphate pathway on 1-labeled glucose as substrate. (b) First order component function for m+1 mass traces for fructose-6-phosphate. $\|f(\mathbf{x}) - y\|_{2,n} = 0.0017$, $\|f(\mathbf{x}) - y\|_{\infty} = 0.0158$, $S_T^{tot} = 0.999$.

This proof-of-principle example was rather obvious. However, *can we also gain new insight in reactions which are not so obvious?* Next, the m+2 mass trace in fructose-6-phosphate is taken into account.

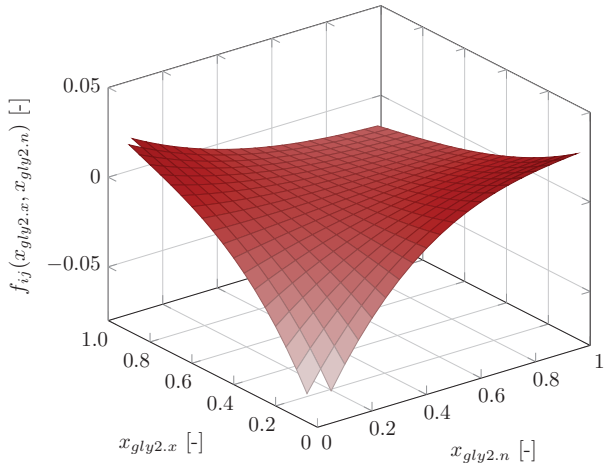


figure 6.9: Second order component function built for fructose-6-phosphate m+1 MS measurement. Elements in surface plots were removed, if the they were not inside of the convex hull spanned by the sampling points. $\|f(\mathbf{x}) - y\|_{2,n} = 0.0017$, $\|f(\mathbf{x}) - y\|_{\infty} = 0.0158$, $S_T^{tot} = 0.999$.

Fructose-6-Phosphate m+2 Mass Trace

The labeling mixture used for simulation of the experiment contains fully, 1-labeled, and unlabeled glucose. Thus, m+2 labeled fructose-6-phosphate has to be generated by reactions in the network. We will now infer its origin by using the participating reactions.
In figure 6.10 m+2 trace's first order component functions are shown. As it was not contained in the substrate mixture, the appearance of 2-labeled species has to be explained by the reactions in the network. Transketolase (*PPP7.x*) and transaldolase (*PPP6.x*) show high influences by their first order component functions (besides flux through glucose-6-phosphate isomerase, *gly2.n*). Thus, the respective reactions are involved in the formation of these species. Two possible routes for its formation will be outlined here.
Fully labeled glucose in the substrate mixture can be transported into pentose phosphate pathway to yield fully labeled xylulose-5-phosphate. The transketolase is forming fructose-6-phosphate. When xylulose-5-phosphate was fully labeled and erythrose unlabeled, we obtain m+2 labeled fructose-6-phosphate (see also figure 6.11, left):

$$\text{X5P\#11111} + \text{E4P\#0000} \xrightleftharpoons{transketolase/PPP7} \text{GAP\#111} + \text{F6P\#110000}$$

This is indicated also by the RS-HDMR first order component function for flux transketolase/*PPP7.x*, see figure 6.10.
Additionally sedoheptulose-7-phosphate can be formed from fully labeled xylulose-5-phosphate and unlabeled ribose-5-phosphate (X denotes arbitrary labeling).

$$\text{R5P\#00 XXXX} + \text{Xu5P\#11111} \xrightleftharpoons{transketolase/PPP5} \text{S7P\#110 XXXX} + \text{GAP\#111}$$

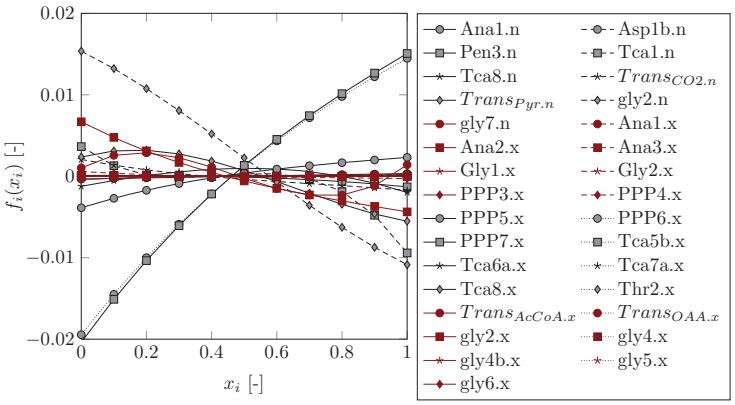
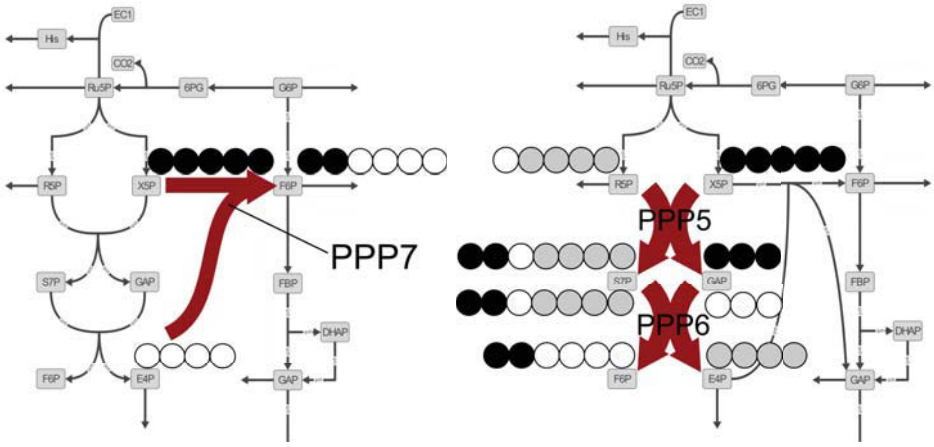


figure 6.10: RS-HDMR built for fructose-6-phosphate m+2 MS measurement. RS-HDMR for m+2: $\|f(\mathbf{x}) - y\|_{2,n} = 4.3 \cdot 10^{-4}$, $\|f(\mathbf{x}) - y\|_{\infty} = 0.0089$, $S_T = 0.998$.



(a) Transketolase (b) Transaldolase

figure 6.11: Two possible metabolic routes for the formation of m+2 labeled fructose-6-phosphate by (a) transketolase and (b) transaldolase/transketolase.

Sedoheptulose-7-phosphate labeled at the first two carbon atoms is then used to form m+2 labeled fructose-6-phosphate by action of transaldolase (see figure 6.11b).

$$\text{S7P\#110 XXXX} + \text{GAP\#000} \overset{transaldolase/PPP6}{\rightleftharpoons} \text{E4P\#XXXX} + \text{F6P\#110000}$$

Nevertheless, further routes for the formation of m+2 labeled fructose-6-phosphate can be found. The described reactions seem to explain the strong dependency of m+2 labeled fructose-6-phosphate on these two exchange fluxes.
The component function of the flux through glycolysis (*gly2.n*) is counter-acting the effects of the first order component functions of fluxes through transaldolase and transketolase (*PPP6.x* and *PPP7.x*). This dependency can be easily resolved. In the network net fluxes through the respective reactions of transaldolase and transketolase (*PPP6.n* and *PPP7.n*) mainly depend on the net flux through glucose-6-phosphate isomerase (*gly2.n*). Decreasing the flux through glycolysis results in increased flux over *PPP6.n* and *PPP7.n* and, thus, in an increased formation of m+2 trace ($PPP6.n = 0.3194 - 0.3333 \cdot gly2.n$, $PPP7.n = 0.3084 - 0.3333 \cdot gly2.n$).

Summing up, it was shown that an application of RS-HDMR is not limited to global sensitivity indices. It can be applied to visualize interplay of measurements and flux distributions. It increases the value of a pure global sensitivity analysis by providing readily interpretable visual representations of processes in the network context. Additionally, it enables the ranking of effects by fluxes acting on labeling patterns even in large networks. This is a complex and extensively discussed topic in literature (see for example figure 3 in Metallo *et al.* [156]). However, this method simplifies the shaping of hypthesis for the explanation of labeling patterns significantly.

6.7 Discussion

RS-HDMR was proven to be a reliable tool for estimation of sensitivity indices based on analytical test functions. The technique was established to perform the first global sensitivity analysis for ^{13}C-MFA models. The analysis was conducted for MIMO models with correlated input samples enabling the identification of key influential factors. For this, visualization of scaled sensitivity indices in a MIMO sensitivity matrix was used. Thereby, it was shown that fluxes (in pentose phosphate pathway, citrate cycle and glycolysis), important for the outcome of the study, were highly sensitive. Fluxes, for example in anaplerosis, were found to be in-sensitive towards the measurements. The approach proved to identify the set of measurements that are most influential for the response of the model input to the model output.
Interpretation of the resulting component functions was possible and they contained detailed information about the interplay of fluxes and labeling measurement. Thus, tedious tracking of labeling through networks for shaping of hypotheses about their origin is simplified and the changes in labeling can be quantified with respect to the fluxes.
The RS-HDMR method using non-extended bases and backfitting procedure could reproduce 96 % of the overall variance and acceptable fit was obtained over a wide range of input parameter values. Besides, it was shown that more than 5,000 samples are needed to build a RS-HDMR with non-orthogonal inputs. A good compromise between run-time and accuracy were 50,000 samples. Importantly, evaluation of the build meta-model on reference sample set was found to be highly recommended, preventing under-estimated errors of the meta-model (see appendix A).

Further Improvements

By further tests it was confirmed that regions with higher deviates ($\|f(\mathbf{x}) - y\|$) of RS-HDMR and model output were situated in input parameter regions with low sample densities. Improved sampling strategies are needed to increase sample size in these regions.
Besides, interpretability of the sensitivities is hampered by influence of correlated samples. Clearly, the decomposition into structural, correlative and total sensitivity described is an approximation. Reduction of input correlation can be accomplished by transformation into orthogonal input space, for more information see Mara *et al.* [147]. However, depending on the type of transformation used, the interpretability of the results might be severely hampered and component function may become more complex to approximate.

Removal of Fluxes

Insensitive parameters identified by global sensitivity analysis need to be removed from the given model in order to prevent false interpretation of the later estimated parameter values. Care has to be taken, as setting of fluxes constant or removing reactions might ultimately result in removal of important network parts under investigation. Removing such branches, however, will/can make the model "meaningless" to the question posed in the study. In this case, no inference about the fluxes of interest in the used network model can be made by the current experimental setup. However, global sensitivity analysis can prevent in this case the conduction of useless experiments. Adding further information by additional measurements or conduction of experimental design can help elucidating these pathways. Here, global sensitivity analysis can be applied again to gurantee global sensitivitiy of the parameters in respect to the measurements.

Chapter 7

^{13}C Metabolic Flux Analysis for *P. chrysogenum* BCB1

Although extensive research has been performed on *P. chrysogenum*, highly resolved flux maps for industrially relevant conditions are still rare. In this chapter ^{13}C-MFA is applied to a chemostat process to establish the technique. Afterwards, the obtained method is translated to a close-to-industrial fed-batch process using the sensor bioreactor concept. Details on the experimental setup are given in chapter 4.
This chapter describes the ^{13}C-MFA which was performed based on the focused ^{13}C metabolic network with 140 reactions. The obtained results are qualitatively compared to flux distribution obtained with the large scale ^{13}C metabolic model (chapter 5.1) to verify that the resulting flux maps obtained by the focused network are physiologically feasible and that the assumptions made to construct it were reasonable. Finally, an experimental design study for a chemostat (chemostat 1) experiment was conducted and the results are used to perform the following ^{13}C-MFA experiments (chemostat 2 and 3).

table 7.1: Cultivations performed for carbon labeling experiments

	μ[h^{-1}]	Labeled substrate [%] 1-^{13}C/U-^{13}C/^{12}C
Chemostat 1	0.03	60/20/20
Chemostat 2	0.03	70/30/0
Chemostat 3	0.05	70/30/0
Fed-Batch 1	0.008	60/20/20
Fed-Batch 2	0.007	60/20/20

7.1 ^{13}C-MFA for Chemostat Cultivations

Chemostat experiments were conducted to establish ^{13}C-MFA in *P. chrysogenum*, because they, at least theoretically, provide ideal steady-state conditions compared to fed-batch processes. First, experiments were performed using a specific growth rate of 0.03 h^{-1} and a substrate mixture of 60 % C1-, 20 % uniformly, and 20 % naturally labeled glucose. Following, an experimental design study was performed to increase the system‘s statistical identifiability. Two chemostat experiments were conducted with the optimized substrate mixture of 70 %

C1- and 30 % uniformly labeled glucose. Therefore two different specific growth rates of 0.03 h^{-1} and 0.05 h^{-1} were chosen.
After extrapolation, correction for natural abundance and statistical reconciliation (as described in chapter 4), of the LC-MS/MS measurements, more than 200 MS measurements were used for the ^{13}C-MFA. Measurements with low fit quality in extrapolation were removed beforehand from the data-set. These were more than 50% of the data.
The labeling data and up to 12 rate measurements were introduced into the model obtained by the kinetic process model in chapter 4. The used constraints and networks were previously described in chapter 5.1. The models including data were fitted and worst fitting MS-measurements were removed consecutively. Here, the results of the analysis are collected.

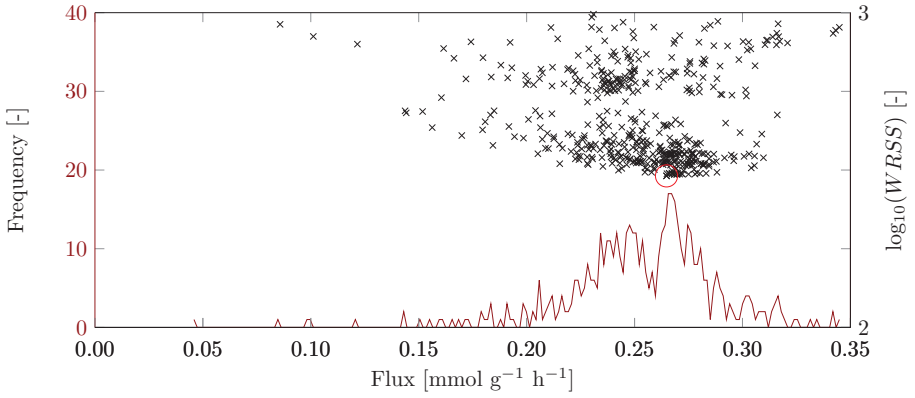
7.1.1 Chemostat 1 with Growth Rate of 0.03 h^{-1}

Flux Estimation

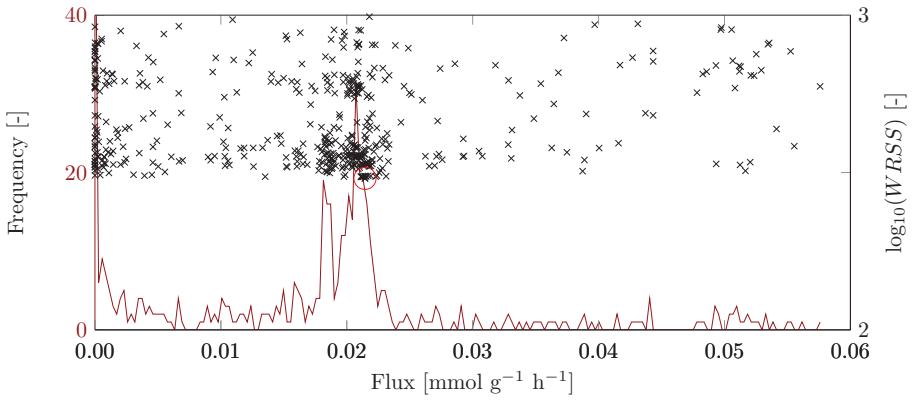
The focused model introduced in chapter 5 was supplemented with all rate and labeling measurements. If for a side-product in the penicillin pathway no rate measurement was available (because its concentration was too low), the measurement value of 0.0 ± 10^{-5} was used. Parameter estimations are performed based on the globalized multi-start strategy. The high performance simulator 13CFLUX2 (version 2.1) was used throughout this study for flux sampling, simulation, parameter estimation, statistics calculation, and experimental design [260]. Based on 500 starting points generated with a Gibbs sampler (using program *ssampler*, net fluxes were sampled in interval $[-1, 1]$ and exchange fluxes in $[0, 1]$), parameter estimation was performed with the program *fitfluxes* (optimizer: NAG-NLP from the commercially available NAGC library, maximal iterations: 350).

Detecting Local Minima

To detect local minima and alternative solutions, the estimated fluxes were compared in density plots (figure 7.1). For most net fluxes the obtained solution was a clearly defined optimum, e.g. for the flux through the pentose phosphate pathway (PPP) via glucose-6-phosphate dehydrogenase (*PPP1*) (figure 7.1a). For this flux, found solutions were ranging from 0.27 to 0.29 $\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$. The flux into the oxidative part of the pentose-phosphate pathway is important for further analysis, because it delivers NADPH for penicillin V formation.
For some fluxes a rather broad distribution of solutions was found, e.g. for the flux to acetyl-CoA via a transporter system. Two predominant clusters at 0.00 $\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$ and 0.02 $\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$ were found, corresponding to either complete formation of acetyl-CoA in the cytosol or in the mitochondria. Kleijn reported formation of acetyl-CoA in the mitochondria, but was not able to provide statistics for the estimated flux distribution [121].
For the model with lowest WRSS of 303, the estimated extracellular flux measurements were within the standard deviation of the experimental data and, thus, correspond to the measured data. However, for some LC-MS/MS measurements, the obtained fitted values were deviating from experimental determined standard deviations. The fit of the LC-MS/MS measurements are discussed below. For this purpose, a visualization was developed to detect measurements with poor fitting quality.

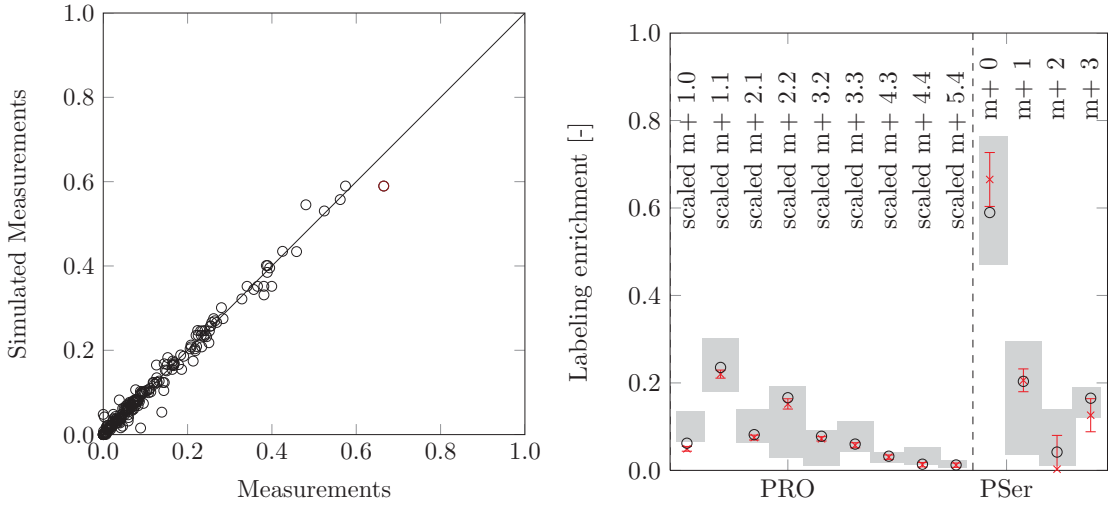


(a) Glucose-6-phosphate dehydrogenase (*PPP1*)



(b) Transporter to acetyl-CoA (*Transporter_AcCoA*)

figure 7.1: Found flux estimates of 500 parameter estimations using the program *fitfluxes* and *ssampler* applied for a globalized multi-start strategy for chemostat 1 with a growth rate of 0.03 h^{-1}: Two examples are given for net fluxes: (a) glucose-6-phosphate dehydrogenase (*PPP1*) and (b) transporter of acetyl-CoA (*Transporter_AcCoA*). The frequency (——) of the found solutions is shown with respect to the flux values. WRSS (×) values found by parameter estimation procedures are shown as scatter plots, the model with lowest residual (◯) is marked.



(a) Parity plot (b)

figure 7.2: (a) Parity plot of chemostat 1 with a growth rate of 0.03 h^{-1}. (b) Plot of minimal/maximal value for the simulated measurement (▭) for 50,000 flux samples, measurement (×) and simulated data (○).

Visualization of Goodness of Fit

Parity plots are used to show the goodness-of-fit of experimental and simulated data (see Kleijn [121], page 92). However, data-sets with differing standard deviations can exhibit strong deviates in the parity plot, although the simulated data lie within the standard deviation of the measurements. An example is given in figure 7.2a indicated by a red marker, which is the PSer m+1 measurement in figure 7.2b. From the second plot it is obvious, that the measurement of PSer possesses a large standard deviation compared to the data of PRO, which is not obvious in the parity plot. Thus, in the parity plot the measurement is an outlier, although by its standard deviation it is not.

Additionally, parity plots imply a possible change of the measurements within the interval $[0, 1]$, which is usually not the case for mass spectrometic labeling measurement in an ^{13}C-MFA experiment. Thus, sampling was performed (50,000 samples) using ssampler for all free fluxes (net and exchange), while measured extracellular fluxes were kept within the interval of 2.5 times their standard deviations. The exchange fluxes were sampled in the interval $[0, 10]$ and net fluxes within $[-100, 100]$. After simulating the MS measurements for each flux distribution, the minimal and maximal value of each labeling measurement was plotted (figure 7.2b, gray rectangles). The new visualization shows, that most measurements cover only a small fraction of the interval $[0, 1]$ in the found samples. Besides the range of the simulated measurements, the measured data, including standard deviation and the simulated LC-MS/MS measurements for the best fit model, are shown. These plots were made for all ^{13}C-MFAs conducted and can be found in the appendix C.

It is important to mention that single measurements have been removed from the measurement groups, because they showed high deviates to the simulated data. In figure 7.2b the m+0.0 mass trace of PRO was removed. This is indicated by adding "scaled" to the mass trace description, given in the plot (here the scaling factor ω was used to fit the data, see equation 2.14).

Discussion of the Flux Map - Glycolysis and Pentose Phosphate Pathway

The flux map for the chemostat 1 at a growth rate of $0.03\,h^{-1}$ (figure 7.3) reveals a high flux through the oxPPP. The flux into the upper part of glycolysis via glucose-6-phosphate isomerase is ~20 % of the absolute substrate uptake flux. High exchange fluxes were found for this enzymatic reaction ($12.03\,mmol\,g_{CDW}^{-1}\,h^{-1}$). In the PPP as well as in glycolysis low to high exchange fluxes were found ranging within 0-30 $mmol\,g_{CDW}^{-1}\,h^{-1}$ and 0-86 $mmol\,g_{CDW}^{-1}\,h^{-1}$, respectively. As the net flux through the oxPPP is high, most 1-labeling in the network is expected to be removed by activity of the 6-phosphogluconate dehydrogenase. Interestingly, this is not the case. Most metabolites in glycolysis show a high content of m+1 traces, e.g. 20-25 % m+1 can be found in 2- and 3-phosphoglycerate. Even in ribose-5-phosphate 25 % of single labeled species are present. That can be explained with the spread of the labeled glucose-6-phosphate pool via glucose-6-phosphate isomerase into the reaction network through glycolysis and PPP by exchange fluxes. Next to the labeled glucose-6-phosphate pool, 40 % m+1 is found in the fructose-6-phosphate pool. Thus, the glucose-6-phosphate pool is highly equilibrated with the fructose-6-phosphate pool. High equilibrating reactions for glucose-6-phosphate isomerase were also found by Kleijn for *P. chrysogenum* [121]. (The same effect was already shown by using RS-HDMR meta-models for fructose-6-phosphate m+1 trace in chapter 6.)

Anaplerosis, Gluconeogenesis and Citrate Cycle

The gluconeogenetic reaction from oxaloacetate to phosphoenolpyruvate by phosphoenolpyruvate carboxykinase was constrained to low values to prevent futile cycles. This reaction has a low statistical significance throughout this study. Kleijn found high values (ca. 50 % flux through lower glycolysis) for this reaction but could not calculate confidence intervals and a statistical analysis is lacking [121]. Flux in anaplerosis by pyruvate carboxylase was found to be high ($0.176\,mmol\,g_{CDW}^{-1}\,h^{-1}$). This can be explained by the activity of mitochondrial malic enzyme, converting $0.054\,mmol\,g_{CDW}^{-1}\,h^{-1}$ of oxaloacetate to pyruvate and, thus, a high amount of TCA replenishing oxaloacetate is needed. The TCA cycle shows high activity and the withdrawn biomass effluxes are comparatively low. As in glycolysis, high exchange fluxes were also detected in the TCA cycle; fluxes through fumarase and malate dehydrogenase were highly equilibrated (exchange fluxes of 18 and $52\,mmol\,g_{CDW}^{-1}\,h^{-1}$).

Alternative Pathways for the Formation of Biomass Components

In the metabolic network, alternative metabolic pathways can be used to form biomass components. The fluxes towards these compounds is now discussed, to identify which pathways are predominantly used for their formation. To give an example, aspartate can be formed within the cytosol via the action of an aspartate transaminase and within the mitochondria via an aminotransferase. Plotting best fits as shown in figure 7.1 results in no clear trend and the flux possessed a high standard deviation.
Acetyl-CoA can also be formed in the cytosol and in the mitochondria. The main formation pathways for acetyl-CoA have been extensively discussed [275]. Possible routes are (**I**) the direct formation in the cytosol (**II**) the formation in the mitochondria and transport by a carnitine transporter to the cytosol, and (**III**) a (citrate) shuttle system using citrate lyase [98]. Kleijn found only mitochondrial formation of acetyl-CoA and transport into the cytosol by ^{13}C-MFA, but did not provide statistical measures for these results [121]. This cannot

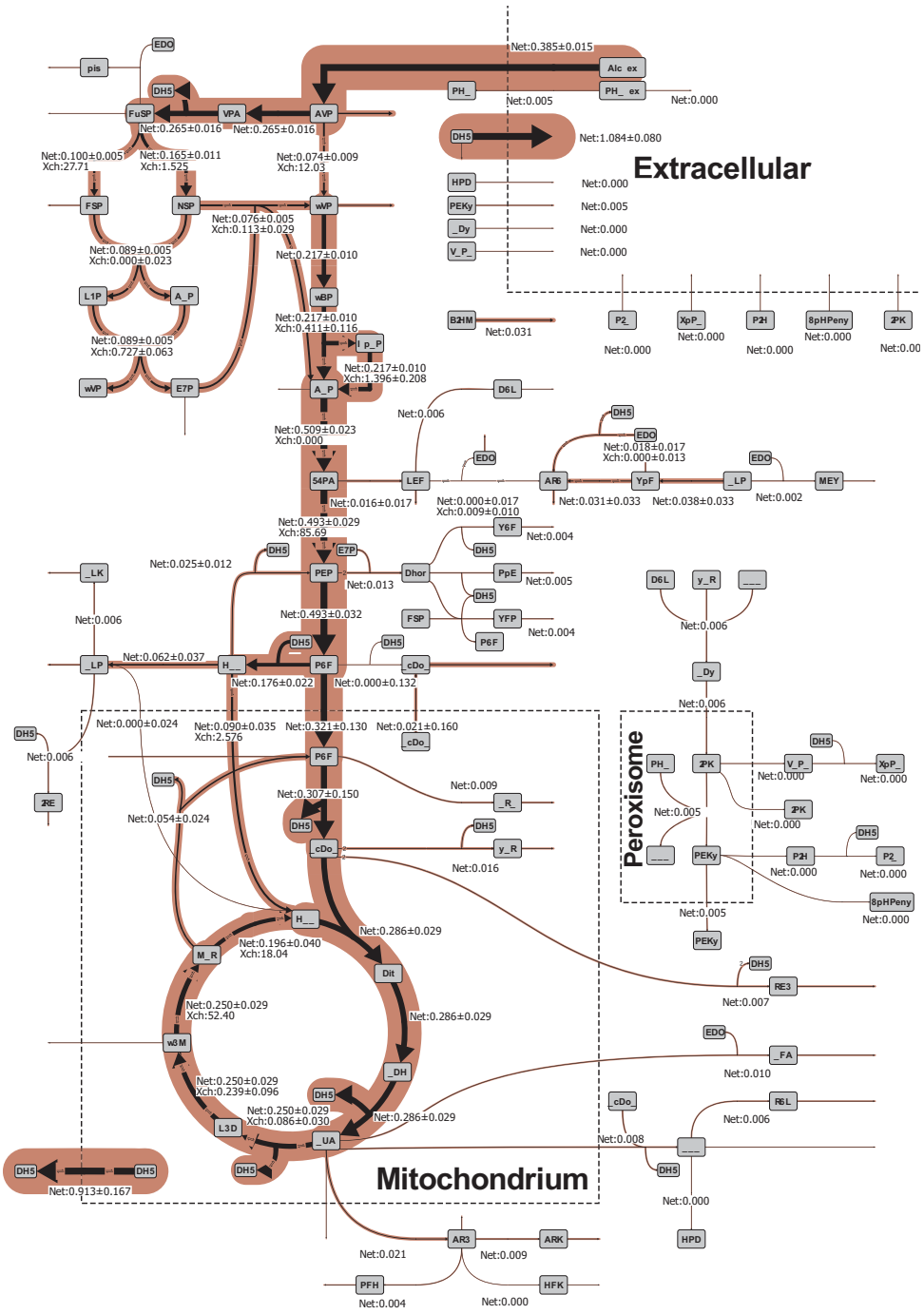
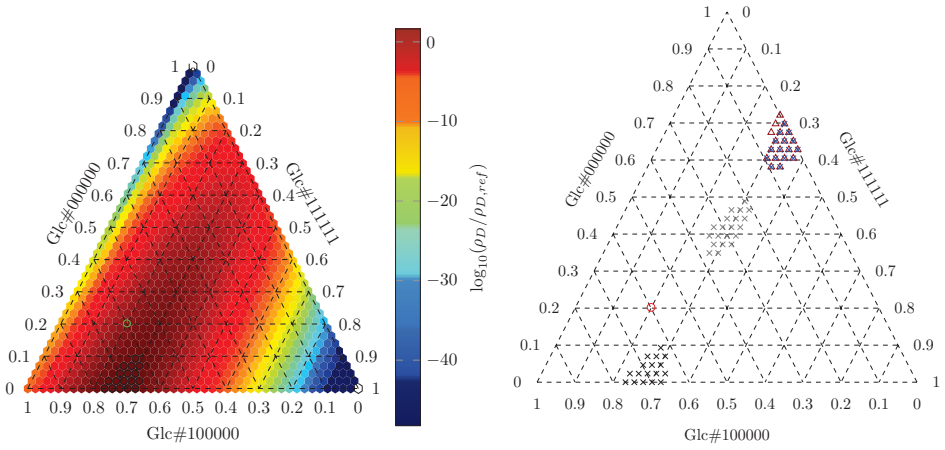


figure 7.3: Flux map for *P. chrysogenum* in chemostat 1 with a growth rate of 0.03 h^{-1} using a mixture of 60 % C1- and 20 % uniformly labeled glucose and 20 % naturally labeled glucose. Fluxes are given in [mmol g$_{\text{CDW}}^{-1}$ h^{-1}]. Biomass formation is given in [h^{-1}]. WRSS= 303.

be verified within this study by the respective fluxes, because the fluxes from cytosol and mitochondria to acetyl-CoA possessed higher standard deviations compared to the estimated flux value.
Furthermore, the formation of glycine was possible via two pathways. Glycine can be formed by serine hdyroxymethyltransferase and threonine aldolase: both possessed higher standard deviations compared to the flux value. In the end it was not possible to proof formation by the one or the other pathway. In fact, Christensen proposed presence of both pathways by ^{13}C-MFA [40].

7.1.2 Optimal Experimental Design of Labeling Mixture

An experimental design study was conducted to increase the statistical identifiability of subsequent chemostat experiments with respect to all fluxes. This was done for the chemostat 1 via using mixtures of C1-, uniformly and naturally labeled glucose. The results of the experimental design study using the D-criterion are plotted in figure 7.4a. A distinct optimal substrate mixture for the flux distribution is 70 % C1- and 30 % uniformly labeled glucose. However, the range for the optimal substrate mixture is broad and the increase in statistical identifiability is rather low compared to the first used mixture. Thus, an already close to optimal substrate mixture was used to conduct the first experiment.
Besides the D-criterion, optimal experimental design studies can be conducted, for example, based on modified A-, E-, and, C- criteria. In figure 7.4b, the best design points are shown for experimental design studies. Besides the previously calculated optimum, a second optimum for (modified) A- and C- optimal design was found at 30 % uniformly and 70 % unlabeled glucose. Interestingly, the E-criterion is showing a separate optimum demonstrating that the outcome of experimental design study is depending on the applied criterion. For in-depth discussion, see chapter 10. The 70 % C1- and 30 % uniformly labeled glucose mixture was used, as it provides the opportunity of elucidating high resolved flux map for PPP-pathways by incorporation of C1-labeled glucose.
For the optimized substrate mixture, a second set of experiments was performed using the same growth rate of $0.03\,\mathrm{h}^{-1}$ (chemostat 2) and a growth rate of of $0.05\,\mathrm{h}^{-1}$ (chemostat 3). The flux map for the chemostat 2 with growth rate of $0.03\,\mathrm{h}^{-1}$ using the optimal substrate mixture can be found in appendix C.4. The statistics in chemostat 2 with the optimized mixture were similar or slightly decreased compared to chemostat 1, although an increase was expected. Reasons may be variations in the labeling measurements causing this effect.



(a) D-criteriom (b) Optimal Experimental Design

figure 7.4: Results of the experimental design study on Fisher information matrix, based on (a) D-criterion, for the chemostat with a growth rate of 0.03 h^{-1}. (b) Optimal experimental designs for A- (△), E- (×), D- (×), and C-criterion (×). Only best 1 % of design points is shown for each criterion. The 13CFLUX2 program *edscanner* was used for calculation of 990 sampling points in the mixture triangle. The calculations are based on the criterion values obtained for the Fisher information matrix. The substrate mixture of the original experiment (○) is shown.

7.1.3 Chemostat with Growth Rate of 0.05 h^{-1}

The chemostat experiment (chemostat 3) was conducted using the optimal substrate mixture of 70 % C1- and 30 % uniformly labeled glucose and a growth rate of 0.05 h^{-1}. Under these conditions a lower specific production of penicillin V compared to the chemostats 1 and 2 was measured (0.05 h^{-1}: $3.0{\cdot}10^{-3}$ mmol g_{CDW}^{-1} h^{-1} and 0.03 h^{-1}: 5.4-8.2·10^{-3} mmol g_{CDW}^{-1} h^{-1}). Thus, a 180-270 % higher penicillin production was obtained at a lower growth rate. Contrary to the lower penicillin production, the specific glucose uptake rate was raised by ~150 % (0.05 h^{-1}) compared to chemostats 1/2. Similar results were shown by van Gulik et al., who found a maximal specific product formation rate at $\mu = 0.03\,h^{-1}$, which was decreasing towards higher growth rates up to 0.125 h^{-1} [242]. In contrast to their results, the maximal specific growth rate for the BCB1 strain under investigation is only ~0.07 h^{-1} (found in batch cultivations by Katja Schmitz at IBG-1:Biotechnologie Forschungszentrum Jülich)[242].

The final flux distribution of chemostat 3 is shown in figure 7.5. The overall flux in the network is strongly increased for a growth rate of 0.05 h^{-1} compared to lower growth rates. However, major changes in the relative fluxes could not be observed compared to the chemostat 1/2. Most interestingly, the flux distribution did change to elevated fluxes through the oxPPP, although the penicillin V production was reduced. This apparently counteracts with the current opinion in literature, claiming that NADPH is predominantly used for penicillin production. Besides, the other fluxes showed a qualitatively similar behavior as the flux map obtained at low growth rate of 0.03 h^{-1}.

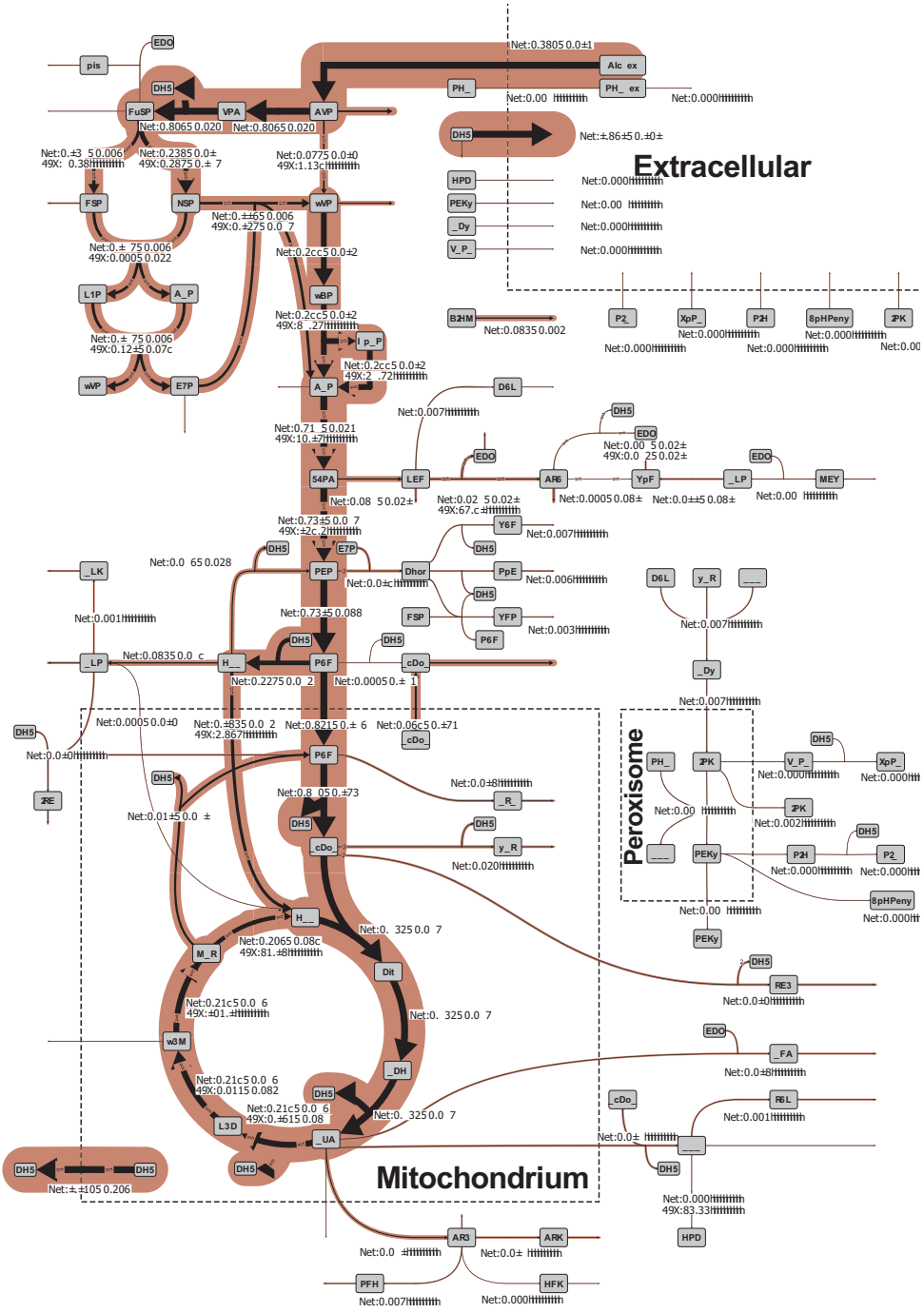


figure 7.5: Flux map for *P. chrysogenum* in chemostat with a growth rate of 0.05 h^{-1} using a mixture of 70 % C1- and 30 % uniformly labeled glucose. Fluxes are given in [mmol g^{-1}_{CDW} h^{-1}]. Biomass formation is given in [h^{-1}]. WRSS = 556.

7.1.4 Statistical Protocol for Assessment of the Flux Estimation

For each obtained flux map the WRSS was obtained from 500 model fits. In table 7.2 the minimal WRSS values corresponding to flux maps in this chapter are shown. The WRSS is in the same order of magnitude for all chemostat experiments; WRSS of 300-350 were found for a growth rate of $0.03\,h^{-1}$. Higher values were obtained for the chemostat at $0.05\,h^{-1}$ with WRSS = 556 . This is due to the fact that more measurement values at steady-state were used for this chemostat and thus, the standard deviations are reduced. The χ^2 value is used to test for the null hypothesis that the simulated data originate from the same distribution as the measured data. With n number of independent measurements and p number of parameters the null hypothesis (evaluated by a χ^2-test) is clearly rejected with a significance level of 0.05 (see table 7.2). Potentially, this indicates either that

1. the standard deviations for the measurements is under-estimated,
2. the metabolic network model is invalid,
3. there were not enough data-points to determine χ^2 accurately as it is also subject to uncertainty,
4. the measurements are effected by systematic errors like non-Gaussian distribution.

The standard deviations for the measurements might be under-estimated as only a limited number of measurements (and only one measurement platform) are available to estimate the measurements standard deviation. To prevent this, a lower boundary (0.005) for the standard deviation was assumed based on the error model obtained for the data. Thus, the data were again statistically reconciled with a lower boundary for the standard deviation of $s = 0.005$. After reconciliation, lower standard deviations were again set to this threshold and thereafter used for parameter estimation.
Additionally, measurement groups with high deviates to simulated data were removed from the model during parameter estimation procedures. The resulting number of measurements are listed in table 7.2. Of 429 MS measurements, only 221-224 single measurements remained in the network. Furthermore, some single measurements (especially m+0 mass traces) showed fluctuations in value throughout the data-sets and high deviates from simulated values. As a consequence, the respective mass traces were removed. Probably, the measurement signal was weak (raw data not shown), which led to a wrong prediction of too low standard deviations. An invalid metabolic model is rather unlikely, because a multitude of model variants were tested. Furthermore, only a low amount of data points were used; significantly more than 1,000 data points are required to obtain good estimates for this test.
Detecting potential systematic errors like non Gaussian distributed measurements is usually addressed to the wet-lab and is not the focus of this work. Development of strategies to reveal systematic errors is subject of future work. Finally, at the point in flux space, the degrees of freedom can be decreaseds. For in-depth discussion see Andrae *et al.* [6].
Rejection of χ^2-test in ^{13}C-MFA was often observed, for example by Kleijn, Leighty and Wiechert [121, 132, 266].

7.1.5 Comparison of Large and Focused ^{13}C-MFA Model

The large scale metabolic network (including 485 reactions, see chapter 5) was built with a complete set of atom transitions in 13CFLUX2 to provide a larger model to test if the found

table 7.2: Fit quality and statistical tests for chemostat cultivations. WRSS for the separate metabolic flux analysis. It is assumed that the WRSS is distributed according to a $\chi^2_{n-p}(\alpha)$-distribution with n independent data-points and p parameters and a significance level α. DOF: degrees of freedom; #Meas.: Number of measurements; #MS groups: number of MS measurement groups; #Flux M.: number of flux measurements; P: probability.

	WRSS	#Meas.	#MS groups	#Flux M.	DOF	$P(\chi^2(DOF) > \text{WRSS})$
Chemostat 1	303	221	26	13	179	$2.0 \cdot 10^{-8}$
Chemostat 2	355	224	27	13	180	$1.5 \cdot 10^{-13}$
Chemostat 3	556	217	25	13	172	$5.2 \cdot 10^{-42}$

flux distribution can be reproduced. The parameter estimation was conducted as described for the focused network using the data of chemostat 1. A comparison of the found solutions with the focused network is used to validate the presented flux maps. A comparison of statistical identifiability of both models was not performed, because they differ in the number of free fluxes significantly are not directly interpretable (88 free fluxes are needed for the large scale network compared to 36 in the focused). Besides, only 30 starting points for parameter optimization were used as the large scale network is computational more demanding for parameter estimation. The results can be seen in appendix C.5. Net fluxes in PPP showed a perfect match. Flux through glycolysis was elevated by ~10% in the large scale network. This is potentially caused by reactions consuming and producing intermediates of glycolysis. Fluxes in TCA showed higher variability; especially fluxes via malate dehydrogenase were increased in the large scale network. Additionally, the large scale network shows a low flux in glyoxylate shunt (~10% of substrate uptake rate), which was not incorporated into the focused network. Besides, fluxes through anaplerosis were reduced in the large scale network. A direct comparison of the fluxes is cumbersome and difficult as multiple pathways are present and compartment specific transport and reactions cannot be mapped easily to the focused network. The residual for the large network was increased by ~200 % (residual= 755).
Thus, ^{13}C-MFA of the focused network can reproduce net fluxes in the large scale metabolic network but is still an approximation of the complex reaction network.

table 7.3: Fit quality and statistical tests for fed-batch cultivations. WRSS for the separate metabolic flux analysis. It is assumed that the WRSS is distributed according to a $\chi^2_{n-p}(\alpha)$-distribution with n independent data-points and p parameters and a significance level α. DOF: degrees of freedom; #Meas.: Number of measurements; #Meas. Groups: number of measurement groups; #Flux Meas.: number of flux measurements; P: probability.

	WRSS	#Meas.	#Meas. Groups	#Flux Meas.	DOF	$P(\chi^2(DOF) > \text{WRSS})$
Fed-batch 1	655	259	29	13	210	$3.1 \cdot 10^{-47}$
Fed-batch 2	573	227	26	13	184	$1.6 \cdot 10^{-41}$

7.2 ^{13}C-MFA for Fed-Batch Cultivations

Fed-Batch cultivations were carried out in the sensor bioreactor with the labeling mixture of 60 % 1-, 20 % uniformly and 20 % unlabeled glucose. The growth rate was reduced (0.007-0.008 h^{-1}) compared to the chemostats. The uptake rate of glucose was also reduced (by 35-45 %) to ~0.25 mmol g_{CDW}^{-1} h^{-1} The flux distribution shows a high flux in oxPPP. The fluxes were significantly decreased in fed-batch compared to chemostats because of the lower substrate uptake rate. Effluxes to biomass were reduced, because of the low growth rate. The specific product formation rate was higher compared to the chemostat process (at least ~20 %). Here, in the close-to-industrial conditions, the flux through the oxPPP is even elevated with a split ratio of 80 % compared to the chemostats. The reaction of malic enzyme shows a high flux in mitochondria and alternative pathways for acetyl-CoA, aspartate, and glycine formation possessed higher standard deviations in the order of their flux value. The estimates for goodness-of-fit for the fed-batch cultivations are in table 7.3. The null hypothesis of the χ^2-test is rejected as in the chemostats.

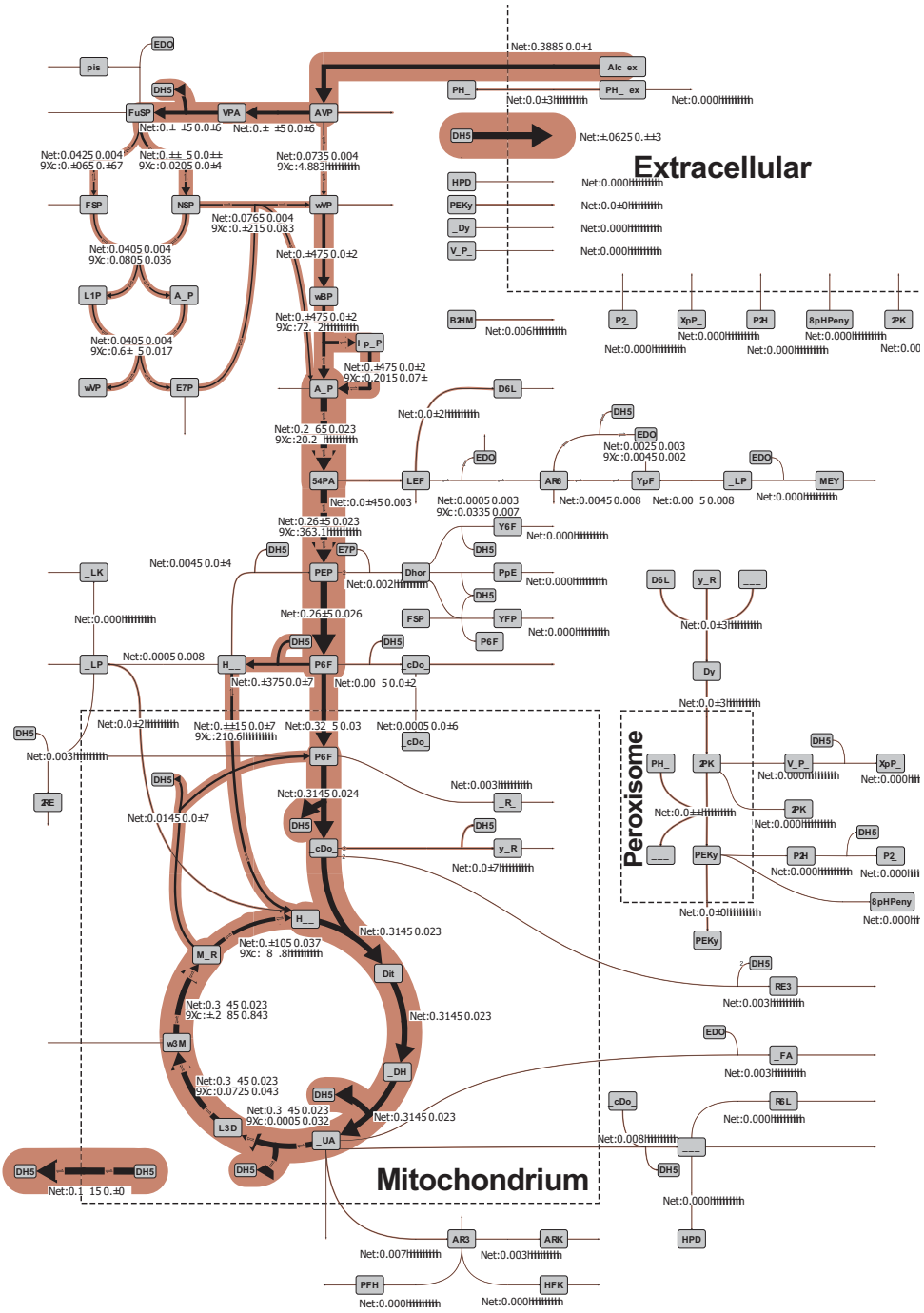


figure 7.6: Flux map for fed-batch 2 with growth rate of $0.007\,\mathrm{h}^{-1}$. Fluxes (except biomass formation rate *BIOM*) are given in $[\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}]$. WRSS=573.

Chapter 8

Discussion of ^{13}C-MFA Results

^{13}C-MFA was used to obtain insights into the intracellular flux distribution of the *P. chrysogenum* strain BCB1. A work-flow for data pre-processing was developed. Rates were estimated by kinetic modeling of the growth, substrate uptake and complex side-product formation. The LC-MS/MS measurements were extrapolated to steady-state and the data were corrected for natural labeling. A large scale network including 500 reactions and atom transitions was built and described. The focused ^{13}C-MFA network was derived from the large one and main assumptions were stated.
Highly resolved flux maps covering not only central carbon metabolism, but also the penicillin production pathway were derived along with statistical analysis of the results. For this, the state-of-the-art fluxomics tool 13CFLUX2 was used. For the first time, a series of experiments was conducted using the same strain. This was not limited to chemostat cultivations frequently applied in scientific studies, but was also extended to a close-to-industrial fed-batch process. This process was specifically designed to establish ^{13}C-MFA in industrial environment. Since the same metabolic network, strain and media were used throughout this study, the results are comparable.

Similar works have been published in the field ^{13}C-MFA for *P. chrysogenum* by Kleijn and Zhao [121, 284]. Kleijn used stationary ^{13}C-MFA for chemostat processes for penicillin G producing and non-producing conditions [121]. Zhao performed non-stationary and stationary ^{13}C-MFA for mixtures of glucose and ethanol [284]. Both were applying gluconate tracer methods to estimate the flux through oxidative pentose phosphate pathway (oxPPP) accurately, but did not provide statistical results for the overall network context [121, 284].

In this chapter, the outcome of the work is concisely summarized and discussed providing strategies for improving penicillin production.

8.1 Split Ratio

^{13}C-MFA was applied to batch, fed-batch and chemostat cultivations. A high flux through oxPPP is beneficial for *P. chrysogenum* producer strains. The oxPPP is known to be the main supplier of the reduction equivalent NADPH. It is used for the biosynthesis of valine and cysteine, where cysteine formation consumes the bulk of NADPH [282]. Both amino acids are needed in high amounts for penicillin production. Thus, observing the flux via oxPPP in

conjunction with penicillin production can provide insights into the usage of NADPH.

Chemostat

A high flux through the oxPPP was found in all chemostat experiments. The oxPPP split ratio is calculated as the flux through oxidative oxPPP divided by the substrate uptake rate. At a growth rate of $0.03\,h^{-1}$ oxPPP split ratios of 69 % were observed. The penicillin V production was varying in terms of carbon efficiency[1] between 2.2-3.4($mol_C\,mol_C^{-1}$) % for chemostats 1 and 2, which was probably due to process variability. The split ratio was also high (75 %) at a specific growth rate of $0.05\,h^{-1}$ while the penicillin V production was decreased to less than ~50 % compared to experiments with $0.03\,h^{-1}$. Importantly, at higher growth rates a decrease of split ratio would be expected because penicillin production was found to be decreased. For example, Kleijn found a decrease from 45% split ratio to 23% for producing and non-producing conditions, respectively [121]. Producing conditions are characterized by addition of the side-chain precursor to the cultivation medium. Additionally, the flux of carbon towards biomass was increased by 12-17 % at a higher growth rate. This behavior points to another effect present during cultivation of *P. chrysogenum* which is not directly linked to production of penicillin. It demonstrates a strong coupling of NADPH metabolism to growth, which was not yet shown in literature for *P. chrysogenum.*
In literature, several ^{13}C-MFA and purely stoichiometric-driven MFA studies have been performed in chemostat or in fed-batch cultivations and split ratios obtained with this technique are ranging from 44 to 75 % for producing conditions (see table 8.1). All in all, literature data were lower or comparable to the found values in this work. To give some examples: Christensen *et al.* measured similar high split ratios in *P. chrysogenum.* They computed stable split ratios for producing (70%) and non-producing conditions (71%) [41]. In this and another study by Christensen *et al.* the production of penicillin was also comparable to the results for the BCB1 strain with a split ratio of 75 %[40]. They speculated that the capability of a strain for realizing a high split ratios may be associated with its potential for penicillin production [41].
A high flux (61%) by stoichiometric metabolic flux analysis through the oxPPP at elevated dilution rates was found by Henriksen *et al.* [98]. The high value in their study is mainly accounting for an assumed presence of cytosolic malic enzyme and valine synthesis in cytosol. In this study, both were assumed to be mitochondrial. Notably, Kleijn and Zheng calculated the split ratio based on a stationary gluconate tracer method and revealed split ratios of about 51 % and 41 % with high statistical significance [284, 122].
However, Kleijn *et al.* and Christensen *et al.* did not provide statistics for a full flux distributions, only for the oxPPP split ratio [122, 40, 41]. Zheng *et al.* provided flux maps including statistics for growth on different glucose ethanol mixture for glycolysis and pentose-phosphate pathway, but not for the whole central carbon metabolism [284].

To summarize, for chemostat cultivations highly resolved flux maps of the central metabolism including statistics were generated for *P. chrysogenum* strain BCB1. Additionally, a strain specific high flux through oxPPP was observed for the BCB1 strain. It was shown for the first time that an elevated flux through oxPPP is, besides penicillin production, significantly effected by growth in *P. chrysogenum.* This was shown with high statistical confidence.

[1] Carbon efficiency gives the ratio of carbon found in substrate (glucose) and product (penicillin nucleus).

table 8.1: Comparison of oxPPP split ratio with literature data for *P. chrysogenum*. INST-^{13}C-MFA=non-stationary ^{13}C-MFA.

Method applied	Process, Strain	$\mu[h^{-1}]$	oxPPP-split ratio	Citation
S-MFA	Chemostat; glucose; DS12975	0.01	18-44 %	van Gulik et al. [242]
^{13}C-MFA; INST-^{13}C-MFA	Chemostat; glucose and gluconate; DS17960	0.06	40.6-41.2 %	Zheng [284]
INST-^{13}C-MFA	Chemostat; glucose and ethanol; DS17960	0.05	52 %	Zheng [284]
^{13}C-MFA	Chemostat; glucose, POA; Wis54-1255, NN P8	0.06-0.07	66-71 %	Christensen et al. [41]
^{13}C-MFA	Chemostat; glucose; Novo Nordisk strain	0.08	75 %	Christensen and Nielsen [40]
S-MFA	Fed-Batch; glucose; Novo Nordisk strain	$\leq$0.1	18-44 %	Jorgensen et al. [111]
^{13}C-MFA	Chemostat; glucose; DS17690	0.02	23-45 %	Kleijn [121]
^{13}C-MFA	Chemostat; glucose and gluconate; DS17690	0.02	51 %	Kleijn et al. [122]
S-MFA	Chemostat; sucrose; Novo Nordisk strain	0.1	61 %	Henriksen et al. [98]
S-MFA	Chemostat; glucose; DS17690 and AFF230	0.05	29-37 %	Nasution et al. [170]

Fed-Batch

The high-producing BCB1 strain was cultivated in a close-to-industrial fed-batch process using the sensor-bioreactor concept. Compared to the chemostat process, even higher fluxes through oxPPP were observed by ^{13}C-MFA, ranging within 74-80 % split ratio. Notably 6.8-6.9 % of the carbon was used in the penicillin formation pathway. However, growth was significantly reduced to 0.007-0.008 h^{-1}. In the fed-batch process the expected high flux through oxPPP demanded by penicillin production is observed.

For the close-to-industrial fed-batch process highly resolved flux maps including statistics could be generated. An increased split ratio for the high-producing conditions was found with high statistical significance. These are the first flux maps using ^{13}C-MFA for a high-producing strain in industrial environment.

8.2 NADPH Balance in the Cytosol

Balancing of NADPH in the cytosol is possible based on cofactor utilization in the separate pathways. For amino acid synthesis 1 $mol_{NADPH}\,mol_{Aa}^{-1}$ is used for transamination. For fatty acid synthesis 2 $mol_{NADPH}\,mol_{AcCoA}^{-1}$ are consumed. Additionally, for the biosynthesis of cysteine 5-8 $mol_{NADPH}\,mol_{CYS}^{-1}$ and valine 1 $mol_{NADPH}\,mol_{VAL}^{-1}$ are consumed. Furthermore, proline (1-2 $mol_{NADPH}\,mol^{-1}$), lysine (2 $mol_{NADPH}\,mol^{-1}$), chorismate (1 $mol_{NADPH}\,mol^{-1}$), and ergosterol formation (19-23 $mol_{NADPH}\,mol_{ES}^{-1}$) consume NADPH. The ranges for NADPH production are founded on the existence of alternative pathways or enzymes utilizing either NADH or NADPH. Those were the main reactions participating in NADPH consumption.

table 8.2: oxPPP split ratio and NADPH balance. Penicillin carbon efficiency is computed by the ratio of carbon flux through penicillin pathway in respect to the carbon uptake.

	Growth Rate	oxPPP-split ratio	NADPH balance closed	Penicillin Carbon Efficiency $[mol_C\, mol_C^{-1}]$
Chemostat 1	0.03 h^{-1}	69 %	47-52 %	2.2 %
Chemostat 2	0.03 h^{-1}	69 %	45-52 %	3.4 %
Chemostat 3	0.05 h^{-1}	75 %	42-46 %	1.3 %
Fed-Batch 1	0.008 h^{-1}	80 %	31-41 %	6.9 %
Fed-Batch 2	0.007 h^{-1}	74 %	35-45 %	6.8 %

Formation of NADPH occurs via the oxPPP reactions and (cytosolic) isocitrate dehydrogenase, whereas the latter was not quantifiable and is neglected for NADPH balancing. What makes the NADPH resolution difficultis that many other pathways are involved in the consumption of NADPH. This oberservation is not limited to NADPH. Zhao also revealed a gap in ATP balance and estimated the influence of substrate cycles in *P. chrysogenum* [284]. Substrate cycles occur, for example, in the formation of polyols and polysaccharide degradation. These substrate cycles accounted for 52.1 % of the ATP gap in metabolic flux analysis [284].

The production of penicillin was high in the fed-batch and chemostat cultivations compared to literature. Surprisingly, most of the carbon is used for growth associated processes. A maximum of 6.9 % of the carbon of glucose was found in penicillin V and side products. Calculating the NADPH balance based on the ^{13}C-MFA results, reveals that it is closed to 45-52 % for chemostats with a growth rate of 0.03 h^{-1} (see table 8.2). Notably, at a higher growth rate of 0.05 h^{-1} the NADPH balance was closed to only 42-46 %. A short overview about possible NADPH sinks is given in the next section. First, the consumption of NADPH in the metabolism during formation of biomass and penicillin V is discussed.

