



Technische Universität München

Fachgebiet für Systembiotechnologie

Development of a Theoretical Workflow  
for Metabolic Engineering and its  
Application to Terpenoid Production in  
*Escherichia coli*

Miguel Ángel Valderrama Gómez

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines Doktor-Ingenieurs genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Dirk Weuster-Botz  
Prüfer der Dissertation: 1. Prof. Dr.-Ing. Andreas Kremling  
2. Prof. Dr. rer. nat. Christoph Kaleta

Die Dissertation wurde am 10.04.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 12.07.2018 angenommen.



# Contents

<b>Acknowledgments</b>	<b>xi</b>
<b>Abstract</b>	<b>xiii</b>
<b>1. Introduction</b>	<b>1</b>
<b>I. Theory</b>	<b>5</b>
<b>2. Metabolic Modeling &amp; Strain Engineering</b>	<b>7</b>
2.1. Constraint-based Methods . . . . .	8
2.1.1. Dynamic Flux Balance Analysis . . . . .	10
2.1.2. Experimental Determination of Reaction Rates . . . . .	12
2.2. Elementary modes-based Methods . . . . .	16
2.3. Kinetic-based Methods . . . . .	18
2.3.1. Ensemble Modeling Approach of Metabolic Networks . . . . .	18
2.4. Strain Design Algorithms . . . . .	24
2.5. Computational Tools . . . . .	28
<b>3. Theoretical Workflow for Metabolic Engineering</b>	<b>31</b>
3.1. Model-guided Strain Engineering . . . . .	31
3.2. Model-guided Process Optimization . . . . .	32
<b>II. Results: Strain Engineering</b>	<b>35</b>
<b>4. Metabolic Burden</b>	<b>39</b>
4.1. Strains & Experimental Data . . . . .	39
4.2. Substrate Quality . . . . .	41
4.2.1. Conversion Factors for Substrate Uptake Standardization . . . . .	42
4.2.2. Data Standardization . . . . .	47
4.3. Yield, Acetate and Formate Lines . . . . .	48
4.4. Process-level Strategy to Reduce Detrimental Effects of Metabolic Burden . . . . .	52
4.5. Discussion . . . . .	54
<b>5. Strain Design Algorithms for Target Identification</b>	<b>57</b>
5.1. Strains & Experimental Data . . . . .	57
5.2. Reduction of Metabolic Burden Through Genomic Integration . . . . .	58
5.3. Constraint-based Assessment of Taxadiene Production Potential . . . . .	61
5.4. Application of Strain-design Algorithms for Target Identification . . . . .	61
5.4.1. Constraint-based Methods: The OptKnock Algorithm . . . . .	62

## Contents

5.4.2. Kinetic-based Methods: The OptEM Algorithm . . . . .	65
5.5. Discussion . . . . .	74
<b>6. Simultaneous Utilization of D-Xylose &amp; Glucose in <i>E. coli</i></b>	<b>77</b>
6.1. Strains & Experimental Data . . . . .	77
6.2. Constraint-based Characterization of Simultaneous Sugar Utilization . . . . .	79
6.2.1. Biomass Yield & Growth Rate . . . . .	80
6.2.2. Acetate Lines & Metabolic Burden . . . . .	82
6.3. Assessing Taxadiene Production Potential . . . . .	85
6.4. Model-based Identification of Optimal Production Strain Candidates . . . . .	90
6.5. Discussion . . . . .	90
<b>III. Results: Process Optimization</b>	<b>95</b>
<b>7. Model-based Medium Optimization</b>	<b>99</b>
7.1. Performance Criteria for Substrate Selection . . . . .	99
7.1.1. Single Substrates . . . . .	100
7.2. Synergy as Design Principle for Pathway Selection . . . . .	102
7.2.1. Three-Substrate Mixtures . . . . .	103
7.3. Discussion . . . . .	107
<b>8. Model-based Semi-Batch Optimization</b>	<b>109</b>
8.1. Strain & Experimental Data . . . . .	109
8.2. Mathematical Identification of Characteristic Process Phases and Rates Calculation . . . . .	111
8.3. Dependence of Model Parameters on Production Temperature . . . . .	113
8.4. <i>In silico</i> Optimization of Production Temperature and Time Point of Induction	116
8.5. Discussion . . . . .	118
<b>9. Product Yield Prediction</b>	<b>121</b>
9.1. Strain & Experimental Data . . . . .	121
9.2. Constraint-based Product Yield Prediction . . . . .	122
9.3. Kinetic-based Product Yield Prediction . . . . .	123
9.4. Application of KAY to Predict Optimal Biomass Yield During Production .	124
9.5. Discussion . . . . .	126
<b>10. Concluding Remarks and Outlook</b>	<b>129</b>
10.1. Lessons From Succinate Production in Engineered Strains . . . . .	129
10.2. Lessons From Taxadiene Production in <i>Escherichia coli</i> ( <i>E. coli</i> ) . . . . .	130
<b>Acronyms</b>	<b>143</b>
<b>A. Appendix for Strain Engineering</b>	<b>145</b>
A.1. Metabolic Burden . . . . .	145
A.1.1. Slope of Acetate Line . . . . .	145
A.1.2. Amino Acid Sequences . . . . .	146
A.1.3. Concentration Time Courses . . . . .	147
A.1.4. Flux Distribution Patters for Three Zones . . . . .	147

A.2. Strain Design Algorithms for Target Identification . . . . .	152
A.2.1. Mevalonate, Non-Mevalonate & Taxadiene Formation Pathways . . .	152
A.2.2. Concentration Time Courses for Taxadiene Producing Strains . . . .	152
A.2.3. $^{13}\text{C}$ Labeling Data . . . . .	152
A.2.4. Metabolic Model Used for the Software 13CFlux2 . . . . .	156
A.2.5. Constraint-based Assessment of Taxadiene Production Potential . .	166
A.3. Simultaneous Utilization of D-Xylose and Glucose in <i>E. coli</i> . . . . .	166
A.3.1. Concentration Time Courses for All Strains . . . . .	166
A.3.2. $^{13}\text{C}$ Labeling Data . . . . .	166
A.3.3. Dependency of Oxygen Uptake Rate on Sugar Co-Utilization . . . .	169
A.3.4. Experimental Protein Content . . . . .	169
<b>B. Appendix for Process Engineering</b>	<b>171</b>
B.1. Model-based Medium Optimization . . . . .	171
B.1.1. Extending Metabolic Models to Allow Taxadiene Synthesis . . . . .	171
B.1.2. Ranking of 180 Substrates According to Yields . . . . .	173
B.1.3. Ranking of 180 Substrates According to Profit . . . . .	175
B.2. Model-based Fed-batch Optimization . . . . .	178
B.2.1. Concentration Time Courses for Process Development . . . . .	178
B.2.2. Temperature Dependencies of Phases Duration and Exchange Rates	178
B.2.3. Modeling Concentration Time Courses for All Fermentations . . . .	178



# List of Tables

2.1. Genome-Scale Metabolic Reconstructions for <i>E. coli</i> . . . . .	9
2.2. Overview of Variables Used to Describe Reactor Dynamics . . . . .	11
2.3. Substrate-level Enzyme Regulation Considered for the <i>E. coli</i> Core Model .	22
2.4. Theoretical Approaches Used in the <i>in silico</i> Strain Optimization for the Succinate Overproduction . . . . .	24
2.5. Commonly Used Computational Tools for the Analysis of Metabolic Networks	29
4.1. Experimentally Calculated Exchange Rates for Three Strains. Growth on Glucose. . . . .	40
4.2. Conversion Factors for Three Quality Criteria . . . . .	43
4.3. Substrate Uptake Rates Used to Assess Quality Criteria . . . . .	45
5.1. Taxadiene Producing Strains . . . . .	58
5.2. Taxadiene Producing Strains: Exchange Rates . . . . .	59
5.3. Objective Functions for OptEM . . . . .	68
5.4. Enzyme Perturbations Identified by OptEM Leading to an Increased Tax- adiene Yield in Strain $\Delta XylE\_1p$ . . . . .	70
5.5. Strain Design Strategy Identified by OptEM for the Maximization of Tax- adiene Yield in Strain $\Delta XylE\_1p$ Using a Solver Approach . . . . .	72
6.1. Strains Used in Sugar Co-utilization Studies . . . . .	78
6.2. Experimental Characterization of Strains Used in Sugar Co-utilization Studies	78
6.3. Strain Performance Parameters . . . . .	79
7.1. Metabolic Models Used for Synergy Studies . . . . .	102
8.1. Experimental Process Parameters for Six Semi-Batch Fermentations . . . . .	110
8.2. Growth, Production and Substrate Uptake Rates for Different Process Phases and Production Temperatures . . . . .	112
9.1. Biomass and Product Yields for Six Semi-Batch Fermentations . . . . .	122
9.2. Average Biomass and Product Yields for Four Different Production Tem- peratures . . . . .	122
A.1. Concentration Time Courses for Galactose as Carbon Source . . . . .	147
A.2. Concentration Time Courses for Glucose as Carbon Source . . . . .	147
A.3. Concentration Time Courses for Mannose as Carbon Source . . . . .	147
A.4. Concentration Time Courses for Fructose as Carbon Source . . . . .	148
A.5. Time Courses for Acetate as Carbon Source . . . . .	148
A.6. Exchange Rates for Three Strains and Five Sugars . . . . .	148
A.7. Concentration Time Courses for Taxadiene Producing Strains . . . . .	154
A.8. $^{13}C$ -labeling Data for Five Taxadiene Producing Strains . . . . .	155

List of Tables

A.9. Confidence Intervals for Taxadiene Production Flux Calculated Considering <sup>13</sup> C Labeling Data and Exchange Rates . . . . .	166
A.10. Concentration Time Courses For Six Strains . . . . .	167
A.11. Concentration Time Courses For Parent Strains: Re-measurement . . . . .	167
A.12. <sup>13</sup> C-labeling Data for Strains Used in Sugar Co-Utilization Studies . . . . .	168
B.1. Maximal Taxadiene Yields Calculated for 180 Carbon Sources . . . . .	173
B.2. Maximal Profit Calculated for 180 Carbon Sources . . . . .	175
B.3. Concentration Time Courses for Batch ID 1 . . . . .	178
B.4. Concentration Time Courses for Batch ID 2 . . . . .	179
B.5. Concentration Time Courses for Batch ID 3 . . . . .	179
B.6. Concentration Time Courses for Batch ID 4 . . . . .	180
B.7. Concentration Time Courses for Batch ID 5 . . . . .	180
B.8. Concentration Time Courses for Batch ID 6 . . . . .	181
B.9. Formulas Used to Describe the Effect of Production Temperature on Phase Duration . . . . .	181
B.10. Formulas Used to Describe the Effect of Production Temperature on Exchange Rates . . . . .	182
B.11. Determination Coefficients for All Batches . . . . .	183

# List of Figures

1.1. Motivation & Project's Aim . . . . .	2
1.2. Modeling Workflow . . . . .	3
2.1. Dynamic Flux Balance Analysis . . . . .	10
2.2. Experimental Determination of Cellular Growth Rates . . . . .	13
2.3. Experimental Determination of Glucose Uptake Rate . . . . .	14
2.4. Statistical Workflow . . . . .	17
2.5. Ensemble Forecasting for Hurricane Path Prediction . . . . .	19
2.6. Parallelization Strategy in Linux Cluster . . . . .	23
3.1. The Workflow . . . . .	33
4.1. Strains Used for the Characterization of the Metabolic Burden. . . . .	40
4.2. Growth Rate vs. Substrate Uptake Rate for Three Strains and Five Substrates. . . . .	41
4.3. Real-Life Analogy for Concept of Substrate Quality . . . . .	42
4.4. Energy Content As a Function of Oxygen Availability for Five Substrates . . . . .	43
4.5. Assessment of Three Different Quality Criteria for Data Standardization. High Oxygen Availability . . . . .	45
4.6. Assessment of Three Different Quality Criteria for Data Standardization at Low Oxygen Availability . . . . .	46
4.7. Standardized versus Raw Experimental Data . . . . .	47
4.8. Load-induced Oxygen Limitation . . . . .	49
4.9. Relationship between Biomass Yield, Acetate and Formate Lines . . . . .	50
4.10. Qualitative Comparison Between Experimental Data and <i>in silico</i> Calculations . . . . .	51
4.11. Process-level Strategy to Reduce Metabolic Burden . . . . .	53
4.12. Emergence of Yield and Acetate Lines By Constraining Flux Through Res- piratory Chain . . . . .	55
5.1. Taxadiene Producing Strains . . . . .	59
5.2. Acetate Lines for Taxadiene Producing Strains . . . . .	60
5.3. Constraint-based Assessment of Production Potential . . . . .	62
5.4. Application of OptKnock to the Production of Taxadiene in <i>E. coli</i> : Basic Idea . . . . .	63
5.5. Application of OptKnock to the Production of Taxadiene in <i>E. coli</i> : De- pendence on Growth Rate and Oxygen Availability. . . . .	64
5.6. Application of OptKnock to the Production of Taxadiene in <i>E. coli</i> : Targets. . . . .	65
5.7. Overview of the Main Steps Comprising the OptEM Strain Design Algorithm . . . . .	66
5.8. Effect of Single Enzyme Up- and Down-regulation Predicted by OptEM for the Maximization of Taxadiene Yield in Strain $\Delta XylE\_1p$ via a Brute Force Approach. . . . .	69

List of Figures

5.9. Strain Design Strategy Identified by OptEM for the Maximization of Taxadiene Yield in Strain $\Delta$ XylE_1p Using a Solver Approach. . . . .	73
6.1. Strain Performance Parameters as a Function of Simultaneous Sugar Co-utilization Extent . . . . .	80
6.2. <i>In silico</i> Calculated Effect of Sugar Co-Utilization on Growth Rate and Biomass Yield: Case Study I . . . . .	81
6.3. <i>In silico</i> Calculated Effect of Sugar Co-Utilization on Growth Rate and Biomass Yield: Case Study II . . . . .	82
6.4. Expected Shift in Acetate Lines as a Result of Plasmid-based XylE-Expression . . . . .	83
6.5. Effect of Glucose and Xylose Co-utilization on Growth Rate and Its Associated Metabolic Burden . . . . .	84
6.6. Effect of Glucose and Xylose Co-utilization on Growth Rate and Its Associated Metabolic Burden. Re-measurement of Selected Strains . . . . .	84
6.7. Assessing Taxadiene Production Potential by a Constraint-based Approach . . . . .	86
6.8. Assessing Taxadiene Production Potential by a Kinetic-based Approach . . . . .	87
6.9. Dependence of Taxadiene Production Potential on % Carbon Xylose . . . . .	88
6.10. Dependence of Taxadiene Production Potential on Growth Characteristics . . . . .	89
6.11. Model-based Identification of Optimal Production Strain Candidates . . . . .	91
6.12. Substrate Quality: Glucose vs Xylose for Different Oxygen to Substrate Ratios . . . . .	92
7.1. Substrate Ranking via Four Different Performance Criteria . . . . .	101
7.2. Synergy as Design Principle for Pathway Selection. Glucose As a Case Study. . . . .	104
7.3. Synergy as Design Principle for Pathway Selection. 180 Substrates . . . . .	105
7.4. Maximal Product Yield for Three-Substrate Mixtures . . . . .	106
8.1. Process Set-up and Experimental Data . . . . .	110
8.2. Modeling Experimental Data: Characteristic Process Times . . . . .	110
8.3. Modeling Experimental Data: Mathematical Identification of Process Phases . . . . .	112
8.4. Modeling Experimental Data: Results . . . . .	112
8.5. Dependence of Characteristic Process Times on Temperature and Time Point of Induction . . . . .	114
8.6. Modeling Experimental Data: Temperature Dependence of Duration of Process Phases . . . . .	114
8.7. Modeling Experimental Data: Temperature Dependence of Exchange Rates . . . . .	115
8.8. Volumetric Productivity for Batches 2 and 5 . . . . .	116
8.9. Model-based Optimization of Process Parameters . . . . .	117
9.1. Constraint-based Product Yield Prediction . . . . .	123
9.2. Kinetically Accessible Yield (KAY) for Experimental Product Yield Prediction . . . . .	125
9.3. Dependence of Maximal Product Yield on Biomass Yield . . . . .	126
9.4. Increasing the Utility of the KAY Concept . . . . .	127
10.1. Application of the Workflow to the Production of Taxadiene in <i>E. coli</i> : Conclusion . . . . .	133
A.1. Determination of the Acetate Line Slope for <i>E. coli</i> HMS . . . . .	145
A.2. Flux Variability Analysis for Phase I . . . . .	149
A.3. Flux Variability Analysis for Phase II . . . . .	150

A.4. Flux Variability Analysis for Phase III . . . . .	151
A.5. Non-Mevalonate Pathway (MEP) . . . . .	152
A.6. Mevalonate Pathway (MVA) . . . . .	153
A.7. Taxadiene Production Pathway . . . . .	153
A.8. Estimation of Oxygen Uptake Rate as a Function of Sugar Co-Uptake . . .	169
A.9. Protein Measurement. . . . .	170
B.1. Modeling Time Course Concentrations: Batch 2 . . . . .	182
B.2. Modeling Time Course Concentrations: Batch 3 . . . . .	182
B.3. Modeling Time Course Concentrations: Batch 4 . . . . .	183
B.4. Modeling Time Course Concentrations: Batch 5 . . . . .	183
B.5. Modeling Time Course Concentrations: Batch 6 . . . . .	183

千里之行，始於足下

A journey of a thousand miles begins  
with a single step

*Laozi*

# Acknowledgments

I *love* sayings. That is why I invented this one: “man promoviert nicht an einem einzigen Tag”, which would be translated into English as: “you don’t do a PhD in just one single day”. Thinking about all the people who supported me in some way in these three and a half years, I realized I need to extend my saying: “you don’t do a PhD in just one single day, and you cannot do it alone, so get some help!”. I do not pretend this saying to be universal, but it describes my situation and the way I see the process of getting a PhD pretty well.

First, I would like to start by thanking my Doktorvater, Prof. Dr.-Ing. Andreas Kremling for providing me the life-changing opportunity of doing a PhD at his group, at the Technical University of Munich (TUM). Thank you for being so patient with me, especially in the first months of my PhD, when I spoke such a broken German. Thank you for letting me discover so many parts of the world, and for allowing me to do the three months research visit at the University of California, Los Angeles, which vastly enriched me personally and professionally. During the time of my PhD, I even learned how to ski by attending our legendary winter seminar in the German Alps once in a year - which is pretty cool, given the fact that in Colombia we do not really have snow-. I feel so grateful for all that, thank you Prof. Kremling!

Next, I would like to thank Michael Frank Panfil for his constant support and guidance. Many of my crazy projects would have not been even thinkable without your unconditional help. From the very first idea of doing my master studies in Germany, to actually finishing my PhD, you have positively shaped much of what I am today. You made Munich my new home town. You were always there for me when I needed you. For that, I will be in debt with you my whole life. Thank you so much Frank!

My special thanks to Boyu Yang -BB- for patiently reading all pages of this dissertation and for providing grammatical advice, for constantly correcting my English. Thank you for sharing your love for Los Angeles with me, and for showing me what a wonderful place California is. Thanks Alberto for the many exciting and interesting discussions, for providing me access to the LRZ computing cluster, which made the calculations for the ensemble modeling approach so interesting. Thank you for introducing me to Michael Savageau and for suggesting him that hiring me for a postdoctoral position would be a good idea. Thank you Kati for clarifying my questions related to biological background and for your valuable feedback when it came to orally presenting research results. I also want to thank my colleagues Sayuri Hortsch, Hannes Löwe, Sabine Wagner, Diana Kreitmayer, Viktoria Kindzersky and Christiana Sehr for making the daily routine at work so enjoyable. My special thanks to Sayuri Hortsch for the constant mathematical support you provided me, for being there to hear my worries and for sharing your point of view about so many things. Thanks to my cooperation partners Lars Janoschek, Max Hirte, Sabine Wagner and Katarina Kemper for the exciting discussions and for sharing your experimental data with

## *Acknowledgments*

me. Thanks to my master students Laura López, Thomas Mainka, Niklas Farke and Darian O’connor and my HiWi students Philipp Schneider and Makis Kolaitis for your great work and valuable input. Thanks to Professor James Liao for allowing me to perform a research visit at your laboratory. Thanks to his lab members Po-Wei Cheng and Mathew Theisen for providing me with Matlab code pre-publication and for answering all of my questions related to the ensemble modeling approach. Thanks to Susanne Kuchenbaur for guiding me through the bureaucracy at TUM.

Finalmente quisiera dedicarle este logro a mis padres Nora y Jinmer; a mis hermanos Adriana y Cesar; a mi tía Norma y a mi primos Luana y Julian. Nada de esto hubiera sido posible sin su apoyo y compañía constante e incondicional. Es difícil expresar en palabras mi gratitud. Me siento infinitamente afortunado de tenerlos a mi lado y espero que la vida nos permita compartir muchos años más de vida para seguir compartiendo nuestros logros y felicidades.

This work was financed by the German Federal Ministry of Education and Research. Most travel expenses were financed by the Joachim Herz Foundation. Research visit at UCLA was partly financed by the TUM Graduate School.

# Abstract

Metabolic Engineering is an emerging science aiming at the development of cellular factories for the overproduction of valuable chemicals. Target molecules can be either naturally produced by endogenous metabolic pathways, or the host metabolism should be complemented by a non-inherent metabolic pathway to enable their heterologous production. Over the past decades, mathematical modeling of cellular metabolism for strain and process optimization has given rise to a more rational, model-based Metabolic Engineering science. However, theoretical workflows providing advice on limitations and proper application of the vast number of available mathematical tools are still scarce. Initially, we review the application of mathematical methods to increase the production of succinate in engineered strains. Succinate is an important building block whose biotechnological production has gained much attention in the last decade. From this initial work, we conclude that direct experimental implementation of model predictions (*in silico knowledge*) is not a straightforward process yet. One of the many reasons is the intrinsic complexity of living systems, which cannot be fully captured by the simplicity of widely used stoichiometric models of metabolism. Additionally, incongruences in the modeling process and the reporting of experimental results hamper a proper assessment of the prediction power of current modeling approaches. Motivated by these observations, we developed a theoretical workflow for metabolic engineering, highlighting capabilities and limitations of each method. The workflow considers the application of not only constraint-based methods like Flux Balance Analysis (FBA), which have been traditionally used to understand optimality principles shaping bacterial metabolism, but also of kinetic-based methods whose spread has been hindered so far by limitations related to model parametrization and to high demand on computational power required to analyze genome-scale kinetic models. While developing the workflow, we paid special attention to consider the so-called metabolic burden, a phenomenon presented in "loaded" cells and characterized by the reduction of both biomass yield and critical growth rate for acetate secretion. The suggested protocol was mainly applied to generate *in silico knowledge*, aimed to guide future experimental efforts towards optimization of taxadiene production in *Escherichia coli* (*E. coli*) at the strain and process level. Taxadiene is a precursor molecule for the anticancer drug taxol and its biotechnological production has gained much attention due to the low yields of the traditional extraction process from the bark of the pacific yew tree. During the development of a flexible taxadiene producing strain, simultaneous utilization of glucose and xylose by *E. coli* was also analyzed. By applying various tools described in the protocol, metabolic load and effects arising from simultaneous sugar uptake were assessed, especially focusing on the production potential of each strain. This analysis should allow the selection of strain candidates for further optimization.



# 1. Introduction

Terpenes are naturally synthesized compounds formed by the ligation of activated isoprene molecules isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). In nature, IPP and DMAPP can be synthesized through two different metabolic routes: the mevalonate (MVA) or the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway. Terpenoids, i.e. functionalized terpenes, have a wide range of applications in the food, cosmetic and pharmaceutical industry. Lycopene, taxol and artemisinin are some of the most prominent examples of terpenoid molecules. Taxol, more specifically, is a potent plant-derived drug, showing anticancer activity against various cancer types. It was first isolated from the slow-growing pacific yew tree (*Taxus brevifolia*) and its early production method required the processing of two to four adult trees to provide dosage for one single patient [140]. Due to its structural complexity (see Figure 1.1.A), total chemical synthesis approaches often require many steps and exhibit low total yields [101]. Current large scale production methods encompass semi-synthetic approaches based on a plant derived precursor (baccatin III, extracted from the needles of the European yew tree *Taxus baccata* [140]) and plant cell cultures [43]. Since these methods still depend on plant-based processes, they exhibit an inherent limited scalability and productivity. Moreover, they offer only marginal economic benefits and are associated with several environmentally hazardous processing steps. By contrast, biotechnological taxol production has been identified as a sustainable and economically interesting alternative to established production methods. Potential advantages of this approach include:

1. High productivity and scalability due to rapid microbial growth rate.
2. Conservation and preservation of natural resources because, in contrast to *Taxus brevifolia*, bacterial host cells like *E. coli* can be sustainably produced in high quantities.
3. Access to novel taxol derivatives by applying tools of synthetic biology and metabolic engineering.

As schematically shown in Figure 1.1.B, the project **SysBioTerp** aims at the successive production of taxol derivatives using an *E. coli*-based production platform. Taxadiene is the first committed intermediate in the synthesis of taxol and therefore, the first molecule to be produced and characterized. Sequential functionalization steps include acetylation, hydroxylation, oxidation and benzylation. Since it has been reported that taxol derivatives might have beneficial effects on human health, biological activity of novel taxoids produced by metabolically engineered *E. coli* strains should be assessed in a final bio-screening step.

Due to the project complexity and scale, different research groups were involved, each one working on different aspects of the project. The study at hand was performed at the

## 1. Introduction

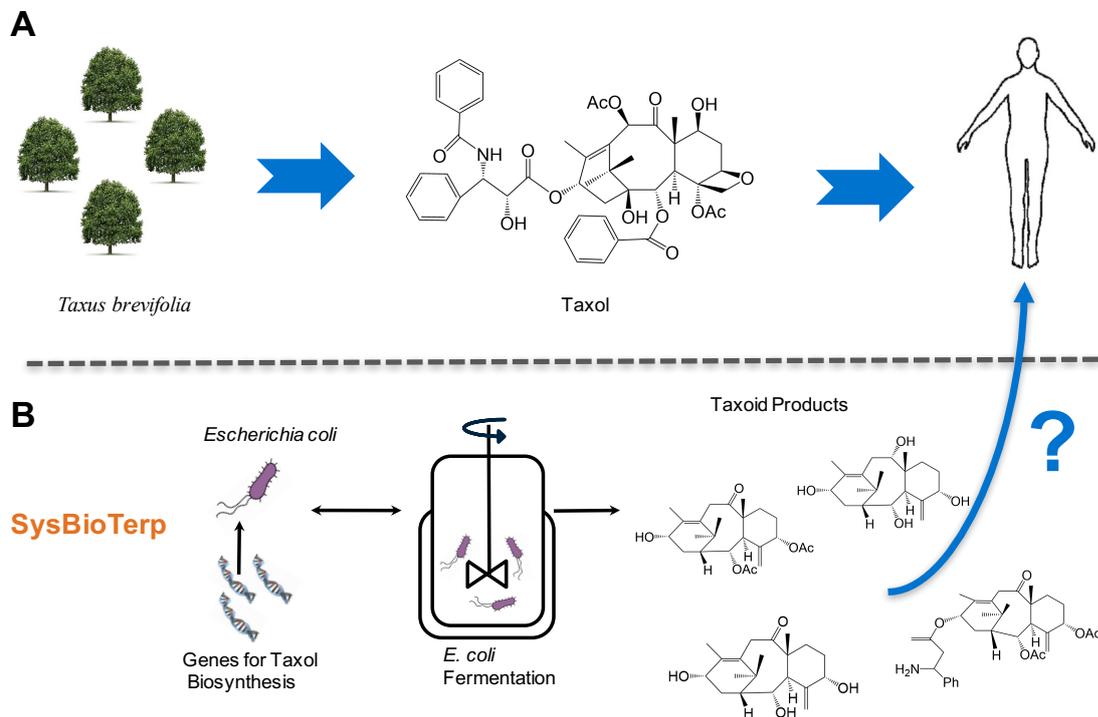


Figure 1.1.: **Motivation & Project's Aim** Motivation and project's aim is shown. **A.** Taxol is a potent plant-derived drug used to treat a number of types of cancer, including ovarian, breast, lung, cervical, and pancreatic cancer. Before current production methods were developed, producing enough taxol to provide dosage for one patient required the processing of two to four adult, slow-growing *Taxus brevifolia* trees. **B.** The SysBioTerp project aims at the development of a *E. coli*-based taxoid production platform, thus contributing to the development of sustainable and environmentally friendly production processes. Additionally, produced taxoids should undergo a bioactivity screening. More specifically, anticancer and antimicrobial activity of each novel taxoid should be determined.

specialty division for systems biotechnology (SBT). The task of our group was the data-driven development and implementation of modeling approaches on both the process and microbial metabolism level. From this task, three goals are derived:

1. *Reduction of metabolic burden caused by heterologous enzyme overexpression.* Metabolic burden is a phenomenon observed in microorganisms supporting plasmid-based enzyme overexpression. It is characterized by a reduction of the overall cellular fitness, which leads to a reduction of cellular growth rate, a reduction of biomass yield and early acetate secretion [11, 15, 58]. Since the native *E. coli* metabolism has to be expanded and re-directed by expressing a number of native and heterologous enzymes to allow for taxoid production, it is expected that metabolic burden will be a major factor limiting the production capabilities of the selected host *E. coli*. In a first step, detrimental effect of enzyme overexpression on cellular fitness should be experimentally characterized. Then, model-driven strategies aiming at the minimization of the metabolic burden should be developed and if possible, experimentally validated.
2. *Development of production strains through rational design.* Over the past years, the emerging science of Metabolic Engineering [7, 139] has allowed the construction of

microbial cellular factories for the overproduction of many relevant target molecules. The development of genome-scale metabolic reconstructions, which are mathematical representations of the cellular metabolism, together with strain-design algorithms, which employ metabolic reconstructions to identify genetic perturbations leading to the overproduction of the target molecule, has given rise to the development of a more rational, model-based Metabolic Engineering. Consequently, state of the art mathematical tools and modeling approaches should be applied to guide the process of strain development, which should ultimately lead to the construction of an optimal taxoid production strain.

3. *Optimization of bio-reactor parameters for maximal process performance.* It is expected that optimal reactor parameters (production temperature, time point of culture induction, aeration level, etc.) that maximize performance indicators, such as yield or productivity, will be highly strain dependent and must therefore be independently identified for each new production strain. Aiming at the reduction of this time-consuming and costly experimental work, modeling approaches should be applied to guide and accelerate experimental process optimization.

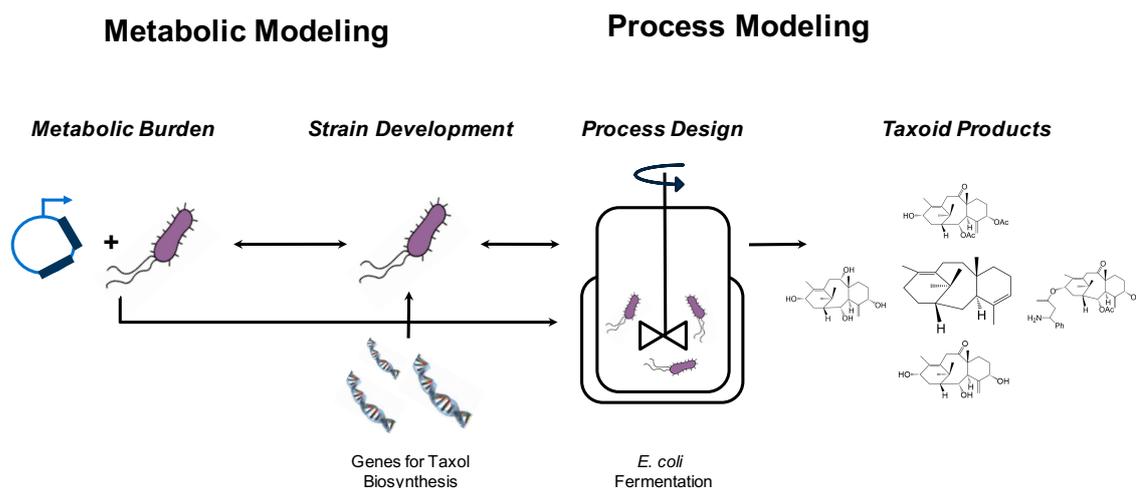


Figure 1.2.: **Modeling Workflow.** Interdependencies between our three main goals, minimization of metabolic burden, strain development and process optimization, should be consistently addressed during the modeling process in order to generate integral optimization strategies leading to maximal overall process performance. Data provided by our cooperation partners for model construction mainly consisted of concentration time courses for biomass, substrate, acetate and product. Additionally, cellular metabolism of selected strains was elucidated by means of  $^{13}\text{C}$  labeling experiments. Applied modeling approaches encompassed both stoichiometric- and kinetic-based methods.

Figure 1.2 shows schematically each of these three goals and the interactions between them. Clearly, all three task - minimization of metabolic burden, strain development and process optimization - are interdependent. For instance, optimal process parameters will necessarily depend on the production strain genotype. Additionally, strategies aiming at reducing the metabolic burden caused by enzyme overexpression will certainly depend on the extend and identity of the enzymes being overexpressed, which in turn is substantially defined by the target molecule. These dependencies should be consistently addressed during the modeling process in order to generate integral optimization strategies leading to

## 1. Introduction

maximal process performance.

The work at hand is not the first model-driven optimization approach ever developed. The same holds true for mathematical tools to be developed. Indeed, the number of existing strain-design algorithms is vast [87]. However, with very few exceptions [170], existing mathematical approaches have been mainly focused on single aspects - strain design, process optimization or minimization of metabolic burden-. Additionally, theoretical workflows providing advice on limitations and proper application of these mathematical tools are still scarce. Consequently, newly developed as well as existing computational methods should be framed into a theoretical workflow addressing the issue of model-driven process optimization in an integral fashion, considering all three previously mentioned aspects and their interdependencies.

This thesis is divided into three parts. We start describing the theoretical methods used in this work and propose a workflow for Metabolic Engineering. The results obtained from the application of this workflow to the terpenoid production in *E. coli* are described in the second and third part.

Part I.  
Theory



## 2. Metabolic Modeling & Strain Engineering

Modeling the cellular metabolism provides a deep understanding of the cellular response to environmental or genetic perturbations and enables the rational design of cellular factories. The aim of strain engineering is to redirect the metabolism of these cell factories in order to maximize the bio-synthesis of natural or non-natural target products. This is achieved through genetic modifications of the host genotype or through adequate culture conditions, i.e. medium composition, culture temperature, etc. Modeling the cellular metabolism can be a challenging task. Additionally, the proper approach greatly depends on the level of detail required for the specific application. Over the past decades, many mathematical approaches have been proposed. Each method is based on different assumptions and consequently, allows for different types of predictions. In general, the cellular metabolism can be mathematically described by setting up a mass balance for all metabolites in the cell:

$$\frac{dc}{dt} = \mathbf{S} r - \mu c, \quad (2.1)$$

where  $r$  is the vector of fluxes through the reactions,  $\mu$  is the growth rate,  $c$  is a vector of intracellular metabolite concentrations and  $\mathbf{S}$  represents the stoichiometric matrix of the reaction network. The first term represents the rates of consumption or production of a specific metabolite, while the second describes the dilution rate caused by cell growth. Since the dilution rate is normally much lower than the reaction rates, the Equation (2.1) can be simplified to:

$$\frac{dc}{dt} = \mathbf{S} r. \quad (2.2)$$

At steady state, there is no accumulation of metabolites in the cell and Equation (2.2) can be further simplified to:

$$0 = \mathbf{S} r. \quad (2.3)$$

In the following sections, the three most commonly used modeling approaches for analyzing Equation (2.3) are presented, namely kinetic-, constraint- and elementary modes-based methods.

## 2.1. Constraint-based Methods

Since the stoichiometric matrix  $\mathbf{S}$  of a real biological network is typically non-square, containing more unknown rates than equations, it does not have an inverse. Consequently, a flux vector  $r$  satisfying Equation (2.3) is not *unique*. Additional constraints on the flux vector  $r$  can be applied to further reduce the number of allowable flux distributions [31]. Limits on the range of individual flux values can be used for this purpose: thermodynamic constraints expressed as reaction reversibility can thus be included by setting one of the boundaries of an irreversible reaction to zero [53]. In a similar way, maximum flux values can be estimated based on enzymatic capacity limitations [13], or for the case of exchange reactions (i.e. reactions that transfer mass between the culture medium and the cell), experimentally determined maximal uptake or production rates can be used. Regulation of gene expression can also be considered in cases where the regulatory effects have a great influence on cellular behavior [32]. Usually, these constraints are not sufficient to reduce the solution space to a single solution. Constraint-based models have been popularly used to calculate a flux vector  $r$  that represents the cellular phenotype at steady-state using different approaches. Flux Balance Analysis (FBA) has been the most widely applied method. It consists of a linear programming formulation that, by imposing an objective function, enables the calculation of a flux distribution that maximizes or minimizes that objective. Typically, the objective function used coincides with an assumed cellular objective, such as growth. Other commonly used cellular objectives include the sum of all intracellular fluxes or ATP generation [128]. Mathematically, the FBA formulation reads:

$$\begin{aligned}
 & \text{Maximize } Z = c \cdot r \\
 & \text{subject to:} \\
 & \quad \mathbf{S} \cdot r = 0 \\
 & \quad lb \leq r \leq ub,
 \end{aligned} \tag{2.4}$$

where  $Z$  is the objective function resulting from a linear combination of selected reactions of the flux vector  $r$ , as determined by the vector  $c$ .  $lb$  and  $ub$  are lower and upper flux boundaries, respectively.

Due to redundancies in the architecture of the cellular metabolism, alternate optimal solutions can exist. Mahadevan et al. [88] introduced the concept of Flux Variability Analysis (FVA) to characterize this issue. The approach begins with determining the optimal value of the objective function by solving the linear optimization problem outlined in Equation 2.4. From this solution, the range of variability that can exist in each flux in the network due to alternate optimal solutions can be calculated through a series of optimization problems. In each problem, the value of the original objective ( $Z_{optimal}$ ) is fixed and each reaction in the network is maximized (Equation 2.5) and minimized (Equation 2.6) to determine the feasible range of flux values for each reaction. The mathematical formulation of the FVA reads:

$$\begin{aligned}
& \text{Maximize } r_i \\
& \text{subject to:} \\
& \quad c \cdot r = Z_{\text{optimal}} \\
& \quad \mathbf{S} \cdot r = 0 \\
& \quad lb \leq r \leq ub,
\end{aligned} \tag{2.5}$$

$$\begin{aligned}
& \text{Minimize } r_i \\
& \text{subject to:} \\
& \quad c \cdot r = Z_{\text{optimal}} \\
& \quad \mathbf{S} \cdot r = 0 \\
& \quad lb \leq r \leq ub,
\end{aligned} \tag{2.6}$$

If for a given reaction it holds that  $r_{max} = r_{min}$  ( $r_{max}$  and  $r_{min}$  are obtained from Equations 2.5 and 2.6 respectively), then no variability is allowed for that reaction and a unique flux value associated to that reaction is required to obtain  $Z_{\text{optimal}}$ .

Constraint-based approaches have been very popular within the modeling community. This popularity can be partly attributed to the simplicity of the mathematical tools required (linear programming), but also to the availability of metabolic reconstructions for many industrially relevant microorganisms [66]. For the specific case of *E. coli*, much work has been done during the past decade towards the development of a comprehensive genome-scale metabolic reconstruction. More specifically, Table 2.1 presents an overview of efforts made by the laboratory of Bernhard Palsson at UC San Diego. Traditionally, metabolic reconstructions have been limited to a stoichiometric description of the cellular metabolism ( $\mathbf{S}$  matrix in Equation 2.3). However, models have been growing in coverage to consider additional aspects of cellular metabolism such as macromolecular expression [102] and protein 3D structures [97]. In this work, we extensively used the *E. coli* core model [107] as well as the *iJO1366* genome-scale metabolic reconstruction [106].

Table 2.1.: **Genome-Scale Metabolic Reconstructions for *E. coli*.** The *E. coli* core model was extensively used in this work. It encompasses central metabolic pathways like glycolysis, pentose phosphate pathway, tricarboxylic acid cycle and pyruvate metabolism. As the metabolic reconstructions grow in scope, it is possible to simulate more complex genotypes. However, the analysis of associated flux distributions become more challenging. The model *iJO1366* was the last genome-scale reconstruction considering only a stoichiometric description of the metabolism. All subsequent models consider additional features such as macromolecular expression and protein 3D structures.

	<i>E. coli</i> core [107]	<i>iJR904</i> [114]	<i>iAF1260</i> [41]	<i>iJO1366</i> [106]	<i>iML1515</i> [97]
Included genes	137	904	1260	1366	1515
Reactions	95	931	2077	2251	2719
Exchange reactions	20	143	298	329	338
Metabolites	72	761	1039	1136	1192

*iML1515* additionally contains 1515 protein 3D structures and 1888 unique protein domains

### 2.1.1. Dynamic Flux Balance Analysis

Equation 2.4 allows the calculation of steady state flux distributions for a given objective function and set of constraints. Assuming that microorganisms are able to rapidly reach steady state in response to changes in the extracellular environment (bio-reactor), it is possible to couple the FBA formulation with a set of differential equations describing the reactor dynamics to simulate time courses of extracellular species within the bio-reactor. This approach is termed dynamic Flux Balance Analysis (dFBA) and is schematically represented in Figure 2.1.

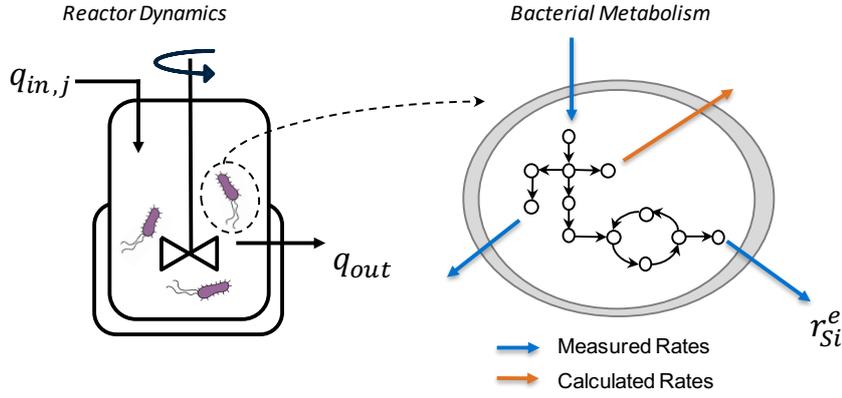


Figure 2.1.: **Dynamic Flux Balance Analysis.** Reactor dynamics can be coupled with bacterial metabolism to simulate concentration time courses of chemical species present in the bioreactor. The coupling is done through the cellular growth rate  $\mu$  and exchange rates  $r_{Si}^e$ . Refer to Equations 2.11 and 2.12

Equations 2.8, 2.11 and 2.12 fully describe the dynamics of reactor volume, biomass concentration and concentration of extracellular metabolites in the bioreactor, respectively. Symbols used in these equations can be found in Table 2.2. Note that the coupling between reactor dynamics and cellular metabolism occurs through the growth rate  $\mu$  and exchange rates  $r_{Si}^e$ .

The equation describing the dynamic of reactor volume is determined as follows. We first set up a mass balance equation for the bio-reactor:

$$\frac{dm_R}{dt} = \sum q_{in,j} \rho - q_{out} \rho \quad (2.7)$$

If the density  $\rho$  can be assumed to be constant and since  $m_R = V_R \rho$ , it follows that:

$$\frac{dV_R}{dt} = \sum q_{in,j} - q_{out} \quad (2.8)$$

In order to model the biomass dynamics, it is assumed that any feed to the bio-reactor do not contain cells. The differential equation describing the change of biomass in the reactor reads:

$$\frac{dm_B}{dt} = \mu m_B - q_{out} c_B \quad (2.9)$$

Table 2.2.: **Overview of Variables Used to Describe Reactor Dynamics.** In order to set up mass balance equations discussed in the main text, multiple volumetric feeds are assumed (one for each substrate). By contrast, one single volumetric efflux is considered. Note that  $c_{S_i}$  refers to the concentration of metabolite  $i$  in the reactor, while  $c_{S_i}^{in}$  corresponds to the concentration of metabolite  $i$  in the corresponding volumetric feed. gDW is used to describe the concentration of biomass and refers to gram *dry weight*.

Name	Symbol	Units
Density	$\rho$	g/l
Reactor volume	$V_R$	l
Growth rate	$\mu$	1/h
Volumetric feed	$q_{in,j}$	l/h
Volumetric efflux	$q_{out}$	l/h
Mass in reactor	$m_R$	g
Biomass	$m_B$	g
Mass of component $i$	$m_{S_i}$	g
Biomass concentration	$c_B$	gDW/l
Concentration of component $i$ in the reactor	$c_{S_i}$	g/l
Concentration of component $i$ in the feed	$c_{S_i}^{in}$	g/l
Molecular weight of component $i$	$w_{S_i}$	g/mol
Exchange reaction for component $i$	$r_{S_i}^e$	mol/gDW h

For convenience, the biomass dynamics is now expressed in terms of biomass concentration. This is done by expressing the biomass in the reactor as a function of the biomass concentration and the reactor volume:

$$\frac{dm_B}{dt} = \frac{d(V_R c_B)}{dt} = V_R \frac{dc_B}{dt} + \frac{dV_R}{dt} c_B \quad (2.10)$$

Substituting Equation 2.10 and 2.8 into Equation 2.9 and solving for biomass concentration leads to:

$$\begin{aligned} V_R \frac{dc_B}{dt} + \frac{dV_R}{dt} c_B &= \mu m_B - q_{out} c_B \\ \frac{dc_B}{dt} &= \mu c_B - \frac{\sum q_{in,j}}{V_R} c_B \end{aligned} \quad (2.11)$$

The mass balance for chemical species (substrates/products) in the liquid phase is derived in a similar way as for the biomass. In this case, exchange reactions  $r_{S_i}^e$  between the cell and culture media have to be considered. Exchange rates will have a positive sign for products secreted by the cell, whereas a negative sign refers to substrates absorbed by the cell.

## 2. Metabolic Modeling & Strain Engineering

$$\begin{aligned}\frac{dm_{Si}}{dt} &= q_{in,j} c_{Si}^{in} - q_{out} c_{Si} + r_{Si}^e c_B V_R w_{Si} \\ \frac{dc_{Si}}{dt} &= \frac{q_{in,j} c_{Si}^{in}}{V_R} - \frac{\sum q_{in,j}}{V_R} c_{Si} + r_{Si}^e c_B w_{Si}\end{aligned}\tag{2.12}$$

The mass balance equations derived for biomass, reactor volume and components in the liquid phase can be used to describe the dynamics of a continuous ( $q_{in,j} \neq 0$ ;  $q_{out} \neq 0$ ), a batch ( $q_{in,j} = q_{out} = 0$ ), or a fed-batch process ( $q_{in,j} \neq 0$ ,  $q_{out} = 0$ ).

### 2.1.2. Experimental Determination of Reaction Rates

As stated before, experimentally determined reaction rates can be used to constrain the space of allowable flux distributions satisfying the linear programming problem defined by Equation 2.4. By doing so, the biological significance of the obtained flux distributions can be increased. Since ordinary measurements of concentration time courses in the fermentation broth allows the estimation of exchange rates (rates describing the exchange of mass between the cell and the culture medium, that is substrate uptake, product and by-product secretion rates), these are normally used to constraint the solution of Equation 2.4. Additionally, more complex data, such as the obtained in  $^{13}\text{C}$  labeling experiments can also be used to estimate a number of intracellular rates. The methodology used in this work to calculate both exchange and intracellular rates from experimental data will be discussed in this section.

#### Determination of Exchange Rates

Concentration time courses for the substrate (or substrates) and the product are usually available for modeling studies, because this data are routinely obtained during experimental strain characterization to assess the production performance of a given strain. Time courses for by-products like acetate, ethanol, pyruvate, lactate, etc. might also be available, since their signals are normally contained in the chromatogram used to quantify the main carbon source if a standard High Performance Liquid Chromatography (HPLC) method is used for sugar quantification. Figures 2.2.A and 2.3.A show exemplary concentration time courses for biomass and glucose, respectively. In the case of biomass, Optical Density (OD) measurements along with a conversion factor have been traditionally used to determine the biomass concentration in the culture in units of gram dry weight (gDW) per liter. It has been shown that the gDW/OD conversion factor strongly depends on the genetic background of the strain [83]. In this study, strain-specific conversion factors were used to obtain biomass concentration in units of gDW/l. For the specific case shown in Figure 2.2.A, a conversion factor of 0.54 was used to calculate the biomass concentration in units of gDW/l.

The first step to calculate any exchange rate is the determination of the cellular growth rate  $\mu$ . As shown in Equation 2.13,  $\mu$  is a proportionality constant used to describe the increase of the biomass concentration over time as a function of the biomass concentration in a certain point in time:

$$\frac{dc_B}{dt} = \mu c_B \quad (2.13)$$

As defined in section 2.1.1,  $c_B$  refers to the biomass concentration. Equation 2.13 is obtained by setting  $\sum q_{in,j}$  to zero in Equation 2.11. This is true for a batch process or when the total volumetric feed  $\sum q_{in,j}$  is low compared to the total volume of the reactor. By assuming that  $\mu$  is not a function of time, Equation 2.13 can be integrated to obtain:

$$\ln(c_{B,f}) = \mu(t_f - t_o) + \ln(c_{B,o}). \quad (2.14)$$

$t_o$  and  $t_f$  refer to the initial and final time points, while  $c_{B,o}$  and  $c_{B,f}$  represent the corresponding biomass concentration. Note that the assumption of constant  $\mu$  is generally valid during the exponential growth phase. By plotting the natural logarithm of experimentally measured biomass concentration as a function of time, it is possible to calculate  $\mu$  as the slope of the resulting straight line. Figure 2.2.B shows this procedure for the biomass concentration time course shown in Figure 2.2.A. In this case, the analyzed culture had a growth rate of 0.49 1/h.

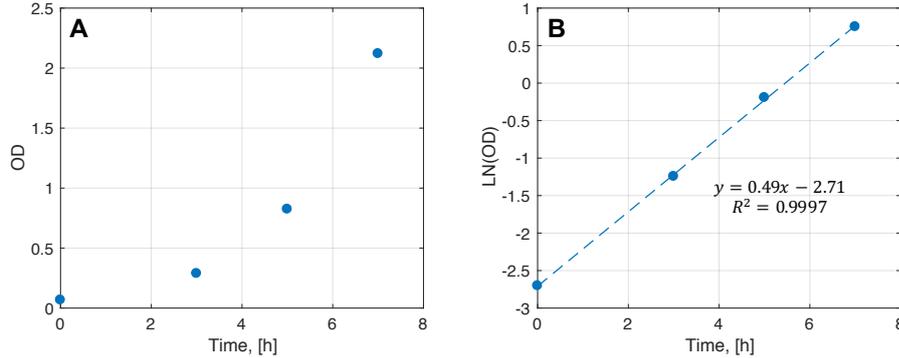


Figure 2.2.: **Experimental Determination of Cellular Growth Rates.** **A.** OD time course **B.** Natural logarithm of OD measurements as a function of time. The slope of the straight line corresponds to the cellular growth rate  $\mu$ .

Once  $\mu$  has been estimated from the experimental OD time course, exchange rates  $r_{S_i}^e$  can be calculated. As indicated before, strain characterization normally occurs in batch processes (shaking flask). By setting  $q_{in,j}$  and  $\sum q_{in,j}$  to zero, Equation 2.12 can be simplified to:

$$\frac{dc_{S_i}}{dt} = r_{S_i}^e c_B w_{S_i}. \quad (2.15)$$

Flux distributions are normally calculated in units of mmol/gDW h. By dividing Equation 2.15 by  $w_{S_i}$  (molecular weight), it is possible to change the units of the balance equation for the metabolite  $S_i$  from g/l to mol/l (or mmol/l). Combining Equations 2.15 and 2.13 and solving the resulting equation for  $r_{S_i}^e$  one obtains:

$$r_{S_i}^e = \mu \frac{dc_{S_i}}{dc_B}. \quad (2.16)$$

## 2. Metabolic Modeling & Strain Engineering

Equation 2.16 states that the exchange rate for metabolite  $r_{S_i}^e$  can be obtained by multiplying the cellular growth rate  $\mu$  by the slope of the curve obtained when plotting  $c_{S_i}$  against  $c_B$ . Figure 2.3.B exemplarily shows this procedure. The glucose uptake rate  $r_{Glucose}^e$  can be then calculated as:  $(-11.75 \text{ mmol/gDW})(0.491/h) = -5.75 \text{ mmol/gDW h}$ . Note that the slopes in Figures 2.2.B and 2.3.B were calculated using least-squares regression.

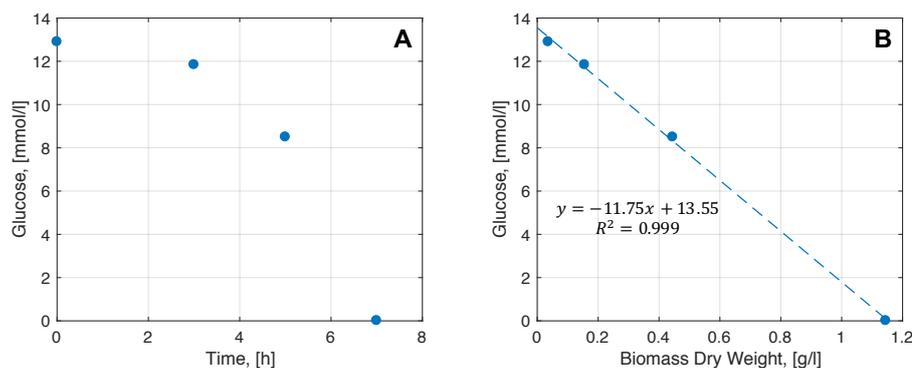


Figure 2.3.: **Experimental Determination of Glucose Uptake Rate.** **A.** Concentration time course for glucose. **B.** Glucose concentration is plotted against biomass concentration. Note that the biomass concentration was obtained by multiplying OD measurements by a conversion factor of 0.54 gDW/OD. The slope of this straight line corresponds to the term  $\frac{dc_{S_i}}{dc_B}$  of Equation 2.16.

### Determination of Intracellular Rates: $^{13}\text{C}$ -Metabolic Flux Analysis

Even though exchange rates can be easily calculated from respective concentration time courses for a number of by-products, they do not provide enough constraints to precisely estimate all fluxes in complex biological systems containing reversible reactions, parallel pathways and internal cycles [5, 127, 157].  $^{13}\text{C}$ -Metabolic Flux Analysis ( $^{13}\text{C}$ -MFA) offers a means for the *indirect* estimation of a number of intracellular fluxes [157]. Typically, cells are grown on  $^{13}\text{C}$ -labeled substrates until the isotope label is distributed throughout the network. The cell reaches isotopic steady state when the labeling patterns do not change over time. At this point, cells are harvested and Mass Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) analysis are implemented to detect the  $^{13}\text{C}$  patterns of either protein-bound amino acids or free metabolic intermediates. The analysis of protein-bound amino acids is preferred because protein is stable and abundant. On the other hand, free metabolic intermediates provides the richest source of information, but high turnover and very low concentrations of intermediates poses serious technical challenges to sample preparation, separation and analytical sensitivity [167]. Specific labeling patterns occur in the metabolic intermediates (free intermediates or protein-bound amino acids) as a function of the particular distribution of fluxes in an organism [157]. *Indirect* estimation means in the context of  $^{13}\text{C}$ -MFA that intracellular fluxes must be extracted from measured labeling patterns using a model-based approach. Comprehensive mathematical models that describe the relationship between metabolite labeling patterns and fluxes are used to simulate isotopic abundances of all metabolites in a network for any set of steady state fluxes. Various mathematical approaches have been developed to describe the relationship between flux distribution and labeling patterns. Usually, these models consist of the complete set of isotopomer (isotope isomer [157]) balances, which may be derived

using a matrix based method as described by Schmidt et al. [126]. Alternative modeling strategies have been proposed based on the concept of cumomer balances [158], bundomer balances [150] and Elementary Metabolite Units (EMU) [6]. Regardless of the approach used, the flux distribution responsible for the measured labeling pattern is identified by minimizing the difference between observed and simulated isotope spectra. In essence, flux determination is a large-scale nonlinear parameter estimation problem:

$$\begin{aligned} & \text{Minimize } \Phi = (x(r) - x^{obs})^T \Sigma_x^{-1} (x(r) - x^{obs}) \\ & \text{subject to:} \\ & \mathbf{S} r = 0 \end{aligned} \tag{2.17}$$

where the objective function  $\Phi$  is the covariance-weighted sum of squared residuals,  $x(r)$  is the vector of simulated measurements (which is a function of the flux vector  $r$ ),  $x^{obs}$  is the vector of experimental data containing both labeling measurements and exchange rates measurements, and  $\Sigma_x$  is the covariance matrix, which contains variances of the measurements on the diagonal. Note that the stoichiometric matrix  $\mathbf{S}$  is a  $m \times k$  matrix, where  $m$  refers to the number of metabolites and  $k$  to the number of reactions.

During the iterative optimization process of Equation 2.17, not all flux values in the vector  $r$  can be freely chosen by the solver. In fact, there are only  $k - \text{rank}(\mathbf{S})$  independent variables, also referred to as free fluxes [126,158]. Independent fluxes can be obtained from the general solution of Equation 2.3:

$$r = \mathbf{N} u \tag{2.18}$$

Where,  $\mathbf{N}$  is the null space matrix of  $\mathbf{S}$  and  $u$  is the vector of independent fluxes. In order to reduce computational time, one can introduce Equation 2.18 into Equation 2.17 to obtain a new optimization problem in terms of the free fluxes vector  $u$ :

$$\begin{aligned} & \text{Minimize } \Phi = (x(u) - x^{obs})^T \Sigma_x^{-1} (x(u) - x^{obs}) \\ & \text{subject to:} \\ & \mathbf{N} u \geq 0 \end{aligned} \tag{2.19}$$

Reversible reactions are usually simulated as two independent reactions. Consequently, the constraint  $\mathbf{N} u \geq 0$  requires that all fluxes are non-negative.

### Calculating Confidence Intervals for Reactions

Confidence intervals for intracellular fluxes obtained from Equation 2.19 are useful to estimate the precision of a certain reaction flux given a certain set of labeling data. Confidence intervals have been calculated from estimated local standard deviations, but it has been shown that these intervals may not accurately describe the true uncertainty due to inherent nonlinearities of isotopomer balances [5]. In this work, we calculate confidence intervals for fluxes using the approach developed by Antoniewicz et al. [5], in which for each reaction, the sensitivity of the minimized sum of squared residuals is determined as a function of the flux value. The approach starts by stating that the difference between the objective function evaluated at the optimal solution  $\hat{u}$  and the objective function when one flux is

## 2. Metabolic Modeling & Strain Engineering

fixed follows a  $\chi^2$ -distribution with one degree of freedom:

$$(\Phi(u)|_{r_i=r_{i0}} - \Phi(\hat{u})) \sim \chi^2(1), \quad (2.20)$$

where  $\Phi(u)|_{r_i=r_{i0}}$  indicates the value of the objective function when the flux  $i$  is fixed at  $r_{i0}$  and the other degrees of freedom are used to minimize the objective function. The  $1 - \alpha$  confidence interval for flux  $i$  is given by the flux values for which following statement is true:

$$\Phi(u)|_{r_i=r_{i0}} \leq (\Phi(\hat{u}) + \chi_{1-\alpha}^2(1)). \quad (2.21)$$

Note that  $\alpha$  refers to the probability of error. The threshold values for  $\chi_{1-\alpha}^2(1)$  corresponding to 80%, 90%, 95% and 99% confidence intervals of fluxes are 1.64, 2.71, 3.84 and 6.63, respectively [5]. Thus, in order to obtain accurate confidence intervals we need to determine the minimized sum of squared residuals as a function of the flux value. Small sensitivities of the objective function with respect to changes in a certain flux indicate that that flux cannot be estimated precisely. Conversely, large sensitivities indicate that the flux is well determined [5]. Figure 2.4 illustrate the three-step implementation of this approach for the estimation of the reaction flux of the reaction catalyzed by the enzyme pyruvate dehydrogenase (PDH) and its confidence interval.

1. Generate multiple initial flux distributions by sampling free fluxes. Then, solve the optimization problem described by Equation 2.19. Note that not all flux distributions will converge to the same value of the objective function. The optimal flux distribution is the one exhibiting the lowest objective function among all sampled initial flux distributions [48]. See Figure 2.4.A
2. In order to estimate the confidence interval for a given reaction, increase and decrease the flux through that reaction starting from the optimal solution obtained in the previous point. For each perturbed flux distribution, solve the optimization problem described by Equation 2.19. See Figure 2.4.B
3. Values for the upper and lower bound of the confidence interval can be identified by applying Equation 2.21. See Figure 2.4.C

All confidence intervals reported in this work were calculated using an  $\alpha$  of 5% (95% confidence intervals). The number of sampled initial flux distributions varied between 500 and 1000.

## 2.2. Elementary modes-based Methods

As discussed before, the number of flux vectors  $r$  satisfying Equation (2.3) is infinite. Elementary mode analysis calculates all the solutions of Equation (2.3) by adding a non-decomposability or genetic independence constraint to the reaction directionality constraint introduced by the FBA formulation. Genetic independence implies that enzymes catalyzing the reactions in one solution, represented by flux vector  $r_1$  are not a subset of another flux vector  $r_2$  [68, 130]. If this condition is satisfied, the flux vectors  $r_1$  and  $r_2$  belong to the set of elementary modes  $e$ . Through this definition all feasible flux distributions  $r$  can be

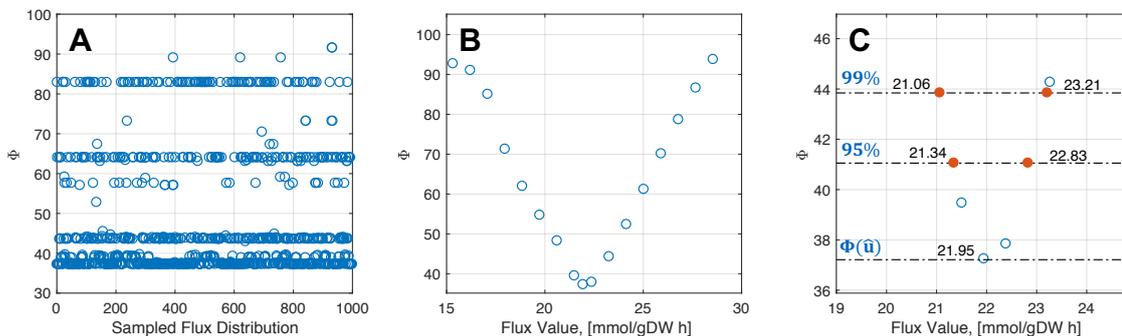


Figure 2.4.: **Constraint-based Assessment of Production Capabilities.** **A.** Due to the non-convex nature of the objective function, the problem must be solved several times using different sets of free fluxes. Each one of the 1000 data points in the figure represents a flux distribution. The flux distribution reaching the minimum value of the objective function is considered to be the solution of the problem described by Equation 2.19. **B.** The sensitivity of the objective function with respect to the flux through the reaction PDH is calculated as explained in the main text. **C.** 95 and 99% confidence intervals for the PDH reaction is calculated using Equation 2.21 and are given in the Figure. Lower and upper boundaries were calculated by linear interpolation. Note that the flux through the PDH providing the best match to the labeling data corresponds to 21.95 mmol/gDW h.

represented by non-negative linear combinations of the elementary modes:

$$r = \sum_{i=1}^n a_i e_i \quad a_i \geq 0. \quad (2.22)$$

In contrast to constraint-based approaches, no assumptions on a cellular objective function are required for flux calculation using elementary mode-based methods. As a consequence, these approaches are classified as unbiased, while constraint-based approaches are classified as biased. Other unbiased approaches for the analysis of metabolic networks include Monte Carlo sampling and extreme pathway analysis [122, 125]. Elementary mode based methods have typically been limited to the analysis of small networks, because the enumeration of all elementary modes is computationally expensive. However recent advances in computational algorithms have extended the applicability of elementary mode based methods to the genome scale [18]. Refer to the original publication for more details on the mathematical definition of elementary modes and their identification [130].

An interesting characteristic of the definition of the elementary modes is that they can be interpreted in terms of biochemical functions [130]. Knowing the set of elementary modes existing in a metabolic network allows, among other things, the identification of the pathway with the highest product yield and the assessment of network stability and functionality after gene deletions [129, 147]. The concepts of elementary modes were successfully applied for the first time to engineer a bacterial strain by Liao et al. [76], who engineered an *E. coli* strain for the high-efficient production of aromatic amino acid precursors.

## 2.3. Kinetic-based Methods

Kinetic models can describe the rates of intracellular reactions as a function of enzyme dependent kinetic parameters as well as metabolite and enzyme concentrations participating in the reaction. The rate expressions are then used to describe concentration changes by a set of ordinary differential equations. In terms of Equation (2.2), this means that the rate vector  $r$  has the form:  $r = f(p, c)$ , where  $p$  represents kinetic parameters and  $c$  metabolite concentrations.

Usually, kinetic approaches have been used to describe the dynamics of small to medium size systems, such as the central metabolism of *E. coli* [14, 22, 69]. Despite the high potential of kinetic based methods for strain development, the construction of large-scale kinetic models has been hindered by many difficulties mainly related to unambiguous parameter estimation. This is due to the need for big sets of kinetic parameters and the fact, that the values of individual kinetic parameters and even the form of the kinetic rate laws may need to be adjusted in response to genetic or environmental perturbations [63]. These difficulties have been tackled by many authors [4, 21, 82, 146] and call for alternative modeling strategies which require less parameters or no parameters at all: *'Qualitative and quantitative understanding and corresponding methodologies for designing desired properties of many complex systems have been successfully achieved in the fields of chemistry, physics, and the associated engineering disciplines without knowing all aspects of systems structure and certainly without knowing all parameter values involved. The same must be possible for biology.'* [8]. Two approaches popularly used in metabolic engineering that requires no exhaustive knowledge of kinetic parameter values and rely only on the knowledge of the stoichiometry of the reactions in the network, i.e. the stoichiometric matrix  $\mathbf{S}$ , were already presented in sections 2.2 and 2.1. In this section, we summarize some of the main aspects of the pioneering work done by the laboratory of James Liao at the University of California (UC) Los Angeles to overcome the issues related to unknown parameter values in large-scale kinetic models. In Chapter 5, 6 and 9, we demonstrate the prediction power of some of these tools and develop new applications based on the theoretical foundation of the Ensemble Modeling approach of metabolic networks [71, 74, 144, 146].

### 2.3.1. Ensemble Modeling Approach of Metabolic Networks

In a first paper, Tran et al. [146] applied the idea of Ensemble Modeling (EM) to the central metabolism of *E. coli*. The approach builds an ensemble of dynamic models that reach the same reference steady state in terms of flux distribution and metabolite concentrations. Within the ensemble, all models share the same kinetic structure but differ in the specific parameter values. Rate laws for each reaction can be assigned to match known mechanisms and regulations [146] or can be automatically generated based in rules involving the number of substrates, products and reversibilities [71, 74, 144] by applying the concept of modular rate laws developed by Liebermeister et al. [78, 79].

The idea of building an ensemble of models instead of using one single model to describe complex dynamic systems is not entirely new. In fact, ensemble forecasting is commonly used in numerical weather prediction [12, 23, 96, 168], where an ensemble of typically 50 models are used to account for mainly two sources of uncertainty, namely errors introduced

by the use of imperfect initial conditions and errors related to imperfections in the model formulation, such as the approximate mathematical methods to solve the equations [12]. Figure 2.5 illustrates the application of ensemble forecasting for path prediction of the hurricane Irma, one of the strongest hurricanes ever observed in the Atlantic. For a given time point, a set of models is parameterized based on the actual hurricane location and satellite data available at that moment (see point ① in Figure 2.5). Then, a set of trajectories is generated, each representing the potential path of the hurricane. In the ideal case, the future location of the hurricane is contained within the prediction of the ensemble (see point ② in Figure 2.5).

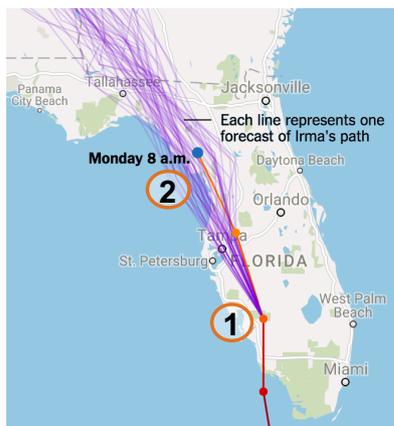


Figure 2.5.: **Application of Ensemble Forecasting for Path Prediction of Hurricane Irma.** A set of models is parameterized based on experimentally measured initial conditions (point ①). Typically, weather models employ up to  $10^7$  observations per day to derive physically consistent initial conditions [12]. Each model generates a forecast regarding the path of the hurricane. In the ideal case, the future path of the hurricane is contained within the ensemble prediction (point ②).

In the case of metabolic systems, the ensemble can be parameterized using experimental steady state flux data determined for a certain strain. As discussed in Section 2.1.2, exchange fluxes can be easily determined from measured concentration time courses for metabolites absorbed or secreted by the cell. Using experimentally determined exchange fluxes to constrain the FBA formulation described by Equation 2.4, it is possible to generate a first estimate for the intracellular flux distribution. If more advanced equipment and expertise is available, labeling experiments along with the computational methods previously described can be applied to obtain more accurate estimates of the intracellular flux distribution. Since labeling data obtained using one single tracer - usually uniformly labeled  $^{13}\text{C}$  ([U- $^{13}\text{C}$ ]) glucose - does not suffice to precisely resolve all fluxes in the central metabolism, more advanced techniques, like parallel labeling experiments [75] can be applied. In any case, uncertainties related to the reference flux distribution can be addressed by constructing different ensembles of metabolic models, each using a representative reference flux distribution to populate the ensemble. In order to obtain ensembles with biological significance, a variety of objective functions may be applied (depending on culture conditions) to obtain reference flux distributions [128]. This approach is briefly illustrated in Section 9.3, where two similar objective functions are used to populate two different ensembles and generate predictions.

Once a biologically feasible ensemble has been constructed, mathematical methods can

## 2. Metabolic Modeling & Strain Engineering

be applied to assess the effect of genetic perturbations on the cellular metabolism. At steady state, the concentration of intracellular metabolites does not change over time and we obtain:

$$\frac{dc}{dt} = \mathbf{S} r(c_{ss}, p) = F(c_{ss}, p) = 0, \quad (2.23)$$

where  $c_{ss}$  refers to a vector of steady state metabolite concentrations. Equation 2.23 is analogous to Equations 2.2 and 2.3. The only difference is the definition of the function  $F(c_{ss}, p)$ . Since  $F(c_{ss}, p) = 0$ , it follows that the total derivative with respect to  $p$  is also zero:

$$\frac{dF}{dp} = \frac{\partial F}{\partial c_{ss}} \frac{dc_{ss}}{dp} + \frac{\partial F}{\partial p} = 0. \quad (2.24)$$

Solving Equation 2.24 for  $\frac{dc_{ss}}{dp}$  yields:

$$\frac{dc_{ss}}{dp} = - \left( \frac{\partial F}{\partial c_{ss}} \right)^{-1} \frac{\partial F}{\partial p}. \quad (2.25)$$

Starting from a reference steady state, Equation 2.25 describes the effect of parameter perturbations (for instance enzyme concentration) on the vector of steady state metabolite concentrations  $c_{ss}$ . Since the calculation of the inverse of the matrix  $\frac{\partial F}{\partial c_{ss}}$  is necessary to solve Equation 2.25, it is crucial to detect the point where this matrix becomes singular. Interestingly, this point is also a bifurcation point, beyond which the system no longer reaches a stable steady state [74].

So far, two methods have been developed that analyze different aspects of Equation 2.25. The first method is termed Ensemble Modeling for Robustness Analysis (EMRA) [74] and was designed to estimate the robustness of non-native pathways towards perturbations. By perturbing the activity of a certain enzyme in the pathway and calculating the percentage of models in the ensemble that remained stable, the robustness of the pathway/enzyme can be assessed. In the original publication [74], the robustness of two synthetic central metabolic pathways that achieve carbon conservation (non-oxidative glycolysis [17] and reverse glyoxylate cycle [90]) was compared. In a subsequent paper [144], experimental data of three different cell-free enzymatic systems were used to demonstrate the existing link between production performance (product end titer, productivity) and system robustness. As predicted by EMRA, unstable systems exhibited a lower production performance.

A second method, termed Kinetically Accessible Yield (KAY), uses the maximal flux value through a given pathway before the metabolic system loses stability (or any metabolite concentration becomes negative) to estimate experimentally measured product yields. In the original publication, Lafontaine Rivera et al. [71] used the KAY formulation to successfully predict the isobutanol yield of three different genotypes. Interestingly, the authors demonstrated that KAY can be calculated by either flux or kinetic parameter integration. In both cases, the calculated KAY value is the same [71]. Throughout this work, KAY values were calculated using flux integration. In this case, no specific knowledge of reaction kinetics of the production pathway is required. Instead, a single lumped reaction representing the whole production pathway is used as input.

## Metabolic Reconstructions for the Ensemble Modeling Approach

One of the main advantages of the EM approach is that it allows for a kinetic-based analysis of available metabolic reconstructions (refer to Table 2.1). These reconstructions have been manually curated and are frequently updated, thus becoming more accurate and complete over the years. Instead of automatically generating the metabolic network from the EcoCyc [60,61] database, as done by Lafontaine Rivera et al. [71], we adapted already existing metabolic reconstructions to make them suitable inputs for the EM approach as follows:

1. *Check exchange reactions.* Exchange reactions are required to have positive flux values within the EM framework for both uptake and export reactions. By contrast, traditionally used metabolic reconstructions (refer to Table 2.1) exhibit negative flux values for uptake reactions and positive flux values for export reactions. Therefore, uptake reactions must be converted from the form “*metabolite*  $\rightarrow$  ” into “  $\rightarrow$  *metabolite*”. This can be simply done by multiplying the column  $i$  of the stoichiometric matrix  $\mathbf{S}$  by -1.  $i$  refers to the reaction(s) responsible for the uptake of a given substrate(s).
2. *Split reaction describing biomass production.* Biomass formation is represented within the EM approach by a set of efflux reactions [71]. By contrast, cellular growth is mathematically described by one single reaction in traditionally used metabolic reconstructions. While a stoichiometric representation of a reaction involving over 100 substrates is straightforward, a kinetic representation of such a reaction would not be practical. For that reason, the metabolic reconstruction should be modified by splitting the biomass reaction into many reactions involving one single substrate. Alternative strategies, in which related substrates are grouped in a single reaction are also possible, for instance to represent DNA formation.

For all kinetic-based analyses, we used an extended version of the *E. coli* core metabolism [107]. Refer to Appendix B.1.1 for more details on model modification. Since the EM framework allows to incorporate known substrate-level regulation of enzyme activity, we used the regulatory interactions contained in Table 2.3 for the construction of rate laws for all ensembles. Additionally, we used the rate law described by Equation 2.26 to describe the flux through the phosphotransferase system (PTS) system, as suggested by Lafontaine Rivera et al. [71]. Note that the rate law described in the Equation 2.26 includes known regulatory interactions, such as flux control through the PEP/Pyruvate ratio [22, 33, 109] and product inhibition by glucose-6-phosphate [22, 29, 59].

$$r_{PTS} = \frac{V_{m,PTS} C_{PEP}/C_{PYR} C_{GLC}}{(k_{m,1} + k_{m,2} C_{PEP}/C_{PYR} + k_{m,3} C_{GLC} + C_{GLC} C_{PEP}/C_{PYR})(1 + C_{G6P}/k_{i,G6P})}. \quad (2.26)$$

## Parallelization Strategies in Large Multi-core Computers and Computer Clusters

EM-based analyses are computationally expensive. Luckily, the process of ensemble construction by parameter sampling and the analysis itself of each model within a given

## 2. Metabolic Modeling & Strain Engineering

Table 2.3.: **Substrate-level Enzyme Regulation Considered for the *E. coli* Core Model.** Regulatory interactions were obtained from reference [71] and were used during the automatized rate law construction within the EM framework.

Pathway	Enzyme	Reaction	Regulators
<b>Pyruvate Metabolism</b>	Acetaldehyde dehydrogenase (acetylating)	ACALD	AMP[1], NAD[2]
	D-lactate dehydrogenase	LDH_D	ATP[1], PYR[4]
	Phosphotransacetylase	PTAr	ADP[1], ATP[1], NADH[1], NADPH[1], PEP[2], PYR[2]
<b>Citric Acid Cycle</b>	Citrate synthase	CS	AKG[1], ACCOA[4], ATP[1], NADH[3]
	Fumarase	FUM	CIT[1]
	Isocitrate dehydrogenase (NADP)	ICDH <sub>yT</sub>	OAA[1], PEP[3]
<b>Glycolysis/ Gluconeogenesis</b>	Fructose-bisphosphate aldolase	FBA	CIT[2], PEP[2]
	Fructose-bisphosphatase	FBP	ADP[1], AMP[1], PEP[1]
	Glyceraldehyde-3-phosphate dehydrogenase	GAPD	PEP[1]
	Pyruvate dehydrogenase	PDH	NAD[1], NADH[1], PYR[1]
	Phosphofructokinase	PFK	ADP[4], ATP[3], FRDP[1], F6P[4], PEP[3]
	Glucose-6-phosphate isomerase	PGI	PEP[1]
	Phosphoenolpyruvate synthase	PPS	AKG[1], ADP[1], AMP[1], OAA[1], PEP [1]
	Pyruvate kinase	PYK	AMP[4], ATP[1], F6P[4], SUCCOA[1]
<b>Pentose Phosphate Pathway</b>	Hexokinase	HEX1	PEP[1]
	Glucose 6-phosphate dehydrogenase	G6PDH2r	NADH[1], NADPH[1]
<b>Glutamate Metabolism</b>	Ribose-5-phosphate isomerase	RPI	AMP[1]
	Glutamine synthetase	GLNS	AMP[1]
<b>Anaplerotic reactions</b>	Glutamate dehydrogenase (NADP)	GLUD <sub>y</sub>	GLN[1]
	Isocitrate lyase	ICL	PEP[1], G3P[1]
	Malic enzyme (NADP)	ME2	ACCOA[3], ACTP[2], FUM[1], G6P[2], NADH[3], NADP[1], OAA[3]
	Phosphoenolpyruvate carboxylase	PPC	ACCOA[4], FDP[4], MAL-L[3]

Substrate-level enzyme regulation is coded as follows: [1]: competitive inhibition; [2]: non-competitive inhibition; [3]: allosteric inhibition; [4]: allosteric activation.