In figure 8.1 the consumption of NADPH is quantified in more detail for the different cultivations. The ratio of NADPH consumption to NADPH production is shown. Ranges are given for NADPH consumption for some biomass building blocks. For these biomass constituents alternative pathways exists, e.g. cysteine consumes 5-8 $mol_{NADPH}\, mol_{CYS}^{-1}$ depending on the pathway used. In the fed-batch process, NADPH is predominantly used in biosynthesis of penicillin V, especially in the cytosolic formation of cysteine. The formation of valine consumes cytosolic NADPH by transamination of cytosolic formed glutamate. Additionally, NADPH is used for amino acid transamination, fatty acid synthesis and *de-novo* synthesis of other compounds. In the chemostats with a growth rate of 0.03 h^{-1} a higher fraction of the produced NADPH is used to produce biomass components, whereas a significant amount is consumed by penicillin production. The amount of NADPH used for penicillin formation is reduced for the chemostat cultivation with a growth rate of 0.05 h^{-1}: high amounts of NADPH are required for growth instead. Also other authors have described a gap in NADPH-balance for chemostat cultivations [40, 41].

Besides, NADPH is also consumed by reactions in the mitochondria, especially for the formation of the amino acids valine (1 $mol_{NADPH}\, mol^{-1}$) and isoleucine (2 $mol_{NADPH}\, mol^{-1}$). In mitochondria NADPH is produced by isocitrate dehydrogenase and malic enzyme [2]. Additionally, NAD(P) transhydrogenases may deliver NADPH as well (see published *P. chrysogenum* GEM [2]). Thus, in mitochondria NADPH balancing is not possible as alternative routes for the production are present.

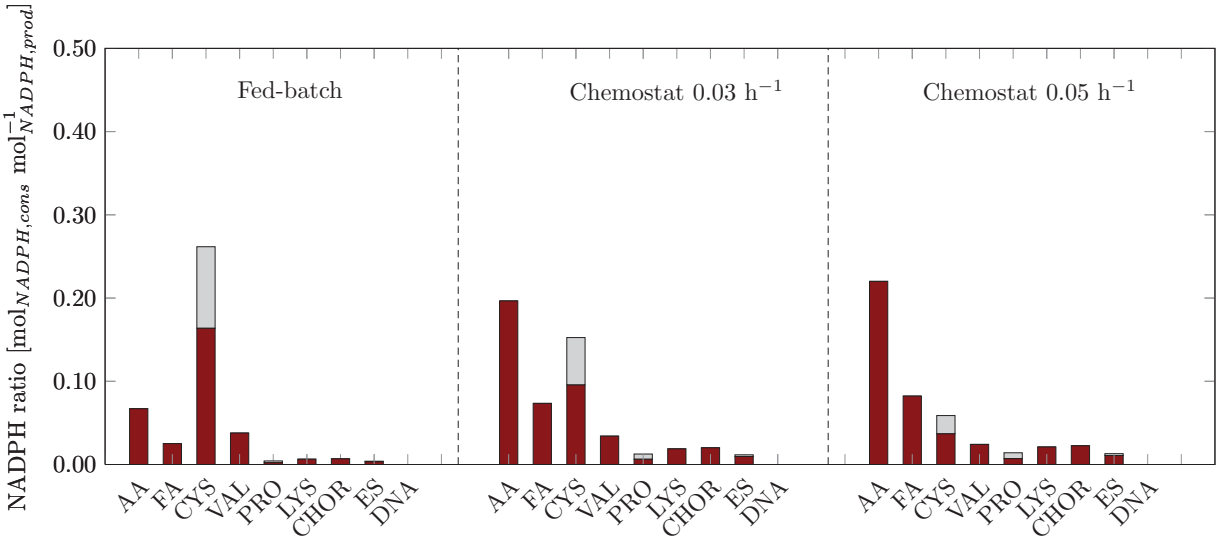


figure 8.1: NADPH balancing based on ^{13}C-MFA results for fed-batch and chemostats (data for chemostat 2 shown). The ratio of NADPH producing to NADPH consuming reactions was visualized. The minimal (▬) and maximal (▬) NADPH consumption is shown. AA: Transamination of amino acids (without CYS and Valine); FA: fatty acids; CYS: cysteine, VAL: valine; PRO: proline; LYS: lysine; CHOR: chorismate; ES: ergosterol.

8.2.1 Sinks for NADPH

Flux balance analysis, which was described in chapter 5.2, revealed that theoretically a higher penicillin formation is possible compared to experimentally derived data. By NADPH balancing in the cytosol it was shown that about 40-50 % of the NADPH is used for unknown cellular processes. This motivates to have a closer look on the NADPH consumption in *P. chrysogenum*. Several hypotheses are formulated for the utilization of NADPH, which could not be assigned to biomass formation or penicillin production.

Phenoxyacetate (POA) Degradation After addition of POA to cultivations of *P. chrysogenum*, increased activity of enzymes for glutathione (GSH) producing reactions were detected [70]. Thus, addition of POA causes an increase of enzymatic activity in pathways using NADPH to break down reactive oxygen species (ROS). POA addition induces oxidative stress response, even without formation of penicillin and a growth reduction [70]. POA is subject to degradation to HOPOA, a side-product found in several studies cultivating *P. chrysogenum* [42, 69]. Emri *et al.* suggested that upon this reaction toxic intermediates like epoxides are formed. In their study, the author could also show that the addition of HOPOA to the cultivation did not result in increased level of intracellular reactive oxygen species. An up-regulation of oxidative stress as in case of POA addition was not observed [70].

Thus, after addition of POA a toxic intermediate may be formed, which causes an increased activity of glutathione dependent pathways. This product is finally converted to HOPOA, which is not further degraded. To break down the formed reactive oxygen species NADPH is used.

Uncoupling Effect of POA POA is known to exhibit protonophoric uncoupling effect of membrane potentials by passive diffusive uptake. This effects was known to be present at pH 5.0. The cultivations were performed at pH 6.5, resulting in negligible protonophoric uncoupling described by Henriksen *et al.* [100]. Also the oxidative phosphorylation pathway might be subject to physiological uncoupling effects like mitochondrial unspecific channels, uncoupling proteins, or electron slipping (variations of H^+/e^- stoichiometry) known in yeast (see Guerrero-Castillo *et al.* for discussion [91]). However, little data is available for *P. chrysogenum* in this field of research.

Penicillin Production Thykaer *et al.* claimed that unknown processes in the consumption of NADPH in the penicillin pathway can be present [235]. Importantly, the flux maps and NADPH balances show no clear correlation of penicillin production and the NADPH gap, besides the major gap is also present if low amount of penicillin is formed (at growth of 0.05 h^{-1}).

NADPH-Dependent Dehydrogenase The presence of a cytosolic NADPH-dependent dehydrogenase was recently indicated by Harris *et al.* [94]. The dehydrogenase reduces the efficiency of the electron transport chain and consumes NADPH for the production of ATP. This may explain why NADPH balances cannot be closed.

Summing up, the hypotheses and known pathways are summarized in figure 8.2. Further experimental evidence is needed to verify which of the effects contribute most to the NADPH gap.

8.3 Hints for Strain Improvement

In general, NADPH availability seems to be the bottleneck for penicillin production. Based on the results of flux balance analysis (chapter 5.2), it can be assumed that a high flux through oxPPP will support higher penicillin formation by increasing NADPH supply. Additionally, it was shown by FBA that increasing flux through cytosolic NADPH producing isocitrate dehydrogenase will further increase the penicillin production.
The results of ^{13}C-MFA have shown that, besides penicillin production, growth has a major impact on the NADPH consumption in the cell. Thus, growth needs to be carefully controlled to avoid high usage of NADPH for growth associated processes. In the end, strain manipulation is highly coupled to the process conditions. Thus, complex protocols are required to balance growth in cultivation process and further strain improvement by metabolic engineering.

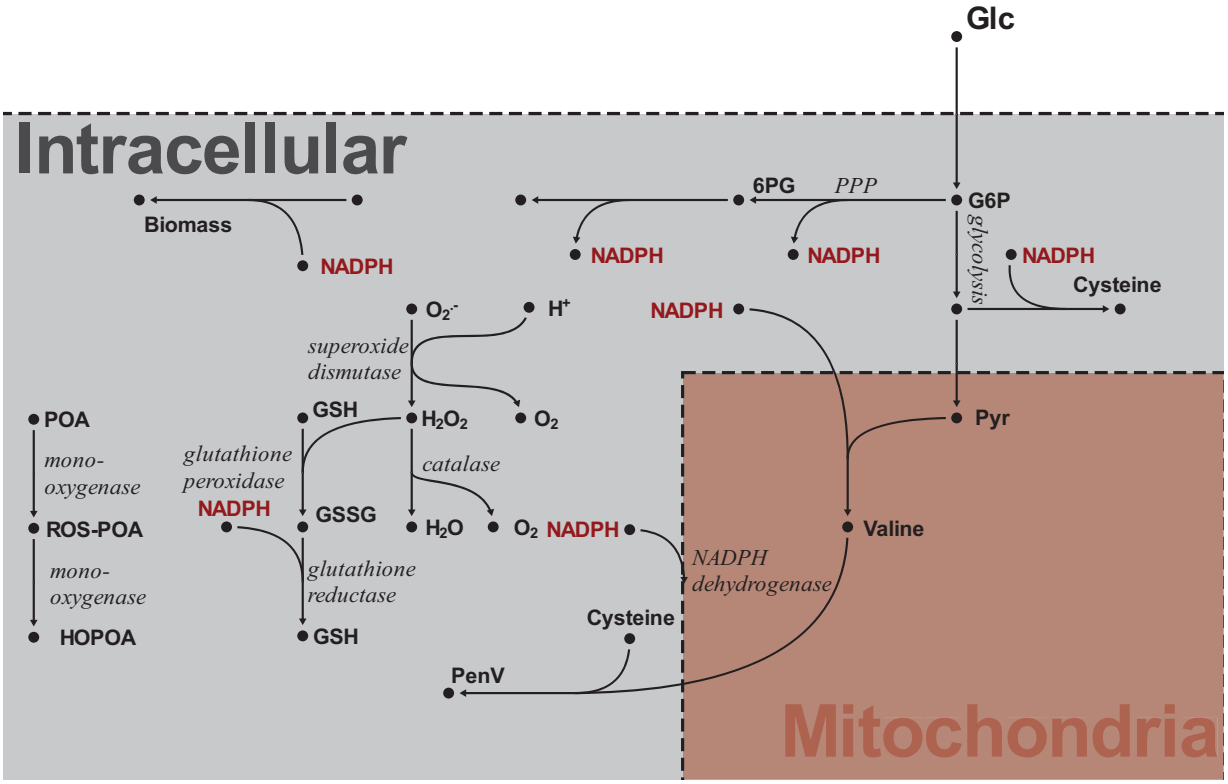


figure 8.2: Hypothetical NADPH utilization in *P. chrysogenum* cultivations

Part III

Design of Experiments for ^{13}C-MFA

Chapter 9

Diversification-Driven Experimental Design

Rarely, single experiments deliver enough information to identify all parameters of interest in a model. Multiple experiments can be conducted to improve accuracy of parameters. Besides, already conducted experiments contain information which can be integrated into the evaluation process. This was already shown by Crown *et al.* for mixtures of fully labeled and unlabeled glucose in parallel experiments [47]. For the integration of multiple experimental data-sets into one analysis, main influencing factors to improve flux accuracy are

- a complementary labeling mixture for the separate experiments and
- a multiplied number of measurements.

The experiments to obtain these data can be conducted sequentially and in parallel, the basic principle was already described by Franceschini and Macchietto [85]. The basic assumption is, that the fluxes are the same in each the experiment. In this chapter we are aiming at a sequential, learning-based, experimental design. The basic idea of Crown *et al.* to use multiple experiments for flux estimation will be extended later by experimental design to obtain assisted integration of multiple data-sets into one analysis [47]. This new strategy is called diversity-driven experimental design. In comparison to conventional experimental designs, it aims at step-by-step increasing statistical identifiability by sequential conduction of experiments. Additionally, the method can be easily extended for planning of parallel experiments.
In this chapter, the basic idea to combine multiple measurements of carbon labeling experiments in one analysis is described. Next, the method is used to increase the statistical significance of an experiment and, finally, optimal diversification-driven experimental design is introduced.

9.1 Combining Experimental Data in ^{13}C-MFA

After the first cycles of experimental design, several fluxes may still be non-identifiable. To increase the statistical identifiability of the fluxes further, integration of measurements from different experiments is needed. The increase in statistical identifiability is measured by design criteria ρ (introduced in chapter 2.2.7). To incorporate multiple labeling data-sets

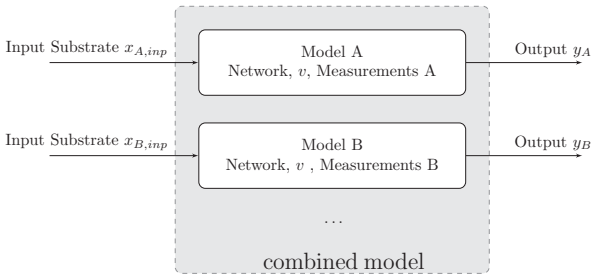


figure 9.1: Construction of a combined model to combine multiple measurement sets of experiments into one analysis. The input substrate mixture was changed for each experiment.

into one model, a suitable algorithm is needed to extend the model structure, which will be given below. First, the new model structure is described.
Two independent replicates of a carbon labeling experiment are used here, each possesses another substrate mixture. Incorporation of both data-sets into one analysis will result in an information gain. Only the known parameters (the substrate mixture x_{inp}) of the model under investigation are varied, resulting in a different model output behavior of the system (labeling measurements). By varying the known parameters, the information about the unkown parameters (fluxes) is maximized. For the given data, a so-called combined model is generated. Each of the sub-models A and B contain the respective input substrate mixture of the experiment and simulate an own set of measurements y_A and y_B. The sub-models contain thus separate measurement specifications for the experiments and simulate them at a time (see figure 9.1). It is important to note that both models share the same flux values v and network topology. Next, the generation of such a combined model is described for ^{13}C-MFA.

9.2 Preparing Models for Combined Experiments

To incorporate labeling measurements emerging from separate experiments with varied labeling of the substrates, the combined model needs to be constructed from a normal ^{13}C-MFA model. A normal model contains the metabolic reaction network including atom transitions, input substrate specification, constraints and measurements (henceforth called root model). The combined model has to account for separate input labeled substrates and measurements of labeling patterns in the independently conducted experiments. Therefore, the model structure is cloned. Basically, two possibilities exist to multiply the model structure. First, all reactions can be inserted twice and the cloned reaction is renamed. This results also in a multiplied constraints section. The free fluxes of the original models are set equal to the second model. Another possibility is to duplicate the metabolites in each reaction, this is described in the appendix D.

9.2.1 Results for Duplicated Models

In chapter 7.1 two chemostat experiments were conducted with a specific growth rate of $0.03\,\mathrm{h}^{-1}$. The statistical identifiability of the fluxes was tried to be increased by conventional experimental design (cf. chapter 7.1.2). However, choosing an optimal substrate mixture resulted not in a significantly higher statistical identifiability of the fluxes. The labeling

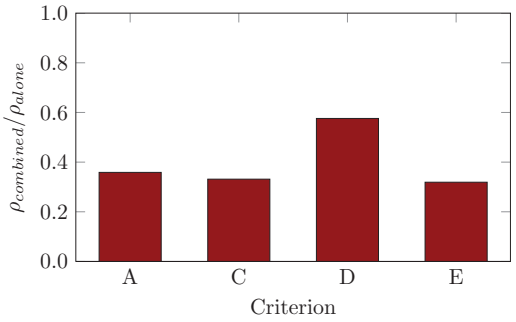


figure 9.2: Comparison of criterion values of chemostat 1 (ρ_{root}) and combined data of chemostat 1 and 2 ($\rho_{combined}$). For comparison of statistics, the same free fluxes were set constant in both models. 17 free fluxes remained. A criterion: $trace(Cov)/p$; C criterion: $max(trace(Cov))$; D criterion: $det(Cov)^{1/p}$; E criterion: $\lambda_{max}(Cov)$

mixture in the experiments was 60 % Glc#100000, 20 % Glc#111111 and 20 % Glc#000000 in chemostat 1 and 70 % Glc#100000, 30 % Glc#111111 in chemostat 2. Here, both data-sets can be integrated in one analysis using a combined model as previously described.

The combined model comprised of 147 pools and 185 reactions was built by applying above mentioned algorithm. It contains 445 labeling measurements and 26 flux measurements (with 382 degrees of freedom). After parameter estimation using *fitfluxes* of 13CFLUX2 (using 500 samples for globalized multi-start strategy), a WRSS of 893 was obtained, thus increased compared to chemostat 1 and 2 with a WRSS of 303 and 355, respectively.

Now, it is possible to compare the results of single and combined experimental data. Chemostat 2 showed less optimal statistics and and is thus not chosen as comparison. The information criteria were evaluated for chemostat 1 and for the combined model based on the covariance matrix. Clearly a substantial decrease in A,C,E-criteria to ~40% of the root model (chemostat 1 alone) was found, the D-criterion decreased to 60%. The increase in statistical identifiability was also reflected in the standard deviations of the fluxes. The corresponding flux map including standard deviations is given in the appendix C.

In the newly generated flux map of the combined model fluxes towards aspartate became better statistical identified. Two possible pathways, one via mitochondrial and one via cytosolic oxaloacetate were incorporated. Using solely the data of chemostat 1, the main flux was directed via cytosolic oxaloacetate towards aspartate with $0.062 \pm 0.037\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$ and from the mitochondria with $0.000 \pm 0.024\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$. With the combined model we receive better statistical estimates for cytosolic and mitochondrial production with $0.047 \pm 0.013\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$ and $0.000 \pm 0.008\,\mathrm{mmol\,g_{CDW}^{-1}\,h^{-1}}$. The flux is even decreased, because aspartate is used less for threonine production. Thus, this is a strong indication that the compartment specific formation of aspartate proposed by Kleijn was correct, which was not possible using the single data-sets [121].

From these finding a new experimental design method was developed, called optimal diversification-driven experimental design.

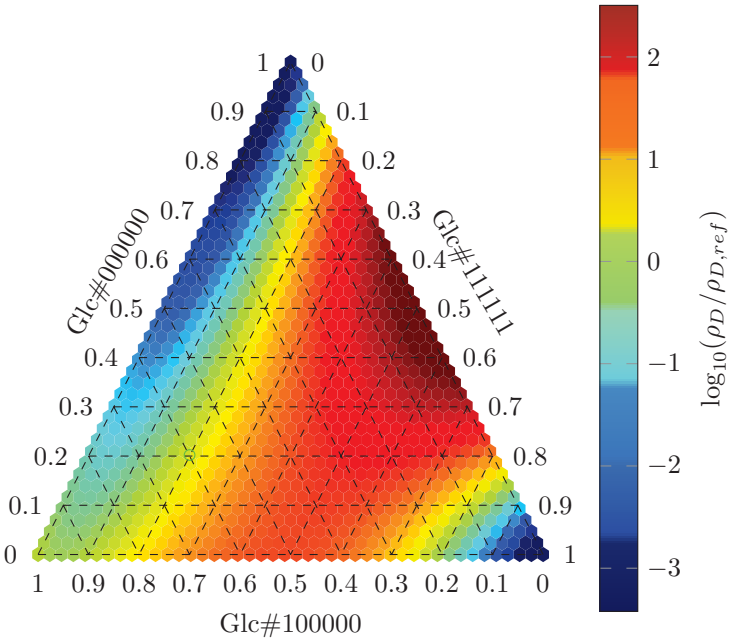


figure 9.3: Results of diversification-driven experimental design for the focused *P. chrysogenum* network used in chapter 7.1 for a chemostat 1 with 0.03 h^{-1} based on D-criterion using Fisher Information matrix.

9.3 Optimal Diversification-Driven Experimental Design

Optimal diversification-driven experimental of design (ODED) aims at increasing the statistical significance of a chosen objective by a second experiment. Thus, it is basically a sequential design, where a sequence of experimental design steps and experiments are conducted. The results of each experimental run are used in the combined model. Basic underlying assumption is that the process has to be performed reproducible, so that the fluxome is not changing. After conduction of the first experiment with a mixture $x_{A,inp}$, the estimated flux distribution $\hat{v}$ is obtained. A design criterion ρ is used to find a new mixture $x_{B,inp}$ with high statistical identifiability. Additionally, a distance operator δ is maximized to increase the difference labeling pattern of the substrate (for example Euclidean distance can be used). This is done to increase the diversity in the used substrate mixtures.

$$\max_{x_{B,inp}} Z = [\rho, \delta(x_{A,inp}, x_{B,inp})] \tag{9.1}$$

This was formulated as a multi-objective optimization, because more than one design point can be suitable. Introducing of further criteria is possible, as outlined later in this study (chapter 10). By maximizing the distance operator and the information criterion, ODED can be planned.

9.3.1 Application Example

To plan an optimal ODED for the chemostat 1, experimental design principles are used on the built combined model. The measurement values of chemostat 1 are incorporated for both

the root model A and model B. An experimental design study is performed using *edscanner* of 13CFLUX2. For each point in a ternary mixture triangle the statistics and design criteria are calculated. Only for model B the substrate mixture is changed.
In figure 9.3 the result for an ODED for chemostat 1 is shown in a ternary mixture diagram. The used mixture for the first experiment was 60 % 1-labeled glucose, 20 % uniform and 20 % unlabeled glucose. All calculated criterion values for the D-criterion ρ_D were normalized to the criterion value $\rho_{D,ref}$ of the first conducted experiment (green circle). To search for an ODED, high statistical identifiability of the fluxes and, at the same time, high distance from the old mixture are aimed at. Thus, a design points possessing a high statistical identifiability and high distance to the first conducted experiment was found for 50 % fully labeled and 50 % unlabeled glucose for the D-criterion. Besides, there are further design points available as we are dealing with a multi-objective optimization problem. However, we assume that an optimal diversification-driven design point was chosen at 50 % fully labeled and 50 % unlabeled glucose. Using this point, the statistical identifiability can be further increased compared to 70 % 1-labeled glucose and 30 % uniform glucose used beforehand for chemostat 2.
Notably, ODED found deviating points compared to conventional experimental design in chapter 7.1.2. Conventional experimental design proposed 70 % 1-labeled glucose, 30 % uniform glucose used in chemostat 2. Thus application of the ODED can result in counter-intuitive solutions, deviating from conventional experimental design design points.

9.4 Findings and Conclusions

Combining several data-sets in a combined model leads to the finding that statistical identifiability is significantly increased. This gives rise to a new strategy called ODED. It extends the applicability of optimal experimental design by introducing independently generated measurements of separate experiments into one combined model. The algorithms to built these combined models were introduced. By choosing the experimental conditions based on the reformulated optimality criterion for ODED, new information content is maximized.

Chapter 10

Economic Multi-Objective Experimental Design: Exploration of Design Space

In the last decade optimal experimental design (ED) has emerged as a powerful toolbox with application in science and industry [85, 93, 186]. The field of application of the methodology is broad and comprises maximizing quality, productivity, or reduction of variance of characteristics of interest. Several techniques have been applied (screening, response surface, and model validation techniques) in many fields of biotechnological and pharmaceutical research [93, 17]. One field of application is ^{13}C-MFA; which provides insights into the metabolism of organisms. The methods aims at identifying reaction rates (fluxes) in living cells by introducing and measuring labeling patterns in them, providing indirect measures for flux values. The result is a high dimensional system. Finally one can pose questions regarding the labeling experiments:

- What is possible (in terms of identifiability) within a measurements platform?
- How much does it cost?
- Is it worth the effort?

The heavy use of expensive analytics and process control techniques, strengthens the need for robust and economic ED.
Several researchers have conducted optimal ED for stationary and in-stationary carbon labeling experiments to elucidate metabolic fluxes. Möllney *et al.* were introducing the basic principles to ^{13}C-MFA [161]. But the major problems in statistical and structural identifiability have never been addressed in depth so far [213, 46, 256, 181, 272, 10]. Besides, a method to decide which of the available measurement techniques provides better statistics has never been proposed.
This chapter focuses on identifiability of metabolic fluxes in the context of ED. A large number of so called alphabetical optimality criteria have been proposed to measure and compare the statistical identifiability of estimated parameters. They all provide a single-number-measure for describing the statistical identifiability of the system, expressed by the covariance matrix. Calculating these criteria is possible when prior knowledge of the parameters, by experimental or theoretical investigations, is available. From the first ED a next guess for optimization

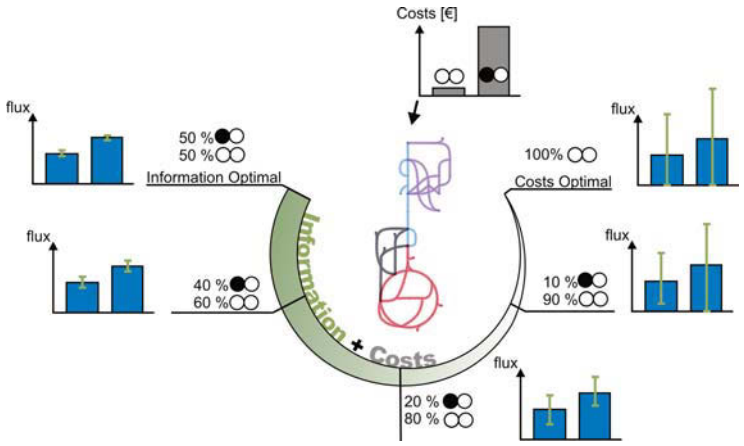


figure 10.1: Economic multi-objective experimental design for ^{13}C-MFA.

of measurement procedure is given, leading to adaption cycles. These criteria are a drastic reduction of the information contained in the covariance matrix condensed into a single value. This may also lead to unintentional effects, as the modeler is not capable of directly assessing the design space. Thus, as frequently indicated in literature, false-predictions based on these criteria are possible [88, 16]. Especially the D-criterion seems to be prone to errors as it might lead to needle-shaped confidence ellipsoids, which are unfavourable because still large confidence intervalls are possible for a parameter. However, this criterion is most frequently applied [88, 16].

Additionally, the number of usable ED criteria is steadily growing. Besides the alphabetical optimality criteria, there exists also a growing number of non-alphabetical criteria dealing with covariance matrix independent measures for an optimal model. Some have been mentioned in literature like local measures of Bates and Watts for non-linearity [21]. Additionally, the cost of an experiment is crucial in biotechnological fields as experimental setups can be expensive. Thus, usage of non-economic demands will lead to unfeasibly high cost, because it is possible that other experimental setups can provide the same statistical significance at lower costs.

In this chapter an approach is presented to circumvent the drawbacks of single criteria in ED and a work-flow is given to analyze the resulting high dimensional data of design criteria, providing insights into the interplay of substrate mixture, measurement setup, identifiability, and costs (see figure 10.1). It enables the researcher to integrate criteria of interest and plan economic, yet informative, experiments thereby exploring optimal ED space and circumvent drawbacks of single or fixed ratio-objective designs. The presented work-flow is directly transferable to other fields of studies and gives general recommendations for reliable EDs and extends it by non-alphabetical criteria which are crucial for their evaluation.

First it is described how to integrate multi-objective optimization into ED studies. This is used to explore the designe space for D-criterion and costs. Next, the same is repeated for several design criteria (costs, D-, E-, and A-criterion) and the results are compared to the D-criterion and costs results. The robustness of the calculations is tested and the studies are repeated for several measurement devices. By these results, general recommendations for ED are derived.

10.1 Theory

10.1.1 ^{13}C Metabolic Flux Analysis

In ^{13}C-MFA isotopic labeled substrates are used with distinct labeling patterns. The stable isotope ^{13}C of carbon atoms are incorporated in the substrates fed to a living cell. After the cell and its environment reaches a steady-state, the external rate measurements of substrates, products and biomass are direct estimates for extracellular fluxes. Importantly for the method, indirect measurements of intracellular fluxes are given by measuring the labeling state of intracellular metabolites. To obtain the intracellular fluxes the cell‘s biochemical system has to be modeled. First, a stoichiometric system of equations is given by building a model of biochemical reactions representing a cell‘s state as close as possible. In stationary ^{13}C-MFA, which will be the main focus of this chapter, stationarity assumptions for metabolic and isotopic conditions are applied.

Several of the given stoichiometric reactions contain reactants with carbon atoms. Atom transitions, the carbon fate from educt to product, can be specified [263, 163]. By choosing a set of free parameters (free fluxes) v_{free} and using the same input substrate labeling state x_{inp} as in the conducted experiment, the system can be solved to calculate the isotopomers x for each metabolite. In the following variable w corresponds to direct influencing factors like the experiment specific state variables (temperature T, dilution rate D in a chemostat, concentration of media components like glucose C_{Glc} , ...) and unknown effects:

$$f(x, v_{free}, w) \tag{10.1}$$

The measurable response variable y (for example a mass spectrometer measurement) depends on the state variables x in the network and is computed. In an experiment, y is measured (by MS or NMR techniques) and is subject to uncertainty. These uncertainties are propagated to the free fluxes, resulting in estimated parameters $v_{free} \pm std$ with a standard deviation. The standard deviation of a parameter is influenced by the used model, available measurements and their respective measurement error (which is, in turn, influenced by w).

In an ED study for ^{13}C-MFA, one aims at improving the statistics of the parameters (v_{free}), usually by changing the input substrates while knowing the flux distribution. The state variables and a measurement matrix M of the measured quantities can be calculated, which are labeling measurements and extracellular fluxes. Basically, the measurements can be represented by scaled sums of isotopomer fractions for all measurement types and are, thus, linear equations. The measurement matrix sums isotopomers x to form MS or NMR measurement vector y. The structure of the measurement matrix is influenced by the measurement setup and metabolites measured. Some examples for measurement matrices are given in the appendix E.1.

$$y_M = M \cdot x(x_{inp}, v_{free}) = g(x, M) \tag{10.2}$$

Additionally, in real life data some measurement error $\varepsilon(y_M)$ is given for each measurement data point. Changing the appearance of Matrix M is done by adding or removing labeling measurements. In the appendix examples are given for the measurement matrix of separate techniques.

10.1.2 Statistics and Design Criteria

Next, the statistics of the free fluxes is to be calculated. For this an estimate of the confidence interval is needed.
It is assumed that the measurements are independent normally distributed and uncorrelated. To calculate the entries in the measurement covariance matrix Σ_y and to finally also perform an ED study reliable values for the standard deviations of the measurements are needed, compare Tillack *et al.* [236]. The measurement error shows a heteroscedastic behavior throughout literature (the error on the measurement ε depends on the measurement value y_M). Thus, an error model is needed. The shape of the function can be arbitrary and depends on the measurement technique and samples/experiment.
To integrate the influence of the number of measurements φ_{meas} on the estimation of the measurements‘ standard deviation, we will modify the above mentioned matrix Σ_y by exchanging the measured variance s^2 to their respective confidence intervals c^2 of the variance. Thus we calculate the confidence interval of the variance as [131]

$$c^2 = \frac{(\varphi_{meas} - 1) \cdot s^2}{\chi^2(1 - \alpha/2, \varphi_{meas} - 1)} \tag{10.3}$$

for a given confidence level α. Hence, there is a direct influence of the number of measurements φ_{meas} (replicates) on the confidence interval. This is done for each measurement group separately, i.e. each single MS measurement group that contains m+0,1,2...n_C mass lanes. Finally, the statistics of the model are calculated by linearization at the design point (Calculation of Jacobian J). The Fisher Information FIM is calculated from the Jacobian and the matrix Σ_y. Inverting this matrix leads to the covariance matrix.

$$FIM = J \cdot \Sigma_y \cdot J^T = Cov^{-1} \tag{10.4}$$

ED studies are based either on the Fisher Information FIM or the covariance matrix Cov. For each of these matrices separate ED criteria are available (see below: Design Criteria).

10.1.3 Identifiability

Identifiability is describing the determination of parameters in a model and it is distinguished between structural and practical identifiability. If a redundant parametrization (over-parametrization) can be found, the model is structurally non-identifiable. This is the case if we can find two flux maps/parameter sets which will lead to the same model output/measurements y [199].
A parameter is practically non-identifiable if its standard deviation becomes very high (or even infinite). Thus changing this parameter values of the model will result in a negligible effect on the output of the model and the measurement values will still be in their standard deviations/confidence intervals. Nevertheless, we possess a clearly defined optimum for the parameter value [199].
Besides these two definitions we define the term statistically non-identifiability. We define that a parameter is statistically non-identifiable if its standard deviation exceeds a certain threshold. Note that there is a slight difference to the definition of practical identifiability and statistical identifiability.

table 10.1: Alphabetical criteria I_{alph} taken from [192, 250] and non-alphabetical criteria I_{non}

I_{alph}	Criterion	Criterion, normalized
Mod. A-criterion	$\rho_{A,mod} = \max trace(FIM)$	$\rho_{A,mod,n} = \rho_{A,mod}/p$
D-criterion	$\rho_D = \max det(FIM)$	$\rho_{D,mod,n} = \rho_{D,mod}^{1/p}$
Mod. E-criterion	$\rho_{E,mod} = \max \frac{\lambda_{min}(FIM)}{\lambda_{max}(FIM)}$	$\rho_{E,mod,n} = \rho_{E,mod}$
E-criterion	$\rho_E = \max \lambda_{min}(FIM)$	$\rho_{E,mod,n} = \rho_E$

I_{non}	Criterion
Costs	I_{costs}
Dimension	p

10.1.4 Design Criteria

Several classical design criteria are described in literature, they are usually denoted by capital letters and are thus summarized as "alphabetical" optimal design criteria
$I_{alph} \in \{\rho_{A,mod}, \rho_D, \rho_E, \dots\}$, the most frequently used are given in table 10.1 [192].
These criteria aim at certain geometrical aspects of the covariance matrix/Fisher information matrix of parameters by reducing the matrix to a single number. To give some examples: The modified A-criterion is reducing the bounding box for the ellipsoid and the D-criterion is reducing the volume of the ellipsoid. To normalize these criteria, we correct them for the dimension of the model p given by the set of free parameters. Besides, the alphabetical criteria we define also non-alphabetical optimality criteria (see table 10.1) like measures for dimensionality of the model, non-linearity, robustness, bias, and costs which are not linked to the covariance matrix. They can also be introduced into the objective vector I. We will focus on the costs and dimensionality here, but other afore mentioned criteria can be applied for ED studies to make their results more reliable.

Several parameters can become non-identifiable during ED studies. This has a direct effect on the information criteria. For calculation of the criteria values, the minimal Eigenvalue λ_{min} needs to be greater than machine precision E:

$$\lambda_{min}(FIM) > Threshold_1 \geq E \tag{10.5}$$

For a numerically stable inversion of covariance matrices a threshold for the condition number is applied:

$$\frac{\lambda_{max}(FIM)}{\lambda_{min}(FIM)} < Threshold_2 \tag{10.6}$$

If these constraints on the Fisher information matrix are not fulfilled, fluxes become numerically or statistically non-identifiable and are thereby excluded from the analysis, resulting in a statistical model with differing dimensionality. The worst determined flux is removed first by convention until both constraints hold for the given model. Care has to be taken, as a non-identifiable parameter can become identifiable in a different ED. Thus, changing adaptively the model's size is needed. This results in information criteria which can no longer be directly compared. As some parameters become non-identifiable, the size (dimensionality) of the covariance matrix is changing and also the statistics of the parameters. Thus, we are using the normalized criteria mentioned in table 10.1 for the multi-objective optimization. Besides, we need to account also for the non-alphabetical design criteria like dimensionality of the system N. This will result in statistical models differing in the non-alphabetical criterion dimensionality, thus, they are not directly comparable to each other.

10.1.5 Multi-Objective Optimization

In a multi-objective optimization problem several objectives $Z = [\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \dots \mathbf{z}_n]^\intercal$ are to be minimized subject to constraints. Each obtained feasible solution φ_1 in the decision space $\mathbf{B}$ is dominated by another feasible solution φ_2, if $\mathbf{z}_j(\varphi_1) < \mathbf{z}_j(\varphi_2)$ for at least one objective function j in solution space (and $\mathbf{z}_j(\varphi_1) \leq \mathbf{z}_j(\varphi_2)$ for all j). A point is called Pareto optimal if there is no other point in solution space found which is better in (at least) one of the objective values. An (usually) infinite set of Pareto optimal points φ represents the solution space. One can obtain a solution close to this Pareto optimal front, so that their solutions are Pareto optimal in respect to all other found solutions (best-known Pareto set) [124]. For the objectives $\mathbf{z}$ all alphabetical criteria and non-alphabetical criteria I can be used.
In literature, also combinations of alphabetical criteria for ED were taken into account [67, 171, 16], but they were usually used as fixed ratios for criteria values by some weight coefficients. The downside of this fixed-ratio approach is that an exploration of design space is difficult as it is *a priori* unknown. Thus, the ratios of the objectives need to be guessed without knowing their ranges [252, 23]. Suboptimal experiments with respect to the one or the other criteria are the result. Instead, we reformulated the problem as an multi-objective optimization problem subject to constraints of upper and lower bound [21]:

$$\min_{\varphi \epsilon B} [z_1, z_2, \dots z_n] \quad \text{such} \quad \text{that} \quad lb_i < z_i < ub_i \, i = 1, 2, \dots n \tag{10.7}$$

The upper and lower bound for the objectives are given by the user, for example an upper bound for the costs‘ objective. The alphabetical and non-alphabetical criteria are used in this study together to circumvent the drawbacks of single objectives or weighted sums of them. Without loss of generality, several objectives can be explored at a time. In order to do this, we tested frequently applied criteria: the A-/E-/D-criterion. Additionally costs and dimensionality of the system were used to make the solutions comparable and rank them.

10.1.6 Measurements

There are basically five different classes of measurement principles recently applied in literature. For each of them several devices and protocols are available, differing in their accuracy/sensitivity (error model), potential measurements and throughput. In this section a brief overview will be given about their application in ^{13}C-MFA.
In figure 10.2 the mainly used measurement devices are shown. Mass spectrometer (MS) and tandem mass spectrometer (MS/MS) are measuring sums of isotopomers sharing equal masses (“cumulative enrichment”), whereas tandem mass spectrometry enhances fragmentation and can potentially extract more information about the isotopomers. H-NMR is measuring the fractional enrichment of a certain carbon (“local enrichment information”), in the figure 10.2 only the local enrichment of C2 carbon is measured. GC-C-IRMS is measuring the overall fraction of labeled to unlabeled specie in all carbon atoms of a metabolite. ^{13}C-NMR measures local labeling patterns around certain labeled carbon atoms. In figure 10.2 only the C2 carbon is measurable by ^{13}C-NMR, thus, four different distinct labeling pattern can be measured and are discriminable.
These measurement techniques provide different measurement principles and also information about the labeling pattern of the metabolite. Nevertheless, the applied methods differ in their applicability to certain classes of metabolites by the used separation techniques (LC/GC) and analyte concentration [128].

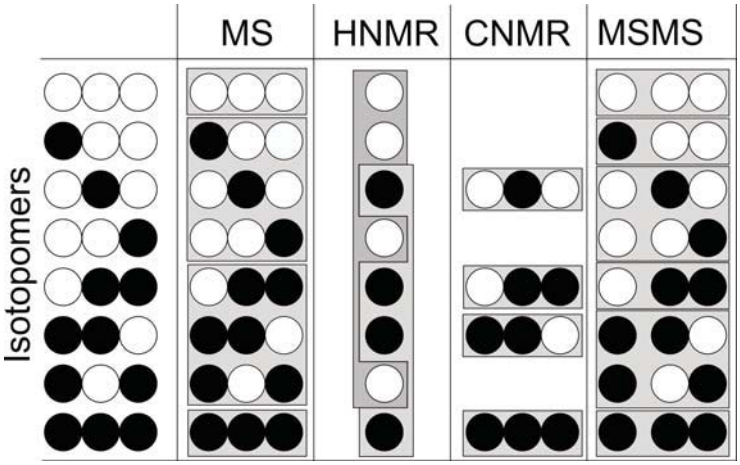


figure 10.2: Measurement principles for three carbon metabolite (C1-C2-C3): HNMR and C-NMR measurement for C2 shown. MS/MS measurement for C2-C3 fragment and C1-C2-C3 mother ion shown (multiple reaction monitoring). Connected grey panes represent a single measurement. They are comprised of sums of isotopomers measured by the respective measurement principle.

To give an overview on the used analytical devices in literature their analyte spectra are shown in table 10.2. The altered information content and analyte spectrum in the measurement does also effect the statistical identifiability of the fluxes. This was already previously published by Rühl and coworkers. They compared LC-MS with LC-MS/MS and showed that measurement of metabolites with fragmentation results in better statistics [203].

10.1.7 Multi-Platform Application

As there are several drawbacks of the separate techniques, like sensitivity and restriction to a certain class of analytes, their combination will result in higher coverage and informative value. GC-MS and NMR were used in combination by McKinlay *et al.* [152]. Kleijn used GC-MS, LC-MS and NMR for ^{13}C-MFA in *Saccharomyces cerevisiae* and NMR and LC-MS in *Penicillium chrysogenum*[123]. Toya *et al.* used CE-TOFMS for measurement of metabolite in TCA, glycolysis and PPP and GC-MS for amino acids and used both in metabolic flux analysis [238].
Besides the analyte spectrum, the mentioned techniques differ severely in their sensitivity for flux elucidation. Until now, it is an ongoing debate which analytical platform provides the maximum information and how one can gain from combination of multiple devices. We have seen that the coverage is different and that the measurement techniques differ in their delivered information about the isotopomers.

10.1.8 Current Status in Literature

As we have seen, there is a broad range of applied techniques in literature, nevertheless GC-MS, LC-MS(MS) and CNMR are widespread for ^{13}C-MFA. To describe the measurement error of the separate techniques an error model for calculation of the confidence intervals is needed. From several published papers the measurement errors on labeling enrichment measurements were extracted. We assumed linear behavior and the results are shown in

table 10.2: Devices used in literature

	Amino acids	Organic acids	Sugars	Comments
H-NMR	[149]	Too low concentration, except secreted products	Too low concentration	
2D/C-NMR	[231, 230, 68, 50, 214, 82]	Too low concentration, except secreted products	Too low concentration	Nucleosides measurable [82]
GC-MS	[39, 49, 270, 20, 229, 7, 79, 77, 68]	[78]	[20]	Derivatization and correction needed [255], mainly used for amino acids
LC-MS	same as LC-MS/MS	[181, 248, 123]	[248, 123]	
LC-MS/MS	[248, 107, 203]	[203]	[203]	
GC-C-IRMS	[278, 97]			
Maldi-TOF-MS	[273]		[273]	

figure 10.3. It is obvious that most devices show similar standard deviations. For GC-MS the standard deviation were considerably higher in literature. For HNMR only a few data-points in low range of labeling were available.

10.1.9 Substrate

First a short overview about the used substrates in literature is given. Walther *et al.* found mixtures of 1,2- labeled glucose and mixtures of 3- and 3,4- labeled glucose to increase statistical identifiability, always mixed with fully labeled glutamate for lung cell carcinoma [256]. Crown *et al.* were using a small scale network with two degrees of freedom. They found 3,4- and 2,3,4,5,6-labeled glucose to be favorable to elucidate reaction rates in the oxidative phosphate pathway and pyruvate carboxylase flux, respectively [48]. Metallo *et al.* also searched for optimal tracers for lung carcinoma cell line. They identified best tracers for each reaction: 1,2-labeled glucose was optimal for most fluxes in pentose phosphate pathway and glycolysis. For TCA-cycle uniform labeled glutamine gave optimal results. However, they also presented that there is a high diversity in the found standard deviations of the fluxes depending on the used substrate [260]. Schellenberger *et al.* used a *Escherichia coli* network to perform experimental design and found 1- or 6- labeled superior to a mixture 20 % uniform 80 % unlabeled glucose mixtures [213]. They also stated that multiple labeling positions are superior to others. Besides they found a high redundancy of information in the measurement data and claimed that frequently used single/fully labeled species (1-Glc, U-Glc, ...) show potentially a medium number of identifiable fluxes. Besides, exotic labeling resulted in a higher identifiability. This was the case for 5,6-, 1,2,5-, 1,2-, 1,2,3- and 2,3-labeled glucose.
In this study ten commercially available substrates are incorporated (Glc#000000, Glc#100000, Glc#010000, Glc#001000, Glc#000010, Glc#000001, Glc#111111, Glc#110000,

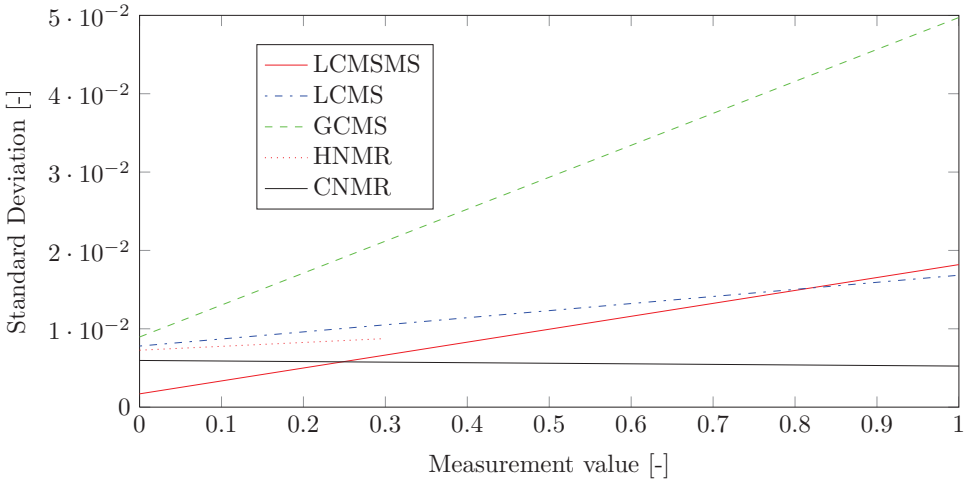


figure 10.3: Error model derived from literature data [121, 107, 247, 211, 4, 232, 149]. Error models were derived by least square fit of $s = m \cdot y_M + b$ to the data published.

and Glc#100001). The purity of the substrate has to be taken into account as it will alter the labeling pattern in the used substrate considerably.

10.2 Materials and Methods

^{13}C-MFA

The high-performance simulator 13CFLUX2 was extended for multi-objective optimization using jMetal as multi-objective library [260, 63].

Multi-Objective Optimization

In simulation studies the multi-objective optimization toolbox jMetal version 1.0 was used [63]. The parameters for the optimization were the the fractions of each substrate used in the substrate mixture and the number of measurements per measurement group (i.e. number of measurements for a complete MS or NMR measurement vector). The output of the program is the found Pareto optimal solutions as a list of the parameter-sets and the corresponding values for each objective. The convergence of the solution (Pareto front) was measured by its hyper volume of best-known Pareto optimal set in the objective space [64]. Maximum iteration number was 10^6 model evaluations, the swarm and archive size was chosen to be 1,000 using a polynomial mutation. The parameters were adjusted as stated in Nebro *et al.* [64]. Each optimization per model was executed in duplicate to circumvent false predictions of results.
The optimizer SMPSO, SPEA2, and NSGAII available in jMetal were compared with respect to run time on a data set. SMPSO was providing, measured on the Pareto volume, fastest convergence, whereas SPEA2 failed by a major misconception of the algorithm. SPEA2 showed strong decrease in hyper-volume after reaching an optimum, thus it decreased the hyper-volume of the solution space after an optimal solution was found; for more information see literature [172, 286, 285]. After testing these algorithms, SMPSO was chosen as it provided speed up of about 300 % compared to SPEA2. Function evaluation was measured to

reach 98 % of the maximal Pareto frontier hyper volume (NSGA2: $1.7 \pm 0.9 \cdot 10^5$ and SMPSO: $4.4 \pm 2.8 \cdot 10^4$ iterations needed). After 10^6 iterations the volume was constant.
All optimality criteria were evaluated based on Fisher information matrix, inversion (if necessary) was done using SVD decomposition. The minimal number of free parameters (free fluxes) in a model was constrained to 21, models with constraint violation were not used to build a Pareto set. The standard deviation of the measurements, given by the error model, is converted into confidence intervals of standard deviations as a function of the number of replicates measured. Jacobian was calculated and the above mentioned constraints equation 10.5-10.6 were applied.

Visualization

Visualization was conducted using CIRCOS tool [126] and OMIX [60].

Model

The network was built for an eukaryotic organism (*P. chrysogenum*). Biologically feasible fluxes were chosen based on previous obtained flux maps in this work. The model describes glycolysis, pentose-phosphate pathway, TCA cycle and formation of all measurable amino acids. The model comprises 34 free fluxes: 12 net and 22 exchange fluxes. It is assumed that the free flux values and the external flux measurements are known (possess no standard deviation). In order to guarantee comparability within the given study, we set non-identifiable fluxes constant. Each used model was previously checked for non-identifiable fluxes. All constrained fluxes can be found in the appendix E.3. This leads to models with differing dimensionality. LC-MS/MS had the highest dimensions with 26 free fluxes. CNMR the lowest with 22 free fluxes. We identified the common set of non-identifiable fluxes for all devices and set them constant in all models ending up with 21 free fluxes. The free net fluxes which needed to be fixed were anaplerosis and product formation rate in the model. All other net fluxes were identifiable. Besides, high exchange fluxes also became statistical non-identifiable and needed to be fixed, e.g. *gly2* exchange flux. The full list of all constrained fluxes can be found in the appendix.
GC-C-IRMS and HNMR were excluded from the study as they could not meet the requirements. For both at least ten additional fluxes compared to GC-, LC-MS (/MS) and CNMR had to be fixed, which made the results incomparable. Besides, the low amount of data available from literature makes the construction of error models difficult for these devices. In the appendix E.3 the constrained fluxes are given. These devices (GC-C-IRMS and H-NMR) showed the worst elucidation of the fluxes as they provide the least information about the given flux distribution. As many fluxes needed to be fixed, the dimensionality of the model was also strongly reduced; 23 and 29 out of 34 free fluxes had to be constrained to evaluate resulting statistics for HNMR and GC-C-IRMS measurements, respectively. We searched literature for used measurement devices in metabolic flux analysis and collected used measurements (see table in appendix).
Non-redundant measurements were filtered out from literature and displayed in the appendix in the context of the metabolic network. In the appendix, non-redundant measurement specifications are shown and their respective number of single measurements. From the mentioned literature 38 measurement for metabolites were incorporated in the metabolic network for LC-MS/MS (314) and LC-MS (226, in brackets the number of single measurement values

are given). CNMR provides measurements for 17 (156) and HNMR for 10 (64) metabolites. GC-C-IRMS was using 17 (32) metabolites and GC-MS 19 (213).

10.3 Results

10.3.1 Cost Function

The cost function is of central importance in this study. We describe in a cost function the major steps of ^{13}C-MFA: Conducting the labeling experiments, sampling, analytics, and modeling. The costs need to be calculated for each *in silico* planned experiment. By introducing them, expensive experiments can be identified and are replaced by designs which are more economic but provide the same (or better) alphabetical criteria. We determined working times to perform most steps of ^{13}C-MFA and found that variable costs factors are, besides purchasing the analytical device, usually measurement time on the device, peak integration/post processing like manual correction of peak areas, additional sample measurements, and model building. We quantified these steps in terms of cost and time by a in-house user study and wrote protocols for each measurement device used as input for calculation of the cost function in figure 10.4. However, as more metabolite measurements and replicates will result in potential use of multiple protocols for measurements the working time and analytical effort will increase for sampling, analysis, peak integration, and model setup. As the number of measurements is a free variable for the algorithm to choose, we set most of the results free for optimization, thereby assessing the question: Which measurements are important?
We will now describe the costs function. The used substrates for the ^{13}C-MFA will determine the statistical identifiability of the system. These substrates range from neglectable cost per gram substrate (unlabeled glucose) to several thousand euros per gram. Some mixture may be beneficial over others but it can be higher in costs $\mathbf{C}_{sub}$. The costs for an experiment consists of labor time $t_{work,exp}$ and costs for the experimental setup $\mathbf{C}_{exp}$ (i.e. labeled substrate, media, number of samples taken...).

$$\mathbf{C}_{exp} = \mathbf{C}_{sub} + \mathbf{C}_{exp} + t_{work} \cdot \mathbf{C}_{wage} \tag{10.8}$$

The resulting labeling patterns of the metabolites are then analyzed in the respective analytical device, evaluated and transcribed into the model. These analytical devices can cost up to 1000 k€, the respective costs are allocated to 5 years in this study. By using run-time per sample we can calculate the costs of the device per sample $\mathbf{C}_{meas}$. For the analytics $\mathbf{C}_{ana}$, we obtain:

$$\mathbf{C}_{ana} = n_{sample}(max(\varphi_{meas})) \cdot \mathbf{C}_{meas} + t_{work,ana}(\varphi_{meas}) \cdot \mathbf{C}_{wage} \tag{10.9}$$

Finally we can calculate the costs objective for one experiment:

$$I_{costs} = \mathbf{C}_{exp}(\varphi_{subst}) + \mathbf{C}_{ana}(\varphi_{meas}) \tag{10.10}$$

Notice: the number φ_{meas} of measured samples can be different for every measurement, as not every measurements' respective peaks need to be integrated. The number of measured samples n_{sample} is determined by the maximum number of samples measured for any of the metabolites. Thus, we can measure all samples, but just evaluate a fraction of their labeling patterns to safe costs.

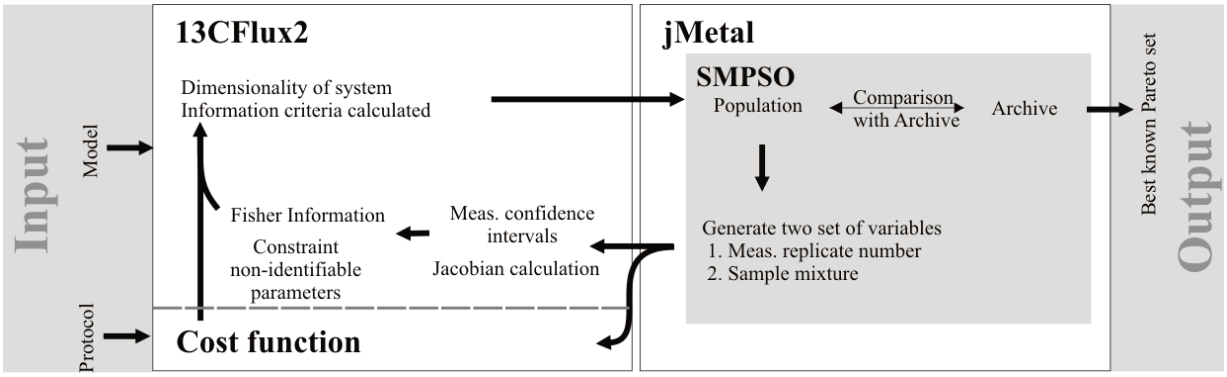


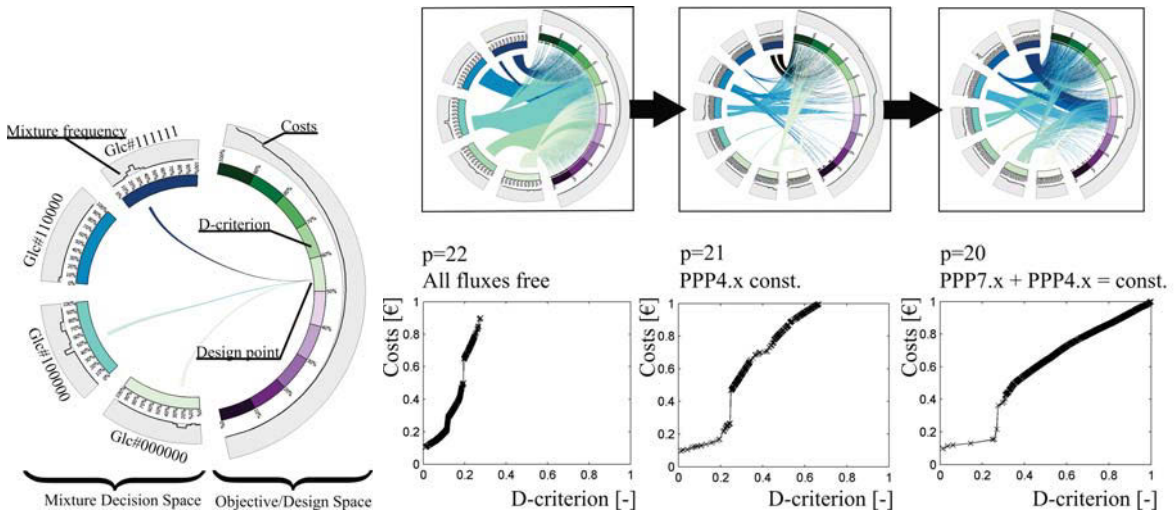
figure 10.4: Program flow diagram for multi-objective experimental design study using jMetal toolbox for optimization. The algorithm is executed until the maximal model evaluation is exceeded.

10.3.2 Implementation

After initializing 13CFLUX2 the optimizer SMPSO is started within jMetal (see figure 10.4), generating a population of variables in design space. Each individual in the population is accounting for measurement replicate number vector and sample mixture. The lower boundary for which a measurement is counted is at least 2 replicates φ_{meas}, as this is the minimal number of measurements for which a finite confidence interval/standard deviation can be calculated. Measurements with lower number of replicates are excluded from the analysis. In 13CFLUX2 the model is simulated using the sample mixture and the Jacobian J is built. The Fisher information matrix FIM is calculated by equation 10.4. The information criteria are evaluated based on the Fisher information matrix, if the constraints (equation 10.5-10.6) are met. In case of non-identifiable parameters these need to be fixed to a constant value specifically for the set of given variables. From the given protocol for the measurement device and the mixture and number of measurements, the costs for the experiments are calculated. Finally, the desired information criteria are given to jMetal as objective values.
From this population of solutions Pareto optimal solutions are selected by the specific algorithm to build the archive (best known Pareto set) and for the next iteration a new set of points in design space is generated. The archive is automatically updated by jMetal based on crowding distance to achieve optimal coverage of the solution space [52]. The archive/best known Pareto set is returned as output at the end.

10.3.3 Visualization

The visualization of the high dimensional substrate mixtures was of central importance in this study. The scientific user needs to be aware of the interplay of costs, measurements, mixture composition and criterion value in order to choose his/her optimal point in decision space. We used clustering algorithms based on shortest Euclidean distance of mixtures to generate dendrograms and projections of the high dimensional data to visualize them (data can also be seen in the appendix). Thereby problems arise. The overall complexity of the data is lost by using clustering and projections are abstract. Thus we established a new tool to visualize the high dimensional mixture plots. In figure 10.5a an example of these mixture plots created by CIRCOS tool is given [126]. On the left hand side the used substrates are shown and on the right hand side the objective values (costs and alphabetical information



(a) Visualization (b) Effect of dimension on experimental design

figure 10.5: (a) Visualization for high dimensional mixture plots with regard to objective value. (b) Effect on the design and objective space by setting fluxes constant and thereby changing the dimensionality of the covariance matrix. For comparison, the D-criterion and costs were scaled to the interval $[0, 1]$ with respect to highest found values. Normalized D-criterion is plotted.

criterion), scaled between 0 and 100 %, are shown. One design point in objective space can possess a complex substrate mixture, thus arrows from the separate substrate indicate the amount they contribute to the overall mixture in this design point. To allow quantitative remarks on the usage of a substrate, a histogram plot shows for each substrate the frequency of the fraction used in separate used mixtures.
As we need to constrain fluxes to their value if they become non-identifiable, we are dealing with systems of differing dimensionality. An example is shown in figure 10.5b. Here, fluxes in a model are consecutively set constant in order to evaluate the statistics of the model. We can see that the mixtures and number of used substrates is changing in the circular plots.
While setting fluxes constant, the model becomes more and more statistical identifiable as the normalized D-criterion is increasing in its value. For simplicity we are discussing in this work only the highest dimensional case (p=22), as the designs for other dimensionality are becoming either harder to interpret or only a low number of data-points was found for them. The dimensionality of the model is incorporated into the objectives: First the used criteria incorporate a correction for the dimensionality. Second, the dimension is a separate objective in the multi-objective optimization. Thus, solutions with a all variables free possess a obective value for the dimension of $I_p = 0$ (p=22), if one flux is set constant one obtains $I_p = 1$ (p=21) and so on. By this, models with higher dimensionality (more free variables) dominate models with lower dimensionality (less free variables).

10.3.4 Results for Economic D-Criterion

Three objectives were taken into account for multi-objective ED: the costs, D-criteria and the dimensions of the system (**c**osts-**D**-criterion-**d**imensionality "cDd" design). The D-criterion is most frequently applied in literature for planning of experiments. Only five out of poten-

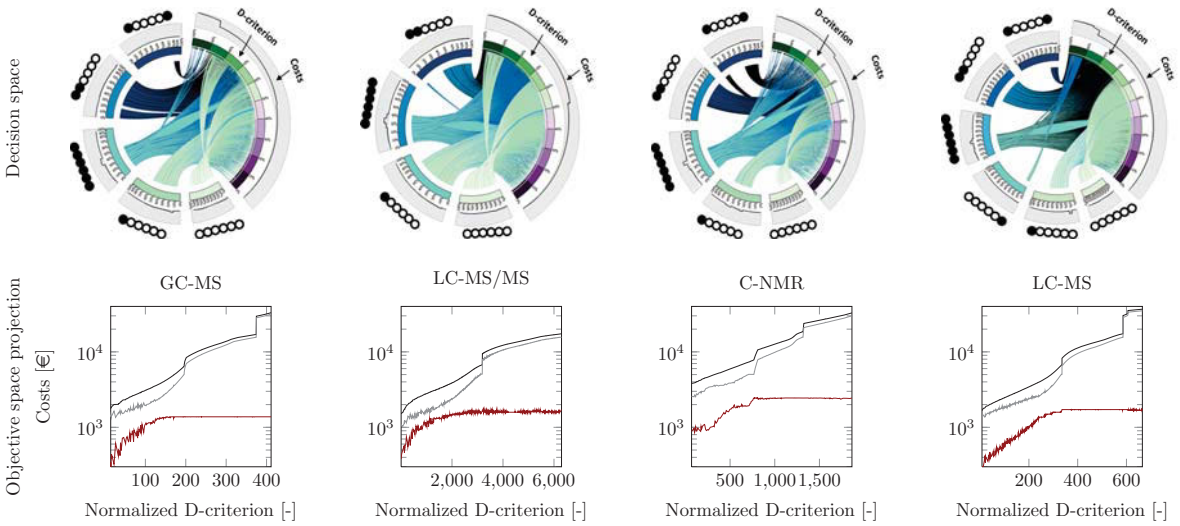


figure 10.6: Results of a multi-objective ED run using CNMR, GC-MS, LC-MS and LC-MS/MS measurement devices. Substrates below 1 % in mixture were removed for visualization purpose. Costs resolved for experiment and analytics. Overall costs (——), experimental costs (——, including costs for substrate) and analytical costs (——).

tially 10 used substrates are predominantly found (*Glc#000000*, *Glc#100000*, *Glc#111111*, *Glc#100001*, and *Glc#110000*). Unlabeled glucose is primarily present in mixtures with low information criteria (see figure 10.6). Interestingly, expensive substrates (*Glc#100001*, *Glc#110000*) were only present in highly informative mixtures. These substrates can increase the statistical measure by about 50 %. One can see that LC-MS/MS uses, besides *Glc#110000*, only cheap substrates compared to the other devices. In the other measurement devices *Glc#100001* is found in the most expensive/most informative substrate mixture and increases their costs for ED strongly as it is one of the most expensive substrates (>1000 €/g).