Substrate abbreviations as follows: AKG: 2-Oxoglutarate; PYR: Pyruvate; CIT: Citrate; OAA: Oxaloacetate; FRDP: D-Fructose 1,6-biphosphate; F6P: Fructose 6-phosphate; MAL-L: L-Malate; FUM: Fumarate; ACTP: Acetyl phosphate; G6P: Glucose 6-phosphate; G3P: Glyceraldehyde 3-phosphate; GLN: Glutamine; SUCCOA: Succinyl-CoA;

ensemble are highly parallelizable. This means that the computation time can be dramatically reduced through the implementation of either multi-core computers or computer clusters (See Figure 2.6). In this work, both parallelization strategies were applied to perform kinetic-based analysis. Therefore, we will briefly discuss both strategies using a Matlab implementation of the EM approach. In the case of large multi-core computers, parallelization can be achieved within the code, by simply replacing the *for*-command with a *par for*-command. By doing so, Matlab will automatically use all available cores to analyze the ensemble. When using computer clusters, a similar strategy can be applied. The only difference is that instead of running the EM code on a single computer, it is run on multiple multi-core nodes (see Figure 2.6). This is conveniently done by instructing Simple Linux Utility for Resource Management (SLURM) to run the EM code on  $k$  different nodes. Consequently, each node will independently sample and analyze  $n/k$  models. Here,  $n$  refer to the total number of models in the ensemble. There are two peculiarities of this approach that require special attention.

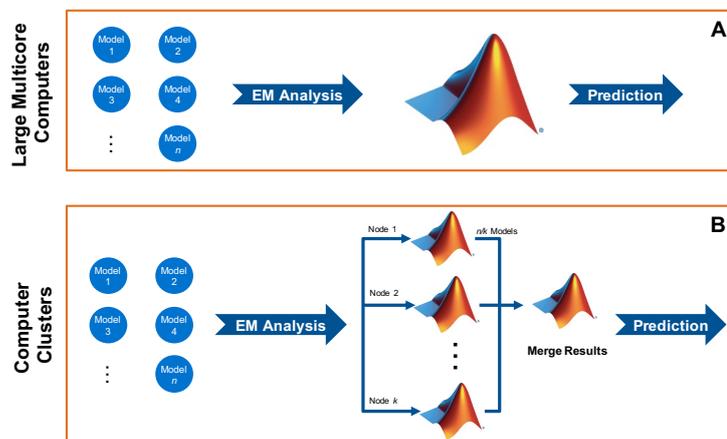


Figure 2.6.: **Parallelization Strategy in Linux Cluster.** **A.** Computational strategy for large multi-core computers. Parallelization is achieved easily by using the Matlab command *par for* **B.** Computational strategy for computer clusters with multiple or single cores. Notice that the command *rng('shuffle')* is necessary to guarantee that each model has a unique set of kinetic parameters. Also, an additional step to merge individual results from each node into one single result might be necessary.

1. *Control random number generation.* Place the command *rng('shuffle')* before the ensemble is populated by randomly sampling the parameter space to ensure that each node is generating and analyzing *unique* models. Failure to do so will cause all nodes to sample exactly the same parameter values. As a consequence, the effective size of the ensemble will be reduced from  $n$  to  $n/k$  *unique* models.
2. *Individual results need to be merged.* When using the parallelization strategy shown in Figure 2.6 for computer clusters, an additional step is required to merge the results from  $k$  different nodes into one single result. Depending on the kind of analysis being performed, this step might require a large amount of Random Access Memory (RAM). Consider using the *hugemem* partition of the LRZ Linux Cluster, which offers up to 240 GBytes in a single shared memory node.

## 2.4. Strain Design Algorithms

One of the main goals of constructing mathematical models of cellular metabolism is the identification of non-intuitive genetic interventions aiming at the over-production of a target molecule. OptKnock [20] was one of the first algorithms developed to systematically identify gene knockout candidates for strain optimization. This strain design algorithm identifies growth-coupled, non-intuitive reaction deletions strategies by solving a bi-level optimization problem using the linear programming duality theory. Many strain design algorithms were developed based on the ideas introduced by OptKnock during the past decade. Table 2.4 presents an overview of some of those and other strain design algorithms. Interestingly, all these approaches used the overproduction of succinate as a case study. Since there is a large body of successful experimental optimization strategies to overproduce succinate in various strains, this knowledge has been repetitively used to test the predictions of newly introduced strain design algorithms. Table 2.4 was reproduced from [149]. For more information on additional strain design algorithms, refer to Machado et al. [87] or Maia et al. [89]

Table 2.4.: **Theoretical Approaches Used in the *in silico* Strain Optimization for the Succinate Overproduction.** A total of 26 theoretical studies were performed in the period of time from year 2003 to 2016.

Model type	Algorithm Name	Year	Organism	Description	Ref.
<i>Kinetics-based</i>	k-OptForce	2014	<i>E. coli</i>	k-OptForce integrates available kinetic information with stoichiometric metabolic models. It enables identification of both enzymatic parameter changes and reaction flux changes. Data for model parametrization can limit the prediction capabilities of the algorithm.	[28], [62]
	OptKnock	2003	<i>E. coli</i>	OptKnock identifies growth-coupled, non-intuitive gene deletion strategies for the over-production of target chemicals by solving a bi-level optimization problem using the linear programming duality theory.	[20]
<i>Constraint-based</i>	OptGene	2005	<i>S. cerevisiae</i>	OptGene is an evolutionary programming based algorithm that permits solving large gene knockout problems in relatively short computational time. It allows the optimization of non-linear objective functions, such as the productivity.	[108]
	GDLS	2005	<i>E. coli</i>	GDLS overcomes the computational burden of bi-level optimization problems, such as OptKnock, by employing an heuristic approach based on local search with multiple search paths.	[85]

Table 2.4 Continued from previous page

Model type	Algorithm Name	Year	Organism	Description	Ref.
	MOFBA	2009	<i>E. coli</i>	MOFBA is a multiobjective flux balance analysis method. The non-inferior set estimation method is used to generate an approximation of the pareto curve for conflicting objectives. MOFBA reduces the computation time required to obtain the pareto curve and its analysis.	[104]
	RobustKnock	2010	<i>E. coli</i>	RobustKnock accounts for the presence of competing pathways in the metabolic network. This method extends OptKnock to identify reactions that should be removed so that the production of the desired product becomes an obligatory byproduct of growth.	[143]
	OptForce	2010	<i>E. coli</i>	OptForce makes use of available flux measurements for the wild-type strain. The algorithm ultimately identifies sets of fluxes that must actively be forced through genetic manipulations in order to ensure the overproduction of the target molecule.	[112]
	SimOptStrain	2011	<i>E. coli</i>	SimOptStrain is a bi-level strain design approach that uses mixed-integer programming. The approach simultaneously considers gene deletion and non-native reaction addition, which enabled the identification of novel strategies with higher predicted production levels.	[64]
	EMILiO	2011	<i>E. coli</i>	EMILiO uses successive linear programming to increase the scope of strain design. The algorithm can quantitatively predict the optimal flux ranges that maximize production. EMILiO was used to generate over 200 strain designs for succinate production.	[161]
	BAFBA	2012	<i>E. coli</i>	BAFBA is an hybrid approach of the Bees Algorithm and Flux Balance Analysis. The algorithm prevents falling into local minima while searching for optimal strain designs.	[25]
	GDMO	2012	<i>E. coli</i>	GDMO is a heuristic and combinatorial multi-objective optimization method that globally searches for genetic manipulations that optimize multiple cellular functions. The algorithm implements a genetic algorithm to find pareto-optimal solutions.	[30]
	GDBB	2012	<i>E. coli</i>	GDBB is an approach developed to find near-optimal gene knockout strategies. It implements an adaptation of the branch and bound algorithm and can be applied to find near-optimal solutions in seconds to minutes instead of days or more.	[35]
	OptSwap	2013	<i>E. coli</i>	OptSwap is a computational method which identifies optimal modifications of the cofactor specificities (NAD(H) and NADP(H)) of oxidoreductase enzymes and complementary reaction knockouts.	[65]

## 2. Metabolic Modeling & Strain Engineering

Table 2.4 Continued from previous page

Model type	Algorithm Name	Year	Organism	Description	Ref.
	ReacKnock	2013	<i>E. coli</i>	ReacKnock is a computational algorithm that uses the Karush-Kuhn-Tucker method to reformulate a bilevel linear programming to a single level programming problem. The algorithm does not return a single solution but all alternative deletion strategies.	[159]
	MOMAKnock	2013	<i>E. coli</i>	MOMAKnock is a bi-level optimization framework that identifies knockout strategies under the MOMA flux distribution approximation, which considers that knockout mutants do not reach a steady state with maximum growth rates.	[115]
	DySScO	2013	<i>E. coli</i>	DySScO integrates dynamic flux balance analysis with existing strain design algorithms. This allows to consider process-level variables such as productivity and titer in the optimization process.	[170]
	FastPros	2013	<i>E. coli</i>	FastPros is an iterative screening algorithm that is based on shadow price analysis. It allows fast identification of strain designs with a large number of reaction knockouts.	[105]
	DBFBA	2014	<i>E. coli</i>	DBFBA improves the performance of BAFBA by using a differential evolution algorithm. DBFBA has shorter computational time and identifies strain designs with higher production yields. DBFBA was validated with OptKnock.	[27]
	FSA	2015	<i>E. coli</i>	Flux-Sum Analysis (FSA) aims at identifying engineering targets from a metabolite-centric perspective. The idea is to identify metabolites that will force the overproduction of a target molecule after attenuation or intensification of their corresponding turnover or "flux-sum".	[72]
	BHFBA	2015	<i>E. coli</i>	Bees Hill Flux Balance Analysis (BHFBA) is an hybrid of Hill climbing and the neighborhood searching strategy of BAFBA [25]. It integrates optKnock for automatic results validation. BHFBA maximizes the production yield of desired phenotypes while sustaining the growth rate.	[26]
	HyMeP	2016	<i>Synechosystis</i> sp. 6803	Hybrid Metabolic Pathway Design (HyMeP) allows exogenous reactions to be considered in the optimization process. It selects individual candidate reactions from the metabolic reactions in the Kyoto Encyclopedia of Genes and Genomes (KEGG) to construct a computerized metabolic model of a hybrid cell.	[137]

Table 2.4 Continued from previous page

Model type	Algorithm Name	Year	Organism	Description	Ref.
	SolveME	2016	<i>E. coli</i>	A Metabolism and macromolecular Expression (ME) model describes not only the metabolic network but also the transcription and translation machinery associated with the genes considered in the model. ME models enable the computation of proteome allocation at different conditions and predict the macromolecular composition of the cell.	[162]
<b>Elementary Modes-based</b>	CASOP	2010	<i>E. coli</i>	CASOP is based on reaction importance measures derived from weighted elementary modes. The method allows the identification of knockout and overexpression candidates considering explicitly the product yield and network capacity.	[51]
	CASOP GS	2010	<i>E. coli</i>	CASOP GS is an improvement of the CASOP method and allows it to be used in genome-scale networks. CASOP GS uses a sampling procedure of the elementary flux modes to estimate the CASOP-scores at the genome-scale.	[18]
	iStruF	2013	<i>E. coli</i> and <i>S. cerevisiae</i>	iStruF introduces the concept of structural fluxes as good predictors of <i>in vivo</i> flux measurements. The algorithm is able to find growth-coupled strain designs while considering optimal and sub-optimal routes and their efficiencies.	[138]
	SSDesign	2014	<i>E. coli</i>	The Solution Space Design (SSDesign) algorithm identifies growth-coupled and non-growth-coupled strain designs. SSDesign permits to visually set the desirable solution space in advance and identifies the necessary knockouts.	[145]

For the sake of simplicity, we have categorized the algorithms presented in Table 2.4 into three main categories: kinetic-, constraint- and elementary modes-based methods. In reality, this distinction is not always practicable, since there are many algorithms combining different types of methods. For instance, thermodynamics-based methods can be embedded into constraint-based approaches to generate thermodynamically feasible flux distributions on a genome scale. This is done by considering an additional set of linear thermodynamic constraints [56]. More recently, a method called thermodynamic elementary flux modes analysis (tEFMA) was introduced to avoid the enumeration of thermodynamically infeasible elementary flux modes [46]. In this way, tEFMA strongly reduces memory usage and overall computational time of the enumeration process, while focusing on biologically relevant elementary flux modes.

The development of frameworks integrating constraint-, kinetic- and thermodynamic-based methods is becoming popular. For instance, the Ensemble Modeling (EM) approach

## 2. Metabolic Modeling & Strain Engineering

[141,142,146], which introduced an elegant solution for difficulties related to unambiguous parameter estimation of large-scale kinetic models, has the potential to integrate constraint-based methods like FBA to identify biologically meaningful steady-state flux distributions, if no  $^{13}\text{C}$ -labeling data is available. Additionally, thermodynamic constraints can be applied not only to complement FBA calculations [56], but also to perform thermodynamically safe parameterization using independent parameter sets [77–79]. In this way, kinetic parameters will satisfy Haldane relationships and Wegscheider conditions, necessary to ensure thermodynamical consistency of the kinetic model [79]. In a more recent work, Sehr et al. [135] proposed a modeling workflow that integrates FBA with Thermodynamic Feasibility Analysis [94]. The results of both approaches are then translated to parameters for a dynamic model. The dynamic model is then formulated and analyzed according to Biochemical Systems Theory [118–120] and Mathematically Controlled Comparison [3, 121]. The authors used this approach to analyze three different case studies. We refer interested readers to the original publication for further details on mathematical implementation and results [135].

### 2.5. Computational Tools

As shown in Table 2.1, the size of the stoichiometric matrix  $\mathbf{S}$  [metabolites x reactions] of commonly used metabolic reconstructions can rapidly increase as the metabolic reconstruction becomes more comprehensive. In order to effectively handle the size and complexity of genome-scale metabolic reconstructions, many computational approaches have been developed. Some prominent examples are listed in Table 2.5. Out of those methods, the COntstraint-Based Reconstruction and Analysis (COBRA) Toolbox has emerged as one of the preferred modeling tools. The COBRA toolbox consist of a compilation of various Matlab scripts designed to allow for a rapid implementation of basic and advanced analyses. The current functionality allows for network gap filling analysis, determination of intracellular flux distributions by analyzing  $^{13}\text{C}$  labeling data, implementation of strain design algorithms, omics-guided analysis, visualization of calculated flux distributions, among others.

Throughout this work, the functionality provided by the COBRA Toolbox v2.0 [124] was extensively implemented. For instance, FVA was used in Chapter 4 to identify changes in by-product secretion pattern as the substrate uptake rate was increased. In the same chapter, relative quality for various substrates was quantitatively assessed by calculating the maximal flux through the ATPM reaction of the *E. coli* core model [107]. In Chapter 5, the OptKnock implementation provided by the COBRA toolbox was used to analyze the effect of oxygen availability on genetic targets improving taxadiene production. In order to demonstrate the simplicity of application of the COBRA toolbox, we show how the maximal theoretical taxadiene yield on glucose can be calculated using some basic commands. The code shown encompasses the following steps:

1. Load a metabolic reconstruction. In this case, we use the *E. coli* core model. See code line 2.
2. Expand the scope of the metabolic reconstruction by adding a set of new reactions. In this case, 12 new reactions along with their corresponding substrates and products are added to the model. See code lines 6 to 46.

Table 2.5.: **Commonly Used Computational Tools for the Analysis of Metabolic Networks.** Due to the implementation of a large number of different constraint-based analysis, the COBRA toolbox has been widely used within the metabolic modeling community.

Toolbox	Description	Reference
<b>COBRA for Matlab</b>	Matlab package for implementing COBRA (constraint-based reconstruction and analysis) methods to simulate, analyze and predict a variety of metabolic phenotypes using genome-scale models.	[55,124]
<b>COBRA for Python</b>	Python package that provides support for basic COBRA methods. COBRApy includes parallel processing support for computationally intensive processes.	[34]
<b>CellNetAnalyzer</b>	Matlab toolbox that provides a graphical user interface and various computational methods and algorithms for exploring structural and functional properties of metabolic, signaling and regulatory networks.	[67]
<b>SNA: Stoichiometric Network Analysis</b>	Interactive, high-performance toolbox for analyzing steady-state behavior of metabolic networks. The toolbox is mainly implemented in Mathematica.	[148]
<b>YANA</b>	Platform-independent, dedicated toolbox for metabolic networks with graphical user interface to calculate, edit, visualize, centralize and compare elementary flux modes.	[131]

3. Set an objective function. Since we are interested in calculating the maximal taxadiene yield, the taxadiene export reaction is set as objective function. See code line 49.
4. Constrain exchange reactions to reflect medium composition or experimentally determined rates. In this case, we set the glucose uptake rate to a value of 1 mmol/gDW h and allow the network an unlimited oxygen availability. See code line 52 and 53.
5. Calculate flux distribution optimizing the objective function set in step 3. Since we constrained the glucose uptake rate to a value of 1 mmol/gDW h, the maximal value of the objective function would correspond to the maximal theoretical taxadiene yield on glucose. See code line 57.

```

1 % Load E.coli core metabolic reconstruction
2 model=readCbModel('ecoli_core_model.xml');
3
4 % Add Non-Mevalonate Pathway:
5 % Reaction 1. - dxs
6 model=addReaction(model,'dxs','pyr[c] + g3p[c] + h[c] -> dxy15p[c] + co2[c]');
7
8 % Reaction 2. ispC
9 model=addReaction(model,'ispC','dxy15p[c] + nadph[c] + h[c] -> nadp[c] + 2me4p[c]');
10
11 % Reaction 3. ispD
12 model=addReaction(model,'ispD','2me4p[c] + atp[c] + h[c] -> 4c2me[c] + 2 pi[c] '); %Eigentlich
    ist die Reaktion mit CTP.
13
14 % Reaction 4. ispE
15 model=addReaction(model,'ispE','4c2me[c] + atp[c] -> 2p4c2me[c] + adp[c] + h[c] ');
16
17 % Reaction 5. ispF
18 model=addReaction(model,'ispF','2p4c2me[c] -> 2mecdp[c] + amp[c] ');
19
20 % Reaction 6. ispG
21 model=addReaction(model,'ispG','2mecdp[c] + nadh[c] -> h2mb4p[c] + h2o[c] + nad[c] ');
22
23 % Reaction 7. ispH

```

## 2. Metabolic Modeling & Strain Engineering

```
24 model=addReaction(model,'ispH',' h2mb4p[c] + nadh[c] + h[c] -> dmpp[c] + h2o[c] + nad[c] ');
25
26 % Reaction 8. ispH2
27 model=addReaction(model,'ispH2',' h2mb4p[c] + nadh[c] + h[c] -> ipdp[c] + h2o[c] + nad[c] ');
28
29 % Reaction 9. idi
30 model=addReaction(model,'idi',' dmpp[c] <=> ipdp[c] ');
31
32 % Add Taxadiene Synthesis:
33 %Reaction 10. - Geranyl diphosphate formation.
34 model=addReaction(model,'DMAT',' dmpp[c] + ipdp[c] -> grdp[c] + 2 pi[c] ');
35
36 %Reaction 9. - Farnesyl diphosphate formation
37 model=addReaction(model,'GRIT',' grdp[c] + ipdp[c] -> frdp[c] + 2 pi[c] ');
38
39 %Reaction 10. - GGPP formation
40 model=addReaction(model,'ggpps',' frdp[c] + ipdp[c] -> ggpp[c] + 2 pi[c] ');
41
42 %Reaction 11. - Taxadiene formation
43 model=addReaction(model,'txs',' ggpp[c] -> txdn[c] + 2 pi[c] ');
44
45 %Reaction 12. - Taxadiene export
46 model=addReaction(model,'txdnx',' txdn[c] -> '); %txdn[e]
47
48 % Set Taxadiene export as objective function.
49 model=changeObjective(model,'txdnx');
50
51 % Set availability of glucose and oxygen.
52 model=changeRxnBounds(model,'EX_glc(e)',-1,'1');
53 model=changeRxnBounds(model,'EX_o2(e)',-100,'1');
54
55 % Solve the linear problem. Flux through taxadiene export reaction
56 % corresponds to molar yield on glucose.
57 s=optimizeCbModel(model);
```

Although the COBRA toolbox offers an implementation for dFBA simulations, this code is not suitable for large-scale calculations due to its limited speed. Many computational tools have been introduced to perform dFBA analysis [47,92,124,169]. Among those, we used the Matlab-based DFBALAB [47] code to perform all dFBA simulations presented in Chapter 8 due to its velocity and simplicity of use. Similarly, even though COBRA also offers an implementation for  $^{13}\text{C}$ -MFA, it requires the use of the commercial nonlinear solver TOMLAB/SNOPT. For that reason, we implemented the 13CFlux2 software instead [155], which uses the large-scale nonlinear open source solver Interior Point OPTimizer (IPOPT) [153]. Refer to Appendix A.2.4 for an exemplary model used for  $^{13}\text{C}$ -MFA calculations using the 13CFlux2 software. As for all EM analyses implemented in this work, the Matlab code provided in the reference [71] was used.

## 3. Theoretical Workflow for Metabolic Engineering

One of the main objectives of this work was the development a theoretical workflow for Metabolic Engineering. Here, we propose such a workflow, incorporating state of the art constraint- and kinetic-based modeling approaches along with the concept of metabolic burden. Our workflow consist of two main steps. The first one covers model-guided development of production strains, while the second step targets model-driven process optimization, as shown in Figure 3.1. Even though the proposed workflow exhibits a sequential nature, with strain engineering preceding process development, other scenarios are also conceivable. For instance, process parameters, e.g., medium composition, production temperature, aeration levels, etc., could be previously fixed so that strain optimization should be performed to reach high strain performance under such conditions. The workflow was designed to reflect current trends of rational strain engineering and systems biotechnology [73,117], characterized by cyclic steps of modeling (dry-lab) and experimental work (wet-lab). In line with the main structure of the workflow, this thesis consist of two results parts, one for strain engineering - **Part II** - and one for process optimization - **Part III** -, each being comprised of three chapters. Note that the application of the methods described in each chapter is not limited by its assignment to one of the two main topics, meaning that some model predictions can be applied to guide both strain and process optimization.

### 3.1. Model-guided Strain Engineering

The workflow starts with the selection of a metabolic reconstruction for the organism of interest. In the case of *E. coli*, different models of diverse scope and complexity are available. Some of the most commonly used are listed in Table 2.1. Once a metabolic reconstruction has been chosen, adapted or constructed, an initial *in silico* strain characterization can is performed. By calculating theoretical maximum product yield on single carbon sources or substrate mixtures, optimal medium composition for the synthesis of the desired product can be identified. Analogously, constraint-based approaches can also be applied to aid pathway selection, when more than one metabolic routes are available for the desired bio-transformation. The results of this initial *in silico* characterization can be applied to inform following steps: decide whether the selected host strain meets minimum requirements on product yield to allow for an economically viable process; guide the selection of initial process parameters for process optimization and, define experimental conditions in early steps of strain characterization. Refer to **Chapter 7** for more details on the specific application of these methods to the taxadiene overproduction in *E. coli*. The results of an initial experimental strain characterization are then used to construct a strain-specific metabolic model. This is done by updating model boundaries of exchange reactions corresponding to substrate uptake, by-products secretion and if available, product

### 3. Theoretical Workflow for Metabolic Engineering

formation. Stoichiometric coefficients of the reaction describing biomass synthesis can also be updated if protein, DNA, RNA or lipids content measurements are available. Calculations regarding *in silico* medium optimization can be repeated and, if necessary, actual medium composition should be adjusted. Additionally, constraint- and kinetic-based strain design algorithms can be implemented to guide genetic interventions aiming at the overproduction of the molecule of interest. Refer to **Chapters 5** and **6** for more details on the specific application of these methods to the taxadiene overproduction in *E. coli*. The next step is experimental strain construction and characterization. Here, genetic interventions suggested by strain design algorithms are experimentally implemented and the resulting strain is characterized to assess its production performance. If the mutant strain meets performance requirements, it is categorized as optimal production strain and the process of strain engineering can be concluded. In the case that strain production performance is not satisfactory, predictive power of the algorithms and methods used for strain design needs to be reviewed. This can be done by thoroughly comparing model predictions with actual strain behavior. The idea is to identify and correct potential model flaws causing discrepancies between *in silico* and *in vivo* behavior. Once model and methods have been reviewed, new predictions are generated. The process of hypothesis generation and experimental validation is repeated until generated mutant strains meet minimal production performance and an optimal production strain is generated.

## 3.2. Model-guided Process Optimization

In the first step of the model-driven process optimization workflow, the production performance of a given production strain is experimentally assessed in a fermentation system for the first time, using a non-optimal initial set of process parameters (e.g., medium composition, production temperature, time point of system induction, growth rate/dilution rate, etc.). The choice of concrete process parameter values can be guided by theoretical tools described in this work. More specifically, the cellular growth rate during the production phase can be selected to consider theoretical aspects related to minimization of metabolic burden (refer to **Chapter 4**) and/or optimal biomass yield resulting in a maximal product yield (refer to **Chapter 9**). Empirical models can also be applied for process optimization, as demonstrated in **Chapter 8**. In this case, a simple model, based on the dynamic flux balance framework, allowed the identification of optimal production temperature and time point of induction, which *in silico* increased the maximal taxadiene productivity by 25%. Additionally, theoretical studies regarding optimal carbon source can be used to guide the process of medium composition optimization, as described in **Chapter 7**. If observed strain performance does not fulfill minimal production requirements, an iterative process, similar to the one applied for strain optimization, can be applied, in which metabolic models are updated so that their predictions match experimental data. Strain performance is then characterized using an improved set of process parameters and the optimization cycle is repeated until strain performance meets minimum production requirements.

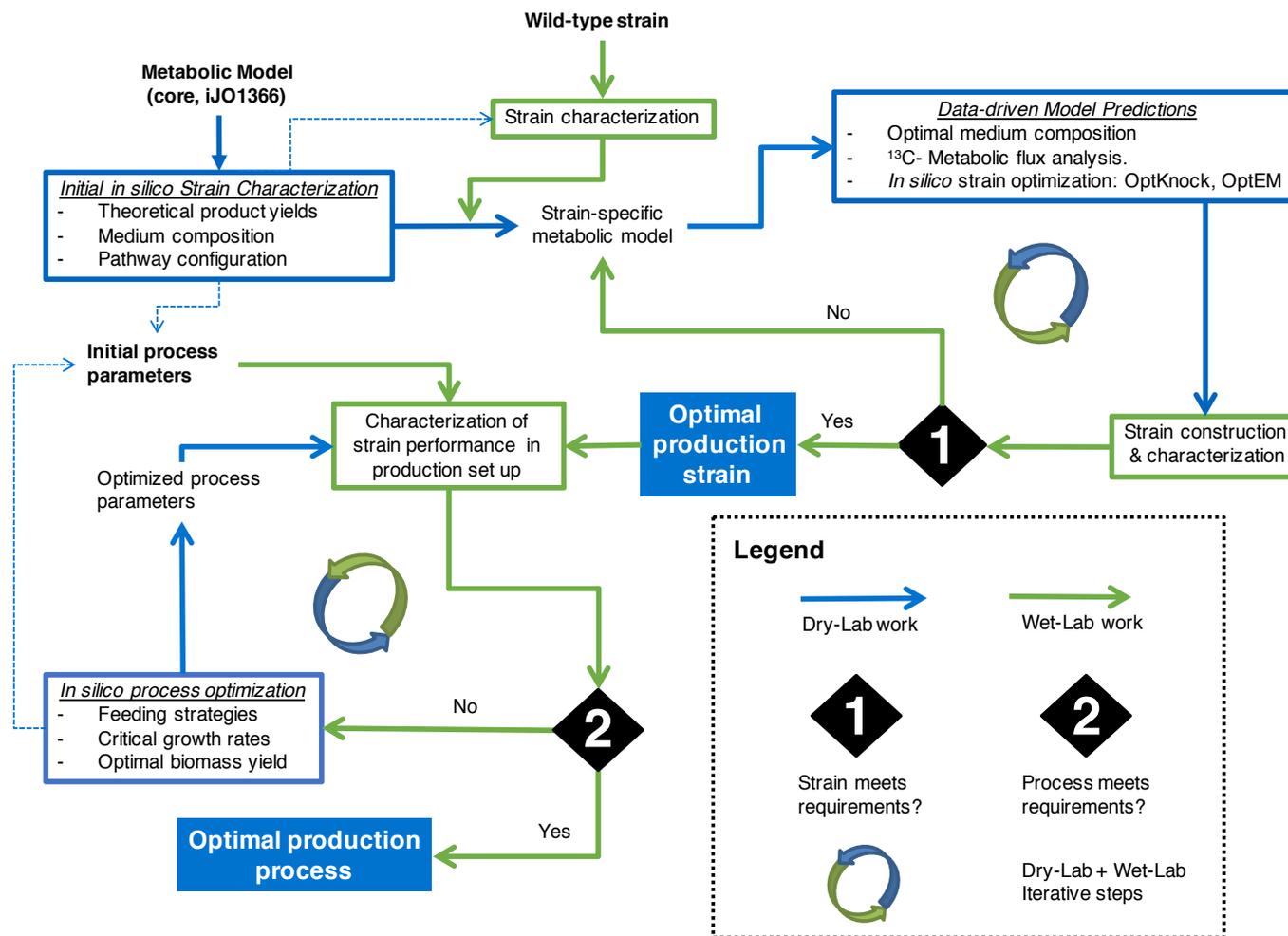


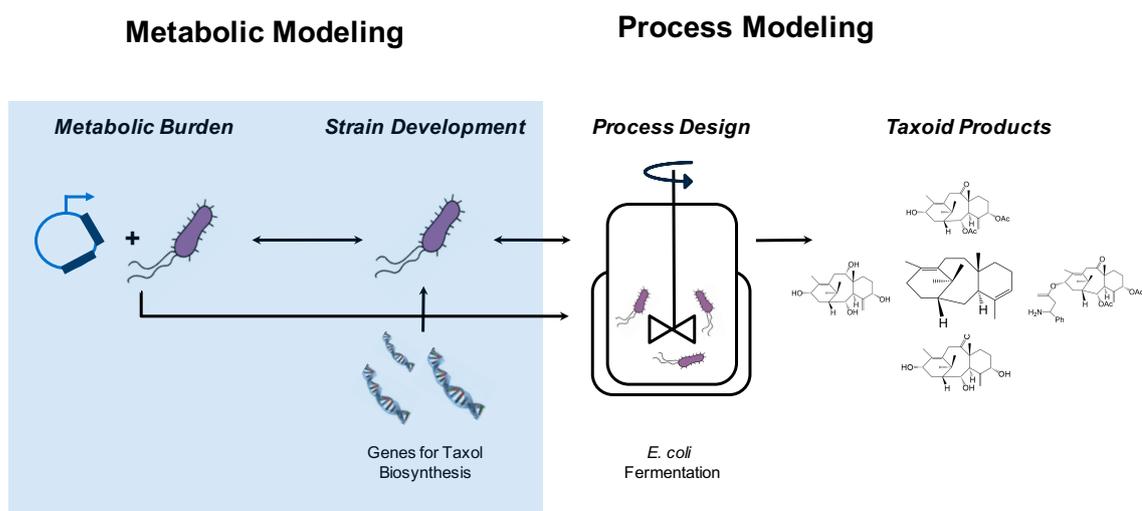
Figure 3.1.: **Workflow for Model-based Metabolic Engineering.** The proposed workflow covers both model-guided strain engineering and model-driven process optimization. Cutting edge constraint- and kinetic-based modeling approaches are implemented in each stage. The workflow was designed to reflect current trends of rational strain engineering and systems biotechnology, characterized by cyclic steps of modeling (dry-lab) and experimental work (wet-lab).



Part II.

Results: Strain Engineering





The second part of this thesis is composed of three chapters, which cover topics related to model-driven strain optimization. **Chapter 4** analyzes the effect of plasmid-based foreign gene expression on strain performance. It is known that loaded cells, i.e., supporting heterologous gene expression, exhibit a decreased growth rate and biomass yield. This phenomenon is commonly termed Metabolic or Protein Burden. We start by describing the concept of acetate and biomass yield lines and by showing how it can be used to monitor the effect of the load on cellular metabolism. Subsequently, we extend this concept and apply it to develop strategies to overcome the metabolic burden caused by heterologous gene over-expression. In **Chapter 5**, we analyze  $^{13}\text{C}$  labeling data for different taxadiene production strains and use them to generate reference flux distributions for each strain, which are subsequently used as input to a kinetic-based approach we termed optEM. OptEM allowed the identification of genetic targets for strain engineering. Finally, in **Chapter 6** we analyze simultaneous utilization of glucose and xylose by *E. coli*. To this end, six different strains exhibiting two different genetic backgrounds were experimentally characterized. All strains were subsequently assessed regarding production potential, metabolic burden and sugar co-uptake level by applying constraint- and kinetic-based approaches.



## 4. Metabolic Burden

In order to analyze the detrimental effect that gene over-expression has on strain performance, our cooperation partner Sabine Wagner at Fachgebiet Systembiotechnologie (SBT) designed, constructed and experimentally characterized three strains, each of which exhibited a different protein load, as shown in Figure 4.1. Strain cultivation was conducted in shaking flasks. Five different carbon sources were used to obtain a relatively wide range of growth rates, ranging from approximately 0.2 1/h for acetate, to almost 0.5 1/h for glucose. We used a constraint-based approach to qualitatively reproduce observed acetate secretion patterns in the three strains. As previously described [11], acetate lines can be used as an indicator of cellular fitness. We found a connection between biomass yield lines and acetate lines. The point where the slope of the biomass yield line is decreased for the first time defines a threshold beyond which cellular growth is always accompanied by acetate secretion under aerobic conditions. By analyzing acetate secretion patterns of strains exhibiting different burden levels, we hypothesize that cells growing below a certain threshold can fulfill the imposed burden, while simultaneously growing with wild-type yields. Our *in silico* analysis also suggests that acetate lines do not have a unique slope as initially suggested, but are rather composed of at least two straight lines with different slopes.

### 4.1. Strains & Experimental Data

Main genetic features of the strains analyzed in this chapter are shown in Figure 4.1. The strains were designed by Sabine Wagner to exhibit an increasing protein burden. The fluorescent protein mCherry was used as a monitor of the cellular capacity, by placing the transcription of the mCherry gene under the control of a constitutive promotor. On the other hand, the protein eGFP was used to load the cell. Since both mCherry and eGFP are fluorescent proteins, their synthesis can be easily determined by measuring light emission at different wavelengths.

Five different carbon sources - glucose, mannose, fructose, galactose and acetate - were used to allow for the analysis of the metabolic burden (caused by plasmid-based heterologous gene expression) at different growth rates. Experimental strain characterization was conducted in shaking flask at 37°C and 200 revolutions per minute (rpm). Additional to light emission at mCherry- and eGFP-specific wavelengths, time courses for OD, substrate and acetate concentration were experimentally determined (see Appendix A.1.3 for raw data). Calculated rates for glucose as carbon source are exemplarily provided in Table 4.1. Refer to Appendix A.1.3, Table A.6 for a complete overview of the results.

In order to analyze the effect of gene over-expression on growth rate and biomass yield, we plot growth rates versus substrate uptake rates for all strains and substrates (Figure

#### 4. Metabolic Burden

Strain ID	31	3150	3152
Plasmid Content	1 	2 	2 
Antibiotics	Kanamycin	Kanamycin Gentamycin	Kanamycin Gentamycin
Inducer	1 mM 3-MB	1 mM 3-MB	1 mM 3-MB
Protein Load			

Figure 4.1.: **Main Genotypic Characteristics of Strains Used for the Characterization of the Metabolic Burden.** Plasmid present in strain 31 correspond to a medium copy number plasmid. Additional plasmid in strain 3150 is a high copy number plasmid. Amino acid sequences of proteins encoded in each plasmid are given in Appendix A.1.2. 3-methylbenzoate (**3-MB**) was used to induce eGFP expression in strain 3152. In order to guarantee identical medium composition for all strains, the culture medium was always supplemented with 1 mM 3-MB

Table 4.1.: **Experimentally Calculated Exchange Rates for Three Strains. Growth on Glucose.** Rates for growth, substrate uptake and acetate secretion were calculated from measured time courses as described in Chapter 2.1.2.

	Strain		
	31	3150	3152
$\mu$ , [1/h]	0.487	0.111	0.086
$r_{uptake}$ , [mmol/gDW h]	14.111	4.449	4.540
$r_{acetate}$ , [mmol/gDW h]	6.074	4.707	3.336

4.2). For a fixed uptake rate of a given substrate, for instance galactose (symbol \* in Figure 4.2), we observe a reduction of growth rate as the protein load is increased, as denoted by the grey arrow. One can explain this observation from a stoichiometric point of view. Heterologous protein production directly competes with cellular growth by reducing the amount of effective substrate available for biomass production. Thus, a reduction of the growth rate will be expected, if heterologous protein production increases at a given constant substrate uptake rate. Of course, mass balance is not the only explanation for the observed behavior. Heterologous protein production can also reduce cellular growth rate by reducing the effective amount of ribosomes available for production of host protein. Additionally and depending on the amino acid composition and identity of the foreign protein, cellular metabolism can be directly affected. For instance, the intracellular pool of a given amino acid can be dramatically reduced or toxic pathway intermediates can accumulate as a consequence of the enzymatic activity of the heterologous protein. In both cases, one would observe a reduction in the cellular growth rate.

Considering the results for strain 31, Figure 4.2 evidences an unexpected, inverse relationship between growth rate and substrate uptake rate: as the value of substrate uptake rate increases, the growth rate seems to decrease. Taking a closer look at Figure 4.2, one can conclude that this behavior mainly arises due to the data point for acetate (symbol +). This observation suggests that in order to properly analyze the correlation between substrate uptake rate and growth rate, we need to standardize the substrate uptake rate,

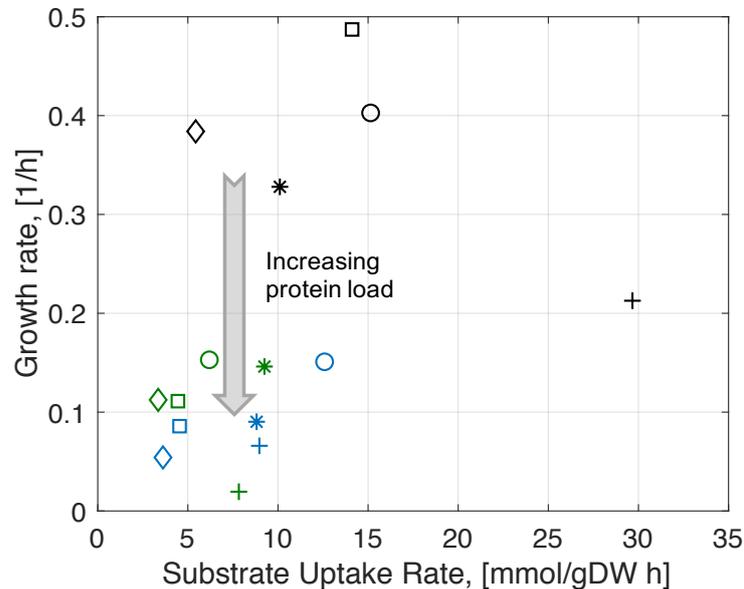


Figure 4.2.: **Growth Rate vs. Substrate Uptake Rate for Three Strains and Five Substrates.** Black symbols correspond to strain **31**, green symbols to strain **3150** and blue symbols to strain **3152**. Substrate symbols are as follows: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose.

to consider differences in the molecular weight and quality of the various substrates used.

## 4.2. Substrate Quality

A real-life analogy for the concept of substrate quality is presented to justify its introduction and to understand its utility. Let us imagine for a moment we are interested in analyzing the correlation between the Gross Domestic Product (GDP) of a given group of countries and the life expectancy of their citizens. One option to graphically analyze this relationship would be to plot the two variables for all countries and determine the slope of the resulting line. Figure 4.3.A shows this relationship for six different countries. Data were collected from the webpage of the World Bank for the year 2016.

Based on Figure 4.3.A, one would conclude that living in rich countries, where the GDP per capita is high, reduces life expectancy. From the figure, we also observe that Colombia appears to be the richest country. Since we know this is not the case, we conclude we first need to standardize the data, more specifically the currency of the GDP per capita, before we can perform any kind of analysis. Figure 4.3.B shows the data after converting the GDP per capita from each local currency to US dollars. After data standardization, we obtain the expected relationship between GDP and life expectancy: the richer the country, the higher the life expectancy. This real-life analogy points out the necessity of adequate conversion factors for each member of the data set. In this case, exchange rates allowed a correct data analysis by adjusting the value of the GDP per capita so that it became comparable within the group. Note that the conversion to another currency different from US\$ is also possible.

#### 4. Metabolic Burden

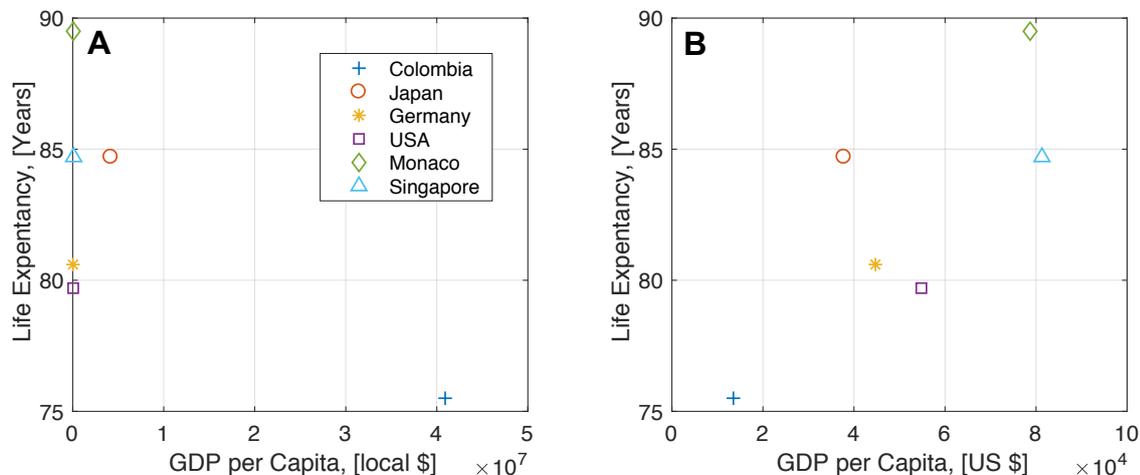


Figure 4.3.: **Real-Life Analogy for Concept of Substrate Quality.** **A.** Life expectancy vs. GDP in local currency. An inverse, unexpected relationship between these variables is observed. **B.** Life expectancy vs. GDP in US\$. As expected, there exist a direct relationship between the two variables

##### 4.2.1. Conversion Factors for Substrate Uptake Standardization

After taking a closer look at the data for strain 31 (Figure 4.2, black symbols), we hypothesize that expressing the substrate uptake rate of each substrate in units of “mmol/gDW h” does not allow a correct data analysis. This is because energy and mass content (or number of carbon atoms) per millimol (mmol) differ for each substrate, especially when acetate is compared with the other substrates. Inspired by this observation, we define three different criteria to describe substrate quality: mass content, carbon content and energy content. In a following step, we set up a theoretical framework to assess the utility of each criterion. In this section, we refer to substrate quality as the potential of a given substrate to allow growth.

##### Determination of Energy Content Using Linear Programming

Conversion factors for the first two quality criteria (molecular weight and carbon content) are solely defined by the chemical identity of each substrate, i.e., they do not depend on culture conditions, for instance, oxygen availability. On the other hand, energy content is a function of both chemical identity of the substrate and oxygen availability. We make use of constraint-based modeling to calculate the energy content of each substrate for various oxygen availability levels by simply maximizing the flux through the ATPM reaction of the iJO1366 model [106]. This reaction describes the dephosphorylation of ATP:  $\text{ATP} + \text{H}_2\text{O} \longrightarrow \text{ADP} + \text{Pi} + \text{H}^+$ . Calculations were performed using the COBRA toolbox and gurobi 5 as the linear solver. Refer to section 2.5 for an overview of the main commands. A constant uptake rate of 1 mmol/gDW h was set for each substrate, while the oxygen uptake rate was varied to obtain oxygen ratios ranging from zero to 20. Throughout this work, we will use the term *ATP content* to refer to the ratio between the flux through the ATPM reaction and the substrate uptake rate. Strictly speaking, this ratio rather refers to the number of high-energy phosphoanhydride bonds that can be

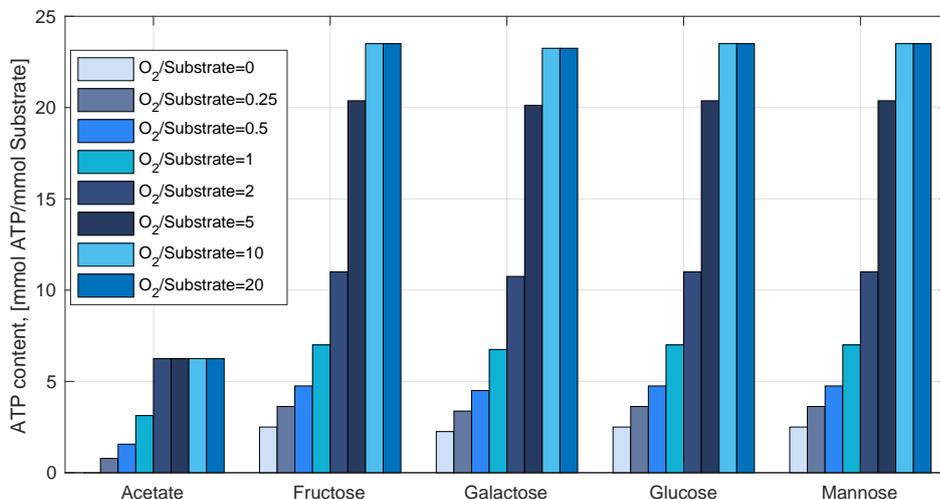


Figure 4.4.: **Energy Content As a Function of Oxygen Availability for Five Substrates.** The *E. coli* metabolic reconstruction iJO1366 was used for all calculations. Energy content is represented by the potential of each substrate to generate energy in form of ATP. Note that anaerobic energy production from acetate is not possible.

Table 4.2.: **Conversion Factors for Three Quality Criteria.** Conversion factors for energy content as quality criteria correspond to maximal ATP content for each substrate, as shown in Figure 4.4. Factors for carbon content correspond to the number of carbon atoms in each molecule and factors for mass content correspond to molecular weight of each substrate in units of g/mmol

	Energy Content [mmol ATP/mmol Substrate]	Carbon Content [C-mmol/mmol Substrate]	Mass Content [g/mmol Substrate]
<b>Acetate</b>	6.25	2	0.06005
<b>Fructose</b>	23.5	6	0.18016
<b>Galactose</b>	23.25	6	0.18016
<b>Glucose</b>	23.5	6	0.18016
<b>Mannose</b>	23.5	6	0.18016

be generated from the given substrate [80]. However, since these phosphate bounds are contained in form of ATP, the term *ATP content* is equally precise.

Figure 4.4 shows the energy content of all five carbon sources. As expected, the ATP content strongly depends on the identity of the substrate and oxygen availability. For all substrates, the ATP content exhibits a saturation behavior as a function of the oxygen ratio. Fructose, galactose, glucose and mannose all reach the same ATP content at an oxygen ratio of 10 mmol O<sub>2</sub>/ mmol substrate, while acetate reaches its maximal ATP content at a oxygen ratio of 2 mmol O<sub>2</sub>/mmol substrate. Assuming that the cell strives to maximize the amount of produced energy by aerobically metabolizing the absorbed substrate, the maximal ATP content should serve as a proper conversion factor when considering energy content as quality criterion. The validity of this assumption will be discussed in section 4.2.2. Table 4.2 contains all conversion factors for the previously defined quality criteria.

Note that for the three quality criteria considered, the conversion factor for acetate is always the lowest out of the five substrates tested. This suggests that the cell should absorb a comparatively higher amount of acetate (measured in mmol/gDW h) in order to

#### 4. Metabolic Burden

support a given growth rate. Considering energy as quality criterion, the cell will have to absorb acetate with a rate 3.76 times higher than the corresponding rate for glucose. This trend is visible for the strain 31 in Figure 4.2, with acetate exhibiting the highest uptake rate. However, this trend is not as clear in strains 3150 and 3152. Having defined three different quality criteria, the natural question following is: which of these criteria is the most suitable to perform standardization of the substrate uptake rate?

#### Assessing the Suitability of Different Quality Criteria

To answer this question, we designed a constraint-based algorithm to assess the suitability of the different quality criteria. The rationale is the following: if uptake rates of a given set of substrates can be correctly standardized by a potential quality criterion, then a given standardized value of substrate uptake rate should lead to a single growth rate, no matter the nature of the substrate. In other words, we require the biomass yield, calculated as the ratio between growth rate and *standardized* substrate uptake rate, to have the same numerical value for all substrates. The proposed algorithm consists of the following steps:

1. Select a quality criterion and determine its corresponding conversion factors. Refer to Table 4.2.
2. Define a numerical value for the standardized uptake rate. For instance, 117.5 mmol ATP/ gDW h.
3. Using corresponding conversion factors, convert the *standardized* uptake rate, into an uptake rate having units of “mmol substrate/gDW h”. In this case, 117.5 mmol ATP/gDW h correspond to: 18.8 mmol acetate/gDW h; 5.05 mmol galactose/gDW h and 5 mmol/gDW h for all other substrates.
4. Set upper and lower boundaries for the uptake reaction of a defined substrate in a given stoichiometric model, for instance iJO1366, and calculate the maximal growth rate supported by this substrate using linear programming.
5. Repeat step 4. for all other substrates to generate a growth vector containing growth rates for all substrates.
6. Calculate the maximal error in the growth vector as  $\frac{\max(\text{growthvector}) - \min(\text{growthvector})}{\min(\text{growthvector})}$
7. Repeat step 1. to 6. for each quality criterion and select the one with the lowest error.

After applying these steps, we obtain the results shown in Figure 4.5. Additionally to the previously mentioned quality criteria, this figure also contains the assessment of a negative control (Figure 4.5.A), where no standardization is used. From Figure 4.5 it results clear that energy content (expressed as ATP) represents the best quality criterion for uptake rate standardization, exhibiting an error of only 1.78%. The performance of mass and carbon content as quality criteria is identical under the tested conditions. An unlimited oxygen supply, id est (i.e.), completely aerobic metabolism, was used to calculate all growth rates shown in Figure 4.5. If the oxygen availability is reduced (from 20 mmol O<sub>2</sub>/ gDW h to 5 mmol O<sub>2</sub>/ gDW h), different results are obtained, as shown in Figure 4.6. While the error of the negative control is reduced to 189%, the error of the three quality criteria becomes identical and increases to 185%.

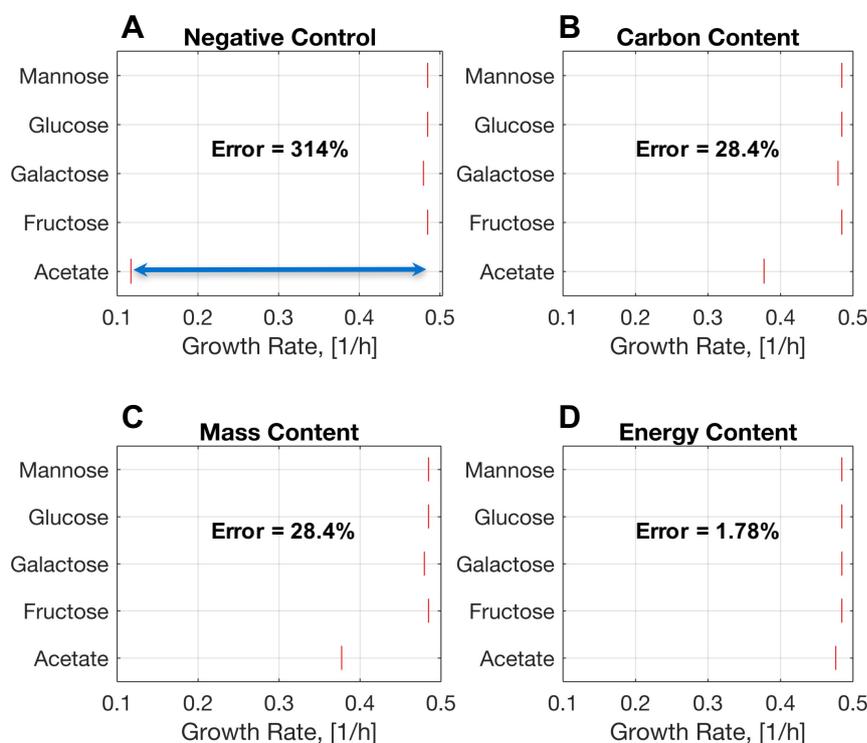


Figure 4.5.: **Assessment of Three Different Quality Criteria for Data Standardization. High Oxygen Availability.** A maximal oxygen uptake rate of 20 mmol  $O_2$ /gDW h was used for all calculations. Uptake rates used to estimate growth rates via linear programming can be extracted from Table 4.3. All calculations were performed using the *E. coli* metabolic reconstruction iJO1366. At a relatively high oxygen to substrate ratio, standardization using energy content as quality criterion seems to be the most accurate approach for standardizing the substrate uptake rate. **A.** Negative control. All substrates had a substrate uptake rate of 5 mmol/gDW h. **B.** Standardization using carbon content as quality criterion **C.** Standardization using mass content as quality criterion **D.** Standardization using energy content as quality criterion.

Table 4.3.: **Substrate Uptake Rates Used to Assess Quality Criteria.** Substrate uptake rates were calculated using conversion factors listed in Table 4.2 and assuming following input fluxes: Negative control: 5 mmol/gDW h for all substrates; Carbon content: 30 C-mmol/gDW h; Mass content: 0.9008 g/ gDW h; Energy content: 117.5 mmol ATP/gDW h. These reference fluxes were selected so that after conversion into molar flux, 5 mmol glucose/gDW are obtained. All uptake rates are given in units of mmol substrate/ gDW h.

	Negative Control	Carbon Content	Mass Content	Energy Content
Acetate	5	15	15.00030	18.8
Fructose	5	5	5	5
Galactose	5	5	5	5.0538
Glucose	5	5	5	5
Mannose	5	5	5	5

#### 4. Metabolic Burden

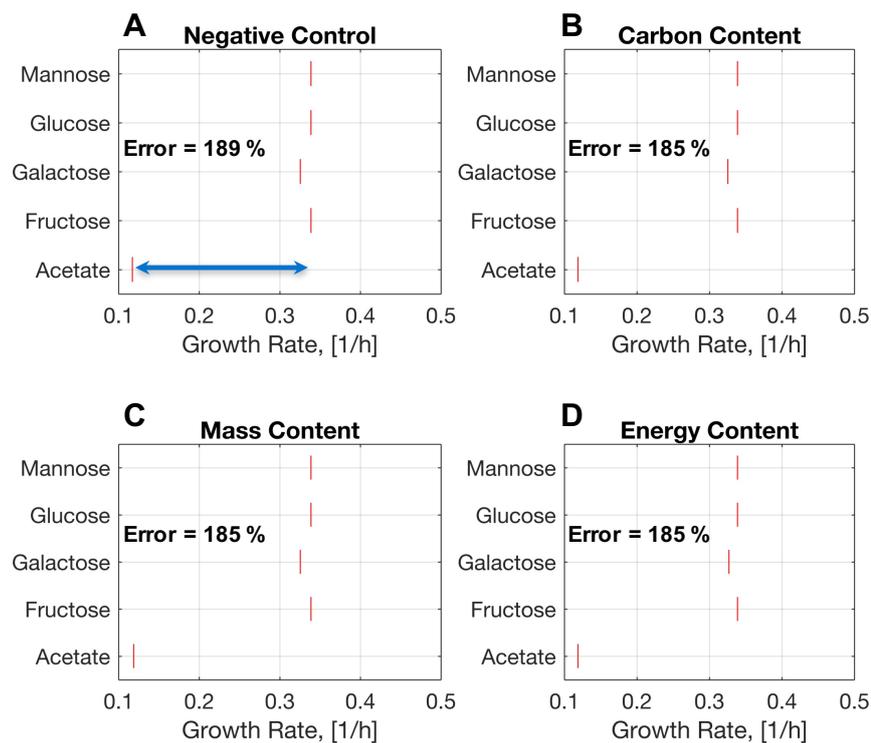


Figure 4.6.: **Assessment of Three Different Quality Criteria for Data Standardization at Low Oxygen Availability.** A maximal oxygen uptake rate of 5 mmol  $O_2$ /gDW h was used for all calculations. Uptake rates used to estimate growth rates via linear programming can be extracted from Table 4.3. All calculations were performed using the *E. coli* metabolic reconstruction iJO1366. At a relatively low oxygen to substrate ratio, none of the three standardization approaches tested seems to reduce the error, when compared to the error of the negative control. **A.** Negative control. All substrates had a substrate uptake rate of 5 mmol/gDW h. **B.** Standardization using carbon content as quality criterion **C.** Standardization using mass content as quality criterion **D.** Standardization using energy content as quality criterion.

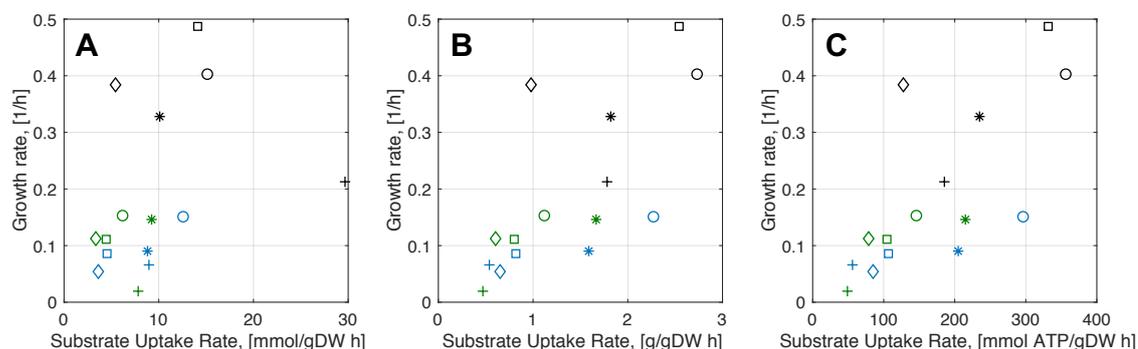


Figure 4.7.: **Standardized versus Raw Experimental Data.** Black symbols correspond to strain 31, green symbols to strain 3150 and blue symbols to strain 3152. Substrate symbols are as follows: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose. **A.** Raw experimental data **B.** Substrate uptake rate was standardized using mass content as quality criterion **C.** Substrate uptake rate was standardized using energy content (ATP) as quality criterion. Note that the relative position of the acetate data point is the only appreciable difference between standardization using mass (**B.**) or energy content (**C.**)

#### 4.2.2. Data Standardization

As previously shown, oxygen availability, commonly expressed as oxygen to substrate ratio, has a big impact on numerical values of conversion factors, when energy content is used as quality criterion (refer to Figure 4.4). As it turned out, oxygen availability also impacts the performance of a given set of conversion factors, no matter which quality criteria was used (refer to Figure 4.5 and Figure 4.6). In order to understand the nature and implication of these observations, consider one more time the real life-analogy presented at the beginning of section 4.2. The relative value of a given currency to another, that is, its exchange rate, is normally determined by market forces of supply and demand, which in turn depend on plenty of factors. In an ever-changing world, exchange rates are constantly changing to represent the relative value of a given currency accurately. Trying to predict their exact behavior has proven to be a challenging, if not impossible endeavor. Contextualizing these observations, we infer that in the case of energy as quality criterion, conversion factors need to be adjusted in order to reflect the effect of variable factors like for instance, oxygen availability. However, oxygen availability is normally unknown, especially when culturing in shaking flasks. Additionally, it has been shown that factors like medium composition, specific growth rate or heterologous protein over-expression also have a profound effect on the overall cellular fitness, consequently affecting the amount of oxygen that the cell can absorb and effectively use to fully metabolize the absorbed substrate (refer to [11] and [110], Figure 6 therein). In other words, the calculation of conversion factors should not only be process-specific, but also strain-specific. Due to the impossibility of an accurate determination of conversion factors using energy content as quality criterion, we opt for choosing mass content as quality criterion.

Figure 4.7 (B and C) shows the effect that standardizing substrate uptake rate has on data interpretation. By changing the units in which substrate uptake rate is plotted, the correlation between growth rate and substrate uptake rate went from being negative to become positive. Note that although the numerical value for the uptake rate is changed for all substrates as a consequence of standardization, the data point for acetate underwent

#### 4. Metabolic Burden

the biggest change in its relative position. Acetate, exhibiting the highest uptake rate *before* standardization, became the substrate with the second lowest uptake rate *after* standardization. This behavior is valid for strain 31, however, strains 3150 and 3152 show a similar pattern.

### 4.3. Yield, Acetate and Formate Lines

From previous constraint-based analysis, in which the dependence of growth rates on glucose uptake rates at varying oxygen levels was investigated, we learned that at a constant oxygen availability, an increase in the glucose uptake rate is always followed by an increase in growth rate. Similarly, at a constant glucose uptake rate, a decrease in the oxygen ratio is accompanied with a reduction in the growth rate, as shown in Figure 4.8.B and .D. We refer to the lines shown in both figures as yield lines, since their slopes correspond to the biomass yield. A comparison between *experimentally* obtained yield lines for strains 31, 3150 and 3152 (Figure 4.8.A and .C) with *in silico* generated ones - using glucose as carbon source at different oxygen levels - evidences an intriguing similarity (compare Figure 4.8.A with .B and Figure 4.8.C with .D). Thus, we venture to hypothesize that heterologous protein over-expression in strains 3150 and 3152 might originate a load-induced oxygen limitation, which is experimentally evidenced as a downward displacement of the biomass yield lines.

One of the advantages of metabolic modeling is that it allows the analysis of complex responses of the cellular metabolism to changes in culture conditions. We used FVA in combination with a core metabolic model [107], describing the central metabolism of *E. coli*, to investigate the reason for the systematic reduction of biomass yield at higher substrate uptake rates observed in Figure 4.8.B and .D. For our analysis, we focused on one specific *in silico* yield line, shown in Figure 4.9. This yield line is divided into three phases, with each phase exhibiting decreasing slopes (biomass yield) as the substrate uptake rate increases. The analysis was performed as follows:

1. Select a combination of growth and substrate uptake rate lying in one of the three phases.
2. Fix the flux through those reactions in the model by adjusting lower and upper boundaries correspondingly.
3. For each reaction in the model, calculate the maximal and minimal flux supported, while fulfilling constraints set in the previous step.
4. Repeat step 1 to step 3 for the other two phases.
5. Compare flux distribution ranges for all three phases, paying special attention to by-product secretion patterns.

This analysis identified clear differences in the by-product secretion patterns of the three phases. While CO<sub>2</sub> is the sole by-product secreted in the first phase, its secretion is complemented by acetate in the second phase and by acetate and formate in the third phase (refer to Appendix A.1.4, Figures A.2 to A.4, for a detailed overview of results). The changing by-product secretion pattern can be understood as a consequence of decreasing

### 4.3. Yield, Acetate and Formate Lines

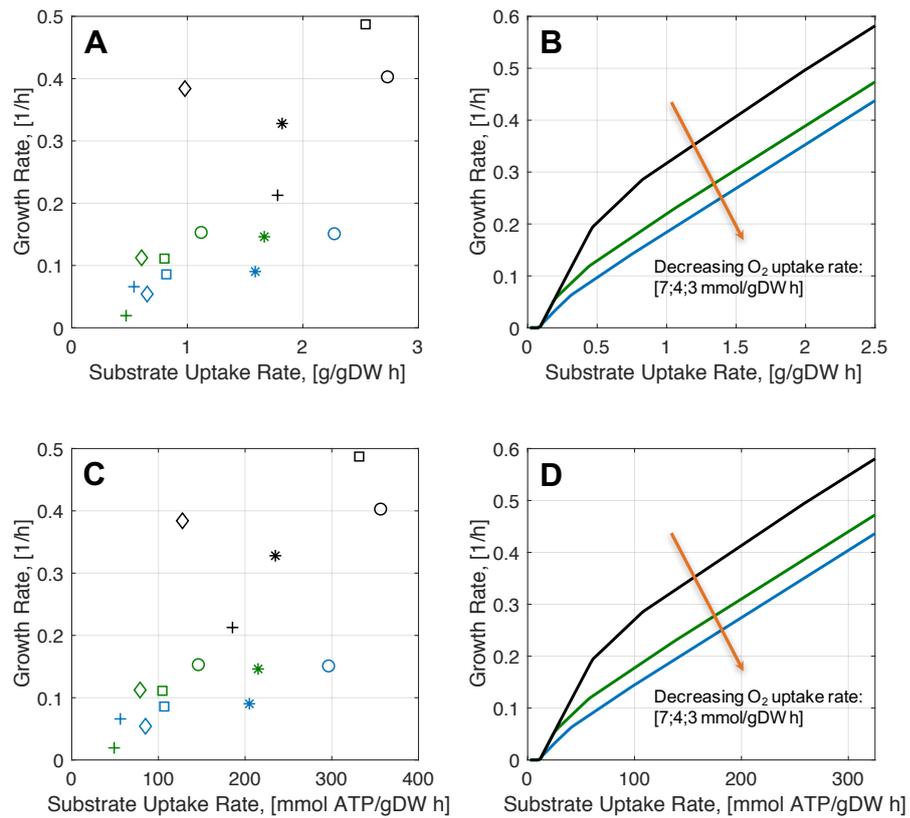


Figure 4.8.: **Load-induced Oxygen Limitation.** **A.** and **C.** correspond to standardized experimental data using mass and energy content as quality criteria, respectively. **B.** and **D.** correspond to *in silico* simulations using glucose as carbon source and the *E. coli* core model. Growth rates were calculated using constraint-based modeling with maximization of growth rate as the objective function for glucose uptake rates ranging from 0 to 14 mmol/gDW and three different oxygen levels. Black symbols correspond to strain 31, green symbols to strain 3150 and blue symbols to strain 3152. Substrate symbols are as follows: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose.

#### 4. Metabolic Burden

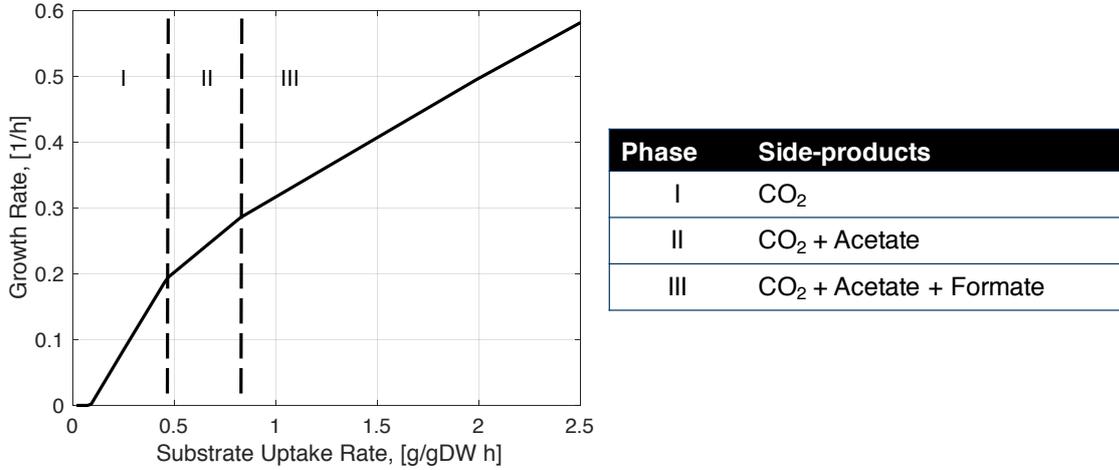


Figure 4.9.: **Relationship between Biomass Yield, Acetate and Formate Lines.** *In silico* yield line for glucose as carbon source is shown. Maximal oxygen uptake rate was set to 7 mmol/gDW h. Yield line is divided into three phases, denoted by I, II and III. Each phase is characterized by a single biomass yield and by-product secretion pattern, as shown in the adjacent table.

oxygen to substrate ratios. This is because the substrate uptake rate increases from phase I to phase III, while the maximal oxygen uptake rate, allowed to the cell to use, is kept constant. Since the substrate uptake rate in phase I is comparatively low, the cell has enough oxygen to allow for a complete oxidation of the substrate to CO<sub>2</sub> via tricarboxylic acid (TCA) cycle and electron transport chain. As the substrate uptake rate keeps increasing, while the capacity of the electron transport chain is kept constant (since O<sub>2</sub> availability is kept constant), the cell cannot completely redirect the formed pyruvate to the TCA cycle and part of the pyruvate is fermented to form acetate in the second phase and acetate and formate in the third phase. Thus, we extend our analysis to the *in silico* acetate and formate secretion patterns.

Figure 4.10 directly compares *in silico* calculations with experimental data. Upper row correspond to *in silico* predictions, while lower row corresponds to experimental data. *In silico* calculations were performed using the *E. coli* core metabolic reconstruction along with the COBRA toolbox. Glucose uptake rate was varied from 0 to 14 mmol/gDW h, while just two oxygen levels were considered: 4 and 7 mmol/gDW h, corresponding to the black and green line of the upper row of Figure 4.10, respectively. Growth rate was used as the objective function and was maximized using linear programming. For a given pair of substrate uptake rate ( $r_{substrate}$ ) and growth rate ( $\mu$ ), FVA analysis was used to estimate acetate secretion ( $r_{acetate}$ ) by minimizing and maximizing flux through the acetate export reaction in the *E. coli* core model, while keeping  $r_{substrate}$  and  $\mu$  fixed. Formate secretion was estimated in a similar way, by maximizing and minimizing the flux through the formate export reaction, while keeping  $r_{substrate}$ ,  $\mu$  and minimal  $r_{acetate}$  constant.

As previously inferred, transitions between different phases of the yield lines are accompanied by changes in the by-product secretion pattern. Consider for instance the black *in silico* yield line of Figure 4.10. As the growth rate increases, cellular metabolism undergoes a transition, from phase I to phase II. This is accompanied by a reduction in the biomass yield and acetate secretion for the first time (4.10.B, point ①). The growth rate in

### 4.3. Yield, Acetate and Formate Lines

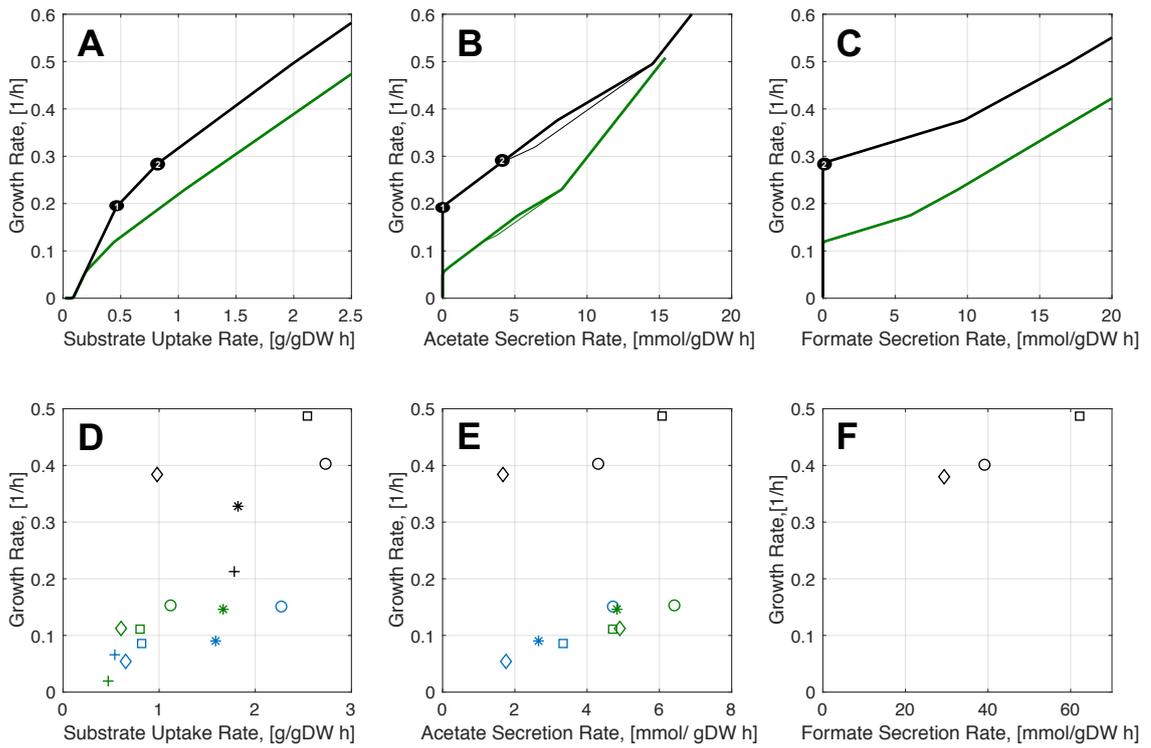


Figure 4.10.: **Qualitative Comparison Between Experimental Data and *in silico* Calculations.** **A.** to **C.** represent *in silico* calculations. Upper boundary for oxygen uptake rate was constrained to 7 mmol/gDW h for the black lines and to 4 mmol/gDW h for the green lines. ① and ② corresponds to transitions from phase I to II and from phase II to III, respectively (see Figure 4.9). Thin lines correspond to maximal acetate secretion rates, while thick lines correspond to minimal acetate secretion rates. **D.** to **F.** represent experimental data. In line with a previous publication [11], we refer to the lines shown in **B.** and **E.** as *acetate lines*. Analogously, we introduce the concept of *formate line* to refer to the line shown in **F.** Legend: black symbols correspond to strain 31, green symbols to strain 3150 and blue symbols to strain 3152. Substrate symbols are as follow: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose.

#### 4. Metabolic Burden

which this transition occurs is termed critical growth rate for acetate,  $\mu_{acetate}$ , and corresponds in this case to 0.19 1/h. As the growth rate continues to increase, acetate secretion keeps linearly increasing. At a certain point (②), cellular metabolism undergoes a second transition, accompanied again, by a reduction in the biomass yield and a change in the by-product secretion pattern. In this case, formate is secreted for the first time (4.10.C, ②). Analogously, we refer to the growth rate in which the second transition occurs as the critical growth rate for formate,  $\mu_{formate}$ , which in this case corresponds to 0.28 1/h. Interestingly, the transition from phase II to phase III is not only characterized by a reduction in the biomass yield, but also by differing minimal and maximal acetate secretion values, as shown in Figure 4.10.C by the thin line. A similar behavior during phase transitions is observed for the green *in silico* yield line.

Initially, similarities between *experimental* and *in silico* yield lines at different oxygen levels led to the hypothesis formulation of load-induced oxygen limitation (Figure 4.8) and motivated the previous analysis of secretion patterns *in silico*. Figure 4.10.E und .F shows experimental acetate and formate secretion for all strains as a function of growth rate. A comparison between experimental and *in silico* secretion patterns reveals further, interesting resemblances. First, we observe a parallel, downwards displacement of the acetate line for strains 3150 and 3152, when compared to the parent strain 31 (Figure 4.10.E). Remarkably, the same behavior can be obtained *in silico* by reducing the maximal oxygen uptake rate, as shown in Figure 4.10.B for low acetate secretion rates. Second, the existence of formate lines is predicted by our *in silico* analysis, as shown in Figure 4.10.C. Experimentally, this can be qualitatively confirmed for strain 31, as shown in Figure 4.10.F. Missing formate lines for strains 3150 and 3152 could be explained by low growth rates exhibited by these strains. As a consequence, experimentally observed metabolic states might be limited to phase I and II. Since actual values for oxygen uptake rate and level for each of the three strains are unknown, mentioned similarities have only a qualitative nature. However, we believe that the proposed hypothesis of load-induced oxygen limitation is supported by our experimental observations and partly by experimental evidence presented by Peebo et al. [110], as discussed in Section 4.5.

### 4.4. Process-level Strategy to Reduce Detrimental Effects of Metabolic Burden

One of the objectives motivating the analysis of the metabolic burden generated by heterologous protein over-expression in *E. coli* was the development of strategies to reduce its negative effect on biomass yield and acetate secretion. Here, we propose a process-level strategy designed to recover the wild-type biomass yield by adjusting the growth rate during heterologous protein production. By using different substrates, it was possible to study gene over-expression at various growth rates in shaking flasks. Assuming that standardization of substrate uptake via molecular weight is reasonably accurate, it is possible to draw straight lines representing experimental yield and acetate lines for all strains as, shown in Figure 4.11.