The minimum and maximum criterion values are shown in figure 10.7, where LC-MS/MS shows least costs. The D-criterion is higher compared to the other devices leading to highest statistical identifiability in the present study. CNMR provides higher minimal costs than the other devices, as the analytic device is more expensive and measurement time is considerably longer. The other devices are similar in costs, as they require also similar costs for the analytics. The main part of the costs is composed of the substrate costs. CNMR shows higher costs for analytics because it has long measurement times and expensive analytical devices (both up to ten times higher compared to the other devices). Besides, LC-MS/MS shows highest effort and costs for manual peak evaluation.

The previously shown results indicate the power of the method. Nevertheless, it might happen that the found designs are spread over almost all possible mixtures. Thus, we clustered the results by their substrate mixture, cost and D-criteria. For LC-MS/MS three large distinct clusters were found, the first contains the standard mixtures of 1- and fully labeled glucose and correspond to the cheapest EDs (<5000 € overall costs; data can be seen in the appendix E.3). The second cluster contains [1,2] labeled glucose and fully labeled glucose. In the third cluster [1,2] and fully labeled glucose can be found. Both clusters are more expensive and yet more informative. The other devices show similar results (data in appendix E.3 shown).

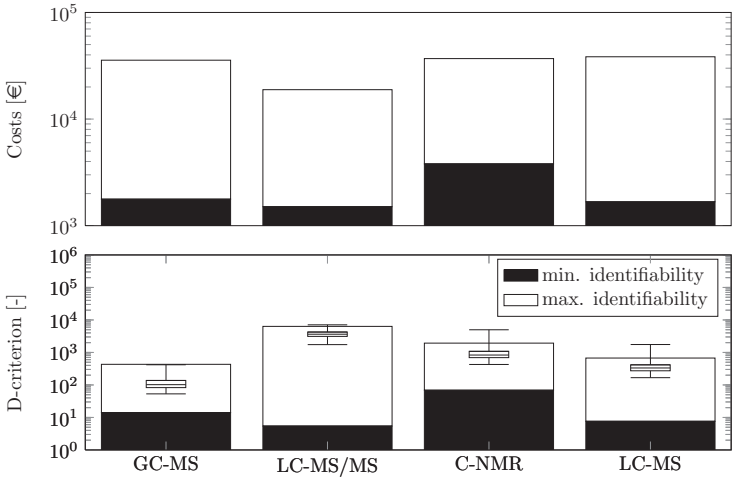


figure 10.7: Comparison of costs and D-criterion value for the fisher information matrix of the cDd design. Whiskers are showing min/max value for D-criteria found for 10^4 sampled flux distributions. Boxes show 25th and 75th percentiles. Median is indicated by a dot.

Robustness Study

We have performed the presented multi-objective ED study for a given flux distribution which we estimated based on published flux distributions. This flux distribution is one point in flux/parameter space and thus a local solution. As our fluxes may vary within their found standard deviation, severe deviations from the experimental design values obtained are possible in this region of parameters. Thus, we sampled our parameters in the bounding box given by the confidence ellipsoids from the parameters covariance matrix. This was performed for the best found experimental design for the D-criterion for each measurement device separately. Finally, we evaluated the D-criterion based on the best known optimal design point to get an impression of the variation for the optimal D-designs if the fluxes are changed. The optimal design point for the D-criterion was using an input substrate mixture of 100 % *Glc#110000* for LC-MS/MS. The results for 10^4 samples are shown in figure 10.7 for optimal cDd design points.
We can see that the median value of the found D-criterion for the newly sampled fluxes is below the solution for the original flux distribution. Thus, for our flux distribution the design obtained in cDd study is optimal, but not mandatory for the other flux distributions. LC-MS/MS shows narrower distributed D-criterion values compared to other devices. This is consistent with the fact that LC-MS/MS results provide the lowest standard deviations on the fluxes and, thus, similar flux values are sampled compared to the original flux distribution.

Influence of Specific Measurements and Replicate Number

If we analyze the cluster of substrate mixtures for LC-MS/MS with regard to the used measurements, we can see that from low to high costs ED there is a growing number of measurements included in the simulated experiments (data shown in appendix E.3, figure E.6-E.5).
Using more measurements is usually not expensive because the substrate mixture contributes more significantly to the overall costs.

In the cluster with the cheapest design points for LC-MS/MS (data shown in appendix E.3, figure E.6) the measurements incorporated in the model vary with respect to replicate number. All measurements are varied and show a broad distribution in their replicate number. A few measurements are less often used like 2,3-phosphoglycerate.
The more expensive clusters differ not only in the cost and usage of substrates but also in their number of measurements. Almost all measurements are used for statistical evaluation of the model. Only measurements for malate, glyceraldehyde-3-phosphate, 2,3-phosphoglycerate, tryptophane, tyrosine, phosphoenolpyruvate, and dihydroxyacetonephosphate show variations. Most likely, because these measurements provide minimal extra information for statistical identifiability.

10.3.5 Results for Economic A-/E-/D-Criteria

We can perform the same study as before (cDd design) also with several alphabetical criteria: A-, E-, D- criteria, dimension and cost called from now on called (**c**osts-**A/E/D**-criterion-**d**imensionality) "cAEDd" criteria. The results of the cDd and cAEDd ED can be seen in figure 10.8.
We are dealing now with a multi-objective optimization in five dimensions. Thus, projections of the results are shown. The previously generated results of the cDd designs are also presented within the plot (black marks), the other criteria values (A- and E-citeria) were calculated for these design points as well.
The found cDd ED solutions are just a distinguished trail through the overall objective space which is represented by the cAEDd objective design space. The solutions in five dimensions for cAEDd showed same or lower D-criterion compared to the cDd results, they even tend to less optimal criteria values, as they are better in terms of the other objectives. The costs per experiments in case of cAEDd ED are higher compared to the cDd ED and the used substrates show a higher variation (see appendix).
In figure 10.8 we can also see that there is a large objective space for A, E and D-criteria based on the point cloud of cAEDd designs. The objectives of the other devices (CNMR, LC-MS, GC-MS) show strong variations in E and A-criteria of the cDd design (see appendix, results for other devices). Thus, there are usually more informative experiments possible then predicted solely by the D-criteria as the other criteria can change substantially in their value.
Several design points for LC-MS/MS show a remarkably higher A-criteria for increased costs compared to the cDd results. This holds true for, at least, CNMR and GC-MS. The respective designs are even more expensive than the found three objective results, whereas the D-criteria value was not becoming better. Again, we can see that experiments with higher statistical identifiability are possible.

10.3.6 Results for Multi-Platform Applications

From the first study, we see that LC-MS/MS outperforms the other devices. Next, we want to test if combining devices in one study will result in additional value for the statistics. We combined GC-MS with LC-MS as they are most frequently used (see figure 10.9). Additionally, HNMR and CNMR can be both performed on one device but with slightly different experimental setups. It was already discussed that HNMR is providing absolute values for the labeling at certain positions, whereas CNMR gives information about labeling patterns

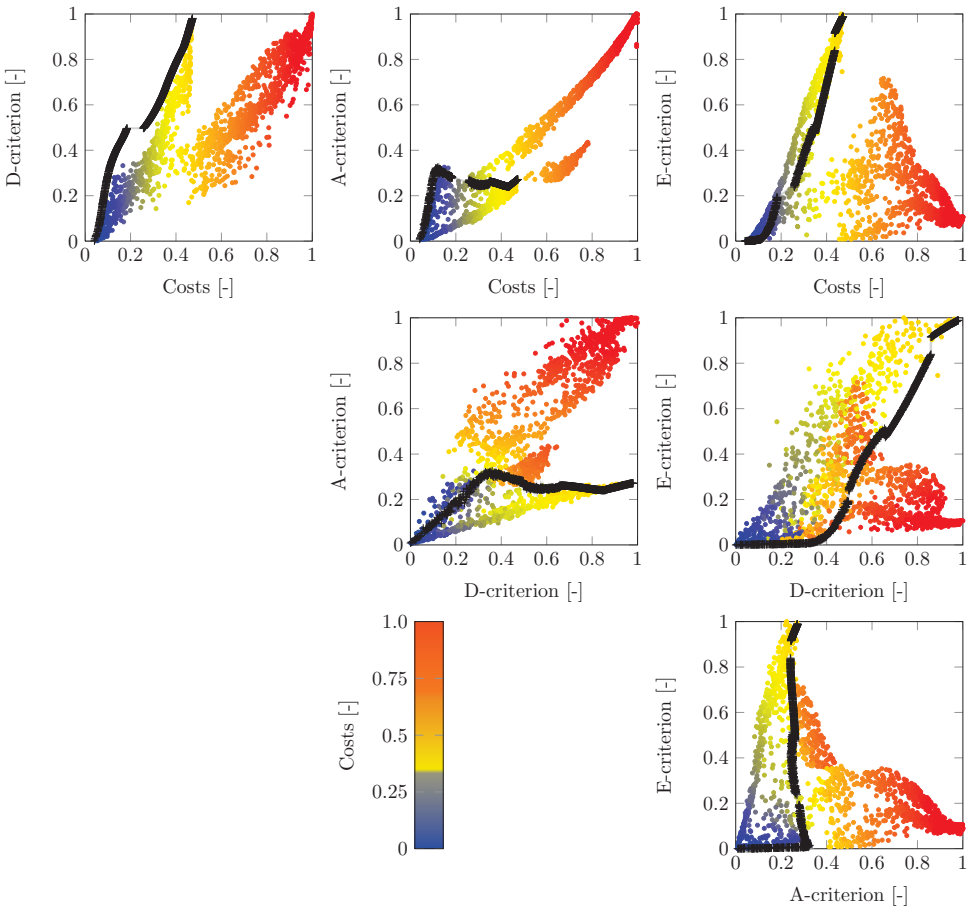


figure 10.8: Multi-objective optimization results for LC-MS/MS. Color codes for costs of design point. The results of three objectives (D-criterion, dimensions and costs) are shown (plus signs). The points in the separate plots are projections from five dimensions into two dimensional space.

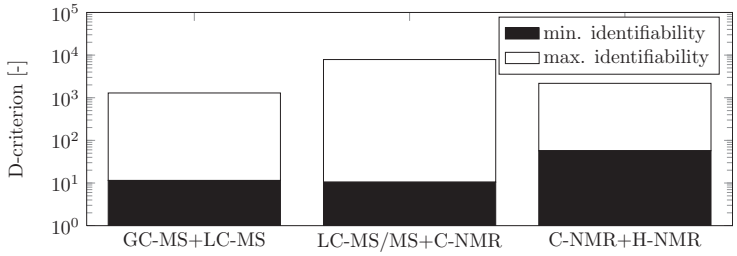


figure 10.9: Comparison of costs and D-criterion value for the fisher information matrix of the cDd design for combinations of measurement devices.

around a labeled carbon atom. Finally, we added CNMR to LC-MS/MS together in one model. For each of those combinations we repeated above study. We will now compare the results of the single devices to multiple devices, number given in brackets are D-criteria values for the devices in a cDd design. GC-MS (417) and LC-MS (665) increase significantly to a D-criterion of 1282. CNMR (1867) and HNMR(-) increased to about 2119. LC-MS/MS (6333) and CNMR (1867) increased together to 7858. All in all, GC-MS and LC-MS showed the highest increase of statistical identifiability.

10.4 Conclusions

We have performed multi-objective optimizations for ^{13}C-MFA ED to increase the methods‘ statistical significance. To this end we introduced the new concept of non-alphabetical criteria into multi-objective ED. As there are several potential measurement setups possible (MS, NMR-based techniques), we extracted linear error models from literature data for them and build specific cost functions. For each of the published devices the work-flow was performed, aiming at answering the following questions:

Which Fluxes are Identifiable?

We could show that the measurement techniques differ in the number of identifiable fluxes. For HNMR and GC-C-IRMS we found a low number of identifiable fluxes and excluded them for further analysis.
LC-MS/MS showed the highest amount of statistical identifiable fluxes. High exchange fluxes were usually statistically weak identified. Nevertheless, at least 21 free fluxes (9 net and 12 exchange fluxes) out of 34 free fluxes were identifiable with CNMR, LC-MS/MS, GC-MS and LC-MS. As the same flux distribution and the same dimensionality for all models was used, a direct comparison of the results was possible.
Basically, all measurement techniques were capable of statistically resolving these fluxes. However, optimal statistical elucidation can be achieved with LC-MS/MS, which could identify these fluxes with highest statistical identifiability. Optimal ED points differ in number of used measurements and costs, which will be discussed in the next section.

How much does it Cost?

We extended the concept of alphabetical design criteria for ED by non-alphabetical design criteria (like costs and dimensionality of the model), aiming at other properties of the system

than the covariance matrix to increase comparability and reliability of ED studies. As we relied heavily on multi-objective optimization, it was possible to search for economic EDs by introducing costs for devices, substrates and measurements as non-alphabetical design criteria.
As expected, one can deduce from the results that costs for experiments are mainly influenced by substrates of choice. They possess a severe impact on the statistical identifiability of the system. For normal substrate mixtures like 1-, uniform- and naturally labeled glucose, only moderate statistical identifiability is found with a moderate number of needed measurements. For higher statistical identifiability, also more expensive substrate mixtures can be used. From the result one can see that there are the basic substrate mixtures used on the one hand (1-/uniform/naturally labeled glucose) where the number of replicates show strong variation. Thus, not all measurements are important for those mixtures as the measurement do not utilize the maximum of potential information by the poor substrate mixture. On the other hand there are highly informative mixtures utilizing almost all measurement/ information content. Thus, saving measurements for best results was only possible to a small extent, as a high demand for measurements strengthens the fact that most measurements are indispensable for high statistical identifiability (using costly substrates).
Nevertheless, the choice of substrate seems to be the most crucial point for elucidation of fluxes. Choosing the right device, like LC-MS/MS and measurement may lead to EDs with 50 % less costs than other devices (CNMR, LC-MS, GC-MS). A more economic design of experiments using LC-MS/MS results also in 300 % higher criteria values compared to the other techniques in this study. We have also seen that measurements are tremendously cheaper than substrates. A potential option to enhance the statistical significance of the measurements is introducing new measurements or using differing measurement principles in one ED.
Thus, increasing the number of measurements can result in higher statistical identifiability as shown for the case of LC-MS/MS. For cheap substrate the measurements were strongly varied. In the end, only a better (and usually more costly) substrate will increase statistical identifiability and will positively effect the outcome of the study.

Is it Worth the Effort?

We could show by comparing different measurement devices that CNMR and LC-MS/MS show best performance in terms of statistical identifiability. Indeed, we can see that omitting measurements is not crucial for frequently used substrates (1- and fully labeled glucose), they possess some sort of robustness or redundancy for the given flux distribution. In this case we can easily save measurements without a dramatic effect on statistical identifiability. Using high priced substrates makes sense if a suitable measurement spectrum is available. the found designs show a relative broad distribution in substrate mixture space. Thus, changing the mixture slightly, the statistical identifiability was found to be stable/robust in this study.
Previous studies have addressed potential problems by using the D-criterion. One example are needle-shaped confidence ellipsoids with large confidence intervals for the parameters who are usually undesired. This can be prevented by observing simultaneously the A-criterion. By introducing the multi-objective optimization with A, E and D-criteria we were able to efficiently explore the design space and compare them to the results for using only the D-criterion. Thus, false predictions like described in literature are prevented [16, 88].
In the end it was evident that D-optimal designs (cDd designs) tend to show strong fluc-

tuations in other criteria. These results strengthen the need to explore the design space in depth. This was also accomplished by multi-objective optimization for ED by comparing the results to the economic A, E and D-optimal designs (cAEDd designs). Besides, for experimentalists, the A-criteria is most important as the covariances described by the D-criteria are not of direct interest. By applying this technique we found A-optimal designs which were more costly than the pure D-criteria but served better in statistical identifiability. They were characterized by equal D-criterion value and higher A-criterion. All in all, the A-criterion is more robust than the D-criterion, resulting in more reliable experimental designs.

10.4.1 Comparison of Techniques

We can, finally, also compare the separate techniques directly. By measuring of the D-criteria we can say that LC-MS/MS performs best for the given flux distribution, measurements and substrates. One could argue that the standard deviations of the measurement devices strongly depend on the experimental procedure, used protocols and strain. Nevertheless, the error models built from literature show similar behavior for all measurement devices. Only GC-MS shows higher errors, which was consistent throughout literature [4, 232]. Although lower standard deviations were reported to be reliable for a limited number of measurements by Antoniewicz *et al.* we did not use them in this study, because they do not reflect current status in literature [8].
For these reasons, the presented approach for optimal ED prevents false predictions and gives researchers an efficient tool to plan experiments based on *a priori* knowledge. The principles applied are universal and can be used throughout many fields of study. Here, for the first time non-alphabetical criteria were introduced in multi-objective ED. The method enables the user to explore design space and thereby evaluate already found design points to choose a labeled substrate. Additionally, economic demands can be faced if the experimental procedure is expensive. We could also compare measurement techniques *in silico* by their statistical identifiability and could show that more economic and most informative experiments were possible using LC-MS/MS.

Part IV

Summary and Outlook

Chapter 11

Summary and Outlook

The aim of this thesis was to establish state-of-the-art fluxomics technologies for *P. chrysogenum* in the industrial environment. To derive quantitative flux maps, the established work-flow of ^{13}C-MFA was extended and new tools developed. In the end it was the aim to elucidate the fluxome of the *P. chrysogenum* strain BCB1 in an industrial environment. Initially, assumptions for ^{13}C-MFA were stated and the changes in process conditions to transfer ^{13}C-MFA to the industrial environment were discussed. Close-to-industrial process conditions for fed-batch cultivations were established using the sensor-bioreactor concept [61, 62]. The main results are briefly summarized and an outlook is given.

11.1 Summary

In this thesis, several new tools were developed to cope with the experimental constraints and limitations of stationary ^{13}C-MFA.

1. NoReTo: A Nonlinear Regression Toolbox and work-flow was implemented to calculate nonlinear, linearized and Monte Carlo bootstrap statistics in MATLAB and applied to estimate steady state labeling data and confidence intervals for ^{13}C-MFA models.

2. GloSA toolbox: Global Sensitivity Analysis can be performed by using random-sampling high dimensional model representation (RS-HDMR). A work-flow was developed and a toolbox comprised of most available techniques was implemented in MATLAB and applied to ^{13}C-MFA models.

3. For the correction of natural abundance in tandem MS measurements an algorithm and concise notation was introduced.

4. Finally, a procedure for multi-objective experimental design was developed and introduced in 13CFLUX2.

These tools were used in the work-flow of ^{13}C-MFA.

^{13}C-MFA for *P. chrysogenum* BCB1

For analysis, three chemostat experiments (with growth rates $0.03\,\mathrm{h}^{-1}$,$0.05\,\mathrm{h}^{-1}$) and, additionally, two fed-batch cultivations for close-to-industrial process conditions were provided by Sandoz GmbH. For these datasets, stationary ^{13}C-MFA was established in this work.

The ^{13}C-MFA technology was introduced and the state-of-the-art work-flow was established, finally, it was applied to the experimental datasets for *P. chrysogenum* BCB1.

Pre-processing A pipeline for pre-processing of the provided measurements was built. First, time resolved extracellular concentration and process data were used to derive extracellular reaction rates. Additionally, time resolved mass spectrometric labeling measurements were available.

- Rates were estimated based on a unified complex kinetic model (24 parameters, implemented in the software Dymola) taking into account growth and the complete measured side-product formation of *P. chrysogenum* BCB1.
- Isotopic steady-state labeling measurements were estimated by extrapolation of transient data. Additionally, uncertainties of the labeling enrichment were propagated to extrapolated values by nonlinear statistical methods applying NoReTo.
- The extrapolated mass spectrometric measurements were finally corrected for natural isotopic labeling and statistical reconciliation was performed by the introduced algorithm.

Network Modeling For conduction of ^{13}C-MFA, a large scale metabolic network model was built (485 reactions). From this a focused metabolic network (140 reactions) was derived and assumptions were thoroughly stated. Strain specific effluxes for the models were calculated from hydrolyzed biomass components. Measurement specification and input substrates were integrated into the ^{13}C-MFA models.

Global Sensitivity Analysis Proceeding in the work-flow for ^{13}C-MFA, a global sensitivity method was chosen (RS-HDMR). Multiple techniques for RS-HDMR were compared and a suitable procedure was selected (non-extended bases with backfitting). Global sensitivity analysis was introduced into the ^{13}C-MFA work-flow and applied to the focused model. Thereby, it was ensured that important fluxes (in pentose phosphate pathway, citrate cycle and glycolysis) were sensitive for the provided labeling measurements. Moreover, deep insights into the complex interplay of changes in labeling measurements and fluxes were given and visualized. Important labeling measurements were identified and insensitive fluxes were removed from the model for further analysis.

Parameter Estimation and Statistical Analysis To conduct ^{13}C-MFA, globalized multi-start strategies were applied to guarantee optimal parameter estimates. Statistics of the resulting flux map were calculated and experimental design was conducted to improve statistical identifiability.

Experimental Design In the last part of the work, the conventional (*a posteriori*) experimental design strategies for ^{13}C-MFA were extended by diversification-driven experimental design to plan sequential experiments and increase statistical identifiability of the fluxes. It was shown that, by applying multiple data-sets in one combined model, statistical identifiability was significantly increased. To balance the shortcomings of single-criterion experimental

designs, multi-objective experimental design was established to plan carbon labeling experiments. By this technique, exploration of design space became possible and comparison of measurement devices was performed, whereas LC-MS/MS was revealed to provide an economic measurement setup for the conduction of experimental design.

Fluxome of *P. chrysogenum* BCB1 Finally, ^{13}C-MFA was conducted for chemostats and fed-batch datasets. This is the first time that a ^{13}C-MFA was conducted in both cultivation processes with the same strain, comparable media and analytics. The steps from data pre-processing, network modeling, global sensitivity analysis, parameter estimation and statistical analysis described above were performed in this thesis for flux estimation.

First, chemostat experiments were used to establish the ^{13}C-MFA for the non-model organism *P. chrysogenum* and revealed a high flux through oxidative pentose phosphate pathway (split ratio of 69-75%). The resulting flux maps possessed high statistical identifiability in most reactions, although biomass production by alternative pathways were not resolved. The data indicate that the flux through oxidative pentose phosphate pathway is not only strongly controlled by penicillin production, but also by biomass formation. By comparing the results to a large scale ^{13}C-MFA network, it was revealed that major tendencies (glycolysis and pentose-phosphate pathway) are correctly predicted by the focused network. Nevertheless, the focused network remains an approximation.
By application of the combined model to the data-sets of two chemostats, it was revealed that aspartate is mainly formed in the cytosol with high statistical significance.

Data for fed-batch cultivations using the sensor-bioreactor concept were provided. First, FBA was conducted with the large scale network and it revealed that penicillin production in experiments is strongly decreased compared to *in-silico* data.
In fed-batch cultivations the strain BCB1 showed a reduced growth rate (0.007-0.008 h^{-1}) and even elevated fluxes through oxidative pentose phosphate pathway (74-80 %) compared to the chemostats. These are the highest fluxes for oxidative pentose phosphate pathway observed in *P. chrysygenum* in respect to literature. Finally overall 6.8-6.9 % of the substrate‘s carbon were used in penicillin pathway.
Upon NADPH balancing it became obvious, that a major fraction of NADPH cannot be assigned to biomass formation or penicillin production. This provided a plausible explanation for the false-prediction by FBA for the penicillin yield.

Summarizing, the main parts of the established state-of-the-art work-flow for ^{13}C-MFA was tested and automatized and in turn applied for the generation of flux maps. By adapting the industrial production-process for penicillin V, flux maps could be provided with high statistical significance for close-to-industrial cultivation conditions.

11.2 Outlook

P. chrysogenum has been subject of extensive research since the 1920ies with a strong focus on biochemical level [81]. Despite, fundamental questions regarding the penicillin V formation are still not answered. With the focus on the phenotypic level, ^{13}C-MFA is a powerful

technology established in the last decades, see Wiechert *et al.* [262]. Here, future directions for new projects will be given, facing both: ^{13}C-MFA and *P. chrysogenum.*

Process Inhomogeneities and Cell Heterogeneity

In large scale cultivations of *P. chrysogenum*, the inhomogeneities of the bio-processes (due to mixing, local feed addition, ...) are known from large scale cultivations in an industrial environment [130]. Besides that, on a different scale, the heterogeneities in *P. chrysogenum* cultures like the hyphae structure, the spatially resolved metabolism, aging and changing product formations were characterized in literature [281, 249, 59]. By both, inhomogeneities in bioprocess and heterogeneities in *P. chrysogenum* cultures, a multitude of cellular states are present within a cultivation. Resolving the separate cellular states by metabolic flux analysis remains an unsolved problem because of the underlying assumptions (i.e. homogeneous population and metabolic steady-state). In the course of an experiment, the measured labeling pattern depends nonlinearly on the fluxes. Averaging the labeling pattern for a population of cells and using them in a ^{13}C-MFA study to resolve fluxes will not necessarily yield the "average" flux distribution. In the end, deviates in the labeling pattern might be caused by population heterogeneities.
Summarizing, the industrial processes are subject to macroscopic inhomogeneities. Besides, the compartmentalization, population heterogeneities and dynamic effects caused by cellular state (e.g. cell cycle, degeneration,...) can result in a multitude of metabolic states. These complex interactions will significantly effect the product formation and strain development processes and, thus, need to be explored in depth.

NADPH balance

The main problem in the interpretation of FBA results in this work was the high gap in the NADPH balance. Further experimental data need to be generated in the wet-lab to test the hypotheses that may explain the gap (see chapter 8). The main fraction of NADPH is utilized for cysteine biosynthesis, which is formed either by sulfhydrylation and transsulfuration pathway. Both pathways were not resolvable by applying stationary ^{13}C-MFA, because they share the same atom transitions. They differ in the NADPH consumption (transsulfuration pathway: 7-8 NADPH and sufhydrylation pathway: 5 NADPH). Applying non-stationary ^{13}C-MFA may be an option to quantify the carbon flow through these pathways.
The investigation of NADPH metabolism has the potential to significantly improve industrial strains further by increasing their efficiency in NADPH utilization.

^{13}C-MFA Work-Flow

The state-of-the-art work-flow for ^{13}C-MFA was proposed in chapter 3.1. Not investigated are global uncertainty analysis and (*a priori*) experimental design. By *a priori* experimental design, a significant increase in information content of the measurement can be achieved. This is especially important for non-model organisms, where an initial knowledge about fluxes can hardly be obtained from literature.
A few works in literature are concerned with the topic of global experimental design. Chu *et al.* proposed a method for global uncertainty analysis and experimental design based on FAST (Fourier amplitude sensitivity test) [45]. Furthermore, Kontoravdi *et al.* and Rodriguez-Fernandez *et al.* used global sensitivity analysis for experimental designs [125, 202], whereas

Rodriguez-Fernandez *et al.* aimed at increasing first order effects [202]. In contrast to practical considerations contributing to experimental design, global structural identifiability was not addressed in the context of ^{13}C-MFA [45].

List of Tables

List of Figures

Bibliography

[1] Cambridge Isotope Laboratories, Inc. ; http://www.isotope.com/cil/products/-listproducts.cfm?cat_id=69&market=research&prodtypeid=40; accessed at 28.01.2011.

[2] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Computational Biology*, 9(3):e1002980, 2013.

[3] G. Aguilar-Osorio, P. van Kuyk, B. Seiboth, D. Blom, P. S. Solomon, A. Vinck, F. Kindt, H. A. B. Wösten, and R. P. de Vries. Spatial and developmental differentiation of mannitol dehydrogenase and mannitol-1-phosphate dehydrogenase in Aspergillus niger. *Eukaryotic Cell*, 9(9):1398–402, 2010.

[4] A. P. Alonso, D. L. Val, and Y. Shachar-Hill. Central metabolic fluxes in the endosperm of developing maize seeds and their implications for metabolic engineering. *Metabolic Engineering*, 13(1):96–107, 2010.

[5] A. Amiri-Simkooei. *Least-squares variance component estimation - theory and GPS applications*. PhD thesis, 2007.

[6] R. Andrae, T. Schulze-Hartung, and P. Melchior. Dos and dont's of reduced chi-squared. *ArXiv e-prints 1012.3754v1*, pages 1–12, 2010.

[7] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos. Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic Engingeering*, 9(1):68–86, 2007.

[8] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos. Accurate assessment of amino acid mass isotopomer distributions for metabolic flux analysis. *Analytical Chemistry*, 79(19):7554–9, 2007.

[9] M. R. Antoniewicz, D. F. Kraynie, L. A. Laffend, J. González-Lergier, J. K. Kelleher, and G. Stephanopoulos. Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metabolic Engineering*, 9(3):277–292, 2007.

[10] M. Araúzo-Bravo. An improved method for statistical analysis of metabolic flux analysis using isotopomer mapping matrices with analytical expressions. *Journal of Biotechnology*, 105(1-2):117–133, 2003.

[11] H. R. V. Arnstein and J. C. Crawhall. The biosynthesis of penicillin. *Biochem Journal*, 67(2):180–187, 1957.

[12] S. Athreye and A. Godley. Internationalization and technological leapfrogging in the pharmaceutical industry. *Industrial and Corporate Change*, 18(2):295–323, 2009.

[13] O. Bañuelos, J. Casqueiro, S. Steidl, S. Gutiérrez, A. Brakhage, and J. F. Martín. Subcellular localization of the homocitrate synthase in *Penicillium chrysogenum*. *Molecular Genetics and Genomics*, 266(5):711–9, 2002.

[14] Z. A. Bainbridge, R. I. Scott, and D. Perry. Oxygen Utilisation by Isopenicillin N Synthase from *Penicillium chrysogenum*. *Journal of Chemical Technology and Biotechnology*, 55(3):233–238, 1992.

[15] A. Ballio, V. Di Vittorio, and S. Russi. The isolation of trehalose and polyols from the conidia of *Penicillium chrysogenum* Thom. *Archives of Bioechmistry and Biophysics*, 107, 1964.

[16] E. Balsa-canto, A. A. Alonso, and J. R. Banga. Computational procedures for optimal experimental design in biological systems. *IET Systems Biology*, 2(4):163–172, 2008.

[17] S. Bandara, J. P. Schlöder, R. Eils, H. G. Bock, and T. Meyer. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Computational Biology*, 5(11):e1000558, 2009.

[18] J. L. Barredo, B. Diez, E. Alvarez, and J. E. Martin. Large amplification of a 35-kb DNA fragment carrying two penicillin biosynthetic genes in high penicillin producing strains of *Penicillium chrysogenum*. pages 453–459, 1989.

[19] D. M. Bates and D. G. Watts. *Nonlinear regression analysis and its applications*. John Wiley & Sons, 1st editio edition, 1988.

[20] J. Becker, C. Klopprogge, and C. Wittmann. Metabolic responses to pyruvate kinase deletion in lysine producing *Corynebacterium glutamicum*. *Microbial Cell Factory*, 7: 8, 2008.

[21] L. Benabbas, S. P. Asprey, and S. Macchietto. Curvature-based methods for designing optimally informative experiments in multiresponse nonlinear dynamic situations. *Industrial Engineering Chemistry Research*, 44(18):7120–7131, 2005.

[22] M. Berglund and M. E. Wieser. Isotopic compositions of the elements 2009 (IUPAC Technical Report). *Pure and Applied Chemistry*, 83(2):397–410, 2011.

[23] R. Berkholz, D. Röhlig, and R. Guthke. Data and knowledge based experimental design for fermentation process optimization. *Enzyme and Microbial Technology*, 27 (10):784–788, 2000.

[24] B. K. Bhattacharyya and S. K. Sen. Antibiotics business: a glimpse. 5(October): 471–476, 2006.

[25] B. K. Bhuyan and M. J. Johnson. The effect of medium constituents on penicillin production from natural materials. *Applied Microbiology*, 5(4):262–7, 1957.

[26] A. Biwer, S. Griffith, and C. Cooney. Uncertainty analysis of penicillin V production using Monte Carlo simulation. *Biotechnology and Bioengineering*, 90(2):167–79, 2005.

[27] T. Blum, S. Briesemeister, and O. Kohlbacher. MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, 10:274, 2009.

[28] E. Boles, P. D. Jong-Gubbels, and J. T. Pronk. Identification and characterization of MAE1 , the *Saccharomyces cerevisiae* structural gene encoding mitochondrial malic enzyme. *Journal of Bacteriology*, 180(11):2875–2882, 1998.

[29] S. Briesemeister, J. Rahnenführer, and O. Kohlbacher. Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics*, 26(9):1232–1238, 2010.

[30] S. Briesemeister, J. Rahnenführer, and O. Kohlbacher. YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, 38(Web Server issue):W497–502, 2010.

[31] R. Bud. *Penicillin: Triumph and Tragedy*. Oxford University Press, New York, 2007.

[32] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–555, 1989.

[33] F. Campolongo, A. Saltelli, and J. Cariboni. From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, 182(4):978–988, 2011.

[34] S. Challinor. Production of Penicillin. *Nature*, 150:688, 1942.

[35] L. T. Chang, E. L. McGrory, R. P. Elander, and D. J. Hook. Decreased production of para-hydroxypenicillin V in penicillin V fermentations. *Journal of Industrial Microbiology*, 7(3):175–179, 1991.

[36] A. G. Chapman, L. Fall, and D. E. Atkinson. Adenylate energy charge in *Escherichia coli* during growth and starvation. *Journal of Bacteriology*, 108(3):1072–1086, 1971.

[37] M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag New York Inc., 1st edition, 2000. ISBN 0-387-98935-8.

[38] M. K. Choudhary, J. M. Yoon, R. Gonzalez, and J. V. Shanks. Re-examination of metabolic fluxes in *Escherichia coli* during anaerobic fermentation of glucose using ^{13}C labeling experiments and 2-dimensional nuclear magnetic resonance (NMR) spectroscopy. *Biotechnology and Bioprocess Engineering*, 16:419–437, 2011.

[39] B. Christensen and J. Nielsen. Isotopomer analysis using GC-MS. *Metabolic Engineering*, 1(4):282–90, 1999.

[40] B. Christensen and J. Nielsen. Metabolic network analysis of *Penicillium chrysogenum* using ^{13}C-labeled glucose. *Biotechnololgy and Bioengineering*, 68(6):652–659, 2000.

[41] B. Christensen, J. Thykaer, and J. Nielsen. Metabolic characterization of high- and low-yielding strains of *Penicillium chrysogenum*. *Applied Microbiology and Biotechnology*, 54(2):212–217, 2000.

[42] L. Christensen, G. Mandrup, J. Nielsen, and J. Villadsen. A robust liquid chromatographic method for measurement of medium components during penicillin fermentations. *Analytica Chimica Acta*, 296(1):51–62, 1994.

[43] L. H. Christensen, J. Nielsen, and J. Villadsen. Degradation of penicillin-V in fermentation media. *Biotechnology and Bioengineering*, 44:165–169, 1994.

[44] L. H. Christensen, C. M. Henriksen, J. Nielsen, J. Villadsen, and M. Egel-Mitani. Continuous cultivation of *Penicillium chrysogenum*. Growth on glucose and penicillin production. *Jorunal of Biotechnology*, 42(2):95–107, 1995.

[45] Y. Chu and J. Hahn. Quantitative optimal experimental design using global sensitivity analysis via quasi-linearization. *Industrial & Engineering Chemistry Research*, 49:7782–7794, 2010.

[46] S. B. Crown and M. R. Antoniewicz. Selection of tracers for ^{13}C-metabolic flux analysis using elementary metabolite units (EMU) basis vector methodology. *Metabolic Engineering*, 14(2):150–61, 2012.

[47] S. B. Crown and M. R. Antoniewicz. Parallel labeling experiments and metabolic flux analysis: Past, present and future methodologies. *Metabolic Engineering*, 2012.

[48] S. B. Crown, W. S. Ahn, and M. R. Antoniewicz. Rational design of ^{13}C-labeling experiments for metabolic flux analysis in mammalian cells. *BMC systems biology*, 6: 43, 2012.

[49] M. Dauner and U. Sauer. GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnology Progress*, 16(4):642–649, 2000.

[50] M. Dauner, J. E. Bailey, and U. Sauer. Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. *Biotechnology and Bioengineering*, 76(2):144–56, 2001.

[51] M. de la Maza and D. Yuret. Dynamic Hill Climbing. *AI Expert*, 9(3):26–31, 1994.

[52] K. Deb, A. Member, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transaction on Evolutionary Computation*, 6(2):182–197, 2002.

[53] A. L. Demain. Inhibition of penicillin formation by amino acid analogs. *Archives of Bioechmistry and Biophysics*, 64:74–79, 1956.

[54] A. L. Demain and R. P. Elander. The β-lactam antibiotics: past, present, and future. *Antonie van Leeuwenhoek*, 75(1-2):5–19, 1999.

[55] A. D. Deshpande, K. G. Baheti, and N. Chatterjee. Degradation of β-lactam antibiotics. *Current Science*, 87(12):1683–1695, 2004.

[56] A. Diano, J. Dynesen, and J. Nielsen. Polyol synthesis in Aspergillus niger: influence of oxygen availability, carbon and nitrogen sources on the metabolism. *Biotechnology and Bioengineering*, 94(5):899–908, 2006.

[57] B. Dixon. Sulfas true significance. *Microbe*, 1(11):500–501, 2006.

[58] K. D. Dolan, L. Yang, and C. P. Trampel. Nonlinear regression technique to estimate kinetic parameters and confidence intervals in unsteady-state conduction-heated foods. *Journal of Food Engineering*, 80(2):581–593, 2007.

[59] R. D. Douma, J. M. Batista, K. M. Touw, J. A. K. W. Kiel, A. M. Krikken, Z. Zhao, T. Veiga, P. Klaassen, R. A. L. Bovenberg, J.-M. Daran, J. J. Heijnen, and W. M. van Gulik. Degeneration of penicillin production in ethanol-limited chemostat cultivations of *Penicillium chrysogenum*: A systems biology approach. *BMC Systems Biology*, 5 (1):132, 2011.

[60] P. Droste, S. Miebach, S. Niedenführ, W. Wiechert, and K. Nöh. Visualizing multi-omics data in metabolic networks with the software Omix: a case study. *Biosystems*, 105(2):154–161, 2011.

[61] A. Drysch, M. E. Massaoudi, C. Mack, R. Takors, A. A. de Graaf, and H. Sahm. Production process monitoring by serial mapping of microbial carbon flux distributions using a novel sensor reactor approach: II - ^{13}C-labeling-based metabolic flux analysis and L-lysine production. *Metabolic Engineering*, 5(2):96–107, 2003.

[62] A. Drysch, M. El Massaoudi, W. Wiechert, A. A. de Graaf, and R. Takors. Serial flux mapping of *Corynebacterium glutamicum* during fed-batch L-lysine production using the sensor reactor approach. *Biotechnology and Bioengineering*, 85(5):497–505, 2004.

[63] J. J. Durillo and A. J. Nebro. jMetal: a java framework for multi-objective optimization. *Advances in Engineering Software*, 42(10):760–771, 2011.

[64] J. J. Durillo, A. J. Nebro, F. Luna, C. A. C. Coello, and E. Alba. Convergence speed in multi-objective metaheuristics: efficiency criteria and empirical study. *International Journal for Numerical Methods in Engineering*, 84:1344–1375, 2010.

[65] J. A. Egea, R. Martí, and J. R. Banga. An evolutionary method for complex-process optimization. *Computers & Operations Research*, 37(2):315–324, 2010.

[66] R. P. Elander. Industrial production of β-lactam antibiotics. *Applied Microbiology and Biotechnology*, 61(5-6):385–92, 2003.

[67] A. F. Emery and A. V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864–876, 1998.

[68] M. Emmerling, M. Dauner, A. Ponti, J. Fiaux, M. Hochuli, T. Szyperski, K. Wüthrich, J. E. Bailey, U. Sauer, and K. Wu. Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *Journal of Bacteriology*, 184(1):152–164, 2002.

[69] T. Emri, G. Bartók, and A. Szentirmai. Regulation of specific activity of glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase in *Penicillium chrysogenum*. *FEMS Microbiology Letters*, 117:67–70, 1994.

[70] T. Emri, E. Leiter, and I. Pócsi. Effect of phenoxyacetic acid on the glutathione metabolism of *Penicillium chrysogenum*. *Journal of Basic Microbiology*, 40(2):93–104, 2000.

[71] T. Emri, B. Oláh, L. Sámi, and I. Pócsi. Does the detoxification of penicillin side-chain precursors depend on microsomal monooxygenase and glutathione S-transferase in *Penicillium chrysogenum*? *Journal of Basic Microbiology*, 43(4):287–300, 2003.

[72] M. E. Evers, H. Trip, M. A. van den Berg, R. A. L. Bovenberg, and A. J. M. Driessen. Compartmentalization and transport in β -lactam antibiotics biosynthesis. In *Advances Biochemical Engineering and Biotechnology*, pages 111–135. 2004.

[73] B. Feil, S. Kucherenko, and N. Shah. Comparison of Monte Carlo and quasi Monte Carlo sampling methods in high dimensional model representation. In *2009 First International Conference on Advances in System Simulation*, pages 12–17. IEEE, 2009. ISBN 978-1-4244-4863-0.

[74] X. Feng, L. Page, J. Rubens, L. Chircus, P. Colletti, H. B. Pakrasi, and Y. J. Tang. Bridging the gap between fluxomics and industrial biotechnology. *Journal of Biomedicine & Biotechnology*, 2010(Article ID 460717):460717, 2010.

[75] X. Feng, Y. Xu, Y. Chen, and Y. J. Tang. Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of Shewanella oneidensis MR-1. *PLoS computational biology*, 8(2):e1002376, 2012.

[76] S. J. Ferguson. ATP synthase: from sequence to ring size to the P/O ratio. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39):16755–6, 2010.

[77] J. Fiaux, Z. P. Çakar, M. Sonderegger, K. Wüthrich, T. Szyperski, and U. Sauer. Metabolic-flux profiling of the yeasts *Saccharomyces cerevisiae* and *Pichia stipitis*. *Eukaryotic Cell*, 2(1):170–180, 2003.

[78] E. Fischer and U. Sauer. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891, 2003.

[79] E. Fischer, N. Zamboni, and U. Sauer. High-throughput metabolic flux analysis based on gas chromatography - mass spectrometry derived ^{13}C constraints. *Analytical Biochemistry*, 325(2):308–316, 2004.

[80] V. Fitton, M. Rigoulet, R. Ouhabi, and B. Guérin. Mechanistic stoichiometry of yeast mitochondrial oxidative phosphorylation. *Biochemistry*, 33(32):9692–8, 1994.

[81] A. Fleming. On the antibacterial action of cultures of a *Penicillium* with special reference to their use in the isolation of B. influenza. *The British Journal of Experimental Pathology*, 10:226–236, 1929.

[82] S. Flores, G. Gosset, N. Flores, A. A. de Graaf, and F. Bolívar. Analysis of carbon metabolism in *Escherichia coli* strains with an inactive phosphotransferase system by ^{13}C labeling and NMR spectroscopy. *Metabolic Engineering*, 4(2):124–137, 2002.

[83] H. W. Florey. Penicillin: a survey. *British Medical Journal*, 2:169–171, 1944.

[84] S. S. Fong, A. Nanchen, B. O. Palsson, and U. Sauer. Latent pathway activation and increased pathway capacity enable *Escherichia coli* adaptation to loss of key metabolic enzymes. *The Journal of Biological Chemistry*, 281(12):8024–33, 2006.

[85] G. Franceschini and S. Macchietto. Model-based design of experiments for parameter precision: state of the art. *Chemical Engineering Science*, 63(19):4846–4872, 2008.

[86] H. C. Frey and S. R. Patil. Identification and review of sensitivity analysis methods. *Risk Analysis*, 22(3):553–78, 2002.

[87] C. Gancedo and K. Schwerzmann. Inactivation by glucose of phosphoenolpyruvate carboxykinase from *Saccharomyces cerevisiae*. *Archives of Microbiology*, 109:221–225, 1976.

[88] T. Goel, R. T. Haftka, W. Shyy, and L. T. Watson. Pitfalls of using a single criterion for selecting experimental designs. *International Journal for Numerical Methods in Engineering*, 75:127–155, 2008.

[89] L. H. Grimm, S. Kelly, R. Krull, and D. C. Hempel. Morphology and productivity of filamentous fungi. *Appl ied Microbiology and Biotechnology*, 69(4):375–384, 2005.

[90] S. Gudmundsson and I. Thiele. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11(489), 2010.

[91] S. Guerrero-Castillo, D. Araiza-Olivera, A. Cabrera-Orefice, J. Espinasa-Jaramillo, M. Gutiérrez-Aguilar, L. A. Luévano-Martínez, A. Zepeda-Bastida, and S. Uribe-Carvajal. Physiological uncoupling of mitochondrial oxidative phosphorylation. Studies in different yeast species. *Journal of Bioenergetics and Biomembranes*, 43(3):323–331, 2011.

[92] B. Hamad. The antibiotics market. *Nature Reviews. Drug Discovery*, 9(9):675–676, 2010.

[93] J. Harms, X. Wang, T. Kim, X. Yang, and A. S. Rathore. Defining process design space for biotech products: case study of *Pichia pastoris* fermentation. *Biotechnology Progress*, 24(3):655–662, 2008.

[94] D. M. Harris, J. A. Diderich, Z. A. van der Krogt, M. A. H. Luttik, L. M. Raamsdonk, R. A. L. Bovenberg, W. M. van Gulik, J. P. van Dijken, and J. T. Pronk. Enzymic analysis of NADPH metabolism in β-lactam-producing *Penicillium chrysogenum*: presence of a mitochondrial NADPH dehydrogenase. *Metabolic Engingeering*, 8(2):91–101, 2006.

[95] B. D. Heavner, K. Smallbone, B. Barker, P. Mendes, and L. P. Walker. Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Systems Biology*, 6(55), 2012.

[96] J. J. Heijnen. Impact of thermodynamic principles in systems biology. *Advances Biochemical Engineering and Biotechnology*, 121:139–162, 2010.

[97] E. Heinzle, Y. Yuan, S. Kumar, C. Wittmann, M. Gehre, H.-H. Richnow, P. Wehrung, P. Adam, and P. Albrecht. Analysis of ^{13}C labeling enrichment in microbial culture applying metabolic tracer experiments using gas chromatography-combustion-isotope ratio mass spectrometry. *Analytical Biochemistry*, 380(2):202–210, 2008.

[98] C. M. Henriksen, L. H. Christensen, J. Nielsen, and J. Villadsen. Growth energetics and metabolic fluxes in continuous cultures of *Penicillium chrysogenum*. *Jorunal of Biotechnology*, 45(2):149–164, 1996.

[99] C. M. Henriksen, S. S. Holm, D. Schipper, H. S. Jø rgensen, J. Nielsen, and J. Villadsen. Kinetic studies on the carboxylation of 6-amino-penicillanic acid to 8-hydroxy-penillic acid. *Process Biochemistry*, 32(2):85–91, 1997.

[100] C. M. Henriksen, J. Nielsen, and J. Villadsen. Modelling of the protonophoric uncoupling by phenoxyacetic acid of the plasma membrane potential of *Penicillium chrysogenum*. *Biotechnology and Bioengineering*, 60(6):761–767, 1998.

[101] C. M. Henriksen, J. Nielsen, and J. Villadsen. Cyclization of alpha-aminoadipic acid into the the delta-lactam 6-oxo-piperidine-2-carboxylic acid by *Penicillium chrysogenum*. *The Journal of Antibiotics*, 51(2):99–106, 1998.

[102] C. M. Henriksen, J. Nielsen, and J. Villadsen. High exogenous concentrations of phenoxyacetic acid are crucial for a high penicillin V productivity in *Penicillium chrysogenum*. *Microbiology*, 144:2001–2006, 1998.

[103] L. J. I. Horspool and Q. A. McKellar. Disposition of penicillin G sodium following intravenous and oral administration to equidae. *British Veterinary Journal*, 151(4): 401–412, 1995.

[104] Q. Hua, A. R. Joyce, B. O. Palsson, and S. S. Fong. Metabolic characterization of *Escherichia coli* strains adapted to growth on lactate. *Applied and Environmental Microbiology*, 73(14):4639–4647, 2007.

[105] M. J. Hynes, O. W. Draht, and M. A. Davis. Regulation of the acuF gene, encoding phosphoenolpyruvate carboxykinase in the filamentous fungus *Aspergillus nidulans*. *Journal of Bacteriology*, 184(1):183–190, 2002.

[106] N. Ishii and M. Tomita. Systems Biology and Biotechnology of *Escherichia coli*. chapter Multi-omic, pages 41–57. Springer Netherlands, Dordrecht, 1st edition, 2009. ISBN 978-1-4020-9393-7.

[107] S. Iwatani, S. V. Dien, K. Shimbo, K. Kubota, N. Kageyama, D. Iwahata, H. Miyano, K. Hirayama, Y. Usuda, K. Shimizu, and K. Matsui. Determination of metabolic flux changes during fed-batch cultivation from measurements of intracellular amino acids by LC-MS/MS. *Journal of Biotechnology*, 128(1):93–111, 2007.

[108] H. P. J. Bonarius, G. Schmid, and J. Tramper. Analysis of underdetermined metabolic networks: the quest of the missing constraints. *Trends in Biotechnology*, 15(8):308–314, 1997.

[109] W. M. Jaklitsch and C. P. Kubicek. Homocitrate synthase from *Penicillium chrysogenum*. Localization, purification of the cytosolic isoenzyme, and sensitivity to lysine. *The Biochemical Journal*, 269(1):247–253, 1990.

[110] M.-S. Jami, C. Barreiro, C. García-Estrada, and J.-F. Martín. Proteome analysis of the penicillin producer *Penicillium chrysogenum*: characterization of protein changes during the industrial strain improvement. *Molecular & Cellular Proteomics : MCP*, 9 (6):1182–1198, 2010.

[111] H. Jorgensen, J. Nielsen, J. Villadsen, and H. Mollgaard. Metabolic flux distributions in *Penicillium chrysogenum* during fed-batch cultivations. *Biotechnology and Bioengineering*, 46(2):117–131, 1995.

[112] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447–455, 2006.

[113] A. R. Joyce and B. O. Palsson. The model organism as a system: integrating 'omics' data sets. *Nature Reviews. Molecular Cell Biology*, 7(3):198–210, 2006.

[114] P. Jüsten, G. C. Paul, A. Nienow, and C. R. Thomas. Dependence of *Penicillium chrysogenum* growth, morphology, vacuolation, and productivity in fed-batch fermentations on impeller type and agitation intensity. *Biotechnology and Bioengineering*, 59 (6):762–775, 1998.

[115] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database issue):D277–80, 2004.

[116] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–7, 2006.

[117] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480–4, 2008.

[118] N. Kardos and A. L. Demain. Penicillin: the medicine with the greatest impact on therapeutic outcomes. *Applied Microbiology and Biotechnology*, 92(4):677–687, 2011.

[119] J. A. K. W. Kiel, I. J. van der Klei, M. A. van den Berg, R. A. L. Bovenberg, and M. Veenhuis. Overproduction of a single protein, Pc-Pex11p, results in 2-fold enhanced penicillin production by *Penicillium chrysogenum*. *Fungal Genetics and Biology*, 42(2): 154–164, 2005.

[120] P. P. King. Biotechnology. An industrial view. *Journal of Chemical Technology and Biotechnology*, 32(1):2–8, 1982.

[121] R. Kleijn. *Development and Application of ^{13}C-Labeling Techniques Analyzing the Pentose Phosphate Pathway of Penicillium chrysogenum*. PhD thesis, Delft University of Technology, 2007.

[122] R. J. Kleijn, W. A. van Winden, C. Ras, W. M. van Gulik, D. Schipper, and J. J. Heijnen. ^{13}C-labeled gluconate tracing as a direct and accurate method for determining the pentose phosphate pathway split ratio in *Penicillium chrysogenum*. *Applied Environmental Microbiology*, 72(7):4743–4754, 2006.

[123] R. J. Kleijn, J.-M. A. Geertman, B. K. Nfor, C. Ras, D. Schipper, J. T. Pronk, J. J. Heijnen, A. J. A. van Maris, and W. A. van Winden. Metabolic flux analysis of a glycerol-overproducing *Saccharomyces cerevisiae* strain based on GC-MS, LC-MS and NMR-derived C-labelling data. *FEMS Yeast Research*, 7(2):216–231, 2007.

[124] A. Konak, D. W. Coit, and A. E. Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, 91(9):992–1007, 2006.

[125] C. Kontoravdi, S. P. Asprey, E. N. Pistikopoulos, and A. Mantalaris. Application of global sensitivity analysis to determine goals for design of experiments: an example study on antibody-producing cell cultures. *Biotechnology Progress*, 21(4):1128–35, 2005.

[126] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[127] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides, and N. Shah. Monte Carlo evaluation of derivative-based global sensitivity measures. *Reliability Engineering & System Safety*, 94(7):1135–1148, 2009.

[128] M. Kusano, A. Fukushima, H. Redestig, and K. Saito. Metabolomic approaches toward understanding nitrogen metabolism in plants. *Journal of Experimental Botany*, 62(4): 1439–1453, 2011.

[129] M. Lamas-Maceiras, I. Vaca, E. Rodríguez, J. Casqueiro, and J. F. Martín. Amplification and disruption of the phenylacetyl-CoA ligase gene of *Penicillium chrysogenum* encoding an aryl-capping enzyme that supplies phenylacetic acid to the isopenicillin N-acyltransferase. *The Biochemical Journal*, 395(1):147–155, 2006.

[130] A. Lapin, J. Schmid, and M. Reuss. Modeling the dynamics of *E. coli* populations in the three-dimensional turbulent field of a stirred-tank bioreactor. A structured-segregated approach. *Chemical Engineering Science*, 61(14):4783–4797, 2006.

[131] I. Lee, K. K. Choi, and Y. Noh. Comparison Study Between Probabilistic and Possibilistic Approach for Problems With Correlated Input and Lack of Input Statistical Information. *Proceedings of the ASME 2009 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, pages 1137–1148, 2009.

[132] R. W. Leighty and M. R. Antoniewicz. Parallel labeling experiments with [U-^{13}C]glucose validate *E. coli* metabolic network model for ^{13}C metabolic flux analysis. *Metabolic Engineering*, 14(5):533–541, 2012.

[133] G. Li and H. Rabitz. General formulation of HDMR component functions with independent and correlated variables. *Journal of Mathematical Chemistry*, 50(1):99–130, 2012.

[134] G. Li, S.-w. Wang, and H. Rabitz. Practical Approaches To Construct RS-HDMR Component Functions. *Society*, (x):8721–8733, 2002.

[135] G. Li, J. Hu, S.-W. Wang, P. G. Georgopoulos, J. Schoendorf, and H. Rabitz. Random sampling-high dimensional model representation (RS-HDMR) and orthogonality of its different order component functions. *The Journal of Physical Chemistry A*, 110(7): 2474–2485, 2006.

[136] G. Li, H. Rabitz, J. Hu, Z. Chen, and Y. Ju. Regularized random-sampling high dimensional model representation (RS-HDMR). *Journal of Mathematical Chemistry*, 43(3):1207–1232, 2007.

[137] G. Li, H. Rabitz, P. E. Yelvington, O. O. Oluwole, F. Bacon, C. E. Kolb, and J. Schoendorf. Global sensitivity analysis for systems with independent and/or correlated inputs. *The journal of physical chemistry. A*, 114(19):6022–32, 2010.

[138] G. Li, R. Rey-de Castro, and H. Rabitz. D-MORPH regression for modeling with fewer unknown parameters than observation data. *Journal of Mathematical Chemistry*, 50 (7):1747–1764, 2012.

[139] X.-B. Li, G.-R. Zhao, and Y.-J. Yuan. A strategy of phosphorus feeding for repeated fed-batch fermentation of penicillin G. *Biochemical Engineering Journal*, 27(1):53–58, 2005.

[140] F. Llaneras and J. Picó. Stoichiometric modelling of cell metabolism. *Journal of bioscience and bioengineering*, 105(1):1–11, 2008.

[141] E. J. Luard. Effect of Osmotic Shock on Some Intracellular Solutes in Two Filamentous Fungi. *Journal of General Microbiology*, 128(1 982):2575–2581, 1982.

[142] J. M. Luengo, G. Revilla, J. R. Villanueva, and J. F. Martín. Lysine regulation of penicillin biosynthesis in low-producing and industrial strains of *Penicillium chrysogenum*. *Journal of General Microbiology*, 115(1):207–211, 1979.

[143] B. Luo, K. Groenke, R. Takors, C. Wandrey, and M. Oldiges. Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography-mass spectrometry. *Journal of Chromatography. A*, 1147(2):153–64, 2007.

[144] K. D. MacDonald, J. M. Hutchinson, and W. A. Gillett. Formation and segregation of heterozygous diploids between a wild-type strain and derivatives of high penicillin yield in *Penicillium chrysogenum*. *Journal of General Microbiology*, 33:385–394, 1963.

[145] M. Mallick. Differential geometry measures of nonlinearity with applications to ground target tracking. *Proceedings of the 8th International Conference on Information Fusion*, pages 219–226, 2005.

[146] M. Mallick and B. la Scala. Differential geometry measures of nonlinearity for ground moving target indicator (GMTI) filtering. *2005 8th International Conference on Information Fusion*, 1:219–226, 2005.

[147] T. A. Mara and S. Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107:115–121, 2012.

[148] J. R. R. A. Martins, P. Sturdza, and J. J. Alonso. The complex-step derivative approximation. *ACM Transactions on Mathematical Software*, 29(3):245–262, 2003.

[149] A. Marx, A. A. de Graaf, W. Wiechert, L. Eggeling, and H. Sahm. Determination of the fluxes in the central metabolism of *Corynebacterium glutamicum* by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnology and Bioengineering*, 49(2):111–129, 1996.

[150] M. E. Massaoudi, J. Spelthahn, A. Drysch, A. de Graaf, and R. Takors. Production process monitoring by serial mapping of microbial carbon flux distributions using a novel sensor reactor approach: I - Sensor reactor system. *Metabolic Engineering*, 5(2): 86–95, 2003.

[151] D. P. Maxwell, M. D. Maxwell, G. Hänssler, V. N. Armentrout, G. M. Murray, and H. C. Hoch. Microbodies and glyoxylate-cycle enzyme activities in filamentous fungi. *Planta*, 124(2):109–123, 1975.

[152] J. B. McKinlay, Y. Shachar-Hill, J. G. Zeikus, and C. Vieille. Determining Actinobacillus succinogenes metabolic pathways and fluxes by NMR and GC-MS analyses of ^{13}C-labeled metabolic product isotopomers. *Metabolic Engineering*, 9(2):177–92, 2007.

[153] F. W. McLafferty. Tandem mass spectrometry. *Science*, 214:280–287, 1981.

[154] W. H. Meijer, L. Gidijala, S. Fekken, J. A. K. W. Kiel, M. A. van den Berg, R. Lascaris, R. A. L. Bovenberg, and I. J. van der Klei. Peroxisomes are required for efficient penicillin biosynthesis in *Penicillium chrysogenum*. *Applied and Environmental Microbiology*, 76(17):5702–5709, 2010.

[155] S. Meinert. *Optimization of a fast sampling set up and two cultivation systems for ^{13}C metabolic flux analysis with Penicillium chrysogenum*. PhD thesis, Universität Innsbruck, 2012.

[156] C. M. Metallo, J. L. Walther, and G. Stephanopoulos. Evaluation of ^{13}C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of Biotechnology*, 144 (3):167–174, 2009.

[157] G. Michal and D. Schomburg. *Biochemical pathways: an atlas of biochemistry and molecular biology*. John Wiley & Sons, 2nd editio edition, 2012. ISBN 978-0-470-14684-2.

[158] S. Miebach. *Charakterisierung und Validierung der ^{13}C-Stoffflussanalyse im Parallelansatz*. PhD thesis, Universität Bielefeld, 2011.

[159] E. A. Miles and A. P. J. Trinci. Effect of pH and temperature on morphology of batch and chemostat cultures of *Penicillium chrysogenum*. *Transactions of the British Mycological Society*, 81(2):193–200, 1983.

[160] P. Millard, F. Letisse, S. Sokol, and J.-C. Portais. IsoCor: correcting MS data in isotope labeling experiments. *Bioinformatics*, 28(9):1294–1296, 2012.

[161] M. Möllney, W. Wiechert, D. Kownatzki, and A. A. de Graaf. Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnol Bioengineering*, 66(2):86–103, 1999.

[162] H. Motulsky and A. Christopoulos. *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* Oxford University Press, 2004.

[163] F. Mu, R. F. Williams, C. J. Unkefer, P. J. Unkefer, J. R. Faeder, and W. S. Hlavacek. Carbon-fate maps for metabolic reactions. *Bioinformatics*, 23(23):3193–3199, 2007.

[164] W. H. Müller, T. P. van der Krift, A. J. Krouwer, H. A. Wösten, L. H. van der Voort, E. B. Smaal, and A. J. Verkleij. Localization of the pathway of the penicillin biosynthesis in *Penicillium chrysogenum*. *The EMBO Journal*, 10(2):489–95, 1991.

[165] W. H. Müller, R. A. Bovenberg, M. H. Groothuis, F. Kattevilder, E. B. Smaal, L. H. M. Van der Voort, and A. J. Verkleij. Involvement of microbodies in penicillin biosynthesis. *Biochimica et Biophysica Acta*, 1116(2):210–213, 1992.

[166] R. O. Mumma, C. L. Fergus, and R. D. Sekura. The lipids of thermophilic fungi: lipid composition comparisons between thermophilic and mesophilic fungi. *Lipids*, 5 (1):100–103, 1970.

[167] A. Nakaya, T. Katayama, M. Itoh, K. Hiranuka, S. Kawashima, Y. Moriya, S. Okuda, M. Tanaka, T. Tokimatsu, Y. Yamanishi, A. C. Yoshizawa, M. Kanehisa, and S. Goto. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Research*, 41(Database issue):D353–7, 2013.

[168] U. Nasution. *A dynamic and steady state metabolome study of central metabolism and its relation with the penicillin biosynthesis pathway in Penicillium chrysogenum.* PhD thesis, Delft University of Technology, 2007.

[169] U. Nasution, W. M. van Gulik, A. Proell, W. A. van Winden, and J. J. Heijnen. Generating short-term kinetic responses of primary metabolism of *Penicillium chrysogenum* through glucose perturbation in the bioscope mini reactor. *Metabolic Engineering*, 8 (5):395–405, 2006.

[170] U. Nasution, W. M. van Gulik, C. Ras, A. Proell, and J. J. Heijnen. A metabolome study of the steady-state relation between central metabolism, amino acid biosynthesis and penicillin production in *Penicillium chrysogenum*. *Metabolic Engingeering*, 10(1): 10–23, 2008.

[171] M. H. Nathanson and G. M. Saidel. Multiple-objective criteria for optimal experimental design: application to ferrokinetics. *The American Journal of Physiology*, 248(3 Pt 2): R378–86, 1985.

[172] A. Nebro, J. Durillo, J. Garcia-Nieto, C. Coello Coello, F. Luna, and E. Alba. SMPSO: A new PSO-based metaheuristic for multi-objective optimization. *2009 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, (2):66–73, 2009.

[173] D. L. Nelson and M. M. Cox. Glycolysis, gluconeogenesis, and the pentose phosphate pathway. In *Lehninger Principles of Biochemistry*, pages 521–559. Wh Freeman, 5th edition, 2008.

[174] J. Nielsen. *Physiological engineering aspects of Penicillium chrysogenum*. World Scientific Publication Co. Pte. Ltd., 1997. ISBN 981-02-2765-5.

[175] J. Nielsen and M. C. Jewett. Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*. *FEMS Yeast Research*, 8(1):122–31, 2008.

[176] J. Nielsen and H. S. Jorgensen. Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of *Penicillium chrysogenum*. *Biotechnology Progress*, 11(3):299–305, 1995.

[177] J. Nielsen, C. L. Johansen, M. Jacobsen, P. Krabben, and J. Villadsen. Pellet formation and fragmentation in submerged cultures of *Penicillium chrysogenum* and its relation to penicillin production. *Biotechnology Progress*, 11(1):93–98, 1995.

[178] J. H. Nielsen and J. Villadsen. *Bioreaction engineering principles*. Kluwer Academic Pub, New York, 1994.

[179] J. G. Nijland, B. Ebbendorf, M. Woszczynska, R. Boer, R. A. L. Bovenberg, and A. J. M. Driessen. Nonlinear biosynthetic gene cluster dose effect on penicillin production by *Penicillium chrysogenum*. *Applied and Environmental Microbiology*, 76(21): 7109–7115, 2010.

[180] K. Nöh and W. Wiechert. The benefits of being transient: isotope-based metabolic flux analysis at the short time scale. *Applied Microbiology and Biotechnology*, 91(5): 1247–1265, 2011.

[181] K. Nöh, A. Wahl, and W. Wiechert. Computational tools for isotopically instationary ^{13}C labeling experiments under metabolic steady state conditions. *Metabolic Engineering*, 8(6):554–577, 2006.

[182] H. Ogata, S. Goto, W. Fujibuchi, and M. Kanehisa. Computation with the KEGG pathway database. *BioSystems*, 47(1-2):119–128, 1998.

[183] J. D. Orth, I. Thiele, and B. O. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–8, 2010.

[184] M. Papagianni. Fungal morphology and metabolite production in submerged mycelial processes. *Biotechnology Advances*, 22(3):189–259, 2004.

[185] G. C. Paul, C. A. Kent, and C. R. Thomas. Hyphal Vacuolation and Fragmentation in *Penicillium chrysogenum*. *Biotechnology and Bioengineering*, 44(5):655–660, 1994.

[186] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews. Drug Discovery*, 9(3):203–14, 2010.

[187] S. J. Pirt and R. C. Righelato. Effect of Growth Rate on the Synthesis of Penicillin by *Penicillium chrysogenum* in Batch and Chemostat Cultures. *Applied microbiology*, 15 (6):1284–90, 1967.

[188] P. D. N. Pissara, J. Nielsen, and M. J. Bazin. Pathway kinetic and metabolic control analysis of a high-yielding strain of Pencillium chrysogenum during fed batch cultivations. *Biotechnology and Bioengineering*, 51(2):168–176, 1996.

[189] C. F. Poole, S. Smghawangcha, F. HSU, A. Zlatkis, A. Arfwkdsson, and J. Vessman. New electron-capturing pentafluorophenyldialkylchlorosilanes as versatile derivatizing reagents for gas chromatography. *Journal of Chromatography*, 199:123–142, 1980.

[190] D. Porro, E. Martegani, B. M. Ranzi, and L. Alberghina. Oscillations in continuous cultures of budding yeast: a segregated parameter analysis. *Biotechnology and Bioengineering*, 32(4):411–417, 1988.

[191] G. Pujol. Simplex-based screening designs for estimating metamodels. *Reliability Engineering*, 94(7):1156–1160, 2009.

[192] F. Pukelsheim. *Optimal design of experiments*. John Wiley & Sons, 1993. ISBN 0-471-61971-X.

[193] S. Queener and R. Swartz. Penicillins: biosynthetic and semisynthetic. *Economic Microbiology*, 3:35–123, 1979.

[194] H. Rabitz. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2-3):197–233, 1999.

[195] H. Rabitz, O. E. Alis, J. Shorter, and K. Shim. Efficient input-output model representations. *Computer Physics Communications*, 117(1-2):11–20, 1999.

[196] B. Raper. The Development of improved penicillin- producing molds. *Annals of the New York Academy of Sciences*, 48(2):41–52, 1941.

[197] D. A. Ratkowsky. *Nonlinear regression modeling - a unified practical approach*. Marcel Dekker, Inc., 1983.

[198] M. Ratto, A. Pagano, and P. Young. State Dependent Parameter metamodelling and sensitivity analysis. *Computer Physics Communications*, 177(11):863–876, 2007.

[199] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics (Oxford, England)*, 25(15):1923–1929, 2009.

[200] G. Revilla, M. J. Lopez-nieto, J. M. Luengo, and J. F. Martin. Carbon Catabolite Repression of Penicillin Biosynthesis by *Penicillium chrysogenum*. *The Journal of Antibiotics*, XXXVII(7):781–789, 1984.

[201] R. C. Righelato, A. P. Trinci, S. J. Pirt, and A. Peat. The influence of maintenance energy and growth rate on the metabolic activity, morphology and conidiation of *Penicillium chrysogenum*. *Journal of General Microbiology*, 50(3):399–412, 1968.

[202] M. Rodriguez-Fernandez, S. Kucherenko, C. Pantelides, and N. Shah. Optimal experimental design based on global sensitivity analysis. *17th European Symposium on Computer Aided Process Engineering - ESCAPE17*, 1:1–6, 2007.

[203] M. Rühl, B. Rupp, K. Nöh, W. Wiechert, U. Sauer, and N. Zamboni. Collisional fragmentation of central carbon metabolites in LC-MS/MS increases precision of ^{13}C metabolic flux analysis. *Biotechnology and Bioengineering*, 109(3):763–771, 2012.

[204] D. D. Y. Ryu and J. Hospodka. Quantitative physiology of *Penicillium chrysogenum* in penicillin fermentation. *Biotechnology and Bioengineering*, 22(2):289–298, 1980.

[205] A. Saltelli and P. Annoni. How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12):1508–1517, 2010.

[206] A. Saltelli and R. Bolado. An alternative way to compute Fourier amplitude sensitivity test (FAST). *Computational Statistics & Data Analysis*, 26(4):445–460, 1998.

[207] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis - The Primer*. John Wiley & Sons Ltd, West Sussex, chichester edition, 2008.

[208] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270, 2010.

[209] Sandoz GmbH. Die Geschichte des ersten säurestabilen Oralpenicillins (Penicillin V); http://www.sandoz.at/assets/content/de/produkte/antibiotikamonitor/pdfs/Antibiotikamonitor_3_2000-Mit_Penicillin_V_ins_dritte_Jahrtausend.pdf, accessed at 25.06.2013.

[210] K. Sanford, P. Soucaille, G. Whited, and G. Chotani. Genomics to fluxomics and physiomics - pathway engineering. *Current Opinion in Microbiology*, 5(3):318–22, 2002.

[211] J. Schaub. *Isotopisch instationäre ^{13}C-Stoffflussanalyse in Escherichia coli*. PhD thesis, Universität Stuttgart, 2007.

[212] J. Schaub, K. Mauch, and M. Reuss. Metabolic flux analysis in *Escherichia coli* by integrating isotopic dynamic and isotopic stationary ^{13}C labeling data. *Biotechnology and Bioengineering*, 99(5):1170–1185, 2008.

[213] J. Schellenberger, D. C. Zielinski, W. Choi, S. Madireddi, V. Portnoy, D. A. Scott, J. L. Reed, A. L. Osterman, and B. O. Palsson. Predicting outcomes of steady-state ^{13}C isotope tracing experiments with Monte Carlo sampling. *BMC Systems Biology*, 6 (9), 2012.

[214] K. Schmidt, J. Nielsen, and J. Villadsen. Quantitative analysis of metabolic fluxes in *Escherichia coli*, using two-dimensional NMR spectroscopy and complete isotopomer models. *Journal of Biotechnology*, 71(1-3):175–190, 1999.

[215] T. Scott and C. Oldenhof. High level group on chemicals a 25-year landslide in the manufacture and business of active pharmaceutical ingredients (APIs) in europe between 1983 - 2008.

[216] G. A. F. Seber and C. J. Wild. *Nonlinear regression*. John Wiley & Sons, 2003.

[217] J. M. Sedivyt and D. G. Fraenkelt. Fructose bisphosphatase of *Saccharomyces cerevisiae* cloning , disruption and regulation of the FBPI structural gene. *Journal of Molecular Biology*, 186:307–319, 1985.

[218] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26 (10):1135–45, 2008.

[219] K. Singh and M. J. Johnson. Evaluation of precursors for penicillin G. *Journal of Bacteriology*, 56(3), 1948.

[220] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001.

[221] I. M. Sobol and S. S. Kucherenko. Global sensitivity indices for nonlinear mathematical models, Review. *Wilmott Magazine*, (2000):2–7, 2005.

[222] P. S. Solomon, O. D. C. Waters, and R. P. Oliver. Decoding the mannitol enigma in filamentous fungi. *Trends in Microbiology*, 15(6):257–62, 2007.

[223] F. V. Soltero and M. J. Johnson. The effect of the carbohydrate nutrition on penicillin production by *Penicillium chrysogenum* Q-176. *Applied Microbiology*, 1(1):52–7, 1953.

[224] H.-S. Song and D. Ramkrishna. Reduction of a set of elementary modes using yield analysis. *Biotechnology and Bioengineering*, 102(2):554–568, 2009.

[225] G. Sriram and J. V. Shanks. Improvements in metabolic flux analysis using carbon bond labeling experiments: bondomer balancing and Boolean function mapping. *Metabolic Engineering*, 6(2):116–132, 2004.

[226] O. Srour, J. D. Young, and Y. C. Eldar. Fluxomers: A new approach for ^{13}C metabolic flux analysis. *BMC Systems Biology*, 5(129), 2011.

[227] C. M. Stevens, P. Vohra, and C. W. de Long. Utilization of valine in the biosynthesis of penicillins. 211(1):297–300, 1954.

[228] L. Stols and M. I. Donnelly. Production of succinic acid through overexpression of NAD(+) -dependent malic enzyme in an *Escherichia coli* mutant. *Applied and Environmental Microbiology*, 63(7):2695–2701, 1997.

[229] P. F. Suthers, A. P. Burgard, M. S. Dasika, F. Nowroozi, S. Van Dien, J. D. Keasling, and C. D. Maranas. Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. *Metabolic Engineering*, 9(5-6):387–405, 2007.

[230] T. Szyperski. Biosynthetically directed fractional ^{13}C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism. *European journal of biochemistry / FEBS*, 232(2):433–48, 1995.

[231] T. Szyperski. ^{13}C-NMR, MS and metabolic flux balancing in biotechnology research. *Quarterly Reviews of Biophysics*, 31(1):41–106, 1998.

[232] Y. Tang, F. Pingitore, A. Mukhopadhyay, R. Phan, T. C. Hazen, and J. D. Keasling. Pathway confirmation and flux analysis of central metabolic pathways in Desulfovibrio vulgaris hildenborough using gas chromatography-mass spectrometry and Fourier transform-ion cyclotron resonance mass spectrometry. *Journal of Bacteriology*, 189(3): 940–949, 2007.

[233] S. Tarantola, N. Giglioli, J. Jesinghaus, and A. Saltelli. Can global sensitivity analysis steer the implementation of models for environmental assessments and decision-making? *Stochastic Environmental Research and Risk Assessment*, 16(1):63–76, 2002.

[234] I. Thiele and B. O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010.

[235] J. Thykaer and J. Nielsen. Metabolic engineering of β-lactam production. *Metabolic Engineering*, 5(1):56–69, 2003.

[236] J. Tillack, N. Paczia, K. Nöh, W. Wiechert, and S. Noack. Error Propagation Analysis for Quantitative Intracellular Metabolomics. *Metabolites*, 2(4):1012–1030, 2012.

[237] V. Tiller, J. Meyerhoff, D. Sziele, K. Schügerl, and K. H. Bellgardt. Segregated mathematical model for the fed-batch cultivation of a high-producing strain of *Penicillium chrysogenum*. *Journal of Biotechnology*, 34(2):119–131, 1994.

[238] Y. Toya, N. Ishii, T. Hirasawa, M. Naba, K. Hirai, K. Sugawara, S. Igarashi, K. Shimizu, M. Tomita, and T. Soga. Direct measurement of isotopomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis. *Journal of Chromatography A*, 1159(1-2):134–141, 2007.

[239] M. van de Kamp, A. J. Driessen, and W. N. Konings. Compartmentalization and transport in β-lactam antibiotic biosynthesis by filamentous fungi. *Antonie van Leeuwenhoek*, 75(1-2):41–78, 1999.

[240] M. A. van den Berg, I. Westerlaken, C. Leeflang, R. Kerkman, and R. A. L. Bovenberg. Functional characterization of the penicillin biosynthetic gene cluster of *Penicillium chrysogenum* Wisconsin54-1255. *Fungal Genetics and Biology*, 44(9):830–844, 2007.

[241] M. A. van den Berg, R. Albang, K. Albermann, J. H. Badger, J.-M. Daran, A. J. M. Driessen, C. Garcia-Estrada, N. D. Fedorova, D. M. Harris, W. H. M. Heijne, V. Joardar, J. A. K. W. Kiel, A. Kovalchuk, J. F. Martin, W. C. Nierman, J. G. Nijland, J. T. Pronk, J. A. Roubos, I. J. van der Klei, N. N. M. E. van Peij, M. Veenhuis, H. von Döhren, C. Wagner, J. Wortman, and R. A. L. Bovenberg. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nature Biotechnology*, 26 (10):1161–1168, 2008.

[242] W. M. van Gulik, W. T. de Laat, J. L. Vinke, and J. J. Heijnen. Application of metabolic flux analysis for the identification of metabolic bottlenecks in the biosynthesis of penicillin-G. *Biotechnology and Bioengineering*, 68(6):602–18, 2000.

[243] W. M. van Gulik, M. R. Antoniewicz, W. T. DeLaat, J. L. Vinke, and J. J. Heijnen. Energetics of growth and penicillin production in a high-producing strain of *Penicillium chrysogenum*. *Biotechnology and Bioengineering*, 72(2):185–193, 2001.

[244] J. van Ooyen, S. Noack, M. Bott, A. Reth, and L. Eggeling. Improved L-lysine production with *Corynebacterium glutamicum* and systemic insight into citrate synthase flux and activity. *Biotechnology and Bioengineering*, 109(8):2070–2081, 2012.

[245] W. van Winden, P. Verheijen, and S. Heijnen. Possible pitfalls of flux calculations based on ^{13}C-labeling. *Metabolic engineering*, 3(2):151–62, 2001.

[246] W. A. van Winden, C. Wittmann, E. Heinzle, and J. J. Heijnen. Correcting mass isotopomer distributions for naturally occurring isotopes. *Biotechnology and Bioengineering*, 80(4):477–479, 2002.

[247] W. A. van Winden, W. M. van Gulik, D. Schipper, P. J. T. Verheijen, P. Krabben, J. L. Vinke, and J. J. Heijnen. Metabolic flux and metabolic network analysis of *Penicillium chrysogenum* using 2D [^{13}C, 1H] COSY NMR measurements and cumulative bondomer simulation. *Biotechnology and Bioengineering*, 83(1):75–92, 2003.

[248] W. A. van Winden, J. C. van Dam, C. Ras, R. J. Kleijn, J. L. Vinke, W. M. van Gulik, J. J. Heijnen, W. A. V. Winden, J. C. V. Dam, and W. M. V. Gulik. Metabolic-flux analysis of *Saccharomyces cerevisiae* CEN.PK113-7D based on mass isotopomer measurements of ^{13}C-labeled primary metabolites. *FEMS Yeast Research*, 5(6-7):559–568, 2005.

[249] B. Vanhoutte, M. N. Pons, C. R. Louvel, and H. Vivier. Characterization of *Penicillium chrysogenum* Physiology in Submerged Cultures by Color and Monochrome Image Analysis. *Biotechnology and Bioengineering*, 48(1):1–11, 1995.

[250] P. A. Vanrolleghem, M. van Daele, and D. Dochain. Practical identifiability of a biokinetic model of activated sludge respiration. *Water Research*, 29(11):2561–2570, 1995.

[251] A. Varma and B. O. Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10):994–998, 1994.

[252] K. J. Versyck and J. F. V. Impe. Trade-offs in Design of Fed-Batch Experiments for Optimal Estimation of Biokinetic Parameters. *Proceedings of the 1998 IEEE International Conference on Control Applications*, 1:51–55, 1998.

[253] D. Voet, J. G. Voet, and C. W. Pratt. *Lehrbuch der Biochemie*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1st edition, 2002.

[254] G. von Jagow and M. Klingenberg. Pathways of hydrogen in mitochondria of Saccharomyces carlsbergensis. *European Journal of Biochemistry*, 12(3):583–592, 1970.

[255] S. A. Wahl, M. Dauner, and W. Wiechert. New tools for mass isotopomer data evaluation in ^{13}C flux analysis: mass isotope correction, data consistency checking, and precursor relationships. *Biotechnology and Bioengineering*, 85(3):259–268, 2004.

[256] J. L. Walther, C. M. Metallo, J. Zhang, and G. Stephanopoulos. Optimization of ^{13}C isotopic tracers for metabolic flux analysis in mammalian cells. *Metabolic Engineering*, 14(2):1–10, 2011.

[257] S.-W. Wang, P. G. Georgopoulos, G. Li, and H. Rabitz. Random sampling-high dimensional model representation (RS-HDMR) with nonuniformly distributed variables: application to an integrated multimedia / multipathway exposure and dose model for trichloroethylene. *Society*, 107(23):4707–4716, 2003.

[258] I. N. Watt, M. G. Montgomery, M. J. Runswick, A. G. W. Leslie, and J. E. Walker. Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39):16823–16827, 2010.

[259] M. Weitzel. *High performance algorithms for metabolic flux analysis*. PhD thesis, Universität Siegen, 2009.

[260] M. Weitzel, K. Nöh, T. Dalman, S. Niedenführ, B. Stute, and W. Wiechert. ^{13}CFLUX2-high-performance software suite for ^{13}C-metabolic flux analysis. *Bioinformatics*, 29(1): 143–145, 2013.

[261] W. Wiechert. An introduction to ^{13}C metabolic flux analysis. *Genetic Engineering*, 24: 215–38, 2002.

[262] W. Wiechert. The thermodynamic meaning of metabolic exchange fluxes. *Biophysical journal*, 93(6):2255–64, 2007.

[263] W. Wiechert and A. A. de Graaf. Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. *Biotechnology and Bioengineering*, 55(1):101–117, 1997.

[264] W. Wiechert and K. Nöh. From stationary to instationary metabolic flux analysis. *Advances Biochemical Engineering and Biotechnology*, 92:145–172, 2005.

[265] W. Wiechert, A. A. D. Graaf, and a. a. de Graaf. In Vivo Stationary Flux Analysis by 13 C Labeling Experiments Abstract :. *Advances in biochemical engineering/biotechnology*, 54:1–50, 1996.

[266] W. Wiechert, C. Siefke, A. A. de Graaf, and A. Marx. Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *Biotechnology and Bioengineering*, 55(1):118–135, 1997.

[267] W. Wiechert, M. Möllney, N. Isermann, M. Wurzel, and A. A. de Graaf. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnology and Bioengineering*, 66(2):69–85, 1999.

[268] W. Wiechert, M. Möllney, S. Petersen, and A. A. de Graaf. A universal framework for ^{13}C metabolic flux analysis. *Metab Eng*, 3(3):265–283, 2001.

[269] G. Winter and J. O. Krömer. Fluxomics - connecting 'omics analysis and phenotypes. *Environmental Microbiology*, (DOI: 10.1111/1462-2920.12064), 2013.

[270] C. Wittmann. Fluxome analysis using GC-MS. *Microbial Cell Factories*, 6(6), 2007.

[271] C. Wittmann and E. Heinzle. Mass spectrometry for metabolic flux analysis. *Biotechnology and Bioengineering*, 62(6):739–750, 1999.

[272] C. Wittmann and E. Heinzle. Modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacterium glutamicum* by mass spectrometry. *Metabolic Engineering*, 3(2):173–191, 2001.

[273] C. Wittmann and E. Heinzle. Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum*: a novel approach for metabolic flux analysis. *European Journal of Biochemistry*, 268(8):2441–2455, 2001.

[274] www.nobleprize.org. The Nobel Prize in Physiology or Medicine 1945. Http://www.nobelprize.org/nobel_prizes/medicine/laureates/1945/; accessed at 17.06.2013.

[275] J. P. Wynn, A. A. Hamid, M. Midgley, and C. Ratledge. Short Communication: Widespread occurrence of ATP:citrate lyase and carnitine acetyltransferase in filamentous fungi. *Journal of Microbiology*, 14(1):145–147, 1998.

[276] C. Xu and G. Z. Gertner. Uncertainty and sensitivity analysis for models with correlated parameters. *Reliability Engineering & System Safety*, 93(10):1563–1573, 2008.

[277] J. Yuan, B. D. Bennett, and J. D. Rabinowitz. Kinetic flux profiling for quantitation of cellular metabolic fluxes. *Nature protocols*, 3(8):1328–40, 2008.

[278] Y. Yuan, T. H. Yang, E. Heinzle, and A. Campus. ^{13}C metabolic flux analysis for larger scale cultivation using gas chromatography-combustion-isotope ratio mass spectrometry. *Metabolic Engineering*, 12(4):392–400, 2010.

[279] N. Zamboni, E. Fischer, and U. Sauer. FiatFlux–a software for metabolic flux analysis from ^{13}C-glucose experiments. *BMC Bioinformatics*, 6(209), 2005.

[280] N. Zamboni, S.-M. Fendt, M. Rühl, and U. Sauer. ^{13}C-based metabolic flux analysis. *Nature Protocols*, 4(6):878–92, 2009.

[281] T. C. Zangirolami, C. L. Johansen, J. Nielsen, and S. B. Jø rgensen. Simulation of penicillin production in fed-batch cultivations using a morphologically structured model. *Biotechnology and Bioengineering*, 56(6):593–604, 1997.

[282] Z. Zhao, K. Kuijvenhoven, W. M. van Gulik, J. J. Heijnen, W. a. van Winden, and P. J. T. Verheijen. Cytosolic NADPH balancing in *Penicillium chrysogenum* cultivated on mixtures of glucose and ethanol. *Applied microbiology and biotechnology*, 89(1): 63–72, 2011.

[283] Z. Zhao, K. Kuijvenhoven, W. M. van Gulik, J. J. Heijnen, W. a. van Winden, and P. J. T. Verheijen. Cytosolic NADPH balancing in *Penicillium chrysogenum* cultivated on mixtures of glucose and ethanol. *Applied microbiology and biotechnology*, 89(1): 63–72, 2011.

[284] Z. Zheng. *^{13}C labeling studies in the central metabolism of Penicillium chrysogenum.* PhD thesis, Technische Universiteit Delft, 2011.

[285] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: improving the strength pareto evolutionary algorithm. *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, pages 95–100, 2001.

[286] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Transaction on Evolutionary Computation*, 7(2):117–132, 2002.

Part V

Appendix

Appendix A

Global Sensitivity Analysis

A.1 Global Sensitivity Analysis

A.1.1 RS-HDMR with Orthonormal Polynomials

RS-HDMR for non-uniformly distributed, independent and/or correlated (non-orthogonal) inputs will be described in this section according to Wang et al. [257]. The sampled inputs x_i are transformed to the interval $[0, 1]$, resulting in a n-dimensional unit hypercube $K^n = \{(x_1, x_2, x_3, \dots x_n)\ |0 \leq x_i \leq 1,\ i = 1, 2, 3, \dots, n\}$. Each input possesses the probability density function $w_i(x_i)$, which satisfies

$$\begin{cases} w_i(x_i) \geq 0 & , 0 \leq x_i \leq 1 \\ \int_0^1 w_i(x_i) dx_i = 1 & , i = 1, 2, \dots, n \end{cases} \tag{A.1}$$

Here, $w_i(x_i)$ may be not separable if there exist some dependencies between inputs x_i and x_j. First, we can define that $\mathbf{x}^i$ and $\mathbf{x}^{ij}$ are all sampled input $\mathbf{x}$ except input parameter x_i and $(x_i,\ x_j)$ within range K^{n-1} and K^{n-2}, respectively. Additionally, we introduce conditional variances, for example, for a fixed value of x_i, the probability density function for $\mathbf{x}$ is $w_{\mathbf{x}^i|x_i}(\mathbf{x}^i)$.

$$w_i(x_i) = \int_{K^{n-1}} w(x_i, \mathbf{x}^i) d\mathbf{x}^i \tag{A.2}$$

$$w_{ij}(x_i, x_j) = \int_{K^{n-2}} w(x_i, x_j, \mathbf{x}^{ij}) d\mathbf{x}^{ij} \tag{A.3}$$

$$w_{\mathbf{x}^i|x_i}(\mathbf{x}^i) = w(x_i, \mathbf{x}^i)/w_i(x_i) \tag{A.4}$$

$$w_{\mathbf{x}^{ij}|x_{ij}}(\mathbf{x}^{ij}) = w(x_i, x_j, \mathbf{x}^{ij})/w_{ij}(x_i, x_j) \tag{A.5}$$

Thus, we can write the terms of the hierarchical correlated function expansion of our RS-HDMR model for the function $f(\mathbf{x})$:

$$f_0 = \int_{K^n} w(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \tag{A.6}$$

$$f_i(x_i) = \int_{K^{n-1}} w_{\mathbf{x}^i|x_i}(\mathbf{x}^i) f(\mathbf{x}) d\mathbf{x}^i - f_0 \tag{A.7}$$

$$f_{ij}(x_i, x_j) = \int_{K^{n-2}} w_{\mathbf{x}^{ij}|x_i x_j}(\mathbf{x}^{ij}) f(\mathbf{x}) d\mathbf{x}^{ij} - f_i(x_i) - f_j(x_j) - f_0 \tag{A.8}$$

By extending the method for correlated inputs, the mutual orthogonality from equation 6.2.1 cannot be preserved.
For estimation of the separate component functions, orthonormal polynomials $\phi_r^i(x_i)$ are chosen with respect to the density $w_i(x_i)$.

$$\int_0^1 w_i(x_i)\phi_r^i(x_i)dx_i = 0 \quad \text{for all}\, r, i \tag{A.9}$$

$$\int_0^1 w_i(x_i)\left[\phi_r^i(x_i)\right]^2 dx_i = 1 \quad \text{for all}\, r, i \tag{A.10}$$

$$\int_0^1 w_i(x_i)\phi_p^i(x_i)\phi_q^i(x_i)dx_i = 0 \quad p \neq q \tag{A.11}$$

The component functions consists of polynomials ϕ and the model coefficients α_r^i, β_{rp}^{ij} and γ_{rpq}^{ijk}.

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \phi_r^i(x_i) \tag{A.12}$$

$$f_{ij}(x_i, x_j) \approx \sum_{r=1}^{l}\sum_{p'=1}^{l'} \beta_{rp'}^{ij}\phi_r^i(x_i)\phi_{p'}^j(x_j) \tag{A.13}$$

$$f_{ijk}(x_i, x_j, x_k) \approx \sum_{r=1}^{m}\sum_{p'=1}^{m'}\sum_{q=1}^{m''} \gamma_{rpq}^{ij}\phi_r^i(x_i)\phi_{p'}^j(x_j)\phi_q^k(x_k) \tag{A.14}$$

The degree of the polynomials (k, l, $l^{'}$, m, m', m'') can be chosen specific for each component function. To build the first order component functions f_i of one input, $r = 1, 2, ...k$ expansion coefficients α_r^i are used for the separate orthonormal polynomials ϕ_r^i . For a second order function $l{\cdot}l'$ and for a third order function $m{\cdot}m'{\cdot}m''$ coefficients are used. Usually, polynomials of an order up to three or four are used. Higher order polynomials tend to oscillations and the number of needed coefficients rises [136]. For a model with 30 inputs with zeroth to second order correlations and polynomial‘s order of 3 results in a RS-HDMR meta-model of 1395 expansion coefficients to be determined. We will now focus on the estimations of the separate terms in a RS-HDMR function expansion.

First, the constant term f_0 can be estimated:

$$f_0 \approx \frac{1}{N}\sum_{s=1}^{N} y(\mathbf{x}^{(s)}) \tag{A.15}$$

The coefficients for the single parameters can be efficiently determined by Monte Carlo integration:

$$\begin{aligned}
\alpha_r^i &= \int_0^1 w_i(x_i) f_i(x_i)\phi_r^i(x_i)dx_i && \text{(A.16)}\\
&= \int_{K^p} \prod_{k=1}^{p} w_k(x_k) y(\mathbf{x})\phi_r^i(x_i)d\mathbf{x} && \text{(A.17)}\\
&\approx \frac{1}{N}\sum_{s=1}^{N} y(\mathbf{x}^{(s)})\phi_r^i(x_i^{(s)}) && \text{(A.18)}\\
\beta_{pq}^{ij} &= \int_0^1\int_0^1 w_i(x_i)w_j(x_j) f_i(x_i,x_j)\phi_r^i(x_i)\phi_q^j(x_j)dx_i dx_j && \text{(A.19)}\\
&= \int_{K^p} \prod_{k=1}^{p} w_k(x_k) y(\mathbf{x})\phi_r^i(x_i)\phi_q^j(x_j)d\mathbf{x}\beta_{pq}^{ij} && \text{(A.20)}\\
&\approx \frac{1}{N}\sum_{s=1}^{N} y(\mathbf{x}^{(s)})\phi_r^i(x_i^{(s)})\phi_q^j(x_j^{(s)}) && \text{(A.21)}\\
\gamma_{pqr}^{ijk} &= \int_0^1\int_0^1\int_0^1 w_i(x_i)w_j(x_j)w_k(x_k) f_i(x_i,x_j,x_k)\phi_r^i(x_i)\phi_q^j(x_j)\phi_q^k(x_k)dx_i dx_j dx_k && \text{(A.22)}\\
&= \int_{K^p} \prod_{k=1}^{p} w_k(x_k) y(\mathbf{x})\phi_r^i(x_i)\phi_q^j(x_j)\phi_{p'}^k(x_k)d\mathbf{x}\gamma_{pqr}^{ijk} && \text{(A.23)}\\
&\approx \frac{1}{N}\sum_{s=1}^{N} y(\mathbf{x}^{(s)})\phi_r^i(x_i^{(s)})\phi_q^j(x_j^{(s)})\phi_{p'}^j(x_k^{(s)}) && \text{(A.24)}
\end{aligned}$$

From these separate terms, the hierarchical correlated function expansion can be constructed:

$$\begin{aligned}
f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{p}\sum_{r=1}^{k}\alpha_r^i\phi_r^i(x_i) + \sum_{1\le i<j\le p}\sum_{r=1}^{l}\sum_{p'=1}^{l'}\beta_{rp}^{ij}\phi_r^i(x_i)\phi_{p'}^j(x_j) + \\
\sum_{1\le i<j<k\le p}\sum_{r=1}^{m}\sum_{p'=1}^{m'}\sum_{q=1}^{m''}\gamma_{rp'q}^{ij}\phi_r^i(x_i)\phi_{p'}^j(x_j)\phi_q^k(x_k) + \ldots
\end{aligned} \tag{A.25}$$

A.1.2 Adaptive Damping for Backfitting Method

The backfitting method is applied throughout this thesis for RS-HDMR [32, 137]. It was found in this work that it can diverge if sample densities are low or strong correlations are present. The method was extended by a damping factor η chosen in the interval of $[0.1, 1]$. In each iteration of backfitting, only a fraction to be approximated model output was used to estimate expansion coefficients. Thus, for the first approximation of the expansion coefficients $\eta \cdot y(x)$ was used to determine the coefficients in equation A.16-A.24. If the

solution was still diverging, the damping factor η was reduced until the method converged. In backfitting the approximation error of the meta-model can be calculated by $h(x) = y(x) - f(x)$. In all subsequent iterations the damping factor is used to estimate correction for the expansion coefficients $(\eta \cdot h(x))$. The adaptive damping was only needed for backfitting method. DMORPH and extended bases methods could be directly applied.

A.2 Sobol's g-Function

Sobol's g-function y_{Sobol} is continuous but not differentiable at $x_i = 0.5$ [220]. The samples were generated according to a uniform distribution $U(0, 1)$ using MATLAB's *rand* function. After generation of the samples and computation of the model output, all samples and the output values were scaled to the interval $[0, 1]$. The g-function is the product of separate terms g_i:

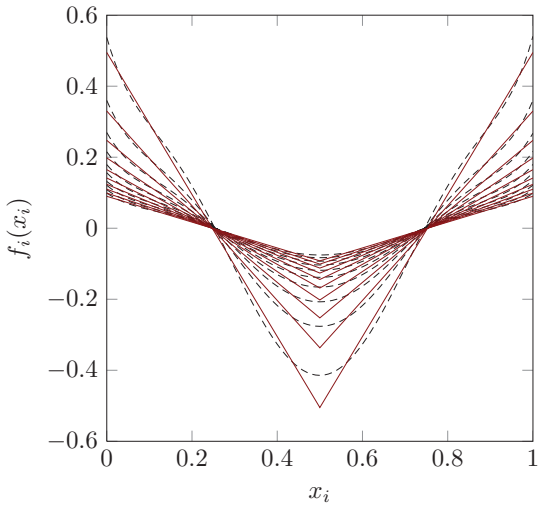
$$y_{Sobol}(x_1, \ldots, x_d) = \prod_{i=1}^{n} g_i(x_i) \tag{A.26}$$

The separate functions g_i are given by

$$g_i(x_i) = \frac{|4 \cdot x_i - 2| + a_i}{1 + a_i} \quad \text{with} \quad a_i \geq 0 \quad i = 1, 2, \ldots p \tag{A.27}$$

The coefficient a_i gives the input parameter a desired sensitivity index. A small value of a_i $(a_i = 0 - 1)$ corresponds to a high influencing factor.

The first order component function estimated based on Sobol's g-function are shown in figure A.1. The functions can be approximated by polynomials. The first order effects contribute 93.4 % and the second order cooperative effect contributes 6.4 % to the overall variance, thus third order component functions are assumed to be of no importance as expected. First and second order polynomials of the order 3-8 were used to generate the RS-HDMR model. Functions with high first order sensitivities (S_i) received also higher polynomial order (k is giving polynomial order for first order component functions). The values for the polynomials order are shown in in figure A.1. Second order component functions were built with the same polynomial order as the first order component function $(k = l, l')$.



x_i	a_i	S_i	k
1	1	0.418	8
2	2	0.186	7
3	3	0.105	7
4	4	0.067	3
5	5	0.047	3
6	6	0.034	3
7	7	0.026	3
8	8	0.021	3
9	9	0.017	3
10	10	0.014	3
$\sum$		0.934	

figure A.1: Sobol's g-function, the analytical (——) values for the first order functions are plotted including their respective first order component function (- - - -) approximation with polynomial order $k = 7$. On the right hand side, the settings for the performed experiments are shown. Second order component functions were built with the same polynomial order as the first order functions.

First the influence of sample size on the estimated component function and sensitivity indices was tested. For this the defined measures $\|f(\mathbf{x}) - y\|_\infty$ and $\|f(\mathbf{x}) - y\|_{2,n}$ were used. The separate techniques in the work-flow are tested for uniform distributed inputs.
In figure A.2 Sample size is varied for building of the meta-model within $5 \cdot 10^3$-10^5 samples (solid lines, "original samples"). The independently generated reference sample set of $5 \cdot 10^5$ samples (dashed lines) is used to re-calculate the measures $\|f(\mathbf{x}) - y\|_\infty$ and $\|f(\mathbf{x}) - y\|_{2,n}$. It is obvious, that the measures calculated based on the reference samples show less optimal measures for goodness-of-fit compared to the originally used samples for building of the RS-HDMR model. This effect vanishes with increasing sample size as the approximation quality by RS-HDMR is increasing.
The upper plot shows the measure $\|f(\mathbf{x}) - y\|_{2,n}$, which is similar for all techniques above 10^4 samples (for reference and original samples). With lower sample size, the results obtained for the separate techniques show differences. Lower order product method showed best performance on low sample size. It approximates second order component functions and their coefficients by using first order functions and is more efficient in of the used expansion coefficients. All other built meta-models possess 856 expansion coefficients for first and second order component functions, whereas lower order product method uses 465 expansion coefficients . The second order coefficients are estimated with an accuracy of roughly 20 %, nevertheless the lower order product method does not reach the accuracy for second order component functions of the other methods (see table A.1 in appendix).
D-MORPH regression, backfitting, correlation and extended basis method had best estimation quality of second order function. Using 5,000 samples large second order ($S_{ij} = 0.1$-0.01) correlation terms were adequately determined by these methods, with a relative error below 5 %. A relative error of less than 10 % for 85 % of the second order sensitivity indices for low influential sensitivity indices in the order of $S_{ij} = 0.01$-0.001 were found if 10^5 samples were used.
Using only 5,000 samples results in 23 % of the second order sensitivity indices in the respec-

tive range. Thus, increasing sample size beyond 5,000 samples can increase approximation quality of RS-HDMR expansions for sensitivity indices estimation significantly, especially for second order indices. The computational effort was low for lower-order product method (in the order of seconds). Correlation and backfitting method were more expensive (order of 1-2 minutes), but a RS-HDMR could be built for 50,000 samples within minutes. DMORPH regression and extended bases method performed similar but lead to high computation time (order of 5-10 minutes).

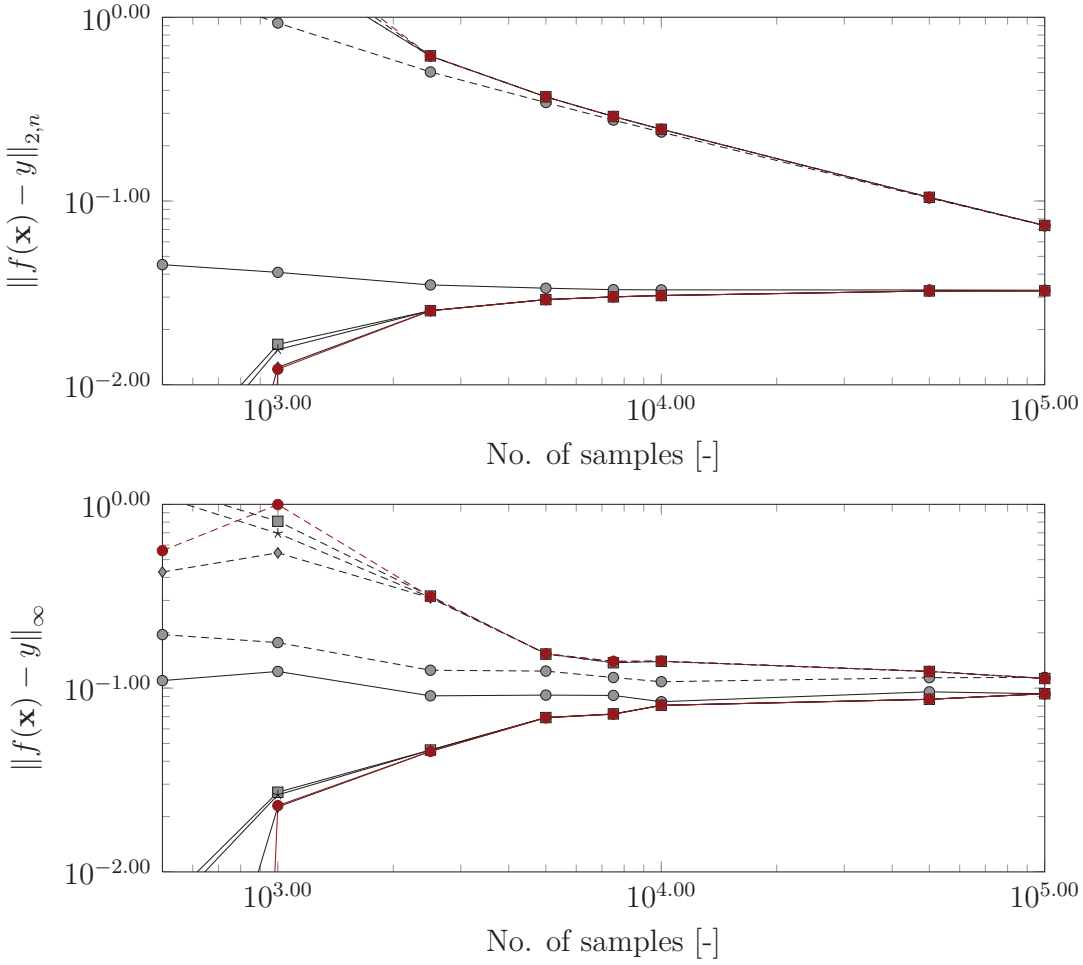


figure A.2: Comparison of RS-HDMR techniques by using Sobol's g-function. Backfitting for first order component function and lower order product for 2nd order component function: original (—●—) and reference samples (- -●- -); backfitting: original (—♦—) and reference samples (- -♦- -); correlation method: original (—■—) and reference samples (- -■- -); extended basis method: original (—⋆—) and reference samples (- -⋆- -); DMORPH regression: original (—●—) and reference samples (- -●- -)

A.3 Preliminary Test of RS-HDMR Techniques for an LC-MS/MS Measurement

One output of the ^{13}C-MFA model of *P. chrysogenum* built in chapter 7 was chosen as a test case to compare the methods for global sensitivity analysis. For the output, the separate HDMR methods were applied. The built RS-HDMR model use zeroth to second order component functions and contains 4279 coefficients for the 31 inputs using orthogonal polynomials of third order. Here, only one specific measurement of alanine, the m+0 trace is approximated by RS-HDMR. The results can be seen in figure A.3. For this model no analytical solution for the sensitivity indices exists, nevertheless the values of the largest sample size (500,000 samples) used as reference point to estimate the goodness of the estimated sensitivity indices.
In figure A.2 sample size used for building of the RS-HDMR meta-model was varied. Sample size from 10^3 to $5 \cdot 10^5$ samples was used (solid lines, "original samples"). The independently generated reference sample set of $5 \cdot 10^5$ samples (dashed lines) is used to re-calculate the measures $\|f(\mathbf{x}) - y\|_\infty$ and $\|f(\mathbf{x}) - y\|_{2,n}$ with the built RS-HDMR model. It is obvious, that the measures calculated based on the reference samples are less optimal for goodness-of-fit compared to the originally used samples for building of the RS-HDMR model. This effect vanishes with increasing sample size as the approximation quality by RS-HDMR is increasing.
The top graph in figure A.3 shows the measure $\|f(\mathbf{x}) - y\|_{2,n}$. DMORPH regression, backfitting and extended bases showed best and comparable results. Around $5 \cdot 10^4$ samples reference and original data show corresponding goodness-of-fit. Lower order product method showed low fit quality on this example, although it was efficiently applied to the simpler case of Sobol's g-function. The maximal approximation error of $\sim 2 \cdot 10^{-1}$ was higher compared to g-function.

Next, the stability of global sensitivity indices prediction was compared (data not shown), as a reference point the approximation at $5 \cdot 10^5$ samples was used of the respective method. The first order component functions are stable predicted by DMORPH regression, extended bases and backfitting. Stable results were obtained for high sensitivity indices with 5,000 samples. For backfitting method, the same was found for increased samples sizes of 50,000 samples.
However, the backfitting method showed stable predictions for the second order functions. Although DMORPH regression and extended bases will provide similar or (slightly) better estimates on lower sample size, backfitting procedure was chosen as the computational effort is lower (5-10 min for backfitting vs. 1-2 hour for DMORPH regression).

table A.1: Accuracy of first order sensitivity index calculation for Sobol's g-function. Sensitivity indices are grouped into their analytical known order of magnitude. 10 first order indices are found in the model. For each sample number the percentage of the sensitivity indices within a relative deviate of 1 %/5 %/10 % were found in this and all higher sample numbers.

Technique	Range of S_i	#Indices	Number of Samples			
			500	100	2,500	5,000
Lower product	0.1-1	3	0/67/100	0/100/100	0/100/100	0/100/100
	0.01-0.1	7	0/0/29	0/0/29	0/0/29	0/0/43
Correlation	0.1-1	3	0/0/0	0/0/0	0/100/100	0/100/100
	0.01-0.1	7	0/0/0	0/0/0	0/0/43	0/0/57
Matching pursuit	0.1-1	3	0/0/0	0/0/33	0/33/100	0/100/100
	0.01-0.1	7	0/0/0	0/0/0	0/14/43	0/43/57
Extended bases	0.1-1	3	0/0/0	0/0/0	0/100/100	0/100/100
	0.01-0.1	7	0/0/0	0/0/0	0/0/43	0/0/57
DMORPH	0.1-1	3	0/0/0	0/33/67	0/100/100	0/100/100
	0.01-0.1	7	0/0/0	0/0/43	0/0/43	0/0/57
Backfitting	0.1-1	3	0/0/67	0/33/100	0/100/100	0/100/100
	0.01-0.1	7	0/0/14	0/0/29	0/0/43	0/0/57

Technique	Range of S_i	#Indices	Number of Samples			
			7,500	10,000	50,000	100,000
Lower product	0.1-1	3	33/100/100	67/100/100	67/100/100	100/100/100
	0.01-0.1	7	0/0/57	0/14/71	0/14/100	0/14/100
Correlation	0.1-1	3	0/100/100	0/100/100	0/100/100	33/100/100
	0.01-0.1	7	0/0/57	0/0/57	0/0/100	0/0/100
Matching pursuit	0.1-1	3	0/100/100	0/100/100	0/100/100	33/100/100
	0.01-0.1	7	14/57/100	14/86/100	29/86/100	29/100/100
Extended bases	0.1-1	3	0/100/100	0/100/100	0/100/100	33/100/100
	0.01-0.1	7	0/0/57	0/0/57	0/0/100	0/0/100
DMORPH	0.1-1	3	0/100/100	0/100/100	0/100/100	33/100/100
	0.01-0.1	7	0/0/57	0/0/57	0/0/100	0/0/100
Backfitting	0.1-1	3	0/100/100	0/100/100	0/100/100	33/100/100
	0.01-0.1	7	0/0/57	0/0/57	0/0/100	0/0/100

table A.2: Accuracy of second order sensitivity index calculation for Sobol's g-function. Sensitivity indices are grouped into their analytical known order of magnitude. 45 first order indices are found in the model, here only 14 are shown (other were lower). For each sample number the percentage of the sensitivity indices within a relative deviate of 1 %/5 %/10 % were found in this and all higher sample numbers.

Method	Range of S_{ij}	Coefficients	Number of Samples			
			500	100	2,500	5,000
Lower product	0.01-0.1	1	0/0/0	0/0/100	0/0/100	0/0/100
	0.001-0.01	13	0/0/0	0/0/8	0/0/15	0/8/15
Correlation	0.01-0.1	1	0/0/0	0/0/0	0/0/100	0/100/100
	0.001-0.01	13	0/0/0	0/0/0	0/0/8	0/0/23
Matching pursuit	0.01-0.1	1	0/0/0	0/0/0	0/0/0	0/0/0
	0.001-0.01	13	0/0/0	0/0/0	0/8/8	0/8/8
Extended bases	0.01-0.1	1	0/0/0	0/0/100	0/0/100	0/100/100
	0.001-0.01	13	0/0/0	0/0/0	0/0/8	0/0/23
DMORPH	0.01-0.1	1	0/0/0	0/0/0	0/0/100	0/100/100
	0.001-0.01	13	0/0/0	0/0/0	0/8/8	0/8/23
Backfitting	0.01-0.1	1	0/0/0	0/0/0	0/0/100	0/100/100
	0.001-0.01	13	0/0/0	0/0/0	0/0/8	0/0/23

Method	Range of S_{ij}	Coefficients				
			7,500	10,000	50,000	100,000
Lower product	0.01-0.1	1	0/0/100	0/0/100	0/0/100	100/100/100
	0.001-0.01	13	0/8/23	0/8/46	0/23/69	8/46/77
Correlation	0.01-0.1	1	0/100/100	0/100/100	0/100/100	0/100/100
	0.001-0.01	13	0/0/31	0/8/31	8/15/69	8/15/85
Matching pursuit	0.01-0.1	1	0/0/100	0/100/100	0/100/100	100/100/100
	0.001-0.01	13	0/8/15	0/15/23	8/38/54	15/46/69
Extended bases	0.01-0.1	1	0/100/100	0/100/100	0/100/100	0/100/100
	0.001-0.01	13	0/0/31	0/8/31	8/15/69	8/15/85
DMORPH	0.01-0.1	1	0/100/100	0/100/100	0/100/100	0/100/100
	0.001-0.01	13	0/8/23	0/8/31	8/15/69	8/15/85
Backfitting	0.01-0.1	1	0/100/100	0/100/100	0/100/100	0/100/100
	0.001-0.01	13	0/0/31	0/8/31	8/15/69	8/15/85

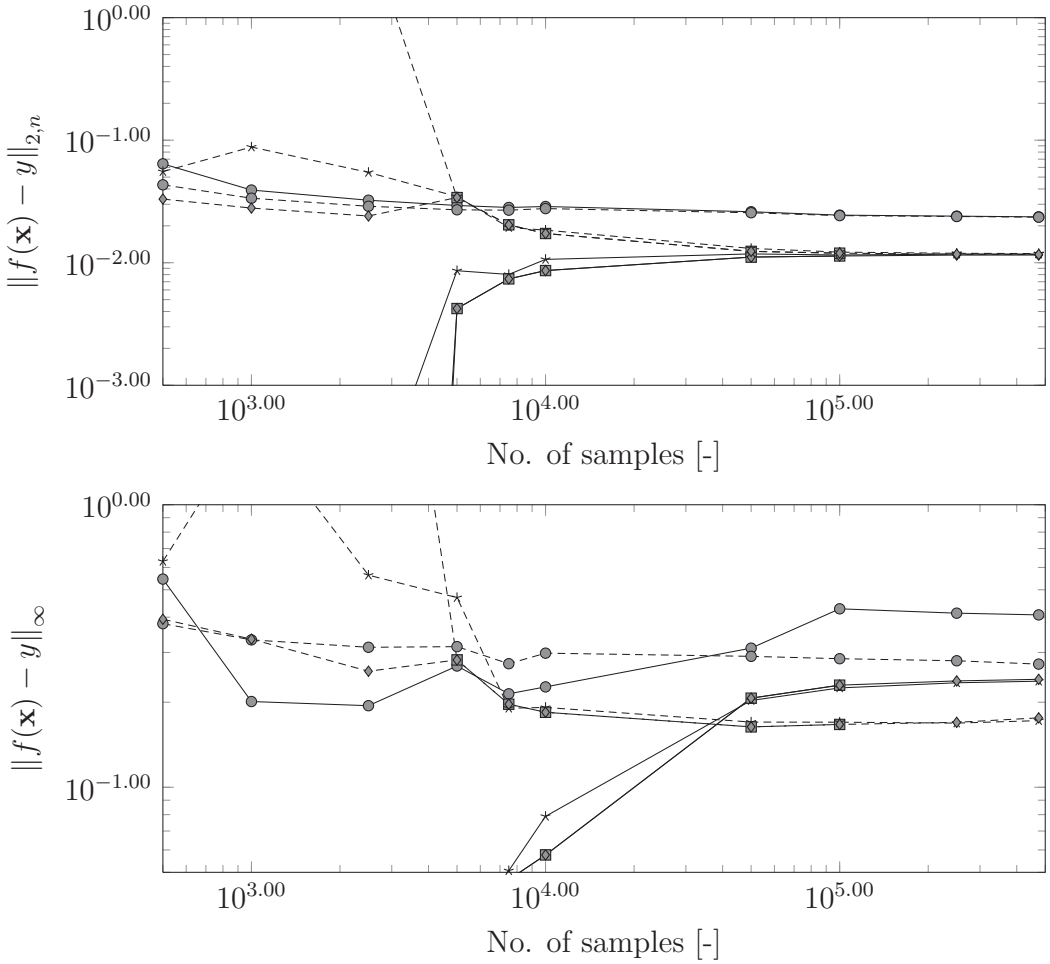


figure A.3: Comparison of RS-HDMR techniques for alanine *m+0* measurement in *P. chrysogenum* network. Backfitting for first order component function and lower order product for 2nd order component function: original (—●—) and reference samples (-●-); backfitting: original (—✶—) and reference samples (-✶-); extended basis method: original (—■—) and reference samples (-■-); DMORPH regression: original (—◆—) and reference samples (-◆-). Test case of alanine m+0 mass spectrometric measurement was taken from the *P. chrysogenum* model later used. The model output was scaled to the interval $[0, 1]$ to compare them to the results with the g-function (data in appendix).

A.4 Focused C-MFA Model for Global Sensitivity Analysis

Reaction name	Reaction
BM_GLUC:	G6P(ABCDEF) →
BM_MAN:	F6P(ABCDEF) →
BM_SER:	SER(ABC) →
BM_GLY:	GLY(AB) →
gly1:	Glc [ex] (ABCDEF) → G6P(ABCDEF)
gly2:	G6P(ABCDEF) ⇌ F6P(AFBCDE)
gly3:	F6P(ABCDEF) → FBP(ABCDEF)
gly4:	FBP(ABCDEF) ⇌ DHAP(EBF) + GAP(DAC)
gly5:	GAP(ABC) ⇌ 23PG(BCA)
gly7:	PEP(ABC) → PYR(ABC)
PPP2:	6PG(ABCDEF) → Ru5P(DBAEC) + CO2(F)
PPP5:	X5P(ABCDE) + R5P(FGHIJ) ⇌ GAP(EBD) + S7P(AFCGJHI)
Ser1:	23PG(ABC) → PHP(ABC)
Gly1:	SER(ABC) ⇌ EC1(A) + GLY(BC)
Gly2:	EC1(A) + CO2(B) ⇌ GLY(AB)
Thr2:	AcCoA(AB) + GLY(CD) ⇌ THR(ABCD)
Trans_Pyr:	PYR(ABC) → PYR [mit] (ABC)
Trans_OAA:	OAA(ABCD) ⇌ OAA [mit] (ABCD)
Ana1:	OAA(ABCD) ⇌ PEP(ABD) + CO2(C)
Ana2:	PYR(ABC) + CO2(D) ⇌ OAA(ABDC)
Thr1:	ASP(ABCD) → HSer(ACBD)
Ac3:	PYR(ABC) → CO2(C) + AcCoA(AB)
BM_AcCoA:	AcCoA(AB) →
Trans_AcCoA:	AcCoA [mit] (AB) ⇌ AcCoA(AB)
CO2 ex:	CO2(A) →
Glc 0:	Glc0(ABCDEF) → Glc [ex] (ABCDEF)
Glc U:	GlcU(ABCDEF) → Glc [ex] (ABCDEF)
Glc 1:	Glc1(ABCDEF) → Glc [ex] (ABCDEF)
PPP1:	G6P(ABCDEF) → 6PG(DACEBF)
Asp1:	OAA(ABCD) → ASP(ABCD)
BM_ASP:	ASP(ABCD) →
His2:	R5P(ABCDE) + EC1(F) → HIS(CEFDBA)
BM_HIS:	HIS(ABCDEF) →
CHOR:	E4P(ABCD) + PEP(EFG) + PEP(HIJ) → Chor(EBDHFICAGJ)
Phe1:	Chor(ABCDEFGHIJ) → PHE(GCHBDAFEI) + CO2(J)
Tyr1:	Chor(ABCDEFGHIJ) → TYR(DBHCAFGEI) + CO2(J)
BM_PHE:	PHE(ABCDEFGHI) →
BM_TYR:	TYR(ABCDEFGHI) →
BM_EC1:	EC1(A) →
PPP4:	Ru5P(ABCDE) ⇌ R5P(BDECA)
gly4b:	DHAP(ABC) ⇌ GAP(ABC)
Ser2:	PHP(ABC) → PSer(ABC)

Met1:	EC1(A) + ASP(BCDE) → MET(ABDCE)
gly6:	23PG(ABC) ⇌ PEP(ABC)
BM_MET:	MET(ABCDE) →
BIOM IN:	BIOM() → BIOMass()
BIOM ex:	BIOMass() →
Trp1:	Chor(ABCDEFGHIJ) → PYR(AEI) + Ant(CGBHFDJ)
Trp2:	Ant(ABCDEFG) + R5P(HIJKL) → Ant2(ABCDLHEFKIJG)
Trp3:	Ant2(ABCDEFGHIJKL) → CO2(L) + Ant3(ABCDEFGIHJK)
Trp4:	Ant3(ABCDEFGHIJK) → TRP(ABCDKEHGJIF)
Tca2:	OAA [mit] (ABCD) + AcCoA [mit] (EF) → Cit [mit] (AECFDB)
Ala:	PYR [mit] (ABC) → ALA [mit] (ABC)
Val1:	PYR [mit] (ABC) + PYR [mit] (DEF) → CO2 [mit] (C) + DHIV [mit] (ADEFB)
Leu:	PYR [mit] (ABC) + PYR [mit] (DEF) + AcCoA(GH) → LEU(DABEGH) + CO2(F) + CO2(C)
Ileu1:	PYR [mit] (ABC) + ASP(DEFG) → CO2(C) + KILE(FADBEG)
BM_VAL:	VAL(ABCDE) →
BM_ALA:	ALA [mit] (ABC) →
BM_ILE:	ILE(ABCDEF) →
BM_LEU:	LEU(ABCDEF) →
BM_LYS:	LYS(ABCDEF) →
BM_THR:	THR(ABCD) →
BM_TRP:	TRP(ABCDEFGHIJK) →
Tca1:	PYR [mit] (ABC) → AcCoA [mit] (AB) + CO2 [mit] (C)
Trans_CO2:	CO2 [mit] (A) → CO2(A)
BM_ERY:	E4P(ABCD) →
BM_GLU:	GLU [mit] (ABCDE) →
BM_PRO:	PRO [mit] (ABCDE) →
BM_ARG:	ARG(ABCDEF) →
Lys1:	AKG [mit] (ABCDE) + AcCoA(FG) → CO2(E) + AAA(BDAFCG)
BM_AAA:	AAA(ABCDEF) →
BM_ORN:	ORN [mit] (ABCDE) →
BM_GALAC:	G6P(ABCDEF) →
BM_MANNOSE:	F6P(ABCDEF) →
BM_GLUCOS:	F6P(ABCDEF) →
BM_ARAB:	Ru5P(ABCDE) →
BM_ICIT:	AKG [mit] (ABCDE) →
BM_FUM:	FUM [mit] (ABCD) →
Pen1:	POA() + IPN(ABCDEF) → PENV() + AAA(ABCDEF)
PEN ex:	PENV() →
Tca4:	ACO [mit] (ABCDEF) → AKG [mit] (CBAED) + CO2 [mit] (F)
Tca7a:	FUM [mit] (ABCD) ⇌ MAL [mit] (BADC)
Glu1:	AKG [mit] (ABCDE) → GLU [mit] (ABCDE)
Glu3:	GLU [mit] (ABCDE) → GLN [mit] (ABCDE)
Glu2:	GLU [mit] (ABCDE) → PRO [mit] (BADCE)
Glu4:	GLU [mit] (ABCDE) → ORN [mit] (BADCE)
Lys2:	AAA(ABCDEF) → LYS(ACBEDF)
Arg1:	AKG [mit] (ABCDE) + EC1(F) → ARG(BADCEF)

Thr3:	HSer(ABCD) → THR(BACD)
Val2:	DHIV [mit] (ABCDE) → KIV(ABECD)
Val3:	KIV(ABCDE) → VAL(ABCDE)
Ileu2:	KILE(ABCDEF) → ILE(ABCDEF)
Ser3:	PSer(ABC) → SER(ABC)
Cys1:	OAcSer(ABCDE) → CYS(BDE) + AcCoA(AC)
Cys2:	CYS(ABC) + CYS(DEF) → Cyst(DAEBFC)
BM_CYS:	CYS(ABC) →
Pen2:	CV() + AAA(ABCDEF) → IPN(ABCDEF)
BM_Cyst:	Cyst(ABCDEF) →
BM_GLN:	GLN [mit] (ABCDE) →
Asp2:	ASP(ABCD) → ASN(ABCD)
BM_ASN:	ASN(ABCD) →
Tca5b:	AKG [mit] (ABCDE) ⇌ SUC [mit] (ABCD) + CO2 [mit] (E)
Tca5a:	AKG [mit] (ABCDE) ⇌ SUC [mit] (BADC) + CO2 [mit] (E)
Poa IN:	POA_IN() → POA()
BM_CO2:	CO2(A) →
BM_R5P:	R5P(ABCDE) →
Pen_Syn_VAL:	VAL(ABCDE) →
Pen_Syn_CYS:	CYS(ABC) →
PPP6:	GAP(ABC) + S7P(DEFGHIJ) ⇌ F6P(BDCAHF) + E4P(JEIG)
PPP7:	E4P(ABCD) + X5P(EFGHI) ⇌ F6P(BEDCAG) + GAP(IFH)
PPP3:	Ru5P(ABCDE) ⇌ X5P(ABCDE)
BM_GAP:	GAP(ABC) →
Tca3:	Cit [mit] (ABCDEF) → ACO [mit] (ABFCDE)
Tca6a:	SUC [mit] (ABCD) ⇌ FUM [mit] (ABCD)
Tca6b:	SUC [mit] (ABCD) ⇌ FUM [mit] (BADC)
Ser4:	SER(ABC) + AcCoA(DE) → OAcSer(DAEBC)
Ana3:	PYR [mit] (ABC) + CO2 [mit] (D) ⇌ MAL [mit] (ABDC)
Tca8:	MAL [mit] (ABCD) ⇌ OAA [mit] (ABCD)
Tca7b:	FUM [mit] (ABCD) ⇌ MAL [mit] (ABCD)
Asp1b:	OAA [mit] (ABCD) → ASP(ABCD)
IPN_ex:	IPN(ABCDEF) →
Pen3:	PENV() → Byprod()
Byprod_ex:	Byprod() →
Help_Ru5P:	Ru5P(ABCDE) → Ru5P_X5P(ABCDE)
Help_X5P:	X5P(ABCDE) → Ru5P_X5P(ABCDE)
Help_Ru5P_X5P:	Ru5P_X5P(ABCDE) →
Pen0:	CV_IN() ⇌ CV()

table A.4: Free net flux values of *P. chrysogenum* model used for sampling

Flux	Interval	Flux	Interval	Flux	Interval
Ana1	$[0, 0.57]$	Asp3	$[0, 0.06]$	Pen3	$[0.32]$
Tca1	$[0.16, 1.39]$	Tca8	$[0, 1.00]$	Transporter_CO2	$[0.67, 4.60]$
Transporter_Pyr	$[0.44, 1.41]$	gly2	$[0, 0.92]$	gly7	$[1.18, 1.67]$

Appendix B

Process Data

B.1 Material and Methods

P. chrysogenum strain BCB1 was used throughout all experiments. This strain (also called P14-B1 strain) was derived from the P2 line of strains generated at Panlabs [193]. All experimental data were kindly provided by Georg Kornfeld. All experiments in wet-lab were performed by Sabine Meinert in 2012 [155]. Simulation studies of the kinetic model for chemostat and fed-batch were performed by Timm Severin at IBG-1. All network visualization in this thesis were made using OMIX® Version 1.5.90 [60].