Our *in silico* analysis suggests that yield lines of a given reference strain and a loaded strain should converge into one single yield line at a sufficiently low growth rate (see 4.10.A). In our specific case, yield lines of the two loaded strains (3150 and 3152) seem to converge with

#### 4.4. Process-level Strategy to Reduce Detrimental Effects of Metabolic Burden

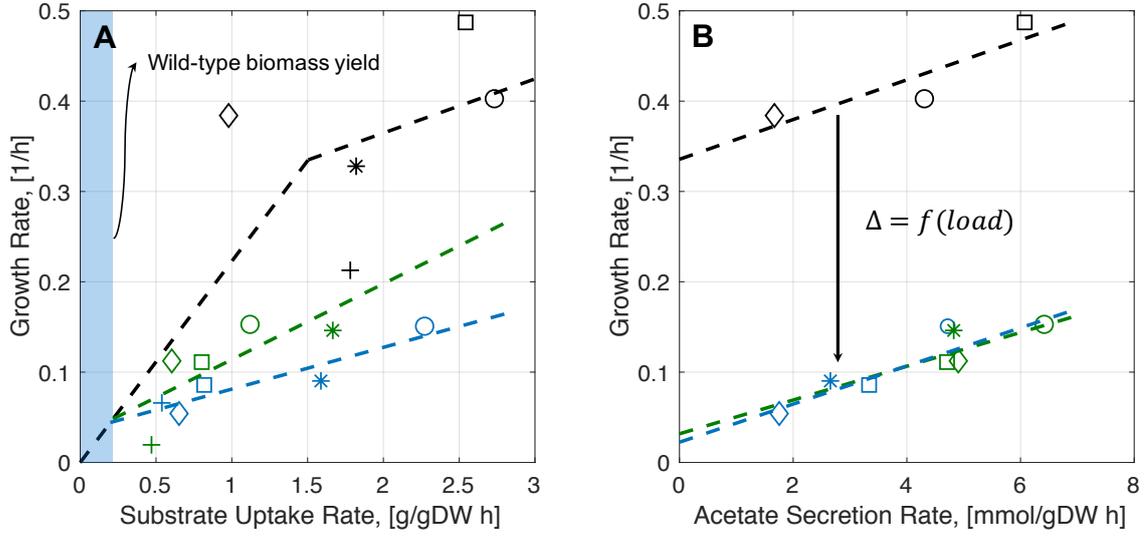


Figure 4.11.: **Process-level Strategy to Reduce Metabolic Burden.** **A.** Experimental yield lines. **B.** Experimental acetate lines. Dashed lines were obtained using linear regression. Black symbols correspond to strain 31, green symbols to strain 3150 and blue symbols to strain 3152. The metabolic burden ( $\Delta$ ), defined as a reduction in  $\mu_{Acetate}$ , is a function of the load exerted on the strain. We predict that by modulating the load in strain 3152, burden-free eGFP expression (blue region) can occur at higher growth rates, when compared to current load level. Substrate symbols are as follows: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose. Critical growth rates for acetate secretion are:  $\mu_{Acetate}^{31} = 0.33$ ,  $\mu_{Acetate}^{3150} = 0.03$ ,  $\mu_{Acetate}^{3152} = 0.022$  1/h.

the reference strain (31) at the same point into one single yield line, as shown in Figure 4.11.A. The growth rate at which yield lines converge can also be obtained from Figure 4.11.B and corresponds to  $\mu_{Acetate}$  of loaded strains. In theory,  $\mu_{Acetate}$  for a given strain defines the transition from phase I to phase II, as discussed in section 4.3. Experimentally determined  $\mu_{Acetate}$  for strains 3150 and 3152 (y-intercept of acetate lines shown in Figure 4.11.B) are in accordance with their respective transition phases, obtained from Figure 4.11.A. This observation provides further support for our load-induced oxygen limitation hypothesis. Additionally, *in silico* predictions generated by this analysis can be used to guide the design of a fed-batch process by providing a threshold for cellular growth rate that should not be exceeded during the production phase, once the plasmid-based system has been induced. In this way, cells producing heterologous proteins will not secrete acetate and their biomass yield will be consequentially identical to that of the wild-type or the reference strain. The range of growth rates allowing protein over-expression with no burden is highlighted in blue in Figure 4.11.A. For the concrete case of the eGFP over-production in strain 3152, a critical growth rate for acetate of 0.022 1/h might cause this burden-free zone be too small to be technically and economically feasible. A potential solution to this problem would be to modulate the load imposed by the eGFP expression by using low-copy plasmids [58] and by tuning the strength of promoter and ribosome binding site. As a result,  $\mu_{Acetate}^{3152}$  should increase, allowing a burden-free eGFP expression at much higher growth rates than the current system.

### 4.5. Discussion

The concept of acetate lines and its relationship with metabolic burden is not entirely new. Basan et al. [11] experimentally showed that increasing LacZ overexpression levels caused parallel shifts of the acetate line, thus leading to a reduction of corresponding values of  $\mu_{Acetate}$  in *E. coli* strains. The authors explained this behavior using a proteomic resource allocation model. The model assumes a limited proteomic capacity, which the authors divided into three fractions. The first fraction is dedicated to biomass synthesis and linearly depends on the growth rate, while the other two fractions are used for energy biosynthesis via respiration and fermentation, respectively. The authors rationalized, under the assumption of growth rate maximization, that at high carbon uptake rates, it is more advantageous for the cell to generate energy by the more proteome-efficient fermentation pathway. In this way, a higher fraction of the proteome can be used to support growth. By contrast, at low carbon uptake rates, the cell should use the more carbon-efficient respiration pathway, so that more carbon flux can be used for biomass synthesis. Since heterologous protein over-expression reduces the effective proteome fractions available for biomass synthesis and energy production, loaded cells should exhibit higher acetate secretion rates at fixed growth rates, which was indeed experimentally observed by the authors (refer to Figure 3.A in reference [11]). Here, we used a constraint-based approach to qualitatively account for the same phenomenon of overflow metabolism. Our modelling strategy is similar to the approach used by Varma et. al [151] to understand glucose catabolism in *E. coli* under various oxygenation rates. By constraining the upper boundary for oxygen uptake rate to a given value, independently of the numerical value of the substrate uptake rate, our constraint-based approach reproduces experimental observations of acetate secretion at high substrate uptake rates. The observed increase in acetate secretion, exhibited by protein-overproducing strains, is reproduced in our model by reducing the numerical value of the oxygen uptake rate constraint, as shown in Figure 4.12. In this figure, orange lines represent the growth behavior of a wild-type strain, while all other lines qualitatively represent strains exhibiting an increasing metabolic burden.

Note that constraining the upper boundary for oxygen uptake directly limits the maximal flux through the respiratory chain, as represented by a saturation-like behavior of the cytochrome bd oxidase activity (Figure 4.12.C). Interestingly, a similar saturation behavior has been experimentally observed for the enzyme ATP synthase, when its expression cost is plotted as a function of the growth rate (refer to Figure 6.A in reference [110]). The same study reported acetate secretion when the ATP synthase reached its maximal expression cost. Both observations provide experimental evidence supporting the key assumption of our constraint-based approach, i.e., that a constraint in the maximal flux through the respiratory chain, implemented in our model by constraining the maximal oxygen uptake rate and experimentally observed by Peebo et al. [110] as a limited ATP synthase expression, gives rise to acetate lines. The work by Peebo also provides indirect experimental evidence to our load-induced oxygen limitation hypothesis. The authors reported that cells grown in amino acids (AA) supplemented medium can support higher ATP synthase expression costs, compared to cells grown in medium without amino acids supplementation. Consequently,  $\mu_{Acetate}$  in the AA supplemented medium increases from 0.35 (no AA supplementation) to 0.55 1/h (refer to Figure 1.A in reference [110]) Applying the proteomic resource allocation theory, used by Basan et al. [11] to explain overflow metabolism in *E. coli*, one can rationalize that a higher ATP synthase expression is possible

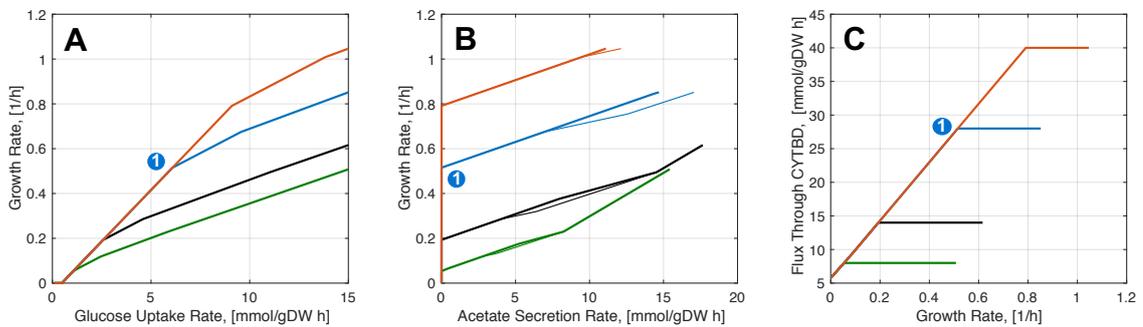


Figure 4.12.: **Emergence of Yield and Acetate Lines By Constraining Flux Through Respiratory Chain.** The effect of constraining the upper boundary for oxygen uptake rate on biomass yield (A.); acetate lines -thin and thick lines correspond to maximal and minimal acetate secretion, respectively- (B.) and the flux through respiratory chain (C.) is shown. Simulations were performed using the *E. coli* core model and a varying upper boundary for the oxygen uptake rate, as follows: green lines: 4 mmol/gDW h; black lines: 7 mmol/gDW h; blue lines: 14 mmol/gDW h; orange lines: 20 mmol/gDW h. **CYTBD** refers to the enzyme cytochrome bd oxidase, which produces a proton motive force in the *E. coli* core model by reducing oxygen to water as follows:  $2h[c] + \frac{1}{2}o_2[c] + q_8h_2[c] \rightarrow 2h[e] + h_2o[c] + q_8[c]$ .  $q_8h_2$  and  $q_8$  represent reduced and oxidized ubiquinone pools, respectively. Note that constraining the upper boundary of the oxygen uptake reaction directly limits the maximal flux through the enzyme CYTBD. Interestingly, the growth rate at which the maximal CYTBD activity is reached correspond to the transition from phase I to phase II. This point is exemplarily marked for the blue lines ①. Since the maximal respiration capacity is exhausted, the cell starts using fermentation pathways to generate energy. As a consequence, acetate is secreted and the biomass yield is reduced.

#### 4. Metabolic Burden

because amino acid supplementation increases proteomic resources by reducing the fraction of the proteome that is dedicated to amino acid biosynthesis. Conversely, heterologous protein overexpression would decrease proteomic resources available for biomass synthesis and energy generation, thus leading to a relative decrease in the maximal ATP synthase expression capacity and a consequent reduction in the flux through the respiratory chain.

Note that our simple constraint-based approach provides a flux-based explanation of the overflow metabolism in *E. coli* and allows for the generation of multiple testable hypothesis:

1. *Acetate lines are not always composed of a single straight line.* Basan et al. [11] reported that plotting growth rates against acetate secretion rates leads to a straight line, which the authors termed acetate line. Here, we showed that this is only the case, if acetate is the only fermentation product, i.e., cellular metabolism is confined to phase II (refer to Figure 4.9 and 4.10). Experimental evidence reported by Peebo et al. [110] (refer to Figure 1.A therein) supports this hypothesis.
2. *Protein over-expression originates a load-induced oxygen limitation,* as discussed in Section 4.3.
3. *There exist a relationship between biomass yield, acetate and formate lines,* as discussed in Section 4.3 and evidenced in Figure 4.10.A to .C.
4. *Burden-free protein over-expression should possible for growth rates lower than  $\mu_{Acetate}$ .* Consequently, a process-level strategy, as discussed in Section 4.11, can be developed to minimize or eliminate the metabolic burden, expressed as a decrease in biomass yield along with an increase in acetate secretion rate.

Further experimental studies can be designed to test the above mentioned hypothesis. Since oxygen uptake rate plays a central role in both defining the cellular metabolic state (phase I, II or III) and estimating conversion factors used to standardize substrate uptake rates of different carbon sources when energy content is used a quality criterion, its experimental determination should be prioritized. Similarly, experimental determination of protein content in loaded strains should be also considered in order to accurately describe the effect of protein over-expression on oxygen uptake capacity and proteome allocation. Finally, it is recommended that modeling efforts are only conducted on high-quality, reproducible data. Only under this circumstances the design-build-test-learn cycle can be optimally executed.

## 5. Strain Design Algorithms for Target Identification

The EM approach developed by the James Liao group at the UC, Los Angeles was used as framework to develop a strain design algorithm that we refer to as optEM. This algorithm uses EM to build an ensemble of kinetic models that fulfill a given reference flux distribution. optEM facilitates the maximization of complex user-tailored objective functions, for instance the weighted sum of volumetric productivity (mmol/l h), product yield (mmol/mmol) and production rate (mmol/gDW h). The kinetic nature of the optEM algorithm allows for consideration of typically inaccessible design parameters such as enzyme over-expression level, overall system stability and metabolite concentration pattern after enzyme perturbation. Using the optEM approach, we identified many genetic targets that were already reported in the literature to have beneficial effects on taxadiene yield, thus confirming the prediction power of our approach. We also identified targets not reported in the literature yet [37].

### 5.1. Strains & Experimental Data

Five production strains were constructed and experimentally characterized by our cooperation partner Monica Fuchs at **IBK**. Table 5.1 describes main genetic features of each of the strains, which are schematically represented in Figure 5.1 for a better overview. All strains are provided with a genetic makeup necessary to synthesize taxadiene. In strains  $\Delta_{xylE\_1p}$  and  $\Delta_{ptsG\_1p}$ , genes contained in plasmid pET28a\_DS (*ispF*, *ispD*, *idi* and *dxs*) were genomically integrated in the *xylE* and *ptsG* position, respectively. In contrast, all enzymes required for the taxadiene biosynthesis are encoded in two plasmids in strains  $\Delta_{xylE\_2p}$  and  $\Delta_{ptsG\_2p}$ . Refer to Table 5.1 for plasmid description. Gene integration aimed at the reduction of the metabolic burden caused by heterologous enzyme expression. Strain cultivation was performed in shaking flask (220 rpm) at 37°C using Riesenberg medium with a mixture of  $^{12}\text{C}$  (4.5 g/l) and  $^{13}\text{C}$  (0.5 g/l) glucose as carbon source. Experimental strain characterization was performed in duplicate and consisted of measurement of concentration time courses for biomass, glucose and acetate. These data were processed as described in Chapter 2.1.2 in order to calculate growth, substrate uptake and acetate secretion rates, respectively. Exchange rates and their standard deviations are reported in Table 5.2. Growth rates exhibit lower variability compared to acetate secretion rates (see Figure 5.2). This is because growth rates were calculated based on OD time courses, whose measurements require a less extent of experimental handling compared to the determination of glucose or acetate concentration time courses via HPLC measurements. Taxadiene could not be detected in the culture medium for any strain, due to low production rates exhibited by all strains combined with low sample volumes. Additional to concentration time courses,  $^{13}\text{C}$ -labeling data were also available for all strains. These measurements

## 5. Strain Design Algorithms for Target Identification

Table 5.1.: **Taxadiene Producing Strains.** Five different strains were constructed and experimentally characterized in biological duplicates. All strains derive from *E. coli* HMS174(DE3). Refer to table notes for a detailed plasmid description. Strains were grown in minimal medium containing 0.5 g/l [U-<sup>13</sup>C] glucose and 4.5 g/l <sup>12</sup>C glucose. Refer to Figures A.5 and A.7 in the Appendix for an overview of the non-mevalonate & the taxadiene production pathway, respectively.

Strain ID	Genotype	Description
HMS_2p	HMS174(DE3) + pET28a_DS + pACY_MEP	HMS174(DE3) strain transformed with two plasmids.
$\Delta$ xylE_2p	HMS174(DE3) $\Delta$ xylE + pET28a_DS + pACY_MEP	Gene <i>xylE</i> is knocked out from strain HMS_2p
$\Delta$ xylE_1p	HMS174(DE3) $\Delta$ xylE + pET28a_DS	Heterologous genes encoded in plasmid pACY_MEP are genomically integrated in the <i>xylE</i> position the control of a Trc promoter.
$\Delta$ ptsG_2p	HMS174(DE3) $\Delta$ ptsG + pET28a_DS + pACY_MEP	Gene <i>ptsG</i> is knocked out from strain HMS_2p.
$\Delta$ ptsG_1p	HMS174(DE3) $\Delta$ ptsG + pET28a_DS	Heterologous genes encoded in plasmid pACY_MEP are genomically integrated in the <i>ptsG</i> position under the control of a Trc promoter.

pET28a\_DS is a high copy number plasmid that contains genes encoding following enzymes: kanamycin resistance, taxadiene synthase (*txs*) and geranylgeranyl diphosphate synthase (*crtE*). Expression of all genes is under the control of a T7 promoter.

pACY\_MEP is a low copy plasmid that contains genes encoding a chloramphenicol resistance and selected enzymes of the non-mevalonate pathway: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (*ispF*), 4-diphosphocytidyl-2C-methyl-D-erythritol synthase (*ispD*), isopentenyl-diphosphate isomerase (*idi*) and 1-deoxy-D-xylulose-5-phosphate synthase (*dxs*). Gene expression is under the control of a T7 promoter.

were conducted by our cooperation partner Claudia Huber at **Bio**. Refer to Appendix A.2.3 for raw data.

## 5.2. Reduction of Metabolic Burden Through Genomic Integration

As mentioned before, the main objective of genomic integration conducted in strains  $\Delta$ ptsG\_1p and  $\Delta$ xylE\_1p was a reduction of the detrimental effect that plasmid-based heterologous gene expression exerts on the host metabolism, experimentally evidenced as a reduction in growth rate and yield. The success of this strategy can be assessed using the previously introduced concept of acetate lines (refer to Chapter 4). An acetate line emerges when plotting growth rates against corresponding acetate secretion rates for a given strain. By comparing the relative position of acetate lines between strains, one can assess the effect that a certain perturbation (in genotype or culture conditions) has on cellular fitness. A relative upwards displacement of the acetate line indicates that the given perturbation increases cellular fitness, thus leading to a higher  $\mu_{Acetate}$ . Since each of the strains listed in Table 5.1 was experimentally characterized for one single growth rate, additional information is needed to plot corresponding acetate lines. An interesting feature

## 5.2. Reduction of Metabolic Burden Through Genomic Integration

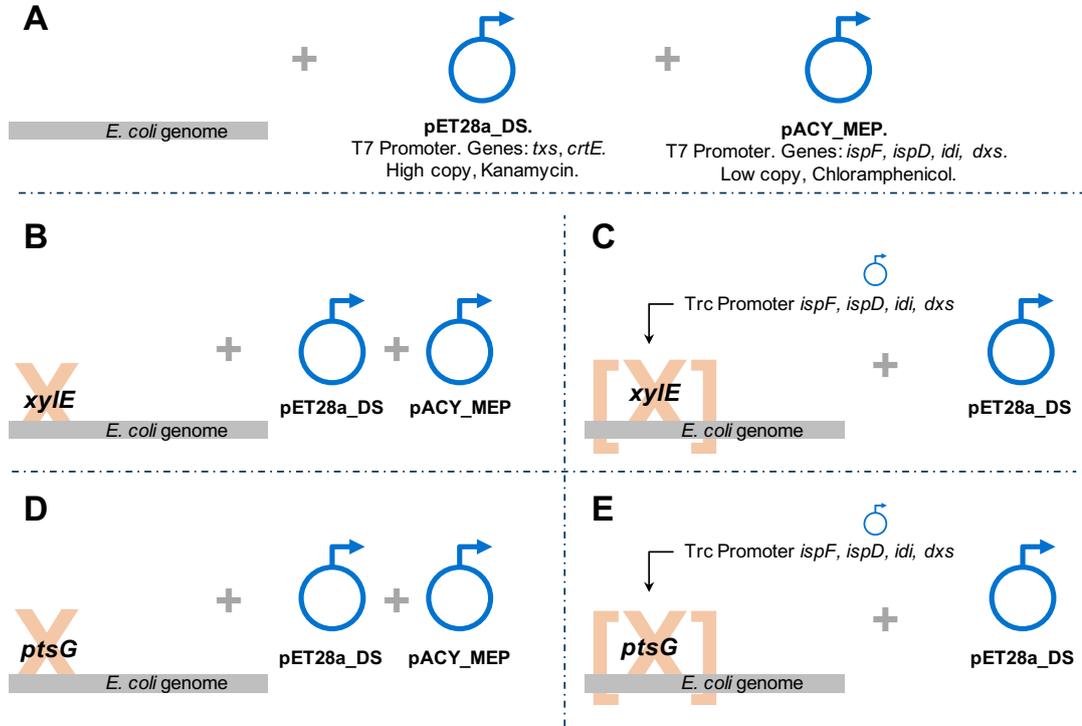


Figure 5.1.: **Taxadiene Producing Strains.** Each of the five taxadiene producing strains described in Table 5.1 is schematically represented in this figure as follows: **A.** HMS\_2p; **B.**  $\Delta xylE$ \_2p; **C.**  $\Delta xylE$ \_1p; **D.**  $\Delta ptsG$ \_2p; **E.**  $\Delta ptsG$ \_1p. Note that strains represented in Figures **C.** and **E.** contain one single plasmid. In both cases, genes encoded in the plasmid pACYC\_MEP were genetically integrated under the control of a Trc promoter in the *xylE* and *dptsG* positions, respectively. The effect that the genomic integration of one plasmid has on the metabolic burden can be nicely observed in Figure 5.2.

Table 5.2.: **Taxadiene Producing Strains: Exchange Rates.** Growth, substrate uptake and acetate secretion rates were calculated from concentration time courses as described in Section 2.1.2. Reported rates correspond to the average of two biological replicates. Errors correspond to one standard deviation.

Strain	Growth Rate	Glucose Rate	Uptake	Acetate Rate	Secretion
HMS_2p	$0.128 \pm 2.20 \times 10^{-3}$	$4.532 \pm 4.242$		$3.013 \pm 6.51 \times 10^{-1}$	
$\Delta xylE$ _2p	$0.131 \pm 4.20 \times 10^{-3}$	$3.923 \pm 5.27 \times 10^{-1}$		$1.197 \pm 1.99 \times 10^{-1}$	
$\Delta xylE$ _1p	$0.185 \pm 5.90 \times 10^{-3}$	$8.527 \pm 3.12 \times 10^{-1}$		$1.024 \pm 4.60 \times 10^{-2}$	
$\Delta ptsG$ _2p	$0.076 \pm 2.76 \times 10^{-2}$	$5.686 \pm 1.412$		$1.659 \pm 4.45 \times 10^{-1}$	
$\Delta ptsG$ _1p	$0.189 \pm 1.00 \times 10^{-4}$	$7.166 \pm 5.49 \times 10^{-1}$		$0.387 \pm 1.03 \times 10^{-1}$	

Growth rates are reported in 1/h; glucose and acetate rates are given in mmol/gDW h.

## 5. Strain Design Algorithms for Target Identification

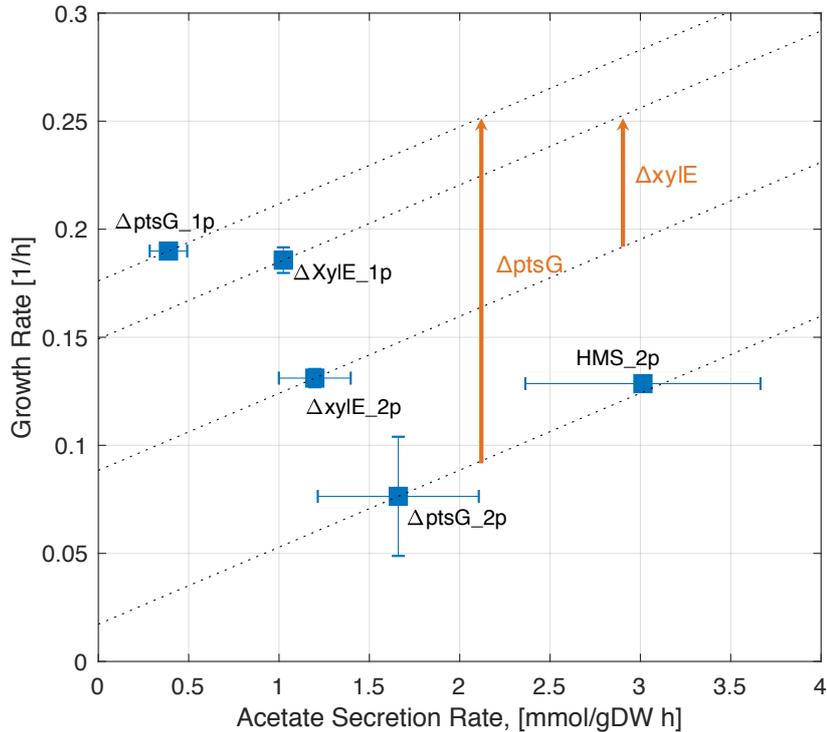


Figure 5.2.: **Acetate Lines for Taxadiene Producing Strains.** Straight dotted lines correspond to acetate lines for each strain. Each line was defined by experimental information contained in Table 5.2 and by a constant slope, obtained as described in Appendix A.1.1. The effect of genetic perturbations (genomic integration, gene knock-outs) on strain fitness can be characterized comparing the relative position of corresponding acetate lines. For instance, orange arrows represent the positive effect that genomic integration of plasmid-encoded genes has on cellular metabolism for both  $\Delta ptsG$  and  $\Delta xylE$  strain families. In this specific case, the extent of burden reduction seems to be a function of the genetic background.

of acetate lines is that under certain conditions, their slope remains constant. We make use of this property to generate acetate lines for all strains, as shown in Figure 5.2. Note that although all five strains contain the necessary genetic makeup to produce taxadiene, the form in which this genetic information is introduced into the cell, i.e., plasmid-encoded or genomically integrated, has an effect on the relative position of each acetate line, and thus, on the relative metabolic burden. For instance, genomic integration of genes encoded in plasmid pACYC\_MEP seems to relieve the metabolic burden by an extent that depends on the host genetic background, as highlighted in Figure 5.2 by the two orange arrows. The effect of gene knock-outs on metabolic burden can also be studied by comparing the relative position of acetate lines for strains HMS\_2p,  $\Delta xylE_2p$  and  $\Delta ptsG_2p$ , as shown in Figure 5.2. While knocking out gene *xylE* in strain HMS\_2p leads to a reduction of metabolic burden, the effect of *ptsG* knockout is not completely clear due to large standard deviations.

### 5.3. Constraint-based Assessment of Taxadiene Production Potential

The main objective of constructing strains listed in Table 5.1 was to generate a pool of production strain candidates. Experimental data generated during strain characterization, including  $^{13}\text{C}$ -labeling data and concentration time courses for biomass, glucose and acetate should now be used to identify strains exhibiting a superior production performance for further engineering rounds. Since taxadiene concentration could not be experimentally measured due to its presumably low concentration in the medium and to low sample volume, our initial modeling efforts were focused on the estimation of production capabilities of each strain. We start our analysis by calculating flux ranges for each reaction in the *E. coli* central metabolism best matching experimentally measured amino acids labeling patterns. For that, we used the software 13CFLUX2 [155] along with a modified version of the *E. coli* metabolic model iRL2013 [75]. Refer to Section 2.1.2 for details on the metabolic model and mathematical implementation. The procedure can be summarized as follows:

for each strain in Table 5.1:

1. Generate input file for the software 13CFLUX2 containing experimentally determined exchange rates, as summarized in Table 5.2 and labeling data (refer to Table A.8 in the Appendix). An exemplary input file can be found in Appendix A.2.4.
2. Calculate intracellular flux distribution that minimizes squared sum of residuals for labeling data.
3. Apply statistical workflow described in Section 2.1.2 in order to calculate flux ranges for free reactions.
4. Calculate flux ranges for each reaction in the metabolic network using flux variability analysis (FVA) considering experimentally determined exchange rates and flux ranges for free reactions calculated in the previous step.
5. Calculate maximal theoretical taxadiene yield as the ratio between taxadiene production rate and glucose uptake rate.

Calculated maximal theoretical taxadiene yields using the previously described steps are plotted in Figure 5.3 for all strains. At this point, it is important to state that product yields shown in Figure 5.3 are not experimental values, but rather a theoretical estimation representing the *potential* of each strain to produce taxadiene with a certain product yield. One can clearly observe that a reduction in the metabolic burden in strains  $\Delta\text{xylE\_1P}$  and  $\Delta\text{ptsG\_1P}$ , achieved by genomic integration of genes encoded in plasmid pACYC\_MEP, leads to a marked increase in the taxadiene production potential.

### 5.4. Application of Strain-design Algorithms for Target Identification

One valuable application of metabolic modeling is the possibility to *in silico* analyze the effect of a large number of perturbations on cellular metabolism in a fast, efficient and

## 5. Strain Design Algorithms for Target Identification

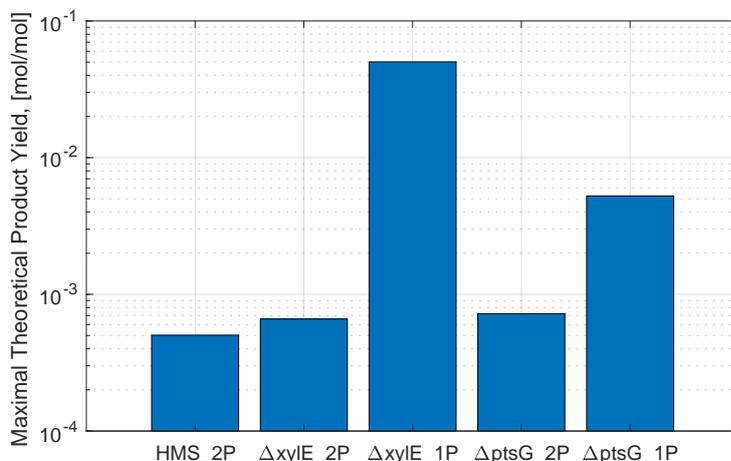


Figure 5.3.: **Constraint-based Assessment of Production Capabilities.** FVA was applied on an extended version of the iRL2013 [75] to estimate the taxadiene production potential of each strain. Shown are maximal theoretical yields, obtained by dividing maximal theoretical flux through the taxadiene pathway by the glucose uptake rate. Refer to Table A.9 in Appendix A.2.5 for an overview of minimal and mean production rates and yields.

cost-effective fashion. Such perturbations can be exerted at the process level (refer to Part III of this work), by modifying medium composition, production temperature, time point of culture induction, dilution rate, etc., or at the genetic level, by modulating the expression of various genes. Product yield, titer and productivity are process parameters commonly chosen for optimization during *in silico* studies. By using *in silico* generated model-based predictions to guide the development of *experimental* strategies for strain and process optimization, time and costs invested in experimental efforts can be greatly minimized.

There are various mathematical approaches used to describe bacterial metabolism, some of which were already presented in Section 2. Here, we apply both a traditionally used constraint-based strain-design algorithm called optKnock [20] and a novel kinetic-based algorithm, we termed OptEM, to identify genetic modifications aiming at the overproduction of taxadiene in *E. coli*.

### 5.4.1. Constraint-based Methods: The OptKnock Algorithm

OptKnock [20] was the first constraint-based strain-design algorithm that used growth-coupling as design principle to systematically identify reaction deletions leading to the overproduction of a target molecule. The basic idea behind growth-coupling is to delete a defined set of reactions, so that the synthesis of a given target molecule become obligatory for the cell to be able to grow, as exemplarily shown in Figure 5.4.B. By deleting three reactions (genes), the cell is forced to change its metabolism to produce taxadiene in order to be able to grow. This is denoted by the arrow in Figure 5.4.B.

As recently reported by Kamp and Klant [152], growth-coupling strategies can be identified for almost all metabolites in genome-scale metabolic models of five major production

#### 5.4. Application of Strain-design Algorithms for Target Identification

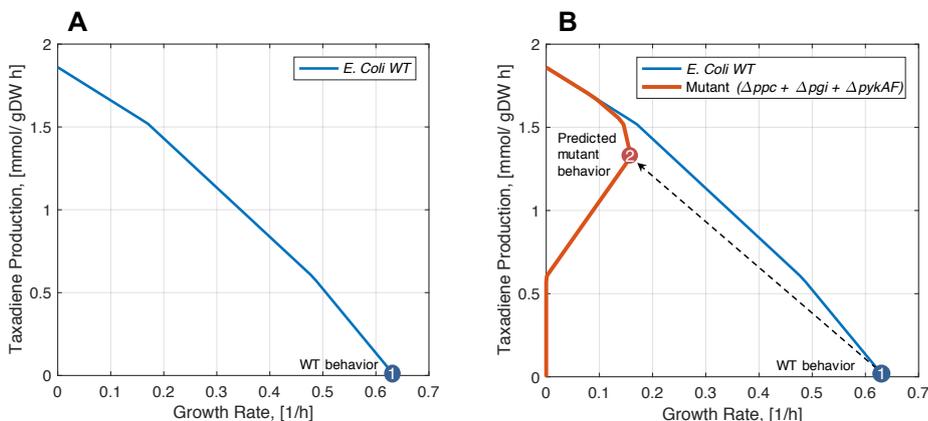


Figure 5.4.: **Application of OptKnock to the Production of Taxadiene in *E. coli*: Basic Idea.** **A.** Maximal theoretical taxadiene yield is calculated and plotted as a function of growth rate. A modified version of the *E. coli* core metabolic model was used for all calculations (refer to Appendix B.1.1). Glucose and oxygen uptake rate were kept at a constant value of 10 mmol/gDW h for all simulations. OptKnock was applied to identify reaction deletion targets shown in **B**. After deleting selected reactions from the network, the cell is forced to produce taxadiene along with biomass in order to maximize its growth rate. Points 1 and 2 correspond to predicted cellular behaviors maximizing growth in the wild-type (WT) and mutant, respectively. Genes *ppc*, *pgi* and *pykAF* encode for the enzymes phosphoenolpyruvate carboxylase, glucose-6-phosphate isomerase and pyruvate kinase, respectively.

organisms: *E. coli*, *S. cerevisiae*, *Corynebacterium glutamicum*, *Aspergillus niger* and *Syn-echocystis* sp. PCC 6803, which demonstrates the wide applicability of growth-coupling as design principle. In its original publication, optKnock was applied to identify different *E. coli* genotypes exhibiting an increased succinate production *in silico*.

Here, we apply OptKnock to identify triple reaction deletion designs leading to growth-coupled taxadiene overproduction. In a first step, an *E. coli* core metabolism model was slightly modified to allow for taxadiene synthesis, as explained in Appendix B.1.1. Then, deletion strategies were identified for various growth rates, ranging from 0 to 0.2 1/h, in accordance with experimental data provided by Monika Fuchs (refer to Table 5.2). Similarly, the upper bound for glucose uptake rate was updated to 10 mmol/gDW h. Since the oxygen uptake rate was not provided, two different rates were assumed in order to consider both oxygen-rich and -poor conditions. All calculations were performed using the Optknock script provided with COBRA-Toolbox [124]. Figure 5.5 shows the dependence of OptKnock predictions on cellular growth rate and oxygen conditions.

While the set of reactions identified by OptKnock seems to remain constant over the range of growth rates tested when oxygen-rich conditions are assumed (Figure 5.5.A), the picture looks different when oxygen-poor conditions are assumed, as shown in Figure 5.5.B. In this case, cellular growth rate seems to affect OptKnock predictions. Figure 5.6 summarizes deletion targets for each condition. A stronger effect of growth rate at poor oxygen conditions on the identity of the deletion targets is evidenced by the larger number of different reactions presented in Figure 5.6.B. Although these results are not unexpected, they suggest that neither growth rate nor oxygen supply alone suffice as scale-up criterion to guarantee good large-scale performance of a production strain developed in the laboratory

## 5. Strain Design Algorithms for Target Identification

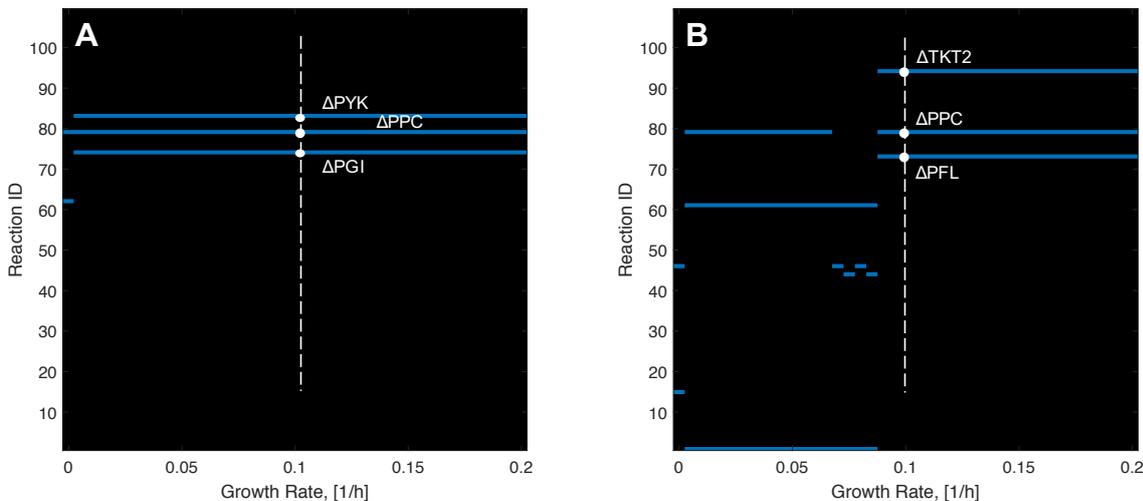


Figure 5.5.: **Application of OptKnock to the Production of Taxadiene in *E. coli*: Dependence on Growth Rate and Oxygen Availability.** Deletion candidates identified by OptKnock are shown in blue. **A.** Oxygen-rich conditions. Oxygen upper boundary was set to 10 mmol/gDW h for all calculations. For a growth rate of 0.1 1/h, the deletion targets identified by OptKnock correspond to pyruvate kinase (PYK), phosphoenolpyruvate carboxylase (PPC) and glucose-6-phosphate isomerase (PGI) **B.** Oxygen-poor conditions. Oxygen upper boundary was set to 1 mmol/gDW h. For a growth rate of 0.1 1/h, the deletion targets identified by OptKnock correspond to transketolase (TKT2), phosphoenolpyruvate carboxylase (PPC) and pyruvate formate lyase (PFL)

at a shaking flask scale. This point will be further analyzed in the discussion section.

Although predictions made by OptKnock have successfully guided a couple of experimental metabolic engineering efforts [42, 164], none of the best production strains was constructed by the sole deletion of genes corresponding to reaction targets identified by OptKnock. Complementary strategies, like adaptive evolution [42] or codon optimization along with optimization of gene expression balance [164], were necessary to achieve the highest production performance. In a recent review [149], our group identified a similar trend for the overproduction of succinate. We found that while a total of 26 theoretical studies, published between 2002 and 2016, reported different *in silico* strain design strategies for the overproduction of succinate, only 10 studies experimentally implemented any kind of the previously generated theoretical knowledge. Similar to the case of OptKnock, none of these publications reported an exact application of the corresponding computational predictions used to guide the strain design process.

Construction, characterization and adaptive evolution of mutant strains is a costly and time consuming process. Aiming at the minimization of experimental efforts, *in silico* predictions of alternative modeling approaches should be first generated, analyzed and compared, before *in silico* identified production genotypes are *experimentally* implemented towards the construction of actual production strain(s). For that reason, we now turn our attention to the application of a kinetic-based approach to identify targets for taxadiene overproduction *in silico*.

#### 5.4. Application of Strain-design Algorithms for Target Identification

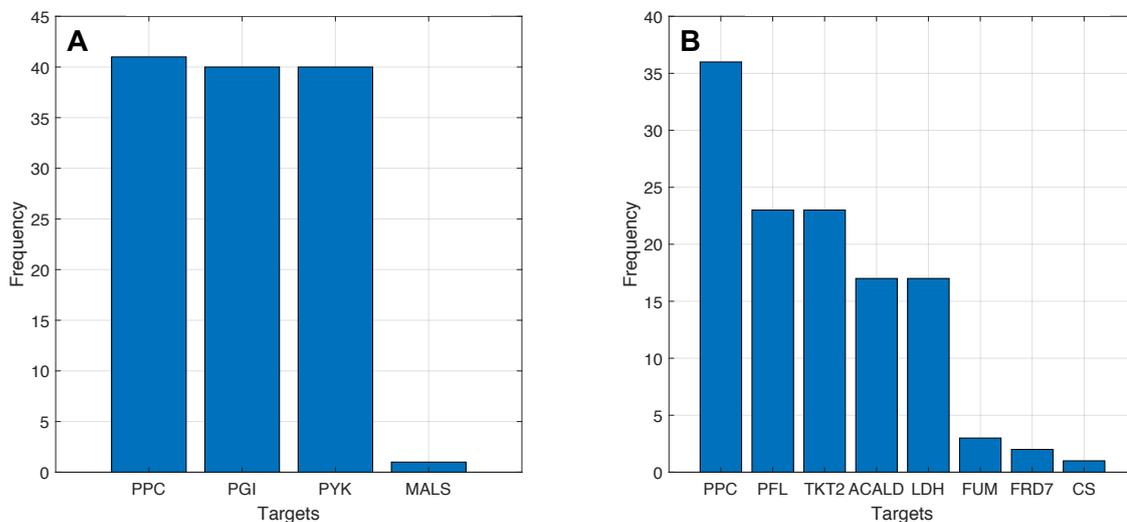


Figure 5.6.: **Application of OptKnock to the Production of Taxadiene in *E. coli*: Targets.** Deletion targets identified by OptKnock over the whole growth rate range are summarized for each oxygen condition. The larger the number of reactions in each histogram, the higher the effect of growth rate on deletion target identity **A.** Oxygen-rich conditions. **B.** Oxygen-poor conditions. Abbreviations: PPC, phosphoenolpyruvate carboxylase; PGI, glucose-6-phosphate isomerase; Pyk, pyruvate kinase; MALS, malate synthase; PFL, pyruvate formate lyase; TKT2, transketolase; ACALD, acetaldehyde dehydrogenase; LDH, L-lactate dehydrogenase; FUM, fumarase; FRD7, fumarate reductase; CS, citrate synthase.

#### 5.4.2. Kinetic-based Methods: The OptEM Algorithm

In many cases, the *modulation* of native and heterologous gene expression in the form of gene up- or down-regulation, rather than gene *knockouts*, is necessary to achieve optimal strain performance [1, 164]. Additionally, genetic perturbations can affect the dynamic stability of cellular metabolism, which in turn set limits on maximal productivity [144] and maximal product yield [71]. Since a kinetic description of cellular metabolism is a prerequisite to properly study these issues, constraint-based methods, like the previously discussed OptKnock, overlook the potential detrimental effect that identified genetic interventions could have on actual strain performance. As a consequence, further genetic interventions are often necessary to achieve high experimental strain performance [42, 149, 164].

Rate laws represent a central component of kinetic models of metabolism. They describe reaction fluxes as a function of metabolite and enzyme concentrations and allow the consideration of metabolite-level enzyme regulation. Since rate laws and their associated kinetic parameters are normally unknown for many enzymes, the application and development of kinetic-based modeling approaches for strain design has been rather moderate [28, 149].

Here, we lay the groundwork for OptEM, a novel kinetic-based strain design algorithm. OptEM uses the Ensemble Modeling formalism EM [146] to build an ensemble of kinetic models that fulfill a given reference flux distribution. OptEM facilitates the maximization of complex user-tailored objective functions, for instance the weighted sum of productivity, product yield and production rate. The kinetic nature of the optEM algorithm allows for consideration of typically inaccessible design parameters such as enzyme over-expression

## 5. Strain Design Algorithms for Target Identification

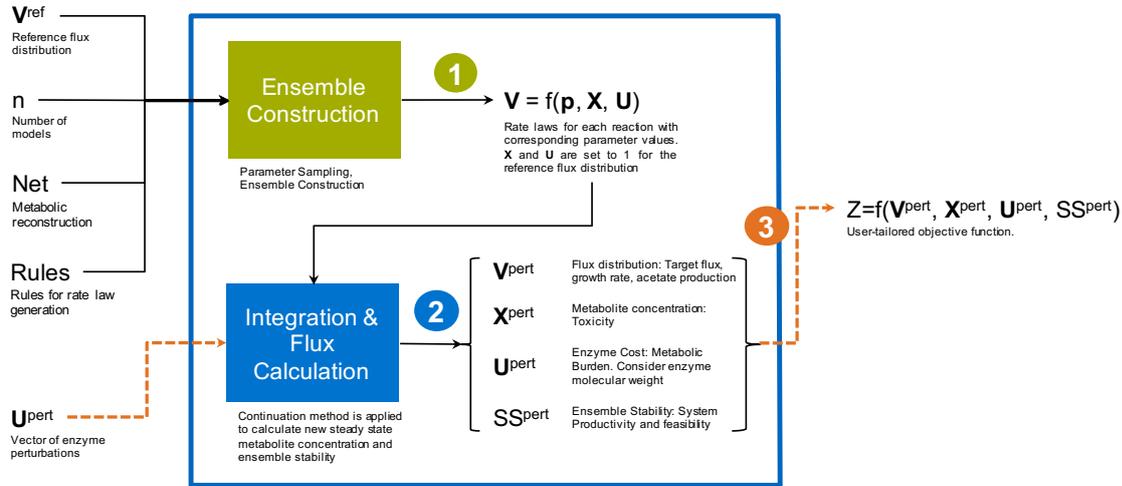


Figure 5.7.: **Overview of the Main Steps Comprising the OptEM Strain Design Algorithm** Ensemble Construction, Parameter Integration & Flux Calculation and Objective Function Evaluation are the three main steps comprising the OptEM algorithm. The effect of genetic perturbations, defined by  $U^{pert}$ , on strain performance can be assessed *in silico* by evaluating an user-defined objective function  $Z$ . This input-output relationship is shown by the orange arrows. Optimal strain design strategies can be identified applying a brute force or a solver approach. Both options are discussed in the main text.

level, overall system stability and metabolite concentration pattern after enzyme perturbation. The idea behind OptEM is described in Figure 5.7. For a given enzyme perturbation vector  $U^{pert}$ , the methodology calculates the corresponding value of an user-defined objective function. Since enzyme and metabolite concentrations are normalized to 1 for the given reference flux distribution, values higher than 1 refer to enzyme overexpression. Analogously, values lower than 1 represent enzyme downregulations. Depending on the size of the metabolic reconstruction used and the number of perturbations allowed, OptEM can extensively explore the design space by employing a brute force approach, as shown in Figure 5.8. If a brute force approach is impractical due to limited computational resources, a solver approach can be used, as explained in the next section.

The OptEM methodology can be summarized in three main steps, as indicated in Figure 5.7:

- 1. Ensemble Construction:** this step requires four inputs, namely a reference flux distribution ( $V^{ref}$ ) for the strain being optimized; an associated metabolic reconstruction consisting of a stoichiometric and a regulatory matrix ( $Net$ ); rate laws or "rules" to generate rate laws for each reaction in the metabolic reconstruction ( $Rules$ ) and the number of models to be generated ( $n$ ). The output of this step is an ensemble of  $n$  models, all fulfilling the given steady state flux distribution. Since metabolite and enzyme concentrations are normally unknown, they are normalized to 1 during the process of ensemble construction.
- 2. Parameter Integration & Flux Calculation:** steady state metabolite concentrations are integrated over an enzyme trajectory, defined by  $U^{pert} - U^{init}$  [74]. Resulting steady state metabolite concentrations can be transformed into fluxes using rate laws generated in the step 1. Note that  $U^{init}$  corresponds to a vector of all-ones

#### 5.4. Application of Strain-design Algorithms for Target Identification

of adequate size. Output of this step are the vectors  $\mathbf{V}^{pert}$ ,  $\mathbf{X}^{pert}$  and  $\mathbf{U}^{pert}$  describing flux distribution, metabolite concentration and enzyme concentration, all of them after conducting enzyme perturbation dictated by  $\mathbf{U}^{pert}$ , respectively. Variable  $SS^{pert}$  describes the percentage of models remaining stable after perturbation.

- 3. Objective Function Evaluation:** output of step 2 can be used to evaluate an user-defined objective function described by  $Z = f(\mathbf{V}^{pert}, \mathbf{X}^{pert}, \mathbf{U}^{pert}, SS^{pert})$ . By doing so, the specific effect of the genetic perturbation described by  $\mathbf{U}^{pert}$  on strain performance can be numerically assessed and compared with alternative genetic perturbations, thus leading to the identification of strain designs maximizing  $Z$ . Note that the evaluation of the objective function should be performed for each model in the ensemble. The effect of a given perturbation on the objective function  $Z$  can be then characterized by calculating the median and 25th and 75th percentiles.

### Objective Functions

The vast majority of Metabolic Engineering efforts to date have aimed at the optimization of product yield. However, it has been recognized that not only high product yields but also high volumetric productivity and product titer, strain robustness and the ability to use low-cost substrates are key factors defining process viability and should therefore be primary objectives for bio-process development [74, 86]. In line with this, OptEM was developed to allow the optimization of user-tailored objective functions, considering the most relevant process parameters. Table 5.3 lists various examples. While each of these objective functions targets product yield maximization, they also maximize or minimize secondary objectives like concentration of toxic pathway intermediates, strain robustness and protein burden.

### Brute Force Approach

In order to illustrate the probabilistic nature of OptEM predictions, the effect of single enzyme up- and down-regulation on product yield for the strain  $\Delta\text{XylE}_1\text{p}$  is exemplarily shown in Figure 5.8. This particular strain was selected as case study out of all production strain candidates (see Table 5.1) because of its superior taxadiene production potential, as identified by a constraint-based analysis -see Figure 5.3 and refer to section 5.3 for details on the estimation of production potential using  $^{13}\text{C}$ -labeling data-. The reference flux distribution used to construct the ensemble of kinetic models was calculated assuming a taxadiene yield of  $2.860 \times 10^{-5}$  mol taxadiene/mol glucose, which corresponds to a mean taxadiene production level, as shown in Table A.9 in Appendix A.2.5. This assumption is justified by the fact that, as mentioned earlier, no experimental taxadiene concentration was available for this set of strains.

As specified in Figure 5.8 by the arrows, the outcome of single up- or down-regulation of a given enzyme with respect to the numerical value of the objective function can be categorized into three different groups. The first group exhibits an increase of the objective function for *all* models in the ensemble, as shown by the green arrows. In this case, the 25th percentile characterizing the effect of the given enzyme perturbation lies above the basal level of the objective function. Conversely, the second group is characterized by

## 5. Strain Design Algorithms for Target Identification

Table 5.3.: **Objective Functions for OptEM.** The proposed algorithm allows for the optimization of user-tailored objective functions. Some examples are provided below.

Rationale	Objective Function
<p><b>Product Yield.</b> It is the simplest and most commonly used objective function in strain optimization efforts. Product yield is defined as the amount of product synthesized per amount of substrate consumed. Depending on the units used to calculate this ratio, product yield can be expressed on a molar, mass or carbon basis, as discussed in Chapter 7.</p>	$Z = \frac{\mathbf{V}_{txdn}^{pert} C_{txdn}}{\sum_i \mathbf{V}_i^{pert} C_i}$
<p>+ <b>Minimization of Toxic Intermediates.</b> Strain designs can be found so that product yield is maximized, while intracellular concentration of toxic intermediates is minimized. For instance, toxicity of high concentrations of phosphorylated pathway intermediates, e.g., isopentenyl pyrophosphate and dimethylallyl pyrophosphate, during taxadiene overproduction in <i>E. coli</i> has been experimentally observed [93].</p>	$Z = \frac{\mathbf{V}_{txdn}^{pert} C_{txd}}{\sum_i \mathbf{V}_i^{pert} C_i} \left( \frac{1}{\sum_j \mathbf{X}_j^{pert} Tox_j} \right)$
<p>+ <b>Maximization Strain Robustness.</b> Heterologous or native enzyme overexpression can generate failure of the metabolic system due to the disappearance of a stable steady state [74]. In turn, this leads to cell death and a reduced productivity [144]. System robustness can be quantitatively estimated as the probability of system failure in an ensemble of models.</p>	$Z = \frac{\mathbf{V}_{txdn}^{pert} C_{txd}}{\sum_i \mathbf{V}_i^{pert} C_i} SS^{pert}$
<p>+ <b>Minimization of Protein Burden.</b> Enzyme overexpression draws energy and biomass building blocks from the cellular metabolism. Depending on various factors, e.g., overexpression strength, enzyme amino acid sequence, carbon source, etc., enzyme overexpression can lead to a reduced cellular growth rate and consequently, a reduced productivity. Strain designs can be identified so that product yield is maximized, while the amount of enzyme invested is reduced.</p>	$Z = \frac{\mathbf{V}_{txdn}^{pert} C_{txd}}{\sum_i \mathbf{V}_i^{pert} C_i} \left( \frac{1}{\sum_k \mathbf{U}_k^{pert} MW_k} \right)$

Vectors  $\mathbf{V}^{pert}$ ,  $\mathbf{X}^{pert}$  and  $\mathbf{U}^{pert}$  and variable  $SS^{pert}$  are obtained during the second step of the OptEM algorithm, as explained in the main text.  $\mathbf{V}_{txdn}^{pert}$  and  $C_{txdn}$  represent taxadiene flux and number of carbon atoms in a taxadiene molecule, respectively. Variables  $i$ ,  $j$  and  $k$  define the set of uptake reactions for carbon sources, toxic intracellular metabolites and perturbed enzymes, respectively. Finally,  $C_i$ ,  $Tox_j$  and  $MW_k$  correspond to the number of carbon atoms in the carbon source absorbed in reaction  $i$ , relative toxicity index for metabolite  $j$  and molecular weight of enzyme  $k$ , respectively.

#### 5.4. Application of Strain-design Algorithms for Target Identification

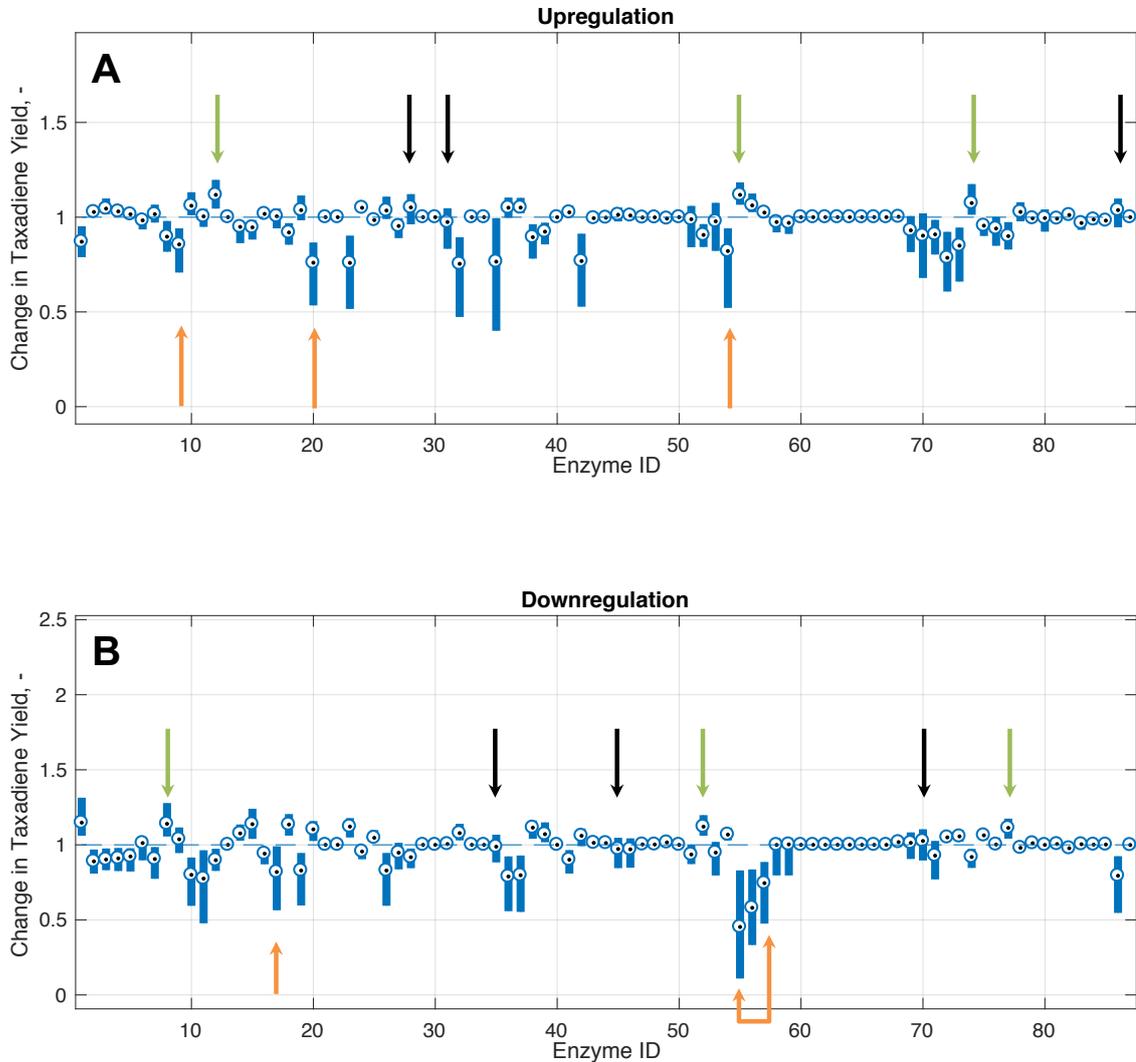


Figure 5.8.: **Effect of Single Enzyme Up- and Down-regulation Predicted by OptEM for the Maximization of Taxadiene Yield in Strain  $\Delta XylE\_1p$  via a Brute Force Approach.** Effect of single enzyme up- (A.) and down-regulation (B.) on taxadiene yield is shown. Ensemble of predictions for each perturbation is characterized by calculating the median and the 25th and 75th percentiles. Product yield after each enzyme perturbation is normalized to that of the wild-type strain, which in this specific case is assumed to exhibit a yield of  $2.860 \times 10^{-5}$  mol taxadiene/mol glucose. Green arrows exemplarily highlight perturbations *increasing* the value of the objective function, while orange and black arrows mark perturbations *decreasing* and *inconclusively* affecting the value of the objective function, respectively. Note that in this particular case, individual enzyme identity is not relevant. Rather, we aimed at showing the nature of the outcome that single enzyme perturbation might originate on the objective function, as predicted by OptEM. The model used corresponds to a modified version of the core *E. coli* core [107], containing the taxadiene production pathway and 17 efflux reactions representing biomass synthesis. Ensemble size corresponds to 100 models.

## 5. Strain Design Algorithms for Target Identification

Table 5.4.: **Enzyme Perturbations Identified by OptEM Leading to an Increased Taxadiene Yield in Strain  $\Delta$ XylE\_1p.** 100 models, three different flux distributions. Note that each genetic perturbation corresponds to a single strain design strategy.

Perturbation	Target	Reaction Name	Reaction Details
Up	dxs	1-deoxyxylulose-5-phosphate synthase	$\text{Pyr} + \text{G3P} \longrightarrow \text{DXYL5P} + \text{CO}_2$
	ispA	Dimethylallyltranstransferase	$\text{DMPP} + \text{IPP} \longrightarrow \text{GRDP} + 2 \text{Pi}$
Down	PTAr	Phosphotransacetylase	$\text{AccoA} + \text{Pi} \longrightarrow \text{ACTP} + \text{CoA}$
	Pyk	Pyruvate kinase	$\text{ADP} + \text{PEP} + \text{H}^+ \longrightarrow \text{ATP} + \text{Pyr}$

Metabolites: **ATP**: adenosine triphosphate; **ADP**: adenosine diphosphate; **Pyr**: pyruvate; **G3P**: glyceraldehyde 3-phosphate; **DXYL5P**: 1-deoxy-D-xylulose 5-phosphate; **DMPP**: dimethylallyl diphosphate; **IPP**: isopentenyl diphosphate; **GRDP**: geranyl diphosphate; **Pi**: phosphate; **AccoA**: acetyl-Coa; **ACTP**: acetyl phosphate; **CoA**: coenzyme A; **PEP**: phosphoenolpyruvate.

a decrease in the value of the objective function in *all* models, as shown by the orange arrows. This outcome is characterized by a 75th percentile lying under the basal level of the objective function. Finally, the third group has an undefined outcome. In this situation and depending on model parametrization, a specific enzyme perturbation can increase the objective function in some models, but also decrease it in others. This situation is characterized by a 25th percentile lying under the basal level of the objective function, while the 75th percentile lies above it. This analysis should help define rules for target identification. In this specific work, a given enzyme perturbation is categorized as positive only if the 25th percentile lies above the basal level of the objective function (first group).

As stated before, experimental concentration time courses for taxadiene were not available for any strain described in this chapter. However, the identification of genetic targets is still possible, if a specific taxadiene production rate is assumed that lies inside the intervals calculated in Section 5.3. Clearly, experimental implementation of the identified genetic designs will only lead to an increased taxadiene yield *in vivo*, if the assumption regarding the taxadiene production level was more or less accurate. Since taxadiene production ranges are rather wide -see Figure 5.3-, assuming a production rate accurately matching the actual, unknown rate becomes a difficult task. A methodology to overcome this problem and identify effective targets is presented as follows. For a strain in Table 5.1:

1. Identify a range of feasible production rates, as described in Section 5.3.
2. Sample the space of feasible production rates.
3. For each sample, use linear programming to generate a reference flux distribution that fulfills the assumed production rate (or yield), experimentally measured exchange rates and available  $^{13}\text{C}$  labeling data and, maximizes/minimizes a defined biologically feasible objective function. ATP maximization or minimization of oxygen uptake rate are some examples.
4. Apply the OptEM methodology to identify genetic perturbations maximizing a defined design objective function. Table 5.3 provides some examples.
5. Identify strain design strategies being independent on the assumed production rate.

Table 5.4 summarizes the results obtained for strain  $\Delta$ XylE\_1p after applying the methodology described above and assuming three different taxadiene production levels:  $1.385 \times 10^{-7}$ ,  $2.431 \times 10^{-4}$  and  $4.26 \times 10^{-1}$  mmol taxadiene/gDW h, which correspond to the minimal,

#### 5.4. Application of Strain-design Algorithms for Target Identification

medium and maximal production rates, as identified by the constraint-based method described in Section 5.3, respectively. OptEM suggests that the individual up-regulation of the enzymes catalyzing the first step of two key pathways, namely non-mevalonate pathway (*dxs*) and taxadiene synthesis (*ispA*), would lead to an increased taxadiene yield. Interestingly, both genes are located in the same operon [81]. Experimental studies supporting the positive effect of overexpressing *dxs* (encoding 1-deoxyxylulose-5-phosphate synthase) and *ispA* (encoding a dimethylallyltranstransferase) on isoprenoid production can be found in the literature [54, 58, 154, 166], thus confirming the prediction power of the OptEM approach. Regarding down-regulation targets, the modulation of Pyk presumably serves the purpose of balancing pool levels of glyceraldehyde 3-phosphate (G3P) and pyruvate, the two precursors of the non-mevalonate pathway. Farmer et al. [39] reported that the simultaneous knockout of genes *pykF* and *pykA*, coding for pyruvate kinase I and II, respectively, increased the lycopene production 2.8-fold when compared to the control strain. The study explored many strategies to balance the level of both metabolites and concluded that: alterations in the central metabolism that redirect flux from pyruvate back to G3P enhance lycopene production, while alterations that channel carbon flux away from the G3P pool have the opposite effect, showing that G3P may be limiting in the biosynthesis of lycopene. Since the synthesis of taxadiene and lycopene differs only in the last reaction step, it can be expected that this observation holds true for taxadiene production and more broadly, for the terpenoid production via the non-mevalonate pathway. The second down-regulation target identified by OptEM corresponds to PTAr. This strategy can be rationalized by the fact that down-regulating PTAr necessarily leads to a reduction in the acetate production. Consequently, more carbon atoms are available for taxadiene overproduction.

The same methodology was applied for all other strains listed in Table 5.1. *dxs* and *ispA* were recurring up-regulation targets, additionally identified for strains HMS\_2p and  $\Delta$ xylE\_2p. Further up-regulation targets corresponded to *idi* for strain  $\Delta$ xylE\_2p and *pgi* for strain  $\Delta$ ptsG\_2p. Regarding down-regulation targets, *pgk* was identified for strain  $\Delta$ ptsG\_2p. Experimental evidence of the positive effect of *idi* amplification on taxadiene [16] and lycopene [24] production can be found in the literature. Additionally, *pgi* over-expression has also been reported to increase lycopene titers 2.3-fold compared to respective control strains [24]. We did not find experimental evidence supporting the positive effect of the *pgk* down-regulation on taxadiene synthesis. However, this strategy can be understood using the rationale provided in the previously mentioned study by Farmer et al. [39]. The gene product of *pgk*, phosphoglycerate kinase, catalyzes following reaction: 3-phospho-D-glycerate + ATP  $\leftrightarrow$  3-phospho-D-glyceroyl-phosphate + ADP. Reducing the flux through this reaction would necessarily lead to a reduction in the upstream reaction catalyzed by the G3P dehydrogenase: D-glyceraldehyde 3-phosphate + NAD<sup>+</sup> + phosphate  $\leftrightarrow$  3-phospho-D-glyceroyl-phosphate + NADH + H<sup>+</sup>. Thus, the expected overall effect of *pgk* down-regulation is an increase of the G3P pool which, as experimentally demonstrated by Farmer et al. [39], enhances terpenoid production.

#### Solver Approach

All strain designs previously discussed consist of one single genetic perturbation. However, experimental evidence strongly suggests that, very often, more than one genetic perturbation must be conducted to efficiently redirect the cellular metabolism towards

## 5. Strain Design Algorithms for Target Identification

Table 5.5.: **Strain Design Strategy Identified by OptEM for the Maximization of Taxadiene Yield in Strain  $\Delta XylE\_1p$  Using a Solver Approach.** A down-regulation of enzymes ALCD2x and TALA along with up-regulation of IDI were identified by the OptEM approach to increase the taxadiene yield 1.4-fold in strain  $\Delta XylE\_1p$ . In the case of ethanol secretion, the efficacy of experimentally down-regulating enzyme ALCD2x is questioned by its wide flux range. Note that flux ranges for all reactions were identified using  $^{13}C$ -labeling data and experimentally determined values for exchange reactions listed in Table 5.2.

Reaction	Reaction Stoichiometry	Reference Flux	Flux Range	Reaction Perturbation Fold
ALCD2x	ETOH + NAD $\leftrightarrow$ ACALD + H + NADH	-11.088	[-15.5 ; 0]	0.1
TALA	G3P + S7P $\leftrightarrow$ E4P + F6P	$-1.815 \times 10^{-2}$	$[-1.815 \times 10^{-2} ; 0]$	0.19
IDI	DMPP $\leftrightarrow$ IPP	$-2.75 \times 10^{-3}$	[-1.353 ; 4.059]	2.21

Metabolites: **ETOH**: ethanol; **ACALD**: acetaldehyde; **S7P**: sedoheptulose 7-phosphate; **E4P**: D-Erythrose 4-phosphate; **F6P**: D-Fructose 6-phosphate; **G3P**: glyceraldehyde 3-phosphate; **DMPP**: dimethylallyl diphosphate; **IPP**: isopentenyl diphosphate;

Reactions (Enzymes): **ALCD2x**: alcohol dehydrogenase ; **TALA**: transaldolase ; **IDI**: isopentenyl-diphosphate delta-isomerase.

Flux units are given in mmol/gDW h.

overproduction of the target molecule. For instance, the sole *idi* over-expression increased the specific taxadiene titer 3.6-fold, from 0.5 mg/gDW in the control strain to 1.8 mg/DW in the strain over-expressing *idi*. By contrast, when combined with *ppk* or *sthA*, *idi* over-expression increased the specific taxadiene titer 11-fold, from 0.5 mg/DW in the control strain, to approximately 7 mg/DW [16]. A similar observation was made for studies reporting the lycopene [24] and abietadiene [98] overproduction in *E. coli*. In both cases, maximal product yield is reached after simultaneously perturbing two and three genes, respectively.

Motivated by this observation, a Genetic Algorithm (GA) was employed along with the OptEM approach to efficiently identify an enzyme perturbation vector maximizing an user defined objective function, in this specific case, product yield. Unlike the *extensive* search performed by the brute force approach to identify optimal single perturbations, the GA *selectively* searches the parameter space. The algorithm was implemented using a GA script, which contained nested FMINCON optimization routines. Figure 5.9 shows the optimal genetic perturbation strategy, identified by the OptEM algorithm for the strain  $\Delta XylE\_1p$ , when a taxadiene production rate of  $2.431 \times 10^{-4}$  mmol/gDW h is assumed and an ensemble of 100 kinetic models is analyzed. In order to increase the success chance of the experimental implementation of the enzymatic perturbation identified by OptEM in silico, the objective function was adjusted to consider the 25th percentile of all responses in the ensemble for the tested perturbation (refer to Figure 5.8, green arrows). Note that alternative approaches, e.g., considering the median or mean response, are also possible.

Table 5.5 and Figure 5.9 summarizes the OptEM predictions for strain  $\Delta XylE\_1p$ . As suggested by the algorithm, down-regulating enzymes ALCD2x and TALA 0.1- and 0.19-fold respectively, together with 2.21-fold up-regulation of enzyme IDI, increased taxadiene yield by at least 40% in 75% of the models encompassing the ensemble of 100 models.

Additional to the perturbation extend of each of the three enzymes, Table 5.5 also includes the reference flux for each reaction used for ensemble construction and its corresponding flux range. Provided flux ranges correspond to the 95% confidence interval calculated using available  $^{13}C$ -labeling data and values for exchange reactions listed in Table 5.2.

#### 5.4. Application of Strain-design Algorithms for Target Identification

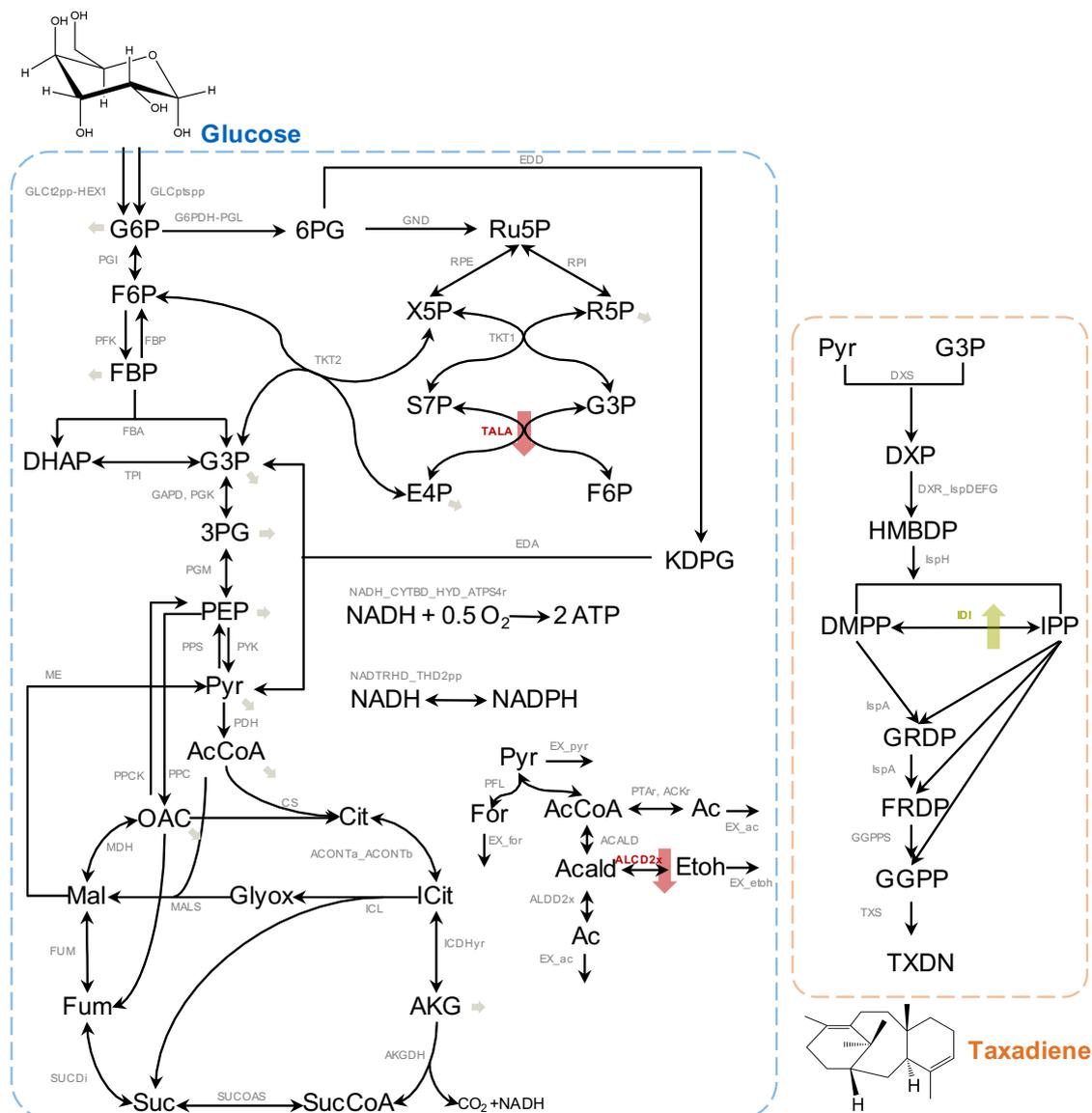


Figure 5.9.: **Strain Design Strategy Identified by OptEM for the Maximization of Taxadiene Yield in Strain  $\Delta$ XylE\_1p Using a Solver Approach.** Model used corresponds to the *E. coli* core metabolism (reactions contained in the blue rectangle), which was extended to allow for taxadiene synthesis (reactions contained in the orange rectangle). The model contains a total of 87 reactions and 69 metabolites. Reference flux distribution was generated considering available  $^{13}\text{C}$ -labeling data, experimentally determined exchange rates and an assumed taxadiene production rate of  $2.431 \times 10^{-4}$  mmol/gDW h. Matlab scripts used correspond to the FMINCON and GA optimization routines. For the GA, a population size of 10 was used, with a maximal number of 5 generations. Additionally, three reactions were allowed to be simultaneously up- and down-regulated 10- and 0.1-fold, respectively. Solver converged after 185 hours and identified strain design shown, which improved the reference taxadiene production 1.4-fold.

## 5. Strain Design Algorithms for Target Identification

Refer to Section 2.1.2 for more details on the calculation of confidence intervals. Reaction ALCD2x, representing ethanol production, has a large confidence interval for its rate. In terms of experimental efficacy, a wide confidence interval for a given reaction suggest that the outcome of perturbing its associated enzyme might not deliver results predicted *in silico*. More specifically, *in vivo* knocking out the gene *adhE*, whose gene product catalyzes the ALCD2x reaction, might not be even necessary, since the reaction ALCD2x, in accordance with available experimental measurements, might also carry a flux of zero under the experimental conditions tested. This rationale can be mathematically implemented by modifying the objective function as follows:

$$Z = \frac{\mathbf{V}_{txdn}^{pert} C_{txdn}}{\sum_i \mathbf{V}_i^{pert} C_i} \left( \frac{1}{\prod_k \frac{\mathbf{V}_{k,max}^{ref} - \mathbf{V}_{k,min}^{ref}}{\mathbf{V}_k^{ref}}} \right)$$

where  $\mathbf{V}_{txdn}^{pert}$  refers to the taxadiene production flux after enzyme perturbation and  $\mathbf{V}_i^{pert}$  to the uptake rates of carbon sources.  $C_i$  and  $C_{txdn}$  are weights used to transform rates from molar flux (mmol/gDW h) to carbon or mass flux. Alternatively, if only one single carbon source is used for production, weights can be ignored. Consequently, product yields are calculated in a molar basis. Finally,  $\mathbf{V}^{ref}$  refers to the vector containing the reference flux distribution. The count variable  $k$  defines the set of perturbed enzymes (reactions). Indices *min* and *max* respectively refer to lower and upper bound of the confidence interval. By using the objective function shown above, OptEM will mainly identify enzyme perturbation strategies whose affected reactions exhibit tight confidence intervals. In this way, the effect of uncertainties related to alternative flux distributions on target selection can be partly diminished. Note that if  $\mathbf{V}_{k,max}^{ref} = \mathbf{V}_{k,min}^{ref}$  for a given reaction  $k$ , the objective function should to be adjusted to avoid numerical problems. In this case, it suffices to replace the expression  $\frac{\mathbf{V}_{k,max}^{ref} - \mathbf{V}_{k,min}^{ref}}{\mathbf{V}_k^{ref}}$  by an arbitrary high number.

## 5.5. Discussion

Our cooperation partner Monika Fuchs at **IBK** constructed and characterized five different taxadiene producing strains. Necessary genes for optimal taxadiene synthesis were provided by plasmids pET28a\_DS and pACY\_MEP. In strains  $\Delta ptsG\_1p$  and  $\Delta XylE\_1p$ , genes contained in plasmid pACY\_MEP were integrated into the *E. coli* genome, in the *ptsG* and *xylE* position, respectively. Experimental characterization encompassed measurement of concentration time courses for biomass, glucose and acetate. Additionally,  $^{13}C$ -labeling patterns of protein-bound amino acids were also provided. Due to low production rates and concentrations in the culture medium, determination of time courses for taxadiene concentration was not possible for any strain for the low sample volume used.

Our first task was the characterization of the reduction of the metabolic burden in strains  $\Delta ptsG\_1p$  and  $\Delta XylE\_1p$ . Using the concept of acetate lines [11], we showed that genomic integration of plasmid-coded genes reduced the metabolic burden. As suggested by the position of the acetate lines for strains  $\Delta ptsG\_1p$  and  $\Delta XylE\_1p$  relative to their corresponding parent strains  $\Delta ptsG\_2p$  and  $\Delta XylE\_2p$ , genomic integration seemed to reduce the metabolic burden within strains carrying *ptsG* deletion in a bigger extent than

within strains carrying  $\Delta xylE$  deletion (Figure 5.2). Conversely, it holds that strains carrying a  $ptsG$  deletion exhibit a higher metabolic burden when loaded with plasmid-based enzyme expression. Interestingly, this trend was also observed within strains used for sugar co-utilization studies (refer to Figure 6.11.B in Chapter 6).

In order to assess the effect of metabolic burden reduction on taxadiene production, we applied a constraint-based approach, namely FVA, on a modified version of the model iRL2013 [75]. Note that this model-based analysis was necessary due to the missing taxadiene concentration time courses. For strains  $\Delta ptsG\_1p$  and  $\Delta XylE\_1p$ , FVA predicted a higher maximal theoretical taxadiene yield, compared to their respective parent strains. This observation suggests that a metabolic burden reduction can *potentially* increase taxadiene production. However, taxadiene production flux for all strains could also be almost zero, as suggested by the lower boundary of the taxadiene export reaction. Refer to Table A.9 in Appendix A.2.5.