Chemostat

Vegetative cells were prepared in a two stage cultivation in stirred bioreactors. For details see Meinert [155]. The first pre-culture was performed in a 15 L bioreactor with 12 L working volume. The medium was based on Li *et al.* and contains complex components like corn steep liquor [139]. After 48 hours the culture was transferred to the second pre-culture.
The medium in the first pre-culture contains a mixture of sucrose and glucose and corn steep liquor: $18\,g\,L^{-1}$ sucrose, $3\,g\,L^{-1}$ glucose $\cdot$ H_2O, $3.8\,g\,L^{-1}$ $CaCO_3$, $26\,g\,L^{-1}$ corn steep liquor, $0.5\,mL\,L^{-1}$ silicone oil and $0.5\,mL\,L^{-1}$ polypropylene glycol. $2.4\cdot10^9$ spores L^{-1} were used for inoculation. The first pre-culture was stirred at 150 rpm (rotations per minute) for 24 hours, afterward 250 rpm were used.
After 48 hours 13.7 % of the first seed culture were used as inoculum for the second stage seed culture in a 50 L bioreactor with 35 L working volume. The medium for the second stage contains a mixture of sucrose and lactose: $30\,g\,L^{-1}$ sucrose, $2\,g\,L^{-1}$ lactose $\cdot$ H_2O, $2.5\,g\,L^{-1}$ urea, $5\,g\,L^{-1}$ $NH_4(SO_4)_2$, $0.5\,g\,L^{-1}$ KH_2PO_4, $4\,g\,L^{-1}$ $CaCl_2$, $10\,g\,L^{-1}$ MOPS (3-morpholinopropane-1-sulfonic acid), $6mL\,L^{-1}$ trace element solution, $1mL\,L^{-1}$ PPG, $1\,mL\,L^{-1}$ silicone oil. The trace element solution contains $0.4\,g\,L^{-1}$ $FeSO_4\cdot 7\,H_2O$, $0.2\,g\,L^{-1}$ $CuSO_4\cdot 5\,H_2O$, $0.8\,g\,L^{-1}$ $ZnSO_4\cdot 7\,H_2O$, $0.4\,g\,L^{-1}$ $MnSO_4\cdot H_2O$, adjusted to pH 2 with H_2SO_4. KOH was used to adjust the medium to pH 5.3. The second seed culture was controlled at 25 °C, 1 vvm aeration rate (1 bar back pressure) and 275 rpm stirrer speed. Dissolved oxygen and pH were not controlled.
The main culture was conducted in a 1.7 L bioreactor with 1.0 L working volume. The volume of cultivation medium was controlled by weight. Reactor, feed medium and removed cultivation broth's weight was measured. For inoculation 15 % of the second seed culture was used. The cultivation was controlled to 25 °C, pH 6.4 $\pm$ 0.1, with 5 % NaOH and 5 % H_2SO_4.

Aeration rate of 1 vvm and stirrer speed of 600 rpm (6-bladed Rushton turbines) were used. The medium used in the main culture is, with adaption in trace element solution, taken from Nasution et al. [169]. 8.25 $g\,L^{-1}$ glucose · H_2O, 5.0 $g\,L^{-1}$ $NH_4(SO_4)_2$, 0.5 $g\,L^{-1}$ $MgSO_4$ · $7\,H_2O$, 1.0 $g\,L^{-1}$ KH_2PO_4, 1.8 $g\,L^{-1}$ sodium phenoxyacetate, 10 $mL\,L^{-1}$ trace element solution. Start of medium feed was applied upon depletion of carbon source in batch phase by an decrease of carbon dioxide in the off-gas. Carbon dioxide and oxygen were measured in the off-gas. Biomass, substrate, penicillin and side products concentrations were measured in samples.
Labeled glucose was added to the medium (instead of unlabeled glucose), U-^{13}C-D-glucose 99% and 1-^{13}C-D-glucose 98-99% were used (Cambridge Isotope Laboratories, Inc., Andover, USA). Samples were taken from the bioreactor throughout cultivations with a fast sampling device and quenched in 60 % cold methanol. The samples were analyzed by LC-ESI-MS/MS and HPLC.

Fed-Batch

Vegetative cells were prepared in a two stage cultivation in stirred bioreactors. For details see Meinert, 2012[155]. The first pre-culture was conducted according to chemostat experiments, see above.

The fed batch process is based on Jüsten *et al.* [114]. The feed medium contains two trace element solutions: 3.5 $g\,L^{-1}$ $(NH_4)_2SO_4$, 0.8 $g\,L^{-1}$ KH_2PO_4, 0.01 $g\,L^{-1}$ $FeCl_3$ · $6H_2O$, 0.05 $g\,L^{-1}$ $MgSO_4$ · $7\,H_2O$, 0.25 $g\,L^{-1}$ KCl, 2.5 $mL\,L^{-1}$ trace element solution I, 2.5 $mL\,L^{-1}$ trace element solution II, 0.025 $mL\,L^{-1}$ polypropylenglycol. After sterilization, the pH of the medium was 4.2-5.8. The trace element solution I contains 0.4 $g\,L^{-1}$ $FeSO_4$ · $7\,H_2O$, 0.2 $g\,L^{-1}$ $CuSO_4$ · $5\,H_2O$, 0.8 $g\,L^{-1}$ $ZnSO_4$ · $7\,H_2O$, 0.4 $g\,L^{-1}$ $MnSO_4$ · H_2O. Trace element solution II contains 13 $g\,L^{-1}$ $CaCl_2$ · H_2O. The main culture was conducted in a 150 L bioreactor with a starting volume of 100 L. The working weight was controlled at 120 kg, adjusted by automatic withdrawal of culture media. The medium was inoculated with 8 % seed culture and grown at 25 °C, pH 6.5 ± 0.1 (15 % NaOH, 20 % H_2SO_4), 0.5 vvm aeration rate, 190-1000 rpm (6-bladed Rushton turbines) to maintain DO between 40 % and 60 %. Overall cultivation lasted 156 hours. Separate feeding solutions for glucose, ammonium sulfate, sodium phenoxyacetate, sodium hydroxide, and polypropylenglycol were used. The separate feeds were composed of 800 $g\,L^{-1}$ glucose · H_2O (feed rate 0 h: 0.75 $L\,h^{-1}$, 16 h: 0.70 $L\,h^{-1}$, 48 h: 0.66 $L\,h^{-1}$) water (feed rate 0 h: 0.181 $L\,h^{-1}$, 56 h: 0.36L h^{-1}), 100 $g\,L^{-1}$ $NH_4(SO_4)_2$ (feed rate 12 h: 0.161 $L\,h^{-1}$, 50 h: 0.131L h^{-1}, 70 h: 0.101 $L\,h^{-1}$, 120 h: 0.081 $L\,h^{-1}$), 40 $g\,L^{-1}$ sodium phenoxyacetate (Feed rate 0 h: 0.121 $L\,h^{-1}$, 56 h: 0.161 $L\,h^{-1}$), 15 % NaOH, 20 % H_2SO_4. Foam was suppressed by addition of polypropylenglycol. The fed-batch was started after 58 h of cultivation using 8 % inoculum of the first pre-culture.

The sensor bioreactor (1.7 L steel reactor) was placed on a weight, as well as the respective feed medium. It is constructed to be geometrical similar to the large (master) bioreactor. The same feeds were added to the small bioreactor. A fast sampling device was included to withdraw samples at well mixed positions. Both reactors were connected by a pipe to transfer broth for a labeling experiment from the large to the small (slave/sensor) bioreactor. After 56-58 hours a transfer of culture broth from master to sensor bioreactor was established. The first 5-10 L were discarded, and afterwards 930 g were than filled into the small reactor.

The small reactor was controlled prior to broth transfer and during cultivation to 25 °C, pH 6.5 ± 0.1 (15 % NaOH, 20 % H_2SO_4), 0.5 vvm aeration rate, 850 rpm to maintain DO between 30 % and 40 %. Feeding solutions in the small bioreactor were: 165 g L^{-1} glucose · H_2O, 75 g L^{-1} $NH_4(SO_4)_2$, 40 g L^{-1} sodium phenoxyacetate, NaOH, polypropylenglycol. The feed rates were adjusted to result in the same rate of substrate addition per broth weight as in the large bioreactor. Water feeding was not required in the slave bioreactor as the glucose feed was less concentrated. H_2SO_4 was only used in the batch phase and, thus, was not required for the sensor bioreactor.

The weight of the sensor bioreactor was changing, due to sampling. Between 30-60 g cultivation broth was withdrawn per sample. The sampling time points for labeling measurements are given in table B.1.

table B.1: Sampling time points for labeling measurements. Time point $t = 0$ is given by starting the labeled feed.

	Sampling time points [h]
Fed-batch 1	0.58; 5.50; 22.67; 29.50; 46.58; 57.91; 70.67; 77.67; 94.67
Fed-batch 2	0.50; 5.50; 17.00; 22.50; 41.00; 46.50; 65.00; 70.50; 89.00; 94.50
Chemostat 1	0.00; 8.00; 30.00; 50.00; 80.00; 95.00; 104.00
Chemostat 2	0.00; 8.00; 30.00; 50.00; 80.00; 95.00; 104.00
Chemostat 3	0.00; 6.00; 23.00; 32.00; 47.00; 56.00; 71.00; 90.00; 95.00; 100.00

Analytics

Quantification of penicillin V and sodium phenoxyacetate were performed by reversed phase chromatography and detected by UV detector. Glucose was analyzed by Hitachi 912 analyzer, enzymatic kits by Roche or Boehringer. LC-ESI-MS/MS was used for determination of relative labeling pattern in free amino acids, respective pathway and central carbon metabolites. It was conducted at IBG-1 at FZ Jülich by Katja Schmitz. The quantification of the metabolites of central carbon metabolism was conducted according to Luo *et al.* [143] via coupling of an UPLC (Jasco-X-LC) to a mass spectrometer (Applied Biosystems/API 4 000). The used column was Synergy Hydro (Phenomex, 1502 mm, 5 µm); an ion par reagent (10 mmol tributylamine and 15 mmol acetic acid, eluent A) and methanol were used (eluent B). 10 µL injection volume with a flow rate of 0.25 mL min^{-1} were used. Measurement specifications were described by van Ooyen *et al.* [244].

Biomass Composition

The culture broth was filtered through a filter and washed with 0.9 % NaCl and de-ionized water. The biomass was shock frozen with liquid nitrogen and lyophilized for 5 days. The cell dry weight, residual water, elementary composition, carbohydrate, protein content, sugar alcohols, organic acids, inorganic acids, RNA, and fatty acids were analyzed, see Meinert, 2012 [155].

Model and Parameter Estimation

Solving the set of differential equations 4.1-4.27 was performed in Dymola version 2013 FD01. Therein the object oriented modeling language Modelica ® version 3.2 was used. This provided the opportunity to build a general library to model all kinds of bio-processes (batch, fed-batch, chemostat). Several modules (representative for feed, reactor, cellular processes and respiration) were built which can be connected to describe a bio-process.
The generated Dymola models were converted into binary files, which allowed to call the model from external software, like MATLAB. From an automatically generated wrapper by Dymola, direct usage of the model in MATLAB is accomplished. For all simulations MATLAB 2013a and for parameter estimation the SSmGO toolbox was used [65]. This toolbox implements global optimization strategies in the *Scatter Search for matlab* (SSm) and local optimization routines. The global optimizer *enhanced Scatter Search (eSS)* and the local optimizer *Direct Search Method (dhc)* were used [51]. This toolbox applies global optimization strategies to search overall parameter space and uses local optimizer to accelerate convergence to optimal solutions. The number of function evaluations was limited to 1,000 and evaluations before first local solver call was set to zero *(opts.local.ni=0).* Solving the model DAE solver *Dassl* was chosen with a tolerance of 10^{-8} with 1,000 output intervals for simulation. Linear interpolation of the simulated values to measurement time points was used to calculate residual values for the measurement data.
The objective function was a weighted residual sum of squares (WRSS). Weights were estimated based on assumed 5 % standard deviation on the measurements $w_{i,std}$ and a lower boundary of $10^{-5}\,\mathrm{g\,kg^{-1}}$ for each measurement i separately. Additionally, the measurement values were re-weighted by manually specified weights $w_{i,man}$.

$$w_i = w_{i,man} \cdot w_{i,std} = w_{i,man} \frac{1}{max(s^2_{meas,i}, 10^{-5})}$$

The weights were used in WRSS, they are used as diagonal elements in matrix Σ^{-1}.

Batch phases of fed-batch and chemostat processes were not included into the model for parameter estimation. After the feed start in the experiment, the model was used to fit the measured concentration profiles. Feed profiles and weight of culture broth were calculated based on weight measurements of feeds, samples and bioreactor.

For the chemostat process, measured side product concentrations of penicillamine, penicilloic acid and penilloic acid were added to the penicillin V pool.
Mean and standard deviations for the net production rates were calculated from the estimated rates of the model. Simulation results of chemostats were used from 50 hours till end of cultivation and of fed-batch from 5-95 hours to calculate the mean and standard deviation.

B.2 Model Parameters and Boundaries

Parameter	Fed-Batch	Chemostat	Units
μ_{max}	$0.02 - 0.15$	$0.02 - 0.15$	$g_{CDW}\, g_{CDW}^{-1}\, h^{-1}$
K_S	$1 \cdot 10^{-8} - 1 \cdot 10^{-6}$	$1 \cdot 10^{-8} - 1 \cdot 10^{-6}$	$kg\, kg^{-1}$
K_{BM}	$1 \cdot 10^{-7} - 1 \cdot 10^{-4}$	$1 \cdot 10^{-6} - 1 \cdot 10^{-4}$	$kg\, kg^{-1}$
Y_{XS}	$0.3 - 0.6$	$0.3 - 0.6$	$kg\, kg^{-1}$
$k_{PenV,max}$	$0.0005 - 0.026$	$0.0005 - 0.026$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
k_{OPC}	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$k_{IPN \rightarrow 6APA}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 2$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$k_{PenV \rightarrow 6APA,max}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$K_{M,6APA}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1$	$kg\, kg^{-1}$
$k`_{8HPA}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 2$	$kg\, kg^{-1}$
$k_{HOPOA,max}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$k`_{PIO}$	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1 \cdot 10^{-3}$	$kg\, kg^{-1}\, h^{-1}$
k_{PIA}	$1 \cdot 10^{-8} - 1 \cdot 10^{-2}$	-	$kg\, kg^{-1}\, h^{-1}$
$k_{HOPenV,1,max}$	$1 \cdot 10^{-8} - 5$	-	$kg\, kg_{CDW}^{-1}\, h^{-1}$
k_{IPN}	$1 \cdot 10^{-8} - 1$	$1 \cdot 10^{-8} - 1$	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$k_{HOPenV,2}$	$1 \cdot 10^{-8} - 1$	-	$kg\, kg_{CDW}^{-1}\, h^{-1}$
$k_{HOPenV,sink}$	$0 - 1$	-	h^{-1}
$k_{sink,max}$	$0 - 1$	-	h^{-1}
K_{OH}	$1 \cdot 10^{-8} - 5$	$1 \cdot 10^{-8} - 5$	$kg\, kg^{-1}$
k_{PenNH2}	$0 - 0.1$	-	$kg\, kg^{-1}\, h^{-1}$
K_{POA}	0.0019	0.0019	$kg\, kg^{-1}$
K_{HOPOA}	0.000415	0.000415	$kg\, kg^{-1}$
Y_{PS}	0.841	0.841	$kg\, kg^{-1}$
m_s	0.015	0.015	$kg\, kg^{-1}\, h^{-1}$

B.3 Molar Mass of Compounds

Compound	$\mathbf{M}[\mathrm{g\,mol^{-1}}]$	v_i
glucose	180.1562	0.400
POA	152.15	0.632
PenV	350.3892	0.548
BM	-	0.4281
OPC	143.141	0.503
6APA	216.261	0.444
8HPA	2601	0.416
IPN	359.41	0.468
HOPOA	168.151	0.571
HOPenV	366.388	0.525
PIO	386.4271	0.497
PIA	342.418	0.526

B.4 Chemostat

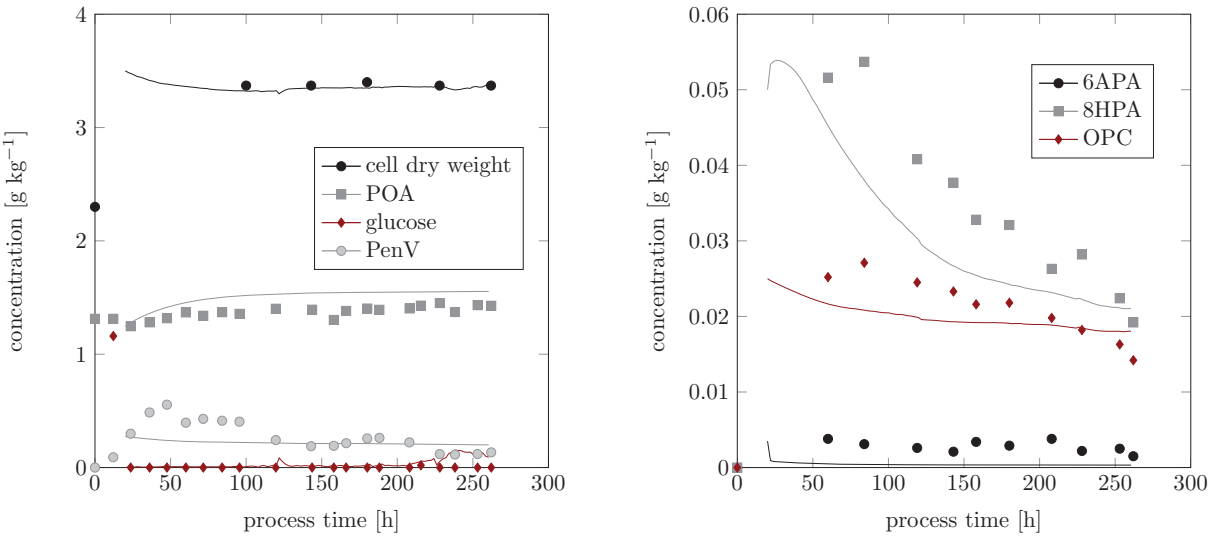


figure B.1: Best fit of the built kinetic model and process data of chemostat 1. Marker indicate concentration measurements. Lines indicate simulation results.

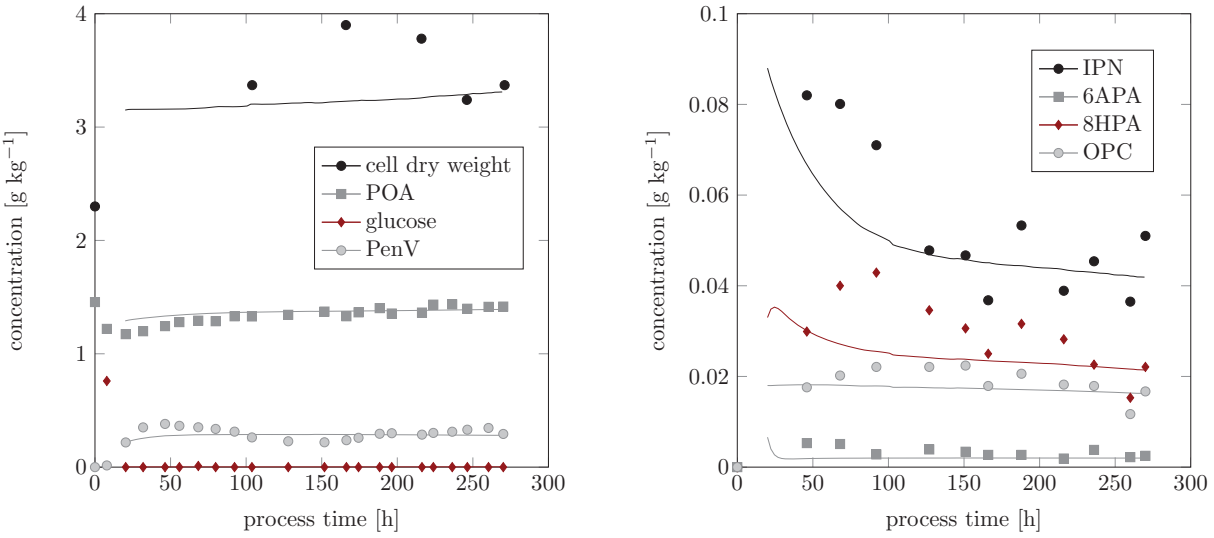


figure B.2: Best fit of the built kinetic model and process data of chemostat 2. Marker indicate concentration measurements. Lines indicate simulation results.

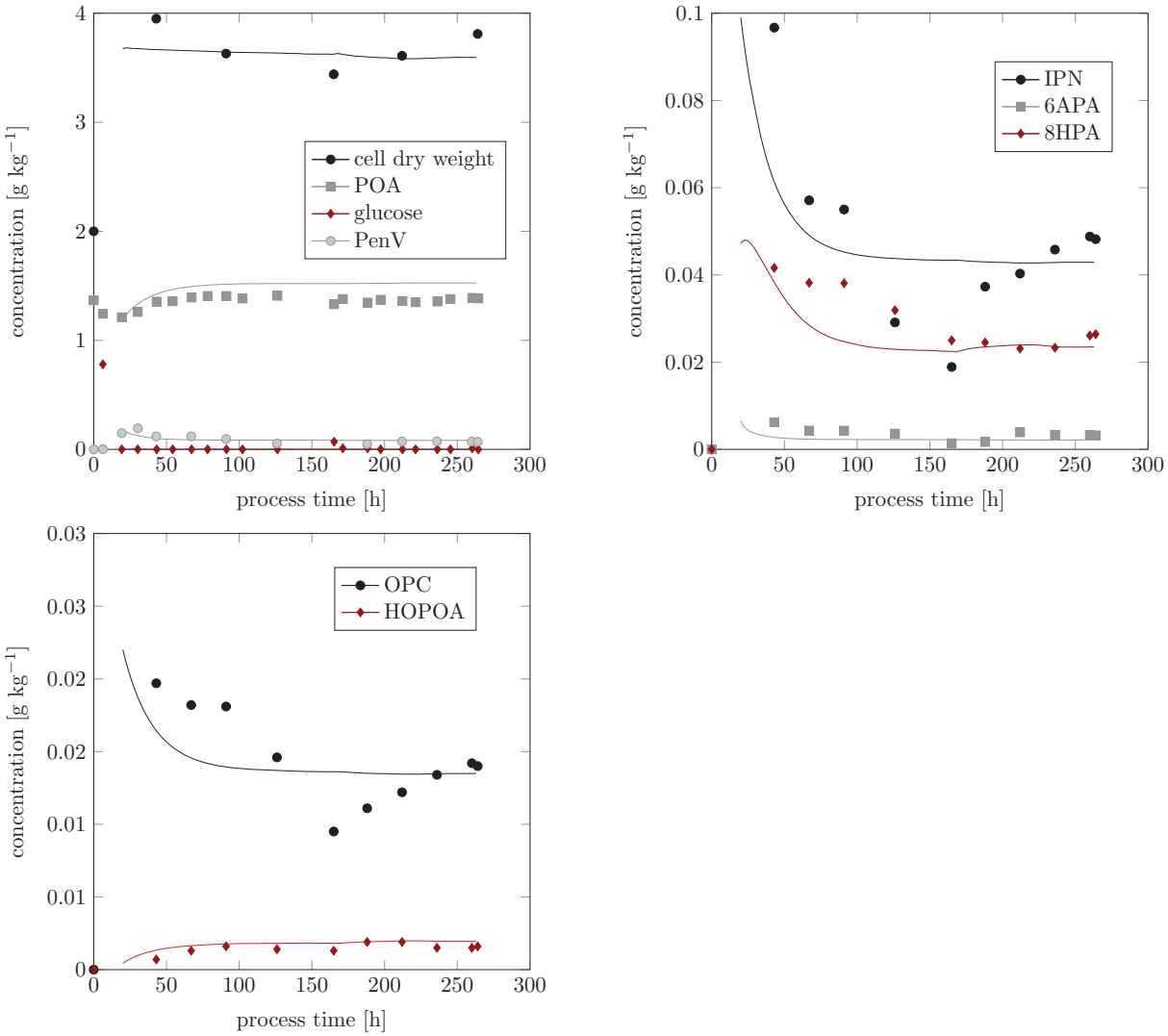


figure B.3: Best fit of the built kinetic model and process data of chemostat 3. Marker indicate concentration measurements. Lines indicate simulation results.

table B.4: Estimated reaction rates for chemostat cultivations. Rates are given in mmol g_{CDW}^{-1} h^{-1} if not stated otherwise. Not detected (n.d.) metabolites are marked.

		Chemostat 1	Chemostat 2	Chemostat 3
μ	h^{-1}	0.030±0.002	0.031±0.002	0.045 ± 0.003
$\mathbf{q}_{Glc}$	mmol g_{CDW}^{-1} h^{-1}	0.390±0.015	0.420±0.018	0.517 ± 0.028
$\mathbf{q}_{PenV}$	mmol g_{CDW}^{-1} h^{-1}	5.362 ±0.871 10^{-3}	17.547 ± 0.099 10^{-3}	4.177 ±0.205 10^{-3}
$\mathbf{q}_{PIA}$	mmol g_{broth}^{-1} h^{-1}	0.000 ±0.000 10^{-3}	0.004 ±0.000 10^{-3}	0.000 ±0.000 10^{-3}
$\mathbf{q}_{PIO}$	mmol g_{broth}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.004 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}
$\mathbf{q}_{HOPOA}$	mmol g_{CDW}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)	0.140 ±0.000 10^{-3}
$\mathbf{q}_{HOPenV,1}$	mmol g_{CDW}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	7.202 ±0.055 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}
$\mathbf{q}_{HOPenV,2}$	mmol g_{CDW}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}
$\mathbf{q}_{IPN\rightarrow 6APA}$	mmol g_{CDW}^{-1} h^{-1}	0.858 ±0.106 10^{-3}	0.000 ± 0.000 10^{-3}	0.032 ±0.002 10^{-3}
$\mathbf{q}_{PenV\rightarrow 6APA}$	mmol g_{CDW}^{-1} h^{-1}	0.000 ±0.000 10^{-3}	0.936 ± 0.007 10^{-3}	1.211 ±0.033 10^{-3}
$\mathbf{q}_{8HPA}$	mmol g_{broth}^{-1} h^{-1}	0.003 ±0.001 10^{-3}	0.003 ±0.000 10^{-3}	0.004 ± 0.000 10^{-3}
$\mathbf{q}_{OPC}$	mmol g_{CDW}^{-1} h^{-1}	1.189 ±0.001 10^{-3}	1.148 ±0.000 10^{-3}	1.186 ±0.061 10^{-3}
$\mathbf{q}_{IPN}$	mmol g_{CDW}^{-1} h^{-1}	5.108 ±0.003 10^{-3}(n.d.)	1.177 ±0.003 10^{-3}(n.d.)	1.534 ±0.079 10^{-3}
$\mathbf{q}_{PIO,sink}$	mmol g_{broth}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)
$\mathbf{q}_{PIA,sink}$	mmol g_{broth}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)
$\mathbf{q}_{HOPenV,sink}$	mmol g_{broth}^{-1} h^{-1}	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)	0.000 ±0.000 10^{-3}(n.d.)

B.5 Fed-Batch

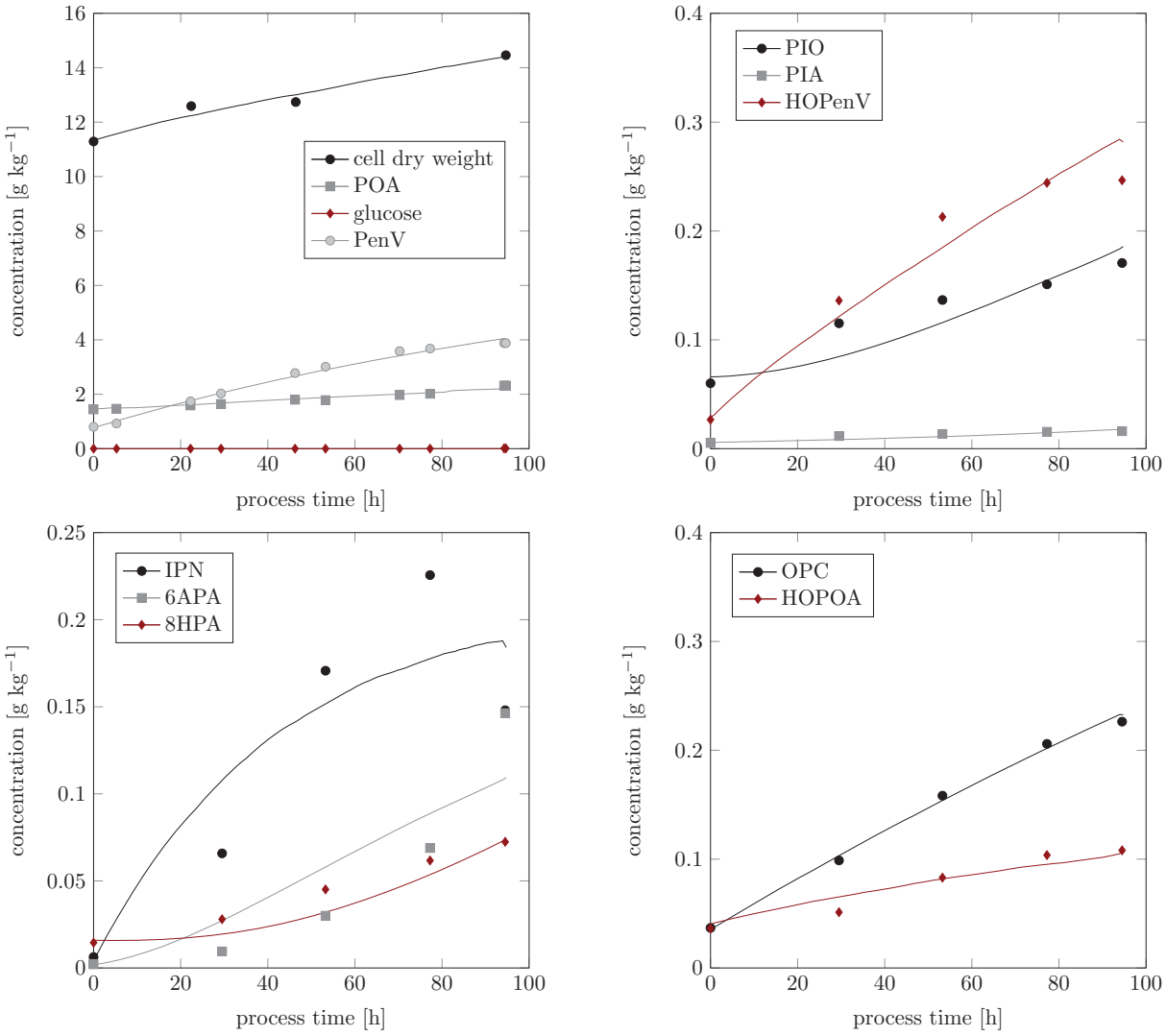


figure B.4: Best fit of the built kinetic model and process data of fed-batch 1. Marker indicate concentration measurements. Lines indicate simulation results.

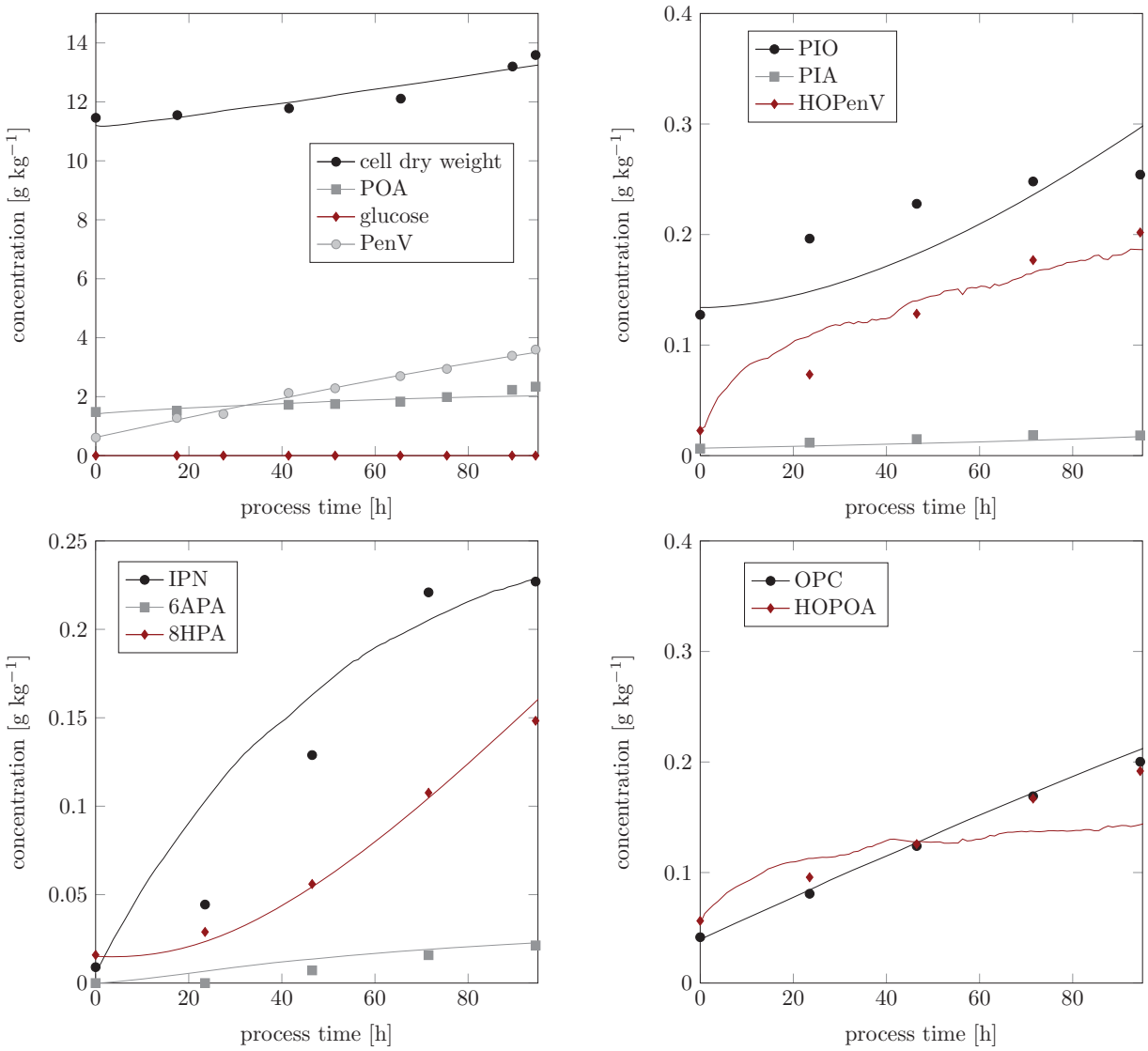


figure B.5: Best fit of the built kinetic model and process data of fed-batch 2. Marker indicate concentration measurements. Lines indicate simulation results.

table B.5: Estimated rates for fed-batch cultivations. Rates are given in mmol $g_{CDW}^{-1}\,h^{-1}$f not stated differently.

		Fed-batch 1	Fed-batch 2
μ	h^{-1}	0.008 ± 0.007	0.007±0.008
$\mathbf{q}_{Glc}$	mmol $g_{CDW}^{-1}\,h^{-1}$	0.251 ± 0.014	0.262±0.021
$\mathbf{q}_{PenV}$	mmol $g_{CDW}^{-1}\,h^{-1}$	16.569 ± 1.263 10^{-3}	12.575 ± 1.257 10^{-3}
$\mathbf{q}_{PIA}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.001± 0.000 10^{-3}	0.000 ± 0.000 10^{-3}
$\mathbf{q}_{PIO}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.007 ± 0.00010^{-3}	0.008 ± 0.003 10^{-3}
$\mathbf{q}_{HOPOA}$	mmol $g_{CDW}^{-1}\,h^{-1}$	1.537 ± 0.170 10^{-3}	8.949 ± 0.776 10^{-3}
$\mathbf{q}_{HOPenV,1}$	mmol $g_{CDW}^{-1}\,h^{-1}$	5.011 ± 1.600 10^{-3}	2.142 ± 0.749 10^{-3}
$\mathbf{q}_{HOPenV,2}$	mmol $g_{CDW}^{-1}\,h^{-1}$	1.005 ± 0.193 10^{-3}	7.915 ± 1.114 10^{-3}
$\mathbf{q}_{IPN\rightarrow 6APA}$	mmol $g_{CDW}^{-1}\,h^{-1}$	0.608 ± 0.206 10^{-3}	0.709 ± 0.255 10^{-3}
$\mathbf{q}_{PenV\rightarrow 6APA}$	mmol $g_{CDW}^{-1}\,h^{-1}$	0.126 ± 0.000 10^{-3}	0.000 ± 0.000 10^{-3}
$\mathbf{q}_{8HPA}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.003 ± 0.002 10^{-3}	0.007 ± 0.004 10^{-3}
$\mathbf{q}_{OPC}$	mmol $g_{CDW}^{-1}\,h^{-1}$	1.526 ± 0.067 10^{-3}	1.415 ± 0.100 10^{-3}
$\mathbf{q}_{IPN}$	mmol $g_{CDW}^{-1}\,h^{-1}$	1.171 ± 0.051 10^{-3}	1.420 ± 0.100 10^{-3}
$\mathbf{q}_{PIO,sink}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.001 ± 0.000 10^{-3}	0.000 ± 0.000 10^{-3}
$\mathbf{q}_{PIA,sink}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.000 ± 0.000 10^{-3}	0.000 ± 0.000 10^{-3}
$\mathbf{q}_{HOPenV,sink}$	mmol $g_{broth}^{-1}\,h^{-1}$	0.070 ± 0.027 10^{-3}	0.117 ± 0.028 10^{-3}

Appendix C

^{13}C Metabolic Flux Maps and Networks

C.1 Biomass Composition for Large Scale Network

table C.1: Biomass equation for the large scale ^{13}C-MFA model of *P. chrysogenum*.

Compound	Flux [mmol g_{CDW}^{-1}]	Compound	Flux [mmol g_{CDW}^{-1}]
Aspartate	0.182	Proline	0.118
Asparagine	0.182	Tryptophane	0.111
Glutamine	0.262	Alanine	0.279
Glutamate	0.262	Erythritol	0.016
α-Aminoadipate	0.028	Arabitol	0.010
Serine	0.264	Mannitol	0.358
Histidine	0.074	Galactose	0.297
Glycine	0.330	Glucose-Units[2]	1.071
Threonine	0.218	Mannose	0.196
Arginine	0.198	Fumarate	0.001
Tyrosine	0.116	Isocitrate	0.001
Cystein	0.008	ATP	0.037
Valine	0.288	GTP	0.027
Methionine	0.062	CTP	0.045
Phenylalanine	0.150	UTP	0.041
Isoleucine	0.195	dATP	0.002
Ornithine	0.004	dGTP	0.001
Leucine	0.292	dCTP	0.002
Lysine	0.177	dTTP	0.001
Glycerol	0.011	Ergosterol[3]	0.001
Cholin[3]	0.022	Fatty acid[1]	0.088
Inositol[3]	0.003		
Chitin	0.095		

1: Fatty acid were measured hydrolyzed, as no concentration of the separate acyl esters formed is known, their formation ratio was set free

2: Glucose was used as summed pool as biomass was hydrolysed, it be can used for formation of glycogen, trehalose or glucan. The fraction of these biomass compounds were set free for optimization

3: Values taken from Nielsen [174] and van den Berg et al. [240].

table C.2: Fatty acid composition of per mol of fatty acids for the focused and large scale ^{13}C-MFA model of *P. chrysogenum* BCB1.

Compound	Flux [$\mathrm{mol\,mol^{-1}_{fattyacid}}$]	Compound	Flux [$\mathrm{mol\,mol^{-1}_{fattyacid}}$]
Myristic acid	0.003	Oleic acid	0.023
Palmitic acid	0.155	Linoleic acid	0.738
Palmitoleic acid	0.004	Linolenic acid	0.014

C.2 Focused ^{13}C Network

Reaction name:	Reaction and atom transitions
BM_GLUC:	G6P(ABCDEF) →
BM_MAN:	F6P(ABCDEF) →
BM_SER:	SER(ABC) →
BM_GLY:	GLY(AB) →
gly1:	Glc [ex] (ABCDEF) → G6P(ABCDEF)
gly2:	G6P(ABCDEF) ⇌ F6P(AFBCDE)
gly3:	F6P(ABCDEF) → FBP(ABCDEF)
gly4:	FBP(ABCDEF) ⇌ DHAP(EBF) + GAP(DAC)
gly5:	GAP(ABC) ⇌ 23PG(BCA)
gly7:	PEP(ABC) → PYR(ABC)
PPP2:	6PG(ABCDEF) → Ru5P(DBAEC) + CO2(F)
PPP5:	X5P(ABCDE) + R5P(FGHIJ) ⇌ GAP(EBD) + S7P(AFCGJHI)
Ser1:	23PG(ABC) → PHP(ABC)
Gly1:	SER(ABC) ⇌ EC1(A) + GLY(BC)
Gly2:	EC1(A) + CO2(B) ⇌ GLY(AB)
Thr2:	AcCoA(AB) + GLY(CD) ⇌ THR(ABCD)
Transporter_Pyr:	PYR(ABC) → PYR [mit] (ABC)
Transporter_OAA:	OAA(ABCD) ⇌ OAA [mit] (ABCD)
Ana1:	OAA(ABCD) → PEP(ABD) + CO2(C)
Ana2:	PYR(ABC) + CO2(D) → OAA(ABDC)
Thr1:	ASP(ABCD) → HSer(ACBD)
Ac3:	PYR(ABC) → CO2(C) + AcCoA(AB)
BM_AcCoA:	AcCoA(AB) →
Transporter_AcCoA:	AcCoA [mit] (AB) → AcCoA(AB)
CO2_ex:	CO2(A) →
Glc_0:	Glc0(ABCDEF) → Glc [ex] (ABCDEF)
Glc_U:	GlcU(ABCDEF) → Glc [ex] (ABCDEF)
Glc_1:	Glc1(ABCDEF) → Glc [ex] (ABCDEF)
PPP1:	G6P(ABCDEF) → 6PG(DACEBF)
Asp1:	OAA(ABCD) → ASP(ABCD)
BM_ASP:	ASP(ABCD) →
His2:	R5P(ABCDE) + EC1(F) → HIS(CEFDBA)
BM_HIS:	HIS(ABCDEF) →
CHOR:	E4P(ABCD) + PEP(EFG) + PEP(HIJ) → Chor(EBDHFICAGJ)
Phe1:	Chor(ABCDEFGHIJ) → PHE(GCHBDAFEI) + CO2(J)
Tyr1:	Chor(ABCDEFGHIJ) → TYR(DBHCAFGEI) + CO2(J)
BM_PHE:	PHE(ABCDEFGHI) →
BM_TYR:	TYR(ABCDEFGHI) →
BM_EC1:	EC1(A) →
PPP4:	Ru5P(ABCDE) ⇌ R5P(BDECA)
gly4b:	DHAP(ABC) ⇌ GAP(ABC)
Ser2:	PHP(ABC) → PSer(ABC)
Met1:	EC1(A) + ASP(BCDE) → MET(ABDCE)

gly6:	23PG(ABC) ⇌ PEP(ABC)
BM_MET:	MET(ABCDE) →
BIOM_IN:	BIOM() → BIOMass()
BIOM_ex:	BIOMass() →
Trp1:	Chor(ABCDEFGHIJ) → PYR(AEI) + Ant(CGBHFDJ)
Trp2:	Ant(ABCDEFG) + R5P(HIJKL) → Ant2(ABCDLHEFKIJG)
Trp3:	Ant2(ABCDEFGHIJKL) → CO2(L) + Ant3(ABCDEFGIHJK)
Trp4:	Ant3(ABCDEFGHIJK) → TRP(ABCDKEHGJIF)
Tca2:	OAA [mit] (ABCD) + AcCoA [mit] (EF) → Cit [mit] (AECFDB)
Ala:	PYR [mit] (ABC) → ALA [mit] (ABC)
Val1:	PYR [mit] (ABC) + PYR [mit] (DEF) → CO2 [mit] (C) + DHIV [mit] (ADEFB)
Leu:	PYR [mit] (ABC) + PYR [mit] (DEF) + AcCoA(GH) → LEU(DABEGH) + CO2(F) + CO2(C)
Ileu1:	PYR [mit] (ABC) + ASP(DEFG) → CO2(C) + KILE(FADBEG)
BM_VAL:	VAL(ABCDE) →
BM_ALA:	ALA [mit] (ABC) →
BM_ILE:	ILE(ABCDEF) →
BM_LEU:	LEU(ABCDEF) →
BM_LYS:	LYS(ABCDEF) →
BM_THR:	THR(ABCD) →
BM_TRP:	TRP(ABCDEFGHIJK) →
Tca1:	PYR [mit] (ABC) → AcCoA [mit] (AB) + CO2 [mit] (C)
Transporter_CO2:	CO2 [mit] (A) → CO2(A)
BM_ERY:	E4P(ABCD) →
BM_GLU:	GLU [mit] (ABCDE) →
BM_PRO:	PRO [mit] (ABCDE) →
BM_ARG:	ARG(ABCDEF) →
Lys1:	AKG [mit] (ABCDE) + AcCoA(FG) → CO2(E) + AAA(BDAFCG)
BM_AAA:	AAA(ABCDEF) →
BM_ORN:	ORN [mit] (ABCDE) →
BM_GALAC:	G6P(ABCDEF) →
BM_MANNOSE:	F6P(ABCDEF) →
BM_GLUCOS:	F6P(ABCDEF) →
BM_ARAB:	X5P(ABCDE) →
BM_ICIT:	AKG [mit] (ABCDE) →
BM_FUM:	FUM [mit] (ABCD) →
Tca4:	ACO [mit] (ABCDEF) → AKG [mit] (CBAED) + CO2 [mit] (F)
Tca7a:	FUM [mit] (ABCD) ⇌ MAL [mit] (BADC)
Glu1:	AKG [mit] (ABCDE) → GLU [mit] (ABCDE)
Glu3:	GLU [mit] (ABCDE) → GLN [mit] (ABCDE)
Glu2:	GLU [mit] (ABCDE) → PRO [mit] (BADCE)
Glu4:	GLU [mit] (ABCDE) → ORN [mit] (BADCE)
Lys2:	AAA(ABCDEF) → LYS(ACBEDF)
Arg1:	AKG [mit] (ABCDE) + EC1(F) → ARG(BADCEF)
Thr3:	HSer(ABCD) → THR(BACD)
Val2:	DHIV [mit] (ABCDE) → KIV(ABECD)
Val3:	KIV(ABCDE) → VAL(ABCDE)

Ileu2:	KILE(ABCDEF) → ILE(ABCDEF)
Ser3:	PSer(ABC) → SER(ABC)
Cys1:	OAcSer(ABCDE) → CYS(BDE) + AcCoA(AC)
Cys2:	CYS(ABC) + CYS(DEF) → Cyst(DAEBFC)
BM_CYS:	CYS(ABC) →
BM_Cyst:	Cyst(ABCDEF) →
BM_GLN:	GLN [mit] (ABCDE) →
Asp2:	ASP(ABCD) → ASN(ABCD)
BM_ASN:	ASN(ABCD) →
Tca5b:	AKG [mit] (ABCDE) ⇌ SUC [mit] (ABCD) + CO2 [mit] (E)
Tca5a:	AKG [mit] (ABCDE) ⇌ SUC [mit] (BADC) + CO2 [mit] (E)
Poa_IN:	POA IN() → POA()
BM_CO2:	CO2(A) →
BM_R5P:	R5P(ABCDE) →
Pen_Syn_VAL:	VAL(ABCDE) →
Pen_Syn_CYS:	CYS(ABC) →
PPP6:	GAP(ABC) + S7P(DEFGHIJ) ⇌ F6P(BDCAHF) + E4P(JEIG)
PPP7:	E4P(ABCD) + X5P(EFGHI) ⇌ F6P(BEDCAG) + GAP(IFH)
PPP3:	Ru5P(ABCDE) ⇌ X5P(ABCDE)
BM_GAP:	GAP(ABC) →
Tca3:	Cit [mit] (ABCDEF) → ACO [mit] (ABFCDE)
Tca6a:	SUC [mit] (ABCD) ⇌ FUM [mit] (ABCD)
Tca6b:	SUC [mit] (ABCD) ⇌ FUM [mit] (BADC)
Ser4:	SER(ABC) + AcCoA(DE) → OAcSer(DAEBC)
Ana3:	PYR [mit] (ABC) + CO2 [mit] (D) ⇌ MAL [mit] (ABDC)
Tca8:	MAL [mit] (ABCD) ⇌ OAA [mit] (ABCD)
Tca7b:	FUM [mit] (ABCD) ⇌ MAL [mit] (ABCD)
Asp1b:	OAA [mit] (ABCD) → ASP(ABCD)
Mix_Ru5P:	Ru5P(ABCDE) → Ru5P_X5P(ABCDE)
Mix_X5P:	X5P(ABCDE) → Ru5P_X5P(ABCDE)
Ru5P_X5P_ex:	Ru5P_X5P(ABCDE) →
Pen4:	POA() + IPN(ABCDEF) → PENV() + AAA(ABCDEF)
PEN_ex:	PENV() →
bisACV_ex:	ACV(ABCDEF) →
Pen5:	PIO() → PIA()
PIA_ex:	PIA() →
Pen1:	CV IN() → CV()
OPC_ex:	OPC(ABCDEF) →
OPC_Reak:	AAA(ABCDEF) → OPC(FDBACE)
Pen3:	ACV(ABCDEF) → IPN(ABCDEF)
6APA_ex:	6APA() →
8HPA_ex:	8HPA(A) →
Pen6:	IPN(ABCDEF) → AAA(ABCDEF) + 6APA()
Pen7:	6APA() + CO2(A) → 8HPA(A)
pHOPLINV_ex:	pHOPLINV() →
PIO_ex:	PIO() →
Pen8:	PENV() → PIO()

Pen9:	PENV() → HOPenV()
HOPOA_ex:	POA() →
Pen2:	CV() + AAA(ABCDEF) → ACV(ABCDEF)
IPN_ex:	IPN(ABCDEF) →

C.3 Fed Batch

Fed-Batch 1

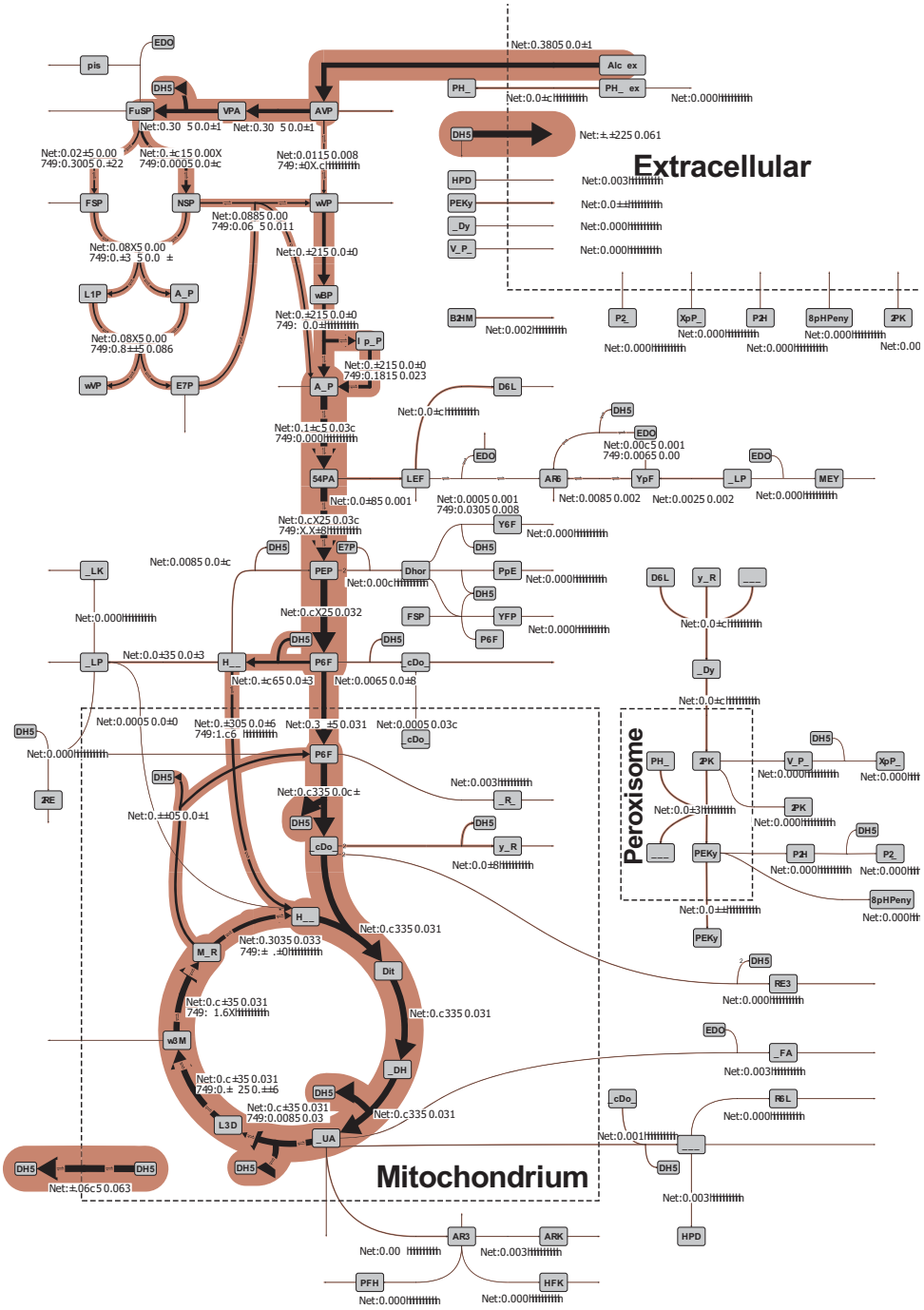


figure C.1: Flux map for fed-batch 1. Specific rates given in [mmol g_{CDW}^{-1} h^{-1}].

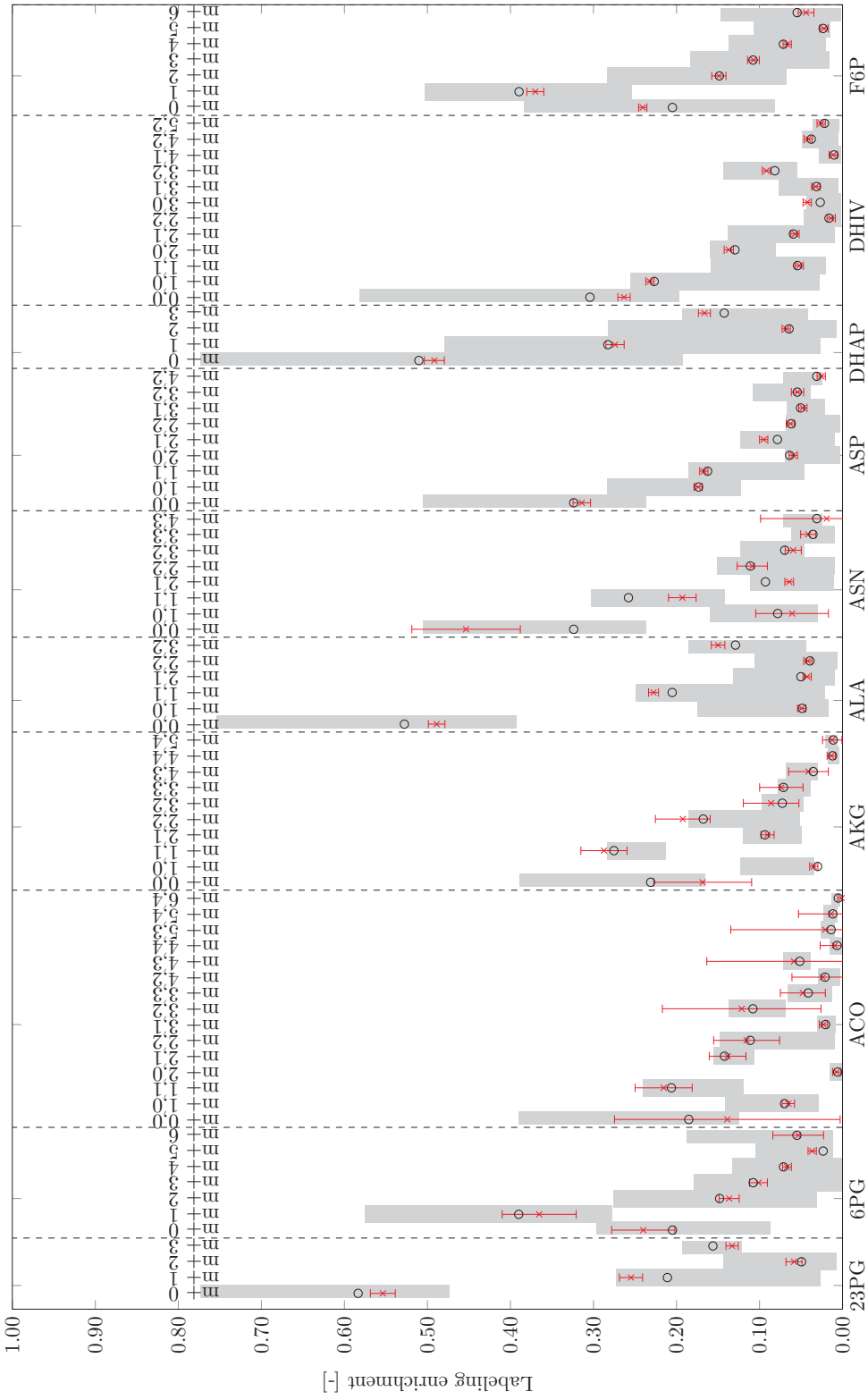


figure C.2: Fed Batch 1. Plot of minimal/maximal value for the simulated measurements (▭) for 50,000 flux sample, measurement (—×—) and simulated (○) data.