One of our major tasks was the model-driven strain optimization. We first applied OptKnock [20] to predict triple reaction knock-outs potentially improving taxadiene production under two different oxygen levels. The motivation behind this was twofold:

1. Basan et al. [11] showed that loading the cells with plasmid-based enzyme expression causes acetate secretion to occur at lower growth rates which, as we shown in Chapter 4, can be *in silico* modeled by reducing the maximal oxygen uptake rate of loaded strains.
2. It has been recognized that oxygen supply is one of the most frequent problems associated with usage of shaking flasks [19]. By contrast, on-line monitoring and control of oxygen supply is possible in stirred bio-reactors.

Since both factors, genetic load and process scale-up (from shaking flask to stirred bio-reactor) are relevant in the construction and screening of optimal production strains, the effect of a variable oxygen supply on optimal reaction knock-outs needed to be assessed. Interestingly, we observed that oxygen availability indeed had an effect on the set of deletion candidates predicted by optKnock. This implies that a given strain design, optimized using shaking flasks, might not exhibit the same optimal production phenotype in a large scale, stirred bio-reactor. Moreover, this also suggests that additional optimization rounds might be necessary if the level of metabolic load is changed, as plasmid-based enzyme expression presumably affects the respiration capabilities of the cell.

In a recent publication [149], we showed that in spite of their multiple flaws, constraint-based strain design algorithms, e.g. OptKnock, clearly outnumber kinetic-based strain-design algorithms. Motivated by this fact, we developed OptEM, a kinetic-based strain-design algorithm that makes use of the ensemble modeling approach to construct an ensemble of kinetic models, all fulfilling a given reference flux distribution. We applied OptEM to identify single and triple enzyme perturbations leading to an increased taxadiene yield. Single enzyme perturbations were identified by a brute-force approach which extensively explored the parameter space. Remarkably, identified single enzyme perturbation strategies agreed well with reported *in vivo* strategies, suggesting a good prediction power of our strain-design algorithm. The triple enzyme perturbation strategy identified by OptEM for the strain  $\Delta XylE\_1p$  (refer to Table 5.5) can be rationalized in the context of the specific flux distribution used for this strain to populate the ensemble of kinetic models.

## 5. Strain Design Algorithms for Target Identification

More specifically, down-regulation of reaction ALCD2x, responsible for ethanol formation, aims at minimizing carbon loss. Note that the reference flux through this reaction corresponds to roughly 40% of the total carbon uptake. Down-regulation of TALA possibly targets at an increase in the G3P pool. Finally, IDI up-regulation is consistent with many experimental reports describing the positive effect of its over-expression on taxadiene [16] and lycopene [24] production in *E. coli*. Its positive effect on terpenoids production possibly points at sub-optimal DMPP to IPP ratios under physiological conditions and to an interdependence in the optimal expression levels of genes *ispH*, *idi* and *ispA*

The modeling approach used here to circumvent unknown kinetic parameters, a limitation commonly associated with kinetic-based descriptions of the cellular metabolism, opens a wide range of valuable applications. As shown in this chapter, one of these applications is the identification of relevant enzyme perturbations increasing yield of a target molecule. Another application, the assessment of production potential based on the flux distribution of a reference strain, will be applied in Chapter 6 and validated in Chapter 9 using experimental data. While specific kinetic parameter values are not required for any of these applications, computational power is mandatory in order to efficiently analyze large ensembles of kinetic models. Since computer clusters are not yet accessible for the majority of the research community, new mathematical methods are required to efficiently sample relevant parameter sets, thus shrinking the size of the model ensemble. Some options include:

- Application of parallel  $^{13}\text{C}$ -labeling experiments to accurately estimate the reference flux distribution [75] and thus reduce the scope of kinetic model.
- Consideration of thermodynamic constraints to reduce the space of feasible parameter sets, as suggested by Liebermeister et al. [78, 79]
- Reducing the scope of sampled parameters by incorporating kinetic information of well characterized enzymes.

Due to difficulties related to the impossibility of determining experimental taxadiene concentrations for low volume samples, and to a re-orientation of the **SysBioTerp** project, none of the model-driven predictions made in this chapter were experimentally implemented. However, experimental data motivated the development of OptEM, a novel kinetic-based strain-design algorithm, which as demonstrated, allows the identification of biologically feasible perturbation targets.

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

Enabling simultaneous utilization of glucose and D-xylose by *E. coli* for the production of valuable chemicals is a key milestone towards a bio-based economy. Moving toward this goal, six different strains exhibiting two different genetic backgrounds were constructed and experimentally characterized by our cooperation partner Katarina Kemper at **IBK**. All strains were assessed regarding sugar co-uptake level, metabolic burden and taxadiene production potential by applying constraint- and kinetic-based approaches. In order to describe the extent of simultaneous sugar utilization, we introduced a parameter we termed “% Carbon Xylose”. This parameter ranges from 0, when all carbon atoms absorbed by the cell correspond to glucose, to 1, when all carbon atoms absorbed by the cell correspond to xylose. Using this parameter in combination with constraint-based modeling, we successfully predicted the effect of increasing co-substrate uptake on the biomass yield and growth rate. We also analyzed the effect that the simultaneous sugar utilization had on the cellular capacity via acetate lines. Based on experimental data, we hypothesized that metabolic burden is a phenomenon not only limited to protein over-expression, but can also arise due to an increasing usage of protein-expensive pathways, like the pentose phosphate pathway. We observed an increase of total protein content as sugar co-uptake increased and a downwards displacement of acetate lines as sugar co-uptake increased. Finally, we generated a reference flux distribution for each strain using  $^{13}\text{C}$  labeling data, which we used to construct an ensemble of models describing the cellular phenotype of each strain and to estimate the taxadiene production potential using a kinetic-based method.

### 6.1. Strains & Experimental Data

All strains presented in this section were constructed and characterized by our cooperation partner Katarina Kemper at **IBK**. Table 6.1 presents an overview of all strains, while Table 6.2 contains experimentally determined exchange rates, estimated as explained in Section 2.1.2. In total, six different strains were constructed and grouped into two strain families for analysis. Strains HMS, HMS\_p1x and HMS\_p3x belong to the family we term HMS, while all other strains belong to the family  $\Delta\text{ptsG}$ , in which as its name suggests, the gene *ptsG* has been knocked out.

In order to increase the degree of simultaneous utilization of glucose and D-xylose, a plasmid-encoded XylE symporter is constitutively expressed in four of the strains listed in Table 6.1. Two different mutants of the XylE protein were implemented: the first enzyme variant contained one single mutation in its amino acid sequence (p13-XylE<sub>1x</sub>), while the second variant contained three mutations in its amino acids sequence (p13-XylE<sub>3x</sub>).

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

Table 6.1.: **Strains for Sugar Co-utilization Studies.** Six strains were constructed and experimentally characterized. The first two strains, HMS and  $\Delta ptsG$  correspond to basis strains for each of the two main strain families. p13-XylE<sub>1x</sub> refers to a plasmid-coded XylE transporter with 1 mutation in its amino acid sequence, while p13-XylE<sub>3x</sub> refers to a plasmid-coded XylE transporter containing 3 mutations in its amino acid sequence.

Strain ID	Genotype	Description
HMS Empty	HMS174(DE3) + p13	Wild-type (WT) strain.
$\Delta ptsG$ Empty	HMS174(DE3), $\Delta ptsG$ + p13	WT background with deleted <i>ptsG</i> gene.
HMS p1x	HMS174(DE3) + p13-XylE <sub>1x</sub>	WT strain + plasmid constitutively expressing XylE transporter mutated in 1 position.
HMS p3x	HMS174(DE3) + p13-XylE <sub>3x</sub>	WT strain + plasmid constitutively expressing XylE transporter mutated in 3 positions.
$\Delta ptsG$ p1x	HMS174(DE3), $\Delta ptsG$ + p13-XylE <sub>1x</sub>	$\Delta ptsG$ strain + plasmid constitutively expressing XylE transporter mutated in 1 position.
$\Delta ptsG$ p3x	HMS174(DE3), $\Delta ptsG$ + p13-XylE <sub>3x</sub>	$\Delta ptsG$ strain + plasmid constitutively expressing XylE transporter mutated in 3 positions.

Table 6.2.: **Experimental Characterization of Strains Used in Sugar Co-utilization Studies.** Reported rates correspond to the mean value of two biological replicates. Errors represent one standard deviation. Numerical values for the exchange rates were calculated as explained in Section 2.1.2 using experimental concentration time courses provided in Table A.10 of the Appendix.

Strain	Growth Rate	Glucose Uptake Rate	Xylose Uptake Rate	Acetate Secretion Rate
HMS Empty	$0.258 \pm 2.88 \times 10^{-3}$	$5.105 \pm 2.92 \times 10^{-1}$	$2.057 \pm 3.45 \times 10^{-1}$	$5.190 \pm 7.26 \times 10^{-2}$
$\Delta ptsG$ Empty	$0.481 \pm 1.42 \times 10^{-2}$	$3.140 \pm 1.46 \times 10^{-1}$	$6.535 \pm 6.60 \times 10^{-2}$	$0.756 \pm 3.22 \times 10^{-2}$
HMS p1x	$0.487 \pm 1.28 \times 10^{-2}$	$5.659 \pm 2.51 \times 10^{-1}$	$2.211 \pm 4.67 \times 10^{-1}$	$6.559 \pm 2.57 \times 10^{-1}$
HMS p3x	$0.316 \pm 9.53 \times 10^{-2}$	$10.642 \pm 7.28$	$11.691 \pm 1.38 \times 10^1$	$3.735 \pm 3.13$
$\Delta ptsG$ p1x	$0.422 \pm 4.29 \times 10^{-3}$	$3.090 \pm 6.45 \times 10^{-1}$	$4.955 \pm 6.74 \times 10^{-1}$	$7.251 \pm 5.78 \times 10^{-1}$
$\Delta ptsG$ p3x	$0.344 \pm 1.32 \times 10^{-2}$	$1.599 \pm 2.34 \times 10^{-1}$	$7.649 \pm 6.77 \times 10^{-1}$	$8.779 \pm 5.51 \times 10^{-1}$

Growth rates are reported 1/h; glucose, xylose and acetate rates are given in mmol/gDW h.

Strains were experimentally characterized in duplicate. Experimental measurements consisted of concentration time courses for OD, glucose, D-xylose and acetate (refer to Appendix A.3.1 for raw data). All strains were cultivated in Riesenbergl medium containing 2.5 g/l D-xylose, 2 g/l glucose and 0.5 g/l uniformly labeled <sup>13</sup>C-glucose. Samples for labeling analysis were taken after 7 hours culture. Raw labeling data can be found in Appendix A.3.2. Exchange rates shown in Table 6.2 were used to calculate three strain performance parameters, namely % Carbon Xylose, total carbon uptake and biomass yield. The first performance parameter quantifies the extent of sugar co-utilization exhibited by each strain and is defined as follows:

$$\% \text{ Carbon Xylose} = \frac{5 r_{xylose}}{5 r_{xylose} + 6 r_{glucose}}$$

where  $r_{xylose}$  and  $r_{glucose}$  refer to uptake rates for xylose and glucose, respectively. These rates are multiplied by 5 and 6 in order to obtain carbon fluxes (C-mmol/gDW h) for each sugar. Clearly, % Carbon Xylose can take values between 0 and 1, and describes all possible scenarios, from a sole glucose utilization to sole xylose utilization, respectively. Strains listed in Table 6.2 exhibit a wide range of simultaneous sugar utilization, ranging from 0.25 for the parent strain HMS to 0.8 for the engineered strain  $\Delta ptsG$  p3x.

Figure 6.1 graphically shows the effect of increasing the extent of simultaneous xylose

## 6.2. Constraint-based Characterization of Simultaneous Sugar Utilization

Table 6.3.: **Strain Performance Parameters.** Extent of sugar co-utilization is quantitatively described by the variable % Carbon Xylose. Total carbon uptake was calculated as the weighted sum of xylose and glucose uptake. Weights correspond to the sugar carbon content, i.e. 5 and 6 respectively. Biomass yield was calculated as the ratio between growth rate and the weighted sum of substrate uptake rate for xylose and glucose. In this case, weights correspond to molecular weights of each substrate. Performance parameters were calculated independently for each of the two biological replicates. Reported values correspond to the mean value. Errors represent one standard deviation.

Strain	% Carbon Xylose	Total Carbon Uptake	Biomass Yield
HMS Empty	$0.250 \pm 2.09 \times 10^{-2}$	$40.912 \pm 3.48$	$0.211 \pm 1.56 \times 10^{-2}$
$\Delta$ ptsG Empty	$0.634 \pm 8.44 \times 10^{-3}$	$51.512 \pm 1.21$	$0.311 \pm 1.91 \times 10^{-3}$
HMS p1x	$0.245 \pm 4.74 \times 10^{-2}$	$45.007 \pm 8.26 \times 10^{-1}$	$0.361 \pm 1.61 \times 10^{-2}$
HMS p3x	$0.380 \pm 2.13 \times 10^{-1}$	$122.308 \pm 1.13 \times 10^2$	$0.170 \pm 1.82 \times 10^{-1}$
$\Delta$ ptsG p1x	$0.574 \pm 1.80 \times 10^{-2}$	$43.316 \pm 7.23$	$0.329 \pm 5.16 \times 10^{-2}$
$\Delta$ ptsG p3x	$0.800 \pm 9.32 \times 10^{-3}$	$47.838 \pm 4.79$	$0.240 \pm 1.49 \times 10^{-2}$

Total carbon uptake rates are given in c-mmol/gDW h; biomass yields are given in g biomass/g substrate.

and glucose utilization on biomass yield, growth rate and total carbon uptake. While the total carbon uptake rate seems to be independent on the extent of sugar co-utilization, both biomass yield and growth rate seem to decrease as the fraction of absorbed xylose is increased. Note that the experimental data corresponding to the strain HMS p3x seems to correspond to an outlier. This is confirmed by large standard deviations in all its performance parameters.

The effect of the *ptsG* gene deletion and mutations in the transporter protein XylE on the extent of xylose and glucose co-utilization can be evidenced and quantified using the introduced variable % Carbon Xylose. Overall, the strain family  $\Delta$ ptsG exhibited higher levels of sugar co-utilization, with growth rates and biomass yields being comparable to those of the HMS strain family. Expressing the single mutated transporter XylE<sub>1x</sub> does not seem to increase sugar co-utilization levels, when compared to the respective control strain. By contrast, expression of the triple mutated transporter XylE<sub>3x</sub> consistently increased the extend of sugar co-utilization within the strain family  $\Delta$ ptsG. Due to large standard deviations in the experimental measurements for strain HMS p3x, the effect of expressing the triple mutated transporter XylE<sub>3x</sub> cannot be precisely quantified. However, the mean value of % Carbon Xylose for this strain suggests a positive effect on sugar co-utilization. (see Figure 6.1)

## 6.2. Constraint-based Characterization of Simultaneous Sugar Utilization

Our first modeling efforts are focused on a constraint-based description of the simultaneous sugar utilization in *E. coli*. More specifically, experimentally obtained values for total carbon uptake along with *in silico* estimated oxygen uptake rates are used to describe the effect of increasing sugar co-utilization on growth rate and biomass yield. Then, the concept of acetate lines presented in Chapter 4 is utilized to estimate the level of metabolic burden in each strain and to unveil its connection with increasing levels of sugar co-utilization.

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

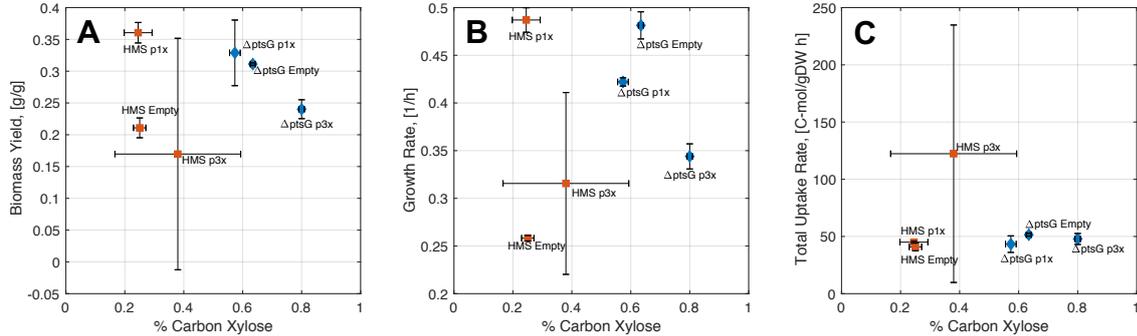


Figure 6.1.: **Strain Performance Parameters as a Function of Simultaneous Sugar Co-utilization Extent.** Effect of simultaneous glucose and xylose utilization on biomass yield. Reported values correspond to the mean value of two biological replicates. Error bars represent one standard deviation. (A.), cellular growth rate (B.) and total carbon uptake rate (C.). For strain families HMS and  $\Delta$ ptsG, an increasing extent of sugar co-utilization seems to reduce both the biomass yield and growth rate. By contrast, total carbon uptake seems to be independent on the extent of sugar co-utilization. Experimental data for HMS p3x seem to represent an outlier.

### 6.2.1. Biomass Yield & Growth Rate

We learned from experimental data that an increase in the degree of xylose and glucose co-utilization, quantified via % Carbon Xylose, seems to reduce the cellular growth rate and the biomass yield in both strain families (see Figure 6.1.A and .B). However, the negative effect of increasing % Carbon Xylose values on growth rate and biomass yield was experimentally observed in a limited range of sugar co-utilization, i.e. from 0.24 to 0.38 for strain family HMS and from 0.57 to 0.8 for the strain family  $\Delta$ ptsG. In order to estimate the effect of sugar co-utilization on growth rate and biomass yield in the whole range of % Carbon Xylose, i.e. from 0 to 1, we employed experimentally determined exchange rates listed in Table 6.2, along with the *E. coli* genome-scale metabolic reconstruction iJO1366 [106].

We start by estimating the oxygen uptake rate for each of the six strains. For that, we minimize the reaction representing oxygen import from the medium to the cell, while simultaneously constraining reactions in the iJO1366 model representing rates for growth, glucose and xylose uptake, and acetate secretion to match experimentally determined values, as listed in Table 6.2. The estimated oxygen uptake rate for each strain is shown in Figure A.8 (in Appendix A.3.3). Overall, we observe a decrease in the oxygen uptake rate, as % Carbon Xylose increases. For strain HMS p3x, its unusually high substrate uptake (refer to Figure 6.1.C) leads to an *in silico* oxygen uptake rate of zero. Since the coefficients of determination  $R^2$  for the linear regression of oxygen uptake rate against % Carbon Xylose have relatively low values, 0.62 for strain family HMS and 0.56 for strain family  $\Delta$ ptsG (for only three data points), we decide to consider two different case studies:

- Case study I: the oxygen uptake rate ( $r_{O_2}$ ) is assumed to depend on % Carbon Xylose ( $x$ ), as described by the functions:  $r_{O_2} = -35.57x + 13.43$  for strain family HMS and  $r_{O_2} = -12.02x + 13.69$  for strain family  $\Delta$ ptsG.
- Case study II:  $r_{O_2}$  is assumed to be independent from % Carbon Xylose and from the strain family. Additionally, we assign it a constant value of 5.2 mmol  $O_2$ /gDW h,

## 6.2. Constraint-based Characterization of Simultaneous Sugar Utilization

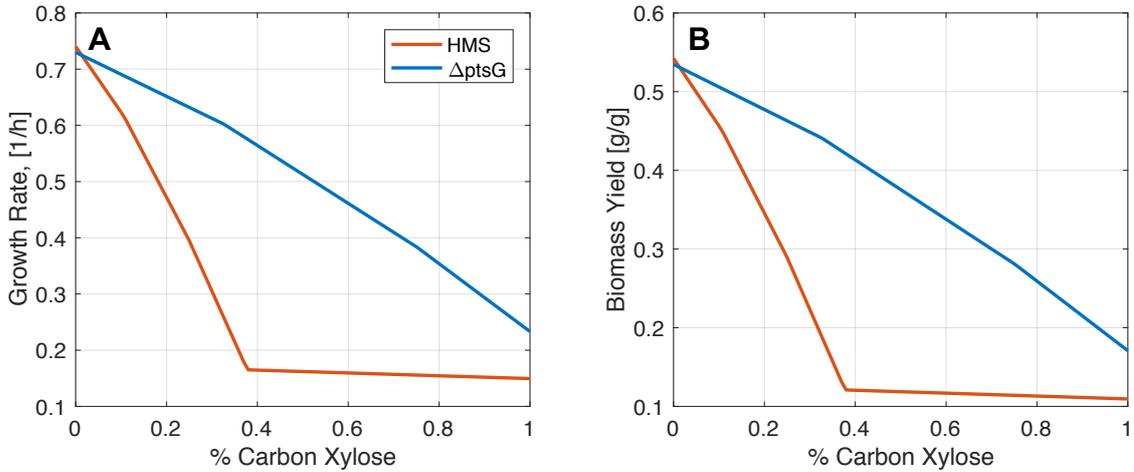


Figure 6.2.: ***In silico* Calculated Effect of Sugar Co-Utilization on Growth Rate and Biomass Yield: Case Study I.** Effect of simultaneous glucose and xylose utilization on growth rate (A.) and biomass yield (B.) as predicted using a constraint-based approach. For all calculations, the oxygen uptake rate ( $r_{O_2}$ ) is assumed to depend on % Carbon Xylose ( $x$ ), as described by the functions:  $r_{O_2} = -35.57x + 13.43$  for strain family HMS and  $r_{O_2} = -12.02x + 13.69$  for strain family  $\Delta$ ptsG. Additionally, a constant total carbon uptake rate of 45 mmol Carbon/gDW h was assumed. Given these two constraints and assuming different sugar co-utilization levels, the maximal growth rate supported by the *E. coli* metabolic reconstruction iJO1366 was calculated using the COBRA toolbox along with the solver Gurobi.

which corresponds to the mean value of all estimated oxygen uptake rates, except the one for strain HMS p3x, which is assumed to be an outlier.

The second step consists on the determination of the total carbon uptake rate. From Figure 6.1.C, we observe that the total carbon uptake rate seems to be independent from both strain family and % Carbon Xylose. Ignoring the data point for strain HMS p3x, we obtain a mean total uptake rate of 45 mmol Carbon/gDW h. Assuming a constant total carbon uptake rate matching this value, one can estimate the effect of varying % Carbon Xylose values on growth rate and biomass yield by means of linear programming. Calculations are done for the two study cases previously introduced and shown in Figure 6.2 for case study I and Figure 6.3 for case study II.

In both cases, a negative effect of increasing % Carbon Xylose values on growth rate and biomass yield is predicted, in line with experimental observations (Figure 6.1.A and .B). As for particular values for growth rate and biomass yield at a given degree of sugar co-utilization, they are greatly influenced by the assumption made on the oxygen uptake rate ( $r_{O_2}$ ). When  $r_{O_2}$  is assumed to be constant, regardless of the strain family and the degree of sugar co-utilization (Case Study II, Figure 6.3), our constraint-based approach predicts the strain family HMS to have higher growth rates and biomass yields than the strain family  $\Delta$ ptsG. This trend is reversed, when a dependency of  $r_{O_2}$  on % Carbon Xylose is assumed. Since the main objective of this *in silico* analysis was to determine the kind of effect of sugar co-utilization on growth rate and biomass yield in the whole range of % Carbon Xylose, rather than calculating specific values for these variables, we conclude that, irregardless of the assumptions made on the oxygen uptake rate, increasing values of

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

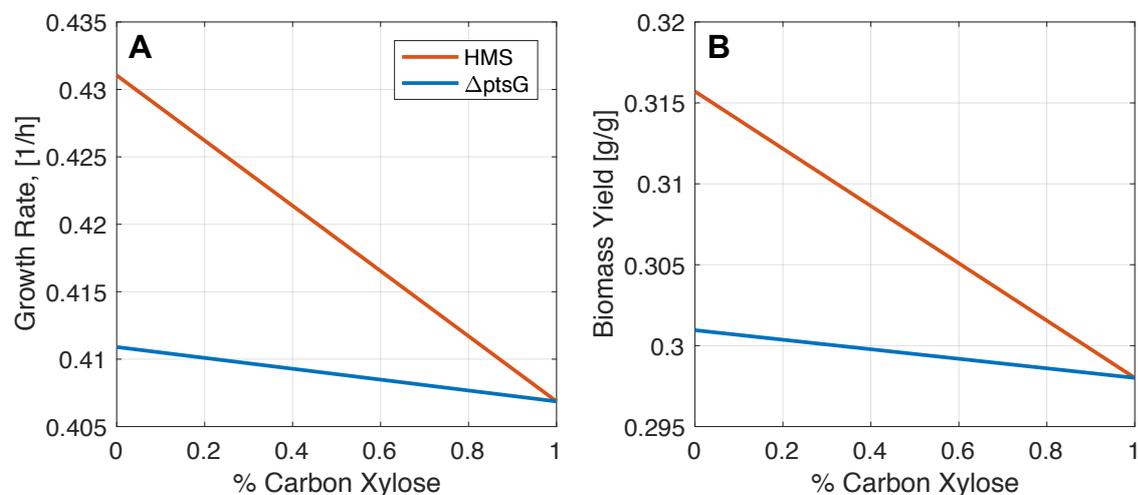


Figure 6.3.: ***In silico* Calculated Effect of Sugar Co-Utilization on Growth Rate and Biomass Yield: Case Study II.** Effect of simultaneous glucose and xylose utilization on growth rate (A.) and biomass yield (B.) as predicted using a constraint-based approach. For all calculations, a constant oxygen uptake rate of 5.2 mmol O<sub>2</sub>/gDW h and a constant total carbon uptake rate of 45 mmol Carbon/gDW h were assumed. Given these two constraints and assuming different sugar co-utilization levels, the maximal growth rate supported by the *E. coli* metabolic reconstruction iJO1366 was calculated using the COBRA toolbox along with the solver Gurobi.

% Carbon Xylose decrease both growth rate and biomass yield. Additionally, we conclude that experimentally measured oxygen uptake rates are indispensable to precisely calculate growth rates and biomass yields as a function of % Carbon Xylose.

### 6.2.2. Acetate Lines & Metabolic Burden

Plasmid-based expression of two XylE mutants, XylE<sub>3x</sub> and XylE<sub>1x</sub>, was employed to alter the degree of xylose and glucose co-utilization within two *E. coli* strain families. As discussed in Chapter 4, protein expression can cause the reduction of cellular fitness, a phenomenon commonly known as metabolic burden. Here, we aim at applying the concept of acetate lines to assess potential fitness costs involved with an increased level of % Carbon Xylose. Based on a high amino acid sequence similarity of both XylE mutants, whose amino acid sequence only differ in 3 positions out of a total of 491 amino acid residues, one can anticipate similar protein expression levels. In terms of relative position and number of acetate lines, this means that the experimental data points for each strain family should be confined to just two acetate lines; one for the parent strain carrying an empty plasmid and one for the two strains supporting the plasmid-based XylE mutant expression. Figure 6.4 schematically shows the expected relative position of the acetate lines within a given strain family.

Experimental acetate lines are obtained by plotting growth rates against acetate secretion rates and employing the previously determined slope (refer to Figure A.1, Appendix A.1.1), as shown in Figure 6.5.B. By plotting growth rate as a function of % Carbon Xylose in the adjacent Figure 6.5.A, one can analyze the effect of simultaneous sugar co-utilization on the localization of the acetate lines.

## 6.2. Constraint-based Characterization of Simultaneous Sugar Utilization

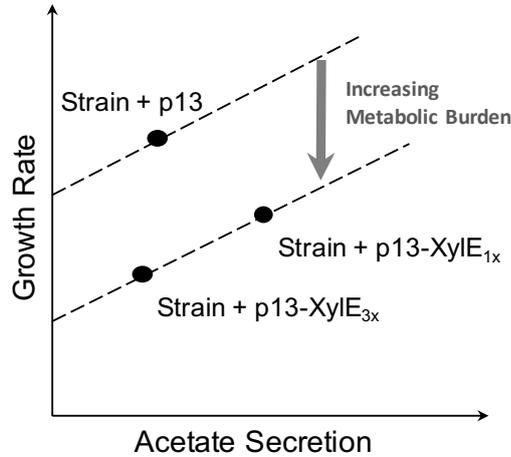


Figure 6.4.: **Expected Shift in Acetate Lines as a Result of Plasmid-based XylE-Expression.** A relative downwards shift of the acetate line is expected in strains supporting plasmid-based expression of XylE mutants, compared to the respective parent strains carrying only an empty plasmid. Since the amino acid sequence of XylE<sub>1x</sub> and XylE<sub>3x</sub> only differs in three amino acid positions, one would expect very similar protein expression levels. Consequently, the data points for strains bearing plasmids p13-XylE<sub>1x</sub> and p13-XylE<sub>3x</sub> should lie on the same acetate line.

Contrary to our expectations, we observe three acetate lines for each strain family instead of just two, as previously hypothesized based on protein sequence similitude. For instance, within strain family  $\Delta$ ptsG, strains can be sorted by increasing burden levels, suggested by the relative position of the corresponding acetate lines, as follows:  $\Delta$ ptsG Empty  $<$   $\Delta$ ptsG p1x  $<$   $\Delta$ ptsG p3x. This observation suggests that an increased degree of sugar co-utilization in strain  $\Delta$ ptsG p3x might originate a metabolic burden that cannot be explained by the sole effect of XylE expression, which corresponds to the shift of the acetate line from  $\Delta$ ptsG Empty to  $\Delta$ ptsG p1x. This trend is not observed within the strain family HMS, in which sorting by burden level leads to: HMS p1x  $<$  HMS p3x  $<$  HMS Empty. By exhibiting the highest burden within its strain family, strain HMS Empty does not fit the observed - and biologically reasonable - trend for the strain family  $\Delta$ ptsG. Consequently, we recommend our cooperation partner Katarina Kemper to experimentally characterize the strain HMS Empty one more time.

Based on its high sugar co-utilization degree and relative position of its acetate line, strain  $\Delta$ ptsG Empty represents an ideal production strain candidate. In order to confirm its outstanding performance, this strain is also suggested for experimental re-measurement. As can be observed in Figure 6.6, performance of strain  $\Delta$ ptsG was experimentally re-confirmed. By contrast, growth rate, acetate secretion rate and % Carbon Xylose of strain HMS Empty were significantly affected by experimental strain re-measurement. New experimental data cause the burden level in strain family HMS to match the one observed in strain family  $\Delta$ ptsG. Consequently, one can generalize the effect of XylE expression and increased % Carbon Xylose on burden level as follows: *Strain* Empty  $<$  *Strain* p1x  $<$  *Strain* p3x, where *Strain* refers to either HMS or  $\Delta$ ptsG. Key aspects learned from Figure 6.6 can be summarized into two points:

1. *Sugar co-utilization increases extent of metabolic burden.* This is evidenced by a

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

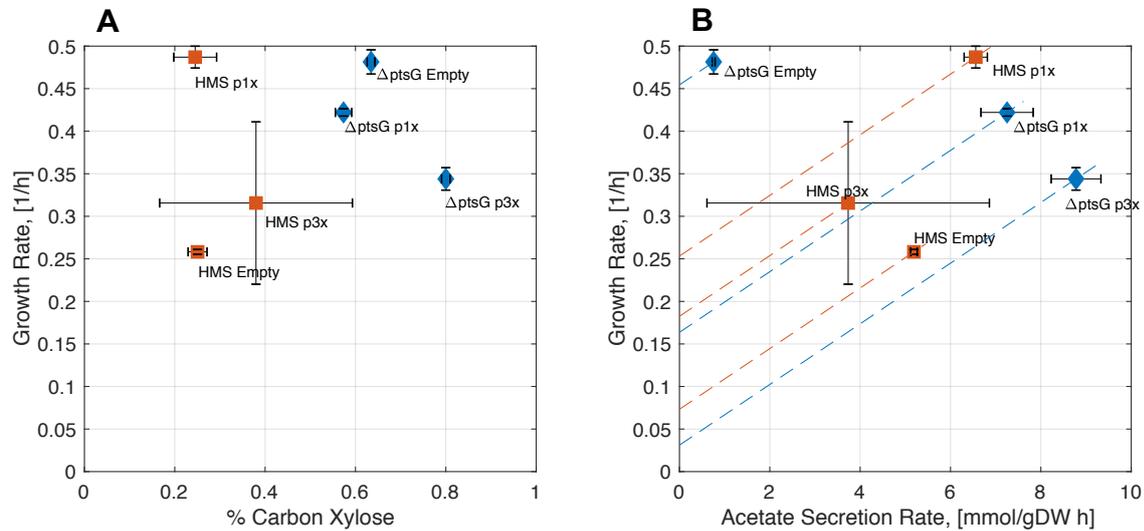


Figure 6.5.: **Effect of Glucose and Xylose Co-utilization on Growth Rate and Its Associated Metabolic Burden.** The detrimental effect of increasing glucose and xylose co-utilization levels on growth rate can be visualized by means of % Carbon Xylose (A.). By plotting growth rate against acetate secretion rates (B.) the metabolic burden associated with an increased sugar co-utilization can be assessed.

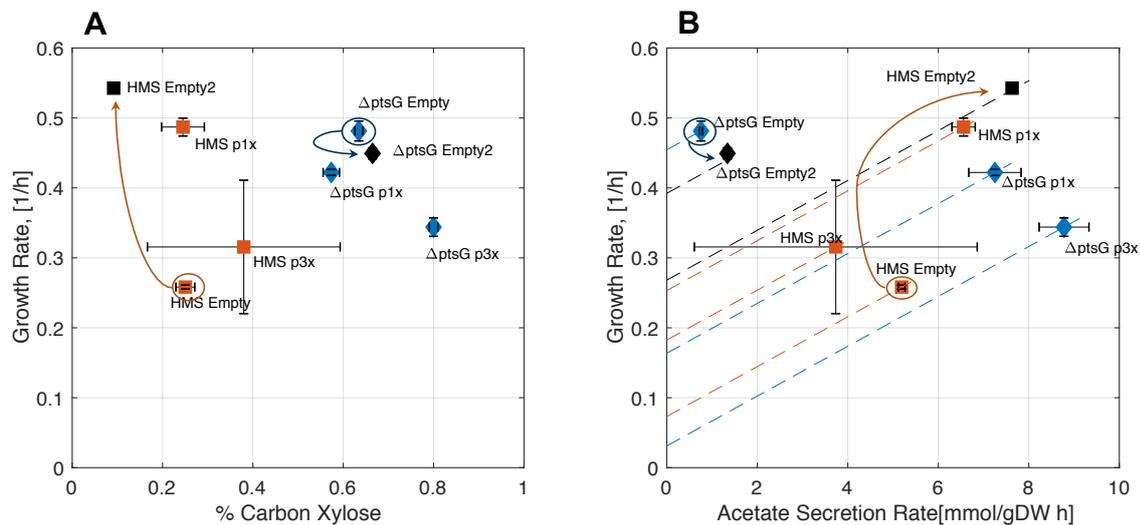


Figure 6.6.: **Effect of Glucose and Xylose Co-utilization on Growth Rate and Its Associated Metabolic Burden. Re-measurement of Selected Strains** Experimental characterization of parent strains HMS Empty and  $\Delta$ ptsG Empty is repeated and represented by the black symbols labeled as HMS Empty2 and  $\Delta$ ptsG Empty2, respectively. Note that the denomination “Empty2” emphasizes that these data points represent experimental repetitions.

downwards displacement of the acetate line for strain HMS p3x and  $\Delta$ ptsG p3x, relative to the acetate line of the corresponding strain expressing the XylE transporter with one single mutation. A tentative explanation for this observation is that xylose metabolism should be associated with an increased expression of protein-expensive pathways.

2. *Strain family  $\Delta$ ptsG does not seem to allow for robust plasmid-based protein expression.* This is evidenced by the different extent in the response to protein expression observed for the two different strain families. Consider for example the expression of XylE<sub>1x</sub>. While the acetate lines for HMS Empty2 and HMS p1x are almost identical, acetate lines for  $\Delta$ ptsG Empty2 and  $\Delta$ ptsG p1x are clearly differentiated, suggesting that the extent of metabolic burden caused by XylE<sub>1x</sub> expression is much higher in the strain family  $\Delta$ ptsG than in the family HMS. Similar trends are observed for the expression of XylE<sub>3x</sub>. Additionally, Figure 5.2 provides further experimental evidence to this observation.

### 6.3. Assessing Taxadiene Production Potential

Does an increased degree of sugar co-utilization have a beneficial effect on taxadiene production? Which strain has the highest taxadiene production potential? Here, we apply both a constraint- and a kinetic-based approach to help answer these questions.

The calculation of maximal theoretical product yields by means of linear programming represents the simplest approach to assess the production capabilities of a given metabolic network under consideration of a given set of constraints. Here, we use the *E. coli* genome-scale metabolic reconstruction iJO1366 along with experimentally determined total carbon and estimated oxygen uptake rate to highlight the effect of increasing degrees of % Carbon Xylose on maximal theoretical taxadiene yields. As shown in Figure 6.7, calculations are done assuming two different growth scenarios. In the first case, a growth rate of zero is assumed (Figure 6.7.A), while in the second case, the same reaction is constrained to have an arbitrary value of 0.2 1/h for all % Carbon Xylose values (6.7.B). Regarding the oxygen uptake rate, it is assumed to be independent from % Carbon Xylose values and the genetic background (refer to case study II, Section 6.2.1). Additionally, a constant value of 7 mmol O<sub>2</sub>/gDW h was assumed in order to ensure the existence of a flux distribution fulfilling all constraints over the whole range of % Carbon Xylose.

Since *E. coli* does not naturally produce taxadiene, the stoichiometric matrix of model iJO1366 is correspondingly expanded to allow for taxadiene biosynthesis from IPP and DMAPP as pathway precursors (refer to Appendix B.1.1). By maximizing the flux through the taxadiene export reaction, under consideration of constraints on growth rate and of glucose, xylose and oxygen uptake rates, one can assess the maximal production capability of the network. Usually, the production capability is expressed in terms of product yield. In this case, since two carbon sources are considered, product yield is expressed on a carbon basis:

$$\text{Taxadiene Carbon Yield} = \frac{20 r_{tdn}}{5 r_{xylose} + 6 r_{glucose}},$$

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

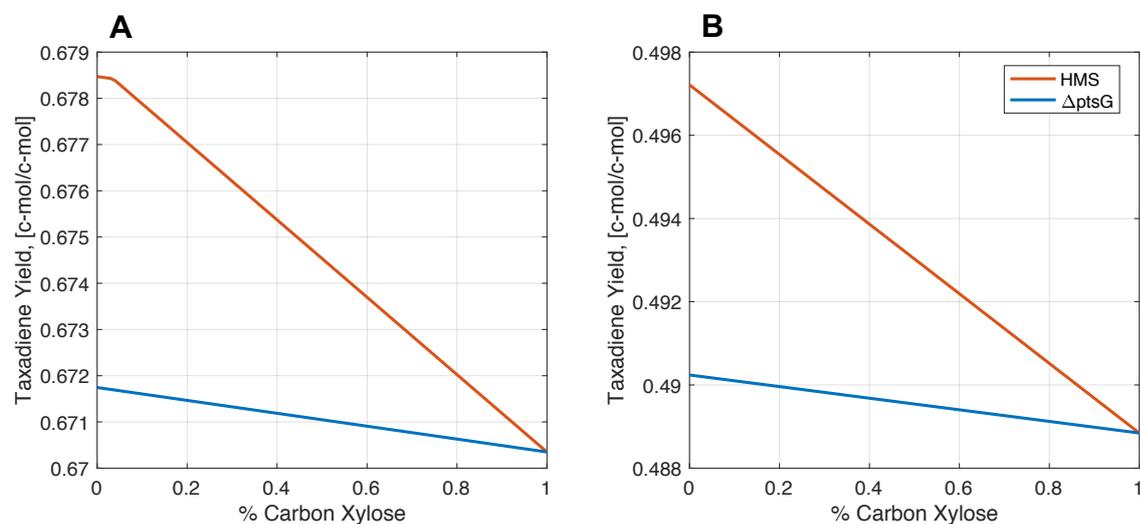


Figure 6.7.: **Assessing Taxadiene Production Potential by a Constraint-based Approach.** The taxadiene production capability of the genome-scale metabolic reconstruction iJO1366 was assessed by means of linear programming assuming a constant growth rate ( $\mu$ ) of zero (**A.**) and of 0.2 1/h (**B.**). For all calculations, the upper boundary for oxygen uptake was set to 7 mmol/gDW h. In line with experimental evidence, a constant, combined carbon flux of 45 mmol C/gDW h was assumed. We observe a reduction in the maximal theoretical taxadiene yield as % Carbon Xylose increases from 0 to 1, as follows: When  $\mu=0$  (**A.**); taxadiene yield exhibits a maximal reduction of 1.2% for strain family HMS and of 0.208% for strain family  $\Delta$ ptsG. Similarly, when  $\mu=0.2$  1/h (**B.**), taxadiene yield exhibits a maximal reduction of 1.68% for strain family HMS and of 0.28% for strain family  $\Delta$ ptsG.

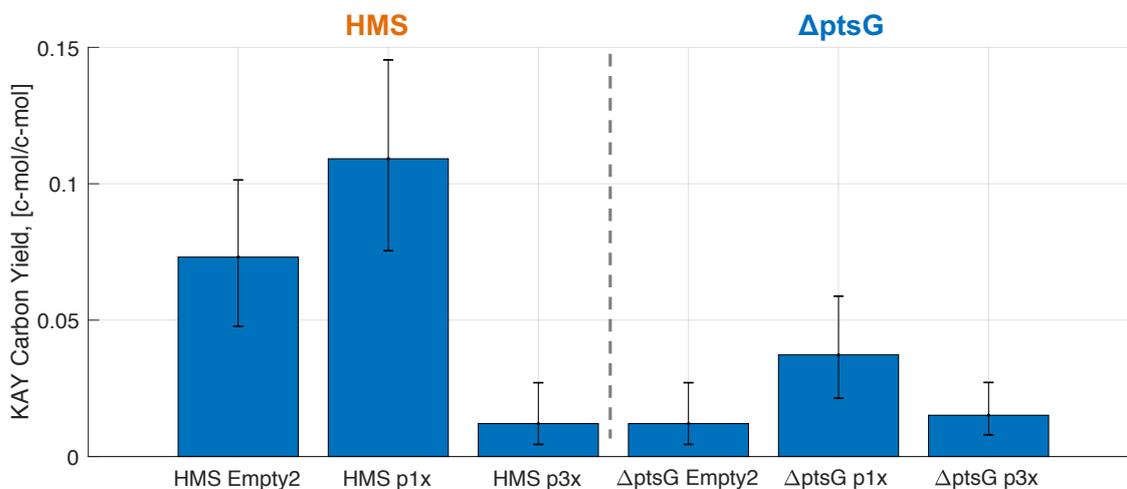


Figure 6.8.: **Assessing Taxadiene Production Potential by a Kinetic-based Approach.** Taxadiene production potential was estimated as suggested in [71]. Lumped reaction representing taxadiene synthesis is given in the main text. A total of 10.000 models were constructed based on  $^{13}\text{C}$ -labeling data and experimentally determined exchange reactions. A modified version of the *E. coli* core model was used as input for the ensemble modeling approach. Given KAY values correspond to the median. Note that error bars represent 25th and 75th percentiles of KAY values within the ensemble.

where  $r_{tdn}$  corresponds to the maximal taxadiene production rate and  $r_{xylose}$  and  $r_{glucose}$  to xylose and glucose uptake rate, respectively. Multiplication by 20, 5 and 6 is used to transform respective rates from units mmol/gDW h into c-mmol/gDW h. Alternatively, rates can be transformed into g/gDW h by multiplication with respective molecular weights. Thus, obtained product yields would be on a mass basis.

Figure 6.7 reveals that for both strain families and the two growth regimes considered, increasing levels of xylose and glucose co-utilization reduce the potential of the *E. coli* metabolic network to produce taxadiene. However, the maximal reduction observed corresponds to just 1.69 %. Additionally, and for % Carbon Xylose values smaller than 1, our constraint-based approach predicts a higher taxadiene production potential for the strain family HMS. Finally, increasing growth rates at constant substrate uptake rate reduces the maximal taxadiene yield that can be reached by the network. This last observation can be rationalized from a mass balance perspective. As the growth rate increases at a constant substrate uptake level, the amount of carbon atoms drawn from the cellular metabolism to support biomass synthesis is incremented. Consequently, less carbon atoms are left that can be potentially used for taxadiene.

Actual product yields of strains in early optimization stages are usually low. Therefore, maximal theoretical product yields seldom allow an accurate assessment of production performance.

It has been recently shown that the dynamic stability of a given metabolic network can impose a further constraint on the maximal product yield that can be achieved by the network. Lafontaine Rivera et al. [71] termed this yield “Kinetically Accessible Yield (KAY)”. In order to estimate KAY for a given molecule, the authors calculated flux values through a lumped reaction expressing the synthesis of the target molecule in one single reaction. Sub-

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

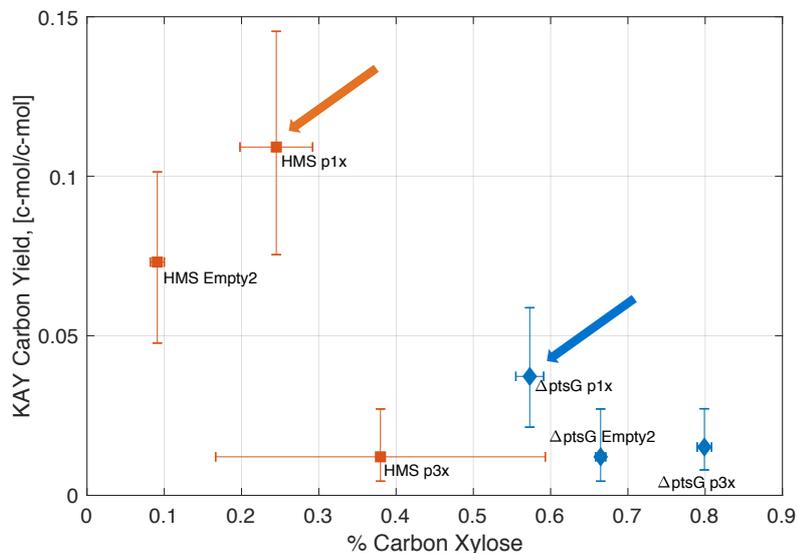


Figure 6.9.: **Dependence of Taxadiene Production Potential on % Carbon Xylose.**

Production potential calculated by a kinetic-based approach is plotted against % Carbon Xylose. Orange and blue arrows highlight the strain exhibiting the best production potential for strain family HMS and  $\Delta$ ptsG, respectively. Note that while vertical error bars represent the 25th and 75th percentiles of KAY values within an ensemble of 10,000 models, i.e. they are *in silico* generated values, horizontal error bars represent the standard deviation of the experimental data.

sequently, product yields are calculated by dividing the flux through the lumped reaction by the corresponding substrate uptake rates. KAY equals to the maximal product yield that can be achieved before the network loses stability. The system is considered stable as long as the real part of eigenvalues of the Jacobian matrix is negative. Since a kinetic model describing cellular metabolism is needed for KAY calculations and, kinetic parameters associated with the particular kinetic model are typically unknown, Lafontaine Rivera et al. calculated KAY for an *ensemble* of kinetic models, rather than for one single model. The ensemble of kinetic models was constructed by sampling kinetic parameters so that the obtained models fulfill a given reference flux distribution. This methodology was applied to accurately predict experimental isobutanol yields in three different *E. coli* strains. Refer to Figure 5.B in [71] and to Section 2.3.1 for more details on the mathematical and computational implementation of the KAY approach.

Here, we used available  $^{13}\text{C}$  labeling data and exchange rates listed in Table 6.2 to generate a reference flux distribution for each strain using the software 13CFLUX2 [155] along with the *E. coli* metabolic model iRL2013 [75]. We extended the iRL2013 model to consider xylose uptake. Then, we estimated the production potential of each strain by calculating its kinetically accessible taxadiene yield by maximizing the flux through the following lumped reaction:



Results are summarized in Figure 6.8. Interestingly, slightly higher product yields are predicted for the strain family HMS, in accordance with our constraint-based estimation of production potential (see Figure 6.7). With a kinetically accessible yield of 0.109 c-

### 6.3. Assessing Taxadiene Production Potential

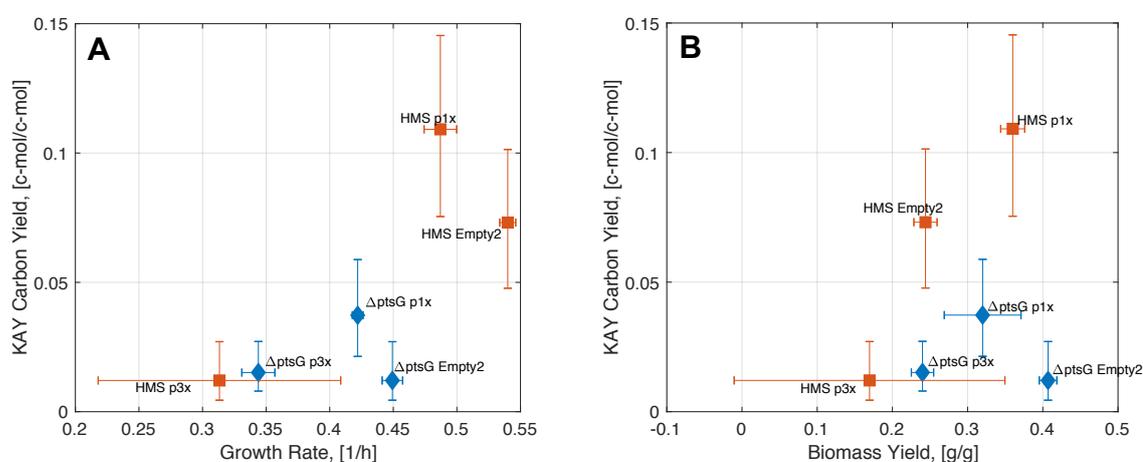


Figure 6.10.: **Dependence of Taxadiene Production Potential on Growth Characteristics.** Production Potential is plotted against growth rate (A.) and biomass yield (B.). In both cases, the existence of a particular value for growth rate and biomass yield maximizing the production potential is suggested by the data. Note that while vertical error bars represent the 25th and 75th percentiles of KAY values within an ensemble of 10.000 models, i.e. they are *in silico* generated values, horizontal error bars represent the standard deviation of the experimental data.

mol/c-mol, strain HMS p1x is predicted to have the highest taxadiene production potential from all considered strains. Analogously and with a kinetically accessible yield of 0.037 c-mol/c-mol, strain ΔptsG p1x is predicted to exhibit the highest production potential within strain family ΔptsG. In order to assess the effect xylose and glucose co-utilization has on production potential, calculated KAYs are plotted against respective % Carbon Xylose values, as shown in Figure 6.9. Orange and blue arrow highlight strains HMS p1x and ΔptsG p1x respectively. For strain family HMS, KAY increases from 0.07 c-mol/c-mol (for strain HMS Empty2) to 0.109 c-mol/c-mol (for strain HMS p1x) as % Carbon Xylose increases from 0.09 to 0.24 to then decreases to 0.012 (for strain HMS p3x) as % Carbon Xylose keeps increasing to reach 0.38. As for strain family ΔptsG, KAY exhibits its maximum value of 0.037 c-mol/c-mol (for strain ΔptsG p1x) at a % Carbon Xylose value of 0.57 and decreases as % Carbon Xylose increases. We conclude from Figure 6.9 that for each strain family, an optimal value of % Carbon Xylose exists. Deviations from this optimal point negatively affect the taxadiene production potential.

From Figure 6.2 we learned that increasing values of % Carbon Xylose led to a reduction in growth rate and biomass yield. Since both of these variables are better suited to globally describe the metabolic status of the cell, we analyze their effect on KAY, as shown in Figure 6.10. Growth rate and biomass yield seem to affect KAY in a similar way. In both cases, the data seem to suggest the existence of particular values for growth rate and biomass yield that maximize KAY. Remarkably, optimal growth rate (or biomass yield) for strain family HMS seems to coincide with optimal growth rate (or biomass yield) for strain family ΔptsG.

## 6.4. Model-based Identification of Optimal Production Strain Candidates

So far, we have independently

- analyzed the effect of XylE mutants expression in two different strain families on the degree of sugar co-utilization using the variable % Carbon Xylose;
- characterized the effect of increasing levels of % Carbon Xylose on metabolic burden using the concept of acetate lines and,
- assessed the taxadiene production potential for each strain using a novel kinetic-based approach.

Now, we step forward to identify optimal production strain candidates considering these three factors at the same time. As suggested by Figure 6.11, strain HMS p1x does not only have the highest taxadiene production potential out of all strains, but it also exhibits a moderate degree of sugar co-utilization and metabolic burden. In order to allow for taxadiene production, strain HMS p1x must be further engineered by the introduction of genes by means of plasmids or genomic integration. As a consequence of this additional load, we expect a reduction in the growth rate of strain HMS p1x. As suggested by Figure 6.11.C, a reduction in its growth rate would even increase its KAY value for taxadiene, since we expect the maximal KAY value to lie in the orange zone. If a strain with a higher degree of sugar co-utilization is desired, strain  $\Delta$ ptsG p1x can be used as parent strain for further engineering rounds. Similarly to strain HMS p1x, we expect a reduction in the growth rate of this strain due to an increased load. Assuming that the maximal KAY value for this strain family lies in the blue zone, a reduction in its growth rate would lead to an increase in the taxadiene production potential too. Note that the positive effect of additional load and subsequent growth rate reduction does not always lead to an increased KAY value. For instance, KAY values for strains  $\Delta$ ptsG p3x and HMS p3x would probably decrease due to the reduction in their growth rates, as suggested by Figure 6.11.C.

## 6.5. Discussion

Xylose and glucose are the two most abundant sugars present in lignocellulosic biomass, a feedstock commonly regarded as renewable [40,52,100]. Unlike glucose-rich substrates such as corn, wheat or sugar cane, plant biomass can be provided in large quantities without ethical concerns related to food prices and land use. Efficient conversion of lignocellulosic hydrolysates requires simultaneous xylose and glucose utilization. However, in most bacteria, carbon sources are used in a hierarchical manner. For instance, *E. coli* preferably metabolizes glucose over xylose. This carbon source preference arises from different regulatory mechanisms, commonly referred to as Carbon Catabolite Repression (CCR). CCR can occur at many levels, including transcription activation and repression, control of translation and, inhibition of enzyme activity [49]. Many strategies have been developed to deactivate CCR to allow for simultaneous sugar utilization in *E. coli*, including deactivation of the phosphotransferase system by knocking out the genes *ptsH*, *ptsI* and *crr* [44], mutation of the gene *ptsG* [100], mutation of the global regulatory protein *crp* [163] and

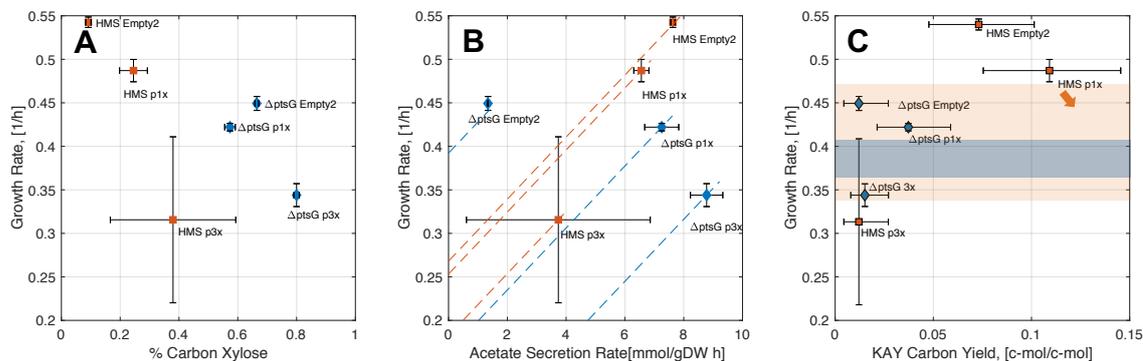


Figure 6.11.: **Model-based Identification of Optima Production Strain Candidates.** Three different criteria, namely sugar co-utilization degree (A.), associated metabolic burden (B.) and production potential (C.) are applied to identify optimal production strain candidates. Error bars in Figures A. and B. represent one standard deviation of the data. In Figure C., vertical error bars represent one standard deviation of the experimental data, while horizontal values represent 25th and 75th percentiles of KAY values within an ensemble of 10.000 models.

deletions of some metabolic genes [45,165]. In *S. cerevisiae*, another biotechnologically relevant microorganism, the mutation of a XylE transporter lead to simultaneous xylose and glucose utilization. All known xylose transporters are competitively inhibited by glucose. Since the mutant XylE transporter exhibited no inhibition by glucose [40], xylose could be transported into the cell by the mutant XylE in presence of glucose.

Our cooperation partner Katarina Kemper at **IBK** combined two strategies, namely knocking out the gene *ptsG* and novel mutations in the xylose specific transporter XylE, to achieve different levels of sugar co-utilization in six *E. coli* strains. We quantified the extent of glucose and xylose co-utilization in those strains by means of a variable we termed % Carbon Xylose. As suggested by the experimental data (Figure 6.1) and confirmed by a constraint-based approach (Figure 6.2 and 6.3), increased % Carbon Xylose levels lead to a reduction in both the observed growth rate and biomass yield. Irregardless of the level of sugar co-utilization, the combined carbon uptake remained constant under the experimental conditions tested, at a level of 45 c-mmol/gDW h. This constant total carbon uptake rate, along with the lower quality of xylose (refer to Figure 6.12) offer an explanation for the decreased growth rates and biomass yields experimentally observed as % Carbon Xylose values increases. Moreover, experimental data suggest a decrease in the oxygen uptake rate as sugar co-utilization increases (refer to Figure A.8 in Appendix A.3.3), which as shown in Figure 6.12, should decrease even more the quality of xylose as substrate, expressed by its specific energy content.

For our strains, increased levels of xylose and glucose co-utilization come at the cost of an increased metabolic burden. This is evidenced in Figure 6.11.B by the downwards shift of the acetate line of a given strain, relative to its parent strain, as % Carbon Xylose increased. Interestingly, strain family  $\Delta$ ptsG, showing overall a higher extent of % Carbon Xylose, seemed to exhibit a higher protein content than strain family HMS, presumably necessary to support high levels of sugar co-utilization (refer to Figure A.9 in Appendix A.3.4 for preliminary total protein content measurements supporting this hypothesis). As a consequence, strains derived from the parent strain  $\Delta$ ptsG Empty showed a stronger

## 6. Simultaneous Utilization of D-Xylose & Glucose in *E. coli*

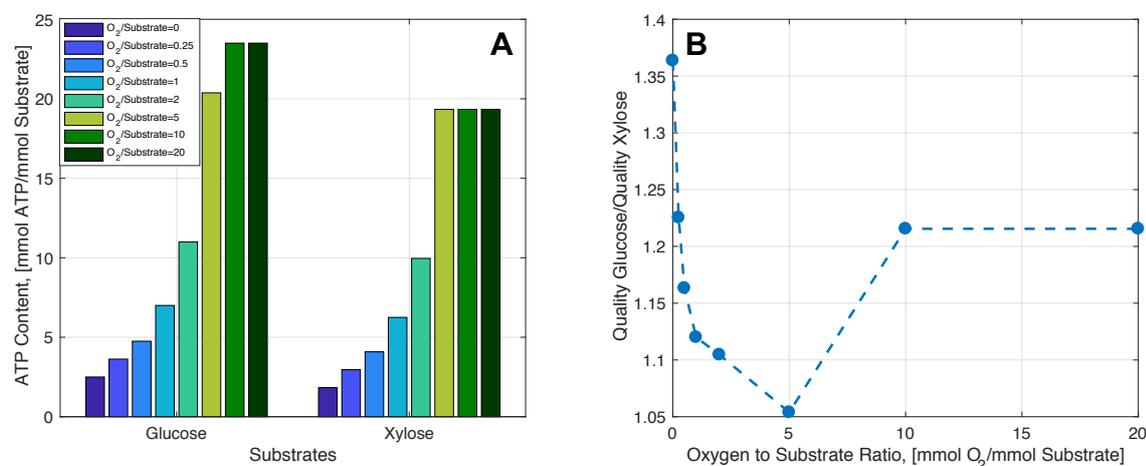


Figure 6.12.: **Substrate Quality: Glucose vs Xylose for Different Oxygen to Substrate Ratios.** Quality of glucose and xylose as substrates was assessed using energy content as quality criteria for different oxygen to substrate ratios (**A.**). The energy content ratio between glucose and xylose (**B.**) shows that glucose exhibits a higher substrate quality. However, this advantage over xylose greatly depends on the oxygen to substrate ratio. All calculations were performed using the genome-scale metabolic reconstruction iJO1366, as explained in Section 4.2.1.

downwards shift in its acetate lines, when compared to the shifts observed within the strain family HMS (see Figure 6.11.B). Since the introduction of further genetic elements into the strains and their expression is necessary to yield a strain capable of producing taxadiene, strain response to an additional load should be considered for the selection of optimal production strain candidates.

In order to accurately select optimal production strain candidates for further engineering rounds and thus to contribute the reduction of *in vivo* experimentation, we applied both a constraint- and a kinetic-based approach to quantitatively assess the taxadiene production potential of all strains. The constraint-based approach indicated a linear reduction in the theoretical maximal taxadiene yield as % Carbon Xylose increased for both strain families. Additionally, strain family HMS exhibited a superior performance when % Carbon Xylose < 1. Considering these predictions, the strain HMS Empty should be further engineered into a production strain, as it exhibits the lowest % Carbon Xylose out of all strains. On the other hand, the kinetic-based approach suggested a more complex relationship between the kinetically accessible taxadiene yield and % Carbon Xylose. This method predicted an overall higher taxadiene production potential for the strain family HMS, just as the constraint-based approach did. However, unlike the constraint-based approach, the kinetic based approach suggested the existence of a maximal taxadiene yield for each strain family, located at % Carbon Xylose values different from zero. Remarkably, the kinetic-based approach suggested that optimal taxadiene production was only indirectly defined by % Carbon Xylose. The true variable determining the kinetically accessible yield for taxadiene was the growth rate or the biomass yield. Experimental evidence supporting this observation was provided by our cooperation partner Lars Janoschek at **STT**; refer to Figure 9.3.A in Chapter 9. Based on three aspects, namely extent of sugar co-utilization, its associated metabolic burden and taxadiene production potential, we suggest strain HMS plx for further engineering rounds towards the construction of an optimal production

strain. Since a further load, exerted by additional genetic elements introduced to allow taxadiene synthesis, is expected to reduce growth rate and biomass yield, we expect the experimental taxadiene yield to surpass the predicted KAY value for this strain, as shown in Figure 6.11.C. To the best of our knowledge, this work represents the first reported effort to guide the selection of strain candidates potentially exhibiting high production performance while considering concepts of metabolic burden and kinetically accessible yields.

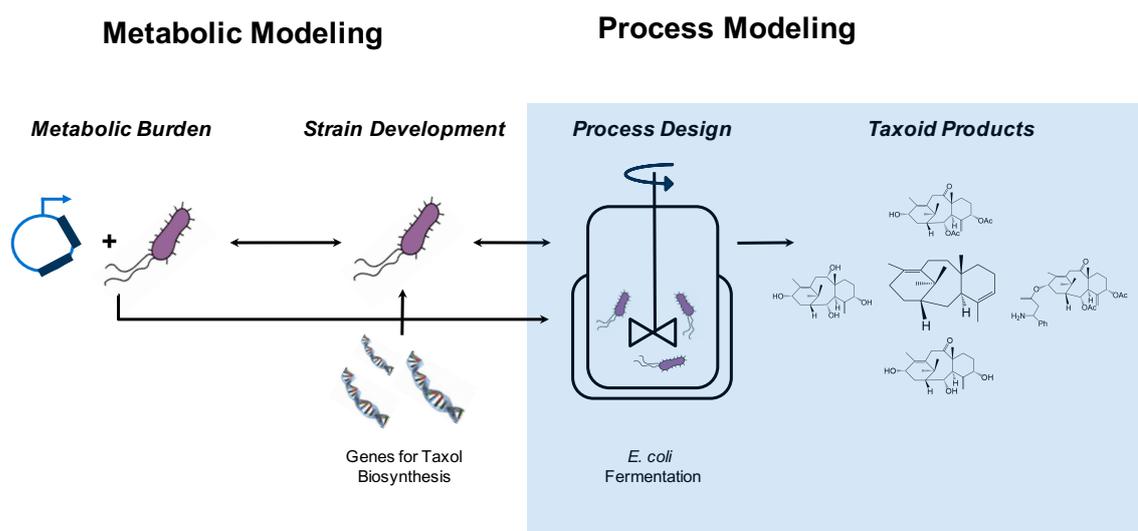
At the moment of the submission of this work, experimental implementation of this model-driven suggestion was still pending.



Part III.

Results: Process Optimization





The third part of this work is composed of three chapters. In **Chapter 7** we address some of the questions arising when designing a production process: what is the best substrate (carbon source) for the production of a given target molecule? Which performance criteria should be defined to identify optimal carbon source? If two different metabolic pathways can be used for production, which one should be preferred? Does the pathway performance depend on the selected carbon source? Then, in **Chapter 8**, experimental data (growth, substrate uptake and production rates) of a semi-batch process for four different temperatures and time points of induction were analyzed. The aim was to develop a model to estimate optimal values for these two process parameters. Finally, in **Chapter 9**, two theoretical approaches for product yield prediction are compared: a constraint-based and a kinetic-based approach. Traditionally, production potential of metabolic networks has been assessed by constraint-based methods. By applying linear programming, it is possible to calculate maximal theoretical yield on a defined carbon source and on a given metabolic network. However, theoretical maximal yield is not an adequate proxy for *experimental* product yield, especially when it comes to describing the production performance of strains in early development stages. By contrast, a novel kinetic-based approach showed promising results, even though numerical values of all kinetic parameters were unknown.



## 7. Model-based Medium Optimization

Typically, the following questions are to be addressed in early design stages of a given biotechnological production process: what is the best substrate for the production of the target molecule? Which performance criteria should be applied to identify the optimal carbon source? If two different metabolic pathways can be used for production, which one should be preferred? Does pathway performance depend on the selected carbon source? Here, we apply constraint-based modeling to answer these questions. First, we defined two different performance criteria to assess the performance of various substrates, namely the maximal profit and the maximal theoretical yield. We calculate theoretical yields using three different bases: a carbon, a molar and a mass basis. Then, we used these criteria to rank 180 different growth-supporting carbon sources. The underlying calculation is the potential of the substrate to be transformed into taxadiene by the *E. coli* metabolic network. Three-substrate mixtures were also considered in the calculations. Additionally, aerobic and anaerobic optimal network performance was compared. Since taxadiene production via the non-mevalonate pathway requires energy in the form of ATP and CTP, our model-driven analysis suggests an aerobic production process. Surprisingly, no single carbon source tested allowed a complete conversion into taxadiene, i.e., all substrates exhibited a maximal theoretical yield lower than 1 c-mol taxadiene/c-mol substrate. The largest maximal theoretical yield was obtained for ethanol as single carbon source, with a value of 0.92 c-mol taxadiene/c-mol ethanol. By contrast, glucose exhibited a maximal carbon yield of 0.65 c-mol taxadiene/c-mol glucose. An analysis of optimal flux distributions identified CO<sub>2</sub> as the main source for carbon loss. Finally, we found out that the endogenous non-mevalonate pathway allows for a higher theoretical yield than the heterologous mevalonate pathway for glucose as the carbon source. However, this is not true for all 180 carbon sources analyzed. When both pathways were active in an optimal ratio, we observed a small degree of synergy (<6%) for some substrates, including glucose.

### 7.1. Performance Criteria for Substrate Selection

Before one can actually select the best substrate from a list of candidates, performance criteria need to be defined. For biotechnological production processes, *product yield* has been traditionally used to assess and to compare the production performance of different substrates and strains. Product yield is defined as the ratio of product gained to substrate consumed. Depending on the units used to calculate the ratio, one can express product yield on three different bases: on a molar basis (mol product/mol substrate), on a carbon basis (c-mol product/c-mol substrate) or on a mass basis (g product/g substrate). When assessing the economic performance of a given set of substrates, profit can be used as a performance criterion. We define profit as the ratio of US dollars gained to US dollars invested.

## 7. Model-based Medium Optimization

The choice of a particular performance criterion to identify the best carbon source depends on the aim of the optimization process. Thus, if the process designer is interested in maximizing revenue, the performance criterion to be applied seems obvious. On the other hand, if network efficiency is to be optimized, product yield on a carbon or mass basis should be preferred over product yield on a molar basis. This is because maximal values of molar yield are not constrained to a defined interval, but rather depend on the substrate used. This leads to difficulties while assessing a given product yield on a molar basis. For instance, is a product yield of 0.1 mol taxadiene/mol glucose good enough? By contrast, values for product yield expressed on a carbon or mass basis will always range from 0 to 1. Substrates exhibiting a product yield of 1 c-mol product/c-mol substrate allow for a *complete* conversion of substrate into product. In other words, the network performance is already optimal and can not be further improved.

### 7.1.1. Single Substrates

We start our constraint-based analysis by assessing 180 different growth-supporting substrates, based on the performance criteria previously described. The *E. coli* metabolic reconstruction iJO1366 [106] was used for all calculations. *Growth-supporting* refers to the ability of the iJO1366 network to *in silico* grow on a given substrate. In order to calculate maximal theoretical yields representing true upper boundaries, both the Non-growth Associated Maintenance (NGAM) term and the cellular growth rate were set to 0. Additionally, oxygen uptake rate was set to an artificially high value of 100 mmol/gDW h to ensure sufficient oxygen availability. Since the *E. coli* metabolism does not allow for an endogenous taxadiene production, the iJO166 model was extended by adding necessary reactions allowing taxadiene production as shown in Appendix B.1.1. Following steps were applied to calculate maximal profit and taxadiene yields using the endogenous non-mevalonate pathway for precursor production.

1. Choose one of the 180 substrates listed in Table B.1 (see Appendix B.1.2).
2. Set a rate of 1 mmol substrate/gDW h by adjusting the lower and upper bound of the respective uptake reaction.
3. Maximize flux through the taxadiene export reaction using linear programming.
4. The numerical value of the maximal flux calculated in previous step coincides with the taxadiene yield on a molar basis, i.e.  $\text{mol } txd / \text{mol substrate}$ . This is because substrate uptake rate was set to 1 mmol/gDW h.  $txd$  refers to taxadiene.
5. Calculate product yield on a *carbon* basis by multiplying yield obtained in step 4 by the factor:  $\left[ \frac{\frac{20 \text{ c-mol } txd}{1 \text{ mol } txd}}{\frac{x \text{ c-mol substrate}}{1 \text{ mol substrate}}} \right]$ ;  $x \text{ c-mol}$  refers to the number of carbon atoms in the substrate chosen in step 1.
6. Calculate product yield on a mass basis by multiplying yield obtained in step 4 by the factor:  $\left[ \frac{\frac{272.48 \text{ g } txd}{1 \text{ mol } txd}}{\frac{x \text{ g substrate}}{1 \text{ mol substrate}}} \right]$ ;  $x \text{ g substrate}$  refers to the molecular weight of the substrate chosen in step 1.

## 7.1. Performance Criteria for Substrate Selection

7. Calculate profit by multiplying yield obtained in step 4 by the factor:  $\left[ \frac{128056.2 \text{ US\$ txd}}{x \frac{1 \text{ mol txd}}{x \text{ US\$ substrate}}} \right]$ ;  $x \text{ US\$ substrate}$  refers to current market price of the substrate chosen in step 1 per mol.

Carbon content, molecular weight and market price for all carbon sources can be found in Table B.2 in the Appendix. Figure 7.1 lists the substrate having the highest performance for each criterion and glucose and glycerol along with their respective position in the ranking. Refer to Table B.1 in the Appendix for a complete overview. From Figure 7.1 we can easily conclude that there is no global optimal substrate exhibiting the highest value for all performance criteria. Rather, optimal substrate differs according to the metrics used. In the case of glucose and glycerol, it is interesting to observe that while glucose seems to be superior when considering product yield on a molar basis, the trend is reversed when product yield is considered on either a carbon or mass basis.

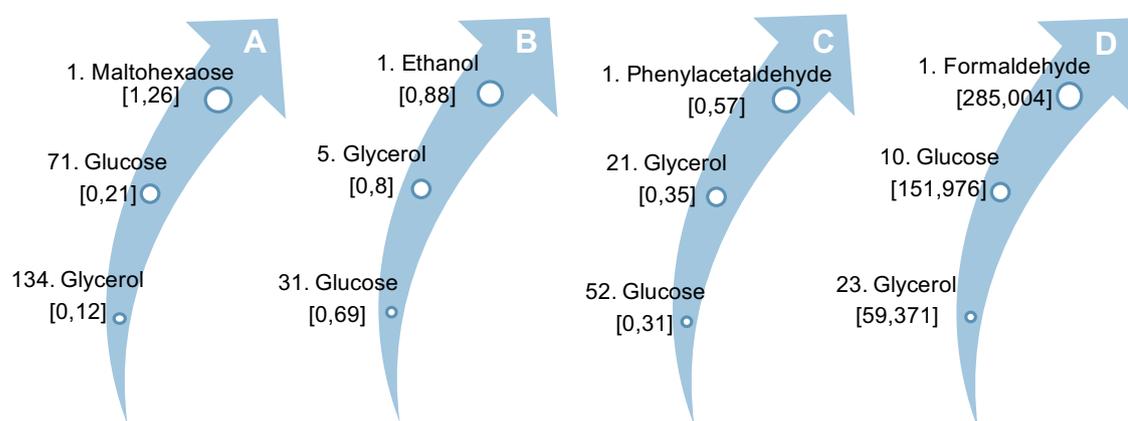


Figure 7.1.: **Substrate Ranking via Four Different Performance Criteria** Performance criteria considered are profit in units of US\$ txd/US\$ substrate (D.) and taxadiene yield on a molar (A.), carbon (B.) and mass basis (C.). For each criterion, the substrate exhibiting the highest performance is shown. Performance and relative position in the ranking for glucose and glycerol is also provided. Profit and yield values are shown in the square brackets. The model used was a modified version of iJO1366. Substrate uptake rate was set to 1 mmol substrate/gDW h, oxygen uptake rate to 100 mmol/gDW h. Growth rate and NGAM were set to zero.

Another interesting observation is the fact that no single substrate can support a complete conversion into taxadiene. Ethanol allows with 0.88 c-mol txd/c-mol ethanol the highest conversion efficiency. This value means that in the best case, 12% of all carbon atoms absorbed by the network are lost in the form of by-products. At this point, it is important to state that these values only represent the *potential* of the metabolic network to produce taxadiene, without guarantee of biological feasibility. Due to regulatory or toxic effects triggered by the substrate or pathway intermediates, which are not considered in maximal theoretical yield calculations, we expect actual product yields to be significantly lower. Nevertheless, theoretical yield calculations are useful to unravel the potential of a given metabolic network.

## 7. Model-based Medium Optimization

Table 7.1.: **Metabolic Models Used for Synergy Studies.** All three models are based on the *E. coli* metabolic reconstruction iJO1366. All models were extended to allow for taxadiene production. IPP and DMAPP are referred to as pathway precursors for taxadiene synthesis.

Characteristics	
Model 1	Mevalonate pathway is used for precursor biosynthesis
Model 2	Non-mevalonate pathway is used for precursor biosynthesis
Model 3	Both mevalonate and non-mevalonate pathways are used for precursor biosynthesis

### 7.2. Synergy as Design Principle for Pathway Selection

Motivated by the impossibility to achieve a complete conversion of any carbon source into taxadiene, we ask following question: can the simultaneous activity of both the mevalonate (MVA) and non-mevalonate (MEP) pathway improve substrate conversion? We introduce our approach to answer this question using glucose as the case study. Let us start by considering three different metabolic models, as shown in Table 7.1. Then, let us calculate the maximal theoretical yield for each of the three models. By comparing theoretical product yields supported by each model, it is possible to identify synergistic effects arising due to the simultaneous utilization of both pathways. Synergy exists, if the theoretical product yield for Model 3 is higher than the theoretical product yield for Model 1 and Model 2.

We quantify synergy as follows:

$$Synergy = \frac{\max(Y_1, Y_2, Y_3) - \max(Y_1, Y_2)}{\max(Y_1, Y_2)},$$

where  $Y_i$  refers to product yield of one of the three models introduced in Table 7.1. Note that the basis used to express product yield (mol, mass or carbon) do not have an effect on the numerical value of synergy. It is possible to identify optimal pathway usage originating the observed synergy via constraint-based modeling, as shown in Figure 7.2. For that, a new constraint, determining the fraction of carbon flux going through each of the two pathways (mevalonate or non-mevalonate pathway), is set on the metabolic model. More specifically, we modulate the activity of the MEP pathway by constraining the flux through its first reaction, termed DXS in Figure 5.9 and DXPS in model iJO1366. By calculating maximal taxadiene yields for various MEP activity levels, ranging from 0 to 1, it is possible to identify an optimal flux fraction ( $MEP^{opt}$ ), as shown in Figure 7.2. Note that  $MEP^{opt}$  will not only depend on the identity of the substrate used, but also on the production conditions. For glucose, it increases from 75% to 91% if the conditions are changed from fully aerobic to fully anaerobic. This observation suggests that the performance of the MEP pathway is less dependent on the value of oxygen uptake than the MVA pathway. In fact, while the maximal taxadiene yield achieved via the MEP pathway is reduced by roughly 15% as a consequence of oxygen deprivation, the maximal taxadiene yield achieved via the MVA pathway drops by 55% (refer to downwards displacement of points ① and ③ in Figure 7.2.A and .B). In order to analyze the effect of growth rate on maximal taxadiene yields, synergy and  $MEP^{opt}$ , we consider a growth rate of 0.2 1/h along with a glucose and oxygen uptake rate of 10 mmol/gDW h. As shown in Figure 7.2.C, cellular growth reduces the maximal taxadiene reached by the *E. coli* network. This results obvious from a mass

## 7.2. Synergy as Design Principle for Pathway Selection

balance perspective. Since growth and taxadiene production compete for a limited supply of carbon atoms, an increase in the growth rate will necessarily lead to a reduction in the maximal taxadiene yield. Interestingly,  $MEP^{opt}$  decreases from 75% to 70% if growth is considered (compare Figure 7.2.C and .A). This trend continues if growth is increased to 0.4 1/h. In this case,  $MEP^{opt}$  amounts to 56%, suggesting that taxadiene production via MVA pathway is less affected by growth than the MEP pathway (results not shown). However, at a growth rate of 0.4 1/h, the synergy between the MEP and MVA pathway only reaches 0.4 %.

Now, we extend synergy calculations to all carbon sources considering fully aerobic conditions and setting NGAM and growth rate to 0. Figure 7.3 summarizes the synergy observed for 180 substrates (refer to Table B.1 in the Appendix for a complete overview of the results). We observe that the simultaneous activity of the MEP and MVA pathway does not lead to an increased taxadiene yield for all substrates, i.e. synergy amounts to 0%. In total, 57 carbon sources belongs to this group, almost the half of which (28 substrates) exhibit higher taxadiene yield using the MEP pathway. With only 14 substrates having a synergy higher than 2.5% but lower than 5.7%, we conclude that simultaneous activity of both the MEP and MVA pathway does not lead to a substantial product yield improvement. A closer look at obtained carbon yields (refer to column “MEP+MVA. Carbon Yield” in Table B.1 in the Appendix) reveals that ethanol, with a carbon yield of 0.92 c-mol txd/c-mol ethanol, still exhibits the highest carbon efficiency. Although the synergy between MEP and MVA pathways increased the carbon yield for ethanol from 0.88 to 0.92 c-mol/c-mol, network performance is not optimal yet, since still 8% of all carbon atoms are lost in the form of by-products.

### 7.2.1. Three-Substrate Mixtures

Determined to find a network configuration that allows a complete substrate conversion into product, we asked the question, if substrate mixtures can potentially increase network performance. In order to keep the problem computationally tractable, we limited our study to two- and three-substrate mixtures. In total, we tested 972,150 different medium compositions:

$$\binom{180}{1} + \binom{180}{2} + \binom{180}{3} = \frac{180!}{1! 179!} + \frac{180!}{2! 178!} + \frac{180!}{3! 177!} \\ = 972,150$$

As in section 7.1.1, we used a constraint-based approach along with the *E. coli* metabolic reconstruction iJO1366 to obtain results shown in Figure 7.4. Since substrate mixtures are considered, product yields calculated on a molar basis do not allow a direct assessment of the network carbon efficiency. Consequently, product yields are reported on a carbon basis. Calculations for profit and product yields on a mass basis can be performed analogously. To guarantee that the maximal possible product yield is identified during calculations, both pathways for precursor production (mevalonate and non-mevalonate) were introduced in the model iJO1366. The applied procedure was the following:

## 7. Model-based Medium Optimization

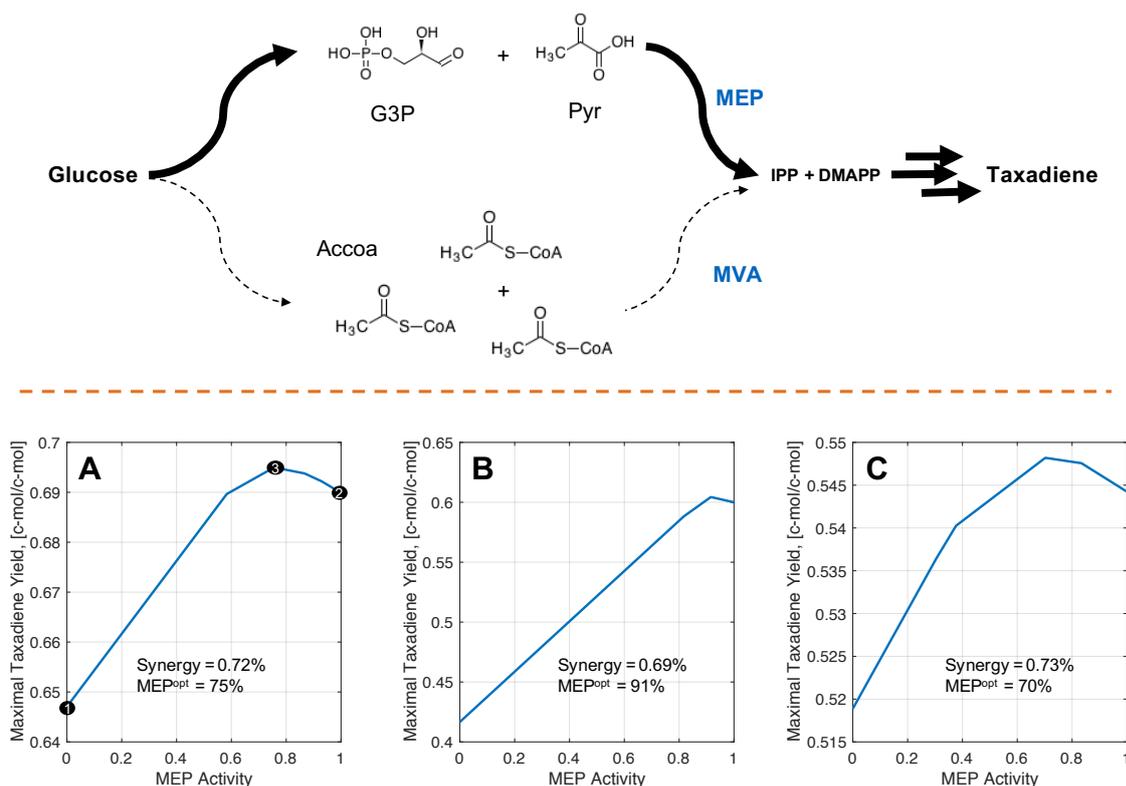


Figure 7.2.: **Synergy as Design Principle for Pathway Selection. Glucose As a Case Study.** All calculations were performed using the model iJO1366. This model was complemented with both the non-mevalonate (MEP) and the mevalonate (MVA) pathway. As shown in the top of the figure, the MVA pathway draws three acetyl-CoA molecules to produce one IPP molecule. By contrast, the MEP pathway draws a molecule of pyruvate and a molecule of glyceraldehyde 3-phosphate to produce a mixture of IPP and DMAPP. Refer to Figures A.5, A.6 and A.7 in the Appendix for a detailed overview of these two pathways and the reactions required for taxadiene production. The activity through the MEP pathway was *in silico* modulated from 0 to 1, which corresponds to its maximal flux. Modulation was done by constraining flux through the reaction DXPS, the first reaction of the MEP pathway. Additional constraints were applied as follows: **(A.)** Fully aerobic taxadiene production. Oxygen uptake rate was set to 100 mmol/gDW h, glucose uptake rate to 1 mmol/gDW h and NGAM to 0. **(B.)** Anaerobic taxadiene production. Rates were set as **(A.)** but oxygen uptake rate was set to 0. **(C.)** Aerobic taxadiene production with a growth rate of 0.2 1/h and NGAM of 3.15 mmol ATP/gDW. Synergy was calculated as explained in the main text. Points ①, ② and ③ correspond to Model 1, 2 and 3 respectively (see Table 7.1).  $MEP^{opt}$  defines the activity of the MEP pathway for which taxadiene flux is maximal.

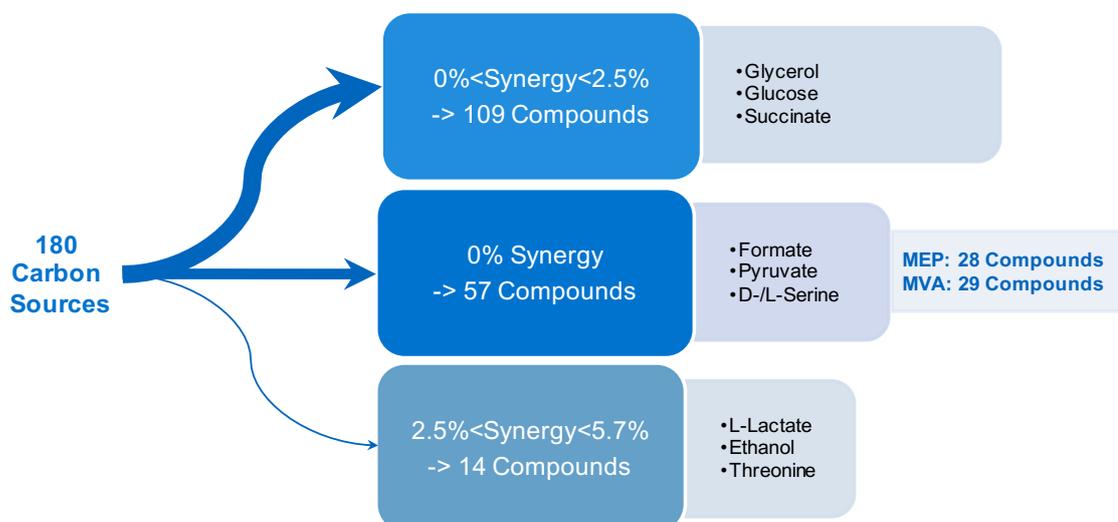
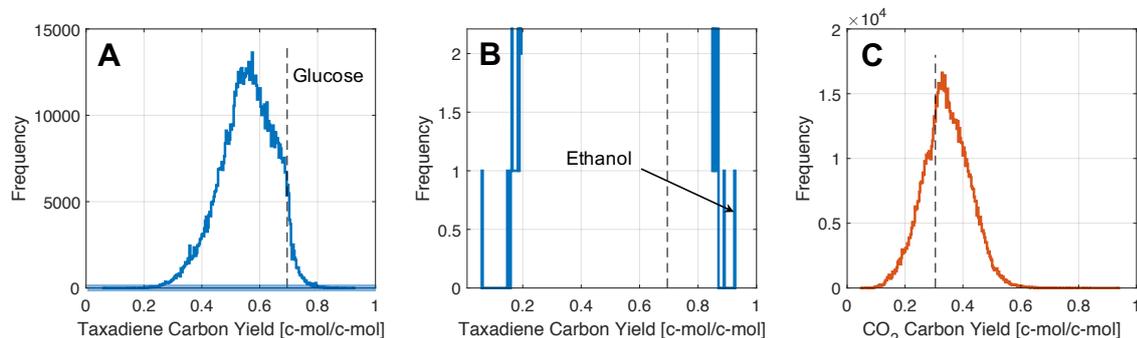


Figure 7.3.: **Synergy as Design Principle for Pathway Selection. 180 Substrates**  
 Synergy was calculated as explained in the main text for 180 different carbon sources. Taxadiene yields were calculated considering no growth rate and a NGAM value of 0. All calculations were made using the model iJO1366. Refer to Table B.1 in the Appendix for a complete overview of the results.

1. Choose one of the 972,150 possible substrate mixtures (or single substrates).
2. Allow the model to consume these substrates by setting the upper bounds for respective uptake reactions to 1 mmol substrate/gDW h.
3. Maximize flux through the taxadiene export reaction using linear programming.
4. Numerical value of product yield on a carbon basis can be calculated as follow:  $\frac{20}{\sum_{i=1}^n r_i c_i} r_{txd}$ , where  $r_{txd}$  refers to taxadiene production rate,  $r_i$  substrate uptake rate for substrate  $i$  and  $c_i$  carbon content of substrate  $i$ . Consequently,  $\sum_{i=1}^n r_i c_i$  refers to the total carbon uptake.
5. Repeat step 1 to 4 for all remaining substrate mixtures.

Figure 7.4 reveals interesting features of the taxadiene production potential of the *E. coli* metabolic network. In average, maximal product yield expressed on a carbon basis lies at 0.57 c-mol txd/c-mol substrate for the substrate mixtures analyzed (refer to Figure 7.4.A). This means that on average, 43% of the absorbed carbon atoms of a given mixture are lost in the form of by-products. Additionally, ethanol still exhibits the highest *in silico* performance, with a product yield on carbon basis of 0.924 c-mol/c-mol (refer to Figure 7.4.B). In a last attempt to push the *E. coli* metabolic network to reach its production limit *in silico*, we turned our attention to analyze by-product secretion during taxadiene production. The motivation was not only to identify a medium composition leading to complete carbon conversion, but also the identification of bottle necks in the taxadiene production. Based on an analysis of optimal flux distributions, we noticed a significant carbon leak towards CO<sub>2</sub>. In order to assess the extent of simultaneous taxadiene and CO<sub>2</sub> production, we calculated corresponding CO<sub>2</sub> carbon yields for the 972,150 substrate mixtures. Results are summarized in Figure 7.4.C. Interestingly, carbon leak in the form of CO<sub>2</sub> can amount to values up to 62% of the total carbon uptake. CO<sub>2</sub> production

## 7. Model-based Medium Optimization



**Figure 7.4.: Maximal Product Yield for Three-Substrate Mixtures** Maximal taxadiene yields were calculated for 972,150 substrate mixtures considering both MEP and MVA for taxadiene production. All yields were calculated on a carbon basis. Shown are histogram of taxadiene yields (**A.**), zoomed histogram of taxadiene yields (**B.**) and CO<sub>2</sub> yields (**C.**). Vertical dotted line correspond to glucose. Calculations were done with the model iJO1366 considering fully aerobic taxadiene production. NGAM and growth rate were set to 0. CO<sub>2</sub> yields were calculated using the CO<sub>2</sub> export flux obtained for each of the 972,150 substrate mixtures.

has many sources, including production of pathway precursors, e.g. acetyl-coA production from glucose by the enzyme pyruvate dehydrogenase, NADH/NADPH regeneration and the first and last reaction of the MEP and MVA pathway, respectively. Even though production of acetyl-CoA and co-factor regeneration (NADH/NADPH) can occur without concomitant CO<sub>2</sub> production by clever substrate selection, CO<sub>2</sub> production associated to the MEP and MVA pathways cannot be avoided without engineering enzymes responsible for CO<sub>2</sub> release, namely 1-deoxy-D-xylulose-5-phosphate synthase (encoded by gene *dxs*) and phosphomevalonate decarboxylase (encoded by gene *mvaD*), respectively (Refer to Figures A.5 and A.6 in the Appendix for a detailed overview of the MEP and MVA pathways). As far as we know, this has not been achieved yet. Thus we conclude that a *carbon neutral* taxadiene production is not possible for any substrate or substrate mixture.

If CO<sub>2</sub> production cannot be avoided, carbon fixation could offer a solution to the *the CO<sub>2</sub> challenge*. Bar-even et al. [10] designed and analyzed different synthetic carbon fixation pathways. The shortest possible fixation pathway identified by the authors corresponded to a pathway fixing two CO<sub>2</sub> molecules to produce glyoxylate, a two carbon compound. Two glyoxylate molecules can be then assimilated by a bacterial-like glycerate pathway to produce glyceraldehyde 3-phosphate (G3P) and one CO<sub>2</sub> molecule. The carbon fixation and glyoxylate assimilation pathway can be lumped into a single reaction as follows:  $3\text{CO}_2 + 3\text{ATP} + 4\text{NADPH} \rightarrow \text{G3P}$ . In order to investigate if the introduction of a synthetic CO<sub>2</sub> fixation pathway would lead to increased taxadiene yields, we introduced this pathway into the iJO1366 model. Then, we calculated the maximal taxadiene yield supported by the network using glucose as carbon source and both MEP and MVA pathways. We observed no improvement in the taxadiene yield for glucose.

### 7.3. Discussion

Even though the conclusion of *the CO<sub>2</sub> challenge* is rather unsatisfactory, we generated an extensive *in silico* knowledge regarding optimal carbon sources for taxadiene production. As such, this knowledge can be used to optimize experimental medium composition by employing substrates exhibiting high *in silico* performance values as carbon source. For instance, substrates potentially increasing the taxadiene production performance of a given main carbon source, for example glucose, can be rapidly identified and experimentally tested. Additionally, profit as performance criterion can be used in early stage techno-economic feasibility studies. By defining a minimum profit level that should be achieved by the process, it is possible to determine if a given host microorganism and a determined substrate have the *potential* to yield an economically feasible production process.

Many studies have assessed and compared the production potential of various substrates using the MEP, MVA or both pathways *in silico* [1,50,95,162]. However, the work reported here surpasses the scope of any of these studies, as we assessed and compared a total of 972,150 unique substrate mixtures. We demonstrated that it is impossible to convert any single substrate or substrate mixture into taxadiene in a carbon neutral fashion. Ethanol, with a taxadiene yield of 0.924 c-mol/c-mol, was identified by our large-scale *in silico* analysis as the substrate allowing the highest carbon efficiency. Interestingly, as identified by Gruchattka et al. [50], ethanol along with the MVA pathway has allowed the highest experimental terpenoid yield (on a carbon basis) reported to date (refer to Table 2 in [50], for original publication refer to [156]). Even though the host microorganism was *S. cerevisiae* and the terpenoid produced corresponded to artemisinic acid, this study strongly points at the potential of ethanol as an efficient carbon source for taxadiene and in general, for terpenoids production.

As for the synergistic effects arising from the simultaneous utilization of the MEP and MVA pathway, a recent publication [160] experimentally reported an improvement in isoprene yield, when both pathways were active in *E. coli*. Yield improvement amounted to 20-fold and threefold respectively, compared to the sole usage of the MEP or the MVA pathway. In the case of the fully aerobic taxadiene production in *E. coli*, our constraint-based approach predicted an increment of 0.72% and 7.38%, compared to the sole performance of the MEP and MVA pathways, respectively (refer to Figure 7.2.A). Even though the extent of the synergy experimentally observed by Yang et al. [160] can not be directly compared to our *in silico* simulations, his study suggests that synergy calculations based on sole stoichiometric constraints do possess a predictive power. An additional study supporting the validity of constraint-based assessment of pathway synergy was provided by Shen et al. [136]. The authors showed that *in silico* predicted synergistic effects arising from the simultaneous activity of two different pathways for the 1-propanol production, namely the threonine and the citramalate pathway, could be also observed *in vivo*.

Unlike our constraint-based calculations and additional reports [1,50,95] predicting a higher *in silico* potential of the MEP pathway over the MVA pathway for taxadiene production, Yang et al. [160] experimentally demonstrated the superiority of the MVA pathway over the MEP pathway for isoprene production. Since both taxadiene and isoprene production depends on the DMAPP production flux, supported by either the MEP or MVA pathway, the report by Yang et al. [160] clearly evidences one of the intrinsic limitations of constraint-based modeling. Since pathway regulation at the transcription-, translation- or

## 7. Model-based Medium Optimization

enzyme activity-level are poorly considered or not considered at all in genome-scale metabolic reconstructions, biologically relevant aspects constraining flux through metabolic pathways can be overseen. Since the MVA pathway is a non-native pathway in *E. coli*, the flux through this pathway is not constrained by any regulatory elements. Consequently, terpenoid yields using the heterologous MVA pathway have been repeatedly reported to surpass the performance of the native MEP pathway [93, 98, 160]. In spite of its flaws, constraint-based modeling represents a valuable modeling tool. This is because large scale constraint-based analyses are computationally tractable (due to the efficiency of available linear-programming solvers) and because high quality genome-scale metabolic reconstructions for many relevant industrial microorganisms are easily accessible to the research community [66, 123]. More importantly, the metabolism of all microorganisms is subjected to mass balance constraints. Hence, theoretical yields calculated using constraint-based approaches are true constraints, representing the maximal potential of a given host metabolism.

## 8. Model-based Semi-Batch Optimization

The performance of a plasmid-based production strain was experimentally assessed in a semi-batch process at different temperatures and time points of induction. Available data consisted of concentration time courses for biomass, substrate and product. The modeling task was to identify optimal production temperature and time point of induction *in silico*. To that end, we applied dFBA as modeling framework to describe the available experimental data. We used Arrhenius-like functions to mathematically describe the temperature dependence of calculated rates (growth, substrate uptake and production). Maximal productivity was used as the performance metric to be optimized by the model. By perturbing both production temperature and time point of induction and analyzing the system response, we identified a process setting that increased the maximal productivity by 56% *in silico* [84]. Unfortunately, Experimental implementation of this model-based prediction by our cooperation partner was not possible due to lacking reproducibility of experimental results. This was potentially caused by strain/plasmid instability. However, we believe that the proposed approach represents a valid proof of concept for process optimization using readily available experimental data in early stages of strain and process development.

### 8.1. Strain & Experimental Data

The strain used in this work was constructed by our cooperation partner Max Hirte at **IBK** and experimentally characterized by Lars Janoscheck at **STT**. Experimental strain characterization was performed in a 1.5 liter bio-reactor and was limited to the measurement of concentration time courses for glucose, biomass and taxadiene. Production was induced by adding Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) to the bio-reactor after a previously defined biomass concentration was reached. We termed this moment the induction time point,  $t_{induction}$ . As shown in Table 8.1, this process parameter remained relatively constant for all processes, ranging from 6.25 to 8.25 hours. Since the main objective of this first characterization round was to identify an optimal production temperature, four different temperatures ranging from 16 to 37°C were tested. Note that while the production temperature ( $T_{production}$ ) was varied, the temperature before induction was kept constant at a value of 37°C to ensure fast biomass formation. Figure 8.1 shows a schematic representation of the process set-up and experimental time courses obtained for a production temperature of 16 °C. Data for all other temperatures can be found in Tables B.3 to B.8, in Appendix B.2.1. The effect of IPTG addition on taxadiene production can be clearly observed in Figure 8.1.D. Shortly after induction, taxadiene is formed and its concentration keeps increasing as the process continues.

In order to use provided experimental data to identify optimal process parameters (production temperature and time point of induction), it is first necessary to develop a process

## 8. Model-based Semi-Batch Optimization

Table 8.1.: **Experimental Process Parameters for Six Semi-Batch Fermentations.**

In order to identify optimal process conditions, six semi-batch fermentations were performed, testing four different production temperatures and time points of induction. Refer to Tables B.3 to B.8, in Appendix B.2.1 for experimental concentration time courses.

Process ID	Temperature before induction [°C]	Production Temperature [°C]	Induction Time Point [h]	Process Duration [h]
1	37	16	8.5	76.5
2	37	16	7	73.75
3	37	22	7.25	75
4	37	37	7.25	75.25
5	37	22	6.25	74.75
6	37	27	6.25	74.75

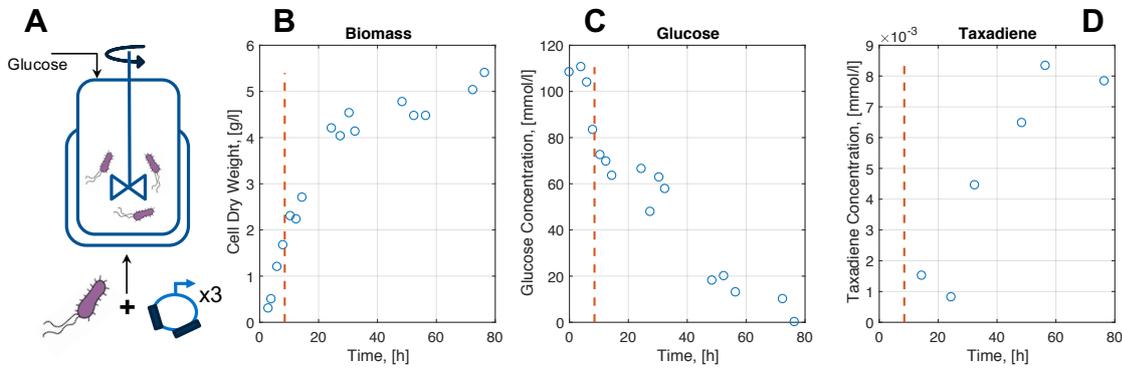


Figure 8.1.: **Process Set-up and Experimental Data** A production strain containing three plasmids was characterized in a bioreactor using four different production temperatures. During the process, glucose was added to the bioreactor in irregular time intervals to prevent the culture from starving (A.). Strain characterization consisted of measurements of concentration time courses for dry weight biomass (B.), glucose (C.) and taxadiene concentration (D.). Dashed vertical line in Figure B. to D. represents the time point of plasmid induction  $t_{induction}$ .

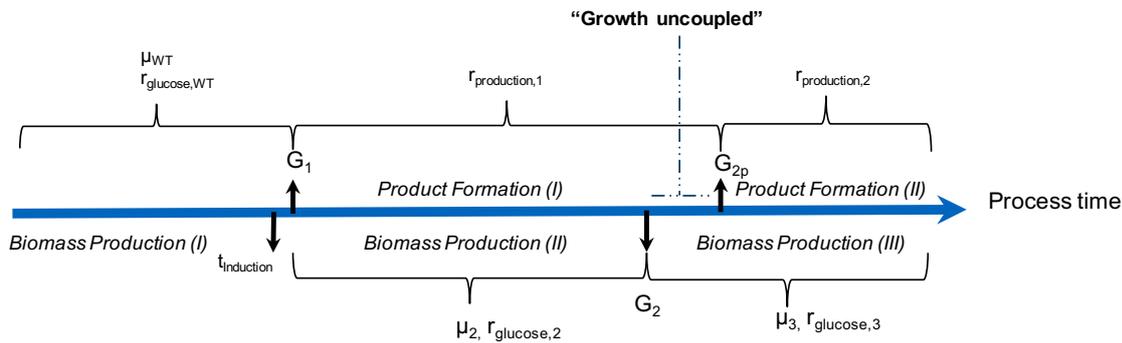


Figure 8.2.: **Modeling Experimental Data: Characteristic Process Times.** Three different biomass production phases arise when characteristic times  $G_1$  and  $G_2$  are introduced. Analogously, two taxadiene formation phases are defined by  $G_1$  and  $G_{2p}$ . Uncoupling taxadiene formation phases from biomass production provided a better fit of experimental data (results not shown). This is done by introducing  $G_{2p}$  (with  $G_{2p} \neq G_2$ ) to mark the end of the first taxadiene formation phase.  $r_{glucose}$ ,  $r_{production}$  and  $\mu$  refer to glucose uptake, taxadiene secretion and growth rates, respectively.  $\mu_{WT}$  refers to the growth rate of the wild-type.

## 8.2. Mathematical Identification of Characteristic Process Phases and Rates Calculation

model capable of describing taxadiene production and its effect on biomass formation and glucose consumption. To describe growth dynamics, we first divide the process time into three different phases, defined by two characteristic times  $G_1$  and  $G_2$ . The rationale behind this strategy is the following: before the system is induced, cells grow with a rate defined by temperature and substrate concentration. This two parameters were kept constant for all six fermentations. Once the system is induced by IPTG, cellular growth is sharply decreased, as the cell reallocates its limited transcription and translation machinery to support the expression of genes encoded in the three plasmids. Thus,  $G_1$  defines the moment, in which a reduction in the growth rate can be experimentally observed. We refer to the difference between  $t_{induction}$  and  $G_1$  as the time response. Since Taxadiene can be produced after all necessary enzymes are synthesized by the cell,  $G_1$  also marks the moment in which taxadiene production begins. Due to the accumulation of toxic by-products and diminishing growth-supporting substrates, the cellular growth rate should decrease one more time as the process continues. We use the characteristic time  $G_2$  to describe this moment. Analogously,  $G_{2p}$  is introduced to describe the moment in which the taxadiene production rate decreases or completely ceases. Figure 8.2 presents an overview of the location of all characteristic process times.

## 8.2. Mathematical Identification of Characteristic Process Phases and Rates Calculation

An optimization-based approach is introduced to guide the identification of characteristic process times for biomass production as follows:

Find  $G_1$  and  $G_2$  maximizing  $Z$ :

$$Z = \sum_i^3 c_i R_i^2(G_1, G_2) = c_1 R_1^2 + c_2 R_2^2 + c_3 R_3^2,$$

with  $R_i^2$  being the determination coefficient corresponding to the least squares regression of the natural logarithm of dry weight measurements against time, for phase  $i$  (see Figure 8.3.A), and:

$$c_i = 1 \text{ if } \#Points \geq 3, \text{ else } c_i = 0.$$

$\#Points$  refers to the number of data points used for the linear regression.

Optimal  $G_1$  and  $G_2$  are identified using a brute force approach, in which  $Z$  values are calculated for all possible  $G_1$  and  $G_2$  combinations within a predefined interval. Since  $G_1$  is defined to be located shortly after  $t_{induction}$ , we scan for optimal  $Z$  values between  $G_1 = t_{induction}$  and  $G = t_{induction} + 4$  h. As for  $G_2$ , preliminary calculations showed optimal values to be located between 15 h and 60% of the total process time. Figure 8.3 shows the results obtained after applying this methodology for the identification of characteristic process times for batch 1. Optimal values for  $G_1$  and  $G_2$  can be obtained from Figure 8.3.B or C. Additionally, Figure 8.3.C, suggests the existence of many combinations of  $G_1$  and  $G_2$  values reaching at least 95% of the maximal  $Z$  value obtained for the whole simulation.

The identification of characteristic times implicitly allows the calculation of rates for growth, glucose uptake and taxadiene production. Growth rate in each of the three bio-

## 8. Model-based Semi-Batch Optimization

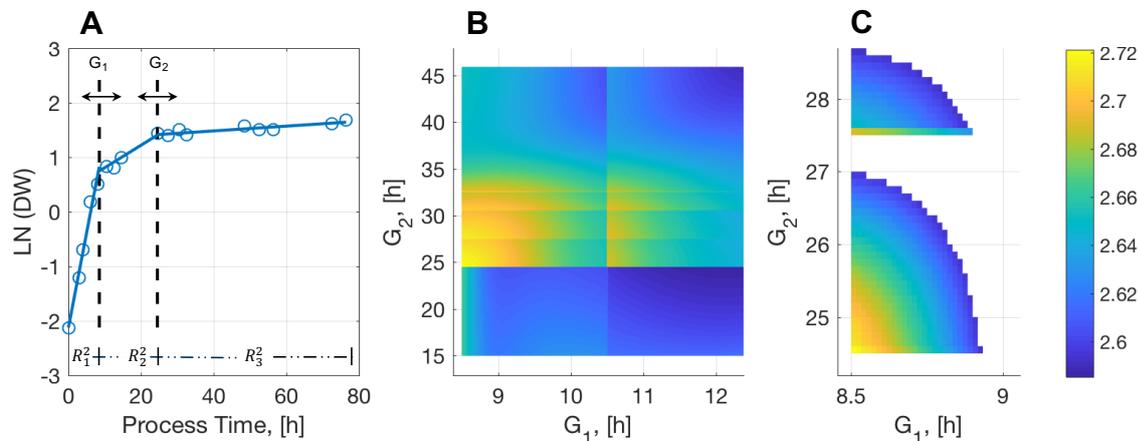


Figure 8.3.: **Modeling Experimental Data: Mathematical Identification of Process Phases** Results are shown for batch 1. **A.** Natural logarithm of biomass dry weight measurements (DW) is plotted against time. Our optimization-based approach identifies characteristic times  $G_1$  and  $G_2$  such that the sum of the three resulting  $R^2$  is maximal. **B.** Values of the objective function  $Z$  for different  $G_1$  and  $G_2$  combinations. **C.** Characteristic process times achieving at least 95% of the maximal  $Z$ . As can be observed from this figure, Optimal characteristic process time correspond to  $G_1 = 8.5$  h and  $G_2 = 24.5$  h.

Table 8.2.: **Growth, Production and Substrate Uptake Rates for Different Process Phases and Production Temperatures.** Growth rates were calculated as described in Section 2.1.2 for three different phases identified by the optimization-based approach described in the main text. The same phases were used for the determination of substrate uptake rates. By contrast, only two phases were used for the calculation of production rates. Additionally, the identification of the characteristic time  $G_{2p}$  was done manually. However, an optimization-based approach could be analogously implemented.

Batch ID	Characteristic Times [h]				Growth Rate [1/h]			Glucose Uptake Rate [mmol/gDW h]			Txd Production Rate [mmol/gDW h]	
	tind	G1	G2	G2p	Phase I	Phase II	Phase III	Phase I	Phase II	Phase III	Phase I	Phase II
1	8.5	8.5	24.5	52.5	$3.402 \times 10^{-1}$	$4.010 \times 10^{-2}$	$4.320 \times 10^{-3}$	5.024	0.652	0.305	$4.770 \times 10^{-5}$	0.000
2	7	7.5	23.8	26.5	$2.983 \times 10^{-1}$	$3.230 \times 10^{-2}$	$3.030 \times 10^{-4}$	3.272	0.678	0.251	$7.858 \times 10^{-5}$	$1.114 \times 10^{-5}$
3	7.25	7.65	25.2	28.7	$3.091 \times 10^{-1}$	$6.149 \times 10^{-2}$	$3.297 \times 10^{-3}$	6.636	1.486	0.347	$4.412 \times 10^{-4}$	0.000
4	7.25	8	27.25	19.1	$4.265 \times 10^{-1}$	$3.070 \times 10^{-2}$	$2.060 \times 10^{-3}$	4.675	2.569	1.022	$8.401 \times 10^{-5}$	0.000
5	6.25	6.5	30.2	38	$5.642 \times 10^{-1}$	$2.220 \times 10^{-2}$	$4.020 \times 10^{-3}$	5.878	0.474	0.216	$1.280 \times 10^{-4}$	$1.657 \times 10^{-5}$
6	6.25	6.9	24.2	29.7	$4.807 \times 10^{-1}$	$4.140 \times 10^{-2}$	$3.840 \times 10^{-3}$	8.491	1.295	0.355	$3.867 \times 10^{-4}$	0.000

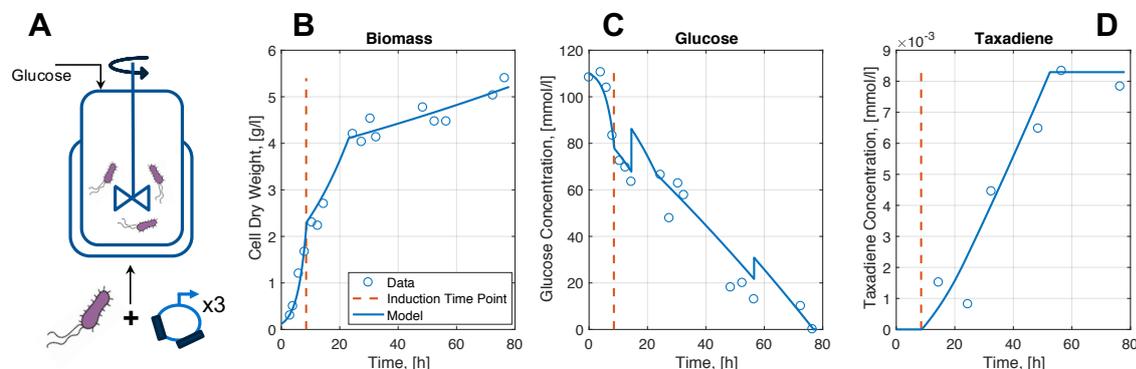


Figure 8.4.: **Modeling Experimental Data: Results** Experimental process set-up used for semi-batch fermentations (**A.**). Measured time courses for biomass (**B.**), glucose (**C.**) and taxadiene (**D.**) are compared with their respective *in silico* simulated time course for the Batch 1. Dashed vertical line in Figure **B.** to **D.** represents the time point of plasmid induction  $t_{induction}$ .

### 8.3. Dependence of Model Parameters on Production Temperature

mass production phases equals the slope of the corresponding line shown in Figure 8.3.A. Glucose uptake and taxadiene secretion rates are calculated by multiplying the growth rate of each of the phases by the slope of the straight line that results from plotting the glucose (or taxadiene) concentration as a function of biomass dry weight. Refer to Section 2.1.2 for more details on the calculation of rates from concentration time courses. Table 8.2 summarizes characteristic process times and respective rates for all six data sets. Note that taxadiene production rate in the second product formation phase (see Figure 8.2) was set to zero for Batches 1, 3 4 and 6. While this is a valid assumption in Batches 1 and 6 (refer to Figures 8.4 and B.5, respectively) because the taxadiene concentration remains relatively constant during this phase, we clearly observe a reduction in the taxadiene concentration for Batches 3 and 4 (refer to Figures B.2 and B.3 in Appendix B.2.3). This means that the taxadiene production rate in the second product formation phase should be negative rather than zero. Since the number of experimental points for taxadiene concentration in this phase is very limited, we assume a zero production rate to be more accurate than a rate calculated using just two data points. Independently of the validity of this assumption, it does not have an effect on predicted optimal  $T_{production}$  or  $t_{induction}$  values, as we will discuss later.

Exchange rates and characteristic process times are parameters that allow for a complete description of respective experimental concentration time courses. To demonstrate this, we applied dFBA to generate *in silico* time courses for each semi-batch fermentation process using parameters listed in Table 8.2 along with information regarding glucose pulses and initial biomass and glucose concentrations for each batch (refer to Tables B.3 to B.8, in Appendix B.2.1 for numerical values). As explained in Section 2.1.1, dFBA allows for the combination of reactor dynamics with cellular metabolism, thus allowing the simulation of concentration time courses of extra-cellular metabolites in the reactor. Figure 8.4 exemplarily shows the performance of dFBA simulations for batch 1 (refer to Figures B.1 to B.5 in Appendix B.2.3 for a complete results overview). As observed, *in silico* generated concentration time courses closely match experimental data (see Table B.11 in Appendix B.2.3 for  $R^2$  values). All dFBA simulations reported in this work were performed using MATLAB functions provided by DFBA Lab [47].

### 8.3. Dependence of Model Parameters on Production Temperature

The main objective of the modeling effort presented in this chapter is the identification of optimal values for two process parameters, namely production temperature and time point of induction. So far, we have developed a modeling approach that allows the simulation of concentration time courses, closely matching experimental data. Since not all parameters listed in Table 8.2 are *directly* accessible, meaning that their numerical values cannot be freely adjusted experimentally, we venture to explore the effect of both production temperature ( $T_{production}$ ) and time point of induction ( $t_{induction}$ ) on them. Note that numerical values of growth and glucose uptake rate during the first production phase,  $\mu_{WT}$  and  $r_{glucose,WT}$  in Figure 8.2, are considered to be constant, since strain, initial medium composition and temperature in the first biomass production phase were not changed and are not subject to optimization in this work.

## 8. Model-based Semi-Batch Optimization

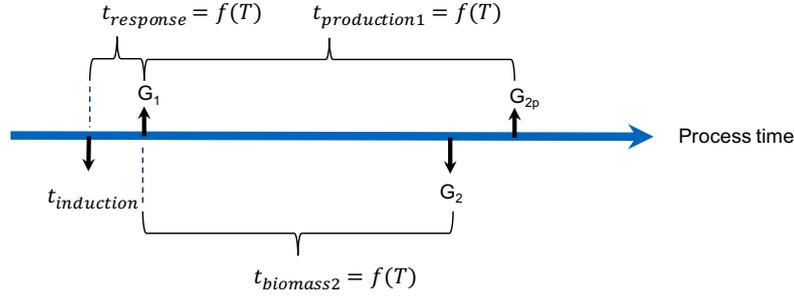


Figure 8.5.: **Dependence of Characteristic Process Times on Temperature and Time Point of Induction.** The dependence of characteristic process times on temperature is studied indirectly, by analyzing the relationship between duration of phases and temperature. This approach would allow the estimation of characteristic process times obtained at a non-measured  $T_{production}$  and  $t_{induction}$ .

We first study the dependence of characteristic process times on  $t_{induction}$  and  $T_{production}$ . For that, we apply the schema shown in Figure 8.5. Instead of directly analyzing the effect of  $T_{production}$  on  $G_1$ ,  $G_2$  and  $G_{2p}$ , we analyze its effect on  $t_{response}$  and the duration of the second biomass production phase ( $t_{biomass2}$ ) and first taxadiene formation phase ( $t_{production1}$ ). Figure 8.6 shows these relationships. While  $t_{response}$  linearly increases with  $T_{production}$ , meaning that the cell needs more time to synthesize enzymes required for taxadiene production at high temperatures, the duration of the main production phase,  $t_{production1}$  linearly decreases as  $T_{production}$  increases. As for  $t_{biomass2}$ , we do not observe a clear dependence on  $T_{production}$ . For that reason, we assume a constant value of 18.3 h, which corresponds to the average duration of the second biomass formation phase, as calculated from characteristic times listed in Table 8.2.

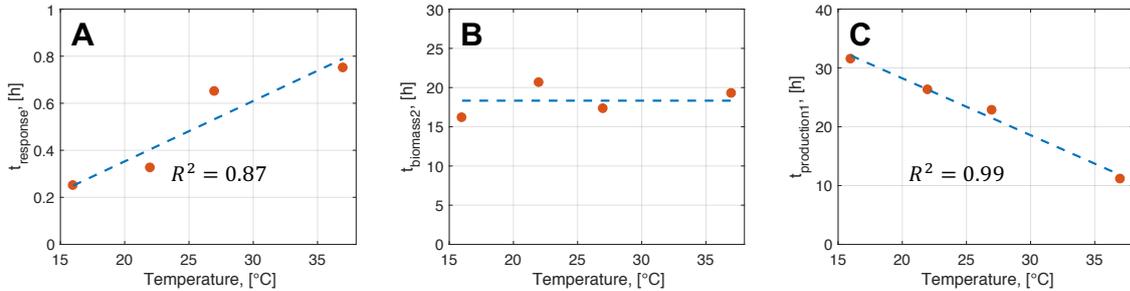


Figure 8.6.: **Modeling Experimental Data: Temperature Dependence of Duration of Process Phases** Linear relationships are used to describe the effect of  $T_{production}$  on  $t_{response}$  (A.), on the duration of the second biomass production phase (B.) and first product formation phase (C.).

The effect of  $t_{induction}$  on  $G_1$ ,  $G_2$  and  $G_{2p}$  can be described as follows:

$$\begin{aligned} G_1 &= t_{response} + t_{induction} \\ G_2 &= t_{biomass2} + G_1 \\ G_{2p} &= t_{production1} + G_1 \end{aligned}$$

Using these equations and the relationships shown in Figure 8.6, it is possible to calculate

### 8.3. Dependence of Model Parameters on Production Temperature

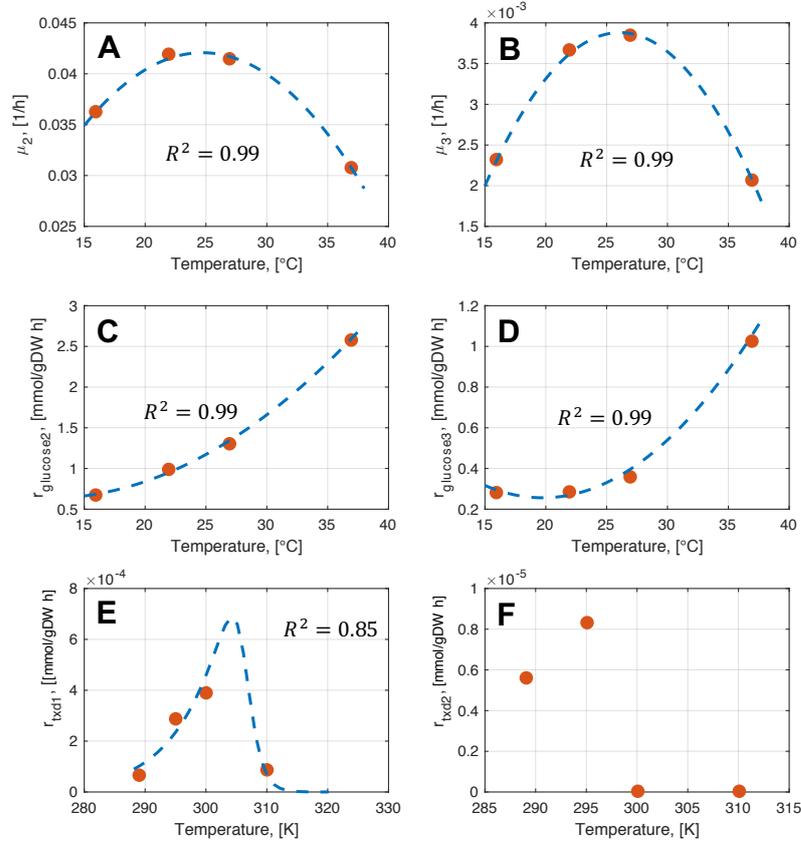


Figure 8.7.: **Modeling Experimental Data: Temperature Dependence of Exchange Rates** Rates shown for a  $T_{\text{production}}$  of 16 and 22°C correspond to average values. A quadratic polynomial is used to mathematically describe the effect of temperature on growth (A. and B.) and glucose uptake rates (C. and D.) in the second and third biomass production phase. As for the taxadiene production rate in the first product formation phase, an Arrhenius-like formula is applied to mathematically describe the observed temperature dependency (E.). Plot F. shows taxadiene production rate obtained for each  $T_{\text{production}}$ . Refer to Table B.10 in Appendix B.2.2 for more details on the formulas used for regression.

numerical values for  $G_1$ ,  $G_2$  and  $G_{2p}$  for a given set of  $T_{\text{production}}$  and  $t_{\text{induction}}$ .

To describe the temperature dependence of growth and glucose uptake rate in the second and third biomass production phases, we used a second-degree polynomial. This approach was motivated by the work by Ratkowsky et al. [113], which suggested a linear relationship between the square root of growth rate and temperature, as follows:  $\sqrt{\mu} = b(T - T_o)$ . As shown in Figure 8.7 (plots A to D), a quadratic polynomial allows a good description of the experimentally observed temperature dependence of both growth and glucose uptake rates. Previous to the work by Ratkowsky et al. [113], Arrhenius-like formulas were suggested to describe the effect of temperature on biological phenomena (growth and bacterial luminescence) [36,57]. We use the formula suggested by Johnson et al. [57], developed to describe the effect of temperature on cellular growth, to account for the effect of temperature on the production rate during the first product formation phase. As evidenced in Figure 8.7.E, this approach provides a good fit to the data.

## 8. Model-based Semi-Batch Optimization

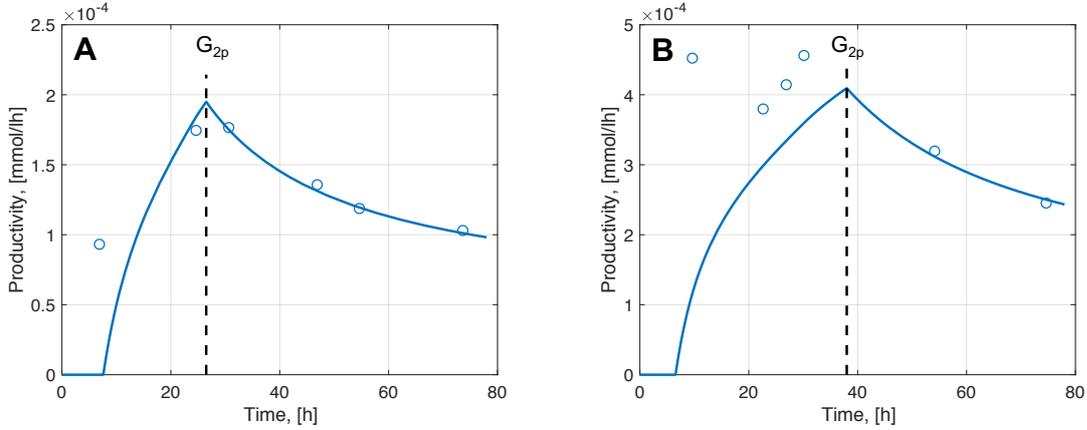


Figure 8.8.: **Volumetric Productivity for Batches 2 and 5** Experimentally observed taxadiene productivity (circles) is plotted as a function of process time for Batches 2 (A.) and 5 (B.). Blue line represents predictions made with our dFBA-based approach.

### 8.4. *In silico* Optimization of Production Temperature and Time Point of Induction

So far, we have identified numerical values for characteristic process times and exchange rates (see Table 8.2), which used in combination with dFBA, allow the generation of concentration time courses for biomass, glucose and taxadiene, closely matching experimental data. Additionally, we introduced formulas to describe the effect of two process parameters, namely  $T_{production}$  and  $t_{induction}$ , on them (characteristic process times and exchange rates). In this way, it is possible to *in silico* generate concentration time courses for  $T_{production}$  and  $t_{induction}$  values experimentally not tested so far and evaluate their impact on a given performance metric. In this section, we show how the developed model can be applied to identify values for  $T_{production}$  and  $t_{induction}$  that potentially increase the maximal volumetric productivity of the process. This metric is defined as follows:  $P_{max} = \max(\frac{txd_{concentration}}{t})$ , where  $txd_{concentration}$  refers to a vector containing taxadiene concentrations at certain time points, contained in the time vector  $t$ .

In order to reduce the simulation time and to provide a reliable prediction regarding optimal process parameters, we assume that for a given batch process,  $P_{max}$  is contained in the first product formation phase. This assumption was done to avoid inaccuracies in the calculation of the taxadiene time course during the second product formation phase, arising due to the incapacity of accurately describe the observed decrease in the taxadiene concentration during this phase for Batches 3 and 4 (refer to Figures B.2 and B.3, respectively). Taking a look at the taxadiene time course for Batches 2 and 5 (Figures B.1 and B.4, respectively), one might argue that this assumption negatively affects the prediction power of the model, since the taxadiene concentration keeps increasing for both batches during the second product formation phase. Interestingly, our simulations (see Figure 8.8) suggest that for both batches,  $P_{max}$  is indeed located within the first product formation phase, indicating that our assumption is accurate and do not affect the prediction power of the model. A similar behavior is observed for all other batches.

In order to identify numerical values for  $T_{production}$  and  $t_{induction}$  increasing  $P_{max}$  we proceed

#### 8.4. *In silico* Optimization of Production Temperature and Time Point of Induction

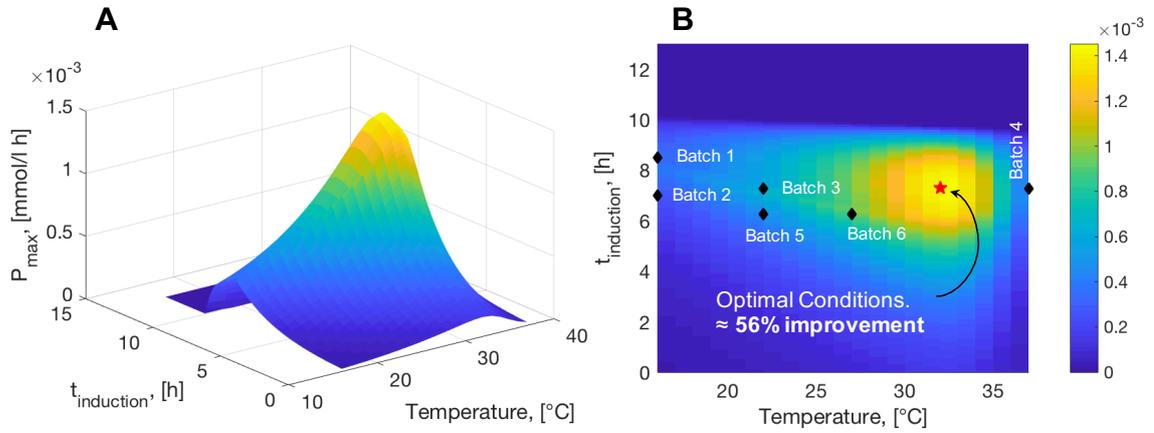


Figure 8.9.: **Model-based Optimization of Process Parameters** A. 3D-View B. Top view. Optimal conditions correspond to  $t_{induction} = 7.3$  h and  $T_{production} = 32^{\circ}\text{C}$ . Under these conditions,  $P_{max}$  amounts to  $1.45 \times 10^{-3}$  mmol/l h.

as follows:

1. Generate a two dimensional grid for  $T_{production}$  ranging from 16 to  $37^{\circ}\text{C}$  and  $t_{induction}$  ranging from 0 to 13 hours, with a defined step, for instance  $1^{\circ}\text{C}$  for  $T_{production}$  and 0.1 h for  $t_{induction}$ . Each position in the grid represents a specific combination of both process parameters.
2. For each position in the grid:
  - a) Calculate exchange rates and characteristic process times using formulas listed in Table B.9 and B.10. Note that growth and glucose uptake rate during the first biomass production phase are kept constant, and set to average observed values:  $\mu_1 = 0.403$  1/h;  $r_{glucose} = 5.662$  mmol/gDW h (refer to Table 8.2)
  - b) Define initial concentrations. For instance 0.12 g/l biomass, 111 mmol/l glucose and 0 mmol/l taxadiene.
  - c) Generate respective concentration time courses using dFBA.
  - d) Compute  $P_{max}$
3. Generate a three dimensional plot to evaluate the effect of  $T_{production}$  and  $t_{induction}$  on  $P_{max}$

Figure 8.9 summarizes the results obtained after applying the procedure described above. A total of 2882 unique combinations of  $T_{production}$  and  $t_{induction}$  were tested *in silico*. As shown in the figure, a production temperature of  $32^{\circ}\text{C}$  and an induction time point of 7.3 hours represent optimal process parameters. Under these conditions, a  $P_{max}$  of  $1.45 \times 10^{-3}$  mmol/l h is reached. Compared to the  $P_{max}$  of  $9.2615 \times 10^{-4}$  mmol/l h for Batch 6, the highest  $P_{max}$  experimentally observed for all batches, this represents an increment of roughly 56 %.

## 8.5. Discussion

In this chapter, we developed a dFBA-based model capable of reproducing experimentally measured concentration time courses for biomass, glucose and taxadiene. Exchange rates and characteristic process times (see Figure 8.2) were used as model parameters. By expressing these parameters as a function of experimentally accessible process variables, namely time point of induction ( $t_{induction}$ ) and production temperature ( $T_{production}$ ), we were able to identify numerical values for these variables, which increased the maximal process productivity,  $P_{max}$ , by 56% *in silico*. Unfortunately and due to reproducibility issues, possibly related to strain instability, our cooperation partner at **STT** was not able to confirm our model predictions.

The introduction of characteristic process times and their associated phases represents a key element of our modeling approach. In order to keep model complexity as low as possible, we introduced three biomass production phases and two product formation phases. We believe that this strategy considers the lowest possible number of phases, while allowing a correct description of the data. Clearly, alternative models considering more phases for biomass production and product formation are also possible. However, these approaches would introduce unnecessary complexity, while at the same time complicating the interpretation of model parameters from a biological point of view.

In order to identify characteristic process times, we applied an optimization-based method. By maximizing an objective function  $Z$ , representing the goodness of fit of the biomass dry weight time course, values for  $G_1$  and  $G_2$  could be identified. By contrast, characteristic time  $G_{2p}$  was obtained by visual inspection. Consequently, our model reproduced the time course for biomass with increased accuracy, while the goodness of fit for glucose and taxadiene was comparatively poor (see Table B.11, in Appendix B.2.3). A possibility to correct this drawback is to consider all characteristic times to maximize a more balanced objective function  $Z_{Total}$ :

$$Z_{Total}(G_1, G_2, G_{2p}) = Z_{Biomass}(G_1, G_2) + Z_{Glucose}(G_1, G_2) + Z_{Taxadiene}(G_{2p}),$$

where  $Z_{Glucose}$  and  $Z_{Taxadiene}$  are defined analogously to  $Z_{Biomass}$ . This approach will certainly allow for data description in a more balanced fashion. However, it will also require more experimental information. In our opinion, this could compromise the biological significance of model parameters, since its determination would be exposed to an increased degree of experimental error. The reason why we decided to only use the biomass time course for the identification of characteristic process times was simple. As biomass formation can be easily monitored through optical density measurements (OD), more data points with less experimental error are available. To demonstrate the feasibility of using OD time courses as input signals defining  $G_1$  and  $G_2$ , we applied a similar approach to the one described in this chapter to identify optimal process conditions. Our OD-based model suggested a production temperature of 30.5°C and a time point of induction of 7.5 hours as optimal, in line with the previously obtained optimal values using a dry weight-based model (refer to Figure 8.9).

In spite of the mentioned drawbacks, numerical values for exchange rates identified by our approach for each phase had a biological meaning. For instance, the detrimental effect of system induction on cellular fitness can be evidenced by comparing the numerical

values for growth and substrate uptake rate between Phase I and II (see Table 8.2). As a consequence of taxadiene production, the cellular growth rate decreased by roughly one order of magnitude. This behavior is in line with the phenomenon of metabolic burden, discussed in Chapter 4. Additionally, the temperature dependence of growth, glucose uptake and taxadiene production rates could be accurately described using established models [36, 57, 113].

Even though we limited our analysis to the identification of optimal values for two different process parameters, the spectrum of potential applications of the developed model is wide, including among others:

1. *Development of substrate feeding strategies.* Using a model-based approach to systematically design feeding strategies, starving phases in the bioreactor along with their associated negative impact on  $P_{max}$ , can be minimized/eliminated. Refer to Figure B.3 in Appendix B.2.3 for an illustration of this issue.
2. *Estimation of optimal batch duration.* After the maximal productivity ( $P_{max}$ ) has been reached, additional process duration generates the process productivity to decrease (see Figure 8.8). Our model allows the identification of optimal values for this process variable, so that no unnecessary production time is wasted.
3. *Estimation of by-product secretion patterns.* For each point of time during the process, dFBA calculates flux distributions fulfilling certain constraints. In our specific case, these constraints correspond to growth, glucose uptake and taxadiene production rates for each phase. By maximizing and minimizing the flux through exchange reactions of interest, for instance acetate secretion, concentration time courses for the desired by-product can be estimated. A similar strategy can be applied to analyze differences in the intracellular flux distributions for the various process phases, thus allowing the identification of potential bottle necks limiting the production flux.



## 9. Product Yield Prediction

A theoretical approach for accurate product yield prediction would reduce experimental efforts during the development of optimal production strains. It would radically boost the pace of the science of Metabolic Engineering. Traditionally, constraint-based methods have been used to assess the production potential of metabolic networks. By applying linear programming, it is possible to calculate the maximal theoretical product yield for a defined carbon source and a given metabolic network (refer to Chapter 7). However, this metric only serves as a maximal bound for product yield and does not always represent an adequate proxy for experimental product yield. This is especially true when it comes to describing the production performance of strains in early development stages (refer to Table 2 in [50]).

Here, a kinetic-based approach was *applied* to predict experimental yields of six different semi-batch processes. The method is based on the EM approach and was introduced by the group of James Liao [71]. It requires three inputs: a metabolic network, a reference flux distribution and a rate law for each reaction. Unknown kinetic parameters are sampled so that the provided reference flux distribution is fulfilled. We observed that this method predicted experimental yields more accurately than its ad hoc constraint-based counterpart. Since labeling data for the semi-batch experiments were lacking, reference flux distributions were estimated using constraint-based methods along with two different objective functions: minimization of oxygen uptake and minimization of citrate synthase activity. We observed a dependence of calculated yield values on the reference flux distribution. However, regardless of the objective function used, predictions made by the kinetic-based method were by up to two orders of magnitude more accurate than the constraint-based method. We also adjusted traditionally used *E. coli* metabolic reconstructions [106, 107], making them suitable inputs for the EM approach. Finally, we used the COBRA Toolbox [124] to create required Matlab structures for the EM approach [70, 71, 74, 144, 146]. Thus, we created a missing bridge between widespread constraint-based and emerging kinetic modeling platforms [91].

### 9.1. Strain & Experimental Data

Experimental data used correspond to experimental taxadiene yields obtained for the six semi-batch fermentations introduced in Chapter 8. Taxadiene yields shown in Table 9.1 were calculated as the ratio of taxadiene production rate to glucose uptake rate for each of the two production phases. Biomass yields were calculated analogously, as the ratio between growth rate in each phase and corresponding glucose uptake rate. Refer to Section 8.1 and Table 8.2 for a description of experimental set-up and rates determination.

Since taxadiene production ceased during the second phase (between G2p and process end) for the majority of the fermentations (batches 1, 3, 4, 6), we will focus our analysis on the

## 9. Product Yield Prediction

Table 9.1.: **Biomass and Product Yields for Six Semi-Batch Fermentations.** Biomass and Product yields are calculated for all phases. Yields are calculated as the ratio between taxadiene secretion rate (or growth rate) and glucose uptake rate. When calculating biomass yield, glucose uptake rate should be converted to units of g/gDW h.

Batch ID	Biomass Yield [g/g]			Txd Yield [mol/mol]	
	1 Phase	2 Phase	3 Phase	1 Phase	2 Phase
1	$3.758 \times 10^{-1}$	$3.413 \times 10^{-1}$	$7.875 \times 10^{-2}$	$7.313 \times 10^{-5}$	0.000
2	$5.061 \times 10^{-1}$	$2.644 \times 10^{-1}$	$6.701 \times 10^{-3}$	$1.159 \times 10^{-4}$	$4.438 \times 10^{-5}$
3	$2.585 \times 10^{-1}$	$2.296 \times 10^{-1}$	$5.274 \times 10^{-2}$	$2.968 \times 10^{-4}$	0.000
4	$5.064 \times 10^{-1}$	$6.633 \times 10^{-2}$	$1.119 \times 10^{-2}$	$3.270 \times 10^{-5}$	0.000
5	$5.328 \times 10^{-1}$	$2.601 \times 10^{-1}$	$1.031 \times 10^{-1}$	$2.702 \times 10^{-4}$	$7.657 \times 10^{-5}$
6	$3.143 \times 10^{-1}$	$1.775 \times 10^{-1}$	$6.003 \times 10^{-2}$	$2.987 \times 10^{-4}$	0.000

Table 9.2.: **Average Biomass and Product Yields for Four Different Production Temperatures.** Yields were calculated for the respective phases after system induction. Yield values for 16°C were calculated from average values of batches 1 and 2, while yield values for 22°C were calculated from batches 3 and 5. Errors correspond to one standard deviation. Since for production temperatures of 27 and 37°C only one single data set was available, no errors can be reported.

Temperature [°C]	Average Biomass Yield [g/g]	Average Product Yield [mol/mol]
16	$3.029 \times 10^{-1} \pm 5.434 \times 10^{-2}$	$9.451 \times 10^{-5} \pm 3.024 \times 10^{-5}$
22	$2.449 \times 10^{-1} \pm 2.157 \times 10^{-2}$	$2.835 \times 10^{-4} \pm 1.88 \times 10^{-5}$
27	$1.775 \times 10^{-1}$	$2.987 \times 10^{-4}$
37	$6.633 \times 10^{-2}$	$3.270 \times 10^{-5}$

first taxadiene production phase (refer to Figures 8.2 and 8.3 for phase definition). Table 9.2 shows the average taxadiene and biomass yield calculated for the respective phases after system induction. Yield values for 16°C were calculated from batches 1 and 2, while yield values for 22°C were calculated from batches 3 and 5.

## 9.2. Constraint-based Product Yield Prediction

Constraint-based modeling can be applied to calculate maximal theoretical product yields for various carbon sources and genotypes, as already shown in Chapter 7. Due to the simplicity of its calculation, maximal theoretical yield has been traditionally used to assess and compare production potential under a variety of conditions. We start our analysis by calculating maximal theoretical product yield for all six available data sets (corresponding to four production temperatures) using experimentally measured growth and substrate uptake rates (refer to Table 8.2) as constraints. The calculations were performed using an expanded version of the *E. coli* core model (refer to Appendix B.1.1) as follows:

1. Select one of the six data sets.
2. Constraint the *E. coli* core model by setting numerical values of upper and lower bounds of the growth and substrate uptake reactions to match experimentally measured rates. Refer to Table 8.2 for rate values.

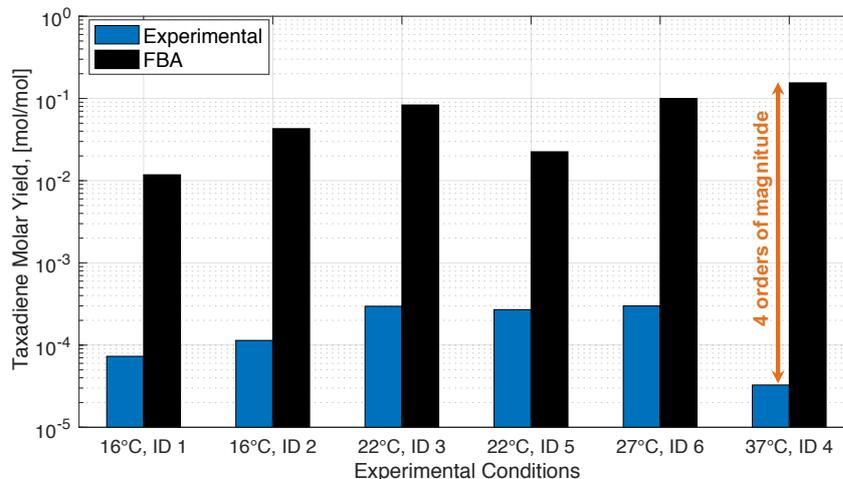


Figure 9.1.: **Constraint-based Product Yield Prediction.** Model used for constraint-based product yield calculations was a modified version of the *E. coli* core model. The model was constrained by setting growth and glucose uptake rates to corresponding experimental values. Maximal theoretical yield was calculated using linear optimization.

3. Maximize flux through the taxadiene export reaction using linear programming.
4. Calculate taxadiene yield as the ratio between production flux calculated in the previous step and experimentally measured substrate uptake rate.

Figure 9.1 compares experimental product yields with maximal theoretical product yields. As it can be observed in the figure, *in silico* predictions differ by up to four orders of magnitude from the experimental target yield.

### 9.3. Kinetic-based Product Yield Prediction

A kinetic-based approach was used to calculate the so-called KAY [71]. KAY is defined as the maximal production flux that can be drawn from the host metabolic system before the metabolic system becomes unstable or the intracellular concentration of any metabolite in the system becomes negative. Dynamic system stability is evaluated by checking the eigenvalues of the system Jacobian matrix. Refer to [71] or Section 2.3.1 for details. Assuming that production performance of the strain under consideration is mostly constrained by kinetic stability issues, KAY should be a suitable proxy to describe observed experimental yields. The mathematical definition of KAY implies that its numerical value depends on the model kinetic structure used to simulate the host metabolism and on its associated parameters. Since both model structure and kinetic parameters are normally unknown, KAY can be calculated for an ensemble of kinetic models, built using the EM approach. Consequently, KAY will not have a unique numerical value but will rather be constrained into a range of probable values.

Here, we will assess the suitability of KAY to predict experimental product yield for six semi-batch data sets. In a first step, an ensemble of models describing a given reference flux distribution is built. Due to the process set-up used to generate experimental data,

## 9. Product Yield Prediction

growth and substrate uptake rates after system induction correspond to the most suitable experimental information to generate corresponding reference flux distributions by using linear optimization. For that, two different objective functions were minimized: oxygen uptake rate and citrate synthase activity. Both objective functions aim to describe a cellular phenotype with limited respiratory capacity. The rationale behind the selection of these objective functions is justified by experimental observations performed when analyzing heterologous gene expression in *E. coli*, as discussed in Chapter 4. For each reference flux distribution, an ensemble of 1000 kinetic models is built and KAY is calculated for each model in the ensemble. For each ensemble, calculated KAY values are characterized by 25th and 75th percentiles and a median value. The procedure can be summarized as follows:

1. Select one of the six data sets.
2. Constraint the *E. coli* core model by setting numerical values of upper and lower bounds of the growth and substrate uptake reactions to match experimentally measured rates. Refer to Table 8.2 for rate values.
3. Generate two reference flux distributions by minimizing either oxygen uptake rate or citrate synthase activity.
4. For each reference flux distribution, apply the EM approach to build an ensemble of 1000 models.
5. Calculate KAY values for each ensemble as described in Section 2.3.1.
6. Characterize results by providing 25th and 75th percentiles and a median KAY value.

Figure 9.2 shows product yield predictions via KAY for all data sets and two objective functions. For both objective functions, experimental product yields (see Table 9.1) are within calculated ranges for KAY values in four out of six data sets. However, incorrectly predicted product yields are different for each objective function. Considering both cases simultaneously, KAY has the *potential* to correctly predict five out of six experimental product yields. The maximal discrepancy between calculated KAY values and experimental product yields does not exceed one order of magnitude. By contrast, product yield prediction using maximal theoretical yield differed by up to four orders of magnitude from the experimental target yield.

### 9.4. Application of KAY to Predict Optimal Biomass Yield During Production

So far, the application of KAY has been limited to predict experimental product yields in engineered strains [71]. Here, we explore further applications by using the KAY concept to analyze the effect of biomass yield on production performance. This analysis was partly motivated by the dependence between product and biomass yield shown in Figure 9.3.A. There, a biomass yield seems to exist, for which product yield is maximal. Even though the effect of varying temperatures on the observed product-biomass yield behavior can not be completely ruled out, it is still interesting to compare *in silico* predictions made by both constraint- and kinetic-based approaches.

#### 9.4. Application of KAY to Predict Optimal Biomass Yield During Production

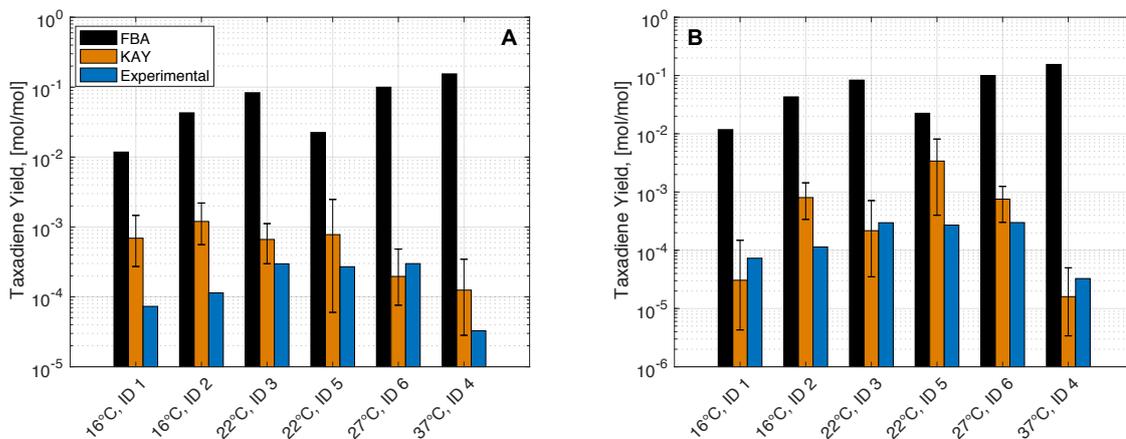


Figure 9.2.: **Kinetically Accessible Yield (KAY) for Experimental Product Yield Prediction.** **A.** KAY for minimization of oxygen uptake rate as objective function. **B.** KAY for minimization of citrate synthase activity as objective function.

We start our analysis by analyzing predictions made by the traditionally used constraint-based approach. We used a modified version of the *E. coli* core model for all calculations (refer to Appendix B.1.1), with glucose uptake rate set to 10 mmol/gDW h. Maximal theoretical product yields supported by the network were calculated while simultaneously considering cellular growth as an additional constraint. Growth rates ranged from 0 to  $0.6 \text{ h}^{-1}$ . Results are represented in Figure 9.3.B by the orange line. As shown in the Figure, the higher the growth rate (biomass yield), the lower the maximal theoretical yield. In other words, if product yield is to be maximized, cells should be prevented from growing. Such predictions will always be obtained when analyzing the effect of cellular growth on the theoretical maximal product yield using constraint-based methods. The rationale behind it is simple. Since constraint-based methods only considers mass balances, any flux increase of a competing reaction (growth in this case) will necessarily lead to a decrease in the production flux (and product yield), if the substrate uptake rate is kept constant. Comparing this prediction with experimental observation shown in Figure 9.3.A, we conclude that it qualitatively describes experimental data only after the maximal product yield is reached (for biomass yields higher than 0.2 g/g). While the constraint-based approach suggests that the maximal product yield should be reached at a biomass yield of zero (Figure 9.3.A), experimental maximal product yield is clearly reached at a non-zero biomass yield (Figure 9.3.B).

Now, we turn our attention to the application of KAY to predict the dependence of product yield on biomass yield. In order to keep predictions of both methods (constraint- and kinetic-based) comparable, we used the same *E. coli* core model to generate an ensemble of models, for which KAY is calculated. Reference flux distributions were generated by minimizing oxygen uptake rate, while keeping substrate uptake rate at a constant value of 10 mmol/gDW h. Growth rates were varied between 0 and  $0.6 \text{ h}^{-1}$ . Refer to Section 2.3.1 for details on mathematical and computational implementation of KAY. Results are represented in Figure 9.3.B by blue diamonds. There are three aspects in which the prediction of the product yield vs. biomass yield relationship using a kinetic-based approach fundamentally differs from the prediction made by the already discussed constraint-based approach. First, the order of magnitude of the calculated KAY values is dramatically lower

## 9. Product Yield Prediction

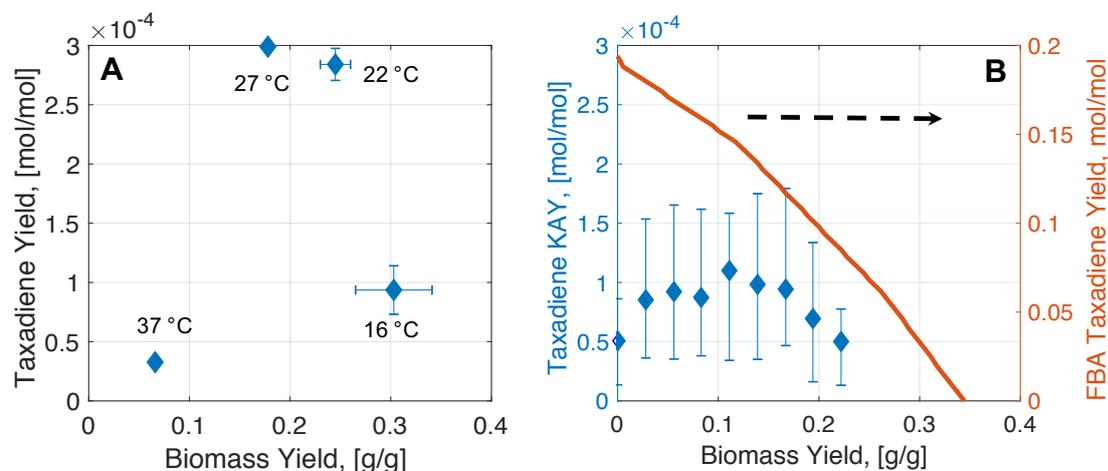


Figure 9.3.: **Dependence of Maximal Product Yield on Biomass Yield** Model used was a modified version of the *E. coli* core model. Glucose uptake rate was set to 10 mmol substrate/gDW h for all calculations. **A.** Experimentally observed dependence. **B.** *In silico* calculated dependence using a constraint based approach (orange line) and the KAY method (data points in blue). Error bars correspond to 25th and 75th percentiles. Diamonds correspond to median KAY of the respective ensemble.

than maximal theoretical product yields and is therefore more in line with experimental product yields (refer to Figure 9.3.A). Second, the kinetic-based approach predicts the maximal product yield to be located at a non-zero biomass yield of 0.11 g/g, surprisingly close to the location of the maximal experimental product yield 0.17 g/g, especially considering that no experimental information was used to generate reference flux distributions. Third, the predicted relationship between product yield and biomass yield using the kinetic-based method shows a non-linear trend, similar to the experimentally observed behavior. In contrast, the constraint-based method predicted a linear relationship between these two quantities. Overall, the consideration of a kinetic layer seems to improve the prediction capabilities of the stoichiometric model, even though exact reference flux distributions and corresponding parameter values were unknown.

## 9.5. Discussion

In Section 9.3, we successfully applied a kinetic-based approach to predict experimental product yields of a semi-batch process. Exchange rates (growth and substrate uptake rate) during production phase (after system induction) were used along with two different objective functions to generate reference flux distributions needed to construct an ensemble of models and subsequently calculate KAY values. We observed a dependence of calculated KAY values on the used reference flux distribution (see Figure 9.2). This highlights the importance of using appropriate reference flux distributions to obtain accurate product yield predictions. Since unambiguous determination of reference flux distributions is in many cases not possible due to technical or economical limitations (refer to Section 2.1.2), an interesting, yet unexplored possibility is the consideration of an ensemble of ensembles. Representative reference flux distributions (each used to populate one ensemble) can be generated by either assuming different (biologically relevant) objective functions, as done in

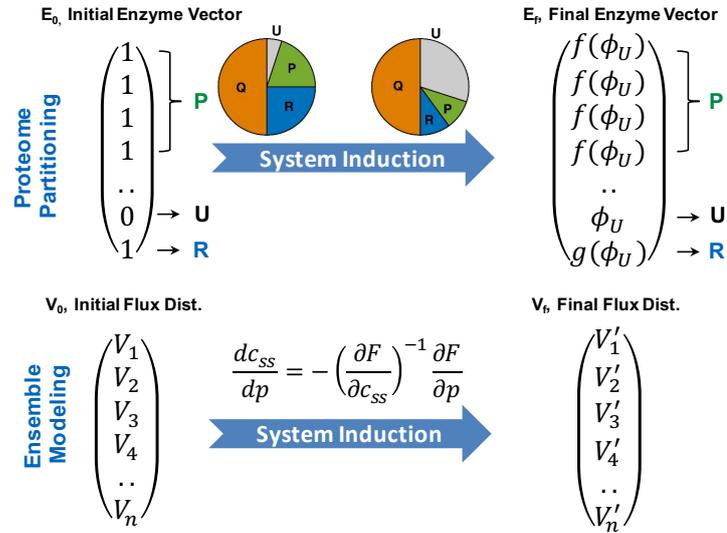


Figure 9.4.: **Increasing the Utility of the KAY Concept** The concept of proteome partitioning can be used to calculate the vector of enzyme concentrations after system (plasmid) induction ( $E_f$ ). By integrating Equation 2.25 between  $E_0$  and  $E_f$ , the vector of steady state metabolite concentrations can be calculated. This vector can be used in a subsequent step to calculate the corresponding flux distribution, which can be in turn used as input for a KAY analysis. Note that in the wild-type strain (pie chart on left side of this figure), the fraction of heterologous protein U amounts to zero. However, it is represented in the pie chart to symbolize that this fraction gains importance as the induction level is increased.

Section 9.3, or by simply covering to some extent the space of mathematically feasible flux distributions by sampling free (independent) fluxes (see Equation 2.18). In a second step, all ensembles are merged into one single ensemble and one single prediction is generated. Clearly, this approach would necessary require the implementation of computer clusters and an efficient parallelization strategy. The approach presented in Figure 2.6.B could be used as an initial cluster set-up.