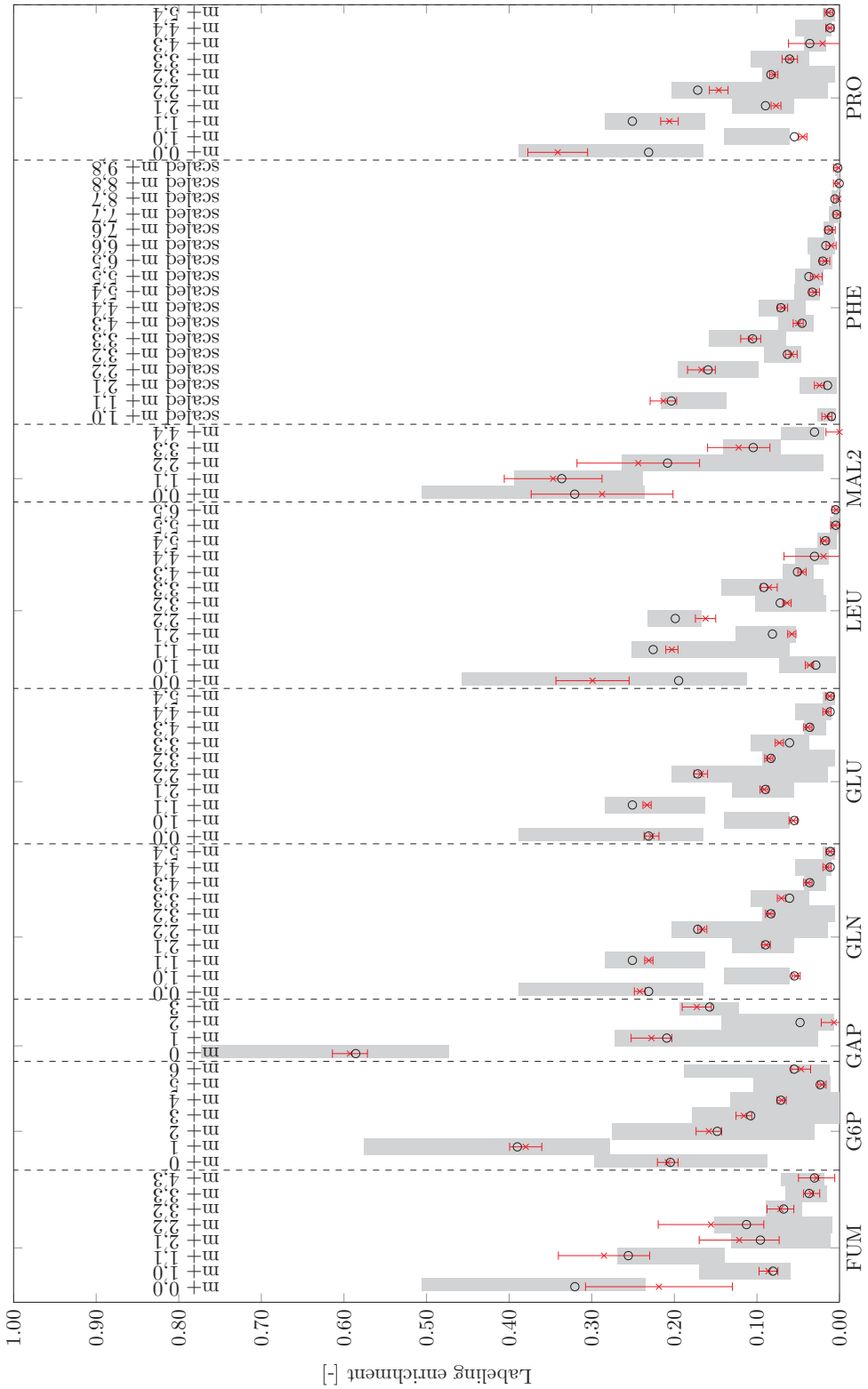


figure C.3: Fed Batch 1. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

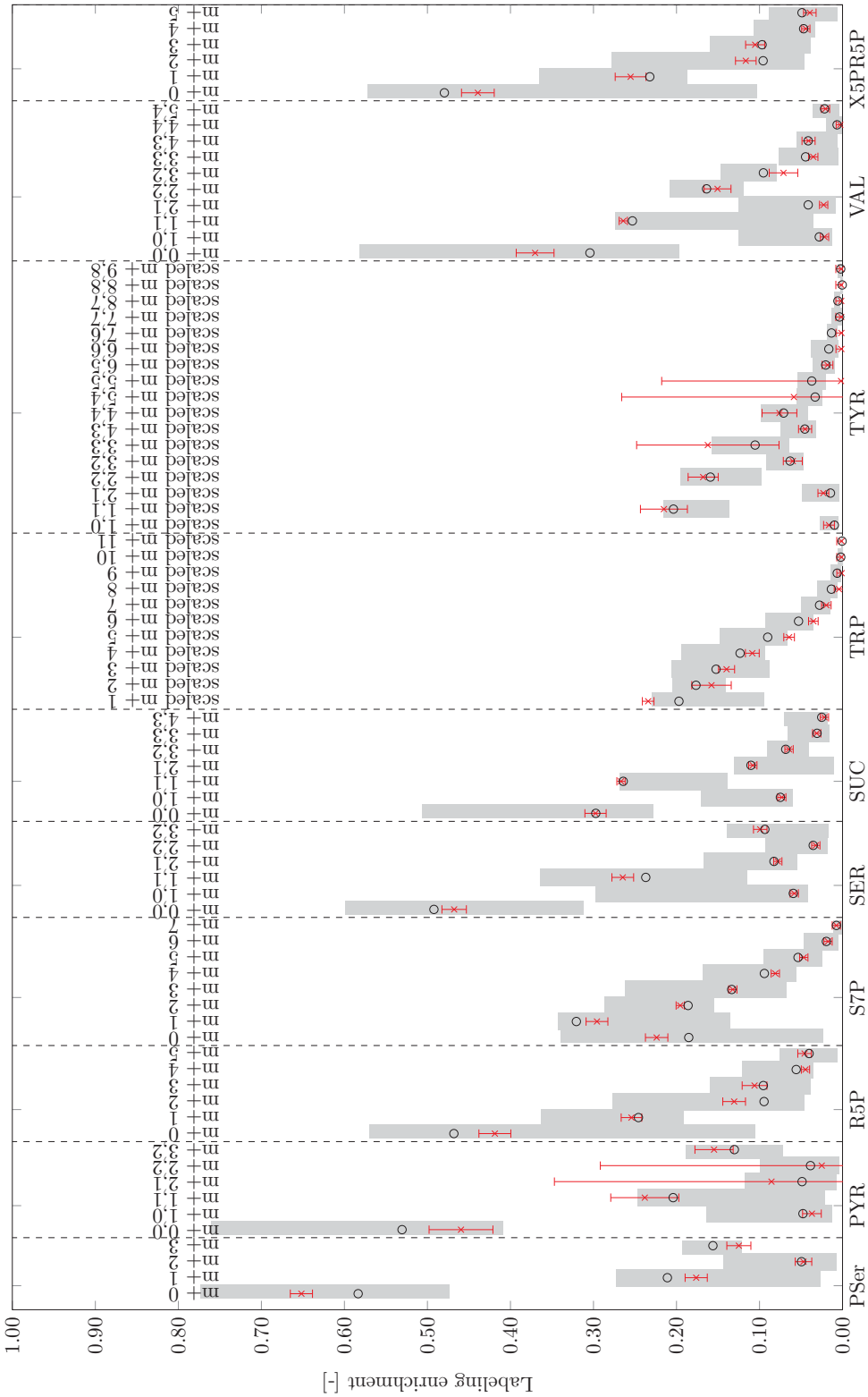


figure C.4: Fed Batch 1. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

Fed-Batch 2

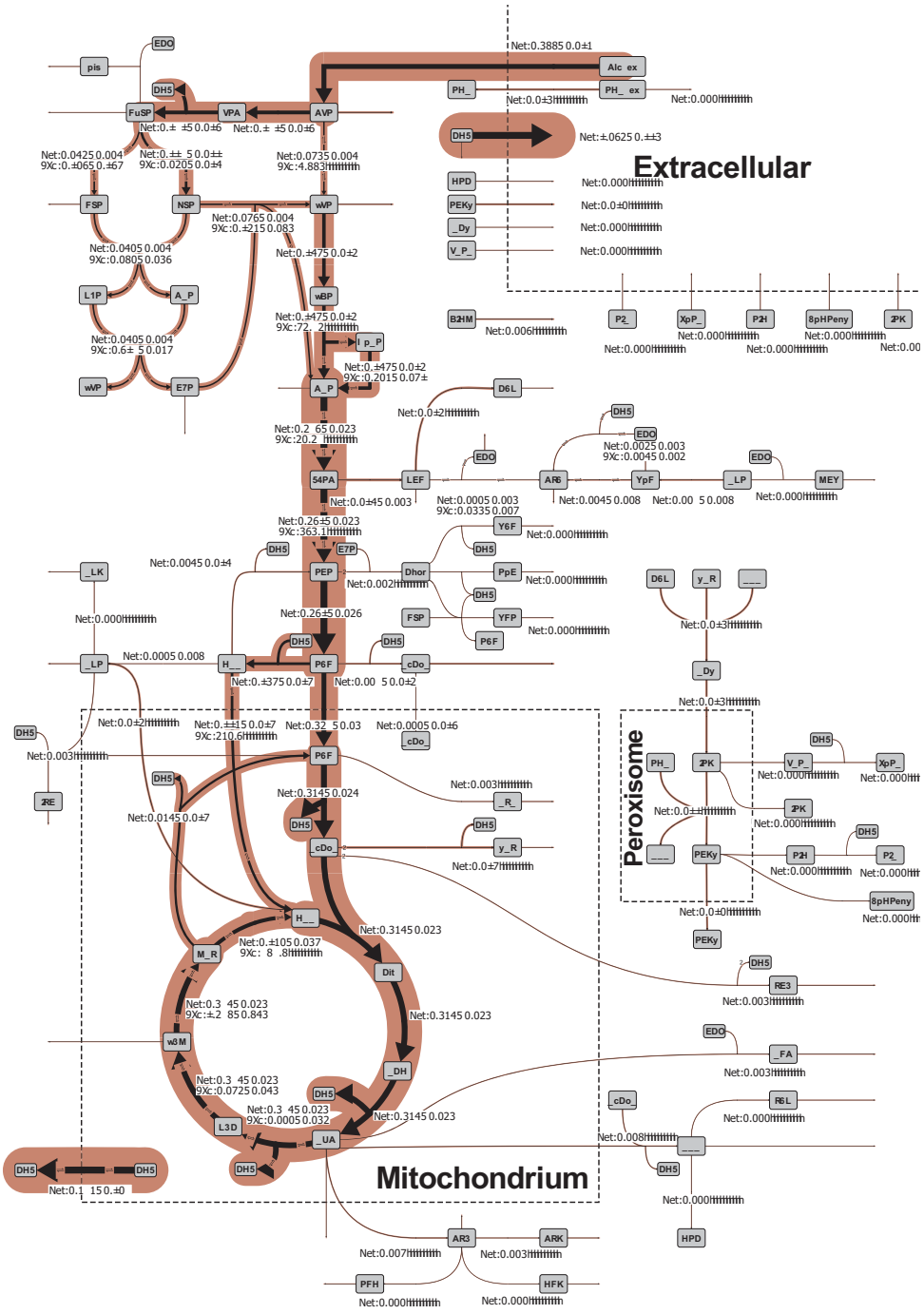


figure C.5: Flux map for fed-batch 2. Specific rates given in [mmol g^{-1}_{CDW} h^{-1}].

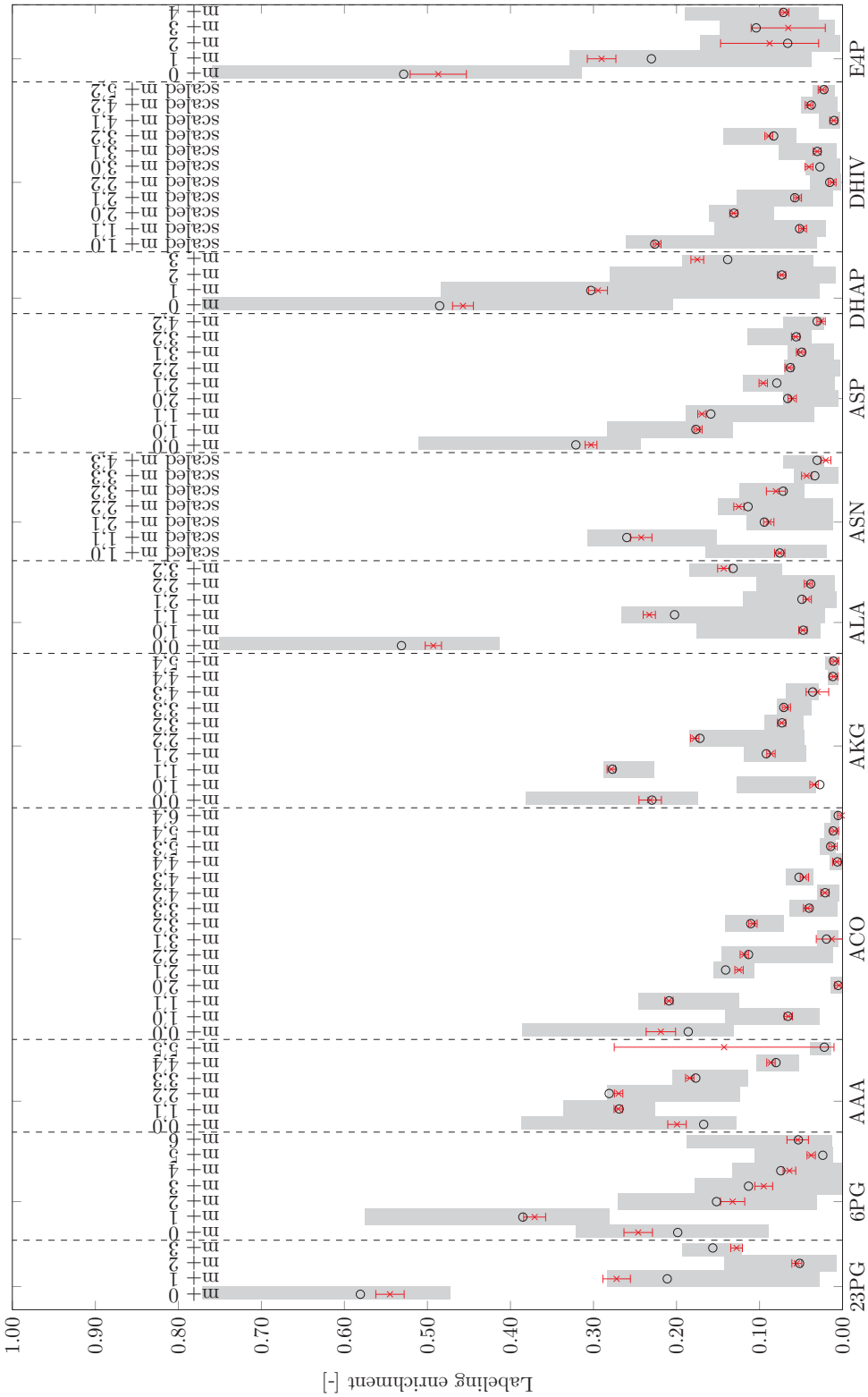


figure C.6: Fed Batch 2. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

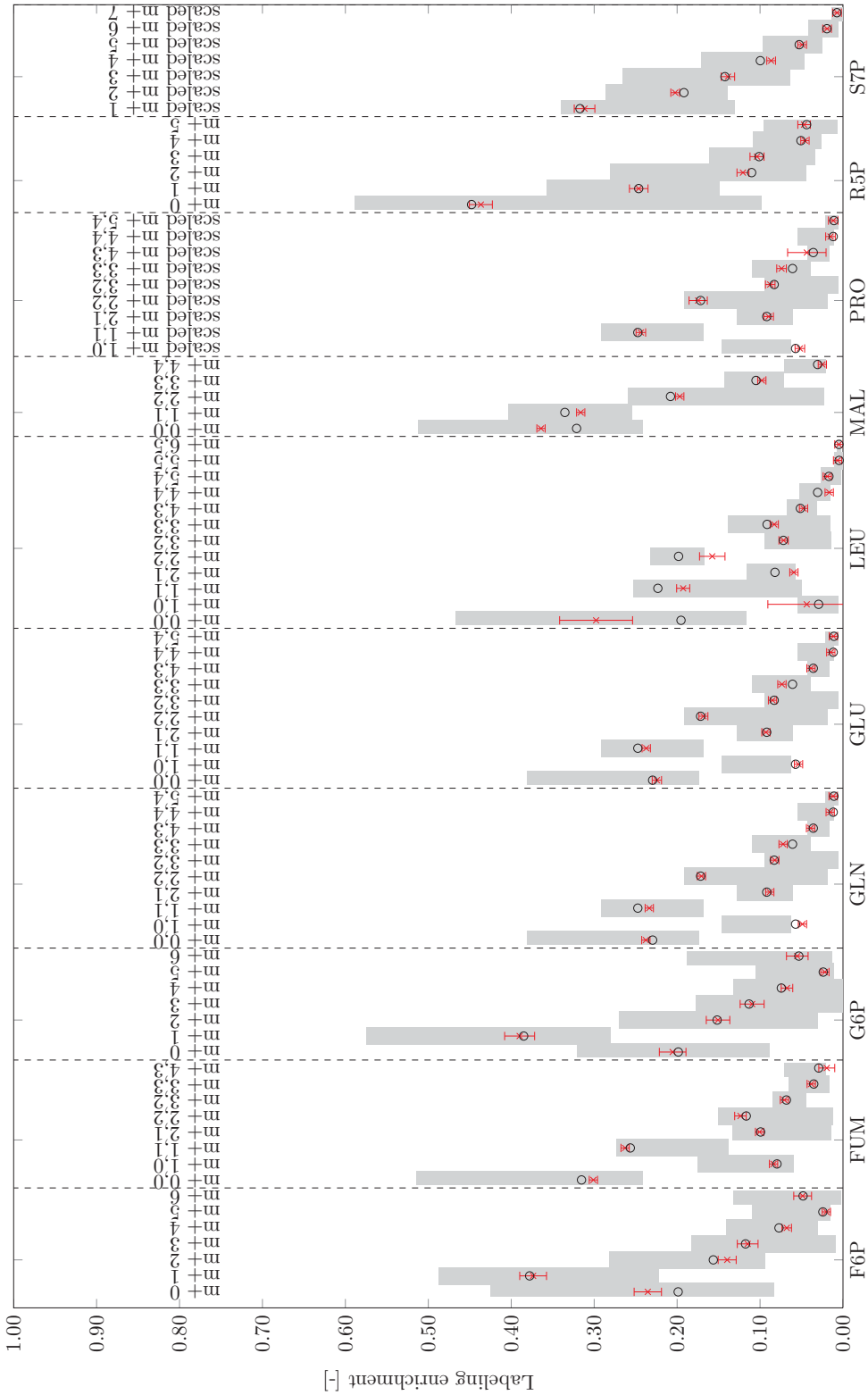


figure C.7: Fed Batch 2 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

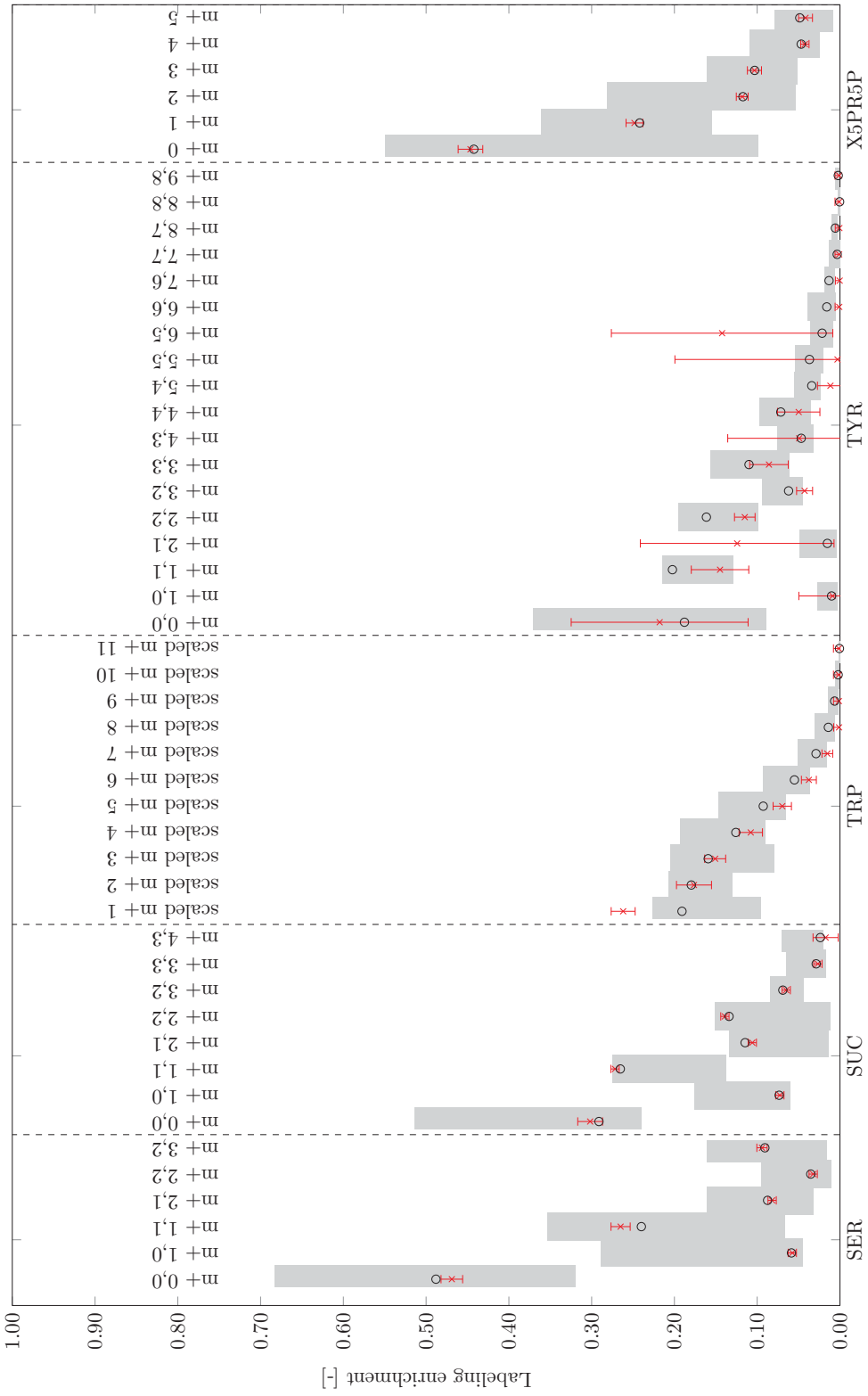


figure C.8: Fed Batch 2. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

C.4 Chemostat

Chemostat 1: 0.03 h^{-1}

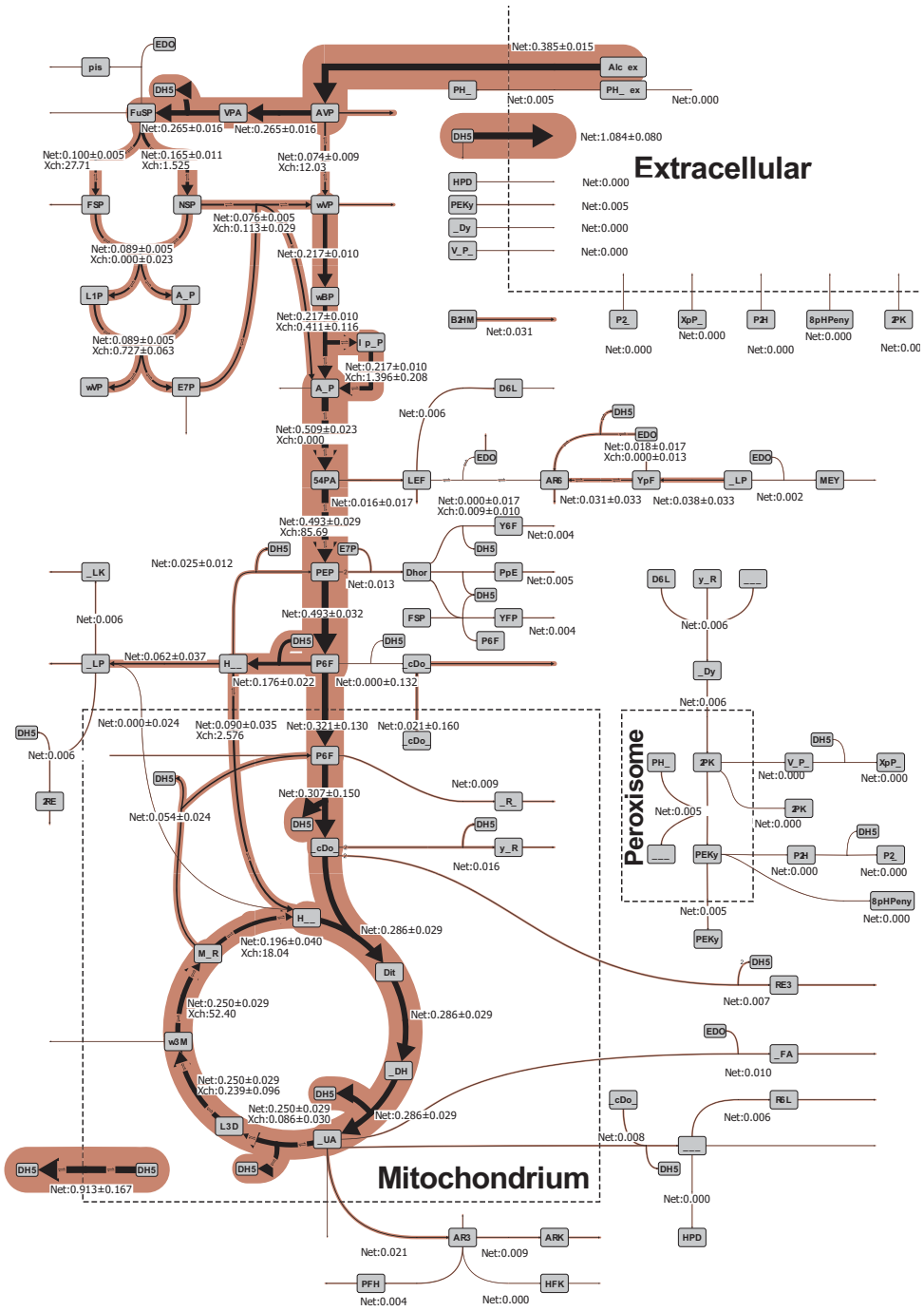


figure C.9: Flux map for chemostat 1 with a growth rate of 0.03 h^{-1}. Specific rates given in [mmol g_{CDW}^{-1} h^{-1}].

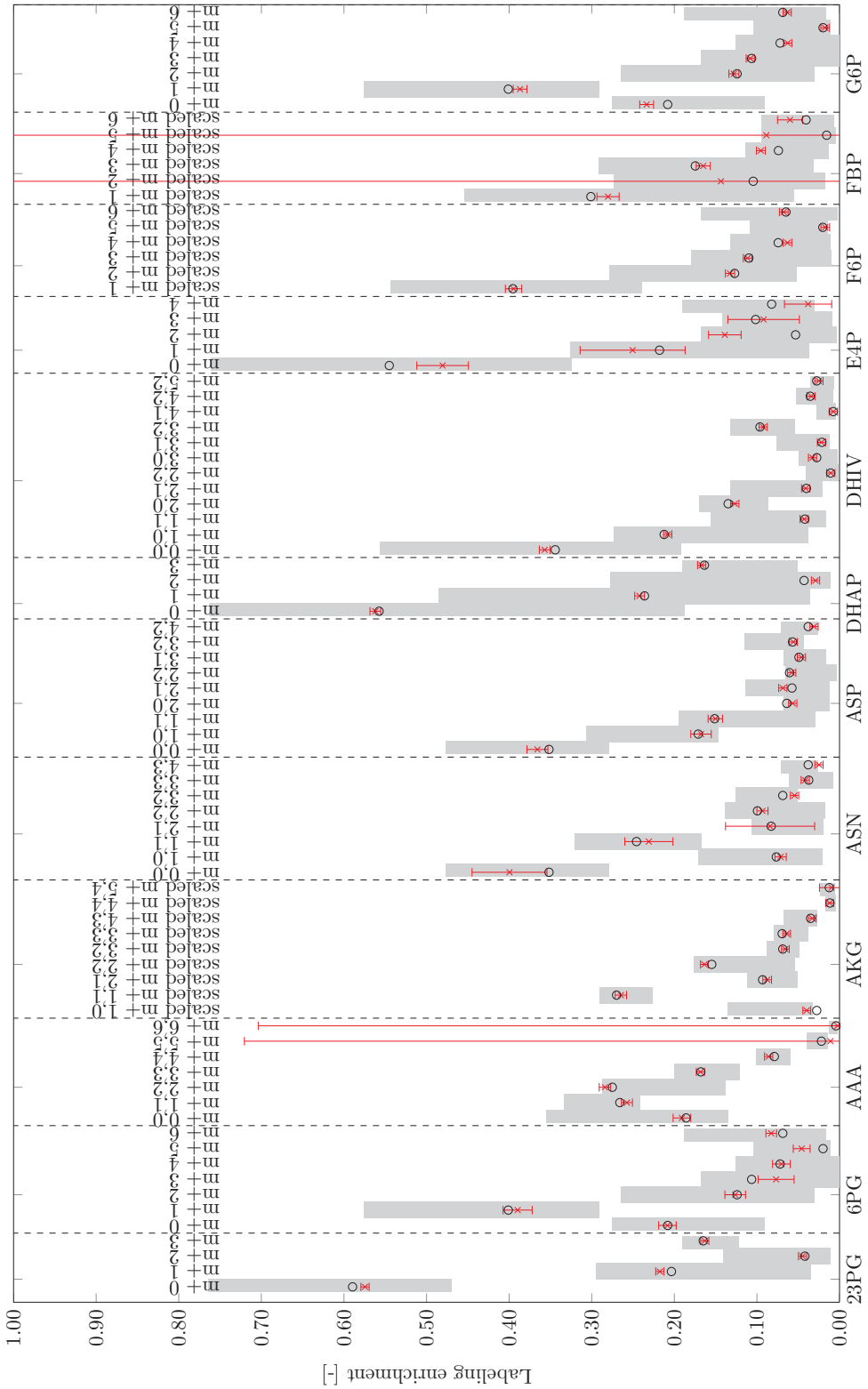


figure C.10: Chemostat 1. Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

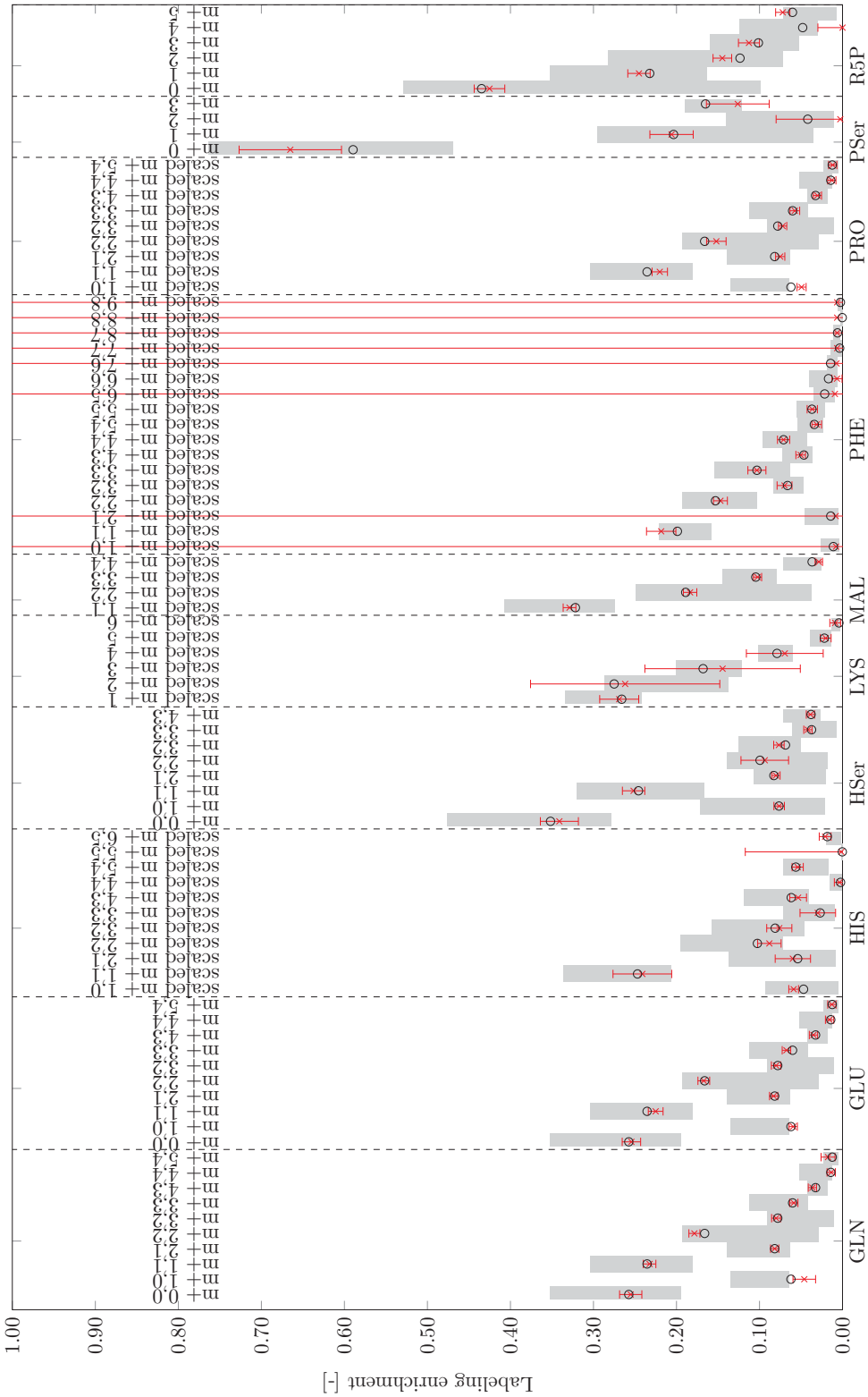


figure C.11: Chemostat 1 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

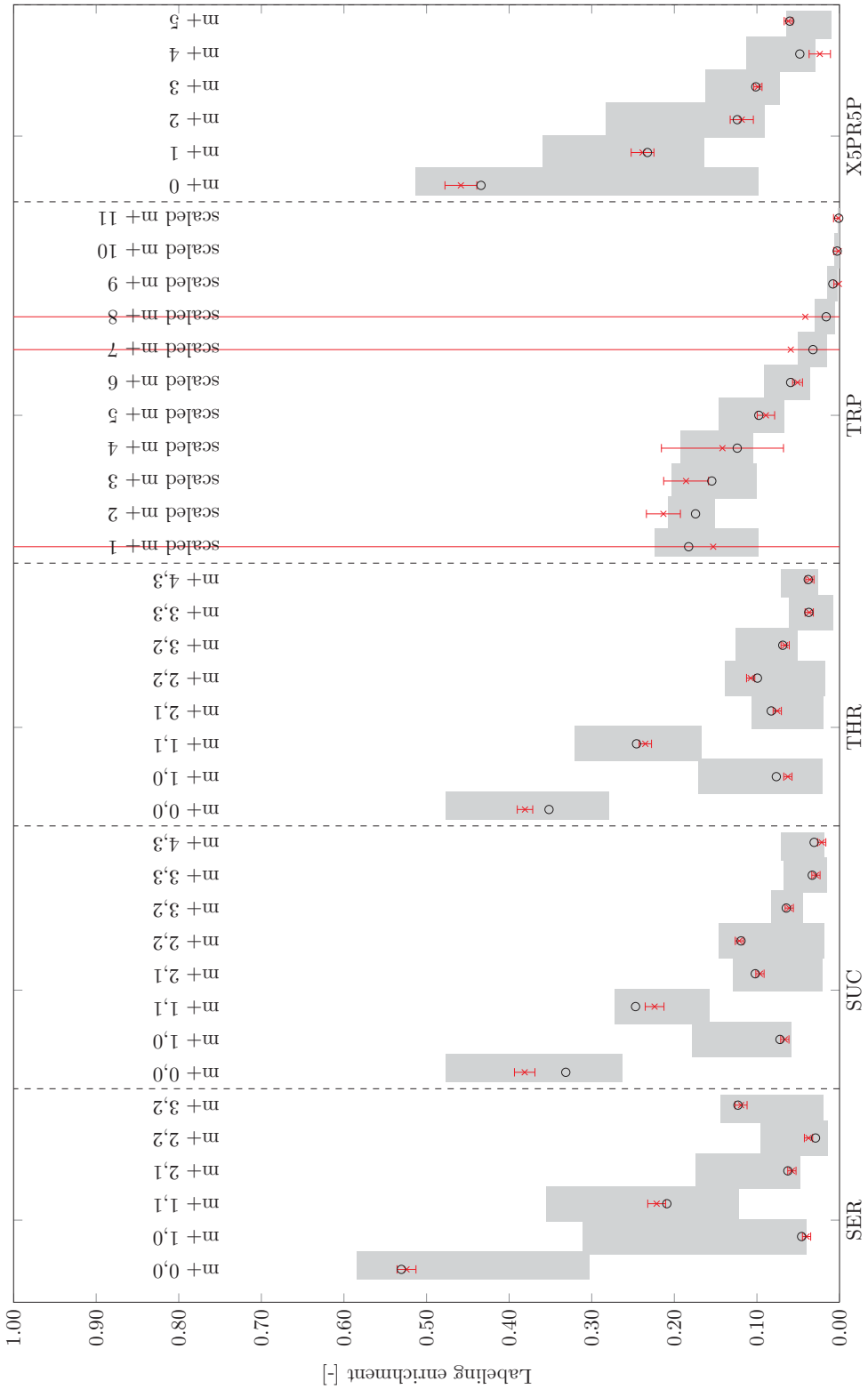


figure C.12: Chemostat 1 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

Chemostat 2: 0.03 h^{-1}

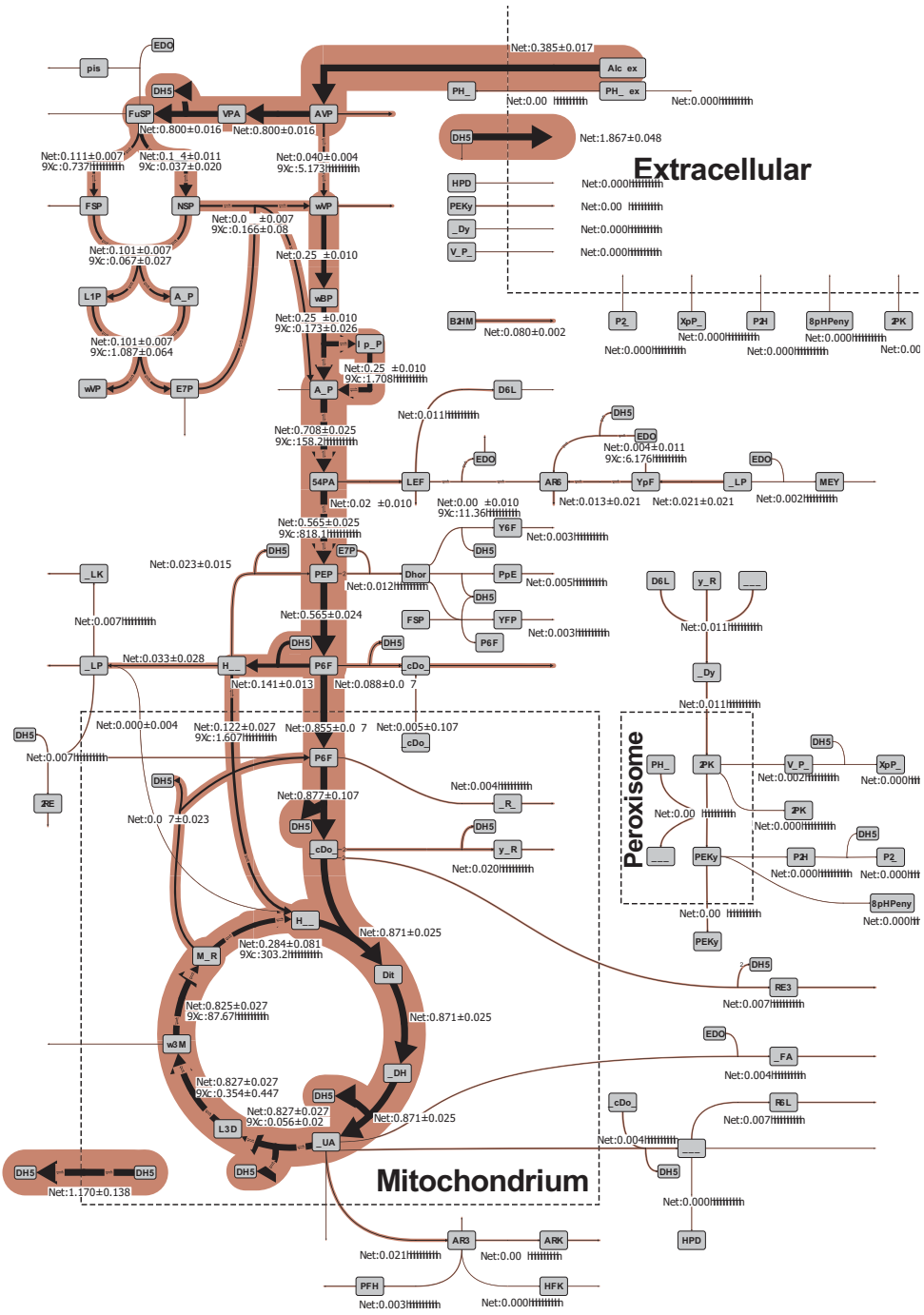


figure C.13: Flux map for chemostat 2 with a growth rate of 0.03 h^{-1}. Specific rates given in [mmol g^{-1}_{CDW} h^{-1}].

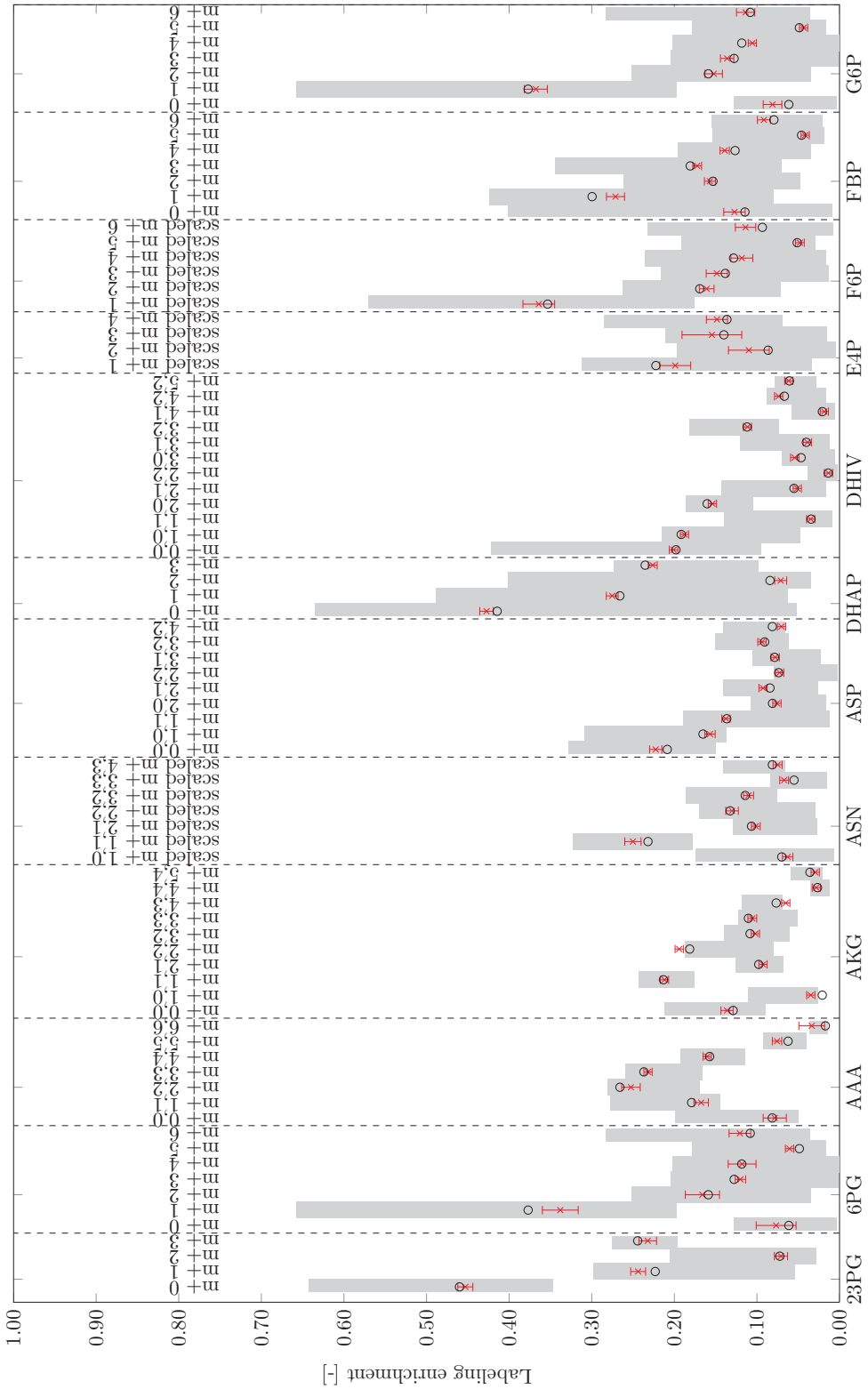


figure C.14: Chemostat 2. Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (–×–) measurement and (○) simulated data.

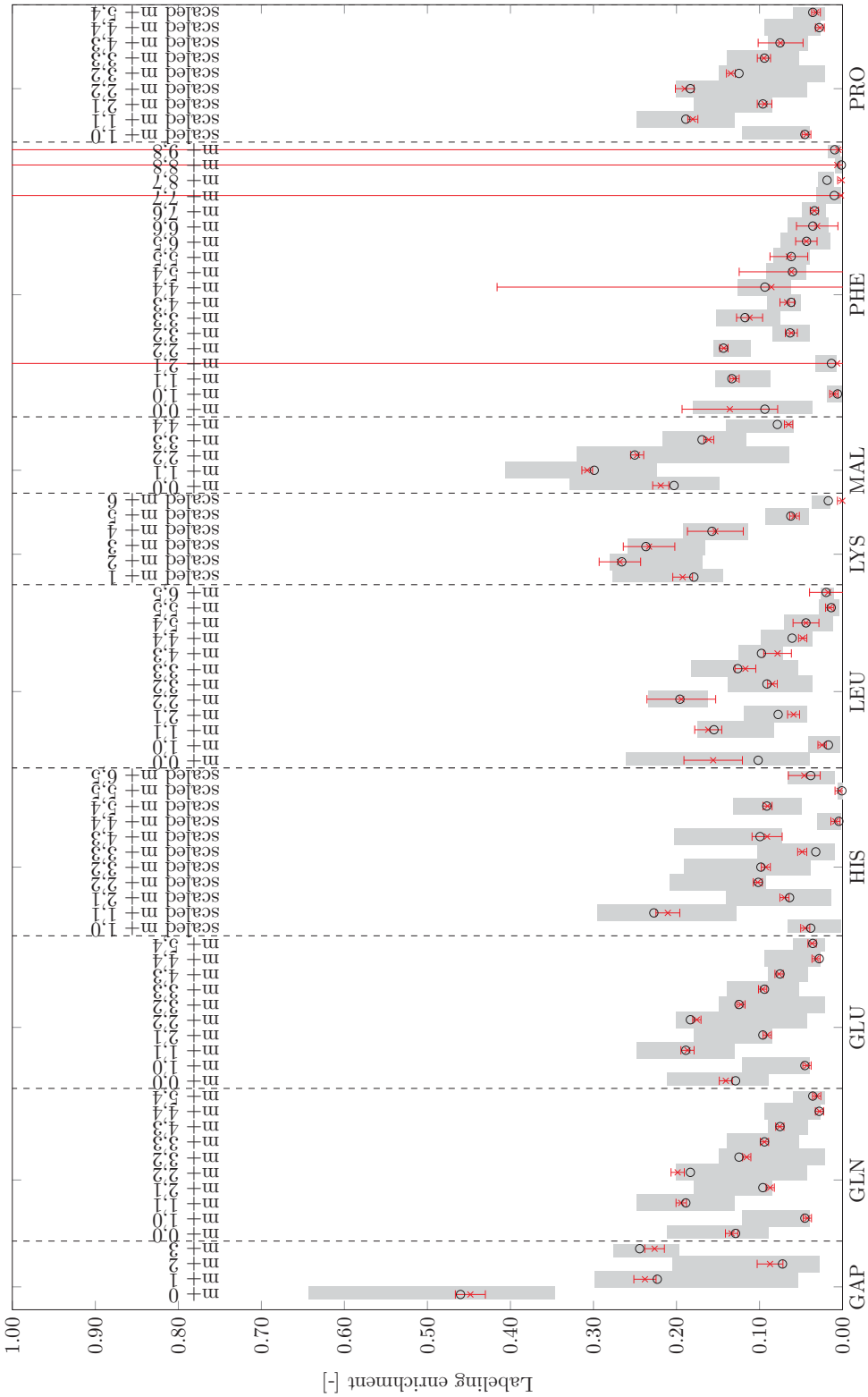


figure C.15: Chemostat 2 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

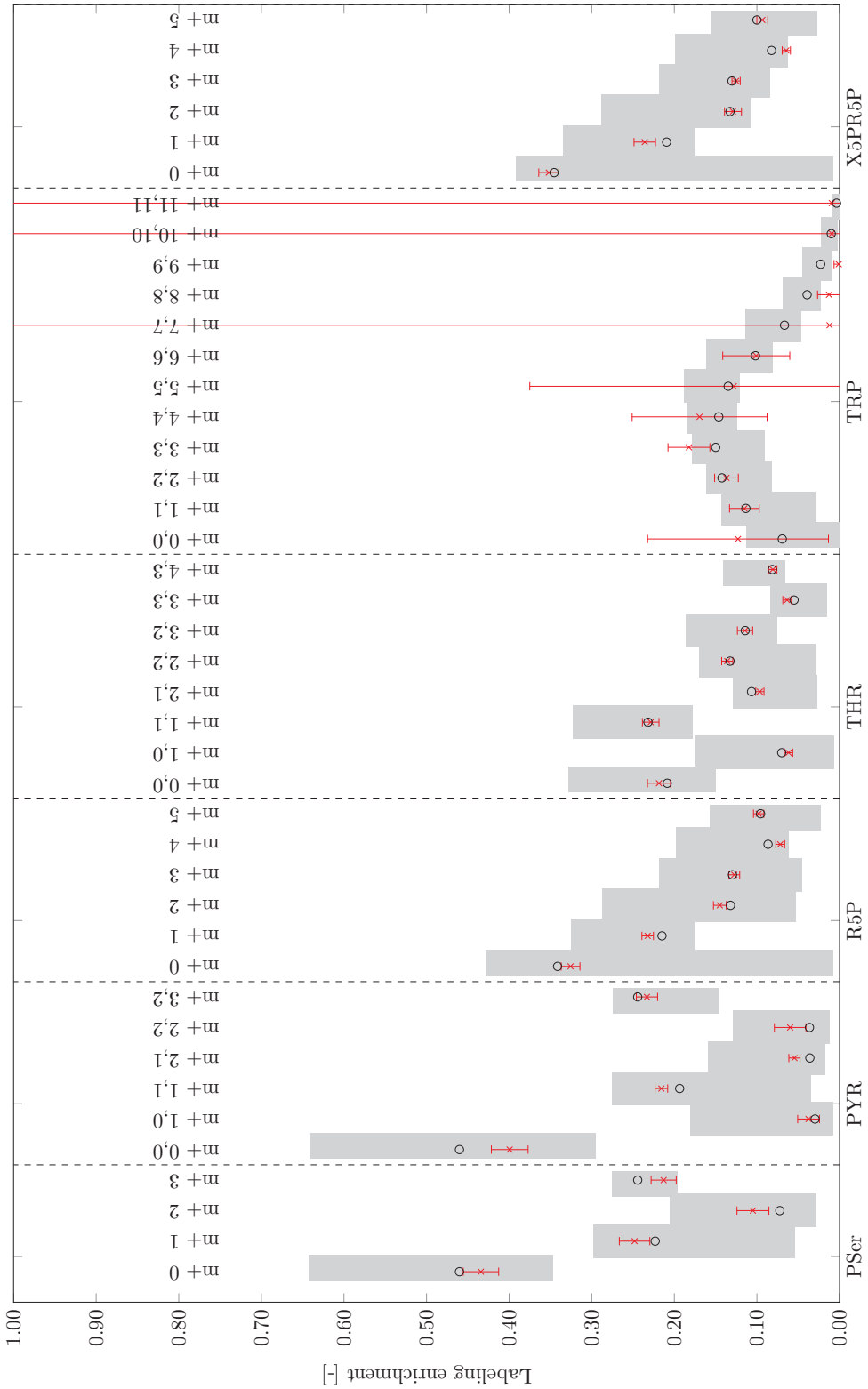


figure C.16: Chemostat 2 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

Chemostat 3: 0.05 h^{-1}

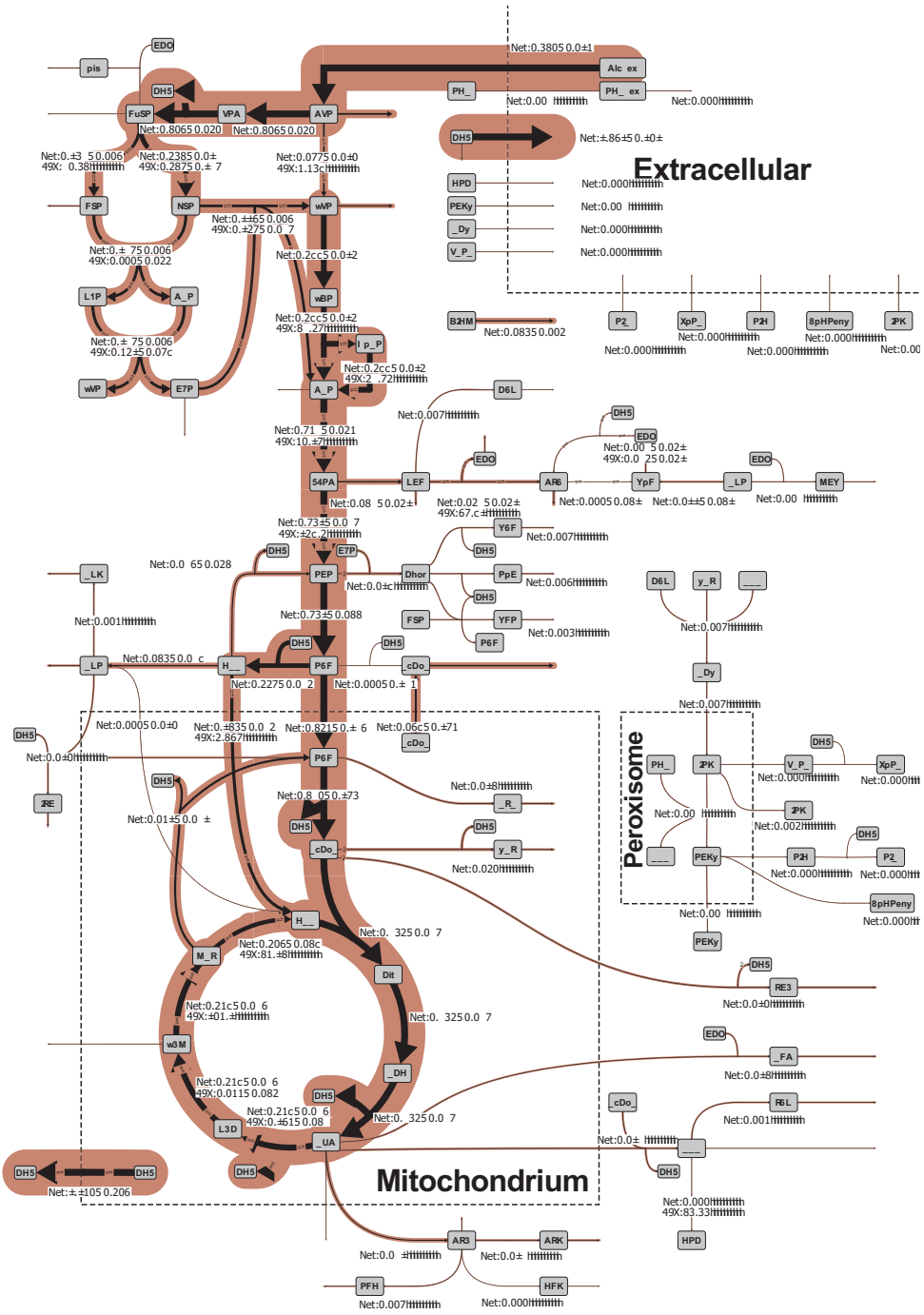


figure C.17: Flux map for chemostat 3 with a growth rate of 0.05 h^{-1}. Specific rates given in [mmol g^{-1}_{CDW} h^{-1}].