Typically, experimental information regarding growth and substrate uptake rates during the production phase are unavailable. However, these data are normally known for un-induced strains or wild-types. In order to increase the utility of KAY calculations, the procedure used in Section 9.3 can be extended to allow for product yield prediction starting directly from the wild-type phenotype. In the original paper, Lafontaine Rivera et al. [71] used this idea. Starting from an estimated flux distribution for the wild-type, they successfully predicted the isobutanol yield for three different strains, one of which exhibited three different gene deletions ( $\Delta adhE$ ,  $\Delta pta$  and  $\Delta ldh$ ). In order to assess the production potential of that strain, the effect of gene deletions on the flux distribution was first assessed. Then, a KAY analysis was performed using only models that retained stability after mathematically knocking out the three genes. The calculated KAY value was consistent with the experimentally determined yield. A similar approach could be applied to estimate the production potential of highly burdened strains. The idea is simple and is motivated by the concept of proteome partitioning developed by the Hwa group at UC, San Diego [132–134]. Consider the pie chart shown on the left side of Figure 9.4. The proteome partitioning model introduced by Hwa and coworkers assumes that the microbial proteome can be split into three different sectors: Q, P and R. Sector Q is fixed and does not depend

## 9. Product Yield Prediction

on the cellular growth rate. R contains all ribosomal proteins and their affiliates, while P represents all proteins responsible for nutrient influx and metabolism. Since the size of the proteome is limited, the expression of an heterologous protein U effectively decreases the fraction allocable to the P and R sectors, as represented in the pie chart on the left side of Figure 9.4. If functions describing the fraction of sectors P ( $\phi_P$ ) and R ( $\phi_R$ ) in dependence of the heterologous fraction U ( $\phi_U$ ) are known, then it should be possible to estimate changes in the flux distribution as  $\phi_U$  increases. For a given  $\phi_U$  fraction, KAY values can be then calculated using the calculated flux distribution as a starting point ( $V_f$  in Figure 9.4) [91].

The results and outlook presented in this chapter evidences the huge potential of kinetic-based modeling, more specifically, of EM-based methods. In our opinion, Figures 9.3 and 9.2 directly support the predictive power of EM-based methods like KAY itself and optEM (see Section 5.4.2). Additional experimental work, including measurements of protein content in both wild-type and induced strains, will be required to further extend and validate the applicability of KAY as proposed in Figure 9.4. An interesting prediction of KAY is the interdependence between biomass and product yield, shown in this Chapter (see Figure 9.3) but also in Chapter 6 (refer to Figure 6.10.B). The validity of this relationship and the ability of KAY to successfully describe it could be exploited not only for the optimal design of fermentation processes but also for the calculation of optimal induction levels, described by the fraction of heterologous protein  $\phi_U$ .

## 10. Concluding Remarks and Outlook

The coverage of this thesis is mostly limited to the taxadiene production in *E. coli*. However, a review on the application of computational methods to increase succinate production in engineered strains evidenced interesting trends of the current model-based, *rational* Metabolic Engineering. Even though the details of that review can be found elsewhere [149], we will highlight some of the key lessons learned from that review in order to understand the way how it motivated and guided parts of the *in silico* studies presented here. Then, we will provide some conclusions and future perspectives concerning the main results of this work.

### 10.1. Lessons From Succinate Production in Engineered Strains

Biotechnological production of succinate has been one of the preferred case studies to test predictions made by strain-design algorithms. Since the introduction of OptKnock in 2003 and until 2016, a total of 26 different strain-design algorithms *in silico* targeting the succinate overproduction (refer to Table 1 in [149] or Table 2.4 of this work) have been published. None of these studies was accompanied by an *in vivo* validation of the algorithm's predictions. Rather, the rich biological knowledge available, gained over decades of research in succinate production in a variety of industrially relevant microorganisms, was used to test the predictive power of the algorithms. An analysis of the nature of mathematical methods used to develop these strain-design algorithms and the way how their predictions are used to guide *in vivo* strain development leads to three key conclusions:

1. *Constraint-based methods predominate.* Roughly 80% of the 26 strain-design algorithms targeting succinate overproduction have a constraint-based nature. This observation is not surprising and is, in our opinion, mainly justified in the relative simplicity of its mathematical formulation, availability of the required information (a genome-scale metabolic reconstruction and an objective function), and its computational tractability.
2. *Kinetic-based methods are underrepresented.* Contrasting the predominance of constraint-based methods, only one single method, out of 26 strain design algorithms reviewed, had a kinetic-based nature. In our opinion, this evidences the difficulties associated with kinetic descriptions of cellular metabolism. In addition to a stoichiometric layer, kinetic information in form of rate laws, describing the flux through each reaction as a function of enzyme-specific parameters and metabolite concentrations, are also required. Since this information is typically unavailable, the construction of large-scale kinetic models have been hindered.

## 10. Concluding Remarks and Outlook

3. *In silico* predictions are not always precise. We found out that only 10 studies reported the model-guided construction of strains exhibiting an enhanced succinate production (refer to Table 2 in [149]). Within these studies, only two used one of the 26 strain-design algorithms previously developed. All other studies reported the construction of production strains based on knowledge gained from FBA calculations. Additionally, none of the 10 studies reported an exact implementation of *in silico* predictions. Often, complementary strategies like directed evolution were necessary to achieve high succinate yields. Altogether, these observations point to a limited predictive power of constraint-based methods, especially when aspects like fine tuning of enzyme expression or dynamic stability of the system become important factors determining the effectiveness of a given genetic intervention [1, 2].

### 10.2. Lessons From Taxadiene Production in *E. coli*

As previously mentioned, advantages of constraint-based methods over kinetic-based methods include the simplicity of use, while simultaneously offering a moderate predictive power. Due to their computational tractability, constraint-based simulations can be performed on a large-scale, thus allowing the identification of non-trivial network properties. In the specific case of taxadiene production in *E. coli*, we demonstrated that a carbon-neutral taxadiene production is not possible for any substrate or substrate mixture. This observation can be easily extended to the production any terpene. This is because carbon leak in form of CO<sub>2</sub> is concomitant to the formation of both IPP and DMAPP, the universal terpene precursors, via either the mevalonate or the non-mevalonate pathway. Additionally, our large-scale constraint-based analysis pointed at a synergy between the mevalonate and the non-mevalonate pathway *in silico* (**Chapter 7**). Remarkably, this synergy was experimentally observed, although in a greater extent, for the production of isoprene in *E. coli* [160].

The predictive power of constraint-based modeling is increased, if biologically meaningful constraints are used. In **Chapter 4**, we applied a constraint-based approach to simulate the experimentally observed effect of protein load on by-product secretion. By reducing the upper boundary of oxygen uptake rate as protein/plasmid load increases, we could qualitatively reproduce the downwards shift of the acetate line experimentally observed in loaded strains. Our assumption on decreasing respiratory capacity is in part justified by proteomic measurements reported by Peebo et al. [110].

One of the main disadvantages of constraint-based methods, which paradoxically has its origin in the simplicity of the mathematical formulation of constraint-based approaches, is the omission of enzyme kinetics. As a consequence, important aspects such as dynamic stability of the cellular metabolism, intracellular concentration of potentially toxic metabolites, metabolite-level enzyme regulation and fine tuning of enzyme expression cannot be adequately considered. By contrast, multiple studies have highlighted the importance of at least one of these factors for the construction of terpenoids overproducing strains [1, 2, 38]. Motivated by this and by the fact that kinetic-based models are underrepresented despite of their application potential in rational Metabolic Engineering, we ventured to explore the utility of the ensemble modeling, an emerging approach for the construction and analysis of kinetic models. The predictive power of the ensemble modeling approach was evaluated

in **Chapter 5** and **9**. Even though kinetic parameters were sampled and reference flux distributions were calculated by means of constraint-based modeling using two experimental exchange rates, namely growth and substrate uptake rate, predicted taxadiene yields using the concept of KAY [71] agreed remarkably well with experimental data (**Chapter 9**). Additionally, predicted targets by our EM-based strain-design algorithm optEM for increased taxadiene production were in line with known, reported beneficial genetic perturbations (**Chapter 5**). We also observed an interesting dependence of experimental taxadiene yields on biomass yields, which, to the best of our knowledge, has not been described before. Interestingly, this dependence could be reproduced using the concept of KAY. In **Chapter 6**, we observed the same dependency for strains exhibiting varying levels of % Carbon Xylose. We combined these predictions and the fact that biomass yield is likely to decrease (as genetic elements necessary for taxadiene production should increase the metabolic burden) to suggest that strain HMS p1x represents the best production strain candidate for further optimization rounds.

Even though the ensemble modeling approach showed an unexpectedly high predictive power and utility, its widespread in the community has been hindered, possibly by its high computational burden. Clearly, this represents a challenge for small research facilities, as computer clusters do not belong to standard equipment. As discussed in **Chapter 5**, computational power can be reduced by increasing the amount of experimental information used to define and characterize the reference state required to populate the ensemble of kinetic models. A related issue is the existence of alternative flux distributions equally fulfilling constraints imposed by experimentally determined exchange rates or available  $^{13}\text{C}$ -labeling data. The dependence of KAY predictions on the reference flux distribution is exemplary shown in Figure 9.2, **Chapter 9**. An option to deal with this additional uncertainty source is to sample the space of possible flux distributions and to perform the desired ensemble modeling-based analysis (KAY [71], EMRA [74, 144] or OptEM) independently. Then, the collective response can be characterized as usual, by providing 25th and 75th percentiles for the whole ensemble of ensembles. Certainly, this imposes an even higher computational burden. Therefore, an optimal trade-off between uncertainty and degree of experimental characterization should be individually found according to equipment availability.

We envision the spectrum of ensemble modeling-related methods and applications to keep widening. For instance, KAY could be used as a soft sensor, calculated in real-time from online OD, substrate and oxygen concentration measurements. This is especially useful when online experimental determination of product concentrations is not possible. By using KAY as an accurate proxy for strain performance, it should be possible to dynamically adjust process conditions such as stirrer velocity, substrate feeding and process time to maximize production and profit. Further potential applications include combination of the ensemble modeling approach with dynamic flux balance analysis for identification of strain designs with increased productivity [170] and guidance while experimentally constructing instability-prone genotypes [2].

High quality, reproducible experimental data play a crucial role in the efficient development and validation of modeling approaches and more broadly, in the advance of science. Nowadays, reproducibility issues do not represent the incapacity of isolated research groups to reproduce own or other group's results, but rather a generalized problem affecting the whole research community, with an extent depending on the specific research field [9,99]. In

## 10. Concluding Remarks and Outlook

this work, model-based improvement strategies for strains (**Chapter 4 & 5**) and a production process (**Chapter 8**) were identified. Unfortunately and due to reproducibility issues, none of these strategies could be validated. This limited the experimental implementation of the theoretical workflow developed (refer to **Chapter 3**) and hindered the desired optimization of strains and their associated production processes through the implementation of model-based predictions in an iterative fashion. However, various theoretical methods for each of the steps of the workflow were developed and applied. Figure 10.1 provides an overview of these methods and summarizes relevant input-output relationships used along this work.

**Chapter 6** was concluded by suggesting strain HMS p1x as an optimal production strain candidate. After genetic elements allowing terpenoid production are introduced into this strain, modifications in its metabolism leading to an increased production performance can be identified by applying the tools discussed in **Chapter 5**. Once the desired production performance has been reached, associated optimal process parameters can be identified as shown in **Chapter 8**. As for culture medium composition, a similar procedure to the one used in **Chapter 7** can be applied to identify promising carbon sources. These steps are illustrated in Figure 10.1 by the dashed lines and represent one possible implementation of the workflow using all tools described in this work.

Two opposite approaches have been suggested to increase reproducibility of results, namely standardization [111] and heterogenization of experimental conditions [116]. In either case, future experimental efforts should aim to keep increasing confidence in experimental data. On the modeling side, strategies should be developed to properly handle uncertainties and experimental error, thus allowing the generation of predictions that are robust to noise.

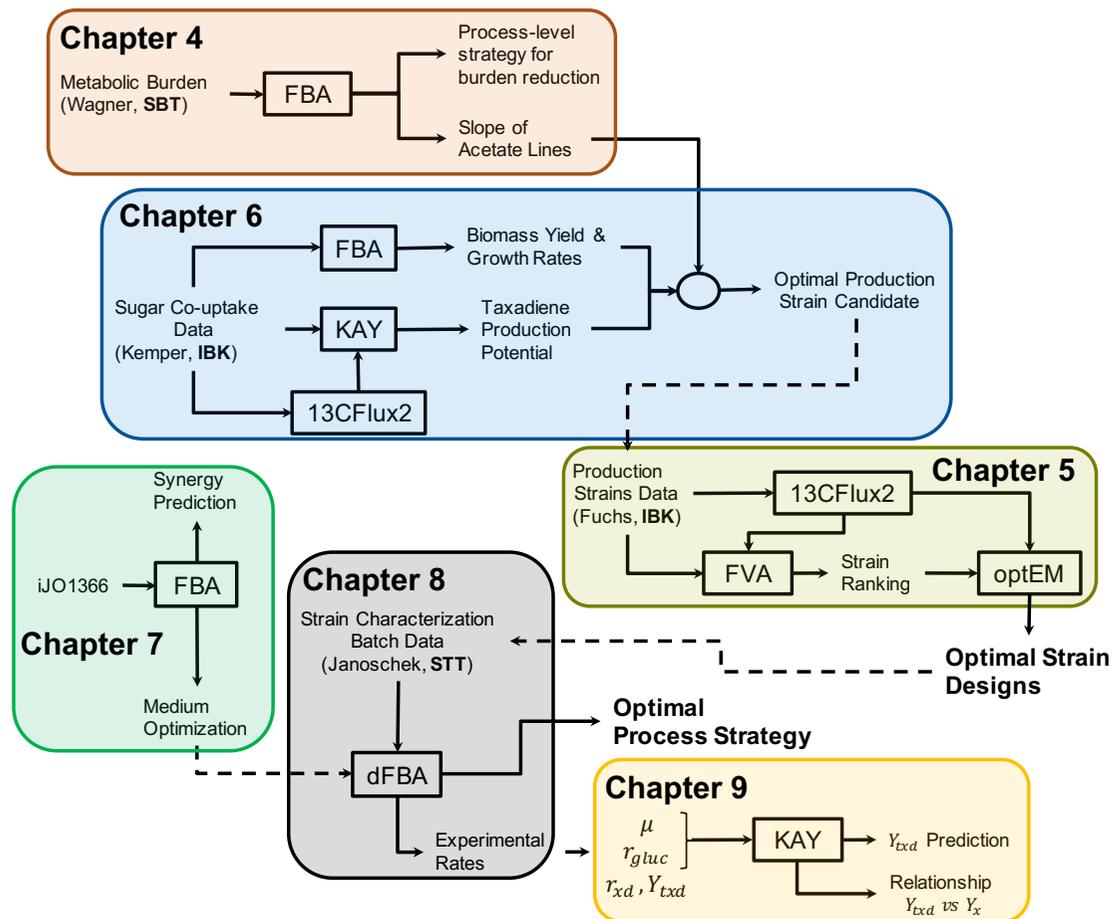


Figure 10.1.: **Application of the Workflow to the Production of Taxadiene in *E. coli*: Conclusion.** The proposed workflow, covering both model-guided strain engineering and model-driven process optimization, was applied to the production of taxadiene in *E. coli*. Even though model-based optimization strategies for strain and production process could not be experimentally validated due to reproducibility issues of experimental data, we demonstrated the utility of a variety *in silico* approaches. Refer to each chapter for more details. Main input-output relationships are shown for each chapter. Applied methods are given in boxes. Dashed lines represent further steps towards the construction of an optimal terpenoid producing strain and the optimization of its performance in a bio-reactor.



# Bibliography

- [1] P. K. Ajikumar, W.-H. Xiao, K. E. J. Tyo, Y. Wang, F. Simeon, E. Leonard, O. Mucha, T. H. Phon, B. Pfeifer, and G. Stephanopoulos. Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science (New York, N. Y.)*, 330(6000):70–4, oct 2010.
- [2] J. Alonso-Gutierrez, D. Koma, Q. Hu, Y. Yang, L. J. G. Chan, C. J. Petzold, P. D. Adams, C. E. Vickers, L. K. Nielsen, J. D. Keasling, and T. S. Lee. Towards industrial production of isoprenoids in *Escherichia coli* : lessons learned from CRISPR-Cas9 based optimization of a chromosomally integrated mevalonate pathway. *Biotechnology and Bioengineering*, dec 2017.
- [3] R. Alves and M. A. Savageau. Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. *Bioinformatics (Oxford, England)*, 16(6):534–47, jun 2000.
- [4] S. Andreozzi, L. Miskovic, and V. Hatzimanikatis. ISCHRUNK - In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. *Metabolic Engineering*, 33:158–168, 2016.
- [5] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metabolic Engineering*, 8(4):324–337, 2006.
- [6] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metabolic Engineering*, 9(1):68–86, 2007.
- [7] J. Bailey. Toward a science of metabolic engineering. *Science*, 252(5013):1668–1675, 1991.
- [8] J. E. Bailey. Complex biology with no parameters. *Nature Biotechnology*, 19(6):503–4, jun 2001.
- [9] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, may 2016.
- [10] A. Bar-Even, E. Noor, N. E. Lewis, and R. Milo. Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8889–94, 2010.
- [11] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson, and T. Hwa. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature*, 528(7580):99–104, 2015.
- [12] P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.
- [13] Q. K. Beg, A. Vazquez, J. Ernst, M. A. de Mezees, Z. Bar-Joseph, A.-L. Barabási, and Z. N. Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12663–8, jul 2007.
- [14] K. Bettenbrock, S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E.-D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *The Journal of Biological Chemistry*, 281(5):2578–84, feb 2006.
- [15] S. Birnbaum and J. E. Bailey. Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnology and Bioengineering*, 37(8):736–745, apr 1991.
- [16] B. A. Boghigian, J. Armando, D. Salas, and B. A. Pfeifer. Computational identification of gene overexpression targets for metabolic engineering of taxadiene production. *Applied Microbiology and Biotechnology*, 93(5):2063–73, mar 2012.
- [17] I. W. Bogorad, T.-S. Lin, and J. C. Liao. Synthetic non-oxidative glycolysis enables complete carbon conservation. *Nature*, 502(7473):693–697, oct 2013.
- [18] K. Bohl, L. F. de Figueiredo, O. Hadicke, S. Klamt, C. Kost, S. Schuster, and C. Kaleta. CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks. *GCB 2010 - German Conference on Bioinformatics 2010 (2010)*, pages 71–80, 2010.
- [19] J. Büchs. Introduction to advantages and problems of shaken cultures. In *Biochemical Engineering Journal*, volume 7, pages 91–98, 2001.
- [20] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–657, dec 2003.
- [21] A. Chakrabarti, L. Miskovic, K. C. Soh, and V. Hatzimanikatis. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnology Journal*, 8(9):1043–1057, 2013.

## Bibliography

- [22] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss. Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1):53–73, jul 2002.
- [23] K. K. Cheung. A review of ensemble forecasting techniques with a focus on tropical cyclone forecasting. *Meteorological Applications*, 8(3):315–332, 2001.
- [24] H. S. Choi, S. Y. Lee, T. Y. Kim, and H. M. Woo. In Silico Identification of Gene Amplification Targets for Improvement of Lycopene Production. *Applied and Environmental Microbiology*, 76(10):3097–3105, mar 2010.
- [25] Y. Choon, M. Mohamad, S. Deris, C. Chong, L. Chai, Z. Ibrahim, and S. Omatu. Identifying gene knockout strategies using a hybrid of bees algorithm and flux balance analysis for in silico optimization of microbial strains. In S. Omatu, J. F. De Paz Santana, S. R. González, J. M. Molina, A. M. Bernardos, and J. M. C. Rodríguez, editors, *Distributed Computing and Artificial Intelligence*, volume 151 of *Advances in Intelligent and Soft Computing*, pages 371–378. Springer Berlin Heidelberg, 2012.
- [26] Y. W. Choon, M. S. Mohamad, S. Deris, C. K. Chong, S. Omatu, and J. M. Corchado. Gene knockout identification using an extension of Bees Hill Flux Balance Analysis. *BioMed Research International*, 2015:124537, 2015.
- [27] Y. W. Choon, M. S. Mohamad, S. Deris, R. M. Ilias, C. K. Chong, L. E. Chai, S. Omatu, and J. M. Corchado. Differential Bees Flux Balance Analysis with OptKnock for *in silico* microbial strains optimization. *PLoS One*, 9(7):e102744, jan 2014.
- [28] A. Chowdhury, A. R. Zomorodi, and C. D. Maranas. k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Computational Biology*, 10(2):e1003487, feb 2014.
- [29] B. CLARK and W. H. HOLMS. Control of the Sequential Utilization of Glucose and Fructose by *Escherichia coli*. *Journal of General Microbiology*, 95(2):191–201, 1976.
- [30] J. Costanza, G. Carapezza, C. Angione, P. Lió, and G. Nicosia. Robust design of microbial strains. *Bioinformatics (Oxford, England)*, 28(23):3097–104, dec 2012.
- [31] M. W. Covert, I. Famili, and B. O. Palsson. Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnology and Bioengineering*, 84(7):763–72, dec 2003.
- [32] M. W. Covert, C. H. Schilling, and B. Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213(1):73–88, nov 2001.
- [33] J. Deutscher. The mechanisms of carbon catabolite repression in bacteria. *Current Opinion in Microbiology*, 11(2):87–93, 2008.
- [34] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. COBRApy: CONstraints-Based Reconstruction and Analysis for Python. *BMC systems biology*, 7(1):74, jan 2013.
- [35] D. Egen and D. S. Lun. Truncated branch and bound achieves efficient constraint-based genetic design. *Bioinformatics (Oxford, England)*, 28(12):1619–23, jul 2012.
- [36] H. Eyring and J. Magee. Application of the theory of absolute reaction rates to bacterial luminescence. *J Cell Comp Physiol*, 20(2):169–177, oct 1942.
- [37] N. Farke. Ensemble Modeling Approach for Metabolic Engineering: Optimising the diterpene production in *Escherichia coli*. *Master Thesis. Technical University of Munich.*, 2017.
- [38] W. R. Farmer and J. C. Liao. Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nature Biotechnology*, 18(5):533–7, may 2000.
- [39] W. R. Farmer and J. C. Liao. Precursor balancing for metabolic engineering of lycopene production in *Escherichia coli*. *Biotechnology Progress*, 17(1):57–61, feb 2001.
- [40] A. Farwick, S. Bruder, V. Schadeweg, M. Oreb, and E. Boles. Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. *Proceedings of the National Academy of Sciences*, 111(14):5159–5164, 2014.
- [41] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3:121, jan 2007.
- [42] S. S. Fong, A. P. Burgard, C. D. Herring, E. M. Knight, F. R. Blattner, C. D. Maranas, and B. O. Palsson. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering*, 91(5):643–8, oct 2005.
- [43] D. Frense. Taxanes: Perspectives for biotechnological production. *Applied Microbiology and Biotechnology*, 73(6):1233–1240, jan 2007.
- [44] J. García, G. Wagner, L. Sörnmo, S. Olmos, P. Lander, and P. Laguna. Characterization of sugar mixtures utilization by an *Escherichia coli* mutant devoid of the phosphotransferase system. *Applied Microbiology and Biotechnology*, 57(1-2):186–191, 2001.
- [45] P. Gawand, P. Hyland, A. Ekins, V. J. Martin, and R. Mahadevan. Novel approach to engineer strains for simultaneous sugar utilization. *Metabolic Engineering*, 20:63–72, 2013.
- [46] M. P. Gerstl, C. Jungreuthmayer, and J. Zanghellini. tEFMA: computing thermodynamically feasible elementary flux modes in

- metabolic networks. *Bioinformatics (Oxford, England)*, 31(13):2232–4, jul 2015.
- [47] J. A. Gomez, K. Höffner, and P. I. Barton. DF-BALab: A fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinformatics*, 15(1), 2014.
- [48] S. Gopalakrishnan and C. D. Maranas. <sup>13</sup>C metabolic flux analysis at a genome-scale. *Metabolic Engineering*, 32:12–22, 2015.
- [49] B. Görke and J. Stülke. Carbon catabolite repression in bacteria: Many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6(8):613–624, 2008.
- [50] E. Gruchattka, O. Hädicke, S. Klamt, V. Schütz, and O. Kayser. *In silico* profiling of *Escherichia coli* and *Saccharomyces cerevisiae* as terpenoid factories. *Microbial cell factories*, 12(1):84, jan 2013.
- [51] O. Hädicke and S. Klamt. CASOP: A Computational Approach for Strain Optimization aiming at high Productivity. *Journal of Biotechnology*, 147:88–101, 2010.
- [52] B. Hahn-Hägerdal, C. F. Wahlbom, M. Gárdonyi, W. H. van Zyl, R. R. Cordero Otero, and L. J. Jönsson. Metabolic engineering of *Saccharomyces cerevisiae* for xylose utilization. *Advances in biochemical engineering/biotechnology*, 73:53–84, 2001.
- [53] J. J. Hamilton, V. Dwivedi, and J. L. Reed. Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. *Biophysical Journal*, 105(2):512–22, jul 2013.
- [54] M. Harker and P. M. Bramley. Expression of prokaryotic 1-deoxy-d-xylulose-5-phosphatases in *Escherichia coli* increases carotenoid and ubiquinone biosynthesis. *FEBS Letters*, 448(1):115–119, 1999.
- [55] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, I. Thiele, J. Orth, A. Feist, D. Zielinski, A. Bordbar, N. Lewis, S. Rahmanian, J. Kang, B. Palsson, and R. M. Fleming. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. *Quantitative Methods*, page arXiv:1710.04038, 2017.
- [56] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical Journal*, 92(5):1792–1805, 2007.
- [57] F. H. Johnson and I. Lewin. The growth rate of *E. coli* in relation to temperature, quinine and coenzyme. *Journal of Cellular and Comparative Physiology*, 28(1):47–75, aug 1946.
- [58] K. L. Jones, S. W. Kim, and J. D. Keasling. Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metabolic engineering*, 2(4):328–338, 2000.
- [59] H. R. Kaback. Regulation of sugar transport in isolated bacterial membrane preparations from *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 63:724–731, 1969.
- [60] P. Karp. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Research*, 24:32–39, 1996.
- [61] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñoz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1), 2013.
- [62] A. Khodayari, A. Chowdhury, and C. D. Maranas. Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive *Escherichia coli* Kinetic Model. *Frontiers in Bioengineering and Biotechnology*, 2(January), 2015.
- [63] A. Khodayari, A. R. Zomorodi, J. C. Liao, and C. D. Maranas. A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metabolic Engineering*, 25:50–62, sep 2014.
- [64] J. Kim, J. L. Reed, and C. T. Maravelias. Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. *PLoS One*, 6(9):e24162, jan 2011.
- [65] Z. A. King and A. M. Feist. Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains OptSwap. *Industrial Biotechnology*, 9(4):236–246, aug 2013.
- [66] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1):D515–D522, 2016.
- [67] S. Klamt, J. Saez-Rodriguez, and E. Gilles. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1(1):2, 2007.
- [68] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2):64–9, feb 2003.
- [69] O. Kotte, J. B. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Molecular Systems Biology*, 6, 2010.
- [70] J. G. Lafontaine Rivera, Y. Lee, and J. C. Liao. An entropy-like index of bifurcational robustness for metabolic systems. *Integrative Biology: quantitative biosciences from nano to macro*, 7(8):895–903, aug 2015.

## Bibliography

- [71] J. G. Lafontaine Rivera, M. K. Theisen, P. W. Chen, and J. C. Liao. Kinetically accessible yield (KAY) for redirection of metabolism to produce exo-metabolites. *Metabolic Engineering*, 41:144–151, may 2017.
- [72] M. Lakshmanan, T. Y. Kim, B. K. S. Chung, S. Y. Lee, and D.-Y. Lee. Flux-sum analysis identifies metabolite targets for strain improvement. *BMC Systems Biology*, 9(1):73, dec 2015.
- [73] S. Y. Lee. *Systems Biology and Biotechnology of Escherichia coli*. Springer, 2009.
- [74] Y. Lee, J. G. Lafontaine Rivera, and J. C. Liao. Ensemble Modeling for Robustness Analysis in engineering non-native metabolic pathways. *Metabolic Engineering*, 25:63–71, sep 2014.
- [75] R. W. Leighty and M. R. Antoniewicz. COMPLETE-MFA: Complementary parallel labeling experiments technique for metabolic flux analysis. *Metabolic Engineering*, 20:49–55, nov 2013.
- [76] J. C. Liao, S. Y. Hou, and Y. P. Chao. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology and Bioengineering*, 52(1):129–40, oct 1996.
- [77] W. Liebermeister and E. Klipp. Biochemical networks with uncertain parameters. *Systems Biology*, 152(3):97–107, sep 2005.
- [78] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theoretical Biology & Medical Modelling*, 3:41, 2006.
- [79] W. Liebermeister, J. Uhlenendorf, and E. Klipp. Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics (Oxford, England)*, 26(12):1528–34, jun 2010.
- [80] F. Lipmann. Metabolic Generation and Utilization of Phosphate Bond Energy. In *Advances in Enzymology and Related Areas of Molecular Biology*, pages 99–162. John Wiley & Sons, Inc., nov 2006.
- [81] L. M. Lois, N. Campos, S. R. Putra, K. Danielsen, M. Rohmer, and A. Boronat. Cloning and characterization of a gene from *Escherichia coli* encoding a transketolase-like enzyme that catalyzes the synthesis of D-1-deoxyxylulose 5-phosphate, a common precursor for isoprenoid, thiamin, and pyridoxol biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2105–2110, mar 1998.
- [82] J. G. Lomnitz and M. A. Savageau. Design space toolbox V2: Automated software enabling a novel phenotype-centric modeling strategy for natural and synthetic biological systems. *Frontiers in Genetics*, 7(JUL), 2016.
- [83] C. P. Long, J. E. Gonzalez, N. R. Sandoval, and M. R. Antoniewicz. Characterization of physiological responses to 22 gene knockouts in *Escherichia coli* central carbon metabolism. *Metabolic Engineering*, 37:102–113, 2016.
- [84] L. Lopez Pineda. *In-silico Process Optimization of the Taxoid Production by Escherichia coli*. Master Thesis. Technical University of Munich., 2016.
- [85] D. S. Lun, G. Rockwell, N. J. Guido, M. Baym, J. A. Kelner, B. Berger, J. E. Galagan, and G. M. Church. Large-scale identification of genetic design strategies using local search. *Molecular Systems Biology*, 5:296, jan 2009.
- [86] L. R. Lynd, C. E. Wyman, and T. U. Gerngross. Biocommodity engineering. *Biotechnology Progress*, 15(5):777–793, 1999.
- [87] D. Machado and M. Herrgård. Co-evolution of strain design methods based on flux balance and elementary mode analysis. *Metabolic Engineering Communications*, 2:85–92, may 2015.
- [88] R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–76, oct 2003.
- [89] P. Maia, M. Rocha, and I. Rocha. *In Silico Constraint-Based Strain Optimization Methods: the Quest for Optimal Cell Factories*. *Microbiology and Molecular Biology Reviews*, 80(1):45–67, 2016.
- [90] S. E. Mainguet, L. S. Gronenberg, S. S. Wong, and J. C. Liao. A reverse glyoxylate shunt to build a non-native route from C4 to C2 in *Escherichia coli*. *Metabolic Engineering*, 19:116–127, 2013.
- [91] T. Mainka. Model-driven Process Optimization of the Taxoid Production with *Escherichia coli*. Master Thesis. Technical University of Munich., 2017.
- [92] L. Mao and W. S. Verwoerd. ORCA: A COBRA toolbox extension for model-driven discovery and analysis. *Bioinformatics*, 30(4):584–585, 2014.
- [93] V. J. J. Martin, D. J. Pitera, S. T. Withers, J. D. Newman, and J. D. Keasling. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature Biotechnology*, 21(7):796–802, 2003.
- [94] T. Maskow and U. von Stockar. How reliable are thermodynamic feasibility statements of biochemical pathways? *Biotechnology and bioengineering*, 92(2):223–30, oct 2005.
- [95] H. Meng, Y. Wang, Q. Hua, S. Zhang, and X. Wang. *In silico* analysis and experimental improvement of taxadiene heterologous biosynthesis in *Escherichia coli*. *Biotechnology and Bioprocess Engineering*, 16(2):205–215, apr 2011.
- [96] F. Molteni, R. Buizza, T. N. Palmer, and T. Petroligis. The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119, 1996.

- [97] J. M. Monk, C. J. Lloyd, E. Brunk, N. Mih, A. Sastry, Z. King, R. Takeuchi, W. Nomura, Z. Zhang, H. Mori, A. M. Feist, and B. O. Palsson. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nature Biotechnology*, 35(10):904–908, 2017.
- [98] D. Morrone, L. Lowry, M. K. Determan, D. M. Hershey, M. Xu, and R. J. Peters. Increasing diterpene yield with a modular metabolic engineering system in *E. coli*: Comparison of MEV and MEP isoprenoid precursor pathway engineering. *Applied Microbiology and Biotechnology*, 85(6):1893–1906, 2010.
- [99] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, jan 2017.
- [100] N. N. Nichols, B. S. Dien, and R. J. Bothast. Use of catabolite repression mutants for fermentation of sugar mixtures to ethanol. *Applied Microbiology and Biotechnology*, 56(1-2):120–125, 2001.
- [101] K. C. Nicolaou, Z. Yang, J. J. Liu, H. Ueno, P. G. Nantermet, R. K. Guy, C. F. Claiborne, J. Renaud, E. A. Couladouros, K. Paulvannan, and E. J. Sorensen. Total synthesis of taxol. *Nature*, 367(6464):630–634, feb 1994.
- [102] E. J. O’Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. Ø. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, 9(1):693, jan 2013.
- [103] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, 1999.
- [104] Y.-G. Oh, D.-Y. Lee, S. Y. Lee, and S. Park. Multiobjective flux balancing using the NISE method for metabolic network analysis. *Biotechnology Progress*, 25(4):999–1008, jan 2009.
- [105] S. Ohno, H. Shimizu, and C. Furusawa. FastPros: screening of reaction knockout strategies for metabolic engineering. *Bioinformatics (Oxford, England)*, 30(7):981–7, apr 2014.
- [106] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Molecular Systems Biology*, 7:535, jan 2011.
- [107] J. D. Orth, B. Ø. Palsson, and R. M. T. Fleming. Reconstruction and Use of Microbial Metabolic Networks: the Core *Escherichia coli* Metabolic Model as an Educational Guide. *EcoSal Plus*, 4(1), sep 2010.
- [108] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen. Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics*, 6(1):308, jan 2005.
- [109] R. Patnaik, W. D. Roof, R. F. Young, and J. C. Liao. Stimulation of glucose catabolism in *Escherichia coli* by a potential futile cycle. *Journal of Bacteriology*, 174(23):7527–7532, 1992.
- [110] K. Peebo, K. Valgepea, A. Maser, R. Nahku, K. Adamberg, and R. Vilu. Proteome reallocation in *Escherichia coli* with increasing specific growth rate. *Mol. BioSyst.*, 11(4):1184–1193, 2015.
- [111] P. Phillips, G. J. Lithgow, and M. Driscoll. A long journey to reproducible results. *Nature*, 548(7668):387–388, aug 2017.
- [112] S. Ranganathan, P. F. Suthers, and C. D. Maranas. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Computational Biology*, 6(4):e1000744, apr 2010.
- [113] D. A. Ratkowsky, J. Olley, T. A. McMeekin, and A. Ball. Relationship between temperature and growth rate of bacterial cultures. *Journal of Bacteriology*, 149(1):1–5, jan 1982.
- [114] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology*, 4(9):R54, jan 2003.
- [115] S. Ren, B. Zeng, and X. Qian. Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints. *BMC Bioinformatics*, 14 Suppl 2(Suppl 2):S17, jan 2013.
- [116] S. H. Richter, J. P. Garner, C. Auer, J. Kunert, and H. Würbel. Systematic variation improves reproducibility of animal experiments. *Nature Methods*, 7(3):167–168, mar 2010.
- [117] Y. L. Sang, D. Y. Lee, and Y. K. Tae. Systems biotechnology for strain improvement. *Trends in Biotechnology*, 23(7):349–358, 2005.
- [118] M. A. Savageau. Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 25(3):365–369, 1969.
- [119] M. A. Savageau. Biochemical systems analysis: II. The steady-state solutions for an n-pool system using a power-law approximation. *Journal of Theoretical Biology*, 25(3):370–379, 1969.
- [120] M. A. Savageau. Biochemical systems analysis: III. Dynamic solutions using a power-law approximation. *Journal of Theoretical Biology*, 26(2):215–226, 1970.
- [121] M. A. Savageau. A theory of alternative designs for biochemical control systems. *Biomed. Biochim. Acta*, 44(6):875–80, 1985.
- [122] J. Schellenberger and B. Ø. Palsson. Use of randomized sampling for analysis of metabolic networks. *The Journal of Biological Chemistry*, 284(9):5457–61, feb 2009.

## Bibliography

- [123] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, jan 2010.
- [124] J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. Ø. Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, 6(9):1290–307, sep 2011.
- [125] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229–48, apr 2000.
- [126] K. Schmidt, M. Carlsen, J. Nielsen, and J. Villadsen. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and Bioengineering*, 55(6):831–840, 1997.
- [127] K. Schmidt, A. Marx, de Graaf AA, W. Wiechert, H. Sahm, J. Nielsen, and J. Villadsen. <sup>13</sup>C tracer experiments and metabolite balancing for metabolic flux analysis: comparing two approaches. *Biotechnology and Bioengineering*, 58(2-3):254–7, apr 1998.
- [128] R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology*, 3:119, jan 2007.
- [129] S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53–60, feb 1999.
- [130] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 02(02):165–182, jun 1994.
- [131] R. Schwarz, P. Musch, A. von Kamp, B. Engels, H. Schirmer, S. Schuster, and T. Dandekar. YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6(1):135, jan 2005.
- [132] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(6007):1099–1102, 2010.
- [133] M. Scott and T. Hwa. Bacterial growth laws and their applications. *Current Opinion in Biotechnology*, 22(4):559–565, 2011.
- [134] M. Scott, S. Klumpp, E. M. Mateescu, and T. Hwa. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Molecular Systems Biology*, 10(8):747–747, 2014.
- [135] C. Sehr, A. Kremling, and A. Marin-Sanguino. Design Principles as a Guide for Constraint Based and Dynamic Modeling: Towards an Integrative Workflow. *Metabolites*, 5(4):601–35, oct 2015.
- [136] C. R. Shen and J. C. Liao. Synergy as design principle for metabolic engineering of 1-propanol production in *Escherichia coli*. *Metabolic Engineering*, 17:12–22, may 2013.
- [137] T. Shirai, T. Osanai, and A. Kondo. Designing intracellular metabolism for production of target compounds by introducing a heterologous metabolic reaction based on a *Synechosystis* sp. 6803 genome-scale model. *Microbial Cell Factories*, 15(1):13, dec 2016.
- [138] Z. I. T. A. Soons, E. C. Ferreira, K. R. Patil, and I. Rocha. Identification of metabolic engineering targets through analysis of optimal and sub-optimal routes. *PloS One*, 8(4):e61648, jan 2013.
- [139] G. Stephanopoulos. Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering*, 1(1):1–11, 1999.
- [140] M. Suffness. *Taxol : science and applications*. CRC Press, 1995.
- [141] Y. Tan, J. G. Lafontaine Rivera, C. A. Contador, J. A. Asenjo, and J. C. Liao. Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. *Metabolic Engineering*, 13(1):60–75, 2011.
- [142] Y. Tan and J. C. Liao. Metabolic ensemble modeling for strain engineers. *Biotechnology Journal*, 7(3):343–53, mar 2012.
- [143] N. Tepper and T. Shlomi. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics (Oxford, England)*, 26(4):536–43, feb 2010.
- [144] M. K. Theisen, J. G. Lafontaine Rivera, and J. C. Liao. Stability of Ensemble Models Predicts Productivity of Enzymatic Systems. *PLoS Computational Biology*, 12(3), 2016.
- [145] Y. Toya, T. Shiraki, and H. Shimizu. SSDesign: Computational metabolic pathway design based on flux variability using elementary flux modes. *Biotechnology and Bioengineering*, 112(4):759–68, apr 2015.
- [146] L. M. Tran, M. L. Rizk, and J. C. Liao. Ensemble modeling of metabolic networks. *Biophysical Journal*, 95(12):5606–17, dec 2008.
- [147] C. T. Trinh, A. Wlaschin, and F. Sreenc. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology*, 81(5):813–26, jan 2009.
- [148] R. Urbanczik. SNA—a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, 7(1):129, jan 2006.

- [149] M. A. Valderrama-Gomez, D. Kreitmayer, S. Wolf, A. Marin-Sanguino, and A. Kremling. Application of theoretical methods to increase succinate production in engineered strains. *Bioprocess and Biosystems Engineering*, 40(4):479–497, 2017.
- [150] W. A. Van Winden, J. J. Heijnen, and P. J. T. Verheijen. Cumulative bondomers: A new concept in flux analysis from 2D [<sup>13</sup>C,<sup>1</sup>H] COSY NMR data. *Biotechnology and Bioengineering*, 80(7):731–745, 2002.
- [151] A. Varma, B. W. Boesch, and B. O. Palsson. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology*, 59(8):2465–73, aug 1993.
- [152] A. von Kamp and S. Klamt. Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms. *Nature Communications*, 8:15956, 2017.
- [153] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
- [154] C. W. Wang, M. K. Oh, and J. C. Liao. Engineered isoprenoid pathway enhances astaxanthin production in *Escherichia coli*. *Biotechnology and Bioengineering*, 62(2):235–241, 1999.
- [155] M. Weitzel, K. Nöh, T. Dalman, S. Niedenführ, B. Stute, and W. Wiechert. 13CFLUX2 - High-performance software suite for 13C-metabolic flux analysis. *Bioinformatics*, 29(1):143–145, jan 2013.
- [156] P. J. Westfall, D. J. Pitera, J. R. Lenihan, D. Eng, F. X. Woolard, R. Regentin, T. Horning, H. Tsuruta, D. J. Melis, A. Owens, S. Fickes, D. Diola, K. R. Benjamin, J. D. Keasling, M. D. Leavell, D. J. McPhee, N. S. Renninger, J. D. Newman, and C. J. Paddon. Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proceedings of the National Academy of Sciences*, 109(3):E111–E118, 2012.
- [157] W. Wiechert. 13C metabolic flux analysis. *Metabolic Engineering*, 3(3):195–206, jul 2001.
- [158] 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.
- [159] Z. Xu, P. Zheng, J. Sun, and Y. Ma. ReactKnock: identifying reaction deletion strategies for microbial strain optimization based on genome-scale metabolic network. *PloS One*, 8(12):e72150, jan 2013.
- [160] C. Yang, X. Gao, Y. Jiang, B. Sun, F. Gao, and S. Yang. Synergy between methylerythritol phosphate pathway and mevalonate pathway for isoprene production in *Escherichia coli*. *Metabolic Engineering*, 37:79–91, 2016.
- [161] L. Yang, W. R. Cluett, and R. Mahadevan. EMILiO: a fast algorithm for genome-scale strain design. *Metabolic Engineering*, 13(3):272–81, may 2011.
- [162] L. Yang, D. Ma, A. Ebrahim, C. J. Lloyd, M. A. Saunders, and B. O. Palsson. solveME: fast and reliable solution of nonlinear ME models for metabolic engineering. *BMC Bioinformatics*, 2016.
- [163] R. Yao, Y. Hirose, D. Sarkar, K. Nakahigashi, Q. Ye, and K. Shimizu. Catabolic regulation analysis of *Escherichia coli* and its *crp*, *mlc*, *mgsA*, *pgi* and *ptsG* mutants. *Microbial Cell Factories*, 10, 2011.
- [164] H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J. D. Trawick, R. E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H. B. Schreyer, S. Andrae, T. H. Yang, S. Y. Lee, M. J. Burk, and S. Van Dien. Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology*, 7(7):445–52, jul 2011.
- [165] L. P. Yomano, S. W. York, K. T. Shanmugam, and L. O. Ingram. Deletion of methylglyoxal synthase gene (*mgsA*) increased sugar co-metabolism in ethanol-producing *Escherichia coli*. *Biotechnology Letters*, 31(9):1389–1398, 2009.
- [166] L. Z. Yuan, P. E. Rouvière, R. A. Larossa, and W. Suh. Chromosomal promoter replacement of the isoprenoid pathway for enhancing carotenoid production in *E. coli*. *Metabolic Engineering*, 8(1):79–90, jan 2006.
- [167] N. Zamboni, S.-M. Fendt, M. Rühl, and U. Sauer. 13C-based metabolic flux analysis. *Nature Protocols*, 4(6):878–892, may 2009.
- [168] H. Zhang and Z. Pu. Beating the Uncertainties: Ensemble Forecasting and Ensemble-Based Data Assimilation in Modern Numerical Weather Prediction. *Advances in Meteorology*, 2010:1–10, 2010.
- [169] K. Zhuang, M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D. R. Lovley. Genome-scale dynamic modeling of the competition between *Rhodospirillum rubrum* and *Geobacter* in anoxic subsurface environments. *ISME Journal*, 5(2):305–316, 2011.
- [170] K. Zhuang, L. Yang, W. R. Cluett, and R. Mahadevan. Dynamic strain scanning optimization: an efficient strain design strategy for balanced yield, titer, and productivity. DySScO strategy for strain design. *BMC Biotechnology*, 13(1):8, jan 2013.



# Acronyms

**<sup>13</sup>C-MFA** <sup>13</sup>C-Metabolic Flux Analysis.

**[U-<sup>13</sup>C]** uniformly labeled <sup>13</sup>C.

***E. coli*** *Escherichia coli*.

**CCR** Carbon Catabolite Repression.

**dFBA** dynamic Flux Balance Analysis.

**DMAPP** dimethylallyl diphosphate.

**EM** Ensemble Modeling.

**EMRA** Ensemble Modeling for Robustness Analysis.

**EMU** Elementary Metabolite Units.

**FBA** Flux Balance Analysis.

**FVA** Flux Variability Analysis.

**G3P** glyceraldehyde 3-phosphate.

**GA** Genetic Algorithm.

**GDP** Gross Domestic Product.

**gDW** gram dry weight.

**HPLC** High Performance Liquid Chromatography.

**i.e.** id est.

**IPOPT** Interior Point OPTimizer.

**IPP** isopentenyl diphosphate.

**IPTG** Isopropyl  $\beta$ -D-1-thiogalactopyranoside.

**KAY** Kinetically Accessible Yield.

## *Acronyms*

**MEP** non-mevalonate.

**mmol** millimol.

**MS** Mass Spectrometry.

**MVA** mevalonate.

**NGAM** Non-growth Associated Maintenance.

**NMR** Nuclear Magnetic Resonance.

**OD** Optical Density.

**PTS** phosphotransferase system.

**RAM** Random Access Memory.

**rpm** revolutions per minute.

**SBT** Fachgebiet Systembiotechnologie.

**SLURM** Simple Linux Utility for Resource Management.

**TCA** tricarboxylic acid.

**UC** University of California.

**WT** wild-type.

# A. Appendix for Strain Engineering

## A.1. Metabolic Burden

### A.1.1. Slope of Acetate Line

The slope of experimental acetate lines was determined for three different strains, namely strain 31, 3150 and 3152. Refer to Table 4.1 for strain description. For strain 31, three different data sets were available for the determination of its slope. Slope of acetate lines was determined using standard linear least square regression for both growth rate as a function of acetate and acetate as a function of growth rate, as shown in Figure A.1. Since experimental acetate secretion rates normally exhibit higher standard deviations than growth rates (refer to Figure 5.2 and 6.11), the slope of the lines acetate vs growth rate is considered to describe more accurately the slope of the acetate lines. Since strains 31, 3150 and 3152 differ only by its metabolic load, the slope of their acetate lines should be constant [11]. The average slope considering all acetate lines correspond to:  $28.063 \pm 3.587$  mol acetate/gDW

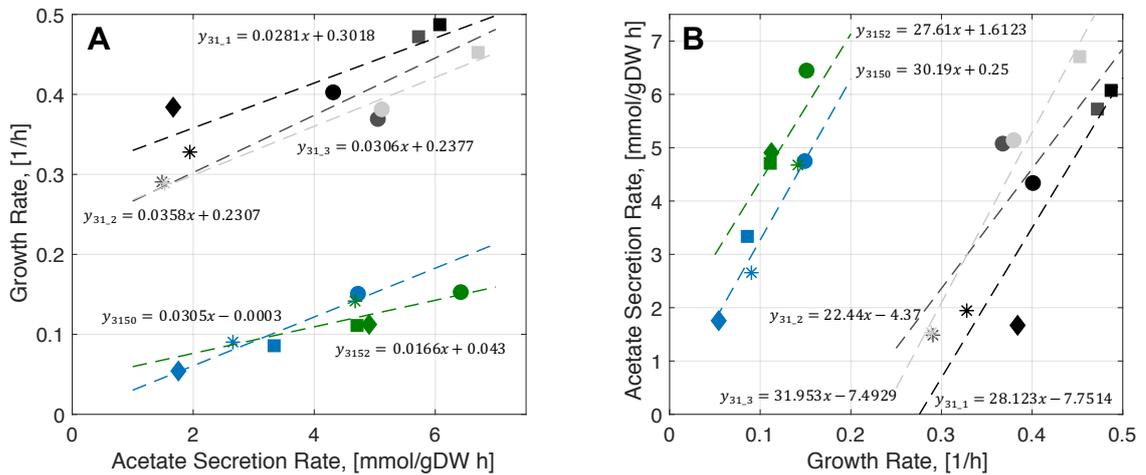


Figure A.1.: **Determination of the Acetate Line Slope for *E. coli* HMS.** Slope of acetate lines for strains 31, 3150 and 3152 were determined using linear regression for growth rate vs. acetate secretion rate (A.) and acetate secretion rate vs. growth rate (B.). Values for coefficient of determination were:  $R_{31,1}^2 = 0.79$ ,  $R_{31,2}^2 = 0.80$ ,  $R_{31,3}^2 = 0.97$ ,  $R_{3150}^2 = 0.92$ ,  $R_{3152}^2 = 0.45$ .  $y_{31,i}$  refers to the regression for one three different data sets. Symbols are as follows: +: acetate, o: fructose, \*: galactose, □: glucose, ◇: mannose.

## A. Appendix for Strain Engineering

### A.1.2. Amino Acid Sequences

Protein sequences for eGFP, mCherry, antibiotic resistances and additional replication proteins are given below. The mCherry-plasmid contains: mCherry, kanamycine resistance and replication protein 237. The eGFP-plasmid contains: eGFP, gentamycine resistance, replication protein 648 and protein XylS. These protein sequences can be used to construct strain specific metabolic reconstructions.

#### Amino acid sequence for eGFP

MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLLKFICTTGKLPVWPPTLVTTTLTYGVQ  
CFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEKDTLVNRIELKGIKDFKEDGNILGH  
KLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKD  
PNEKRDHMLLEFVTAAGITLGMDELYK

#### Amino acid sequence for mCherry

MVSKGEEDNMAIIEKFMRFKVMHEGVSNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQF  
MYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFDGGVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPV  
MQKKTMGWEASSERMYPEDGALKGEIKQRLKLDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSH  
NEDYTIVEQYERAEGRHSTGGMDELYK

#### Protein sequence for kanamycine resistance

MSHIQRETSRRLNSNMDADLYGYKWARDNVGQSGATIYRLYGKPDAPFLKHKGKSVANDVTDEM  
VRLNWLTEFMPLPTIKHFIRTPDDAWLLTTAIPGKTAQVLEEYVPSGENIVDALAVFLRRLHSIPVCNCPF  
NSDRVFRLAQAQSRMNNGLVDASDFDERRNGWPVEQVVKEMHKLLPFPSPDSVVTHGDFSLDNLIFDEGK  
LIGCIDVGRVGIADRYQDLAILWNCLEGFSPSLQKRLFQKYGIDNPD MNKLQFHLMLDEFF

#### Protein sequence for gentamycine resistance

MLRSSNDVTQQGSRPKTKLGGSSMGIIRTCRLGPDQVKS MR AALDFGREFGDVATYSQHQPDS DYLG NL  
LRSKTFIALAAFDQEAVVGALAAAYVLPKFEQPRSEIYIYDLAVSGEHRRQGIATALINLLKHEANALGAYVIY  
VQADYGDDPAVALYTKLGIREEVMHFDIDPSTAT

#### Protein sequence for replication protein 237

MATQSREIGIQAKNKP GHVWVQTERKAHEAWAGLIARKPTAAMLLHHLVAQMGHQNAV VVSQKTL SKLIG  
RSLRTVQYAVKDLVAERWISVVKLNGPGTVSAYVVNDRVAWGQPRDQLRSLVFSAAVVVDHDDQDESLL  
GHGDLRRIPTLYPGEQQLP TGPEPPSQPGIPGMEPDLPAL TETE EWERRGQQRLPMPDEPCFLDDGEP  
LEPPTRVTLPRR

#### Protein sequence for replication protein 648

MSNVASPPMVYKSNALVEAA YRLSVQEQRIVLACISQVKRSEPV TDEVMYSVTAEDIATMAGVPIESSYNQL  
KEAALRLKRREVRLTQEPNGKGRPSVMITGWVQTHIYREGEGRVELRFTKDM LPYLTELTKQFTKYALA  
DVAKMDSTHAIRLYELLMQWDSIGQREIEIDQLRKWFQLEGRYPSIKDFKLRVLDPAVTQINEHSPLQVEV  
AQRKTGRKVTHLLFSFGPKKPAKAVGKAPAKRKAGKISDAEIAKQARPGETWEAARL TQMP

#### Protein sequence for protein XylS

MDFCLLNEKSQIFVHAEPYAVSDYVNQYVGT HSIRLPKGGRPAGRLHHRIFGCLDL CRISYGGSVRVIS PGL  
ETCYHLQIILKGHCLWRGHGQEHYFAPGELLLLNPDDQADLTYS EDC EK FIVKLP SVVLD RACSDNNWHKP  
REGIRFAARHNLQQLDGFINLLGLVCDEAEHTKSM PRVQEHYAGHASKLLEMLG SNVSREIFSKGNPSFERV  
VQFIEENLKRNISLERLAELAMMSPRSLYNLFEKHAGTTPKNYIRNRKLESIRACLNDPSANVRSITEIALDY  
GFLHLGRFAENYRSAFGELPSD TLRQCKKEVA

Table A.1.: Concentration Time Courses for Galactose as Carbon Source.

Time [h]	Biomass, [g/l]			Galactose, [mmol/l]			Acetate, [mmol/l]		
	31	3150	3152	31	3150	3152	31	3150	3152
0	1.597E-02	1.903E-02	1.209E-02	2.346E+01	2.391E+01	2.302E+01	NaN	NaN	7.980E-01
1.5	2.075E-02	1.930E-02	1.362E-02	2.322E+01	2.202E+01	2.150E+01	NaN	7.191E-01	7.597E-01
3	3.707E-02	3.004E-02	1.804E-02	2.088E+01	2.309E+01	2.129E+01	NaN	8.542E-01	7.326E-01
4.5	6.215E-02	3.806E-02	2.183E-02	2.076E+01	2.101E+01	2.125E+01	NaN	1.056E+00	8.901E-01
6	1.043E-01	3.933E-02	1.894E-02	2.016E+01	2.259E+01	2.100E+01	7.349E-01	1.534E+00	1.061E+00
7.5	1.710E-01	5.493E-02	2.462E-02	1.817E+01	2.076E+01	2.142E+01	7.137E-01	1.823E+00	1.205E+00

Table A.2.: Concentration Time Courses for Glucose as Carbon Source.

Time [h]	Biomass, [g/l]			Glucose, [mmol/l]			Acetate, [mmol/l]			Formate, [mmol/l]
	31	3150	3152	31	3150	3152	31	3150	3152	31
0	1.506E-02	1.858E-02	1.660E-02	2.298E+01	2.029E+01	2.049E+01	NaN	NaN	7.703E-01	NaN
1.5	2.697E-02	3.103E-02	2.291E-02	1.931E+01	1.877E+01	1.907E+01	7.026E-01	8.158E-01	7.973E-01	NaN
3	6.765E-02	2.904E-02	1.624E-02	1.819E+01	1.856E+01	1.921E+01	1.320E+00	9.096E-01	9.029E-01	NaN
4.5	1.410E-01	4.303E-02	2.435E-02	1.666E+01	1.861E+01	1.866E+01	2.461E+00	1.164E+00	1.080E+00	5.730E+00
6	3.114E-01	4.844E-02	3.292E-02	1.326E+01	1.885E+01	1.903E+01	4.919E+00	1.600E+00	1.390E+00	2.748E+01
7.5	4.991E-01	4.221E-02	3.031E-02	6.482E+00	1.873E+01	1.899E+01	6.485E+00	1.855E+00	1.466E+00	NAN

### A.1.3. Concentration Time Courses

Experimental time courses for biomass, substrate, acetate and formate for growth on different carbon sources (galactose, glucose, mannose, fructose and acetate) are shown in Tables A.1 to A.5. From these time courses, respective rates were calculated as explained in Section 2.1.2 and are listed in Table A.6. Note that formate was only detected for strain 31, when grown fructose, mannose and glucose.

### A.1.4. Flux Distribution Patters for Three Zones

FVA results for Phase I, II and III are shown in Figures A.2 to A.4, respectively. The *E. coli* core model was used for all calculations. For each phase, glucose and oxygen uptake rate were set to defined values then, maximal growth rate was calculated. After fixing the growth rate to that value, minimal and maximal rates supported by each reaction in the model was calculated. Three rules were used to identify biologically feasible fluxes:

1.  $\text{minFlux} == \text{maxFlux}$  and  $\text{minFlux} == 0$
2.  $\text{minFlux} == 0$  and  $\text{maxFlux} > 900$
3.  $\text{abs}(\text{abs}(\text{minFlux}) - \text{abs}(\text{maxFlux})) > 900$

reactions *not* fulfilling *any* of these rules are identified their flux ranges plotted for each phase. Note that in each phase, exchange reactions are highlighted in blue.

Table A.3.: Concentration Time Courses for Mannose as Carbon Source.

Time [h]	Biomass, [g/l]			Mannose, [mmol/l]			Acetate, [mmol/l]			Formate, [mmol/l]
	31	3150	3152	31	3150	3152	31	3150	3152	31
0	1.696E-02	2.580E-02	1.993E-02	1.987E+01	2.111E+01	2.054E+01	NaN	NaN	1.038E+00	NaN
1.5	2.850E-02	2.688E-02	1.741E-02	2.122E+01	2.053E+01	1.847E+01	7.858E-01	7.244E-01	1.035E+00	NaN
3	5.737E-02	4.131E-02	1.984E-02	1.861E+01	2.053E+01	1.839E+01	1.217E+00	7.244E-01	1.165E+00	NaN
4.5	9.660E-02	4.266E-02	2.237E-02	1.827E+01	1.858E+01	1.874E+01	1.473E+00	9.470E-01	1.312E+00	NaN
6	1.792E-01	5.619E-02	2.246E-02	1.859E+01	2.083E+01	1.874E+01	2.121E+00	1.755E+00	1.312E+00	1.620E+01
7.5	2.859E-01	5.358E-02	2.950E-02	1.613E+01	1.917E+01	1.819E+01	1.913E+00	1.949E+00	1.405E+00	2.437E+01

A. Appendix for Strain Engineering

Table A.4.: Concentration Time Courses for Fructose as Carbon Source.

Time [h]	Biomass, [g/l]			Fructose, [mmol/l]			Acetate, [mmol/l]			Formate, [mmol/l]
	31	3150	3152	31	3150	3152	31	3150	3152	31
0	1.497E-02	1.696E-02	7.487E-03	2.393E+01	2.245E+01	2.426E+01	NaN	NaN	7.192E-01	NaN
1.5	2.517E-02	1.867E-02	9.471E-03	2.466E+01	2.378E+01	2.151E+01	9.693E-01	7.371E-01	7.626E-01	NaN
3	3.906E-02	2.426E-02	1.191E-02	2.147E+01	2.180E+01	2.198E+01	1.102E+00	9.476E-01	7.957E-01	NaN
4.5	8.966E-02	3.346E-02	1.768E-02	2.014E+01	2.166E+01	2.196E+01	1.539E+00	1.275E+00	9.296E-01	NaN
6	1.648E-01	3.987E-02	1.705E-02	1.694E+01	2.302E+01	2.121E+01	2.526E+00	2.231E+00	1.055E+00	1.683E+01
7.5	2.787E-01	4.970E-02	2.345E-02	1.429E+01	2.113E+01	2.206E+01	3.652E+00	1.797E+00	1.224E+00	2.798E+01

Table A.5.: Time Courses for Acetate as Carbon Source.

Time, [h]	Biomass, [g/l]			Acetate, [mmol/l]		
	31	3150	3152	31	3150	3152
0	1.335E-02	1.696E-02	1.146E-02	2.235E+01	2.357E+01	2.122E+01
1.5	1.633E-02	1.488E-02	1.795E-02	2.192E+01	2.139E+01	2.008E+01
3	2.426E-02	1.858E-02	1.479E-02	1.926E+01	2.026E+01	2.000E+01
4.5	3.310E-02	1.732E-02	1.786E-02	1.859E+01	2.034E+01	1.973E+01
6	4.221E-02	1.885E-02	1.651E-02	1.720E+01	2.037E+01	1.928E+01
7.5	6.621E-02	1.831E-02	2.318E-02	1.492E+01	2.048E+01	1.940E+01

Table A.6.: Exchange Rates for Three Strains and Five Sugars. Rates for growth, substrate uptake and acetate secretion were calculated from measured time courses as described in Chapter 2.1.2

	Glucose			Mannose			Fructose			Galactose			Acetate		
	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	Strain	
	31	3150	3152	31	3150	3152	31	3150	3152	31	3150	3152	31	3150	3152
$\mu$	0.487	0.111	0.086	0.384	0.112	0.054	0.402	0.152	0.150	0.328	0.146	0.090	0.213	0.019	0.066
$r_{uptake}$	14.111	4.449	4.540	5.433	3.353	3.619	15.198	6.249	12.643	10.103	9.248	8.814	29.668	7.829	8.967
$r_{acetate}$	6.074	4.707	3.336	1.670	4.905	1.755	4.324	6.432	4.734	-	4.825	2.655	-	-	-

$\mu$  is given in 1/h; both substrate uptake ( $r_{uptake}$ ) and acetate secretion ( $r_{acetate}$ ) rates are given in mmol/gDW h. '-' symbolizes that no acetate secretion was detected.

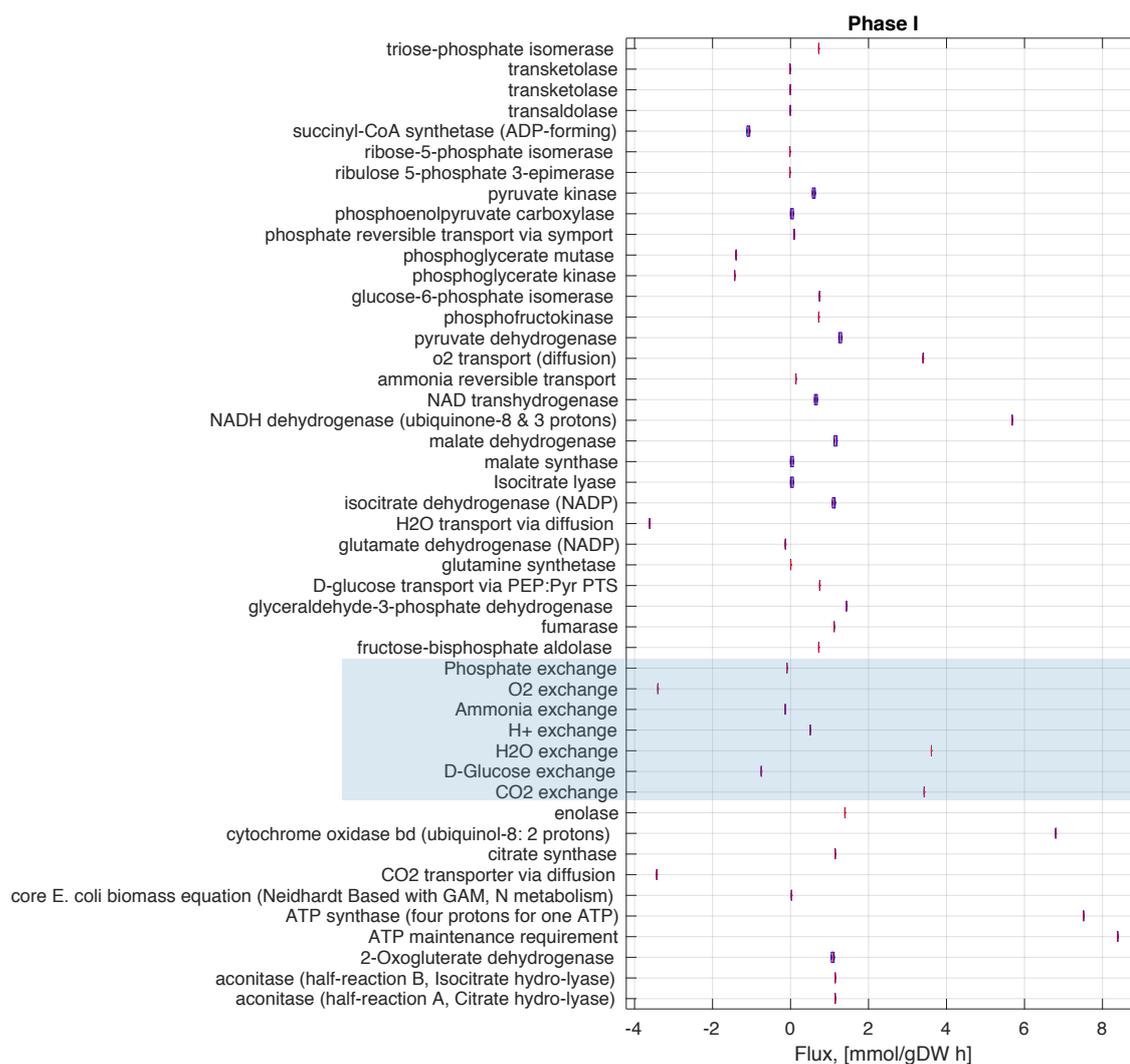


Figure A.2.: **Flux Variability Analysis for Phase I.** Upper boundary for glucose and oxygen uptake rates were set to 0.75 and 4 mmol/gDW h, respectively. CO<sub>2</sub> is the only carbon-based by-product secreted by the cell.

## A. Appendix for Strain Engineering

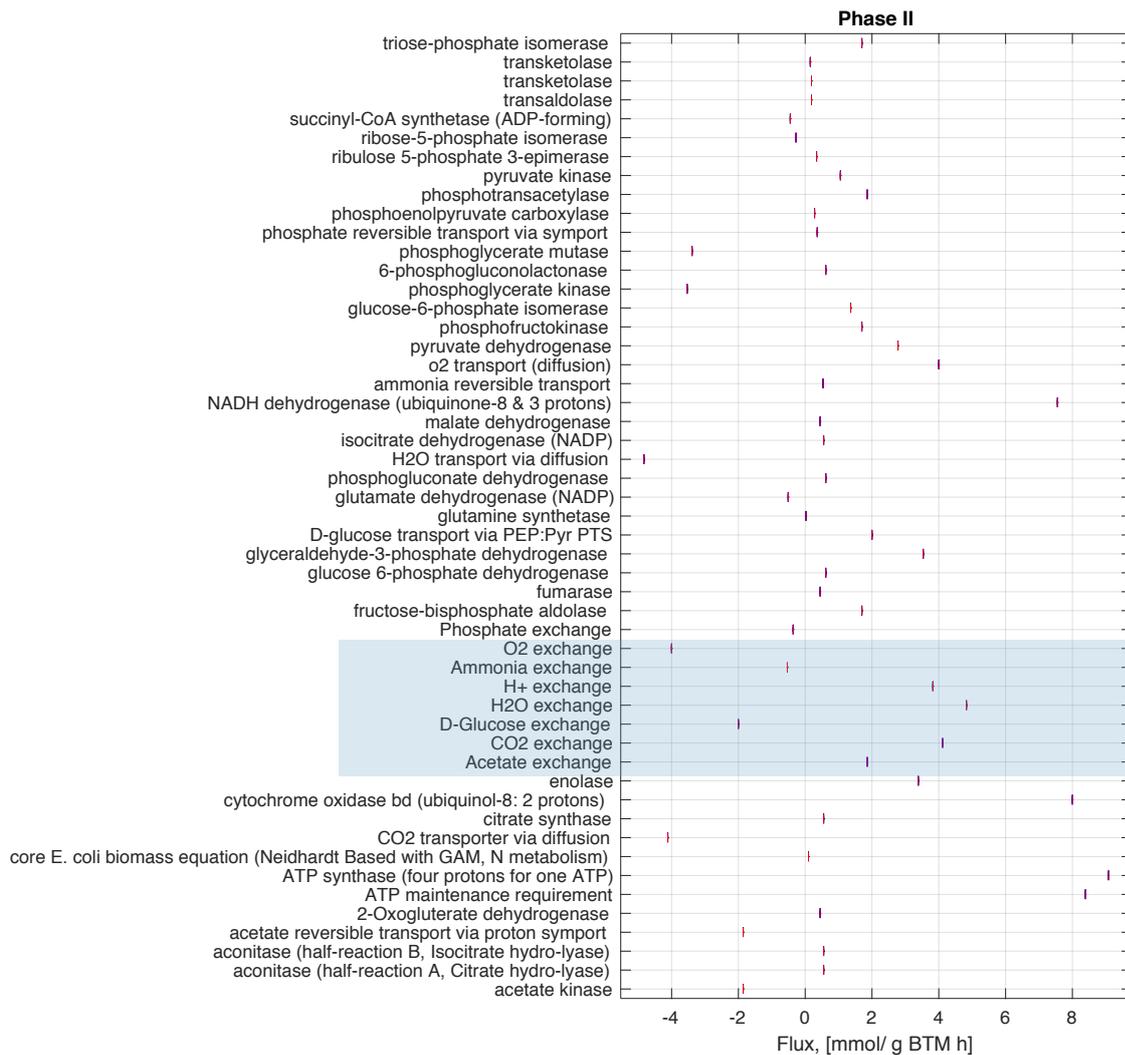


Figure A.3.: **Flux Variability Analysis for Phase II.** Upper boundary for glucose and oxygen uptake rates were set to 2 and 4 mmol/gDW h, respectively. CO<sub>2</sub> and acetate represent carbon-based by-products secreted by the cell.

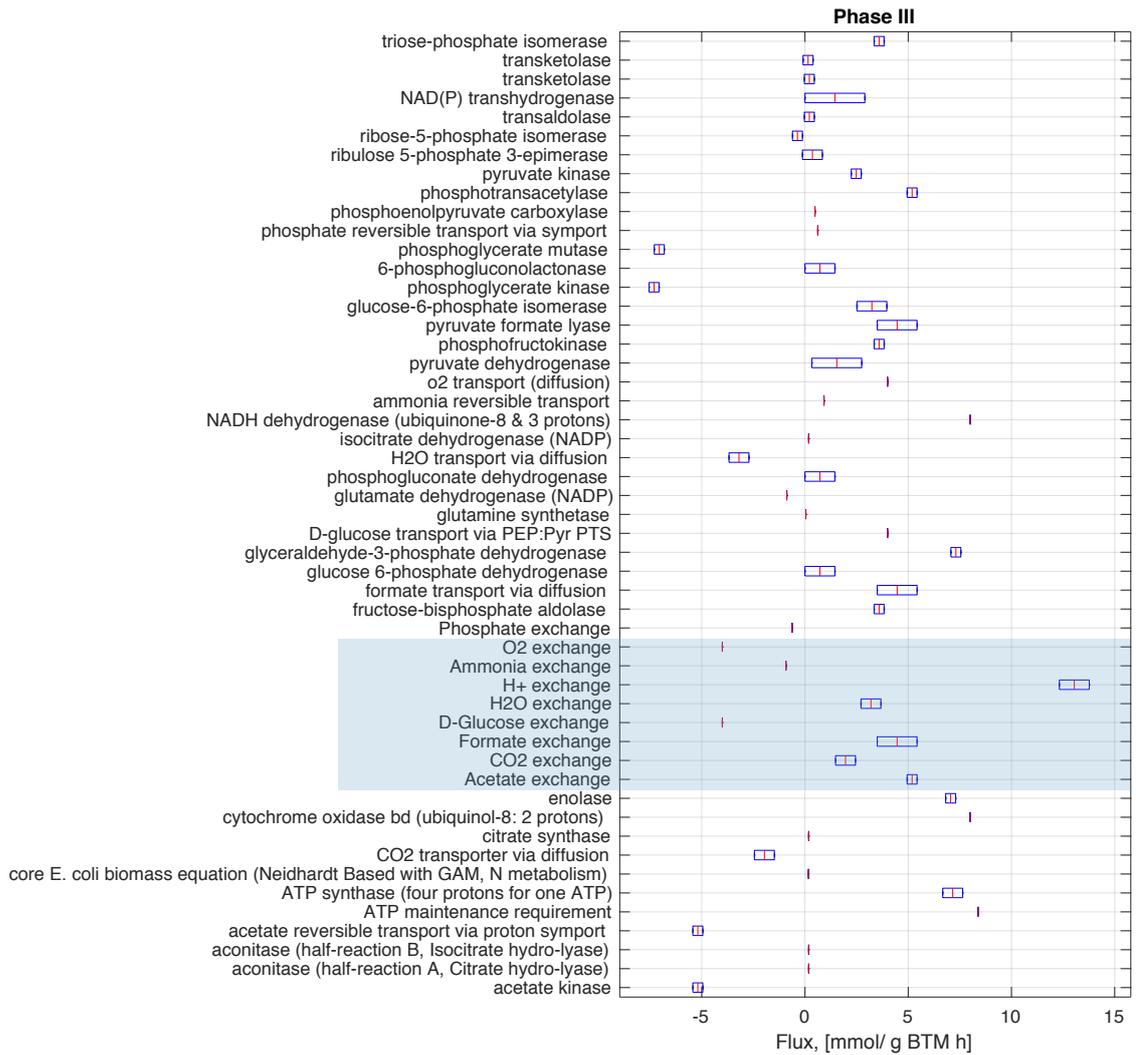


Figure A.4.: **Flux Variability Analysis for Phase III.** Upper boundary for glucose and oxygen uptake rates were set to 4 and 4 mmol/gDW h, respectively. CO<sub>2</sub>, acetate and formate represent carbon-based by-products secreted by the cell.

## A. Appendix for Strain Engineering

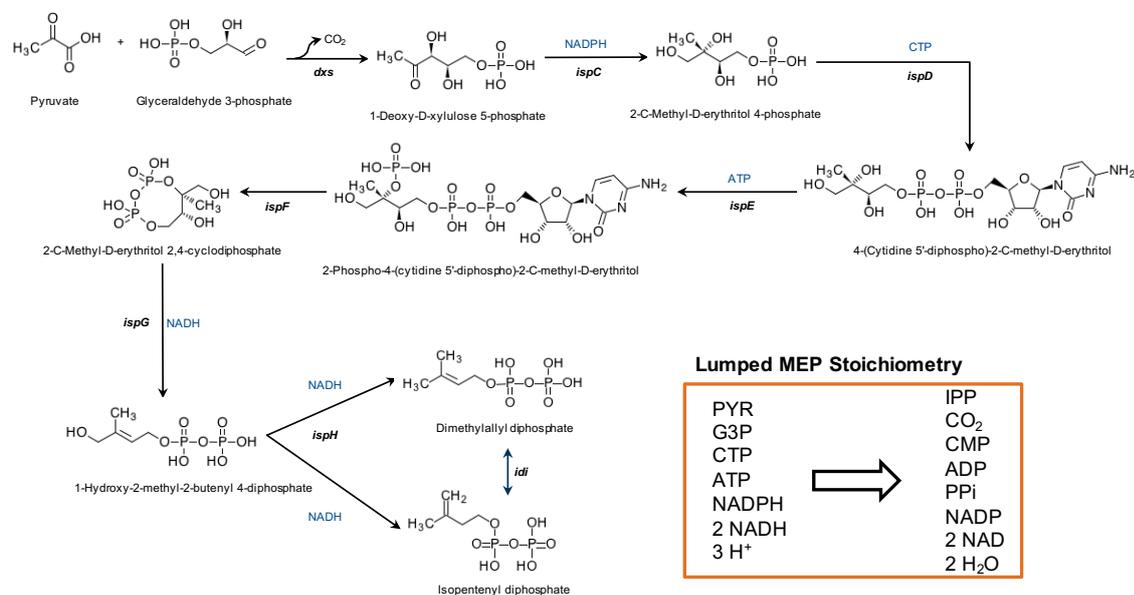


Figure A.5.: **Non-Mevalonate Pathway (MEP)**. The non-mevalonate pathway consists of 8 different reactions. This pathway converts pyruvate and glyceraldehyde 3-phosphate into the universal isoprenoid precursors dimethylallyl diphosphate and isopentenyl diphosphate. All metabolite's figures were taken from KEGG [103]

## A.2. Strain Design Algorithms for Target Identification

### A.2.1. Mevalonate, Non-Mevalonate & Taxadiene Formation Pathways

Terpenes are naturally synthesized compounds formed by the ligation of activated isoprene molecules IPP and its isomer DMAPP. In nature, IPP and DMAPP can be synthesized through two different metabolic routes: the mevalonate (MVA) or the non-mevalonate (MEP) pathway. Figures A.5, A.6 and A.7 schematically show the non-mevalonate, the mevalonate and the taxadiene production pathways, respectively.

### A.2.2. Concentration Time Courses for Taxadiene Producing Strains

Table A.7 summarizes time courses for OD, glucose and acetate concentration for five different strains. Concentrations were measured in duplicate.

### A.2.3. $^{13}\text{C}$ Labeling Data

Table A.8 provides labeling patterns for protein-bound amino acids for five different taxadiene producing strains. Measurements were kindly provided by Claudia Huber at **Bio**

## A.2. Strain Design Algorithms for Target Identification

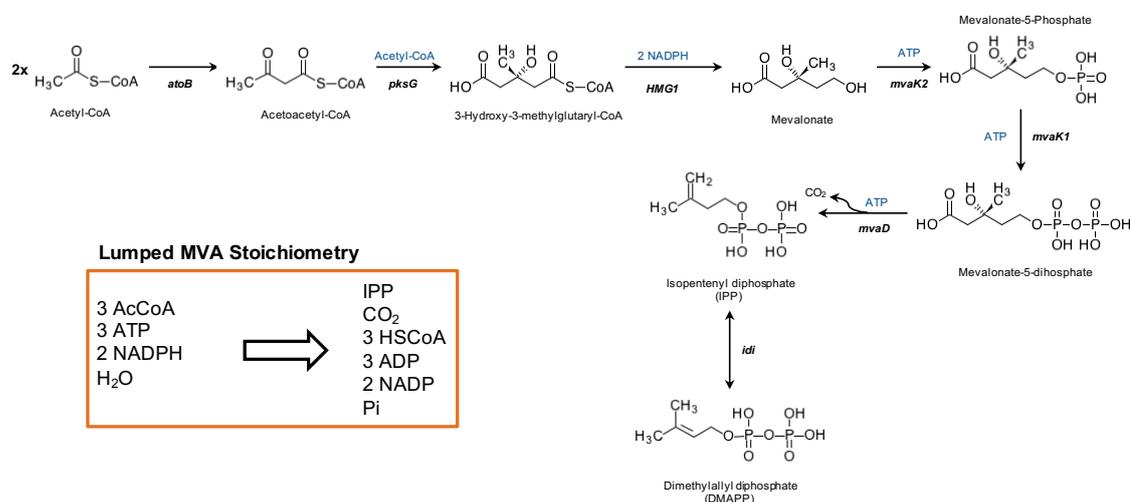


Figure A.6.: **Mevalonate Pathway (MVA)**. The mevalonate pathway consist of 7 different reactions. This pathway converts three molecules of acetyl-coA into one molecule of isopentenyl diphosphate (IPP). Dimethylallyl diphosphate is produced via isomerization of IPP. All metabolite's figures were taken from KEGG [103]

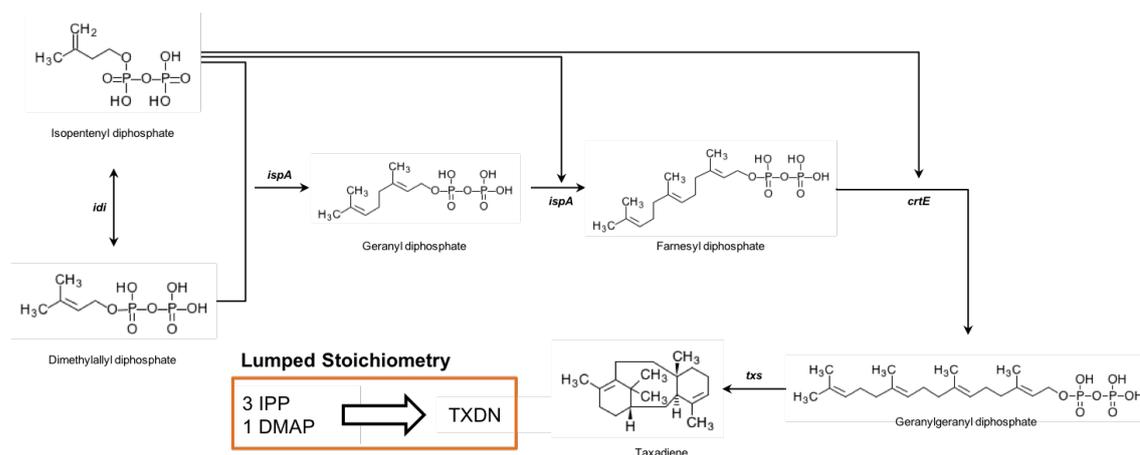


Figure A.7.: **Taxadiene Production Pathway**. Three molecules of isopentenyl diphosphate and one molecule of dimetylallyl diphosphate are converted in a serie of four reactions into taxadiene -a diterpene-. All metabolite's figures were taken from KEGG [103]

A. Appendix for Strain Engineering

Table A.7.: **Concentration Time Courses for Taxadiene Producing Strains.** Duplicate concentration time courses for five taxadiene producing strains are given.

Time, [h]	OD		Glucose, [mmol/l]		Acetate, [mmol/l]	
	1	2	1	2	1	2
<b>HMS 2p</b>						
0	0.574	0.582	25.186	38.874	0.512	0.648
1	0.644	0.650	15.407	36.665	0.940	2.098
3	0.894	0.903	19.173	13.076	1.959	1.767
5	1.113	1.188	23.574	28.184	5.747	0.868
7	1.374	1.392	21.489	20.248	8.232	6.831
<b><math>\Delta</math>ptsG 2p</b>						
0	0.582	0.574	29.437	37.734	0.514	0.451
1	0.631	0.615	26.627	37.706	0.687	0.765
3	0.795	0.735	28.390	19.556	1.431	0.929
5	0.933	0.747	27.945	27.710	2.308	1.961
7	1.134	0.873	16.959	26.575	1.940	4.500
<b><math>\Delta</math>ptsG 1p</b>						
0	0.561	0.564	29.276	44.400	0.412	0.404
1	0.696	0.692	27.753	24.874	0.455	–
3	1.149	1.167	29.740	24.838	0.403	–
5	1.398	1.518	22.468	18.646	1.058	0.895
7	2.208	2.124	14.830	22.124	1.354	1.834
<b><math>\Delta</math>xylE 2p</b>						
0	0.522	0.556	27.468	51.126	0.556	0.535
1	0.624	0.633	31.077	26.852	0.967	0.760
3	0.927	0.876	23.335	15.331	1.662	0.993
5	1.146	1.173	25.100	23.428	3.055	2.392
7	1.308	1.302	18.742	15.412	3.279	2.715
<b><math>\Delta</math>xylE 1p</b>						
0	0.540	0.561	28.647	34.630	0.552	0.501
1	0.622	0.621	31.239	18.537	1.131	4.820
3	0.945	0.972	28.533	17.004	2.124	1.317
5	1.413	1.377	21.021	15.139	4.266	3.044
7	1.971	1.917	6.947	8.740	2.889	3.148

## A.2. Strain Design Algorithms for Target Identification

Table A.8.: <sup>13</sup>C-labeling Data for Five Taxadiene Producing Strains. Labeling patterns of protein-bound amino acids was experimentally determined in duplicate for five taxadiene producing strains.

Amino Acid	Fragment	HMS %	<sup>2</sup> p σ	ΔxyIE %	<sup>2</sup> p σ	ΔxyIE %	<sup>1</sup> p σ	ΔptsG %	<sup>2</sup> p σ	ΔptsG %	<sup>1</sup> p σ
Ala-260	M+0	0.934	2.980E-03	0.944	1.315E-03	0.919	2.968E-03	0.931	3.617E-03	0.912	2.092E-03
Ala-260	M+1	0.009	2.564E-03	0.008	1.291E-03	0.016	2.732E-03	0.019	3.303E-03	0.019	1.962E-03
Ala-260	M+2	0.006	3.178E-04	0.006	6.089E-04	0.010	6.032E-04	0.010	6.044E-04	0.012	5.825E-04
Ala-260	M+3	0.050	6.413E-04	0.042	6.788E-04	0.055	5.422E-04	0.040	3.745E-04	0.057	7.258E-04
Asp-418	M+0	0.891	1.809E-03	0.909	9.501E-04	0.862	4.125E-03	0.888	1.894E-03	0.841	2.554E-03
Asp-418	M+1	0.048	1.843E-03	0.043	8.426E-04	0.064	4.174E-03	0.055	1.680E-03	0.081	2.810E-03
Asp-418	M+2	0.024	1.147E-03	0.020	5.316E-04	0.033	1.330E-03	0.033	2.134E-03	0.050	1.077E-03
Asp-418	M+3	0.035	3.678E-04	0.025	4.796E-04	0.037	8.487E-04	0.022	5.565E-04	0.025	5.196E-04
Asp-418	M+4	0.003	2.635E-04	0.002	3.231E-04	0.004	1.584E-04	0.002	3.570E-04	0.003	8.999E-05
Glu-432	M+0	0.854	2.631E-03	0.872	3.018E-03	0.820	4.462E-03	0.854	4.843E-03	0.804	2.166E-03
Glu-432	M+1	0.042	2.117E-03	0.045	3.079E-03	0.060	3.794E-03	0.054	3.564E-03	0.077	1.841E-03
Glu-432	M+2	0.082	1.566E-03	0.066	6.649E-04	0.095	1.916E-03	0.073	1.695E-03	0.096	2.225E-03
Glu-432	M+3	0.018	1.000E-03	0.013	6.943E-04	0.020	6.734E-04	0.015	6.820E-04	0.019	5.804E-04
Glu-432	M+4	0.003	3.735E-04	0.003	2.509E-04	0.004	4.178E-04	0.003	1.360E-04	0.004	3.837E-04
Glu-432	M+5	0.002	2.339E-04	0.001	6.043E-05	0.002	1.119E-04	0.001	7.015E-05	0.001	7.789E-05
Gly-246	M+0	0.936	6.189E-04	0.947	7.906E-04	0.924	6.366E-04	0.939	2.662E-03	0.918	1.091E-03
Gly-246	M+1	0.013	6.568E-04	0.011	6.325E-04	0.018	6.722E-04	0.019	2.531E-03	0.024	9.768E-04
Gly-246	M+2	0.051	6.939E-04	0.042	2.483E-04	0.058	5.347E-04	0.043	3.970E-04	0.058	3.015E-04
His-440	M+0	0.857	3.009E-03	0.881	1.877E-03	0.827	2.361E-03	0.862	2.795E-03	0.812	4.596E-03
His-440	M+1	0.055	4.232E-03	0.047	9.179E-04	0.071	3.330E-03	0.058	3.612E-03	0.079	4.839E-03
His-440	M+2	0.027	1.822E-03	0.022	2.179E-03	0.036	2.402E-03	0.027	1.590E-03	0.042	2.163E-03
His-440	M+3	0.038	1.244E-03	0.032	9.731E-04	0.046	2.021E-03	0.035	8.171E-04	0.048	1.866E-03
His-440	M+4	0.006	9.074E-04	0.005	7.217E-04	0.007	5.517E-04	0.006	1.471E-03	0.009	6.396E-04
His-440	M+5	0.016	6.389E-04	0.013	3.869E-04	0.014	4.941E-04	0.012	8.099E-04	0.011	4.956E-04
His-440	M+6	0.001	2.831E-04	0.001	2.520E-04	0.001	3.943E-04	0.000	3.612E-04	0.000	1.400E-04
Ile-200	M+0	0.859	1.114E-03	0.885	9.018E-04	0.826	1.307E-03	0.864	2.857E-03	0.807	9.733E-04
Ile-200	M+1	0.038	3.968E-04	0.034	5.744E-04	0.057	1.387E-03	0.050	2.182E-03	0.075	8.657E-04
Ile-200	M+2	0.082	7.766E-04	0.065	3.457E-04	0.094	4.016E-04	0.069	6.309E-04	0.097	6.888E-04
Ile-200	M+3	0.017	2.073E-04	0.013	1.061E-04	0.019	1.738E-04	0.014	3.132E-04	0.017	3.678E-04
Ile-200	M+4	0.003	1.375E-04	0.002	6.190E-05	0.003	7.314E-05	0.002	1.139E-04	0.003	1.291E-04
Ile-200	M+5	0.001	6.075E-05	0.001	4.236E-05	0.001	5.891E-05	0.001	6.891E-05	0.001	5.338E-05
Leu-274	M+0	0.842	8.555E-04	0.875	1.677E-03	0.813	2.007E-03	0.852	5.227E-03	0.809	2.346E-03
Leu-274	M+1	0.051	1.097E-03	0.041	1.899E-03	0.068	2.039E-03	0.056	4.421E-03	0.066	1.840E-03
Leu-274	M+2	0.092	6.231E-04	0.072	7.846E-04	0.101	5.817E-04	0.077	1.007E-03	0.107	8.957E-04
Leu-274	M+3	0.010	3.020E-04	0.009	1.679E-04	0.012	1.312E-04	0.010	4.354E-04	0.013	2.845E-04
Leu-274	M+4	0.004	9.750E-05	0.004	1.787E-04	0.005	1.364E-04	0.004	8.379E-05	0.005	1.128E-04
Leu-274	M+5	0.000	1.074E-04	0.000	6.585E-05	0.001	9.298E-05	0.000	9.645E-05	0.001	7.141E-05
Lys-431	M+0	0.846	1.740E-03	0.875	3.888E-03	0.812	3.891E-03	0.853	3.967E-03	0.794	2.218E-03
Lys-431	M+1	0.046	2.254E-03	0.042	4.034E-03	0.065	3.982E-03	0.056	5.255E-03	0.081	2.019E-03
Lys-431	M+2	0.055	2.517E-03	0.044	1.603E-03	0.063	3.038E-03	0.052	3.271E-03	0.072	1.546E-03
Lys-431	M+3	0.051	1.258E-03	0.039	1.256E-03	0.056	1.139E-03	0.038	1.399E-03	0.048	7.768E-04
Lys-431	M+4	0.001	5.495E-04	0.001	1.074E-03	0.002	4.057E-04	0.001	4.545E-04	0.003	1.205E-03
Lys-431	M+5	0.002	5.656E-04	0.001	5.356E-04	0.002	5.198E-04	0.000	2.647E-04	0.001	4.245E-04
Lys-431	M+6	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00
Met-320	M+0	0.915	1.005E-01	0.933	7.293E-02	0.814	6.578E-03	0.928	7.937E-02	1.000	4.419E-08
Met-320	M+1	0.052	7.160E-02	0.041	4.511E-02	0.102	9.025E-03	0.041	4.593E-02	0.000	4.486E-08
Met-320	M+2	0.010	1.222E-02	0.015	1.610E-02	0.040	6.690E-03	0.018	2.060E-02	0.000	1.359E-08
Met-320	M+3	0.011	1.810E-02	0.007	1.206E-02	0.038	4.471E-03	0.010	1.171E-02	0.000	0.000E+00
Met-320	M+4	0.004	6.758E-03	0.002	2.926E-03	0.004	3.451E-03	0.001	1.870E-03	0.000	0.000E+00
Met-320	M+5	0.008	9.015E-03	0.001	3.268E-03	0.001	1.837E-03	0.001	1.773E-03	0.000	0.000E+00
Phe-336	M+0	0.810	3.311E-03	0.850	2.089E-03	0.770	2.718E-03	0.807	1.068E-02	0.766	1.739E-03
Phe-336	M+1	0.040	3.024E-03	0.032	1.535E-03	0.061	2.605E-03	0.058	1.027E-02	0.056	1.596E-03
Phe-336	M+2	0.046	1.835E-03	0.036	1.319E-03	0.054	1.398E-03	0.044	1.098E-03	0.060	1.062E-03
Phe-336	M+3	0.067	6.878E-04	0.053	4.226E-04	0.074	1.529E-03	0.061	1.856E-03	0.075	9.133E-04
Phe-336	M+4	0.024	7.201E-04	0.020	7.191E-04	0.027	9.748E-04	0.020	5.691E-04	0.030	3.603E-04
Phe-336	M+5	0.006	2.895E-04	0.005	3.894E-04	0.007	2.821E-04	0.006	4.702E-04	0.007	4.034E-04
Phe-336	M+6	0.004	2.618E-04	0.003	2.484E-04	0.004	2.484E-04	0.003	1.238E-04	0.005	1.960E-04
Phe-336	M+7	0.002	2.171E-04	0.001	1.332E-04	0.002	1.873E-04	0.001	2.396E-04	0.002	1.276E-04
Phe-336	M+8	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	1.141E-05
Phe-336	M+9	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00
Pro-286	M+0	0.865	1.491E-03	0.884	1.626E-03	0.834	2.155E-03	0.867	1.683E-03	0.814	1.780E-03
Pro-286	M+1	0.045	1.421E-03	0.045	1.278E-03	0.062	2.114E-03	0.057	1.881E-03	0.081	1.202E-03
Pro-286	M+2	0.065	8.713E-04	0.052	9.787E-04	0.077	1.018E-03	0.056	1.961E-03	0.080	1.978E-03
Pro-286	M+3	0.020	6.241E-04	0.015	6.414E-04	0.021	7.249E-04	0.016	6.849E-04	0.019	6.094E-04
Pro-286	M+4	0.004	5.591E-04	0.003	3.881E-04	0.004	5.576E-04	0.003	5.908E-04	0.005	2.993E-04
Pro-286	M+5	0.001	2.546E-04	0.000	3.194E-04	0.001	2.386E-04	0.000	1.980E-04	0.001	2.314E-04
Ser-390	M+0	0.926	2.752E-03	0.945	2.232E-03	0.914	2.593E-03	0.934	4.186E-03	0.907	2.597E-03
Ser-390	M+1	0.022	2.837E-03	0.013	3.392E-03	0.027	2.418E-03	0.022	4.058E-03	0.033	2.747E-03
Ser-390	M+2	0.021	1.056E-03	0.015	1.349E-03	0.023	1.067E-03	0.018	1.440E-03	0.025	8.984E-04
Ser-390	M+3	0.031	5.418E-04	0.026	3.878E-04	0.036	6.687E-04	0.026	7.491E-04	0.035	7.691E-04
Thr-404	M+0	0.868	2.713E-03	0.888	1.778E-03	0.839	2.511E-03	0.861	8.432E-03	0.819	2.866E-03
Thr-404	M+1	0.068	2.081E-03	0.063	2.445E-03	0.087	3.752E-03	0.083	1.045E-02	0.101	1.860E-03
Thr-404	M+2	0.031	1.712E-03	0.028	1.288E-03	0.038	2.383E-03	0.035	2.951E-03	0.056	1.697E-03
Thr-404	M+3	0.032	1.453E-03	0.021	4.553E-04	0.036	1.413E-03	0.020	1.148E-03	0.024	1.016E-03
Thr-404	M+4	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00	0.000	0.000E+00
Tyr-466	M+0	0.820	4.341E-03	0.858	3.717E-03	0.779	7.490E-03	0.827	4.299E-03	0.775	3.563E-03
Tyr-466	M+1	0.033	3.134E-03	0.027	4.244E-03	0.052	5.207E-03	0.045	4.067E-03	0.046	2.948E-03
Tyr-466	M+2	0.047	1.833E-03	0.036	1.944E-03	0.055	3.081E-03	0.041	1.624E-03	0.061	2.868E-03
Tyr-466	M+3	0.062	1.067E-03	0.050	1.449E-03	0.070	1.065E-03	0.054	2.127E-03	0.070	1.017E-03
Tyr-466	M+4	0.024	9.120E-04	0.018	6.356E-04	0.028	7.953E-04	0.021	1.052E-03	0.031	9.899E-04
Tyr-466	M+5	0.007	7.686E-04	0.007	8.202E-04	0.009	9.056E-04	0.007	6.027E-04	0.010	6.855E-04
Tyr-466	M+6	0.004	4.356E-04	0.003	4.505E-04	0.005	4.586E-04	0.004	2.627E-04	0.005	4.722E-04
Tyr-466	M+7	0.001	1.486E-04	0.001	3.328E-04	0.001	1.458E-04	0.001	4.547E-04	0.002	3.671E-04
Tyr-466	M+8	0.000	4.309E-05	0.001	3.921E-04	0.000	3.580E-04	0.001	2.226E-04	0.001	9.012E-05
Tyr-466	M+9	0.001	3.107E-04	0.000	3.071E-04	0.000	3.830E-04	0.000			

## A.2.4. Metabolic Model Used for the Software 13CFlux2

An exemplary model for the analysis of  $^{13}\text{C}$  labeling data is provided below. All required “.fml” files were generated using a Matlab script. All calculations were performed with the software 13CFlux2 [155]. 13CFlux2 is a Linux-based software, which is operated via the terminal command line. 13CFlux2 offers a great variety of functions. In this work, the software was used to calculate the intracellular flux distribution best matching experimentally determined exchange rates and labeling data. Five different 13CFlux2 functions were used to estimate flux distributions. First, the “.fml” file must be signed using the function *fmlsign*. The resulting file is then initialized using the function *sscanner*, which chooses a set of free fluxes and initialize the flux values with the analytical center of the specified constraints (exchange and growth rates). In a third step, the function *ssampler* is used to sample uniformly distributed, random flux distributions from the stoichiometry of the metabolic network given in the “.fml” file. Each sampled flux distribution is used as a starting point for the optimization problem which is performed by the function *multifitfluxes*. For each flux distribution, *multifitfluxes* generates a “.fwdsim” file, which contains the fitted flux distribution. Finally, the function *collectfitdata* is used for collecting together the results from multiple “.fwdsim” files into a single HDF5 file. Further analysis of the fitted flux distributions is performed in Matlab using the generated HDF5 file. The process described above was fully automatized using a series of python scripts and Matlab workflows.

```

1 <?xml version="1.0" encoding="utf-8"?>
2 <fluxml xmlns="http://www.13cflux.net/fluxml"><info><signature>ICaVRZF1HhE6iJcMsbD7Xoj3+
  GnQWeQkjZD6N9Cgwi+NjEGi9bHLJMxC4RF+6qQaVwhZru4bFW3s4/
  B1R8Xhn1CPGR4Yj6Oick9fogptEEN6NRmO4JVdgpPNBrkyKNCg7y0Vr1W/DEopJZiwNG/
  XBNPKM6aH1kYtVXxgbRnIestwpQ/d2Uufe9ESe4A7UIKrKqVVPeio5zsb1CEhbcA/
  QDWgpwwY0P86PuU4tpEnvAFK4apMNEAX2zSiPrTMcN7bXdidY7lzZk2Kia+UkMOF766wf/XUj+4
  tgapeRLlqH09KCj3BBqVMbhD2ksWr4QMrvuAppIS3qA4RV2otf4dMQ=</signature></info>
3 <reactionnetwork>
4 <metabolitepools>
5 <pool atoms="5" id="glu_DASH_L_c"/>
6 <pool atoms="3" id="pyr_c"/>
7 <pool atoms="4" id="succoa_c"/>
8 <pool id="atp_c"/>
9 <pool atoms="4" id="asp_DASH_L_c"/>
10 <pool id="nadph_c"/>
11 <pool atoms="7" id="26dap_DASH_M_c"/>
12 <pool atoms="4" id="succ_c"/>
13 <pool atoms="5" id="akg_c"/>
14 <pool atoms="6" id="cit_c"/>
15 <pool atoms="6" id="icit_c"/>
16 <pool id="nadh_c"/>
17 <pool atoms="1" id="co2_c"/>
18 <pool atoms="3" id="ala_DASH_L_c"/>
19 <pool atoms="5" id="gln_DASH_L_c"/>
20 <pool atoms="2" id="accoa_c"/>
21 <pool atoms="2" id="ac_c"/>
22 <pool atoms="6" id="arg_DASH_L_c"/>
23 <pool atoms="4" id="fum_c"/>
24 <pool id="nh4_c"/>
25 <pool atoms="4" id="asn_DASH_L_c"/>
26 <pool atoms="4" id="oaa_c"/>
27 <pool atoms="6" id="lys_DASH_L_c"/>
28 <pool atoms="6" id="2ddg6p_c"/>
29 <pool atoms="3" id="g3p_c"/>
30 <pool atoms="6" id="6pgc_c"/>
31 <pool atoms="1" id="co2_c_unlabeled"/>
32 <pool atoms="6" id="glc_DASH_D_e_ext"/>
33 <pool atoms="6" id="glc_DASH_D_e"/>
34 <pool id="nh4_c_ext"/>
35 <pool id="o2_c_ext"/>
36 <pool id="o2_c"/>
37 <pool id="so4_c_ext"/>
38 <pool id="so4_c"/>
39 <pool id="fadh2_c"/>
40 <pool atoms="6" id="fdp_c"/>
41 <pool atoms="3" id="dhap_c"/>
42 <pool atoms="4" id="mal_DASH_L_c"/>
43 <pool atoms="6" id="g6p_c"/>
44 <pool atoms="3" id="3pg_c"/>
45 <pool atoms="3" id="ser_DASH_L_c"/>
46 <pool atoms="2" id="gly_c"/>
47 <pool atoms="1" id="mlthf_c"/>

```

## A.2. Strain Design Algorithms for Target Identification

```

48 <pool atoms="3" id="pep_c"/>
49 <pool atoms="6" id="glc_DASH_D_c"/>
50 <pool atoms="5" id="ru5p_DASH_D_c"/>
51 <pool atoms="1" id="10thf_c"/>
52 <pool atoms="5" id="prpp_c"/>
53 <pool atoms="6" id="his_DASH_L_c"/>
54 <pool atoms="2" id="glx_c"/>
55 <pool atoms="4" id="thr_DASH_L_c"/>
56 <pool atoms="6" id="ile_DASH_L_c"/>
57 <pool atoms="6" id="leu_DASH_L_c"/>
58 <pool atoms="1" id="methf_c"/>
59 <pool atoms="3" id="cys_DASH_L_c"/>
60 <pool atoms="5" id="met_DASH_L_c"/>
61 <pool atoms="6" id="f6p_c"/>
62 <pool atoms="4" id="e4p_c"/>
63 <pool atoms="9" id="phe_DASH_L_c"/>
64 <pool atoms="5" id="pro_DASH_L_c"/>
65 <pool atoms="5" id="r5p_c"/>
66 <pool atoms="5" id="xu5p_DASH_D_c"/>
67 <pool atoms="7" id="s7p_c"/>
68 <pool atoms="3" id="TA_C3_c"/>
69 <pool atoms="2" id="TK_C2_c"/>
70 <pool atoms="11" id="trp_DASH_L_c"/>
71 <pool atoms="9" id="tyr_DASH_L_c"/>
72 <pool atoms="5" id="val_DASH_L_c"/>
73 <pool atoms="6" id="glcU"/>
74 <pool atoms="6" id="glc0"/>
75 <pool atoms="5" id="Xylose"/>
76 <pool atoms="5" id="Xylose_c"/>
77 </metabolitepools>
78 <reaction bidirectional="false" id="26dap_DASH_MSYN">
79 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
80 <reduct cfg="C#1@2 C#2@2 C#3@2" id="pyr_c"/>
81 <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3" id="succoa_c"/>
82 <reduct id="atp_c"/>
83 <reduct cfg="C#1@4 C#2@4 C#3@4 C#4@4" id="asp_DASH_L_c"/>
84 <reduct id="nadph_c"/>
85 <reduct id="nadph_c"/>
86 <rproduct cfg="C#1@4 C#2@4 C#3@4 C#4@4 C#3@2 C#2@2 C#1@2" id="26dap_DASH_M_c"/>
87 <rproduct cfg="C#1@3 C#2@3 C#3@3 C#4@3" id="succ_c"/>
88 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
89 </reaction>
90 <reaction bidirectional="true" id="ACONTa ACONTb">
91 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="cit_c"/>
92 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="icit_c"/>
93 </reaction>
94 <reaction bidirectional="false" id="AKGDH">
95 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
96 <rproduct id="nadh_c"/>
97 <rproduct cfg="C#2@1 C#3@1 C#4@1 C#5@1" id="succoa_c"/>
98 <rproduct cfg="C#1@1" id="co2_c"/>
99 </reaction>
100 <reaction bidirectional="false" id="ALATA L">
101 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
102 <reduct cfg="C#1@2 C#2@2 C#3@2" id="pyr_c"/>
103 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
104 <rproduct cfg="C#1@2 C#2@2 C#3@2" id="ala_DASH_L_c"/>
105 </reaction>
106 <reaction bidirectional="false" id="ArgSYN">
107 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
108 <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="gln_DASH_L_c"/>
109 <reduct cfg="C#1@3" id="co2_c"/>
110 <reduct id="atp_c"/>
111 <reduct id="atp_c"/>
112 <reduct id="atp_c"/>
113 <reduct id="atp_c"/>
114 <reduct id="atp_c"/>
115 <reduct id="nadph_c"/>
116 <reduct cfg="C#1@4 C#2@4" id="accoa_c"/>
117 <reduct cfg="C#1@5 C#2@5 C#3@5 C#4@5" id="asp_DASH_L_c"/>
118 <rproduct cfg="C#1@4 C#2@4" id="ac_c"/>
119 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#1@3" id="arg_DASH_L_c"/>
120 <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="akg_c"/>
121 <rproduct cfg="C#1@5 C#2@5 C#3@5 C#4@5" id="fum_c"/>
122 </reaction>
123 <reaction bidirectional="false" id="ASNn">
124 <reduct id="atp_c"/>
125 <reduct id="atp_c"/>
126 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="asp_DASH_L_c"/>
127 <reduct id="nh4_c"/>
128 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="asn_DASH_L_c"/>
129 </reaction>
130 <reaction bidirectional="false" id="ASPTA">
131 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
132 <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2" id="oaa_c"/>
133 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
134 <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2" id="asp_DASH_L_c"/>
135 </reaction>
136 <reaction bidirectional="false" id="ATPM">
137 <reduct id="atp_c"/>
138 </reaction>
139 <reaction bidirectional="false" id="CS">

```

## A. Appendix for Strain Engineering

```
140 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="oaa_c"/>
141 <reduct cfg="C#1@2 C#2@2" id="accoa_c"/>
142 <rproduct cfg="C#4@1 C#3@1 C#2@1 C#2@2 C#1@2 C#1@1" id="cit_c"/>
143 </reaction>
144 <reaction bidirectional="false" id="DAPDC">
145 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1 C#7@1" id="26dap_DASH_M_c"/>
146 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="lys_DASH_L_c"/>
147 <rproduct cfg="C#7@1" id="co2_c"/>
148 </reaction>
149 <reaction bidirectional="false" id="EDA">
150 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="2ddg6p_c"/>
151 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
152 <rproduct cfg="C#4@1 C#5@1 C#6@1" id="g3p_c"/>
153 </reaction>
154 <reaction bidirectional="false" id="EDD">
155 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="6pgc_c"/>
156 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="2ddg6p_c"/>
157 </reaction>
158 <reaction bidirectional="false" id="EX_ac_LPAREN_e_RPAREN_">
159 <reduct cfg="C#1@1 C#2@1" id="ac_c"/>
160 </reaction>
161 <reaction bidirectional="false" id="EX_co2_LPAREN_e_RPAREN_">
162 <reduct cfg="C#1@1" id="co2_c"/>
163 </reaction>
164 <reaction bidirectional="false" id="EX_co2_LPAREN_e_RPAREN__unlabeled">
165 <reduct cfg="C#1@1" id="co2_c_unlabeled"/>
166 <rproduct cfg="C#1@1" id="co2_c"/>
167 </reaction>
168 <reaction bidirectional="false" id="EX_glc_LPAREN_e_RPAREN_">
169 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_e_ext"/>
170 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_e"/>
171 </reaction>
172 <reaction bidirectional="false" id="EX_nh4_LPAREN_e_RPAREN_">
173 <reduct id="nh4_c_ext"/>
174 <rproduct id="nh4_c"/>
175 </reaction>
176 <reaction bidirectional="false" id="EX_o2_LPAREN_e_RPAREN_">
177 <reduct id="o2_c_ext"/>
178 <rproduct id="o2_c"/>
179 </reaction>
180 <reaction bidirectional="false" id="EX_so4_LPAREN_e_RPAREN_">
181 <reduct id="so4_c_ext"/>
182 <rproduct id="so4_c"/>
183 </reaction>
184 <reaction bidirectional="false" id="FADR_NADH_CYTBD_HYD_ATPS4r">
185 <reduct id="fadh2_c"/>
186 <reduct id="fadh2_c"/>
187 <reduct id="o2_c"/>
188 <rproduct id="atp_c"/>
189 <rproduct id="atp_c"/>
190 </reaction>
191 <reaction bidirectional="true" id="FBA">
192 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="fdp_c"/>
193 <rproduct cfg="C#3@1 C#2@1 C#1@1" id="dhap_c"/>
194 <rproduct cfg="C#4@1 C#5@1 C#6@1" id="g3p_c"/>
195 </reaction>
196 <reaction bidirectional="true" id="FUM">
197 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="fum_c"/>
198 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="mal_DASH_L_c"/>
199 </reaction>
200 <reaction bidirectional="false" id="G6PDH2r_PGL">
201 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="g6p_c"/>
202 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="6pgc_c"/>
203 <rproduct id="nadph_c"/>
204 </reaction>
205 <reaction bidirectional="true" id="GAPD_PGK">
206 <reduct cfg="C#1@1 C#2@1 C#3@1" id="g3p_c"/>
207 <rproduct id="nadh_c"/>
208 <rproduct id="atp_c"/>
209 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="3pg_c"/>
210 </reaction>
211 <reaction bidirectional="true" id="GHMT2r">
212 <reduct cfg="C#1@1 C#2@1 C#3@1" id="ser_DASH_L_c"/>
213 <rproduct cfg="C#1@1 C#2@1" id="gly_c"/>
214 <rproduct cfg="C#3@1" id="mlthf_c"/>
215 </reaction>
216 <reaction bidirectional="false" id="GLCptspp">
217 <reduct cfg="C#1@1 C#2@1 C#3@1" id="pep_c"/>
218 <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2 C#6@2" id="glc_DASH_D_e"/>
219 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
220 <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2 C#6@2" id="g6p_c"/>
221 </reaction>
222 <reaction bidirectional="false" id="GLCt2pp">
223 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_e"/>
224 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_c"/>
225 </reaction>
226 <reaction bidirectional="false" id="GLNS">
227 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
228 <reduct id="atp_c"/>
229 <reduct id="nh4_c"/>
230 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="gln_DASH_L_c"/>
231 </reaction>
```

## A.2. Strain Design Algorithms for Target Identification

```

232 <reaction bidirectional="false" id="GluSYN">
233   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
234   <reduct id="nh4_c"/>
235   <reduct id="nadph_c"/>
236   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
237 </reaction>
238 <reaction bidirectional="false" id="GND">
239   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="6pgc_c"/>
240   <rproduct cfg="C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="ru5p_DASH_D_c"/>
241   <rproduct cfg="C#1@1" id="co2_c"/>
242   <rproduct id="nadph_c"/>
243 </reaction>
244 <reaction bidirectional="false" id="HEX1">
245   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_c"/>
246   <reduct id="atp_c"/>
247   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="g6p_c"/>
248 </reaction>
249 <reaction bidirectional="false" id="HisSYN">
250   <reduct cfg="C#1@1" id="10fthf_c"/>
251   <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="prpp_c"/>
252   <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3 C#5@3" id="gln_DASH_L_c"/>
253   <reduct id="atp_c"/>
254   <reduct id="atp_c"/>
255   <reduct id="atp_c"/>
256   <reduct id="atp_c"/>
257   <reduct cfg="C#1@4 C#2@4 C#3@4 C#4@4" id="asp_DASH_L_c"/>
258   <rproduct id="nadh_c"/>
259   <rproduct id="nadh_c"/>
260   <rproduct cfg="C#5@2 C#4@2 C#3@2 C#2@2 C#1@2 C#1@1" id="his_DASH_L_c"/>
261   <rproduct cfg="C#1@3 C#2@3 C#3@3 C#4@3 C#5@3" id="akg_c"/>
262   <rproduct cfg="C#1@4 C#2@4 C#3@4 C#4@4" id="fum_c"/>
263 </reaction>
264 <reaction bidirectional="true" id="ICDHyr">
265   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="icit_c"/>
266   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
267   <rproduct id="nadph_c"/>
268   <rproduct cfg="C#6@1" id="co2_c"/>
269 </reaction>
270 <reaction bidirectional="false" id="ICL">
271   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="icit_c"/>
272   <rproduct cfg="C#5@1 C#4@1 C#3@1 C#6@1" id="succ_c"/>
273   <rproduct cfg="C#1@1 C#2@1" id="glx_c"/>
274 </reaction>
275 <reaction bidirectional="false" id="IleSYN">
276   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
277   <reduct cfg="C#1@2 C#2@2 C#3@2" id="pyr_c"/>
278   <reduct id="nadph_c"/>
279   <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3" id="thr_DASH_L_c"/>
280   <rproduct cfg="C#1@3 C#2@3 C#2@2 C#3@3 C#4@3 C#3@2" id="ile_DASH_L_c"/>
281   <rproduct cfg="C#1@2" id="co2_c"/>
282   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
283   <rproduct id="nh4_c"/>
284 </reaction>
285 <reaction bidirectional="false" id="LeuSYN">
286   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
287   <reduct cfg="C#1@2 C#2@2 C#3@2" id="pyr_c"/>
288   <reduct cfg="C#1@3 C#2@3 C#3@3" id="pyr_c"/>
289   <reduct id="nadph_c"/>
290   <reduct cfg="C#1@4 C#2@4" id="accoa_c"/>
291   <rproduct id="nadh_c"/>
292   <rproduct cfg="C#1@4 C#2@4 C#2@2 C#2@3 C#3@3 C#3@2" id="leu_DASH_L_c"/>
293   <rproduct cfg="C#1@2" id="co2_c"/>
294   <rproduct cfg="C#1@3" id="co2_c"/>
295   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
296 </reaction>
297 <reaction bidirectional="false" id="MALS">
298   <reduct cfg="C#1@1 C#2@1" id="glx_c"/>
299   <reduct cfg="C#1@2 C#2@2" id="accoa_c"/>
300   <rproduct cfg="C#1@1 C#2@1 C#2@2 C#1@2" id="mal_DASH_L_c"/>
301 </reaction>
302 <reaction bidirectional="true" id="MDH">
303   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="mal_DASH_L_c"/>
304   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="oaa_c"/>
305   <rproduct id="nadh_c"/>
306 </reaction>
307 <reaction bidirectional="false" id="ME1">
308   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="mal_DASH_L_c"/>
309   <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
310   <rproduct cfg="C#4@1" id="co2_c"/>
311   <rproduct id="nadph_c"/>
312 </reaction>
313 <reaction bidirectional="false" id="ME2">
314   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="mal_DASH_L_c"/>
315   <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
316   <rproduct id="nadh_c"/>
317   <rproduct cfg="C#4@1" id="co2_c"/>
318 </reaction>
319 <reaction bidirectional="false" id="MetSYN">
320   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="succoa_c"/>
321   <reduct id="atp_c"/>
322   <reduct cfg="C#1@2" id="methf_c"/>
323   <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3" id="asp_DASH_L_c"/>

```

## A. Appendix for Strain Engineering

```
324 <reduct id="nadph_c"/>
325 <reduct id="nadph_c"/>
326 <reduct cfg="C#1@4 C#2@4 C#3@4" id="cys_DASH_L_c"/>
327 <rproduct cfg="C#1@4 C#2@4 C#3@4" id="pyr_c"/>
328 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="succ_c"/>
329 <rproduct cfg="C#1@3 C#2@3 C#3@3 C#4@3 C#1@2" id="met_DASH_L_c"/>
330 <rproduct id="nh4_c"/>
331 </reaction>
332 <reaction bidirectional="true" id="MlthfSYN">
333 <reduct cfg="C#1@1 C#2@1" id="gly_c"/>
334 <rproduct id="nadh_c"/>
335 <rproduct cfg="C#1@1" id="co2_c"/>
336 <rproduct cfg="C#2@1" id="mlthf_c"/>
337 <rproduct id="nh4_c"/>
338 </reaction>
339 <reaction bidirectional="false" id="MTHFC">
340 <reduct cfg="C#1@1" id="mlthf_c"/>
341 <rproduct cfg="C#1@1" id="10fthf_c"/>
342 <rproduct id="nadph_c"/>
343 </reaction>
344 <reaction bidirectional="false" id="MIHFD">
345 <reduct id="nadh_c"/>
346 <reduct cfg="C#1@1" id="mlthf_c"/>
347 <rproduct cfg="C#1@1" id="methf_c"/>
348 </reaction>
349 <reaction bidirectional="false" id="NADH_CYTBD_HYD_ATPS4r">
350 <reduct id="nadh_c"/>
351 <reduct id="nadh_c"/>
352 <reduct id="o2_c"/>
353 <rproduct id="atp_c"/>
354 <rproduct id="atp_c"/>
355 <rproduct id="atp_c"/>
356 <rproduct id="atp_c"/>
357 </reaction>
358 <reaction bidirectional="true" id="NADTRHD_THD2pp">
359 <reduct id="nadh_c"/>
360 <rproduct id="nadph_c"/>
361 </reaction>
362 <reaction bidirectional="false" id="PDH">
363 <reduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
364 <rproduct id="nadh_c"/>
365 <rproduct cfg="C#2@1 C#3@1" id="accoa_c"/>
366 <rproduct cfg="C#1@1" id="co2_c"/>
367 </reaction>
368 <reaction bidirectional="false" id="PFK">
369 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="f6p_c"/>
370 <reduct id="atp_c"/>
371 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="fdp_c"/>
372 </reaction>
373 <reaction bidirectional="true" id="PGI">
374 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="g6p_c"/>
375 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="f6p_c"/>
376 </reaction>
377 <reaction bidirectional="true" id="PGM">
378 <reduct cfg="C#1@1 C#2@1 C#3@1" id="3pg_c"/>
379 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pep_c"/>
380 </reaction>
381 <reaction bidirectional="false" id="PheSYN">
382 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="e4p_c"/>
383 <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="glu_DASH_L_c"/>
384 <reduct id="nadph_c"/>
385 <reduct id="atp_c"/>
386 <reduct cfg="C#1@3 C#2@3 C#3@3" id="pep_c"/>
387 <reduct cfg="C#1@4 C#2@4 C#3@4" id="pep_c"/>
388 <rproduct cfg="C#1@3 C#2@3 C#3@3 C#2@4 C#3@4 C#1@1 C#2@1 C#3@1 C#4@1" id="phe_DASH_L_c"/>
>
389 <rproduct cfg="C#1@4" id="co2_c"/>
390 <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="akg_c"/>
391 </reaction>
392 <reaction bidirectional="false" id="PPC">
393 <reduct cfg="C#1@1 C#2@1 C#3@1" id="pep_c"/>
394 <reduct cfg="C#1@2" id="co2_c"/>
395 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#1@2" id="oaa_c"/>
396 </reaction>
397 <reaction bidirectional="false" id="PPCK">
398 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="oaa_c"/>
399 <reduct id="atp_c"/>
400 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pep_c"/>
401 <rproduct cfg="C#4@1" id="co2_c"/>
402 </reaction>
403 <reaction bidirectional="false" id="ProSYN">
404 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
405 <reduct id="atp_c"/>
406 <reduct id="nadph_c"/>
407 <reduct id="nadph_c"/>
408 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="pro_DASH_L_c"/>
409 </reaction>
410 <reaction bidirectional="false" id="PRPPS">
411 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="r5p_c"/>
412 <reduct id="atp_c"/>
413 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="prpp_c"/>
414 </reaction>
```

## A.2. Strain Design Algorithms for Target Identification

```

415 <reaction bidirectional="true" id="PTAr_ACKr_ACS">
416 <reduct cfg="C#1@1 C#2@1" id="accoa_c"/>
417 <rproduct id="atp_c"/>
418 <rproduct cfg="C#1@1 C#2@1" id="ac_c"/>
419 </reaction>
420 <reaction bidirectional="false" id="PYK">
421 <reduct cfg="C#1@1 C#2@1 C#3@1" id="pep_c"/>
422 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="pyr_c"/>
423 <rproduct id="atp_c"/>
424 </reaction>
425 <reaction bidirectional="true" id="RPE">
426 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="ru5p_DASH_D_c"/>
427 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="xu5p_DASH_D_c"/>
428 </reaction>
429 <reaction bidirectional="true" id="RPI">
430 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="ru5p_DASH_D_c"/>
431 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="r5p_c"/>
432 </reaction>
433 <reaction bidirectional="false" id="SERAT_CYSS">
434 <reduct id="atp_c"/>
435 <reduct id="atp_c"/>
436 <reduct id="atp_c"/>
437 <reduct cfg="C#1@1 C#2@1 C#3@1" id="ser_DASH_L_c"/>
438 <reduct id="so4_c"/>
439 <reduct cfg="C#1@2 C#2@2" id="accoa_c"/>
440 <reduct id="nadph_c"/>
441 <reduct id="nadph_c"/>
442 <reduct id="nadph_c"/>
443 <reduct id="nadph_c"/>
444 <rproduct cfg="C#1@2 C#2@2" id="ac_c"/>
445 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="cys_DASH_L_c"/>
446 </reaction>
447 <reaction bidirectional="false" id="SerSYN">
448 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
449 <reduct cfg="C#1@2 C#2@2 C#3@2" id="3pg_c"/>
450 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
451 <rproduct id="nadh_c"/>
452 <rproduct cfg="C#1@2 C#2@2 C#3@2" id="ser_DASH_L_c"/>
453 </reaction>
454 <reaction bidirectional="true" id="SUCCOAS">
455 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="succoa_c"/>
456 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="succ_c"/>
457 <rproduct id="atp_c"/>
458 </reaction>
459 <reaction bidirectional="true" id="SUCDi">
460 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="succ_c"/>
461 <rproduct id="fadh2_c"/>
462 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="fum_c"/>
463 </reaction>
464 <reaction bidirectional="true" id="TALA">
465 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1 C#7@1" id="s7p_c"/>
466 <rproduct cfg="C#4@1 C#5@1 C#6@1 C#7@1" id="e4p_c"/>
467 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="TA_C3_c"/>
468 </reaction>
469 <reaction bidirectional="false" id="THRD_GLYAT">
470 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="thr_DASH_L_c"/>
471 <rproduct id="nadh_c"/>
472 <rproduct cfg="C#1@1 C#2@1" id="gly_c"/>
473 <rproduct cfg="C#3@1 C#4@1" id="accoa_c"/>
474 </reaction>
475 <reaction bidirectional="false" id="ThrSYN">
476 <reduct id="atp_c"/>
477 <reduct id="atp_c"/>
478 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="asp_DASH_L_c"/>
479 <reduct id="nadph_c"/>
480 <reduct id="nadph_c"/>
481 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="thr_DASH_L_c"/>
482 </reaction>
483 <reaction bidirectional="false" id="TKT1a">
484 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1 C#7@1" id="s7p_c"/>
485 <rproduct cfg="C#1@1 C#2@1" id="TK_C2_c"/>
486 <rproduct cfg="C#3@1 C#4@1 C#5@1 C#6@1 C#7@1" id="r5p_c"/>
487 </reaction>
488 <reaction bidirectional="false" id="TKT1a1">
489 <reduct cfg="C#1@1 C#2@1" id="TK_C2_c"/>
490 <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="r5p_c"/>
491 <rproduct cfg="C#1@1 C#2@1 C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="s7p_c"/>
492 </reaction>
493 <reaction bidirectional="false" id="TKT1b">
494 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="f6p_c"/>
495 <rproduct cfg="C#1@1 C#2@1 C#3@1" id="TA_C3_c"/>
496 <rproduct cfg="C#4@1 C#5@1 C#6@1" id="g3p_c"/>
497 </reaction>
498 <reaction bidirectional="false" id="TKT1b1">
499 <reduct cfg="C#1@1 C#2@1 C#3@1" id="TA_C3_c"/>
500 <reduct cfg="C#1@2 C#2@2 C#3@2" id="g3p_c"/>
501 <rproduct cfg="C#1@1 C#2@1 C#3@1 C#1@2 C#2@2 C#3@2" id="f6p_c"/>
502 </reaction>
503 <reaction bidirectional="true" id="TKT2a">
504 <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="xu5p_DASH_D_c"/>
505 <rproduct cfg="C#1@1 C#2@1" id="TK_C2_c"/>
506 <rproduct cfg="C#3@1 C#4@1 C#5@1" id="g3p_c"/>

```

## A. Appendix for Strain Engineering

```
507 </reaction>
508 <reaction bidirectional="true" id="TKT2b">
509   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="f6p_c"/>
510   <rproduct cfg="C#3@1 C#4@1 C#5@1 C#6@1" id="e4p_c"/>
511   <rproduct cfg="C#1@1 C#2@1" id="TK_C2_c"/>
512 </reaction>
513 <reaction bidirectional="true" id="TPI">
514   <reduct cfg="C#1@1 C#2@1 C#3@1" id="dhap_c"/>
515   <rproduct cfg="C#1@1 C#2@1 C#3@1" id="g3p_c"/>
516 </reaction>
517 <reaction bidirectional="false" id="TrpSYN">
518   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="e4p_c"/>
519   <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="gln_DASH_L_c"/>
520   <reduct id="nadph_c"/>
521   <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3 C#5@3" id="r5p_c"/>
522   <reduct id="atp_c"/>
523   <reduct id="atp_c"/>
524   <reduct id="atp_c"/>
525   <reduct cfg="C#1@4 C#2@4 C#3@4" id="ser_DASH_L_c"/>
526   <reduct cfg="C#1@5 C#2@5 C#3@5" id="pep_c"/>
527   <reduct cfg="C#1@6 C#2@6 C#3@6" id="pep_c"/>
528   <rproduct cfg="C#1@6 C#2@6 C#3@6" id="pyr_c"/>
529   <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="glu_DASH_L_c"/>
530   <rproduct cfg="C#3@3 C#4@3 C#5@3" id="g3p_c"/>
531   <rproduct cfg="C#1@5" id="co2_c"/>
532   <rproduct cfg="C#1@4 C#2@4 C#3@4 C#2@3 C#1@3 C#3@5 C#1@1 C#2@1 C#3@1 C#4@1 C#2@5" id="
trp_DASH_L_c"/>
533 </reaction>
534 <reaction bidirectional="false" id="TyrSYN">
535   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1" id="e4p_c"/>
536   <reduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="glu_DASH_L_c"/>
537   <reduct id="nadph_c"/>
538   <reduct id="atp_c"/>
539   <reduct cfg="C#1@3 C#2@3 C#3@3" id="pep_c"/>
540   <reduct cfg="C#1@4 C#2@4 C#3@4" id="pep_c"/>
541   <rproduct id="nadh_c"/>
542   <rproduct cfg="C#1@3 C#2@3 C#3@3 C#2@4 C#3@4 C#1@1 C#2@1 C#3@1 C#4@1" id="tyr_DASH_L_c"/>
543 >
544   <rproduct cfg="C#1@4" id="co2_c"/>
545   <rproduct cfg="C#1@2 C#2@2 C#3@2 C#4@2 C#5@2" id="akg_c"/>
546 </reaction>
547 <reaction bidirectional="false" id="ValSYN">
548   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="glu_DASH_L_c"/>
549   <reduct cfg="C#1@2 C#2@2 C#3@2" id="pyr_c"/>
550   <reduct cfg="C#1@3 C#2@3 C#3@3" id="pyr_c"/>
551   <reduct id="nadph_c"/>
552   <rproduct cfg="C#1@3" id="co2_c"/>
553   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="akg_c"/>
554   <rproduct cfg="C#1@2 C#2@2 C#3@2 C#2@3 C#3@3" id="val_DASH_L_c"/>
555 </reaction>
556 <reaction bidirectional="false" id="uptU">
557   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glcU"/>
558   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_e_ext"/>
559 </reaction>
560 <reaction bidirectional="false" id="upt0">
561   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc0"/>
562   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1" id="glc_DASH_D_e_ext"/>
563 </reaction>
564 <reaction bidirectional="false" id="Pentose_Uptake">
565   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="Xylose"/>
566   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="Xylose_c"/>
567 </reaction>
568 <reaction bidirectional="false" id="Xylulokinase">
569   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="Xylose_c"/>
570   <reduct id="atp_c"/>
571   <rproduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1" id="xu5p_DASH_D_c"/>
572 </reaction>
573 <reaction id="BM_phe_DASH_L_c">
574   <reduct cfg="C#1@1 C#2@1 C#3@1 C#4@1 C#5@1 C#6@1 C#7@1 C#8@1 C#9@1" id="phe_DASH_L_c"/>
575 </reaction>
576 <reaction id="BM_mlthf_c">
577   <reduct cfg="C#1@2" id="mlthf_c"/>
578 </reaction>
579 <reaction id="BM_oaa_c">
580   <reduct cfg="C#1@3 C#2@3 C#3@3 C#4@3" id="oaa_c"/>
581 </reaction>
582 <reaction id="BM_lys_DASH_L_c">
583   <reduct cfg="C#1@4 C#2@4 C#3@4 C#4@4 C#5@4 C#6@4" id="lys_DASH_L_c"/>
584 </reaction>
585 <reaction id="BM_atp_c">
586   <reduct id="atp_c"/>
587 </reaction>
588 <reaction id="BM_ser_DASH_L_c">
589   <reduct cfg="C#1@5 C#2@5 C#3@5" id="ser_DASH_L_c"/>
590 </reaction>
591 <reaction id="BM_g3p_c">
592   <reduct cfg="C#1@6 C#2@6 C#3@6" id="g3p_c"/>
593 </reaction>
594 <reaction id="BM_tyr_DASH_L_c">
595   <reduct cfg="C#1@7 C#2@7 C#3@7 C#4@7 C#5@7 C#6@7 C#7@7 C#8@7 C#9@7" id="tyr_DASH_L_c"/>
596 </reaction>
597 <reaction id="BM_pep_c">
```

## A.2. Strain Design Algorithms for Target Identification

```

597     <reduct cfg="C#1@8 C#2@8 C#3@8" id="pep_c"/>
598 </reaction>
599 <reaction id="BM_met_DASH_L_c">
600   <reduct cfg="C#1@9 C#2@9 C#3@9 C#4@9 C#5@9" id="met_DASH_L_c"/>
601 </reaction>
602 <reaction id="BM_g6p_c">
603   <reduct cfg="C#1@10 C#2@10 C#3@10 C#4@10 C#5@10 C#6@10" id="g6p_c"/>
604 </reaction>
605 <reaction id="BM_akg_c">
606   <reduct cfg="C#1@11 C#2@11 C#3@11 C#4@11 C#5@11" id="akg_c"/>
607 </reaction>
608 <reaction id="BM_glu_DASH_L_c">
609   <reduct cfg="C#1@12 C#2@12 C#3@12 C#4@12 C#5@12" id="glu_DASH_L_c"/>
610 </reaction>
611 <reaction id="BM_gln_DASH_L_c">
612   <reduct cfg="C#1@13 C#2@13 C#3@13 C#4@13 C#5@13" id="gln_DASH_L_c"/>
613 </reaction>
614 <reaction id="BM_r5p_c">
615   <reduct cfg="C#1@14 C#2@14 C#3@14 C#4@14 C#5@14" id="r5p_c"/>
616 </reaction>
617 <reaction id="BM_f6p_c">
618   <reduct cfg="C#1@15 C#2@15 C#3@15 C#4@15 C#5@15 C#6@15" id="f6p_c"/>
619 </reaction>
620 <reaction id="BM_pyr_c">
621   <reduct cfg="C#1@16 C#2@16 C#3@16" id="pyr_c"/>
622 </reaction>
623 <reaction id="BM_gly_c">
624   <reduct cfg="C#1@17 C#2@17" id="gly_c"/>
625 </reaction>
626 <reaction id="BM_thr_DASH_L_c">
627   <reduct cfg="C#1@18 C#2@18 C#3@18 C#4@18" id="thr_DASH_L_c"/>
628 </reaction>
629 <reaction id="BM_asp_DASH_L_c">
630   <reduct cfg="C#1@19 C#2@19 C#3@19 C#4@19" id="asp_DASH_L_c"/>
631 </reaction>
632 <reaction id="BM_nadph_c">
633   <reduct id="nadph_c"/>
634 </reaction>
635 <reaction id="BM_cys_DASH_L_c">
636   <reduct cfg="C#1@20 C#2@20 C#3@20" id="cys_DASH_L_c"/>
637 </reaction>
638 <reaction id="BM_3pg_c">
639   <reduct cfg="C#1@21 C#2@21 C#3@21" id="3pg_c"/>
640 </reaction>
641 <reaction id="BM_val_DASH_L_c">
642   <reduct cfg="C#1@22 C#2@22 C#3@22 C#4@22 C#5@22" id="val_DASH_L_c"/>
643 </reaction>
644 <reaction id="BM_ala_DASH_L_c">
645   <reduct cfg="C#1@23 C#2@23 C#3@23" id="ala_DASH_L_c"/>
646 </reaction>
647 <reaction id="BM_ile_DASH_L_c">
648   <reduct cfg="C#1@24 C#2@24 C#3@24 C#4@24 C#5@24 C#6@24" id="ile_DASH_L_c"/>
649 </reaction>
650 <reaction id="BM_asn_DASH_L_c">
651   <reduct cfg="C#1@25 C#2@25 C#3@25 C#4@25" id="asn_DASH_L_c"/>
652 </reaction>
653 <reaction id="BM_his_DASH_L_c">
654   <reduct cfg="C#1@26 C#2@26 C#3@26 C#4@26 C#5@26 C#6@26" id="his_DASH_L_c"/>
655 </reaction>
656 <reaction id="BM_leu_DASH_L_c">
657   <reduct cfg="C#1@27 C#2@27 C#3@27 C#4@27 C#5@27 C#6@27" id="leu_DASH_L_c"/>
658 </reaction>
659 <reaction id="BM_accoa_c">
660   <reduct cfg="C#1@28 C#2@28" id="accoa_c"/>
661 </reaction>
662 <reaction id="BM_arg_DASH_L_c">
663   <reduct cfg="C#1@29 C#2@29 C#3@29 C#4@29 C#5@29 C#6@29" id="arg_DASH_L_c"/>
664 </reaction>
665 <reaction id="BM_pro_DASH_L_c">
666   <reduct cfg="C#1@30 C#2@30 C#3@30 C#4@30 C#5@30" id="pro_DASH_L_c"/>
667 </reaction>
668 <reaction id="BM_trp_DASH_L_c">
669   <reduct cfg="C#1@31 C#2@31 C#3@31 C#4@31 C#5@31 C#6@31 C#7@31 C#8@31 C#9@31 C#10@31 C#11
    @31" id="trp_DASH_L_c"/>
670 </reaction>
671 </reactionnetwork>
672 <constraints>
673   <net>
674     <textual>BM_phe_DASH_L_c=0.0882992;BM_mlthf_c=0.2222531;BM_oaa_c=0.170578;
    BM_lys_DASH_L_c=0.1635542;BM_atp_c=16.6800199;BM_ser_DASH_L_c=0.1028485;BM_g3p_c=0.0647193;
    BM_tyr_DASH_L_c=0.0657227;BM_pep_c=0.0255867;BM_met_DASH_L_c=0.0732482;BM_g6p_c=0.1028485;
    BM_akg_c=0.0436479;BM_glu_DASH_L_c=0.125425;BM_gln_DASH_L_c=0.125425;BM_r5p_c=0.3782818;
    BM_f6p_c=0.0356207;BM_pyr_c=0.0416411;BM_gly_c=0.2919894;BM_thr_DASH_L_c=0.1209097;
    BM_asp_DASH_L_c=0.1148893;BM_nadph_c=2.6906171;BM_cys_DASH_L_c=0.0436479;BM_3pg_c=0.3105523;
    BM_val_DASH_L_c=0.2016834;BM_ala_DASH_L_c=0.2448296;BM_ile_DASH_L_c=0.1384692;
    BM_asn_DASH_L_c=0.1148893;BM_his_DASH_L_c=0.045153;BM_leu_DASH_L_c=0.2147276;BM_accoa_c
    =1.259267;BM_arg_DASH_L_c=0.1409777;BM_pro_DASH_L_c=0.105357;BM_trp_DASH_L_c=0.0270918;
    EX_co2_LPAREN_e_RPAREN__unlabeled &lt;=20</textual>
675   </net>
676   <xch>
677     <textual/>
678   </xch>

```

## A. Appendix for Strain Engineering

```
679 </constraints>
680 <configuration name="default">
681   <comment>Eine Belegung bla bla</comment>
682   <input pool="glc0" type="isotopomer">
683     <label cfg="110111">0.0016</label>
684     <label cfg="011111">0.0016</label>
685     <label cfg="101111">0.0016</label>
686     <label cfg="111101">0.0016</label>
687     <label cfg="111111">0.9904</label>
688     <label cfg="111110">0.0016</label>
689     <label cfg="111011">0.0016</label>
690   </input>
691   <input pool="glcU" type="isotopomer">
692     <label cfg="000100">0.0107</label>
693     <label cfg="000000">0.9358</label>
694     <label cfg="000001">0.0107</label>
695     <label cfg="000010">0.0107</label>
696     <label cfg="001000">0.0107</label>
697     <label cfg="010000">0.0107</label>
698     <label cfg="100000">0.0107</label>
699   </input>
700   <input pool="Xylose" type="isotopomer">
701     <label cfg="00000">0.9462</label>
702     <label cfg="10000">0.01076</label>
703     <label cfg="01000">0.01076</label>
704     <label cfg="00100">0.01076</label>
705     <label cfg="00010">0.01076</label>
706     <label cfg="00001">0.01076</label>
707   </input>
708   <input pool="co2_c_unlabeled" type="isotopomer">
709     <label cfg="0">0.989</label>
710     <label cfg="1">0.011</label>
711   </input>
712   <input pool="so4_c_ext" />
713   <input pool="nh4_c_ext" />
714   <input pool="o2_c_ext" />
715 </measurement>
716 <model>
717   <labelingmeasurement>
718     <group id="ms_group_1" scale="auto">
719       <textual>ala_DASH_L_c[1,2,3]#M0,1,2,3</textual>
720     </group>
721     <group id="ms_group_2" scale="auto">
722       <textual>asp_DASH_L_c[1,2,3,4]#M0,1,2,3,4</textual>
723     </group>
724     <group id="ms_group_3" scale="auto">
725       <textual>glu_DASH_L_c[1,2,3,4,5]#M0,1,2,3,4,5</textual>
726     </group>
727     <group id="ms_group_4" scale="auto">
728       <textual>gly_c[1,2]#M0,1,2</textual>
729     </group>
730     <group id="ms_group_5" scale="auto">
731       <textual>his_DASH_L_c[1,2,3,4,5,6]#M0,1,2,3,4,5,6</textual>
732     </group>
733     <group id="ms_group_6" scale="auto">
734       <textual>ile_DASH_L_c[2,3,4,5,6]#M0,1,2,3,4,5</textual>
735     </group>
736     <group id="ms_group_7" scale="auto">
737       <textual>leu_DASH_L_c[2,3,4,5,6]#M0,1,2,3,4,5</textual>
738     </group>
739     <group id="ms_group_8" scale="auto">
740       <textual>lys_DASH_L_c[1,2,3,4,5,6]#M0,1,2,3,4,5,6</textual>
741     </group>
742     <group id="ms_group_9" scale="auto">
743       <textual>met_DASH_L_c[1,2,3,4,5]#M0,1,2,3,4,5</textual>
744     </group>
745     <group id="ms_group_10" scale="auto">
746       <textual>phe_DASH_L_c[1,2,3,4,5,6,7,8,9]#M0,1,2,3,4,5,6,7,8,9</textual>
747     </group>
748     <group id="ms_group_11" scale="auto">
749       <textual>pro_DASH_L_c[1,2,3,4,5]#M0,1,2,3,4,5</textual>
750     </group>
751     <group id="ms_group_12" scale="auto">
752       <textual>ser_DASH_L_c[1,2,3]#M0,1,2,3</textual>
753     </group>
754     <group id="ms_group_13" scale="auto">
755       <textual>thr_DASH_L_c[1,2,3,4]#M0,1,2,3,4</textual>
756     </group>
757     <group id="ms_group_14" scale="auto">
758       <textual>tyr_DASH_L_c[1,2,3,4,5,6,7,8,9]#M0,1,2,3,4,5,6,7,8,9</textual>
759     </group>
760     <group id="ms_group_15" scale="auto">
761       <textual>val_DASH_L_c[1,2,3,4,5]#M0,1,2,3,4,5</textual>
762     </group>
763   </labelingmeasurement>
764   <fluxmeasurement>
765     <netflux id="fm_Glucose">
766       <textual>EX_glc_LPAREN_e_RPAREN_</textual>
767     </netflux>
768     <netflux id="fm_Xylose">
769       <textual>Pentose_Uptake</textual>
770     </netflux>
```

## A.2. Strain Design Algorithms for Target Identification

```
771 </fluxmeasurement>
772 </model>
773 <data>
774 <datum id="fm_Glucose" stddev="0.2552">5.8296</datum>
775 <datum id="fm_Xylose" stddev="0.48">2.277</datum>
776 <datum id="ms_group_1" stddev="0.0098765" weight="0">0.8623</datum>
777 <datum id="ms_group_1" stddev="0.00092696" weight="1">0.015</datum>
778 <datum id="ms_group_1" stddev="0.0009844" weight="2">0.0123</datum>
779 <datum id="ms_group_1" stddev="0.0079651" weight="3">0.1105</datum>
780 <datum id="ms_group_2" stddev="0.0082325" weight="0">0.7323</datum>
781 <datum id="ms_group_2" stddev="0.0017888" weight="1">0.1086</datum>
782 <datum id="ms_group_2" stddev="0.0024468" weight="2">0.0894</datum>
783 <datum id="ms_group_2" stddev="0.002975" weight="3">0.0586</datum>
784 <datum id="ms_group_2" stddev="0.0010218" weight="4">0.0111</datum>
785 <datum id="ms_group_3" stddev="0.0083226" weight="0">0.6666</datum>
786 <datum id="ms_group_3" stddev="0.0034515" weight="1">0.1053</datum>
787 <datum id="ms_group_3" stddev="0.0028649" weight="2">0.1687</datum>
788 <datum id="ms_group_3" stddev="0.0017457" weight="3">0.0432</datum>
789 <datum id="ms_group_3" stddev="0.00014063" weight="4">0.0118</datum>
790 <datum id="ms_group_3" stddev="0.00040113" weight="5">0.0045</datum>
791 <datum id="ms_group_4" stddev="0.0076784" weight="0">0.8572</datum>
792 <datum id="ms_group_4" stddev="0.00088828" weight="1">0.0243</datum>
793 <datum id="ms_group_4" stddev="0.0067902" weight="2">0.1185</datum>
794 <datum id="ms_group_5" stddev="0.020841" weight="0">0.7336</datum>
795 <datum id="ms_group_5" stddev="0.007452" weight="1">0.1165</datum>
796 <datum id="ms_group_5" stddev="0.0040407" weight="2">0.0376</datum>
797 <datum id="ms_group_5" stddev="0.0054956" weight="3">0.0625</datum>
798 <datum id="ms_group_5" stddev="0.00064921" weight="4">0.0124</datum>
799 <datum id="ms_group_5" stddev="0.003035" weight="5">0.034</datum>
800 <datum id="ms_group_5" stddev="0.00016856" weight="6">0.0034</datum>
801 <datum id="ms_group_6" stddev="0.011413" weight="0">0.6671</datum>
802 <datum id="ms_group_6" stddev="0.00015775" weight="1">0.1072</datum>
803 <datum id="ms_group_6" stddev="0.0081215" weight="2">0.1689</datum>
804 <datum id="ms_group_6" stddev="0.0020657" weight="3">0.0428</datum>
805 <datum id="ms_group_6" stddev="0.00078686" weight="4">0.0098</datum>
806 <datum id="ms_group_6" stddev="0.00028126" weight="5">0.0043</datum>
807 <datum id="ms_group_7" stddev="0.015476" weight="0">0.6597</datum>
808 <datum id="ms_group_7" stddev="0.001692" weight="1">0.1075</datum>
809 <datum id="ms_group_7" stddev="0.010628" weight="2">0.1851</datum>
810 <datum id="ms_group_7" stddev="0.0019916" weight="3">0.0306</datum>
811 <datum id="ms_group_7" stddev="0.00085108" weight="4">0.0147</datum>
812 <datum id="ms_group_7" stddev="0.00031291" weight="5">0.0024</datum>
813 <datum id="ms_group_8" stddev="0.011261" weight="0">0.6487</datum>
814 <datum id="ms_group_8" stddev="0.0010794" weight="1">0.1045</datum>
815 <datum id="ms_group_8" stddev="0.0035929" weight="2">0.1243</datum>
816 <datum id="ms_group_8" stddev="0.0042259" weight="3">0.0965</datum>
817 <datum id="ms_group_8" stddev="0.001673" weight="4">0.016</datum>
818 <datum id="ms_group_8" stddev="0.00068325" weight="5">0.0088</datum>
819 <datum id="ms_group_8" stddev="6.5616e-06" weight="6">0.0012</datum>
820 <datum id="ms_group_9" stddev="0.0057048" weight="0">0.6414</datum>
821 <datum id="ms_group_9" stddev="0.00021786" weight="1">0.1828</datum>
822 <datum id="ms_group_9" stddev="0.00067911" weight="2">0.0943</datum>
823 <datum id="ms_group_9" stddev="0.0033826" weight="3">0.0622</datum>
824 <datum id="ms_group_9" stddev="0.0013483" weight="4">0.0178</datum>
825 <datum id="ms_group_9" stddev="7.7032e-05" weight="5">0.0015</datum>
826 <datum id="ms_group_10" stddev="0.020755" weight="0">0.6167</datum>
827 <datum id="ms_group_10" stddev="0.0018601" weight="1">0.0558</datum>
828 <datum id="ms_group_10" stddev="0.0053606" weight="2">0.0902</datum>
829 <datum id="ms_group_10" stddev="0.0074037" weight="3">0.1525</datum>
830 <datum id="ms_group_10" stddev="0.0027818" weight="4">0.0406</datum>
831 <datum id="ms_group_10" stddev="0.0018944" weight="5">0.0224</datum>
832 <datum id="ms_group_10" stddev="0.00071459" weight="6">0.0145</datum>
833 <datum id="ms_group_10" stddev="0.00047423" weight="7">0.0061</datum>
834 <datum id="ms_group_10" stddev="0.00012597" weight="8">0.0008</datum>
835 <datum id="ms_group_10" stddev="0.00013908" weight="9">0.0003</datum>
836 <datum id="ms_group_11" stddev="0.0068072" weight="0">0.6718</datum>
837 <datum id="ms_group_11" stddev="0.0014376" weight="1">0.1094</datum>
838 <datum id="ms_group_11" stddev="0.0031831" weight="2">0.1562</datum>
839 <datum id="ms_group_11" stddev="0.0013378" weight="3">0.0445</datum>
840 <datum id="ms_group_11" stddev="0.00053207" weight="4">0.0138</datum>
841 <datum id="ms_group_11" stddev="0.00031655" weight="5">0.0044</datum>
842 <datum id="ms_group_12" stddev="0.0073209" weight="0">0.8249</datum>
843 <datum id="ms_group_12" stddev="0.00059233" weight="1">0.0503</datum>
844 <datum id="ms_group_12" stddev="0.00095619" weight="2">0.05</datum>
845 <datum id="ms_group_12" stddev="0.0057724" weight="3">0.0748</datum>
846 <datum id="ms_group_13" stddev="0.0064421" weight="0">0.7055</datum>
847 <datum id="ms_group_13" stddev="0.00033316" weight="1">0.1371</datum>
848 <datum id="ms_group_13" stddev="0.0021757" weight="2">0.0856</datum>
849 <datum id="ms_group_13" stddev="0.0038007" weight="3">0.0661</datum>
850 <datum id="ms_group_13" stddev="0.00079882" weight="4">0.0057</datum>
851 <datum id="ms_group_14" stddev="0.026183" weight="0">0.632</datum>
852 <datum id="ms_group_14" stddev="0.0036002" weight="1">0.0422</datum>
853 <datum id="ms_group_14" stddev="0.0067108" weight="2">0.1012</datum>
854 <datum id="ms_group_14" stddev="0.0077757" weight="3">0.1343</datum>
855 <datum id="ms_group_14" stddev="0.0048849" weight="4">0.0439</datum>
856 <datum id="ms_group_14" stddev="0.0011994" weight="5">0.0247</datum>
857 <datum id="ms_group_14" stddev="0.00091415" weight="6">0.0145</datum>
858 <datum id="ms_group_14" stddev="0.0006957" weight="7">0.0055</datum>
859 <datum id="ms_group_14" stddev="0.00012668" weight="8">0.0016</datum>
860 <datum id="ms_group_14" stddev="0.00019422" weight="9">0.0001</datum>
861 <datum id="ms_group_15" stddev="0.0076514" weight="0">0.7464</datum>
862 <datum id="ms_group_15" stddev="0.001291" weight="1">0.0185</datum>
```

## A. Appendix for Strain Engineering

```

863     <datum id="ms_group_15" stddev="0.0044788" weight="2">0.1155</datum>
864     <datum id="ms_group_15" stddev="0.003517" weight="3">0.1017</datum>
865     <datum id="ms_group_15" stddev="0.00026097" weight="4">0.0021</datum>
866     <datum id="ms_group_15" stddev="0.0006857" weight="5">0.0158</datum>
867   </data>
868 </measurement>
869 <simulation>
870   <variables/>
871 </simulation>
872 </configuration>
873 </fluxml>

```

### A.2.5. Constraint-based Assessment of Taxadiene Production Potential

Table A.9 contains confidence intervals for the taxadiene export reaction, calculated using the method described in Section 2.1.2

Table A.9.: **Confidence Intervals for Taxadiene Production Flux Calculated Considering  $^{13}\text{C}$  Labeling Data and Exchange Rates.** Minimal and maximal production flux were calculated using an iterative approach. Taxadiene flux was increased (or decreased) until the squared sum of residuals (SSR) surpassed a boundary defined by the minimal SSR + 3.84. Refer to Section 2.1.2 for details. Mean taxadiene production rate was calculated as follows:  $10^{(\log_{10}(\min)+\log_{10}(\max))/2}$ . *min* and *max* refer to the minimal and maximal taxadiene flux, respectively.

	Production Rate [mmol/gDW h]			Production Yield [mol/mol]		
	min	mean	max	min	mean	max
HMS_2p	$1.464 \times 10^{-7}$	$1.824 \times 10^{-5}$	$2.273 \times 10^{-3}$	$3.240 \times 10^{-8}$	$4.037 \times 10^{-6}$	$5.030 \times 10^{-4}$
$\Delta_{\text{xylE}}_2\text{P}$	$1.739 \times 10^{-7}$	$2.551 \times 10^{-5}$	$3.741 \times 10^{-3}$	$3.068 \times 10^{-8}$	$4.500 \times 10^{-6}$	$6.600 \times 10^{-4}$
$\Delta_{\text{xylE}}_1\text{P}$	$1.385 \times 10^{-7}$	$2.431 \times 10^{-4}$	$4.267 \times 10^{-1}$	$1.629 \times 10^{-8}$	$2.860 \times 10^{-5}$	$5.020 \times 10^{-2}$
$\Delta_{\text{ptsG}}_2\text{P}$	$1.433 \times 10^{-7}$	$2.421 \times 10^{-5}$	$4.092 \times 10^{-3}$	$2.528 \times 10^{-8}$	$4.272 \times 10^{-6}$	$7.220 \times 10^{-4}$
$\Delta_{\text{ptsG}}_1\text{P}$	$1.412 \times 10^{-7}$	$7.279 \times 10^{-5}$	$3.751 \times 10^{-2}$	$1.977 \times 10^{-8}$	$1.019 \times 10^{-5}$	$5.250 \times 10^{-3}$

## A.3. Simultaneous Utilization of D-Xylose and Glucose in *E. coli*

### A.3.1. Concentration Time Courses for All Strains

Concentration time courses for biomass, glucose, xylose and acetate for six different strains are shown in Table A.10. Measurements are provided in duplicates. Table A.11 contains experimental re-characterization of parent strains  $\Delta_{\text{ptsG}}$  Empty and HMS Empty.

### A.3.2. $^{13}\text{C}$ Labeling Data

$^{13}\text{C}$  Labeling data for the six strains used in the sugar cu-utilization studies is given in Table A.12. Experimental data was kindly provided by Claudia Huber at **Bio**.

A.3. Simultaneous Utilization of D-Xylose and Glucose in *E. coli*

Table A.10.: Concentration Time Courses For Six Strains

Time, [h]	OD		Glucose, [mmol/l]			Xylose, [mmol/l]			Acetate, [mmol/l]	
	1	2	1	2	1	2	1	2	1	2
<b>HMS p1x</b>										
0	0.079	0.067	13.138	12.901	17.205	17.057	0.000	0.000		
3	0.338	0.288	11.559	11.841	17.015	17.220	2.597	2.276		
5	0.912	0.824	7.483	8.499	16.072	16.638	7.978	6.811		
7	2.204	2.120	0.000	0.000	11.306	13.110	15.569	15.335		
<b>HMS p3x</b>										
0	0.088	0.092	12.928	12.895	17.201	17.096	0.000	0.000		
3	0.204	0.136	11.798	12.479	16.849	17.310	1.988	1.386		
5	0.474	0.232	9.884	11.877	16.179	17.197	4.351	2.807		
7	1.308	0.544	4.606	0.046	14.260	0.063	9.370	1.944		
<b><math>\Delta</math>ptsG p1x</b>										
0	0.053	0.051	12.916	12.825	17.009	17.057	0.000	0.000		
3	0.156	0.152	12.415	12.549	16.692	16.694	0.930	0.959		
5	0.386	0.354	11.511	11.770	15.295	15.387	3.125	2.808		
7	1.048	0.980	8.906	10.034	11.063	12.359	8.606	7.348		
<b><math>\Delta</math>ptsG p3x</b>										
0	0.082	0.079	12.817	12.772	16.654	16.593	0.000	0.000		
3	0.168	0.175	12.713	12.300	16.446	15.812	1.148	1.277		
5	0.352	0.378	12.308	11.820	14.766	13.920	3.267	3.747		
7	0.868	0.952	11.499	10.921	10.205	8.678	7.939	9.109		
<b>HMS Empty</b>										
0	0.097	0.098	13.847	14.173	18.277	18.701	0.000	0.000		
3	0.239	0.240	12.077	11.781	16.592	16.261	1.008	1.086		
5	0.373	0.379	10.870	10.704	16.425	16.341	2.663	2.751		
7	0.586	0.610	9.025	8.653	16.301	15.993	4.871	5.141		
<b><math>\Delta</math>ptsG Empty</b>										
0	0.096	0.086	12.580	12.698	16.428	16.386	0.000	0.000		
3	0.405	0.389	12.195	12.155	14.577	14.892	0.767	0.830		
5	1.120	1.120	10.652	10.530	10.014	10.415	1.468	1.870		
7	2.518	2.593	5.919	5.578	1.889	1.842	1.838	1.944		

Table A.11.: Concentration Time Courses For Parent Strains: Re-measurements.