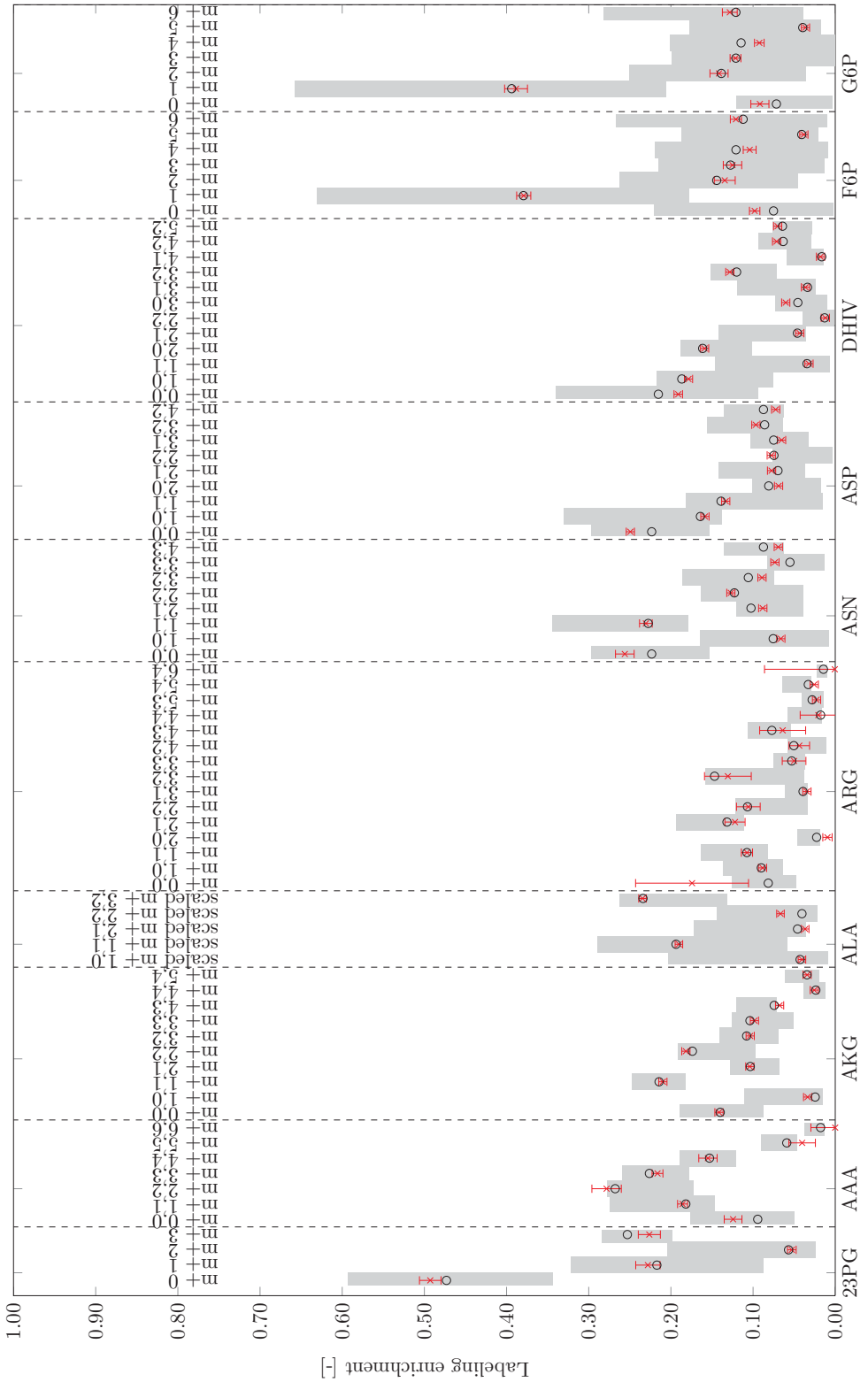


figure C.18: Chemostat 3. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

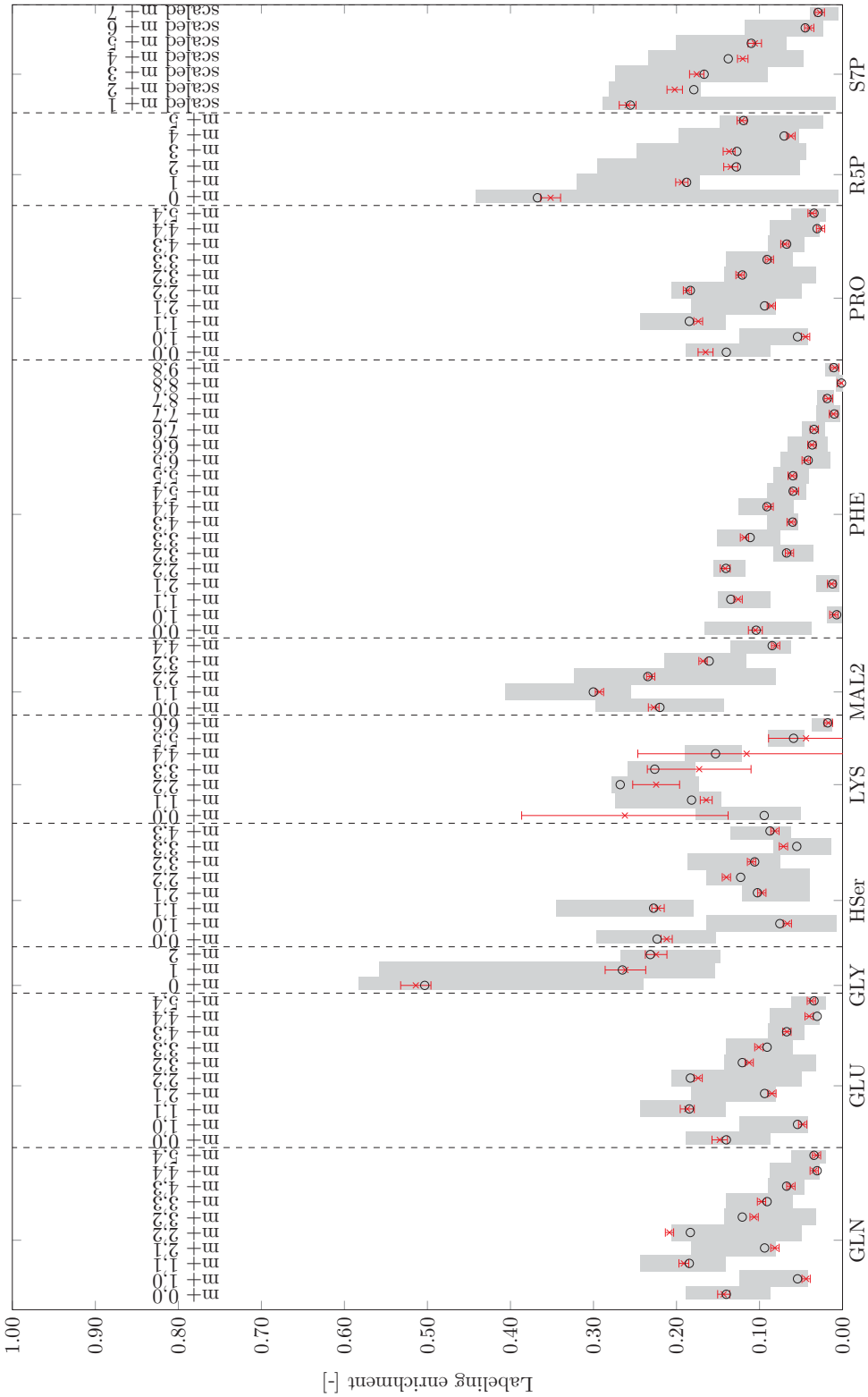


figure C.19: Chemostat 3 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

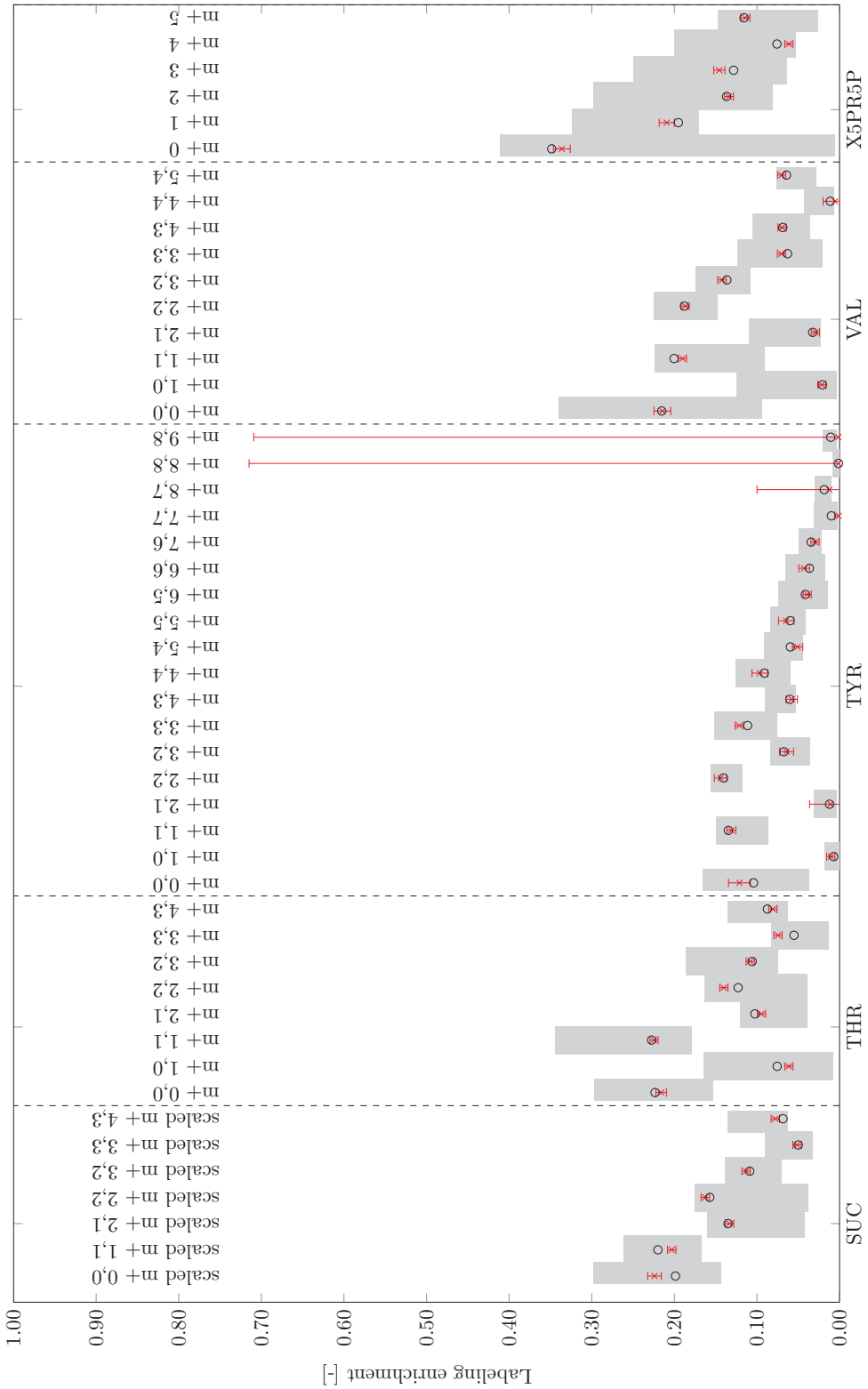


figure C.20: Chemostat 3 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—⁎—) measurement and (○) simulated data.

Chemostat 1 and 2: 0.03 h^{-1}

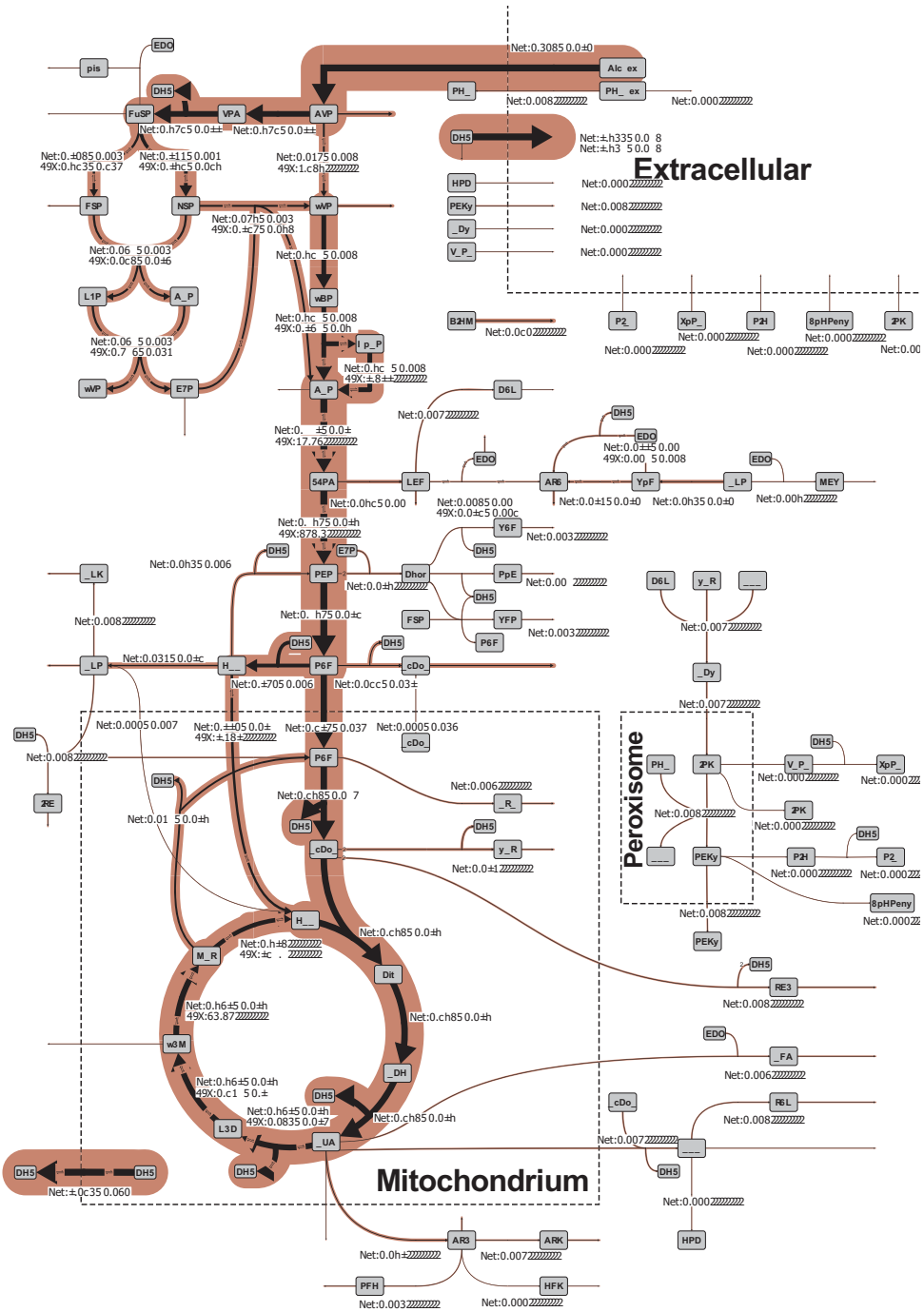


figure C.21: Flux map for chemostat 1 and 2 with a growth rate of 0.03 h^{-1}. Specific rates given in [mmol g^{-1}_{CDW} h^{-1}].

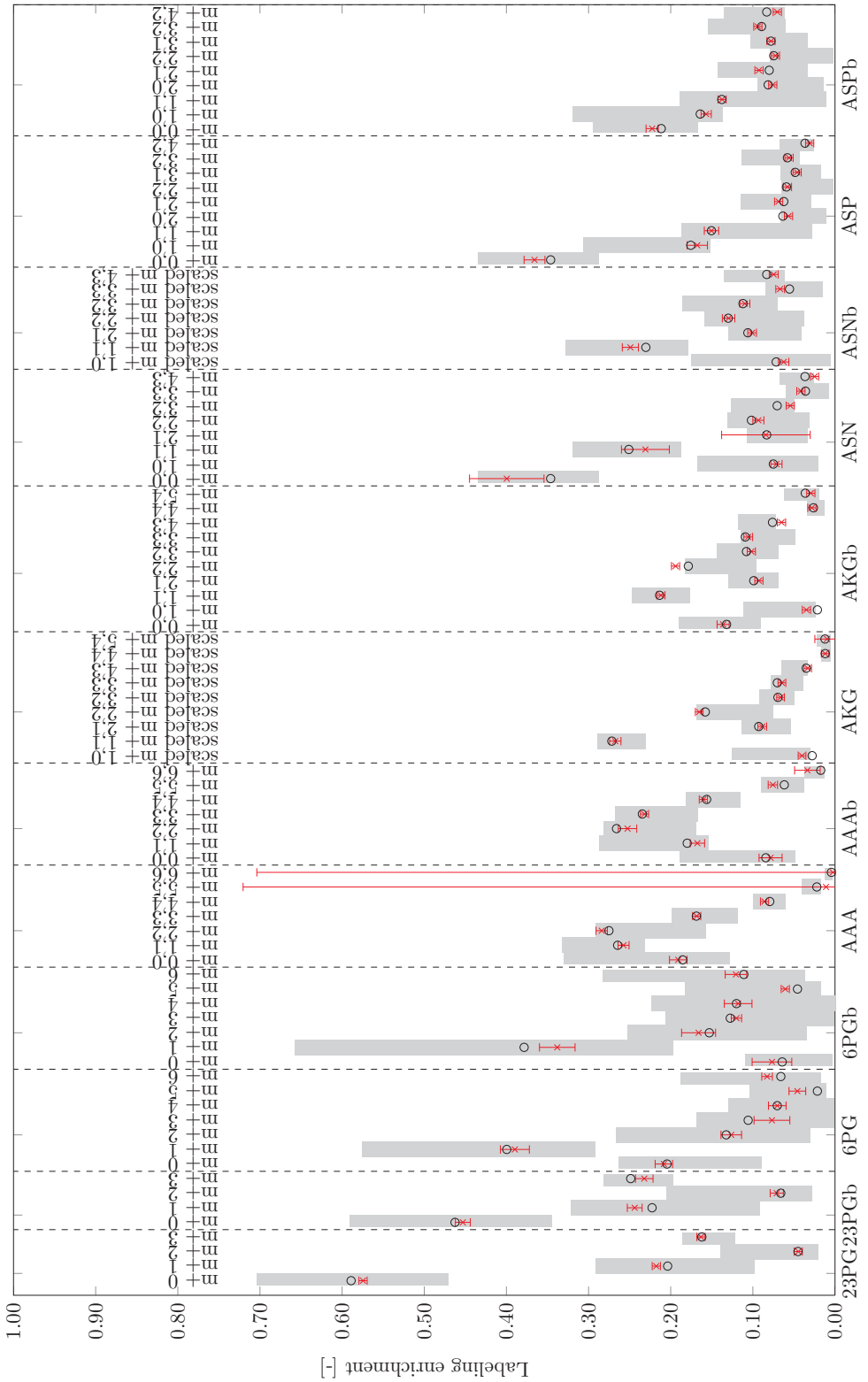


figure C.22: Chemostat 1 and 2. Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

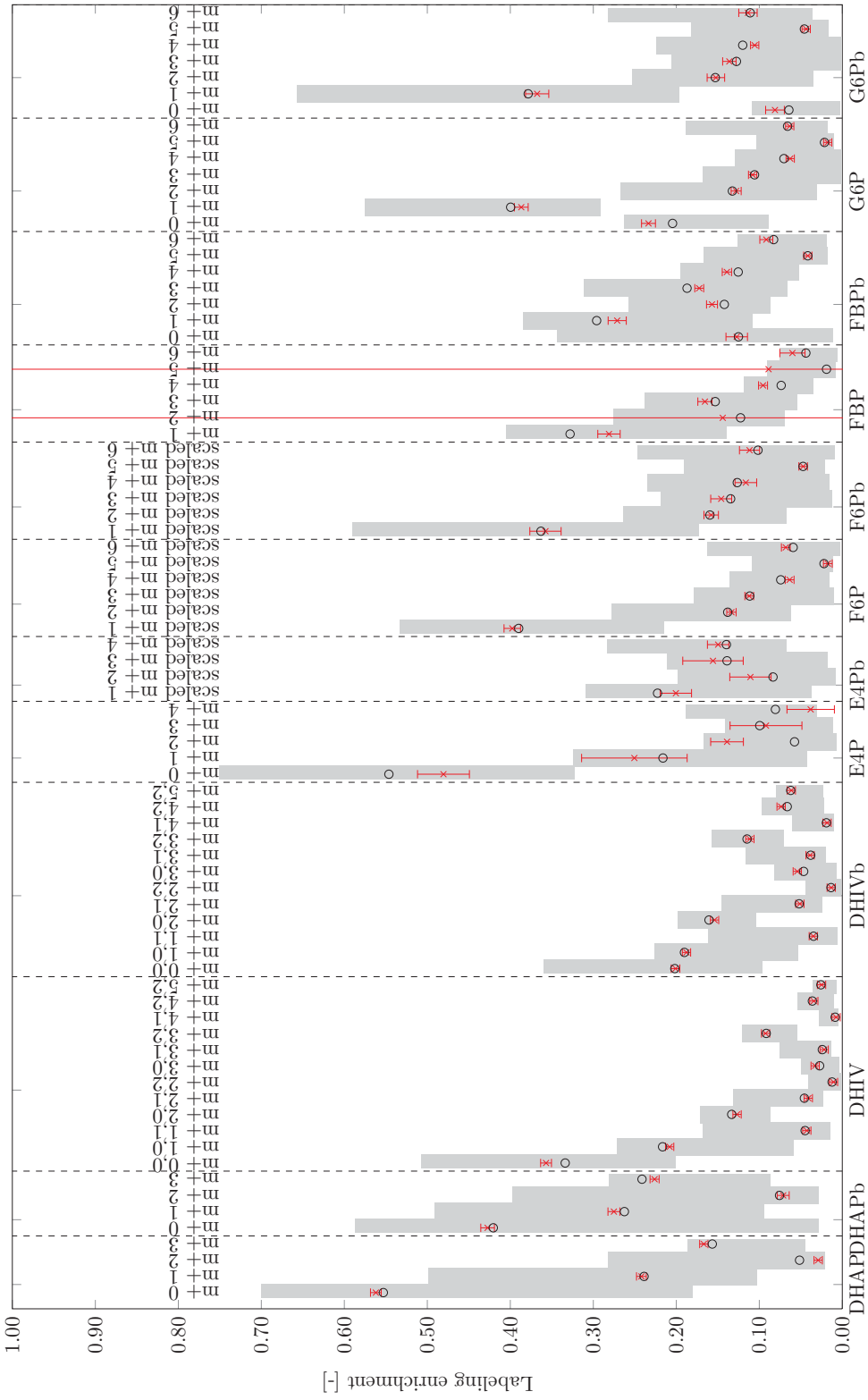


figure C.23: Chemostat 3 (continued). Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

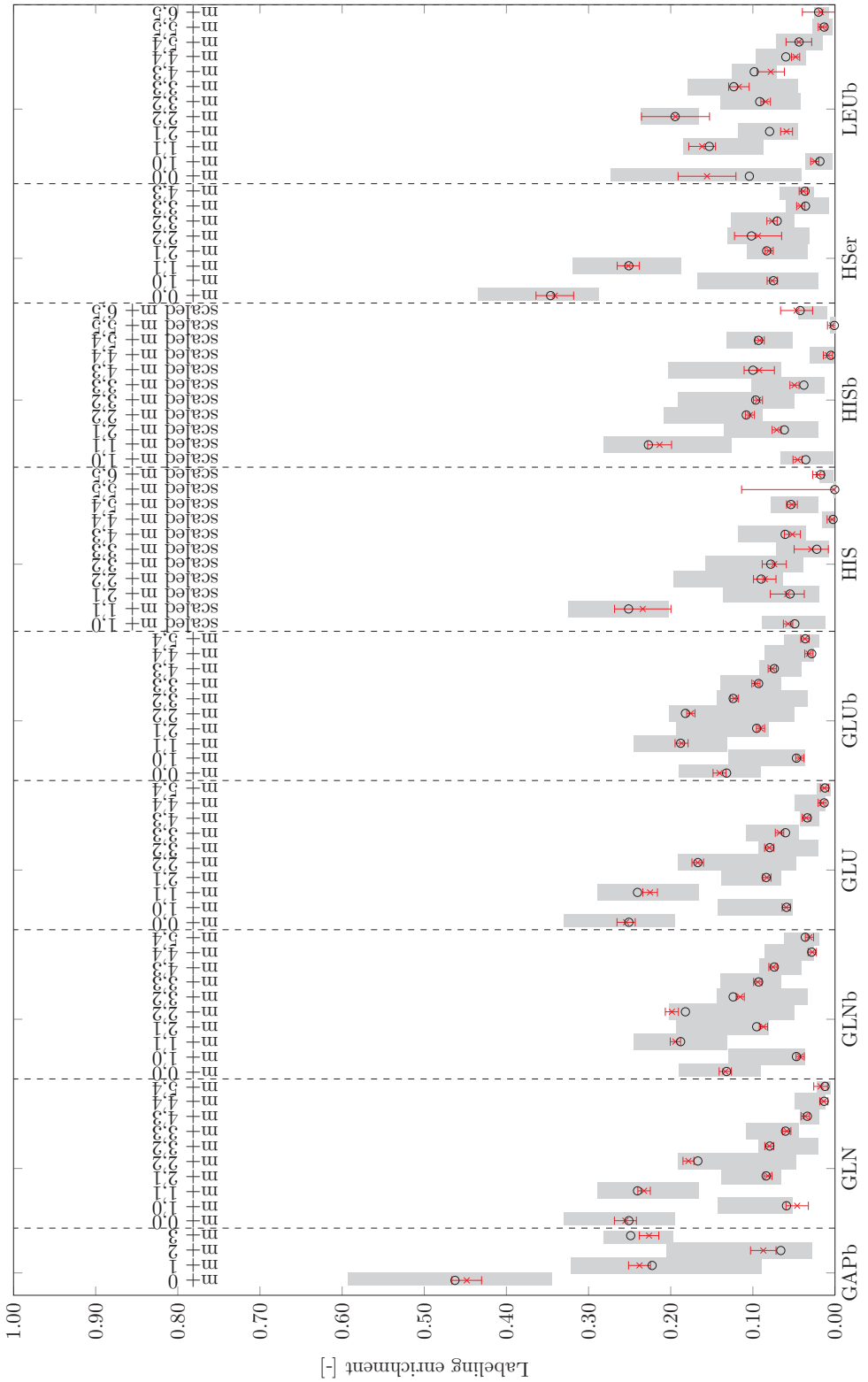
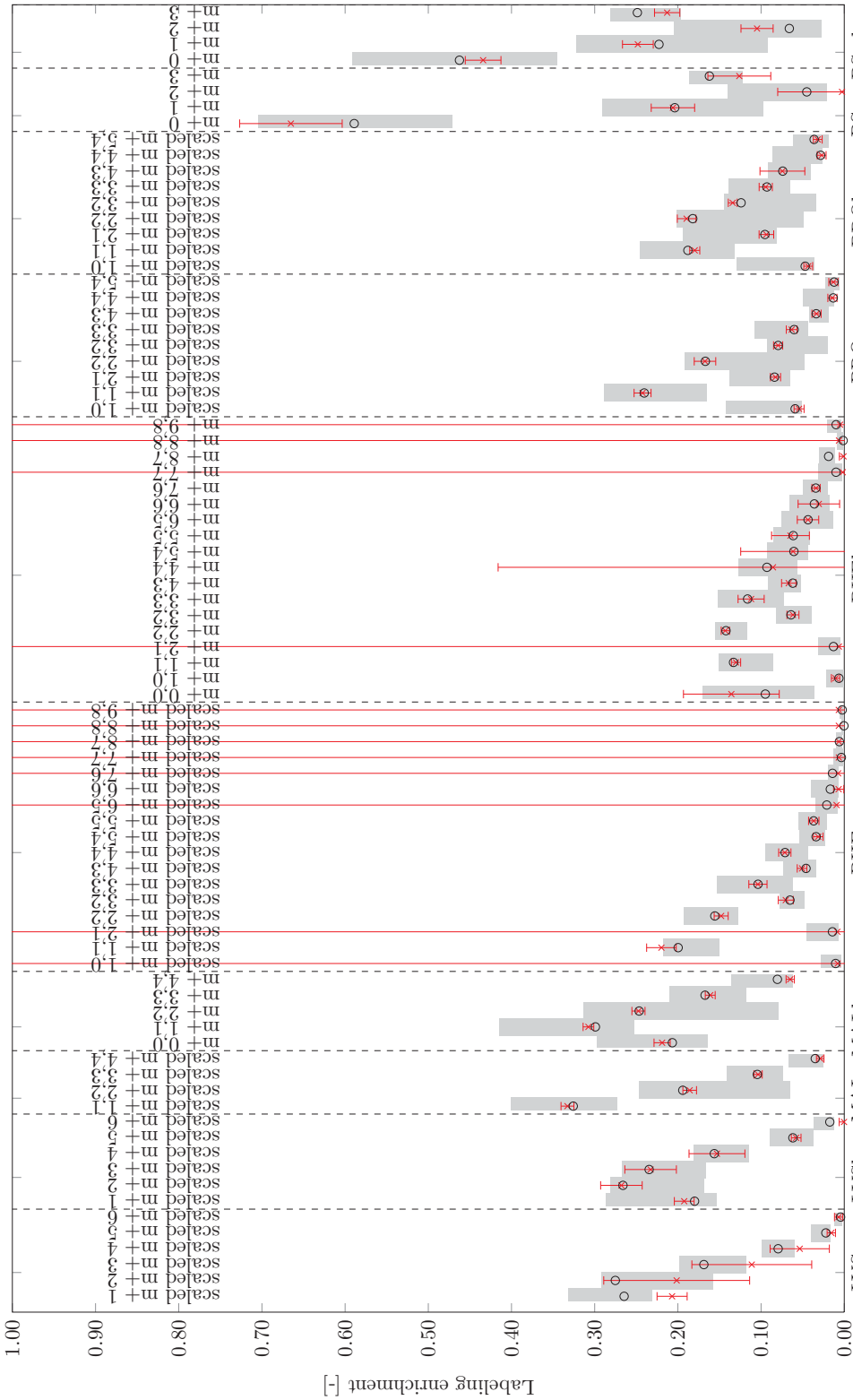


figure C.24: Chemostat 3 (continued). Plot of minimal/maximal value (▭) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.


Labeling enrichment [-]
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
LYS
LYSb
MAL
MALb
PHE
PHEb
PRO
PROb
PSer
PSerb
m+ 0
m+ 1
m+ 2
m+ 3
m+ 4
m+ 5
m+ 6
m+ 7
m+ 8
m+ 9
m+ 10
m+ 11
m+ 12
m+ 13
m+ 14
m+ 15
m+ 16
m+ 17
m+ 18
m+ 19
m+

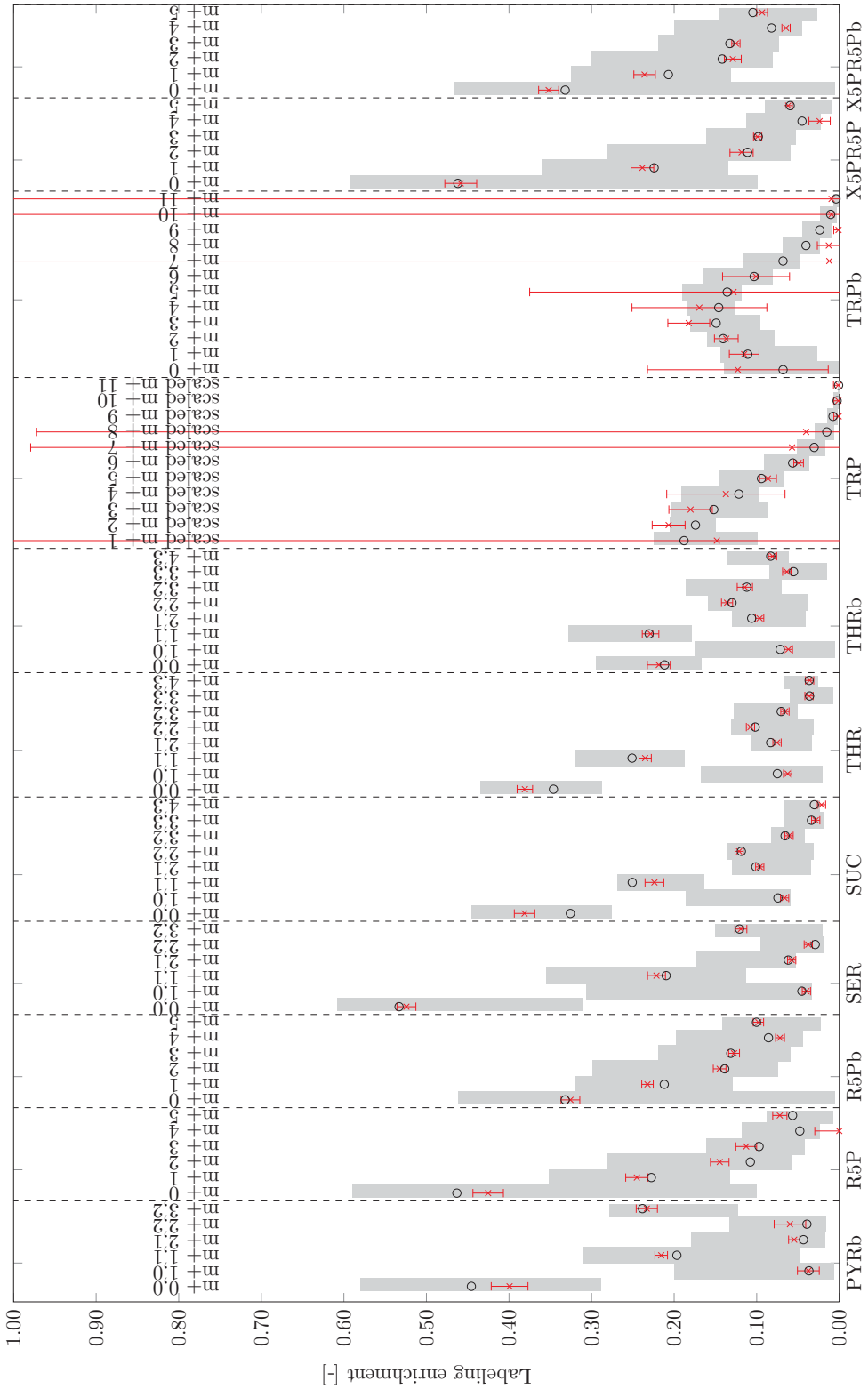
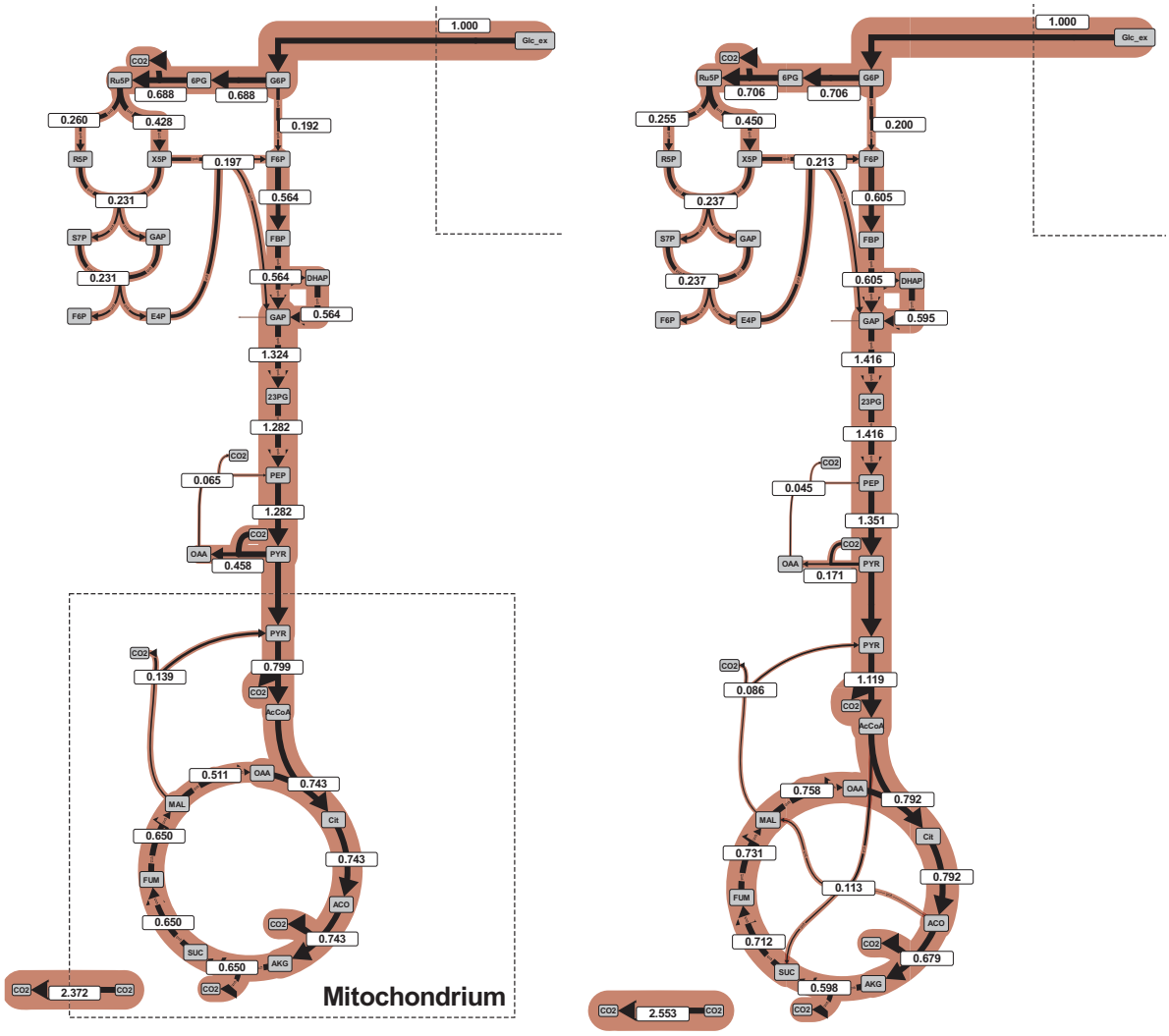


figure C.26: Chemostat 3 (continued). Plot of minimal/maximal value (▬) for the simulated measurements for 50,000 flux sample, (—×—) measurement and (○) simulated data.

C.5 Comparison of Large and Focused Model



(a) Focused Network

(b) Large sclae network

figure C.27: Best found solution for large scale (485 reactions) and focused ^{13}C metabolic network using the same set of measurements of chemostat 1 with a growth rate of $0.03\,h^{-1}$. Influxes were scaled to 1.0. Enzymatic reactions occurring in several compartments were mapped to one reaction for visualization. Fluxes are given in [mmol $g_{CDW}^{-1}\,h^{-1}$].

Appendix D

Diversification Driven Experimental Design

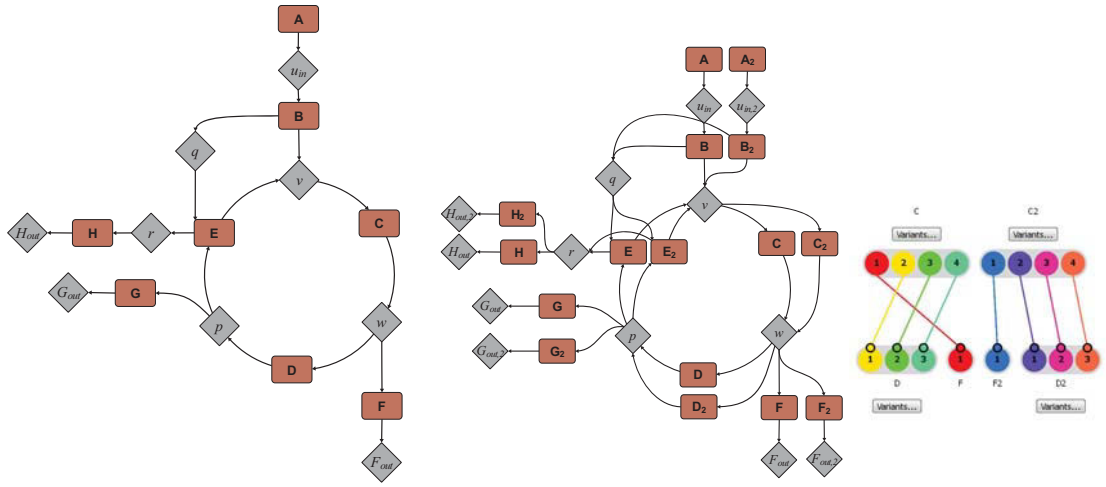
Next, an algorithm for duplicating metabolic networks is introduced.

Algorithm 1 *Duplication of network structure for ^{13}C-MFA to evaluate two labeling data-sets in parallel at the same metabolic steady-state:*

Given is a root model A with reactions $r_A = \{r_{A,1}, r_{A,2}, ...r_{A,n_r}\}$ and metabolites $m_A = \{m_{A,1}, m_{A,2}, ...m_{A,n_m}\}$. The same definitions can be likewise applied for the duplicated model B :

1. *Keep all constraints, input substrates and measurements for model A*
2. *Duplicate all metabolites of all reactions and rename them to m_B, except...*
   (a) *a reaction is an influx $r_{A,j,in}$ $j = 1, 2, ...n_{A,r}$*
   (b) *a reaction is an efflux $r_{A.j,out}$ $j = 1, 2, ...n_{A,m}$*
3. *Duplicate all atom transitions of m_A for the added metabolites m_B*
4. *Duplicate all effluxes $r_{A,out}$ to $r_{B,out}$ and rename the metabolites in $r_{B,out}$ to m_B. If a metabolite m_i contains $n_{i,r,out} > 1$ effluxes, set $n_{i,r,out} - 1$ effluxes of the two system equal ($r_{A,i,out} = r_{B,i,out}$)*
5. *Duplicate all influxes $r_{A,in}$ to $r_{B,in}$ and rename the metabolite in $r_{B,in}$ to m_B. If a metabolite m_i contains $n_{i,r,in} > 1$ influxes, set $n_{i,r,in} - 1$ influxes of the two system equal ($r_{A,i,in} = r_{B,i,in}$)*
6. *Specify input substrate mixture for system m_B*
7. *Introduce new measurements for system m_B*

Next, the proposed algorithm is applied to the Spiral model (chapter 2.2.3). Duplicating the Spiral model, results in the combined model shown in figure D.1. The number of metabolites



(a) Spiral network (b) Spiral network for combined experiments

figure D.1: (a) Spiral network. (b) Duplicated Spiral network. The networks metabolites and atom transitions are duplicated.

were increasing from 8 to 16. In- and effluxes were increasing from 4 to 8. In the end, the model possesses only 2 degrees of freedom as the original Spiral model.
The atom transitions in the reaction are duplicated. Thus, another set of measurement values can be used, to (consistently) estimate the flux distribution by using experimental data derived from two different substrate mixtures.

Appendix E

Multi-Objective Experimental Design

E.1 Measurement Matrices

Measurement matrices are built from isotopomer fractions. A metabolite with n_C carbon atoms possesses 2^n isotopomers. The sum of all isotopomer fractions equals 1. Measurable atoms are specified by their atom numbers written in brackets, i.e. Pyr[2] is measuring the second carbon atom. In MS/MS the first atoms #M specified are measured mother ion carbon atoms and the daughter ion is specified by #d: [#M:#d].

$$\begin{pmatrix} HNMR_{labeled} \\ HNNMR_{unlabeled} \end{pmatrix} = M_{HNMR,Pyr[1]} \cdot m = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} m_{000} \\ m_{001} \\ m_{010} \\ m_{011} \\ m_{100} \\ m_{101} \\ m_{110} \\ m_{111} \end{pmatrix}$$

$$\begin{pmatrix} CNMR_{singulet} \\ CNMR_{doublet,1} \\ CNMR_{doublet,2} \\ CNMR_{double\,doublet} \end{pmatrix} = M_{CNMR,Pyr[2]} \cdot m = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{000} \\ m_{001} \\ m_{010} \\ m_{011} \\ m_{100} \\ m_{101} \\ m_{110} \\ m_{111} \end{pmatrix}$$

$$\begin{pmatrix} MS_{m+0} \\ MS_{m+1} \\ MS_{m+2} \\ MS_{m+3} \end{pmatrix} = M_{CNMR,Pyr[2]} \cdot m = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{000} \\ m_{001} \\ m_{010} \\ m_{011} \\ m_{100} \\ m_{101} \\ m_{110} \\ m_{111} \end{pmatrix}$$

table E.1: Substrate costs, each substrate was assumed to possess a purity of 99 %, unlabeled glucose posses a putiry of 0.9807. Data were taken from www.isotope.com [1].

Substrate	Price [€ g^{-1}]
GLC#000000	0.3
GLC#100000	147.0
GLC#010000	472.0
GLC#001000	912.0
GLC#000100	1218.0
GLC#000010	1293.0
GLC#000001	532.0
GLC#111111	134.0
GLC#110000	494.0
GLC#100001	1141.0

$$\begin{pmatrix} MSMS_{m+0,m+0} \\ MSMS_{m+1,m+0} \\ MSMS_{m+1,m+1} \\ MSMS_{m+2,m+1} \\ MSMS_{m+2,m+2} \\ MSMS_{m+3,m+2} \end{pmatrix} = M_{CNMR,Pyr[2]} \cdot m = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{000} \\ m_{001} \\ m_{010} \\ m_{011} \\ m_{100} \\ m_{101} \\ m_{110} \\ m_{111} \end{pmatrix}$$

$$\begin{pmatrix} GCCIRMS_{labeled} \\ GCCIRMS_{unlabeld} \end{pmatrix} = M_{GCCIRMS,Pyr[1-3]} \cdot m = \begin{pmatrix} 1 & 2/3 & 2/3 & 1/3 & 2/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 2/3 & 1/3 & 2/3 & 2/3 & 1 \end{pmatrix} \begin{pmatrix} m_{000} \\ m_{001} \\ m_{010} \\ m_{011} \\ m_{100} \\ m_{101} \\ m_{110} \\ m_{111} \end{pmatrix}$$

E.2 Cost Calculations

For the cultivation media with 20 g/L substrate is used. It is conducted in a 250 mL bioreactor operating as a chemostat. 6 volumes of the bioreactor are needed to conduct the experiment (resulting in 1.5 L medium). 39 h of observation and preparation for the experiment are assumed. The wage is 20 €/h. A measurement is checked, transferred to the model in 0.0055 h (20 seconds) based on our experiences in LC-MS/MS. The measurement time varies depending on the machine used. 10 samples are assumed as maximum per experiment.

It was assumed that each device is in usage for five years, the resulting costs for acquisition are passed to the device usage time.

table E.2: Costs for devices and measurement time per replicate

Substrate	Price [€]	Measurement time [h]
GC-MS	100 000	0.66
LC-MS	180 000	0.66
LC-MS/MS	350 000	0.66
NMR	1 000 000	6

E.3 Simulation Results

Constrained Fluxes

In the following table all fluxes needed to be kept constant for evaluation of statistics is marked with "X" (to receive a Fisher Information with a condition number below the given threshold, see corresponding chapter). "(X)" marks fluxes which needed to be constant in order to obtain comparable results for the separate simulations (comparability between models is guaranteed).

FLUX	LC-MS/MS	LC-MS	GC-MS	HNMR	CNMR	GC-C-IRMS
Byprod_ex.n	X	X	X	X	X	X
Ana1.n	X	X	X	X	X	X
Ana1.x	X	X	X	X	X	X
Tca8.x	X	X	X	X	X	X
Thr2.x	X	X	X	X	X	X
gly2.x	X	X	X		X	X
Ana2.x	X	X	X	X	X	X
gly4.x	(X)	(X)	X	X	X	X
gly4b.x	(X)	(X)	X	X	X	X
gly5.x	(X)	(X)	(X)	X	X	X
gly6.x	(X)	(X)	(X)	X	X	X
Ppp3.x	(X)	(X)	(X)	X	X	X
Transporter_OAA.x	X	X	X	X	(X)	X
Asp1b.n				X		X
CO2_ex.n						
PPP2.n						X
Tca1.n				X		
Tca8.n						
Transporter_CO2.n				X		X
Transporter_Pyr.n				X		X
Gly1.n						
Gly7.n						
PPP4.x				X		X
PPP5.x						X
PPP6.x				X		X
PPP7.x						X
Tca5b.x				X		X

Tca6a.x	X	X
Tca7a.x	X	X
Transporter_AcCoA.x		X
Help_Ru5P.n	X	X
Gly1.x		X
Gly2.x	X	X
Ana3.x		X

Reaction name:	Reaction and atom transitions
BM GLUC:	G6P(ABCDEF) →
BM MAN:	F6P(ABCDEF) →
BM SER:	SER(ABC) →
BM GLY:	GLY(AB) →
gly1:	Glc ex(ABCDEF) → G6P(ABCDEF)
gly2:	G6P(ABCDEF) ⇌ F6P(AFBCDE)
gly3:	F6P(ABCDEF) → FBP(ABCDEF)
gly4:	FBP(ABCDEF) ⇌ DHAP(EBF) + GAP(DAC)
gly5:	GAP(ABC) ⇌ 23PG(BCA)
gly7:	PEP(ABC) → PYR(ABC)
PPP2:	6PG(ABCDEF) → Ru5P(DBAEC) + CO2(F)
PPP5:	X5P(ABCDE) + R5P(FGHIJ) ⇌ GAP(EBD) + S7P(AFCGJHI)
Ser1:	23PG(ABC) → PHP(ABC)
Gly1:	SER(ABC) ⇌ EC1(A) + GLY(BC)
Gly2:	EC1(A) + CO2(B) ⇌ GLY(AB)
Thr2:	AcCoA(AB) + GLY(CD) ⇌ THR(ABCD)
Transporter Pyr:	PYR(ABC) → PYR [mit] (ABC)
Transporter OAA:	OAA(ABCD) ⇌ OAA [mit] (ABCD)
Ana1:	OAA(ABCD) ⇌ PEP(ABD) + CO2(C)
Ana2:	PYR(ABC) + CO2(D) ⇌ OAA(ABDC)
Thr1:	ASP(ABCD) → HSer(ACBD)
Ac3:	PYR(ABC) → CO2(C) + AcCoA(AB)
BM AcCoA:	AcCoA(AB) →
Transporter AcCoA:	AcCoA [mit] (AB) ⇌ AcCoA(AB)
CO2 ex:	CO2(A) →
PPP1:	G6P(ABCDEF) → 6PG(DACEBF)
Asp1:	OAA(ABCD) → ASP(ABCD)
BM ASP:	ASP(ABCD) →
His2:	R5P(ABCDE) + EC1(F) → HIS(CEFDBA)
BM HIS:	HIS(ABCDEF) →
CHOR:	E4P(ABCD) + PEP(EFG) + PEP(HIJ) → Chor(EBDHFICAGJ)
Phe1:	Chor(ABCDEFGHIJ) → PHE(GCHBDAFEI) + CO2(J)
Tyr1:	Chor(ABCDEFGHIJ) → TYR(DBHCAFGEI) + CO2(J)
BM PHE:	PHE(ABCDEFGHI) →
BM TYR:	TYR(ABCDEFGHI) →
BM EC1:	EC1(A) →
PPP4:	Ru5P(ABCDE) ⇌ R5P(BDECA)
gly4b:	DHAP(ABC) ⇌ GAP(ABC)
Ser2:	PHP(ABC) → PSer(ABC)
Met1:	EC1(A) + ASP(BCDE) → MET(ABDCE)
gly6:	23PG(ABC) ⇌ PEP(ABC)
BM MET:	MET(ABCDE) →
BIOM IN:	BIOM() → BIOMass()
BIOM ex:	BIOMass() →
Trp1:	Chor(ABCDEFGHIJ) → PYR(AEI) + Ant(CGBHFDJ)

Trp2:	Ant(ABCDEFG) + R5P(HIJKL) → Ant2(ABCDLHEFKIJG)
Trp3:	Ant2(ABCDEFGHIJKL) → CO2(L) + Ant3(ABCDEFGIHJK)
Trp4:	Ant3(ABCDEFGHIJK) → TRP(ABCDKEHGJIF)
Tca2:	OAA [mit] (ABCD) + AcCoA [mit] (EF) → Cit [mit] (AECFDB)
Ala:	PYR [mit] (ABC) → ALA [mit] (ABC)
Val1:	PYR [mit] (ABC) + PYR [mit] (DEF) → CO2 [mit] (C) + DHIV [mit] (ADEFB)
Leu:	PYR [mit] (ABC) + PYR [mit] (DEF) + AcCoA(GH) → LEU(DABEGH) + CO2(F) + CO2(C)
Ileu1:	PYR [mit] (ABC) + ASP(DEFG) → CO2(C) + KILE(FADBEG)
BM VAL:	VAL(ABCDE) →
BM ALA:	ALA [mit] (ABC) →
BM ILE:	ILE(ABCDEF) →
BM LEU:	LEU(ABCDEF) →
BM LYS:	LYS(ABCDEF) →
BM THR:	THR(ABCD) →
BM TRP:	TRP(ABCDEFGHIJK) →
Tca1:	PYR [mit] (ABC) → AcCoA [mit] (AB) + CO2 [mit] (C)
Transporter CO2:	CO2 [mit] (A) → CO2(A)
BM ERY:	E4P(ABCD) →
BM GLU:	GLU [mit] (ABCDE) →
BM PRO:	PRO [mit] (ABCDE) →
BM ARG:	ARG(ABCDEF) →
Lys1:	AKG [mit] (ABCDE) + AcCoA(FG) → CO2(E) + AAA(BDAFCG)
BM AAA:	AAA(ABCDEF) →
BM ORN:	ORN [mit] (ABCDE) →
BM GALAC:	G6P(ABCDEF) →
BM MANNOSE:	F6P(ABCDEF) →
BM GLUCOS:	F6P(ABCDEF) →
BM ARAB:	Ru5P(ABCDE) →
BM ICIT:	AKG [mit] (ABCDE) →
BM FUM:	FUM [mit] (ABCD) →
Pen1:	POA() + IPN(ABCDEF) → PENV() + AAA(ABCDEF)
PEN ex:	PENV() →
Tca4:	ACO [mit] (ABCDEF) → AKG [mit] (CBAED) + CO2 [mit] (F)
Tca7a:	FUM [mit] (ABCD) ⇌ MAL [mit] (BADC)
Glu1:	AKG [mit] (ABCDE) → GLU [mit] (ABCDE)
Glu3:	GLU [mit] (ABCDE) → GLN [mit] (ABCDE)
Glu2:	GLU [mit] (ABCDE) → PRO [mit] (BADCE)
Glu4:	GLU [mit] (ABCDE) → ORN [mit] (BADCE)
Lys2:	AAA(ABCDEF) → LYS(ACBEDF)
Arg1:	AKG [mit] (ABCDE) + EC1(F) → ARG(BADCEF)
Thr3:	HSer(ABCD) → THR(BACD)
Val2:	DHIV [mit] (ABCDE) → KIV(ABECD)
Val3:	KIV(ABCDE) → VAL(ABCDE)
Ileu2:	KILE(ABCDEF) → ILE(ABCDEF)
Ser3:	PSer(ABC) → SER(ABC)
Cys1:	OAcSer(ABCDE) → CYS(BDE) + AcCoA(AC)

Cys2:	CYS(ABC) + CYS(DEF) → Cyst(DAEBFC)
BM CYS:	CYS(ABC) →
Pen2:	CV() + AAA(ABCDEF) → IPN(ABCDEF)
BM Cyst:	Cyst(ABCDEF) →
BM GLN:	GLN [mit] (ABCDE) →
Asp2:	ASP(ABCD) → ASN(ABCD)
BM ASN:	ASN(ABCD) →
Tca5b:	AKG [mit] (ABCDE) ⇌ SUC [mit] (ABCD) + CO2 [mit] (E)
Tca5a:	AKG [mit] (ABCDE) ⇌ SUC [mit] (BADC) + CO2 [mit] (E)
Poa IN:	POA IN() → POA()
BM CO2:	CO2(A) →
BM R5P:	R5P(ABCDE) →
Pen Syn VAL:	VAL(ABCDE) →
Pen Syn CYS:	CYS(ABC) →
PPP6:	GAP(ABC) + S7P(DEFGHIJ) ⇌ F6P(BDCAHF) + E4P(JEIG)
PPP7:	E4P(ABCD) + X5P(EFGHI) ⇌ F6P(BEDCAG) + GAP(IFH)
PPP3:	Ru5P(ABCDE) ⇌ X5P(ABCDE)
BM GAP:	GAP(ABC) →
Tca3:	Cit [mit] (ABCDEF) → ACO [mit] (ABFCDE)
Tca6a:	SUC [mit] (ABCD) ⇌ FUM [mit] (ABCD)
Tca6b:	SUC [mit] (ABCD) ⇌ FUM [mit] (BADC)
Ser4:	SER(ABC) + AcCoA(DE) → OAcSer(DAEBC)
Ana3:	PYR [mit] (ABC) + CO2 [mit] (D) ⇌ MAL [mit] (ABDC)
Tca8:	MAL [mit] (ABCD) ⇌ OAA [mit] (ABCD)
Tca7b:	FUM [mit] (ABCD) ⇌ MAL [mit] (ABCD)
Asp1b:	OAA [mit] (ABCD) → ASP(ABCD)
bisACV IPN ex:	IPN(ABCDEF) →
Pen3:	PENV() → Byprod()
Byprod ex:	Byprod() →
Help Ru5P:	Ru5P(ABCDE) → Ru5P X5P(ABCDE)
Help X5P:	X5P(ABCDE) → Ru5P X5P(ABCDE)
Help Ru5P X5P:	Ru5P X5P(ABCDE) →
Pen0:	CV IN() ⇌ CV()

GC-MS

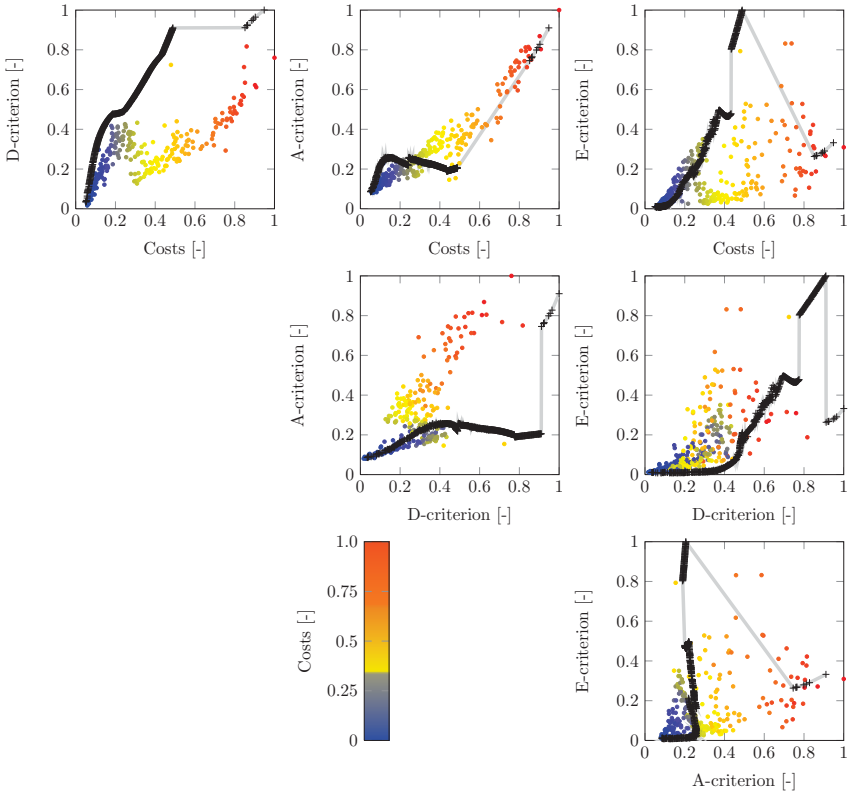


figure E.1: Multi-objective optimization results for GC-MS. Color codes for costs of design point. The results of three objectives (D-criterium, dimensions and costs) are shown (plus signs). Normalized criteria were used and scaled to the interval $[0, 1]$. The five objective optimization is shown in color coded (D-, A-, E-criterium, dimensions and costs).

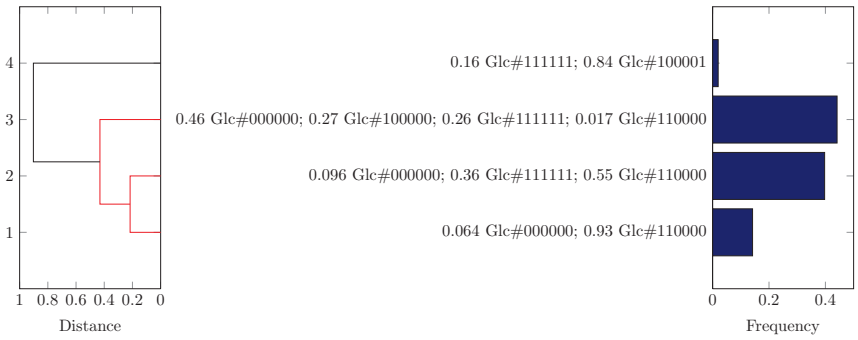


figure E.2: Dendrogram of the found clustered substrate mixtures for GC-MS. The best known Pareto set was clustered by the minimal distance of the composition fractions of labeling. The length of the tree shows dissimilarities in the mixtures found. The average values of the cluster are shown in vertical directions. The frequency is shown in the bar plot on the right side. Basically four different mixtures were found frequently for GCMS.

LC-MS/MS

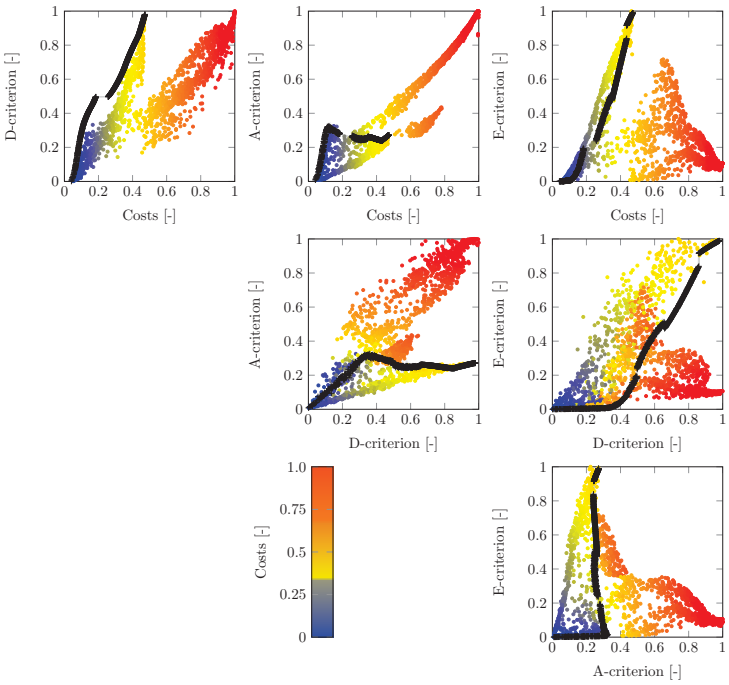


figure E.3: Multi-objective optimization results for LC-MS/MS. Color codes for costs of design point. The results of three objectives (D-criterion, dimensions and costs) are shown (plus signs). Normalized criteria were used and scaled to the interval $[0, 1]$. The five objective optimization is shown in color coded (D-, A-, E-criterium, dimensions and costs).

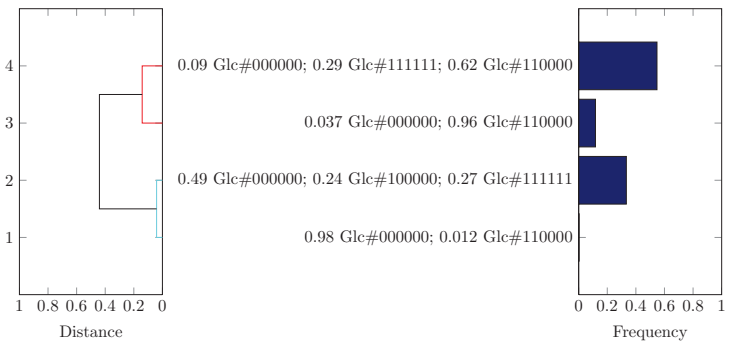


figure E.4: Dendrogram of the found clustered substrate mixtures for LC-MS/MS. The best known Pareto set was clustered by the minimal distance of the composition fractions of labeling. The length of the tree shows dissimilarities in the mixtures found. The average values of the cluster are shown in vertical directions. The frequency is shown in the bar plot on the right side. Basically four different mixtures were found for LC-MS/MS. For these four mixtures, the number of measurements used for the separate metabolites is plotted in the following figures. It starts with the first cluster in the figure.

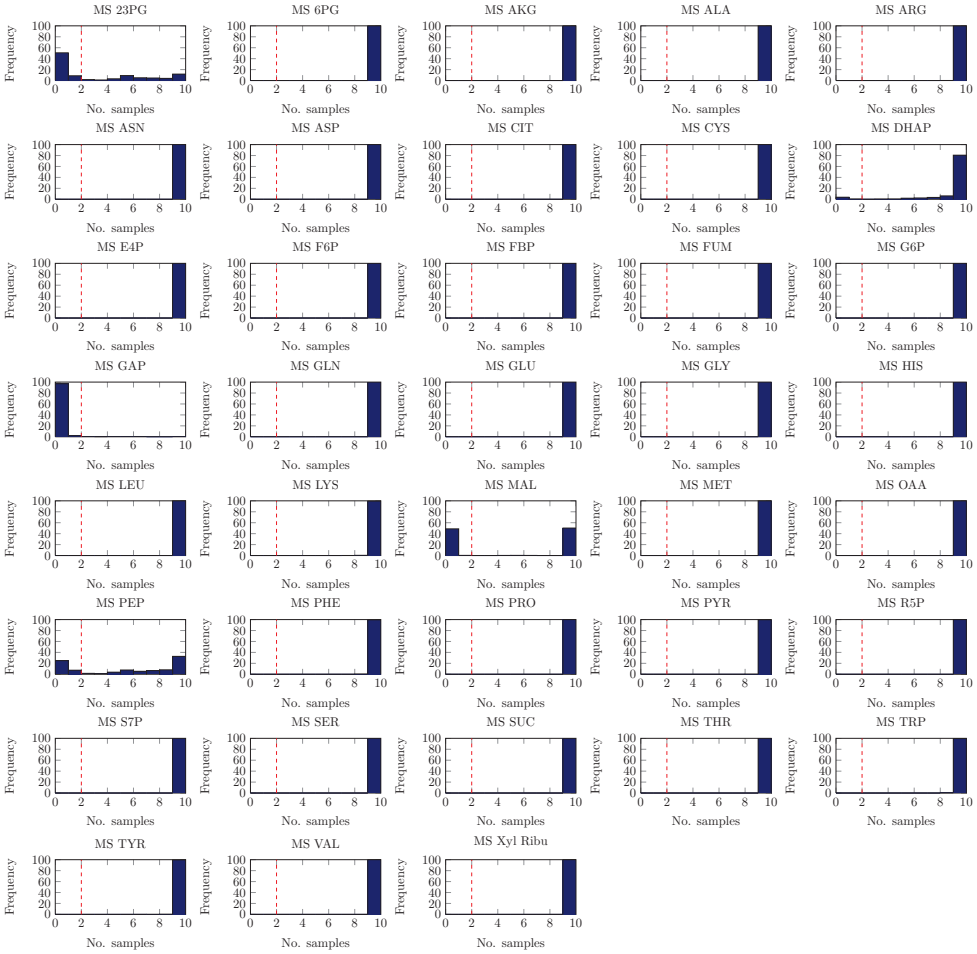


figure E.5: Clustered results of measurement number in multi-objective optimization for LCMSMS (D-criteria, dimension and costs). First cluster is shown. Figure a) Histogram plots are shown for each measurement. The replicate number for the single measurement of the separate measurements is shown. Number of measurements for mixture cluster of shown dendrogram. Red line indicates 2 replicates. Below this threshold measurements were removed from the statistics calculation.

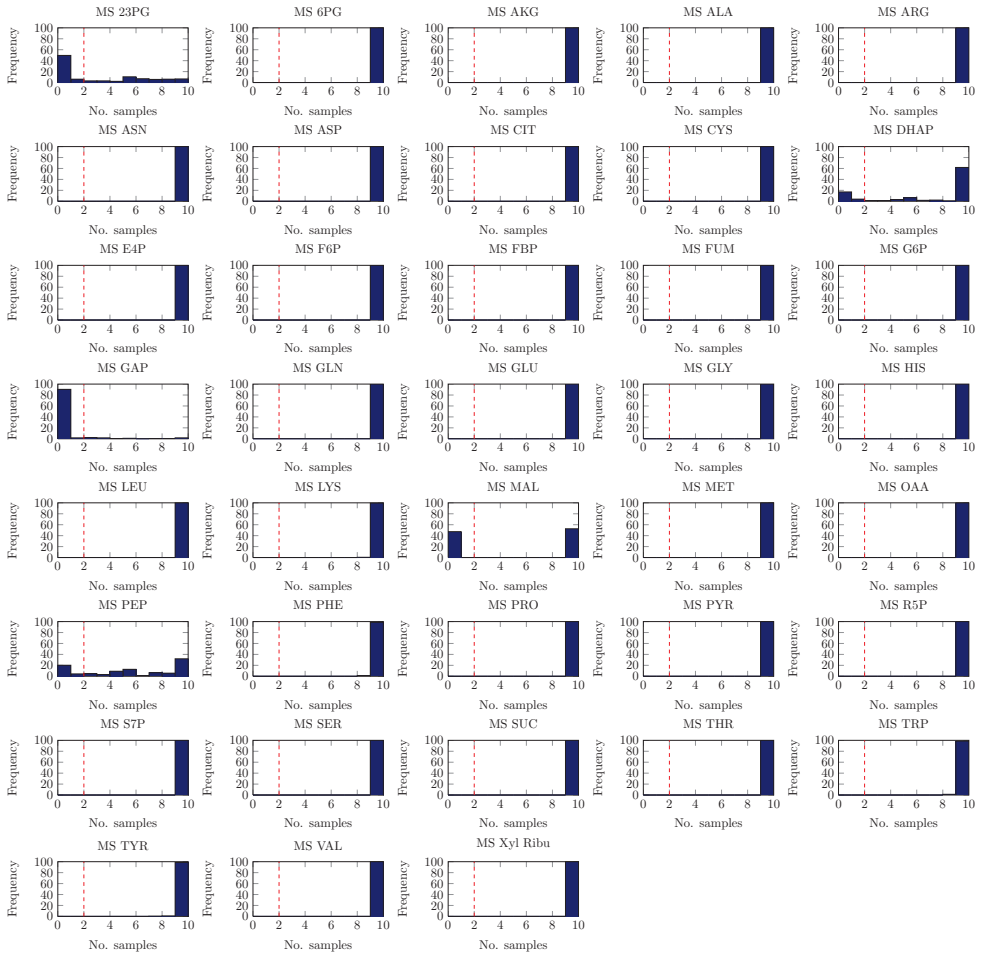


figure E.6: Clustered results of measurement number in multi-objective optimization for LCMSMS (D-criteria, dimension and costs). Second cluster is shown. Figure a) Histogram plots are shown for each measurement. The replicate number for the single measurement of the separate measurements is shown. Number of measurements for mixture cluster of shown dendrogram. Red line indicates 2 replicates. Below this threshold measurements were removed from the statistics calculation.

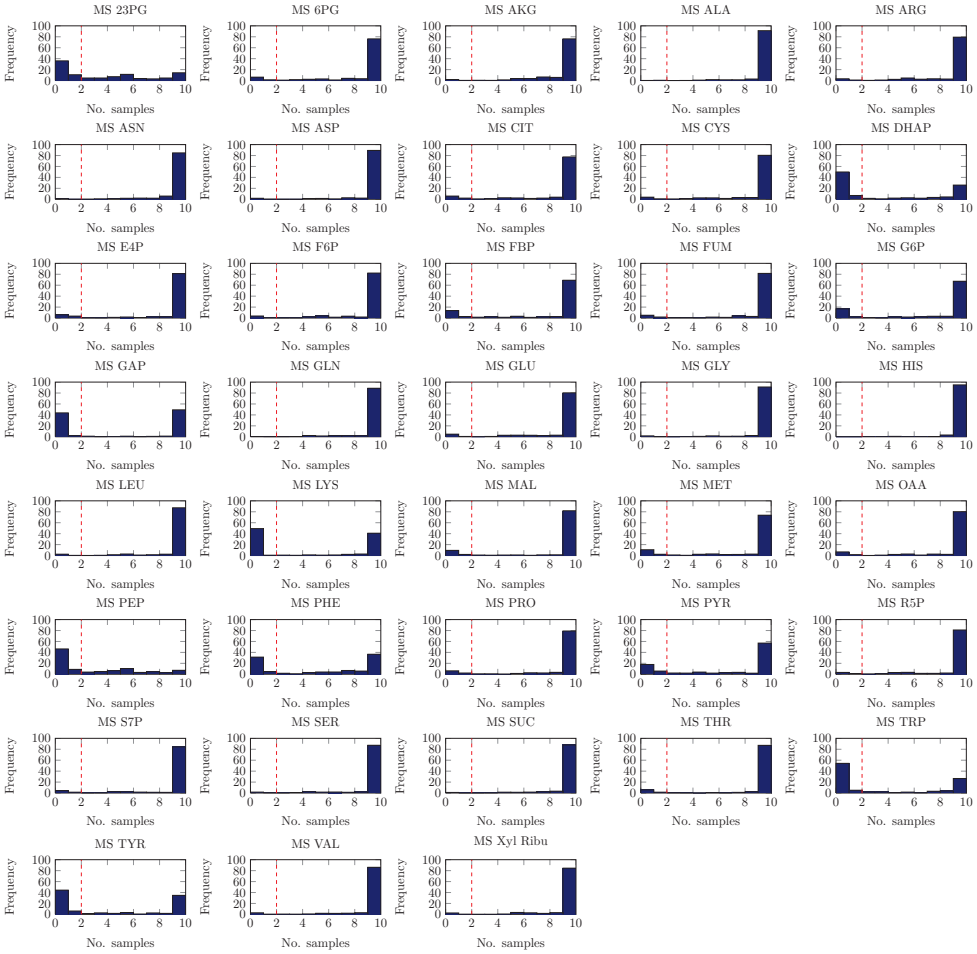


figure E.7: Clustered results of measurement number in multi-objective optimization for LCMSMS (D-criteria, dimension and costs). Third cluster is shown. Figure a) Histogram plots are shown for each measurement. The replicate number for the single measurement of the separate measurements is shown. Number of measurements for mixture cluster of shown dendrogram. Red line indicates 2 replicates. Below this threshold measurements were removed from the statistics calculation.

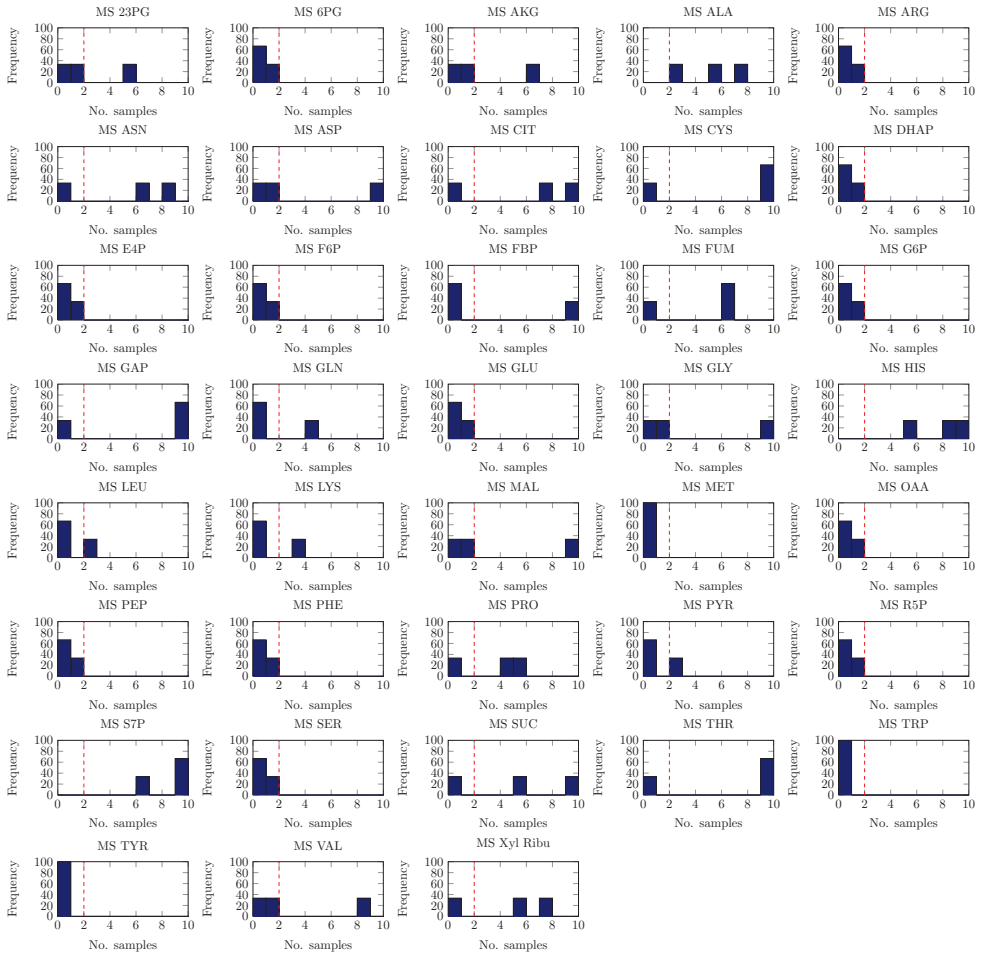


figure E.8: Clustered results of measurement number in multi-objective optimization for LCMSMS (D-criteria, dimension and costs). Fourth cluster is shown. Figure a) Histogram plots are shown for each measurement. The replicate number for the single measurement of the separate measurements is shown. Number of measurements for mixture cluster of shown dendrogram. Red line indicates 2 replicates. Below this threshold measurements were removed from the statistics calculation.

C-NMR

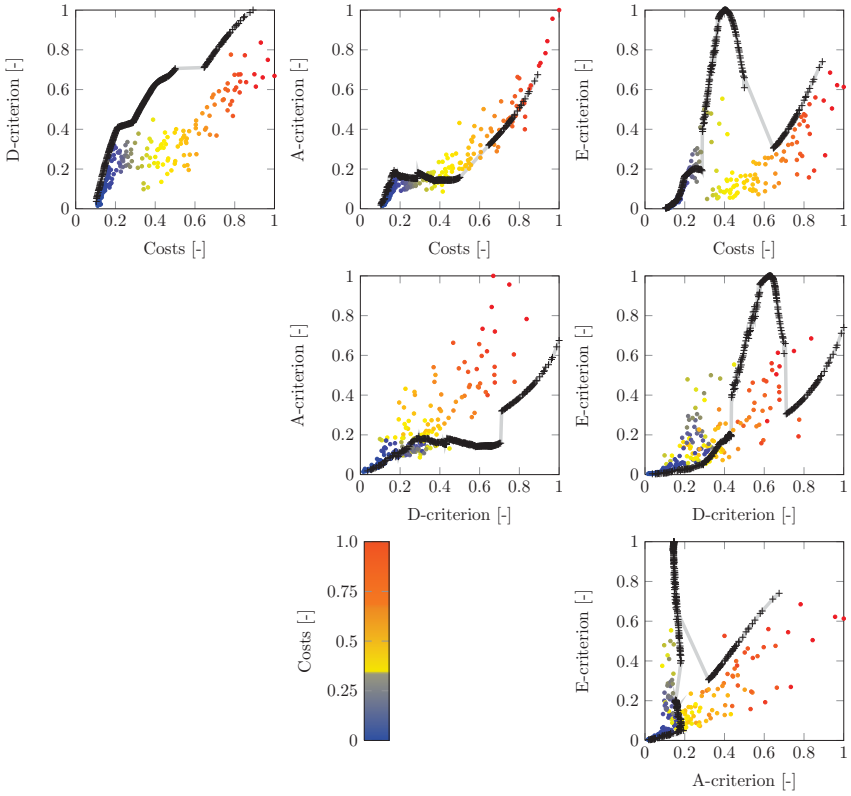


figure E.9: Multi-objective optimization results for C-NMR. Color codes for costs of design point. The results of three objectives (D-criterium, dimensions and costs) are shown (plus signs). Normalized criteria were used and scaled to the interval $[0, 1]$. The five objective optimization is shown in color coded (D-, A-, E-criterium, dimensions and costs).

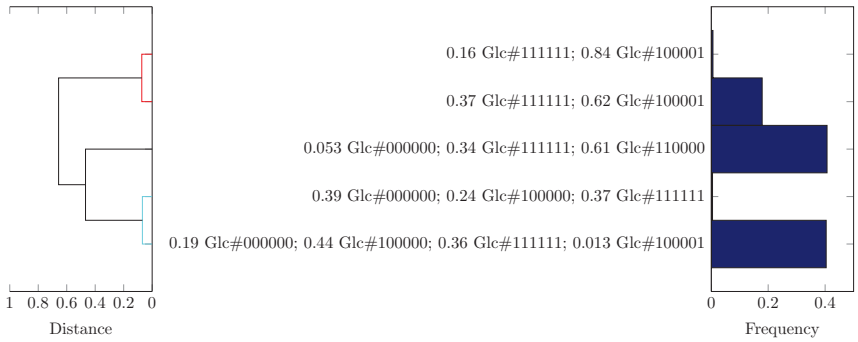


figure E.10: Dendrogram of the found clustered substrate mixtures for C-NMR. The best known Pareto set was clustered by the minimal distance of the composition fractions of labelings. The length of the tree shows dissimilarities in the mixtures found. The average values of the cluster are shown in vertical directions. The frequency is shown in the bar plot on the right side. Basically five different mixtures were found frequently for C-NMR.

LC-MS

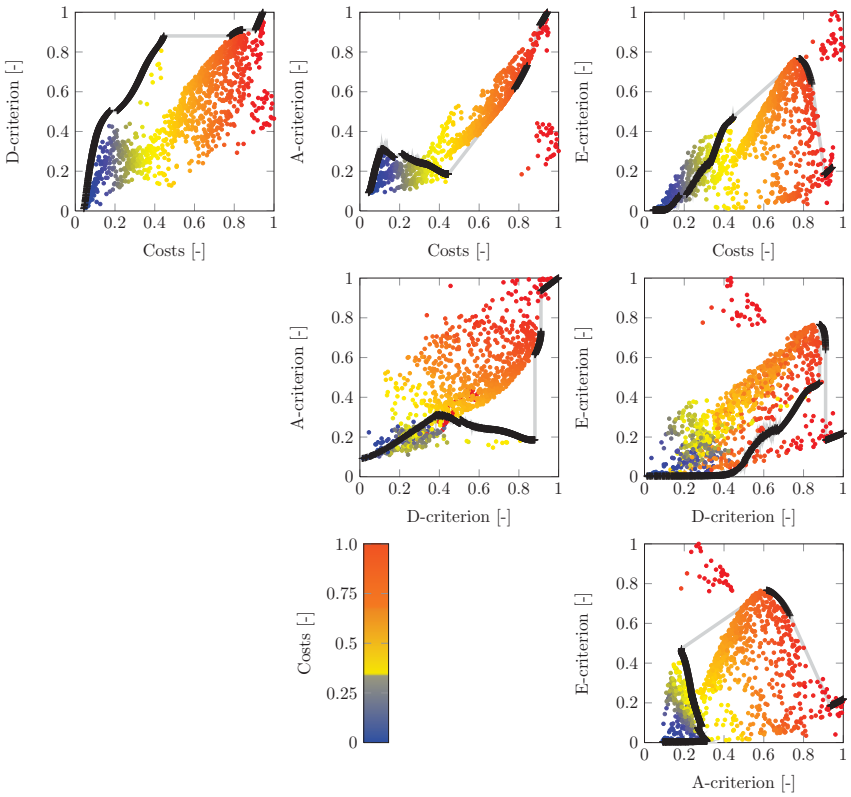


figure E.11: Multi-objective optimization results for LC-MS. Color codes for costs of design point. The results of three objectives (D-criterion, dimensions and costs) are shown (plus signs). Normalized criteria were used and scaled to the interval $[0, 1]$. The five objective optimization is shown in color coded (D-, A-, E-criterium, dimensions and costs).

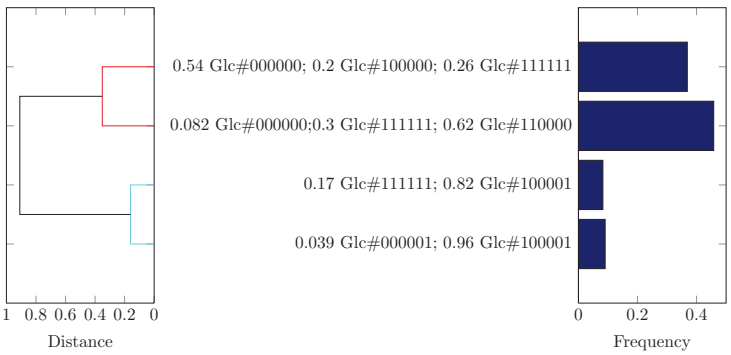


figure E.12: Dendrogram of the found clustered substrate mixtures for LC-MS/MS. The best known Pareto set was clustered by the minimal distance of the composition fractions of labeling. The length of the tree shows dissimilarities in the mixtures found. The average values of the cluster are shown in vertical directions. The frequency is shown in the bar plot on the right side. Basically four different mixtures were found frequently for LC-MS.

E.4 Measurement Specifications

- *#Full* refers to the number of measured mother ions containing all carbon atoms of the metabolite in the respective publication
- #Fragment refers to the number of fragments measured in the respective publications
- *Fragment C* is describing the carbon atom numbers found in the fragment

GC-MS

Metabolite	Network comp.	#Fragment	#Full	Fragment C	Reference
Alanine	Ala	1	1	[2-3]	[20]
Alanin	Ala	1	1	[2-3]	[4]
Alanin	Ala	1	1	[2-3]	[229]
Alanin	Ala	2	1	[2-3]|[2-3]	[49]
Alanin	Ala	1	1	[2-3]	[9]
Alanin	Ala	1	1	[2-3]	[152]
Alanin	Ala	1	1	[2-3]	[121]
Arginin	Arg	1	0	[1-5]	[20]
Aspartate	Asp	1	1	[2-4]	[20]
Aspartate	Asp	2	1	[2-4]|[2-4]	[229]
Aspartate	Asp	2	1	[1-2]|[2-4]	[4]
Aspartate	Asp	3	1	[1-2]|[1-2]|[2-4]	[9]
Aspartate	Asp	2	1	[2-4]|[1-2]	[121]
Asx	-	3	1	[2-4]|[2-4]|[1-2]	[49]
Asx	-	3	1	[2-4]|[2-4]|[1-2]	[152]
Fumarate	Fum	0	1		[152]
Glutamate	Glu	1	1	[2-5]	[20]
Glutamate	Glu	1	1	[2-5]	[4]
Glutamate	Glu	2	1	[2-4]|[2-4]	[229]
Glutamate	Glu	2	1	[2-5]|[2-5]	[9]
Glutamate	Glu	1	1	[2-5]	[121]
Glycine	Gly	1	1	[2]	[20]
Glycine	Gly	1	1	[2]	[4]
Glycine	Gly	1	1	[2]	[229]
Glycine	Gly	1	1	[2]	[49]
Glycine	Gly	1	1	[2]	[9]
Glycine	Gly	1	1	[2]	[152]
Glycine	Gly	1	1	[2]	[121]
Glyoxylate	-	1	1	[2-5]|[2-5]	[152]
Glutamate/ Glutamine	-	3	1	[2-5]|[2-5]|[1-2]	[49]
Histidine	His	1	0	[2-6]	[4]
Isoleucine	Ile	2	0	[2-6]|[2-6]	[229]
Isoleucine	Ile	2	0	[2-6]|[2-6]	[49]
Isoleucine	Ile	2	0	[2-6]|[2-6]	[9]
Isoleucine	Ile	1	1	[2-6]	[4]
Isoleucine	Ile	2	0	[2-6]|[2-6]	[152]
Isoleucine	Ile	1	1	[2-6]	[121]
Leucine	Leu	2	0	[2-6]|[2-6]	[229]
Leucine	Leu	2	0	[2-6]|[2-6]	[49]
Leucine	Leu	1	0	[2-6]	[9]
Leucine	Leu	1	1	[2-6]	[4]
Leucine	Leu	2	0	[2-6]|[2-6]	[152]
Leucine	Leu	1	1	[2-6]	[121]
Leucine	Leu	2	0	[2-6]|[2-6]	[229]
Lysine	Lys	2	1	[2-6]|[1-2]	[49]
Lysine	Lys	1	1	[2-6]	[121]
Lysine	Lys				
Methionine	Met	2	1	[2-5]|[2-5]	[229]
Methionine	Met	2	1	[2-5]|[2-5]	[49]
Methionine	Met	2	1	[2-5]|[2-5]	[9]
Methionine	Met	1	1	[2-5]	[152]

Phenylalanine	Phe	2	1	[2-9]|[1-2]	[20]
Phenylalanine	Phe	2	1	[2-9]|[2-9]	[229]
Phenylalanine	Phe	4	0	[2-9]|[2-9]|[1-2]|[3-9]	[49]
Phenylalanine	Phe	3	1	[2-9]|[1-2]|[2-9]	[9]
Phenylalanine	Phe	2	1	[2-9]|[1-2]	[4]
Phenylalanine	Phe	2	1	[2-9]|[1-2]	[152]
Phenylalanine	Phe	2	1	[2-9]|[1-2]	[121]
Proline	Pro	1	0	[2-5]	[229]
Proline	Pro	2	1	[2-5]|[2-5]	[49]
Proline	Pro	1	1	[2-5]	[4]
Proline	Pro	1	0	[2-5]	[152]
Proline	Pro	1	0	[2-5]	[121]
Serine	Ser	2	1	[2-3]|[2-3]	[20]
Serine	Ser	2	1	[2-3]|[2-3]	[229]
Serine	Ser	2	1	[2-3]|[1-2]	[49]
Serine	Ser	2	1	[2-3]|[1-2]	[4]
Serine	Ser	2	1	[2-3]|[2-3]	[9]
Serine	Ser	3	1	[2-3]|[2-3]|[1-2]	[152]
Serine	Ser	2	1	[2-3]|[1-2]	[121]
Succinate	Suc	0	2		[152]
Threonine	Thr	1	1	[2-4]	[20]
Threonine	Thr	2	1	[2-4]|[3-4]	[4]
Threonine	Thr	2	1	[2-4]	[229]
Threonine	Thr	1	1	[2-4]	[9]
Threonine	Thr	1	1	[2-4]	[152]
Threonine	Thr	1	1	[2-4]	[121]
Trehalose	-	1	0	[1-6]=[7-12]	[20]
Tyrosine	Tyr	1	1	[1-2]	[20]
Tyrosine	Tyr	1	0	[1-2]	[4]
Tyrosine	Tyr	3	0	[2-9]|[2-9]|[1-2]	[49]
Tyrosine	Tyr	1	0	[1-2]	[9]
Tyrosine	Tyr	2	1	[2-9]|[1-2]	[121]
Valine	Val	2	1	[2-5]|[2-5]	[20]
Valine	Val	1	1	[2-5]|[2-5]	[229]
Valine	Val	1	1	[2-5]	[49]
Valine	Val	1	1	[2-5]	[9]
Valine	Val	2	1	[2-5]|[1-2]	[4]
Valine	Val	1	1	[2-5]	[152]
Valine	Val	1	1	[2-5]|[1-2]	[121]

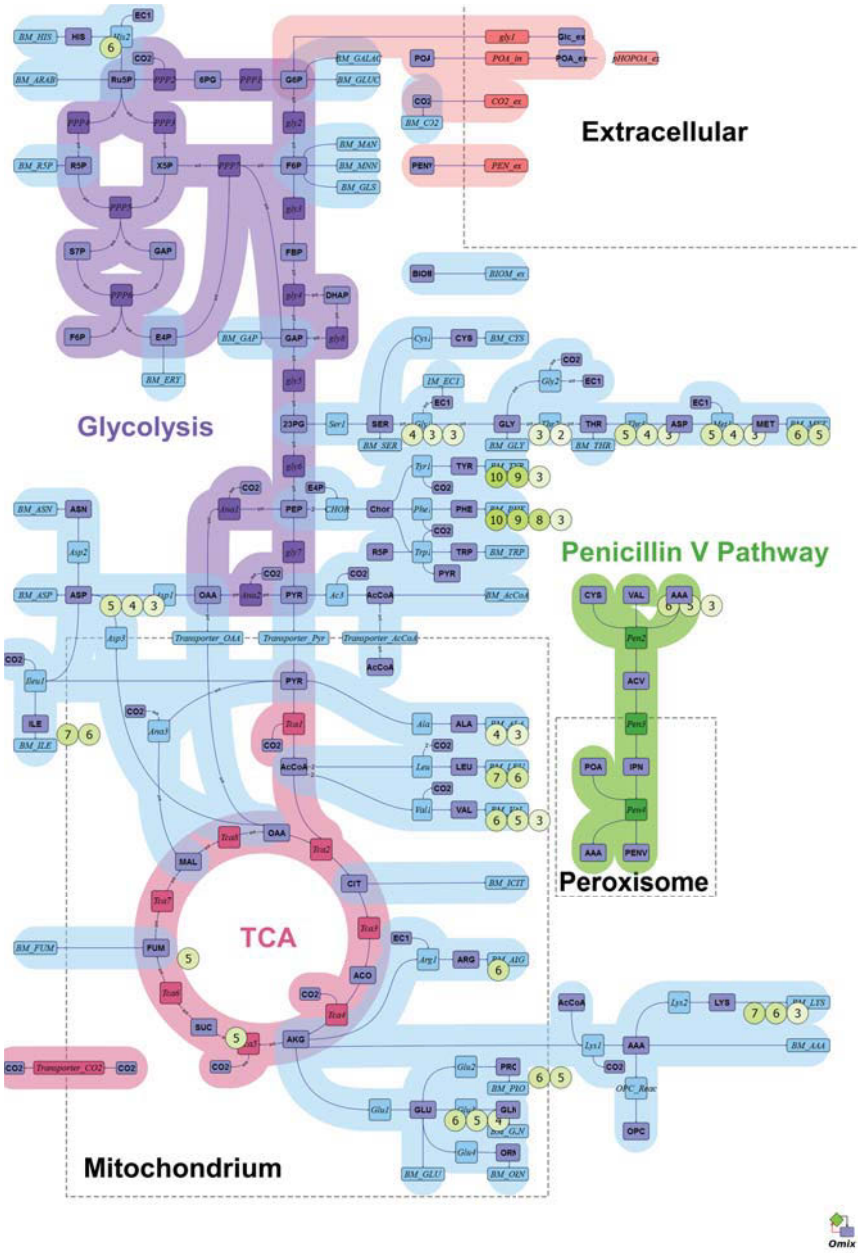


figure E.13: Network with metabolite measurements for GC-MS. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.

LC-MS/MS

Metabolite	Network comp.	Fragment C	Reference
2PG/3PG	2PG3PG		[244]
6-Phosphogluconate	6PG		[244]
Alanine	Ala	[2-3]	[244]
Alanine	Ala		[107]
Arginine	Arg	[1-5]	[244]
Asparagine	Asn	[2-4]	[244]
Asparagine	Asn		[107]
Aspartate	Asp	[1-2]	[244]

Citrate/ Isocitrate	CitIcit	[2-6]	[244]
Cysteine	Cys	[2-3]	[244]
DHAP	DHAP		[244]
Erythrose-4-phosphate	E4P		[244]
Fructose 1,6-bisphosphate	FBP		[244]
Fructose 6-phosphate	F6P		[244]
Fumarate	Fum	[1-3]=[2-4]	[244]
Glucose 6-phosphate	G6P		[244]
GAP	GAP		[244]
Glutamine	Gln		[107]
Glutamine	Gln	[2-5]	[244]
Glutamate	Glu	[2-5]	[244]
Glutamate	Glu		[107]
Glycine	Gly	[1]	[244]
Glycine	Gly		[107]
Glyoxylate	-	[1]=[2]	[244]
Histidine	His	[2-6]	[244]
Homo-serine/ Threonine	Thr	[2-4]	[244]
Isoleucine/ Leucine	Leu	[2-5]	[244]
Lysine	Lys		[244]
Malate	MAL		[244]
Methionine	Met	[2-5]	[244]
Phosphoenol-pyruvate	PEP		[244]
Phenylalanine	Phe	[2-9]	[244]
Phenylalanine	Phe		[107]
Proline	Pro	[2-5]	[244]
Pyruvate	PYR	[2-3]	[244]
Ribose-5-phosphat	R5P		[244]
Sedoheptulose-7-phosphate	S7P		[244]
Serine	Ser	[2-3]	[244]
Serine	Ser		[107]
Succinate	SUC	[1-3]=[2-4]	[244]
Threonine	Thr		[107]
Tryptophan	Trp		[244]
Tyrosine	Tyr	[2-9]	[244]
Tyrosine	Tyr		[107]
Valine	Val	[2-5]	[244]
Valine	Val		[107]
Xylulose-5-phosphat/ Ribulose-5-phosphat	X5P_Ribu5P		[244]
Oxoglutarate	AKG	[1-4]	[244]
Oxaloacetate	OAA	[1-3]	[244]

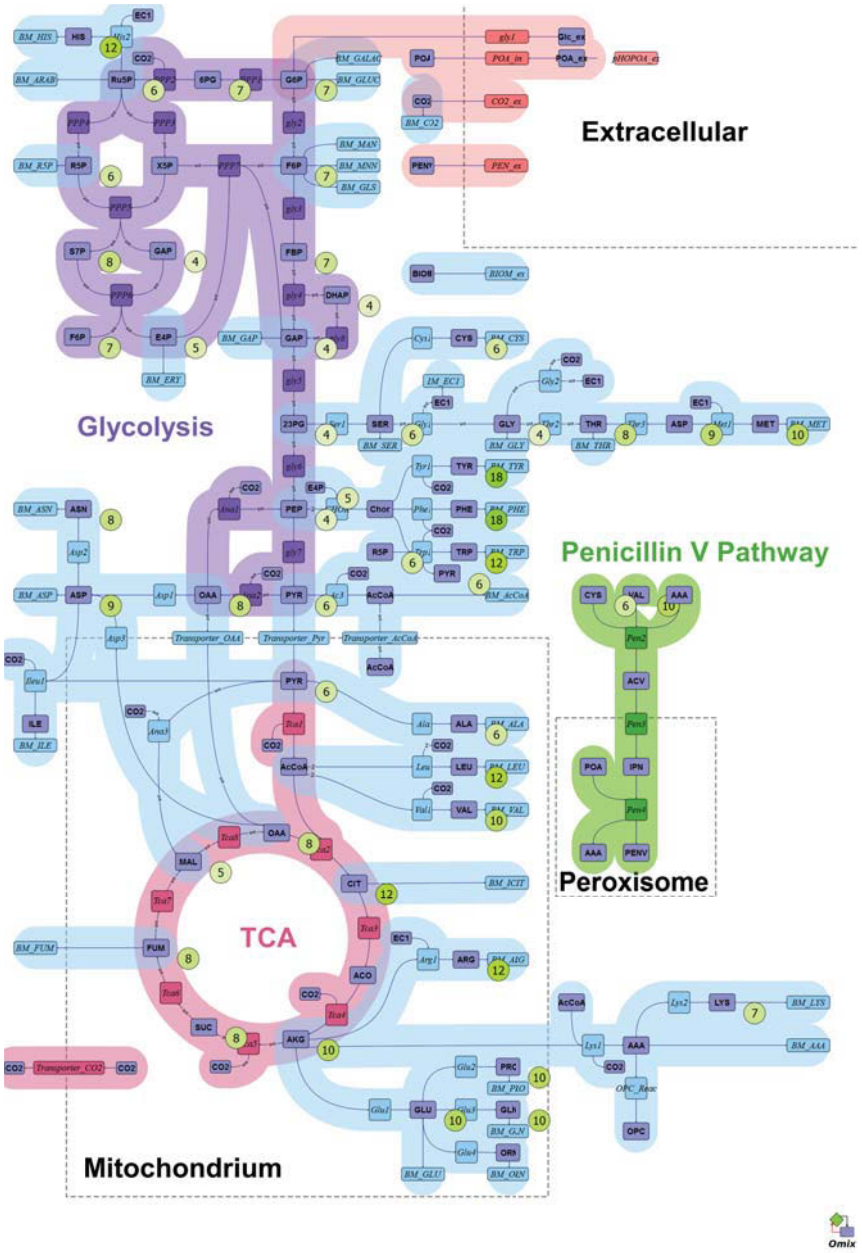


figure E.14: Network with metabolite measurements for LC-MS/MS. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.

2D-NMR

Metabolite	Abbr.	Device	Carbon detected	Measurements	Reference
Alanine	Ala	HSQC NMR	[2] | [3]	[s,da,db,dd] | [s,da]	[82]
Alanine	Ala	13C–13C COSY NMR	[2] | [3]	[s,da,db,dd] | [s,da]	[214]
Alanine	Ala	[13C,1H] COSY NMR	[2] | [3]	[s,da,db,dd] | [s,da]	[121]
Alanine	Ala	[13C, 1H] HSQC NMR	[2] | [3]	[s,da,db,dd] | [s,da]	[38]
Arginine	Arg	13C–13C COSY NMR	[3] | [4] | [5]	[s,da,dd] | [s,da,dd] | [s,da]	[214]
Arginine	Arg	[13C,1H] COSY NMR	[3] | [4]	[s,da,t] | [s,da]	[121]
Arginine	Arg	[13C, 1H] HSQC NMR	[3] | [5]	[s,da,t] | [s,da]	[38]
Aspartate	Asp	[13C, 1H] HSQC NMR	[2] | [3]	[s,da,db,dd] | [s,da,db,dd]	[82]
Aspartate	Asp	[13C,1H] COSY NMR	[2] | [3]	[s,da,db,dd] | [s,da,db,dd]	[121]
Aspartate	Asp	[13C, 1H] HSQC NMR	[2] | [3]	[s,da,db,dd] | [s,da,db,dd]	[38]
Asx	-	13C–13C COSY NMR	[2] | [3]	[s,da,db,dd] | [s,da,db,dd]	[214]

Glutamate	Glu	[13C, 1H] HSQC NMR	[2]|[3]|[4]	[s,da,db,dd]|[s,da,dd]|[s,da,db,dd]	[82]
Glutamate	Glu	[13C,1H] COSY NMR	[2]|[3]|[4]	[s,da,db,dd]|[s,da,t]|[s,da,db,dd]	[121]
Glycine	Gly	[13C, 1H] HSQC NMR	[2]	[s,da]	[82]
Glycine	Gly	13C–13C COSY NMR	[2]	[s,da]	[214]
Glycine	Gly	[13C,1H] COSY NMR	[2]	[s,da]	[121]
Glycine	Gly	[13C,1H] HSQC NMR	[2]	[s,da]	[38]
Glx	-	13C–13C COSY NMR	[2]|[3]|[4]	[s,da,db,dd]|[s,da,dd]|[s,da,db,dd]	[214]
Histidine	His	[13C, 1H] HSQC NMR	[2]|[3]|[5]	[s,da,db,dd]|[s,da,db,dd]|[s,da]	[82]
Histidine	His	13C–13C COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da,db,dd]	[214]
Histidine	His	[13C,1H] COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da,db,dd]	[121]
Histidine	His	[13C, 1H] HSQC NMR	[3]|[5]	[s,da,db,dd]|[s,da]	[38]
Isoleucine	Ile	13C–13C COSY NMR	[2]|[4]|[5]|[6]	[s,da,db,dd]|[s,da,dd]|[s,da]|[s,da]	[214]
Isoleucine	Ile	[13C, 1H] HSQC NMR	[2]	[s,da,db,dd]	[82]
Isoleucine	Ile	[13C,1H] COSY NMR	[3]|[5]	[s,da,t]|[s,da]	[121]
Isoleucine	Ile	[13C, 1H] HSQC NMR	[2]|[4]|[5]|[6]	[s,da,db,dd]|[s,da,t]|[s,da]|[s,da]	[38]
Leucine	Leu	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da,dd]	[82]
Leucine	Leu	13C–13C COSY NMR	[2]|[3]|[5]|[6]	[s,da,db,dd]|[s,da,dd]|[s,da]|[s,da]	[214]
Leucine	Leu	[13C,1H] COSY NMR	[2]|[3]|[5]|[6]	[s,da,db,dd]|[s,da,t]|[s,da]|[s,da]	[121]
Leucine	Leu	[13C, 1H] HSQC NMR	[2]|[3]|[5]|[6]	[s,da,db,dd]|[s,da,t]|[s,da]|[s,da]	[38]
Lysine	Lys	13C–13C COSY NMR	[3]|[4]|[5]|[6]	[s,da,dd]|[s,da,dd]|[s,da,dd]|[s,da]	[214]
Lysine	Lys	[13C,1H] COSY NMR	[3]|[4]	[s,da,t]|[s,da,t]	[121]
Lysine	Lys	[13C, 1H] HSQC NMR	[3]|[4]|[5]	[s,da,t]|[s,da,t]|[s,da,t]	[38]
Methionine	Met	[13C,1H] COSY NMR	[2]	[s,da,db,dd]	[121]
Methionine	Met	13C–13C COSY NMR	[2]	[s,da,db,dd]	[214]
Phenylalanine	Phe	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da,dd]	[82]
Phenylalanine	Phe	13C–13C COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da,db,dd]	[214]
Phenylalanine	Phe	[13C,1H] COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da,db,dd]	[121]
Phenylalanine	Phe	[13C, 1H] HSQC NMR	[2]	[s,da,db,dd]	[38]
Proline	Pro	[13C, 1H] HSQC NMR	[2]|[5]	[s,da,db,dd]|[s,da]	[82]
Proline	Pro	13C–13C COSY NMR	[2]|[3]|[4][5]	[s,da,db,dd]|[s,da,dd]|[s,da,dd]|[s,da]	[214]
Proline	Pro	[13C,1H] COSY NMR	[2]|[4]	[s,da,db,dd]|[s,da]	[121]
Proline	Pro	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da,t]	[38]
Serine	Ser	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da]	[82]
Serine	Ser	13C–13C COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da]	[214]
Serine	Ser	[13C,1H] COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da]	[121]
Serine	Ser	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da]	[38]
Threonine	Thr	13C–13C COSY NMR	[2]|[3]|[4]	[s,da,db,dd]|[s,da,dd]|[s,da]	[214]
Threonine	Thr	[13C,1H] COSY NMR	[2]|[3]|[4]	[s,da,db,dd]|[s,da,t]|[s,da]	[121]
Threonine	Thr	[13C, 1H] HSQC NMR	[4]	[s,da]	[82]
Threonine	Thr	[13C, 1H] HSQC NMR	[4]	[s,da]	[38]
Tyrosine	Tyr	[13C, 1H] HSQC NMR	[5]|[6]	[s,da,dd]|[s,da,dd]	[82]
Tyrosine	Tyr	13C–13C COSY NMR	[2]|[3]	[s,da,db,dd]|[s,da,db,dd]	[214]
Tyrosine	Tyr	[13C,1H] COSY NMR	[2]|[3]	[s,da,db,dd]|[s, da,db,dd]	[121]
Tyrosine	Tyr	[13C, 1H] HSQC NMR	[3]|[5 from pep]	[s,da,t]|[s,da,t]|	[38]
Uridine	-	[13C, 1H] HSQC NMR	[4]|[5]	[s,da]|[s,da]	[82]
Valine	Val	[13C,1H] COSY NMR	[2]|[4]|[5]	[s,da,db,dd]|[s,da]|[s,da]	[121]
Valine	Val	[13C, 1H] HSQC NMR	[2]|[3]	[s,da,db,dd]|[s,da,dd,ddd]	[82]
Valine	Val	13C–13C COSY NMR	[2]|[4]|[5]	[s,da,db,dd]|[s,da]|[s,da]	[214]
Valine	Val	[13C, 1H] HSQC NMR	[2]|[4]|[5]	[s,da,db,dd]|[s,da]|[s,da]	[38]

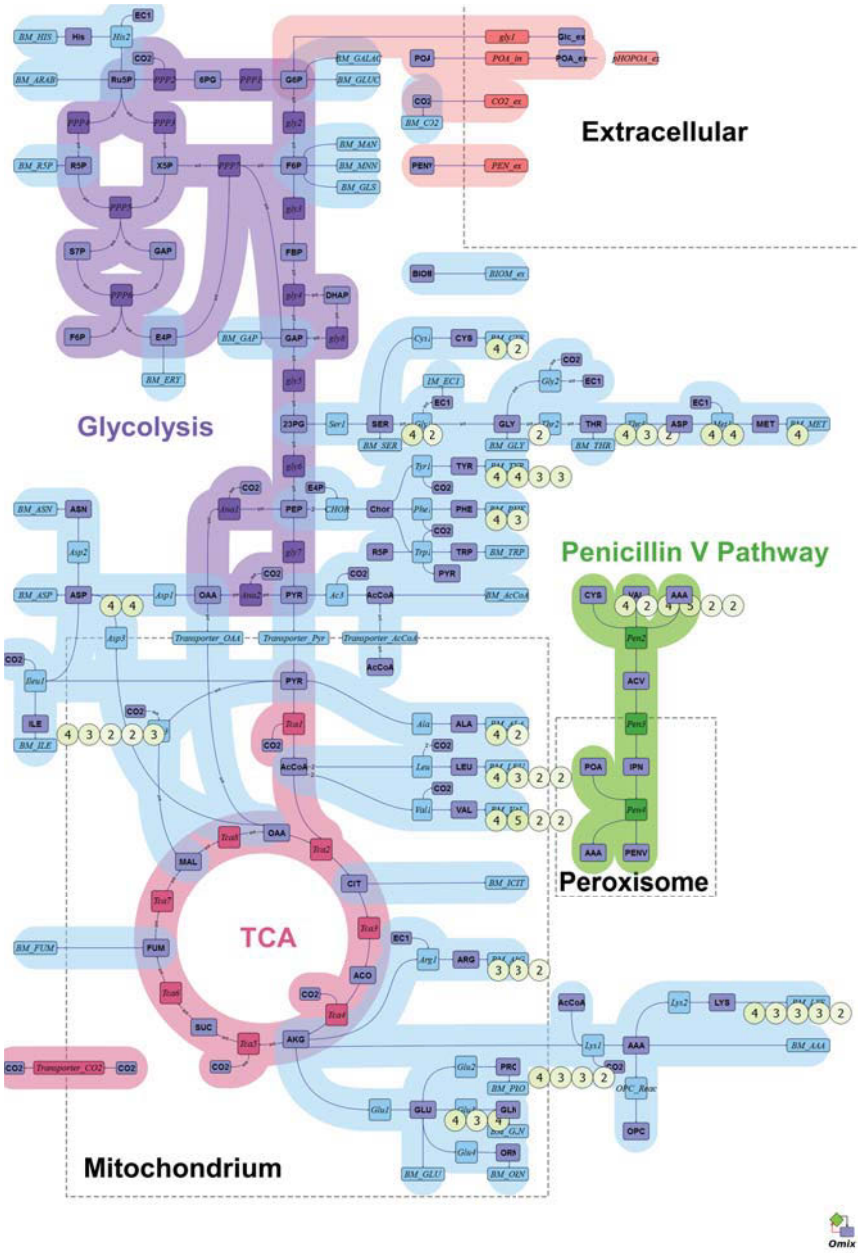


figure E.15: Network with metabolite measurements for CNMR. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.

LC-MS

Metabolite	Network comp.	Device	Reference
2PG/3PG	2PG3PG	LC-MS	[212]
2PG/3PG	2PG3PG	LC-MS	[121]
2PG/3PG	2PG3PG	LC-MS	[248]
6-Phosphogluconate	6PG	LC-MS	[212]
6-Phosphogluconate	6PG	LC-MS	[121]
6-Phosphogluconate	6PG	LC-MS	[248]
DHAP/GAP	DHAPGAP	LC-MS	[212]

Erythrose-4-phosphate	E4P	LC-MS	[248]
Fructose 6-phosphate	F6P	LC-MS	[248]
Fructose 6-phosphate	F6P	LC-MS	[121]
Fructose 1,6-bisphosphate	FBP	LC-MS	[212]
Fructose 1,6-bisphosphate	FBP	LC-MS	[121]
Fructose 1,6-bisphosphate	FBP	LC-MS	[248]
Glucose 1-phosphate	-	LC-MS	[248]
Glucose 6-phosphate	G6P	LC-MS	[121]
Glucose 6-phosphate	G6P	LC-MS	[248]
M6P	-	LC-MS	[121]
P5P	-	LC-MS	[212]
P5P	-	LC-MS	[121]
P5P	-	LC-MS	[248]
Phosphoenolpyruvate	PEP	LC-MS	[212]
Phosphoenolpyruvate	PEP	LC-MS	[121]
Phosphoenolpyruvate	PEP	LC-MS	[248]
Sedoheptulose-7-Phosphat	S7P	LC-MS	[248]
Sedoheptulose-7-Phosphat	S7P	LC-MS	[121]

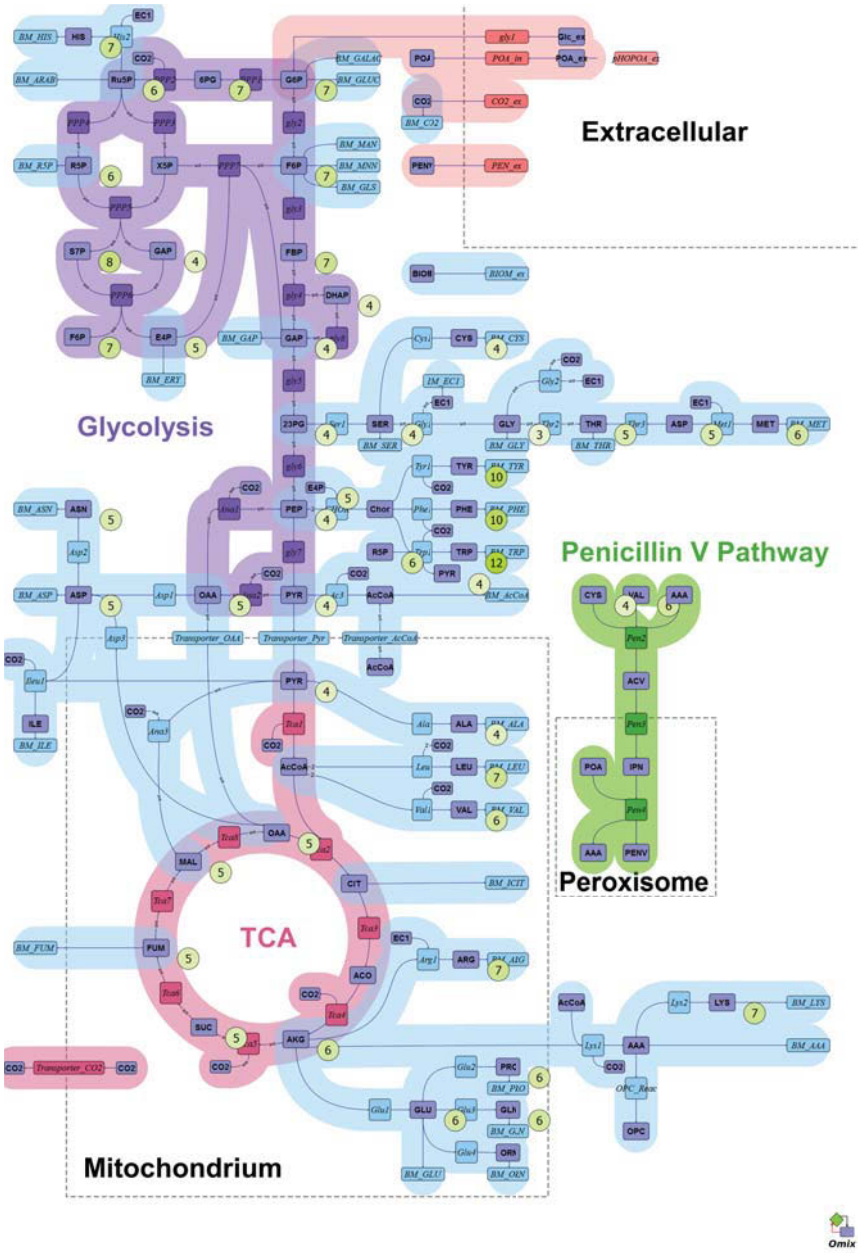


figure E.16: Network with metabolite measurements for LC-MS. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.

HNMR

Metabolite	Network comp.	Carbon detected	Reference
Acetate	-	[1]||[2]	[152]
Alanine	Ala	[2]	[3]	[152]
Glycine	Gly	[1]	[2]	[149]
Serine	Ser	[2]	[3]	[149]
Alanine	Ala	[2]	[3]	[149]
Valine	Val	[4]	[5]	[149]
Isoleucine	Ile	[2]	[5]	[6]	[149]
Leucine	Leu	[2]|[3]|[4]|[5]|[6]	[149]

Glutamate	Glu	[2]|[3]|[4]	[149]
Aspartate	Asp	[2]|[3]	[149]
Succinate	Suc	[1]=[4]|[2]=[3]	[152]
Threonine	Thr	[2]|[3]|[4]	[149]
Phenylalanine	Phe	[2]|[3]|[5]|[6]|[7]|[8]|[9]	[149]
Lysine	Lys	[2]|[3]|[4]|[5]|[6]	[149]

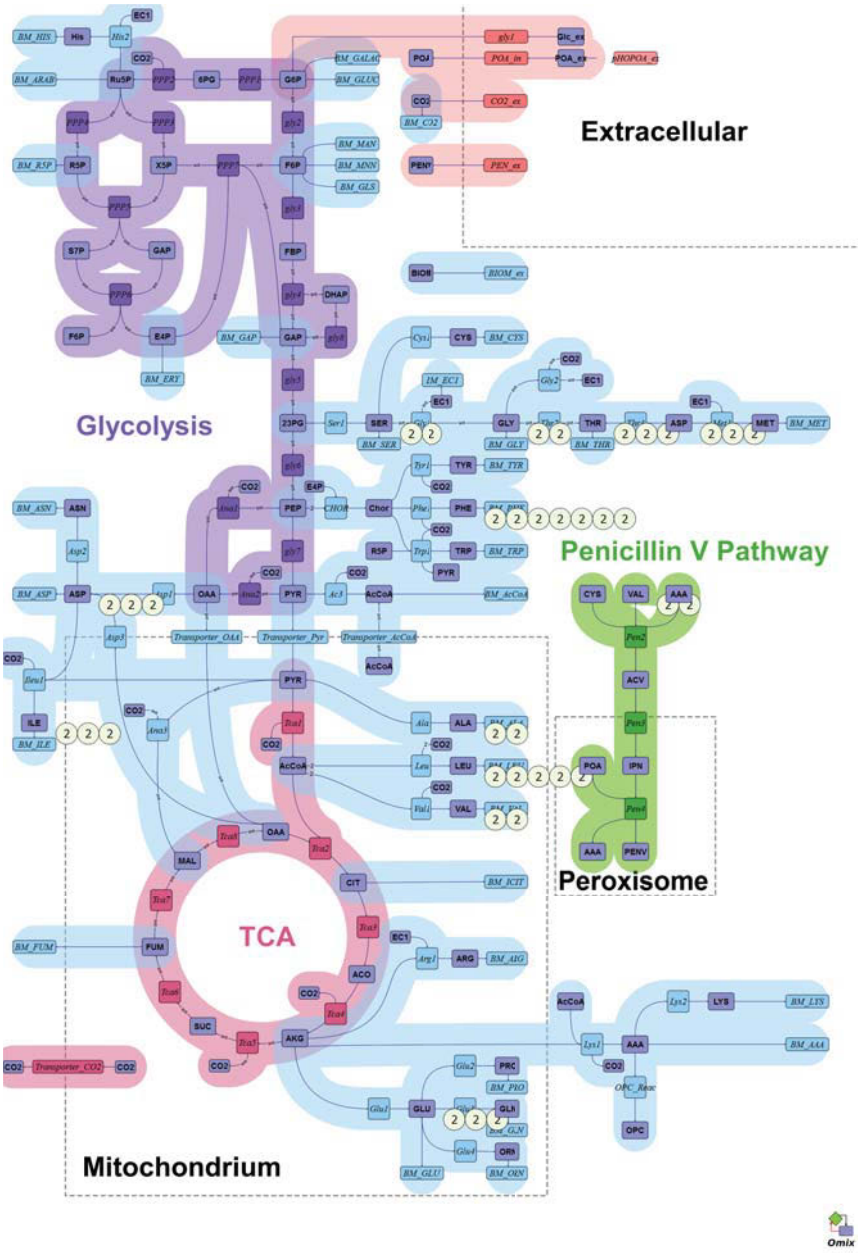


figure E.17: Network with metabolite measurements for HNMR. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.

GC-C-IRMS

Metabolite	Network comp.	Fragment C	Reference
Aspartate	Asp		[278]
Glutamate	Glu		[278]
Glycine	Gly		[278]
Histidine	His		[278]
Isoleucine	Ile	[2-5]	[278]
Leucine	Leu	[2-6]	[278]
Phenylalanine	Phe		[278]
Proline	Pro		[278]
Threonine	Thr		[278]

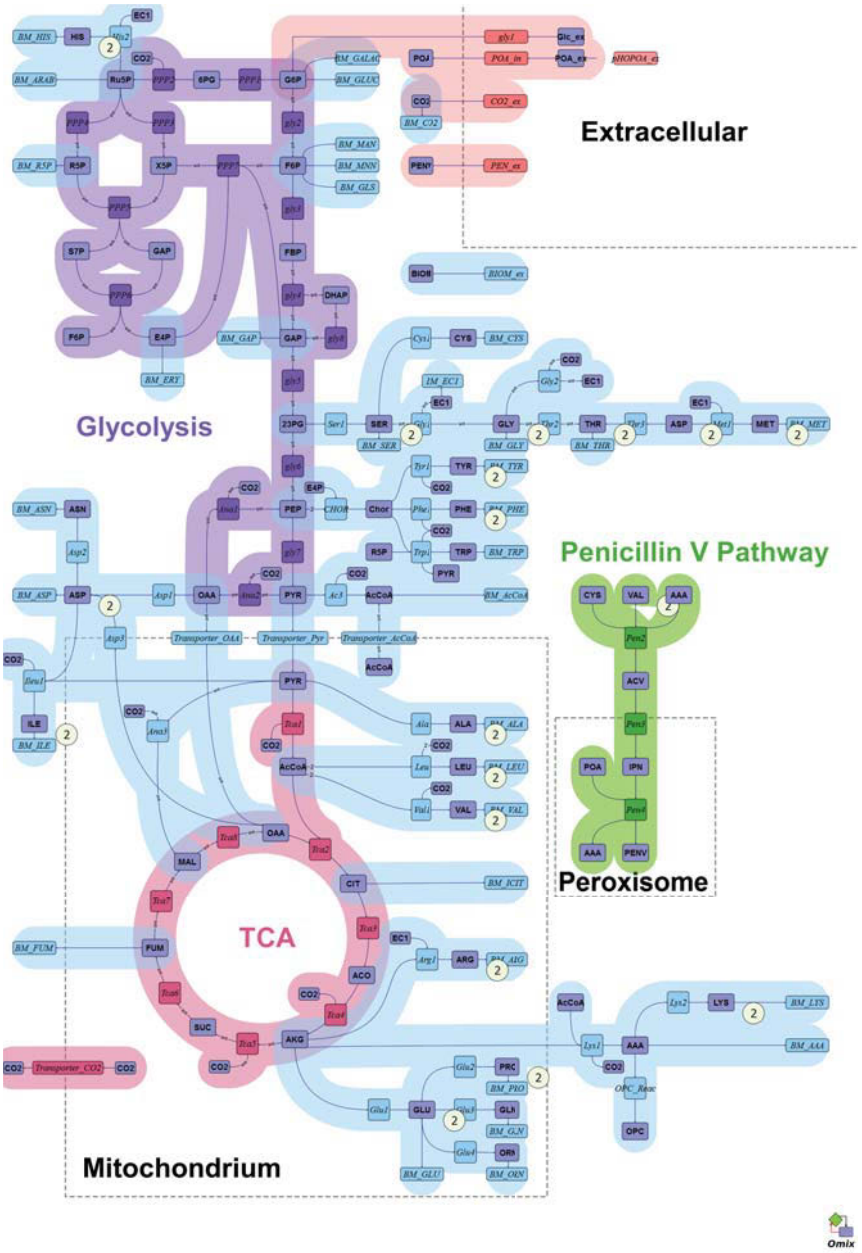


figure E.18: Network with metabolite measurements for GC-C-IRMS. Each circle corresponds to a measurement group. For each measurement group the number of single measurement values is given.