Time, [h]	OD			Glucose, [mmol/l]			Xylose, [mmol/l]			Acetate, [mmol/l]		
	1	2	3	1	2	3	1	2	3	1	2	3
<b>HMS Empty2</b>												
0	0.107	0.111	0.109	13.222	13.197	13.106	16.681	16.719	16.419	0.000	0.000	0.000
1	0.155	0.158	0.161	–	–	–	–	–	–	–	–	–
2	0.268	0.273	0.278	–	–	–	–	–	–	–	–	–
3	0.511	0.519	0.543	10.646	10.676	10.645	15.936	16.007	15.981	2.530	2.363	2.363
4	0.921	0.935	0.934	–	–	–	–	–	–	–	–	–
5	1.464	1.424	1.520	4.220	3.784	3.950	15.508	15.356	15.378	6.406	6.340	6.541
6	2.105	2.138	2.123	–	–	–	–	–	–	–	–	–
7	2.910	2.830	3.090	0.000	0.000	0.000	5.977	5.632	5.914	8.728	8.857	10.051
<b><math>\Delta</math>ptsG Empty2</b>												
0	0.064	0.066	0.060	13.037	13.104	13.156	16.318	16.399	16.354	0.000	0.000	0.000
1	0.080	0.076	0.069	–	–	–	–	–	–	–	–	–
2	0.105	0.101	0.092	–	–	–	–	–	–	–	–	–
3	0.150	0.147	0.138	12.514	12.449	12.492	15.329	15.207	15.322	0.000	0.000	0.000
4	0.242	0.240	0.216	–	–	–	–	–	–	–	–	–
5	0.383	0.360	0.216	11.530	11.514	11.584	12.468	12.740	12.698	0.981	0.705	0.644
6	0.533	0.573	0.331	–	–	–	–	–	–	–	–	–
7	0.930	0.927	0.535	8.852	9.004	9.178	6.660	6.785	7.072	2.412	2.327	1.924

A. Appendix for Strain Engineering

Table A.12.: <sup>13</sup>C-labeling Data for Strains Used in Sugar Co-Utilization Studies. Labeling patterns of protein-bound amino acids was experimentally determined in duplicate for six strains exhibiting different values for % Carbon Xylose.

Amino Acid	Fragment	HMS Empty %	ΔptsG %	Empty σ	HMS p1x %	HMS p3x %	ΔptsG %	p1x σ	ΔptsG %	p3x σ			
Ala-260	M+0	0.857	4.434E-04	0.923	6.815E-04	0.862	7.822E-03	0.872	1.802E-02	0.911	2.463E-03	0.949	3.479E-03
Ala-260	M+1	0.014	4.323E-04	0.010	8.531E-04	0.015	1.615E-03	0.014	3.245E-03	0.012	1.212E-03	0.008	2.467E-03
Ala-260	M+2	0.012	1.982E-04	0.008	1.599E-04	0.012	9.147E-04	0.011	1.547E-03	0.009	8.427E-04	0.006	6.036E-04
Ala-260	M+3	0.117	2.618E-04	0.059	4.262E-04	0.110	6.218E-03	0.104	1.341E-02	0.068	1.532E-03	0.037	6.941E-04
Asp-418	M+0	0.759	2.173E-03	0.859	2.004E-03	0.732	6.553E-03	0.779	3.910E-02	0.839	6.984E-03	0.904	4.898E-03
Asp-418	M+1	0.094	1.504E-03	0.065	2.056E-03	0.109	2.235E-03	0.085	1.946E-02	0.069	5.482E-03	0.045	4.081E-03
Asp-418	M+2	0.085	1.220E-03	0.047	8.150E-04	0.089	1.916E-03	0.070	1.188E-02	0.051	1.532E-03	0.051	1.856E-03
Asp-418	M+3	0.053	4.046E-04	0.026	2.757E-04	0.059	2.413E-03	0.057	5.596E-03	0.038	6.172E-04	0.020	7.195E-04
Asp-418	M+4	0.010	1.885E-04	0.002	2.194E-04	0.011	9.798E-04	0.009	2.474E-03	0.003	4.629E-04	0.001	2.560E-04
Gluc-432	M+0	0.668	9.711E-04	0.817	2.135E-03	0.667	7.060E-03	0.711	3.916E-02	0.794	6.844E-03	0.873	4.763E-03
Gluc-432	M+1	0.102	1.104E-03	0.066	2.763E-03	0.105	3.219E-03	0.080	1.564E-02	0.066	3.889E-03	0.042	2.586E-03
Gluc-432	M+2	0.168	7.077E-04	0.097	7.856E-04	0.169	2.894E-03	0.155	1.501E-02	0.113	2.302E-03	0.073	2.099E-03
Gluc-432	M+3	0.043	3.788E-04	0.016	6.368E-04	0.043	1.735E-03	0.039	6.558E-03	0.021	1.352E-03	0.011	8.033E-04
Gluc-432	M+4	0.014	4.510E-04	0.003	1.425E-04	0.012	6.841E-04	0.010	1.588E-03	0.004	4.772E-04	0.002	2.362E-04
Gluc-432	M+5	0.004	1.502E-04	0.001	9.422E-05	0.004	4.350E-04	0.004	6.825E-04	0.001	1.345E-04	0.001	9.866E-05
Gly-246	M+0	0.855	4.280E-04	0.925	6.503E-04	0.857	6.000E-03	0.871	1.728E-02	0.914	1.366E-03	0.950	8.671E-04
Gly-246	M+1	0.029	2.275E-04	0.016	6.754E-04	0.024	7.335E-04	0.023	3.172E-03	0.017	1.098E-03	0.011	8.399E-04
Gly-246	M+2	0.116	3.736E-04	0.059	3.171E-04	0.119	5.736E-03	0.106	1.414E-02	0.069	1.574E-03	0.039	4.329E-04
His-440	M+0	0.733	1.901E-03	0.884	2.267E-03	0.734	1.688E-02	0.755	3.538E-02	0.862	4.877E-03	0.924	5.207E-03
His-440	M+1	0.101	1.907E-03	0.061	2.900E-03	0.117	7.659E-03	0.100	1.684E-02	0.073	5.377E-03	0.040	4.246E-03
His-440	M+2	0.047	6.245E-04	0.010	8.228E-04	0.038	3.513E-03	0.039	5.358E-03	0.011	3.966E-03	0.005	1.542E-03
His-440	M+3	0.066	5.609E-04	0.034	7.664E-04	0.063	4.311E-03	0.056	7.779E-03	0.034	7.729E-04	0.020	8.922E-04
His-440	M+4	0.014	8.053E-05	0.003	2.328E-04	0.012	6.548E-04	0.011	2.713E-03	0.004	4.954E-04	0.002	2.735E-04
His-440	M+5	0.035	1.938E-04	0.009	1.505E-04	0.034	2.394E-03	0.035	3.920E-03	0.016	6.045E-04	0.009	3.208E-04
His-440	M+6	0.005	2.197E-04	0.000	8.030E-05	0.003	4.935E-04	0.004	8.892E-04	0.001	2.352E-04	0.000	9.922E-05
Ile-200	M+0	0.696	5.368E-04	0.822	2.101E-03	0.667	8.884E-03	0.718	4.259E-02	0.800	5.154E-03	0.880	1.920E-03
Ile-200	M+1	0.091	6.325E-04	0.065	7.628E-04	0.107	8.431E-04	0.080	1.733E-02	0.065	2.342E-03	0.044	1.495E-03
Ile-200	M+2	0.159	1.729E-04	0.095	1.175E-03	0.169	6.299E-03	0.152	1.673E-02	0.110	1.869E-03	0.065	1.264E-03
Ile-200	M+3	0.038	7.809E-05	0.015	1.918E-04	0.043	1.495E-03	0.038	6.337E-03	0.020	9.163E-04	0.010	1.771E-04
Ile-200	M+4	0.011	9.602E-05	0.002	5.270E-05	0.010	6.609E-04	0.009	1.558E-03	0.003	1.829E-04	0.001	4.447E-05
Ile-200	M+5	0.004	7.405E-05	0.001	1.921E-05	0.004	2.223E-04	0.004	6.733E-04	0.001	4.360E-05	0.000	3.346E-05
Leu-274	M+0	0.686	1.351E-03	0.815	1.756E-03	0.660	1.238E-02	0.699	3.836E-02	0.787	6.876E-03	0.872	2.491E-03
Leu-274	M+1	0.091	1.444E-03	0.063	1.025E-03	0.108	2.958E-03	0.090	1.214E-02	0.072	4.123E-03	0.048	2.318E-03
Leu-274	M+2	0.174	7.278E-04	0.110	1.024E-03	0.185	8.551E-03	0.168	2.012E-02	0.123	2.135E-03	0.074	2.662E-03
Leu-274	M+3	0.031	5.213E-04	0.008	3.440E-04	0.031	1.790E-03	0.027	3.714E-03	0.012	7.280E-04	0.004	3.840E-04
Leu-274	M+4	0.015	1.254E-04	0.004	1.005E-04	0.015	7.352E-04	0.013	2.049E-03	0.005	5.188E-04	0.002	2.012E-04
Leu-274	M+5	0.003	5.968E-05	0.000	3.242E-05	0.002	2.514E-04	0.002	4.332E-04	0.001	8.311E-05	0.000	6.790E-05
Lys-431	M+0	0.684	2.154E-03	0.814	3.905E-03	0.649	9.282E-03	0.706	4.289E-02	0.786	5.965E-03	0.878	4.303E-03
Lys-431	M+1	0.088	1.193E-03	0.065	2.772E-03	0.104	5.735E-03	0.078	1.464E-02	0.067	3.613E-03	0.041	2.937E-03
Lys-431	M+2	0.115	9.071E-04	0.071	6.617E-04	0.124	3.645E-03	0.106	1.270E-02	0.081	2.679E-03	0.049	1.403E-03
Lys-431	M+3	0.087	4.226E-04	0.049	6.542E-04	0.097	3.485E-03	0.089	9.894E-03	0.061	8.854E-04	0.036	7.789E-04
Lys-431	M+4	0.016	3.977E-04	0.001	2.095E-04	0.016	1.410E-03	0.013	4.230E-03	0.003	8.854E-04	0.000	0.000E+00
Lys-431	M+5	0.009	1.788E-04	0.001	3.152E-04	0.009	6.625E-04	0.008	1.241E-03	0.002	3.143E-04	0.000	1.804E-04
Lys-431	M+6	0.001	1.761E-04	0.000	0.000E+00	0.001	1.732E-04	0.001	5.612E-04	0.000	0.000E+00	0.000	0.000E+00
Met-320	M+0	0.688	6.278E-03	0.810	1.531E-03	0.641	6.083E-03	0.699	4.676E-02	0.779	2.572E-03	0.867	4.377E-03
Met-320	M+1	0.157	2.158E-03	0.110	8.997E-04	0.183	2.847E-03	0.153	2.554E-02	0.122	3.064E-03	0.077	3.042E-03
Met-320	M+2	0.085	1.476E-03	0.050	7.271E-04	0.094	2.713E-03	0.075	1.216E-02	0.055	1.376E-03	0.034	2.034E-03
Met-320	M+3	0.054	2.044E-03	0.026	5.467E-04	0.062	3.083E-03	0.056	7.303E-03	0.038	9.150E-04	0.020	6.745E-04
Met-320	M+4	0.014	1.527E-03	0.004	1.157E-04	0.018	1.157E-03	0.016	3.445E-03	0.006	4.056E-04	0.002	6.650E-04
Met-320	M+5	0.001	4.198E-05	0.000	1.472E-04	0.002	1.530E-03	0.003	6.385E-03	0.001	2.203E-04	0.000	1.377E-04
Phe-336	M+0	0.642	1.271E-03	0.803	2.399E-03	0.617	1.612E-02	0.654	4.447E-02	0.767	4.030E-03	0.859	1.993E-03
Phe-336	M+1	0.052	1.013E-03	0.030	2.031E-03	0.056	1.962E-03	0.050	5.393E-03	0.038	1.818E-03	0.029	8.042E-04
Phe-336	M+2	0.082	8.989E-04	0.056	1.319E-03	0.090	4.457E-03	0.082	9.137E-03	0.064	1.558E-03	0.039	8.560E-04
Phe-336	M+3	0.133	7.613E-04	0.087	8.181E-04	0.152	5.832E-03	0.135	1.564E-02	0.101	2.051E-03	0.060	5.451E-04
Phe-336	M+4	0.046	3.019E-04	0.014	8.085E-04	0.041	2.588E-03	0.040	7.047E-03	0.017	6.505E-04	0.010	6.233E-04
Phe-336	M+5	0.022	2.690E-04	0.006	2.474E-04	0.022	1.509E-03	0.020	3.410E-03	0.008	2.644E-04	0.003	1.027E-04
Phe-336	M+6	0.015	2.696E-04	0.003	2.875E-04	0.014	6.457E-04	0.013	2.487E-03	0.004	2.963E-04	0.001	1.151E-04
Phe-336	M+7	0.007	2.078E-04	0.001	1.386E-04	0.006	3.811E-04	0.006	1.305E-03	0.001	7.124E-05	0.000	7.881E-05
Phe-336	M+8	0.001	1.836E-04	0.000	0.000E+00	0.001	1.154E-04	0.001	2.278E-04	0.000	0.000E+00	0.000	0.000E+00
Phe-336	M+9	0.000	5.768E-05	0.000	0.000E+00	0.000	1.492E-04	0.000	2.029E-04	0.000	0.000E+00	0.000	0.000E+00
Pro-286	M+0	0.701	3.194E-03	0.824	5.201E-03	0.672	6.054E-03	0.725	4.085E-02	0.800	5.682E-03	0.876	2.853E-03
Pro-286	M+1	0.095	1.727E-03	0.067	1.498E-03	0.109	3.772E-03	0.083	1.548E-02	0.072	2.381E-03	0.049	1.832E-03
Pro-286	M+2	0.147	1.163E-03	0.084	2.745E-03	0.156	3.015E-03	0.137	1.675E-02	0.098	3.110E-03	0.059	1.723E-03
Pro-286	M+3	0.039	6.929E-04	0.017	5.813E-04	0.045	2.061E-03	0.039	6.088E-03	0.023	1.004E-03	0.012	2.514E-04
Pro-286	M+4	0.015	4.270E-04	0.006	9.635E-04	0.014	6.280E-04	0.012	1.992E-03	0.006	2.702E-04	0.004	1.736E-04
Pro-286	M+5	0.004	3.393E-04	0.001	2.084E-04	0.004	3.092E-04	0.004	6.673E-04	0.001	1.323E-04	0.000	4.366E-05
Ser-390	M+0	0.851	2.035E-03	0.915	4.296E-03	0.825	6.231E-03	0.844	1.950E-02	0.903	2.932E-03	0.951	3.158E-03
Ser-390	M+1	0.030	1.424E-03	0.024	3.353E-03	0.050	2.335E-03	0.044	6.015E-03	0.025	2.248E-03	0.010	2.810E-03
Ser-390	M+2	0.027	6.824E-04	0.022	1.241E-03	0.050	1.969E-03	0.044	5.218E-03	0.025	1.007E-03	0.015	9.845E-04
Ser-390	M+3	0.091	5.789E-04	0.039	3.264E-04	0.075	4.638E-03	0.068	9.498E-03	0.047	1.092E-03	0.024	9.457E-04
Thr-404	M+0	0.748	2.069E-03	0.854	2.469E-03	0.706	5.359E-03	0.757	3.921E-02	0.820	5.313E-03	0.884	3.150E-03
Thr-404	M+1	0.107	2.375E-03	0.075	1.708E-03	0.137	1.468E-03	0.109	2.137E-02	0.091	4.376E-03	0.070	3.007E-03
Thr-404	M+2	0.086	9.052E-04	0.051	1.410E-03	0.086	2.303E-03	0.068	9.30				

### A.3. Simultaneous Utilization of D-Xylose and Glucose in *E. coli*

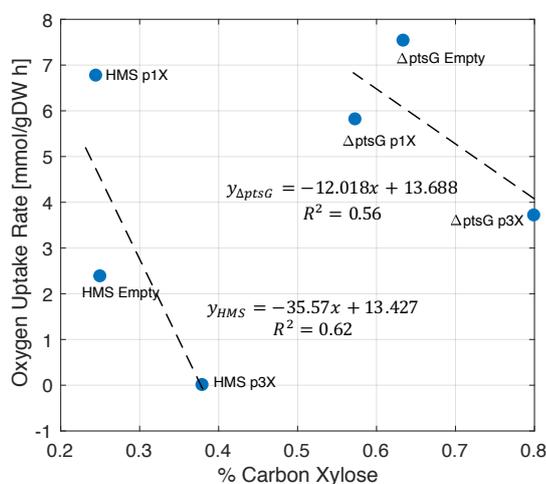


Figure A.8.: **Estimation of Oxygen Uptake Rate as a Function of Sugar Co-Uptake.** Overall, we observe a decrease in the oxygen uptake rate, as % Carbon Xylose increases. For strain HMS p3x, its unusually high substrate uptake (refer to Figure 6.1.C) leads to an *in silico* oxygen uptake rate of zero.

#### A.3.3. Dependency of Oxygen Uptake Rate on Sugar Co-Utilization

Figure A.8 shows the dependence of *in silico estimated* oxygen uptake for strain families HMS and  $\Delta$ ptsG on the extent of experimentally determined xylose and glucose co-utilization. Oxygen uptake rate for each strain is estimated by minimizing the reaction representing oxygen import from the medium to the cell, while simultaneously constraining reactions in the iJO1366 model representing rates for growth, glucose and xylose uptake, and acetate secretion to match experimentally determined values, as listed in Table 6.2. The biological feasibility of minimization of oxygen uptake as objective function is discussed in Chapter 4.

#### A.3.4. Experimental Protein Content

Total protein content was experimentally determined in two rounds. In the first round, strains carrying mutated XylE transporters were characterized. In a second round, total protein content of respective parent strains carrying an empty plasmid was determined. Results are shown in Figure A.9.A and B., respectively. Overall, we observe an increase in the total protein content as % Carbon Xylose increases. Additionally, strain family  $\Delta$ ptsG exhibits a higher total protein content compared to protein contents measured for the strain family XylE. Since the 2-D Quant Kit (provided by GE Healthcare) used for protein measurement is an indirect method, it requires the simultaneous measurement of a protein standard. Thus, absolute protein content might vary between experimental rounds, which is the case for protein levels within the strain family  $\Delta$ ptsG. As such, we expect the protein content within this family to follow the following trend:  $\Delta$ ptsG Empty <  $\Delta$ ptsG p1x <  $\Delta$ ptsG p3x. By contrast, the observed trend is as follows:  $\Delta$ ptsG p1x <  $\Delta$ ptsG p3x <  $\Delta$ ptsG Empty. Since it is very improbable that the parent strain  $\Delta$ ptsG Empty exhibits a higher protein content than both daughter strains, we conclude that observed deviations are originated because protein content for strain  $\Delta$ ptsG Empty was determined

A. Appendix for Strain Engineering

independently from  $\Delta$ ptsG p1x and  $\Delta$ ptsG p3x. Consequently, protein measurements should be repeated for all strains using the same protein standard and experimental set up.

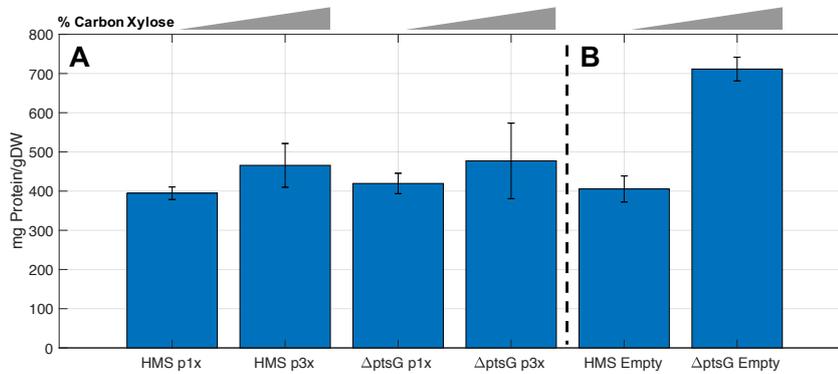


Figure A.9.: **Protein Measurement.** Total protein content for strains carrying Xyle mutant transporters in two different genetic backgrounds (**A.**) and corresponding parent strains (**B.**). Protein content was determined using the 2-D Quant Kit provided by GE Healthcare for the two strain groups separately.

## B. Appendix for Process Engineering

### B.1. Model-based Medium Optimization

#### B.1.1. Extending Metabolic Models to Allow Taxadiene Synthesis

##### *E. coli* Core

The code shown below was used to extend the *E. coli* core model to enable taxadiene production using the non-mevalonate pathway, the mevalonate pathway or both.

```
1 % Load core model.
2 model=readCbModel('ecoli_core_model.xml');
3
4 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
5 %
6 % Add reactions of the MEP pathway.
7 %
8 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
9
10 % Reaction 1. 1-Deoxyululose-5-phosphate synthesis from g3p and pyr.
11 model=addReaction(model, 'dxs', 'pyr[c] + g3p[c] + h[c] -> dxy15p[c] + co2[c]');
12
13 % Reaction 2. ispC
14 model=addReaction(model, 'ispC', 'dxy15p[c] + nadph[c] + h[c] -> nadp[c] + 2me4p[c]');
15
16 % Reaction 3. ispD
17 model=addReaction(model, 'ispD', '2me4p[c] + atp[c] + h[c] -> 4c2me[c] + 2 pi[c] '); %Eigentlich
    ist die Reaktion mit CTP.
18
19 % Reaction 4. ispE
20 model=addReaction(model, 'ispE', '4c2me[c] + atp[c] -> 2p4c2me[c] + adp[c] + h[c] ');
21
22 % Reaction 5. ispF
23 model=addReaction(model, 'ispF', '2p4c2me[c] -> 2mecdp[c] + amp[c] ');
24
25 % Reaction 6. ispG
26 model=addReaction(model, 'ispG', '2mecdp[c] + nadh[c] -> h2mb4p[c] + h2o[c] + nad[c] ');
27
28 % Reaction 7. ispH
29 model=addReaction(model, 'ispH', 'h2mb4p[c] + nadh[c] + h[c] -> dmpp[c] + h2o[c] + nad[c] ');
30
31 % Reaction 8. ispH2
32 model=addReaction(model, 'ispH2', 'h2mb4p[c] + nadh[c] + h[c] -> ipdp[c] + h2o[c] + nad[c] ');
33
34 % Reaction 9. idi
35 model=addReaction(model, 'idi', 'dmpp[c] <=> ipdp[c] ');
36
37 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
38 %
39 % Add reactions of the MVA pathway.
40 %
41 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
42
43
44 % Reaction 1. - synthesis of Acetoacetyl.
45 model=addReaction(model, 'atoB', '2 accoa[c] -> aacoa[c] + coa[c]');
46
47 %Reaction 2. - synthesis of HMGCoA.
48 model=addReaction(model, 'pksg', 'aacoa[c] + accoa[c] + h2o[c] -> h[c] + coa[c] + hmgcoa[c]');
49
50 %Reaction 3 - synthesis of MVA
51 model=addReaction(model, 'hmg1', 'hmgcoa[c] + 2 nadph[c] + 2 h[c] -> 2 nadp[c] + coa[c] + mva[c]');
52
53 %Reaction 4 - synthesis of MVAP
54 model=addReaction(model, 'mvak1', 'mva[c] + atp[c] -> adp[c] + h[c] mvap[c]');
55
56 %Reaction 5 - synthesis of MVAPP
57 model=addReaction(model, 'mvak2', 'mvap[c] + atp[c] -> adp[c] + mvapp[c]');
58
```

## B. Appendix for Process Engineering

```
59 %Reaction 6 - synthesis of IPP
60 model=addReaction(model, 'mvad', 'mvapp[c] + atp[c] -> adp[c] + co2[c] + pi[c] + ipdp[c]');
61
62 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
63 %
64 % Add taxadiene synthesis from IPP (ipdp) and DMAPP (dmpp)
65 %
66 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
67
68 %Reaction 1. - synthesis of Geranyl diphosphate
69 model=addReaction(model, 'DMAT1', 'dmpp[c] + ipdp[c] -> grdp[c] + 2 pi[c] ');
70
71 %Reaction 2. - synthesis of Farnesyl diphosphate
72 model=addReaction(model, 'GRIT1', 'grdp[c] + ipdp[c] -> frdp[c] + 2 pi[c] ');
73
74 %Reaction 3. - synthesis of GGPP
75 model=addReaction(model, 'ggpps', 'frdp[c] + ipdp[c] -> ggpp[c] + 2 pi[c] ');
76
77 %Reaction 4. - synthesis of taxadiene
78 model=addReaction(model, 'txs', 'ggpp[c] -> txdn[c] + 2 pi[c] ');
79
80 %Reaction 5. - export reaction for taxadiene
81 model=addReaction(model, 'txdnx', 'txdn[c] -> '); %txdn[e]
```

## E. coli iJO1366

The code shown below was used to extend the *E. coli* genome-scale metabolic reconstruction iJO1366 to enable taxadiene production using the non-mevalonate pathway, the mevalonate pathway or both. Note that the non-mevalonate pathway is an endogenous pathway and thus already contained in the iJO1366 model. Reactions DXPIi, MEPCT, CDPMEX, MECDPS, MECDPDH2, DMPPS, IPDPS should be deleted if theoretical calculations using the mevalonate pathway for taxadiene synthesis are to be performed.

```
1 % Load Genome Scale Metabolic Reconstruction.
2 model0=readCbModel('iJO1366.xml');
3
4 %NGAM modification
5 model0=changeRxnBounds(model0, 'ATPM', 0, 'b');
6
7 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
8 %
9 %Control activity of endogeneous MEP Pathway.
10 %
11 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
12
13 if 1==0
14 rxnRemoveList={'DXPIi', 'MEPCT', 'CDPMEX', 'MECDPS', 'MECDPDH2', 'DMPPS', 'IPDPS'};
15 model0=removeRxn(model0, rxnRemoveList);
16 end
17
18 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
19 %
20 %Add MVA Pathway. This pathway is composed of 6 reactions.
21 %
22 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
23
24 if 1==1 % MVA aktiv?
25
26 %Reaction 1. - synthesis of Acetoacetyl-coA, is already contained in the
27 %model.
28
29 %Reaction 2. - synthesis of HMGCoA.
30 model0=addReaction(model0, 'pksg', 'accoa[c] + accoa[c] + h2o[c] -> h[c] + coa[c] + hmgcoa[c]');
31
32 %Reaction 3. - synthesis of MVA
33 model0=addReaction(model0, 'hmg1', 'hmgcoa[c] + 2 nadph[c] + 2 h[c] -> 2 nadp[c] + coa[c] + mva[c]');
34
35 %Reaction 4. - synthesis of MVAP
36 model0=addReaction(model0, 'mvak1', 'mva[c] + atp[c] -> adp[c] + h[c] mvap[c]');
37
38 %Reaction 5. - synthesis of MVAPP
39 model0=addReaction(model0, 'mvak2', 'mvap[c] + atp[c] -> adp[c] + mvapp[c]');
40
41 %Reaction 6. - synthesis of IPP
42 model0=addReaction(model0, 'mvad', 'mvapp[c] + atp[c] -> adp[c] + co2[c] + pi[c] + ipdp[c]');
43
44 end
45
46 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
47 %
48 %Taxadiene Synthesis
```

## B.1. Model-based Medium Optimization

```

49 %
50 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
51
52 %Reaktion 1. - synthesis of GGPP from farnesyl diphosphate and isopentenyl
53 %diphosphate.
54 model0=addReaction(model0,'ggpps','frdp[c] + ipdp[c] -> ggpp[c] + ppi[c]');
55
56 %Reaktion 2. - synthesis of taxadiene from GGPP
57 model0=addReaction(model0,'txs','ggpp[c] -> txdn[c] + ppi[c]');
58
59 %Reaktion 3. - taxadiene export reaction
60 model0=addReaction(model0,'txdnx','txdn[c] -> ');

```

### B.1.2. Ranking of 180 Substrates According to Yields

Table B.1 provides maximal theoretical taxadiene yields for 180 different growth supporting carbon sources. Two different network configurations are considered. In the first case, IPP and DMAPP synthesis can occur only via the non-mevalonate pathway. In the second case, both the mevalonate and the non-mevalonate pathway are considered. For each substrate and the non-mevalonate pathway, yields are reported in three different basis. When both pathways are considered, only carbon yield is reported. Synergy is calculated as explained in the main text.

Table B.1.: **Maximal Taxadiene Yields Calculated for 180 Carbon Sources.** The *E. coli* metabolic reconstruction iJO1366 was used for all calculations. Yields are reported for the non-mevalonate pathway (MEP). For synergy calculation, the iJO1366 model was expanded to consider the mevalonate (MVA) pathway. Synergy was calculated as explained in the main text. Refer to Section 7.2. Carbon yield for the simultaneous pathway usage (MEP+MVA) is also provided. For all calculations, NGAM and growth rate were set to 0.

Nr	Carbon Source	Non-Mevalonate Pathway			MEP+MVA. Carbon Yield	Synergy, [%]	Best Pathway
		Molar	Carbon	Mass			
1	Ethanol exchange	0.088	0.879	0.519	0.925	5.205	MEP+MVA
2	Glycerophosphoglycerol exchange	0.250	0.833	0.278	0.833	0.000	MEP
3	Glycerol 3-phosphate exchange	0.125	0.833	0.200	0.833	0.000	MEP
4	sn-Glycero-3-phosphoethanolamine exchange	0.204	0.815	0.258	0.828	1.699	MEP+MVA
5	Glycerol exchange'	0.121	0.804	0.356	0.805	0.188	MEP+MVA
6	Glycerol 2-phosphate exchange	0.121	0.804	0.193	0.805	0.188	MEP+MVA
7	Putrescine exchange	0.159	0.793	0.479	0.793	0.000	MEP
8	(S)-Propane-1,2-diol exchange	0.113	0.755	0.405	0.765	1.275	MEP+MVA
9	Acetaldehyde exchange	0.074	0.742	0.459	0.798	0.000	MVA
10	D-Mannitol exchange	0.220	0.735	0.329	0.740	0.696	MEP+MVA
11	D-Sorbitol exchange	0.220	0.735	0.329	0.740	0.696	MEP+MVA
12	Galactitol exchange	0.220	0.733	0.328	0.740	1.016	MEP+MVA
13	Ethanolamine exchange	0.072	0.723	0.317	0.777	0.000	MVA
14	Propanal exchange	0.108	0.718	0.505	0.731	1.723	MEP+MVA
15	D-Glucose 6-phosphate exchange	0.213	0.711	0.225	0.714	0.420	MEP+MVA
16	D-Mannose 6-phosphate exchange	0.213	0.711	0.225	0.714	0.420	MEP+MVA
17	D-fructose 6-phosphate exchange	0.213	0.711	0.225	0.714	0.420	MEP+MVA
18	D-Glucosamine 6-phosphate exchange	0.213	0.711	0.225	0.714	0.420	MEP+MVA
19	Maltohexaose exchange	1.259	0.700	0.346	0.704	0.555	MEP+MVA
20	1,4-alpha-D-glucan exchange	1.259	0.700	0.346	0.704	0.555	MEP+MVA
21	D-Glyceraldehyde exchange	0.105	0.698	0.316	0.706	1.224	MEP+MVA
22	Glycerophosphoserine exchange	0.209	0.697	0.221	0.705	1.092	MEP+MVA
23	Maltopentaose exchange	1.044	0.696	0.343	0.700	0.600	MEP+MVA
24	Sucrose exchange	0.414	0.690	0.329	0.695	0.667	MEP+MVA
25	Trehalose exchange	0.414	0.690	0.329	0.695	0.667	MEP+MVA
26	Dihydroxyacetone exchange	0.104	0.690	0.313	0.695	0.667	MEP+MVA
28	Maltotriose exchange	0.621	0.690	0.335	0.695	0.667	MEP+MVA
27	Maltose exchange	0.414	0.690	0.329	0.695	0.667	MEP+MVA
29	Maltotetraose exchange	0.828	0.690	0.338	0.695	0.667	MEP+MVA
30	D-Glucose 1-phosphate exchange	0.207	0.690	0.218	0.695	0.738	MEP+MVA
31	D-Glucose exchange'	0.207	0.690	0.313	0.695	0.738	MEP+MVA
32	D-Fructose exchange	0.207	0.689	0.312	0.695	0.871	MEP+MVA
33	D-Mannose exchange	0.207	0.689	0.312	0.695	0.871	MEP+MVA
34	D-Glucosamine exchange	0.207	0.689	0.312	0.695	0.871	MEP+MVA
35	Melibiose exchange	0.412	0.687	0.328	0.692	0.709	MEP+MVA
36	Lactose exchange	0.412	0.687	0.328	0.692	0.709	MEP+MVA
37	beta D-Galactose exchange	0.205	0.683	0.310	0.689	0.753	MEP+MVA
38	D-Galactose exchange	0.205	0.683	0.310	0.689	0.753	MEP+MVA
39	alpha-D-Galactose 1-phosphate exchange	0.205	0.683	0.216	0.689	0.753	MEP+MVA
40	L-Lyxose exchange	0.171	0.682	0.309	0.687	0.770	MEP+MVA
41	L-Xylulose exchange	0.171	0.682	0.309	0.687	0.770	MEP+MVA
42	L-Arabinose exchange	0.171	0.682	0.309	0.687	0.770	MEP+MVA
43	D-Xylose exchange	0.171	0.682	0.309	0.687	0.770	MEP+MVA
44	L-Rhamnose exchange	0.204	0.679	0.338	0.685	0.867	MEP+MVA

B. Appendix for Process Engineering

Table B.1 Continued from previous page

Nr	Carbon Source	Non Mevalonate Pathway Molar	Carbon	Mass	MEP+MVA. Carbon Yield	Synergy, [%]	Best Pathway
45	L-Fucose exchange	0.204	0.679	0.338	0.685	0.867	MEP+MVA
46	4-Aminobutanoate exchange	0.134	0.672	0.355	0.672	0.000	MEP
47	D-Allose exchange	0.199	0.662	0.300	0.668	0.878	MEP+MVA
48	chitobiose exchange	0.528	0.660	0.339	0.672	1.712	MEP+MVA
49	D-Ribose exchange	0.164	0.657	0.298	0.663	0.866	MEP+MVA
50	alpha-D-Ribose 5-phosphate exchange	0.164	0.657	0.196	0.663	0.866	MEP+MVA
51	N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramic acid exchange	0.623	0.656	0.355	0.667	1.575	MEP+MVA
52	N-Acetyl-D-glucosamine 1-phosphate exchange	0.262	0.654	0.238	0.672	2.646	MEP+MVA
53	N-Acetyl-D-mannosamine exchange	0.262	0.654	0.322	0.672	2.646	MEP+MVA
54	N-Acetyl-D-glucosamine exchange	0.262	0.654	0.322	0.672	2.646	MEP+MVA
55	N-Acetylmuramate exchange	0.355	0.646	0.331	0.665	2.963	MEP+MVA
56	2(alpha-D-Mannosyl)-D-glycerate exchange	0.290	0.644	0.295	0.649	0.838	MEP+MVA
57	Ornithine exchange	0.160	0.642	0.328	0.642	0.000	MEP
58	Phenethylamine exchange	0.254	0.635	0.566	0.661	4.111	MEP+MVA
59	Phenylacetaldehyde exchange	0.254	0.635	0.575	0.661	4.111	MEP+MVA
60	Agmatine exchange	0.159	0.634	0.327	0.634	0.000	MEP
61	D-Gluconate exchange	0.190	0.633	0.265	0.636	0.562	MEP+MVA
62	D-Alanyl-D-alanine exchange	0.190	0.632	0.322	0.654	3.427	MEP+MVA
63	N-Acetylneuraminate exchange	0.345	0.627	0.304	0.636	1.441	MEP+MVA
64	L-Idonate exchange	0.187	0.622	0.260	0.624	0.349	MEP+MVA
65	D-Alanine exchange	0.093	0.618	0.283	0.641	3.757	MEP+MVA
66	L-Alanine exchange	0.093	0.618	0.283	0.641	3.757	MEP+MVA
67	L-alanine-L-glutamate exchange'	0.246	0.614	0.308	0.614	0.092	MEP+MVA
68	L-alanine-D-glutamate exchange	0.246	0.614	0.308	0.614	0.092	MEP+MVA
69	tetradecanoate (n-C14:1) exchange	0.430	0.614	0.519	0.660	0.000	MVA
70	hexadecanoate (n-C16:1) exchange	0.490	0.613	0.527	0.659	0.000	MVA
71	Propionate exchange	0.092	0.612	0.342	0.625	2.128	MEP+MVA
72	D-lactate exchange	0.092	0.612	0.281	0.625	2.128	MEP+MVA
73	octadecanoate (n-C18:1) exchange	0.551	0.612	0.533	0.658	0.000	MVA
74	L-Galactonate exchange	0.183	0.612	0.256	0.614	0.447	MEP+MVA
75	2-Dehydro-3-deoxy-D-gluconate exchange	0.183	0.612	0.282	0.614	0.447	MEP+MVA
76	D-Galactonate exchange	0.183	0.612	0.256	0.614	0.447	MEP+MVA
77	L-Threonine O-3-phosphate exchange	0.120	0.601	0.166	0.635	5.691	MEP+MVA
78	L-Threonine exchange	0.120	0.601	0.275	0.635	5.691	MEP+MVA
79	octadecanoate (n-C18:0) exchange	0.537	0.597	0.516	0.642	0.000	MVA
80	Hexadecanoate (n-C16:0) exchange	0.477	0.596	0.508	0.641	0.000	MVA
81	tetradecanoate (n-C14:0) exchange	0.416	0.594	0.498	0.639	0.000	MVA
82	Dodecanoate (n-C12:0) exchange	0.355	0.592	0.486	0.637	0.000	MVA
83	Decanoate (n-C10:0) exchange	0.295	0.590	0.469	0.634	0.000	MVA
84	octanoate (n-C8:0) exchange	0.234	0.586	0.446	0.630	0.000	MVA
85	L-Glutamate exchange	0.146	0.582	0.271	0.583	0.156	MEP+MVA
86	L-Glutamine exchange	0.146	0.582	0.271	0.583	0.156	MEP+MVA
87	Hexanoate (n-C6:0) exchange	0.174	0.579	0.411	0.623	0.000	MVA
88	Phenylpropanoate exchange	0.260	0.577	0.474	0.592	2.551	MEP+MVA
89	5-Dehydro-D-gluconate exchange	0.173	0.575	0.243	0.579	0.667	MEP+MVA
90	L-Proline exchange	0.144	0.575	0.340	0.575	0.047	MEP+MVA
91	L-Ascorbate exchange	0.171	0.571	0.265	0.579	1.472	MEP+MVA
92	Acetoacetate exchange	0.113	0.566	0.305	0.609	0.000	MVA
93	Butyrate (n-C4:0) exchange	0.113	0.566	0.354	0.609	0.000	MVA
94	L-Lactate exchange	0.085	0.564	0.258	0.586	3.940	MEP+MVA
95	L-Arginine exchange	0.169	0.563	0.262	0.563	0.000	MEP
96	D-Glucuronate exchange	0.168	0.561	0.237	0.565	0.720	MEP+MVA
97	D-Fructuronate exchange	0.168	0.561	0.237	0.565	0.720	MEP+MVA
98	D-Glucuronate 1-phosphate exchange	0.168	0.561	0.169	0.565	0.720	MEP+MVA
99	D-Galacturonate exchange	0.168	0.561	0.237	0.565	0.720	MEP+MVA
100	3-(3-hydroxy-phenyl)propionate exchange	0.246	0.547	0.406	0.560	2.446	MEP+MVA
101	dAMP exchange	0.273	0.546	0.226	0.549	0.431	MEP+MVA
102	Deoxyadenosine exchange	0.273	0.546	0.296	0.549	0.431	MEP+MVA
103	Deoxyinosine exchange	0.271	0.542	0.293	0.545	0.455	MEP+MVA
104	dTMP exchange	0.271	0.542	0.224	0.545	0.455	MEP+MVA
105	Reduced glutathione exchange	0.269	0.537	0.239	0.545	1.484	MEP+MVA
106	L-Prolinylglycine exchange	0.188	0.536	0.296	0.538	0.405	MEP+MVA
107	(R)-Glycerate exchange	0.080	0.536	0.208	0.536	0.143	MEP+MVA
108	3-hydroxycinnamic acid exchange	0.238	0.529	0.397	0.542	0.000	MVA
109	Acetate exchange	0.053	0.527	0.243	0.567	0.000	MVA
110	AMP exchange	0.263	0.527	0.208	0.527	0.084	MEP+MVA
111	2',3'-Cyclic AMP exchange	0.263	0.527	0.219	0.527	0.084	MEP+MVA
112	3'-AMP exchange	0.263	0.527	0.208	0.527	0.084	MEP+MVA
113	Adenosine exchange	0.263	0.527	0.268	0.527	0.084	MEP+MVA
114	UDP-N-acetyl-D-glucosamine exchange	0.445	0.524	0.200	0.530	1.278	MEP+MVA
115	IMP exchange	0.262	0.523	0.206	0.524	0.106	MEP+MVA
116	Inosine exchange	0.262	0.523	0.265	0.524	0.106	MEP+MVA
117	Succinate exchange	0.105	0.523	0.245	0.530	1.326	MEP+MVA
118	Deoxyguanosine exchange	0.260	0.519	0.264	0.522	0.608	MEP+MVA
119	dGMP exchange	0.260	0.519	0.205	0.522	0.608	MEP+MVA
120	UDPglucose exchange	0.388	0.517	0.187	0.520	0.486	MEP+MVA
121	2,3-diaminopropionate exchange	0.078	0.517	0.201	0.532	0.000	MVA
122	D-Serine exchange	0.078	0.517	0.201	0.532	0.000	MVA
123	O-Phospho-L-serine exchange	0.078	0.517	0.115	0.532	0.000	MVA
125	L-Serine exchange	0.078	0.517	0.201	0.532	0.000	MVA
124	Pyruvate exchange	0.078	0.517	0.243	0.532	0.000	MVA
126	UDPgalactose exchange	0.386	0.515	0.186	0.517	0.530	MEP+MVA
127	2-Oxoglutarate exchange	0.127	0.509	0.241	0.518	1.791	MEP+MVA
128	3'-GMP exchange	0.250	0.500	0.188	0.500	0.000	MEP
129	2',3'-Cyclic GMP exchange	0.250	0.500	0.198	0.500	0.000	MEP
130	Guanosine exchange	0.250	0.500	0.240	0.500	0.000	MEP
131	GMP exchange	0.250	0.500	0.188	0.500	0.000	MEP
132	Xanthosine exchange	0.248	0.496	0.238	0.498	0.274	MEP+MVA

## B.1. Model-based Medium Optimization

Table B.1 Continued from previous page

Nr	Carbon Source	Non Mevalonate Pathway			MEP+MVA. Carbon Yield	Synergy, [%]	Best Pathway
		Molar	Carbon	Mass			
133	Xanthosine 5'-phosphate exchange	0.248	0.496	0.186	0.498	0.274	MEP+MVA
134	L-Asparagine exchange	0.098	0.491	0.202	0.498	1.387	MEP+MVA
135	L-Aspartate exchange	0.098	0.491	0.202	0.498	1.387	MEP+MVA
136	D-Glucarate exchange	0.147	0.489	0.192	0.494	1.008	MEP+MVA
137	D-Galactarate exchange	0.147	0.489	0.192	0.494	1.008	MEP+MVA
138	Citrate exchange	0.146	0.485	0.210	0.486	0.156	MEP+MVA
139	L-Malate exchange	0.096	0.480	0.198	0.487	1.537	MEP+MVA
140	Fumarate exchange	0.096	0.480	0.229	0.487	1.537	MEP+MVA
141	Cys-Gly exchange	0.120	0.478	0.183	0.492	0.000	MVA
142	L-Cysteine exchange	0.071	0.474	0.160	0.490	0.000	MVA
143	D-Cysteine exchange	0.071	0.474	0.160	0.490	0.000	MVA
144	UDP-D-glucuronate exchange	0.349	0.465	0.165	0.471	1.280	MEP+MVA
145	Fe(III)dicitrate exchange	0.278	0.464	0.174	0.464	0.000	MEP
146	Glycine exchange	0.046	0.459	0.166	0.469	2.128	MEP+MVA
147	D-Malate exchange	0.092	0.459	0.189	0.469	2.128	MEP+MVA
148	Deoxycytidine exchange	0.189	0.421	0.227	0.431	2.453	MEP+MVA
149	dCMP exchange	0.189	0.421	0.169	0.431	2.453	MEP+MVA
150	Deoxyuridine exchange	0.188	0.417	0.224	0.426	2.351	MEP+MVA
151	dUMP exchange	0.188	0.417	0.167	0.426	2.351	MEP+MVA
152	L-tartrate exchange	0.080	0.402	0.148	0.413	0.000	MVA
153	2',3'-Cyclic CMP exchange	0.179	0.397	0.160	0.397	0.000	MEP
154	3'-cmp exchange	0.179	0.397	0.151	0.397	0.000	MEP
157	Cytidine exchange	0.179	0.397	0.200	0.397	0.000	MEP
159	Uridine exchange	0.179	0.397	0.199	0.397	0.000	MEP
158	2',3'-Cyclic UMP exchange	0.179	0.397	0.159	0.397	0.000	MEP
155	UMP exchange	0.179	0.397	0.151	0.397	0.000	MEP
160	CMP exchange	0.179	0.397	0.151	0.397	0.000	MEP
156	3'-UMP exchange	0.179	0.397	0.151	0.397	0.000	MEP
161	D-tartrate exchange	0.078	0.391	0.144	0.401	0.000	MVA
162	L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine exchange	0.334	0.371	0.198	0.376	1.434	MEP+MVA
163	Thymidine exchange	0.183	0.367	0.206	0.376	2.339	MEP+MVA
164	dTMP exchange	0.183	0.367	0.156	0.376	2.339	MEP+MVA
165	Glycolate exchange	0.036	0.357	0.130	0.357	0.000	MEP+MVA
166	Formaldehyde exchange	0.017	0.345	0.156	0.345	0.000	MEP
167	fructoselysine exchange	0.203	0.338	0.179	0.341	0.840	MEP+MVA
168	psicoselysine exchange	0.203	0.338	0.179	0.341	0.840	MEP+MVA
169	L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate exchange	0.239	0.318	0.167	0.319	0.238	MEP+MVA
170	sn-Glycero-3-phosphocholine exchange	0.125	0.313	0.132	0.313	0.000	MEP
171	Adenine exchange	0.078	0.312	0.157	0.312	0.000	MEP
172	Hypoxanthine exchange	0.078	0.312	0.156	0.312	0.000	MEP
173	sn-Glycero-3-phospho-1-inositol exchange	0.125	0.278	0.102	0.278	0.000	MEP
174	Guanine exchange	0.069	0.276	0.125	0.276	0.000	MEP
175	butanesulfonate exchange	0.055	0.273	0.109	0.294	0.000	MVA
176	Xanthine exchange	0.068	0.271	0.121	0.271	0.000	MEP
177	UDP-N-acetyl-D-galactosamine exchange	0.179	0.210	0.080	0.210	0.000	MEP
178	Allantoin exchange	0.036	0.179	0.061	0.179	0.000	MEP+MVA
179	L-Tryptophan exchange	0.080	0.145	0.106	0.149	0.000	MVA
180	Formate exchange	0.003	0.060	0.018	0.061	0.000	MVA

### B.1.3. Ranking of 180 Substrates According to Profit

Table B.2 provides a ranking of 180 different growth supporting substrates using profit as performance criteria. Note that for 21 substrates, the molar price was not available.

Table B.2.: **Maximal Profit Calculated for 180 Carbon Sources.** The *E. coli* metabolic reconstruction iJO1366 was used for all calculations. Profits are reported for the non-mevalonate pathway (MEP). For all calculations, NGAM and growth rate were set to 0. Additionally, molecular weight, prices on a molar basis and carbon content for each carbon source are provided.

Nr	Carbon Source	Molecular Weight, [g/mol]	Nr. Carbon Atoms	Price, [US\$/mol]	Profit MEP, [US\$ Txd/US\$ Substrate]
1	Formaldehyde exchange	30.026	1	7.75E-03	285,004
2	Maltohexaose exchange	990.870	36	7.07E-01	223,319
3	Ethanol exchange	46.069	2	5.10E-02	220,493
4	Maltose exchange	342.300	12	2.61E-01	203,365
5	Acetaldehyde exchange	44.053	2	5.21E-02	182,370
6	Hexadecanoate (n-C16:0) exchange	255.421	16	3.12E-01	167,338
7	D-Gluconate exchange	195.149	6	1.49E-01	163,707
8	octadecanoate (n-C18:0) exchange	283.475	18	3.72E-01	158,154
9	D-Sorbitol exchange	182.174	6	1.81E-01	155,600
10	D-Glucose exchange	180.158	6	1.74E-01	151,976

## B. Appendix for Process Engineering

Table B.2 Continued from previous page

Nr	Carbon Source	Molecular Weight, [g/mol]	Nr. Carbon Atoms	Price, [US\$/mol]	Profit MEP, [US\$ Txd/ US\$ Substrate]
11	Propanal exchange	58.080	3	1.03E-01	133,928
12	Sucrose exchange	342.300	12	4.10E-01	129,244
13	tetradecanoate (n-C14:0) exchange	227.367	14	4.06E-01	112,776
14	Citrate exchange	189.102	6	1.70E-01	109,495
15	Decanoate (n-C10:0) exchange	171.260	10	3.47E-01	94,429
16	L-Lactate exchange	89.071	3	1.32E-01	81,920
17	Propionate exchange	73.072	3	1.46E-01	80,436
18	Ethanolamine exchange	62.092	2	1.21E-01	76,469
19	Butyrate (n-C4:0) exchange	87.098	4	1.88E-01	70,362
20	octanoate (n-C8:0) exchange	143.206	8	3.93E-01	66,879
21	Fumarate exchange	114.057	4	1.87E-01	65,709
22	Lactose exchange	342.300	12	8.63E-01	61,135
23	Glycerol exchange'	92.095	3	2.60E-01	59,371
24	L-Threonine exchange	119.120	4	3.25E-01	47,386
25	Hexanoate (n-C6:0) exchange	115.152	6	4.39E-01	44,977
26	Succinate exchange	116.073	4	3.00E-01	44,687
27	D-Mannitol exchange	182.174	6	6.73E-01	41,920
28	L-Glutamate exchange	146.123	5	4.71E-01	39,532
29	D-Xylose exchange	150.131	5	5.57E-01	39,200
30	D-Fructose exchange	180.158	6	7.12E-01	37,175
31	Glycine exchange	75.067	2	1.88E-01	31,256
32	Phenylacetaldehyde exchange	120.151	8	1.08E+00	29,764
33	L-Alanine exchange	89.094	3	4.26E-01	27,872
34	L-Malate exchange	132.073	4	4.53E-01	27,148
35	L-Aspartate exchange	132.096	4	6.22E-01	20,215
36	Uridine exchange	244.204	9	1.34E+00	17,075
37	L-Ascorbate exchange	176.126	6	1.35E+00	16,278
38	L-tartrate exchange	148.072	4	6.51E-01	15,809
39	4-Aminobutanoate exchange	103.121	4	1.11E+00	15,437
40	L-Arginine exchange	175.211	6	1.65E+00	13,109
41	L-Glutamine exchange	146.146	5	1.52E+00	12,225
42	Glycolate exchange	75.044	2	3.97E-01	11,523
43	Formate exchange	45.018	1	3.53E-02	10,871
44	Pyruvate exchange	87.055	3	9.34E-01	10,640
45	Inosine exchange	268.229	10	3.68E+00	9,107
46	D-Mannose exchange	180.158	6	3.49E+00	7,591
47	Phenethylamine exchange	122.190	8	4.95E+00	6,564
48	Dodecanoate (n-C12:0) exchange	199.313	12	6.18E+00	6,355
49	Hypoxanthine exchange	136.113	5	1.75E+00	5,707
50	Allantoin exchange	158.117	4	8.97E-01	5,099
51	L-Cysteine exchange	121.160	3	1.79E+00	5,084
52	L-Serine exchange	105.094	3	2.13E+00	4,669
53	IMP exchange	346.193	10	8.38E+00	3,999
54	Adenine exchange	135.128	5	3.26E+00	3,061
55	L-Proline exchange	115.132	5	4.61E+00	2,947
56	Guanine exchange	151.128	5	3.12E+00	2,837
57	D-Ribose exchange	150.131	5	8.11E+00	2,594
58	Cytidine exchange	243.219	9	9.81E+00	2,332
59	D-Galactose exchange	180.158	6	1.28E+01	2,053
60	(S)-Propane-1,2-diol exchange	76.095	3	7.61E+00	1,906
61	2-Oxoglutarate exchange	144.084	5	1.02E+01	1,593
62	Ornithine exchange	133.171	5	1.29E+01	1,580
63	D-Alanine exchange	89.094	3	7.66E+00	1,549
64	L-Asparagine exchange	132.119	4	8.46E+00	1,487
65	D-Galactonate exchange	195.149	6	1.68E+01	1,400
66	D-lactate exchange	89.071	3	9.00E+00	1,307
67	D-Glucosamine exchange	180.181	6	2.05E+01	1,289
68	Guanosine exchange	283.244	10	2.52E+01	1,270
69	Putrescine exchange	90.169	4	1.65E+01	1,234
70	Adenosine exchange	267.245	10	2.78E+01	1,214
71	Galactitol exchange	182.174	6	2.35E+01	1,197
72	N-Acetyl-D-glucosamine exchange	221.210	8	3.05E+01	1,098
73	Dihydroxyacetone exchange	90.079	3	1.38E+01	962
74	L-Rhamnose exchange	164.158	6	3.09E+01	845
75	L-Tryptophan exchange	204.229	11	1.31E+01	781
76	D-Galactarate exchange	208.125	6	2.60E+01	722
77	D-tartrate exchange	148.072	4	1.85E+01	541
78	3-hydroxycinnamic acid exchange	163.153	9	5.86E+01	521
79	Trehalose exchange	342.300	12	1.22E+02	437
80	Glycerol 2-phosphate exchange	170.059	3	3.83E+01	403
81	D-Serine exchange	105.094	3	2.50E+01	397
82	Xanthine exchange	152.113	5	2.28E+01	380
83	AMP exchange	345.209	10	9.60E+01	352
84	Thymidine exchange	242.232	10	7.02E+01	334
85	CMP exchange	321.183	9	8.22E+01	278
86	D-Malate exchange	132.073	4	4.28E+01	275
87	Reduced glutathione exchange	306.320	10	1.33E+02	258
88	chitobiose exchange	424.405	16	2.81E+02	229
89	D-Cysteine exchange	121.160	3	5.15E+01	177
90	L-Arabinose exchange	150.131	5	1.27E+02	172
91	D-Glucose 6-phosphate exchange	258.122	6	1.86E+02	147
92	D-Mannose 6-phosphate exchange	258.122	6	1.86E+02	147
93	D-Glucose 1-phosphate exchange	258.122	6	1.81E+02	146
94	alpha-D-Galactose 1-phosphate exchange	258.122	6	1.81E+02	145
95	Deoxyuridine exchange	228.205	9	1.71E+02	140
96	(R)-Glycerate exchange	105.070	3	8.22E+01	125
97	butanesulfonate exchange	137.180	4	6.09E+01	115
98	D-fructose 6-phosphate exchange	258.122	6	2.52E+02	109
99	N-Acetylneuraminat exchange	308.265	11	4.10E+02	108
100	alpha-D-Ribose 5-phosphate exchange	228.095	5	1.98E+02	106

## B.1. Model-based Medium Optimization

Table B.2 Continued from previous page

Nr	Carbon Source	Molecular Weight, [g/mol]	Nr. Carbon Atoms	Price, [US\$/mol]	Profit MEP, [US\$ Txd/ US\$ Substrate]
101	Melibiose exchange	342.300	12	5.51E+02	96
102	dIMP exchange	330.194	10	3.68E+02	94
103	L-Fucose exchange	164.158	6	2.77E+02	94
104	D-Galacturonate exchange	193.133	6	2.33E+02	92
105	D-Glucuronate exchange	193.133	6	2.33E+02	92
106	Deoxycytidine exchange	227.220	9	2.74E+02	88
107	L-Xylulose exchange	150.131	5	2.49E+02	88
108	Deoxyguanosine exchange	267.245	10	4.25E+02	78
109	Deoxyadenosine exchange	251.245	10	5.53E+02	63
110	D-Glucarate exchange	208.125	6	3.04E+02	62
111	UDP-N-acetyl-D-glucosamine exchange	605.343	17	9.56E+02	60
112	dAMP exchange	329.209	10	6.34E+02	55
113	3'-cmp exchange	321.183	9	4.23E+02	54
114	D-Glucosamine 6-phosphate exchange	258.145	6	5.28E+02	52
115	Maltotriose exchange	504.442	18	1.61E+03	49
116	sn-Glycero-3-phosphocholine exchange	257.224	8	3.40E+02	47
117	Deoxyinosine exchange	252.230	10	7.45E+02	47
118	dCMP exchange	305.184	9	5.88E+02	41
119	Glycerol 3-phosphate exchange	170.059	3	4.51E+02	35
120	N-Acetyl-D-mannosamine exchange	221.210	8	9.64E+02	35
121	L-Lyxose exchange	150.131	5	6.50E+02	34
122	2',3'-Cyclic CMP exchange	304.176	9	7.75E+02	29
123	dTMP exchange	320.196	10	8.07E+02	29
124	D-Glyceraldehyde exchange	90.079	3	4.74E+02	28
125	2',3'-Cyclic AMP exchange	328.201	10	1.37E+03	25
126	UDP-N-acetyl-D-galactosamine exchange	605.343	17	9.56E+02	24
127	UMP exchange	322.168	9	1.02E+03	22
128	3'-AMP exchange	345.209	10	1.78E+03	19
129	Agmatine exchange	132.209	5	1.29E+03	16
130	GMP exchange	361.208	10	2.03E+03	16
131	D-Allose exchange	180.158	6	1.74E+03	15
132	UDPglucose exchange	564.290	15	4.19E+03	12
133	L-Threonine O-3-phosphate exchange	197.084	4	1.36E+03	11
134	L-alanine-D-glutamate exchange	217.202	8	3.80E+03	8
135	UDPgalactose exchange	564.290	15	7.86E+03	6
136	dGMP exchange	345.209	10	5.29E+03	6
137	Acetoacetate exchange	101.082	4	2.98E+03	5
138	N-Acetylmuramate exchange	292.266	11	1.08E+04	4
139	Maltopentaose exchange	828.727	30	9.61E+04	1
140	Phenylpropanoate exchange	149.169	9	4.30E+04	1
141	O-Phospho-L-serine exchange	183.058	3	1.79E+04	1
142	D-Alanyl-D-alanine exchange	160.173	6	4.76E+04	1
143	beta D-Galactose exchange	180.158	6	5.55E+04	0
144	Acetate exchange	59.045	2	1.75E+04	0
145	5-Dehydro-D-gluconate exchange	193.133	6	5.95E+04	0
146	Maltotetraose exchange	666.585	24	2.91E+05	0
147	Xanthosine exchange	284.229	10	9.46E+04	0
148	Xanthosine 5'-phosphate exchange	362.193	10	1.04E+05	0
149	dUMP exchange	306.169	9	8.82E+04	0
150	D-Fructuronate exchange	193.133	6	8.29E+04	0
151	UDP-D-gluconate exchange	577.266	15	1.78E+05	0
152	Glycerophosphoserine exchange	258.145	6	1.09E+05	0
153	2',3'-Cyclic UMP exchange	305.161	9	9.40E+04	0
154	N-Acetyl-D-glucosamine 1-phosphate exchange	299.174	8	1.41E+05	0
155	2',3'-Cyclic GMP exchange	344.201	10	1.46E+05	0
156	Cys-Gly exchange	178.212	5	6.99E+04	0
157	fructoselysine exchange	309.340	12	1.31E+05	0
158	3'-UMP exchange	322.168	9	1.38E+05	0
159	3'-GMP exchange	361.208	10	1.94E+05	0
160	L-Galactonate exchange	195.149	6	NA	NA
161	D-Glucuronate 1-phosphate exchange	271.097	6	NA	NA
162	1,4-alpha-D-glucan exchange	990.870	36	NA	NA
163	2,3-diaminopropionate exchange	105.117	3	NA	NA
164	2-Dehydro-3-deoxy-D-gluconate exchange	177.134	6	NA	NA
165	3-(3-hydroxy-phenyl)propionate exchange	165.169	9	NA	NA
166	L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate exchange	389.386	15	NA	NA
167	L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine exchange	460.465	18	NA	NA
168	L-alanine-L-glutamate exchange	217.202	8	NA	NA
169	N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramic acid exchange	477.446	19	NA	NA
170	Fe(III)dicitrate exchange	434.050	12	NA	NA
171	sn-Glycero-3-phosphoethanolamine exchange	215.143	5	NA	NA
172	Glycerophosphoglycerol exchange	245.146	6	NA	NA
173	sn-Glycero-3-phospho-1-inositol exchange	333.209	9	NA	NA
174	hexadecenoate (n-C16:1) exchange	253.405	16	NA	NA
175	L-Idonate exchange	195.149	6	NA	NA
176	2(alpha-D-Mannosyl)-D-glycerate exchange	267.213	9	NA	NA
177	octadecenoate (n-C18:1) exchange	281.459	18	NA	NA
178	L-Prolinylglycine exchange	172.184	7	NA	NA
179	psicoselysine exchange	309.340	12	NA	NA
180	tetradecenoate (n-C14:1) exchange	225.351	14	NA	NA

Table B.3.: **Concentration Time Courses for Batch ID 1.** Glucose was added to the bioreactor as follows: 2 g diluted in 4 ml after 14.5 hours; 1 g diluted in 2 ml after 56.5 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0	0.09	–	108.33	–
1	0.11	–	–	–
2	0.20	–	–	–
3	0.39	0.30	–	–
4	0.68	0.50	110.56	–
5	1.13	–	–	–
6	1.72	1.20	103.89	–
6.5	2.00	–	–	–
7	2.34	–	–	–
7.5	2.90	–	–	–
8	3.48	1.67	83.33	–
8.5	3.72	–	–	0.000
10.5	4.00	2.30	72.46	–
12.5	4.30	2.23	69.66	–
14.5	5.15	2.70	63.48	$1.515 \times 10^{-3}$
24.5	8.00	4.20	66.43	$8.137 \times 10^{-4}$
27.5	8.70	4.03	47.84	–
30.5	9.20	4.53	62.77	–
32.5	10.10	4.13	57.75	$4.452 \times 10^{-3}$
48.5	12.20	4.77	18.09	$6.475 \times 10^{-3}$
52.5	11.90	4.47	20.00	–
56.5	11.90	4.47	12.98	$8.332 \times 10^{-3}$
72.5	11.50	5.03	10.06	–
76.5	14.60	5.40	0.11	$7.831 \times 10^{-3}$

## B.2. Model-based Fed-batch Optimization

### B.2.1. Concentration Time Courses for Process Development

Concentration time courses for OD, glucose, and taxadiene for six different semi-batch fermentations are provided in Tables B.3 to B.8.

### B.2.2. Temperature Dependencies of Phases Duration and Exchange Rates

Tables B.9 and B.10 list all formulas used to describe the effect of  $T_{production}$  on phase duration and exchange rates, respectively.

### B.2.3. Modeling Concentration Time Courses for All Fermentations

Simulated concentration time courses for all six fermentation processes are shown in Figure 8.4 for batch 1 and in Figures B.1 to B.5 for all other batches. Determination coefficients are calculated using following formula

$$R^2 = 1 - \frac{\sum(E_i - S_i)^2}{\sum(E_i - \bar{E})^2}, \quad \text{where} \quad \bar{E} = \frac{1}{n} \sum_{i=1}^n E_i.$$

$E$  represents the experimental data set, containing  $n$  measurements,  $\bar{E}$  is the mean of the measured data and  $S$  is the simulated data set.  $S_i$  is the corresponding predicted value of

## B.2. Model-based Fed-batch Optimization

Table B.4.: **Concentration Time Courses for Batch ID 2.** Glucose was added to the bioreactor as follows: 2 g diluted in 4 ml after 8.5 hours; 1.5 g diluted in 3 ml after 33.3 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0	0.07	–	115.61	–
2	0.26	–	–	–
4	0.78	–	–	–
5	1.47	–	–	–
5.5	2.04	1.27	103.27	–
6.5	3.60	–	–	–
7	1.00	–	–	$6.491 \times 10^{-4}$
7.5	4.06	2.30	89.80	–
8.5	4.13	–	–	–
24.75	8.33	4.20	76.98	$4.309 \times 10^{-3}$
27.75	8.13	4.30	77.82	–
30.75	8.68	4.30	66.79	$5.416 \times 10^{-3}$
33.33	9.23	–	–	–
47	10.95	4.37	63.42	$6.359 \times 10^{-3}$
49.75	11.23	4.40	65.20	–
52.25	10.65	4.40	60.07	–
54.75	10.90	4.47	65.10	$6.477 \times 10^{-3}$
71.25	11.45	4.37	37.04	–
73.75	11.10	4.33	34.29	$7.565 \times 10^{-3}$

Table B.5.: **Concentration Time Courses for Batch ID 3.** Glucose was added to the bioreactor as follows: 2.5 g diluted in 5 ml after 9.25, 27.17 and 47.25 hours; 5 g diluted in 10 ml after 55.5 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0.00	–	–	118.00	–
0.75	0.14	–	117.83	–
1.75	0.29	–	–	–
3.75	0.94	0.47	109.13	–
4.75	1.50	–	–	–
5.75	2.40	0.73	98.93	–
6.25	2.72	–	–	–
6.75	3.13	–	–	–
7.25	3.26	1.40	87.82	0.000
9.25	4.32	–	–	–
23.00	10.81	4.13	48.38	$1.792 \times 10^{-2}$
27.17	11.84	4.47	38.65	–
31.00	12.64	4.43	56.56	$2.197 \times 10^{-2}$
47.25	14.00	4.67	25.82	–
51.50	14.95	4.83	38.67	–
55.50	14.74	4.57	37.77	$2.770 \times 10^{-2}$
75.00	17.18	5.27	59.59	$1.604 \times 10^{-2}$

B. Appendix for Process Engineering

Table B.6.: **Concentration Time Courses for Batch ID 4.** Glucose was added to the bioreactor as follows: 5 g diluted in 10 ml after 23 hours; 2.5 g diluted in 5 ml after 27.3, 31.3 and 51.8 hours; 5 g diluted in 10 ml after 55.8 hours; 7.5 g diluted in 15 ml after 75.3 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0.00		0.06	111.01	–
1.00	0.18	0.09	101.92	–
2.00	0.30	–	–	–
4.00	0.84	0.30	98.38	–
5.00	1.26	–	–	–
6.00	1.85	0.83	92.93	–
6.75	2.62	–	–	–
7.25	3.21	1.17	89.80	0.000
9.50	6.22	–	–	–
10.50	7.89	2.43	42.04	$1.922 \times 10^{-3}$
23.00	12.81	–	–	–
27.25	13.13	3.47	24.93	–
31.25	12.28	2.67	31.31	$1.830 \times 10^{-3}$
47.50	12.60	2.87	0.00	–
51.75	14.20	3.80	0.00	–
55.75	13.00	3.87	20.86	$7.340 \times 10^{-4}$
75.25	12.97	4.10	0.00	–

Table B.7.: **Concentration Time Courses for Batch ID 5.** Glucose was added to the bioreactor as follows: 2 g diluted in 4 ml after 9.8 hours; 5 g diluted in 10 ml after 30.3 hours and 1 g diluted in 2 ml after 54.3 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0.00	0.10		111.11	–
0.50	0.15	–	112.23	–
1.50	0.24	–	–	–
2.50	0.44	–	–	–
3.50	0.80	0.49	116.46	–
4.50	1.54	–	–	–
5.50	3.00	1.51	105.80	–
6.25	3.80	–	92.76	–
7.25	4.25	–	–	–
8.25	5.44	2.87	–	–
9.25	5.78	–	–	–
9.75	6.40	3.04	65.48	$4.401 \times 10^{-3}$
22.75	9.70	4.27	57.48	$8.618 \times 10^{-3}$
24.25	10.30	–	–	–
25.50	–	–	–	–
27.00	11.10	4.56	64.93	$1.116 \times 10^{-2}$
30.25	10.80	4.53	55.46	$1.377 \times 10^{-2}$
47.00	13.20	5.07	74.61	–
50.75	–	–	–	–
54.25	14.05	5.18	69.62	$1.729 \times 10^{-2}$
70.50	15.40	5.44	–	–
74.75	15.55	5.44	–	$1.827 \times 10^{-2}$

## B.2. Model-based Fed-batch Optimization

Table B.8.: **Concentration Time Courses for Batch ID 6.** Glucose was added to the bioreactor as follows: 2 g diluted in 4 ml after 9.75 hours; and 5 g diluted in 10 ml after 30.25 and 54.25 hours.

Time [h]	OD	DW [g/l]	Glucose [mmol/l]	Taxadiene [mmol/l]
0.00	0.10	–	111.11	–
0.50	0.16	0.11	126.05	–
1.50	0.24	–	–	–
2.50	0.53	–	–	–
3.50	0.80	0.49	114.45	–
4.50	1.54	–	–	–
5.50	2.60	1.22	105.46	–
6.25	3.40	–	84.93	–
7.25	–	–	–	–
8.25	4.23	3.07	–	–
9.25	5.76	–	–	–
9.75	5.97	3.24	74.97	$2.075 \times 10^{-3}$
22.75	6.40	4.49	39.55	$1.303 \times 10^{-2}$
24.25	11.60	–	–	–
25.50	12.50	–	–	–
27.00	13.00	4.76	26.63	$1.614 \times 10^{-2}$
30.25	14.10	5.16	12.39	$2.802 \times 10^{-2}$
47.00	17.50	5.62	31.19	–
50.75	18.85	–	–	–
54.25	19.35	5.93	8.11	$3.578 \times 10^{-2}$
70.50	22.20	6.24	34.54	–
74.75	23.50	6.51	–	$3.598 \times 10^{-2}$

Table B.9.: **Formulas Used to Describe the Effect of Production Temperature on Phase Duration.**  $x$  represents  $T_{production}$  in °C.

Phase	Symbol	Equation	R2
Response time	$t_{response}$	$0.0257x - 0.1612$	0.87
Duration of second biomass production phase	$t_{biomass2}$	18.3	–
Duration of first product formation phase	$t_{production1}$	$-0.9678x + 47.59$	0.98

B. Appendix for Process Engineering

Table B.10.: Formulas Used to Describe the Effect of Production Temperature on Exchange Rates.

Exchange Rate	Symbol	Equation	$R^2$
Growth	$\mu_2$	$-7.5724 \times 10^{-5}x^2 + 3.7440 \times 10^{-3}x - 4.2090 \times 10^{-3}$	0.99
	$\mu_3$	$-1.5319 \times 10^{-5}x^2 + 7.9919 \times 10^{-4}x - 6.5429 \times 10^{-3}$	0.99
Glucose Uptake	$r_{glucose2}$	$0.0031x^2 - 0.073x + 1.06$	0.99
	$r_{glucose3}$	$0.0027x^2 - 0.1066x + 1.3083$	0.99
Taxadiene Production	$r_{txd1}$	$\frac{c x e^{-\frac{\Delta H}{R x}}}{1 + e^{\frac{\Delta S}{R}} e^{-\frac{\Delta H}{R x}}}$	0.85
	$r_{txd2}$	0	-

$x$  represents  $T_{production}$  in Celsius.

Parameter values for the Arrhenius-like equation providing a good fit to the data were calculated as scaling factors modifying the original parameter values used in [57] as follows:  $c = 0.3612 e^{24.04}$  should be multiplied by 10.36;  $\Delta \tilde{H} = 15.000$  should be multiplied by 1.54;  $R = 1.9872$  (gas constant);  $\Delta S = 476.46$  should be multiplied by 1.11;  $\Delta H = 150.000$  should be multiplied by 1.08.

$E_i$ . Results are listed in Table B.11. Overall, concentration time courses for biomass are better described by our model.

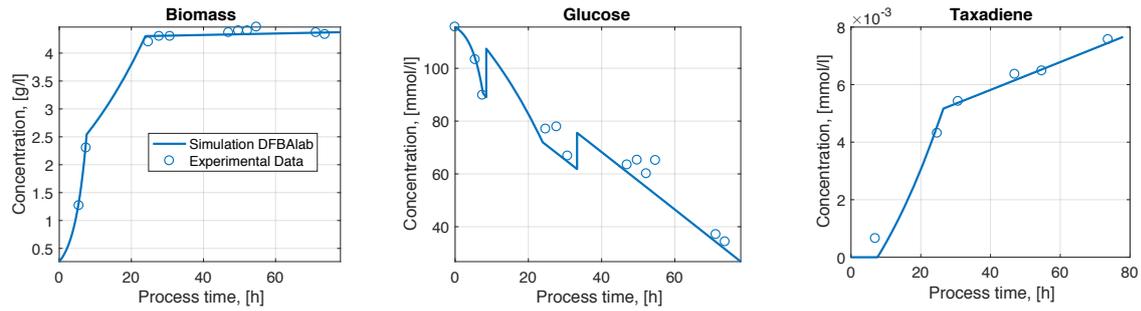


Figure B.1.: Modeling Time Course Concentrations: Batch 2. Production conditions corresponds to:  $T_{production} = 16^\circ C$ ,  $t_{induction} = 1$  h.  $P_{max}$  amounted to  $1.7614 \times 10^{-4}$  mmol taxadiene/1 h

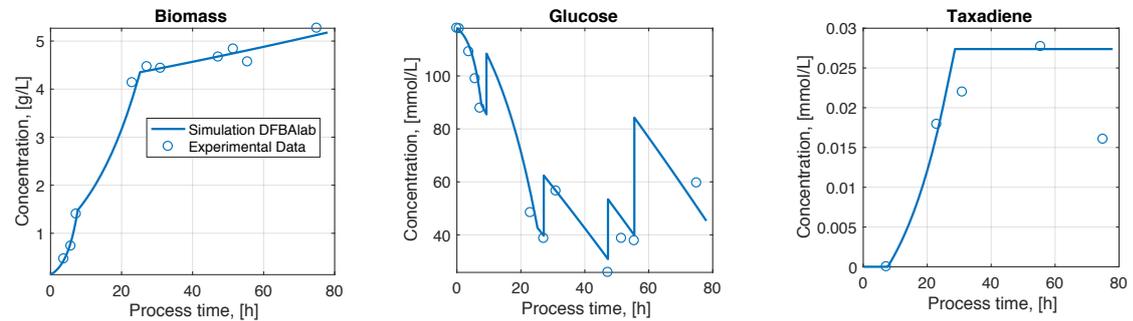


Figure B.2.: Modeling Time Course Concentrations: Batch 3. Production conditions corresponds to:  $T_{production} = 22^\circ C$ ,  $t_{induction} = 7.25$  h.  $P_{max}$  amounted to  $7.7892 \times 10^{-4}$  mmol taxadiene/1 h

## B.2. Model-based Fed-batch Optimization

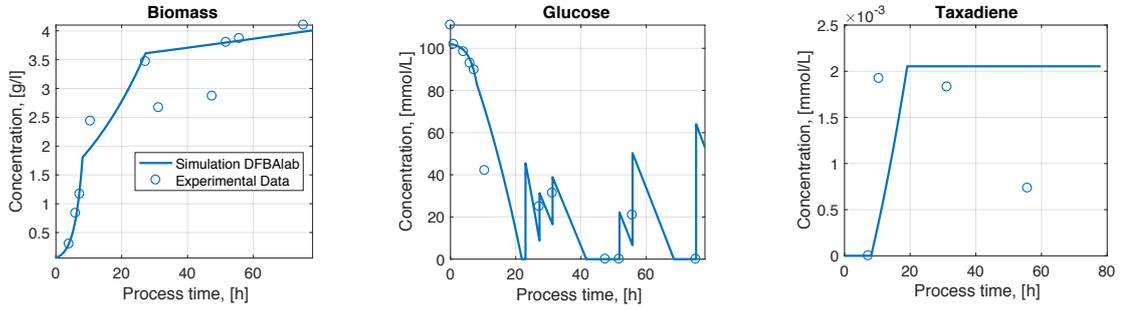


Figure B.3.: **Modeling Time Course Concentrations: Batch 4.** Production conditions corresponds to:  $T_{production} = 37^{\circ}C$ ,  $t_{induction} = 7.25$  h.  $P_{max}$  amounted to  $1.8304 \times 10^{-4}$  mmol taxadiene/1 h

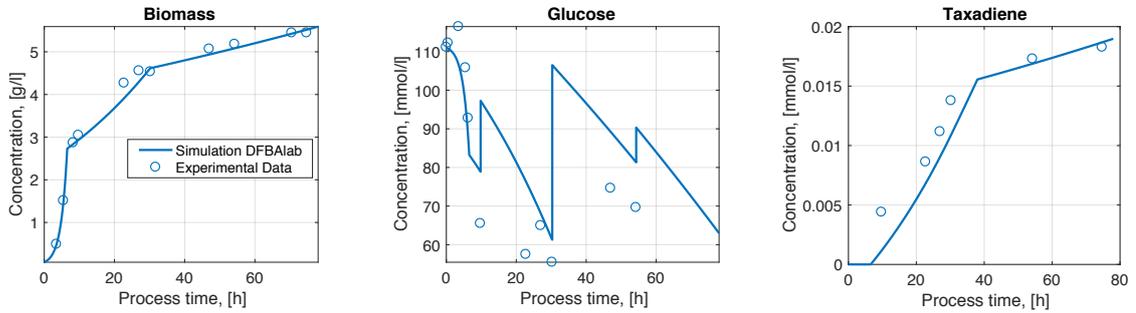


Figure B.4.: **Modeling Time Course Concentrations: Batch 5.** Production conditions corresponds to:  $T_{production} = 22^{\circ}C$ ,  $t_{induction} = 6.25$  h.  $P_{max}$  amounted to  $4.5526 \times 10^{-4}$  mmol taxadiene/1 h

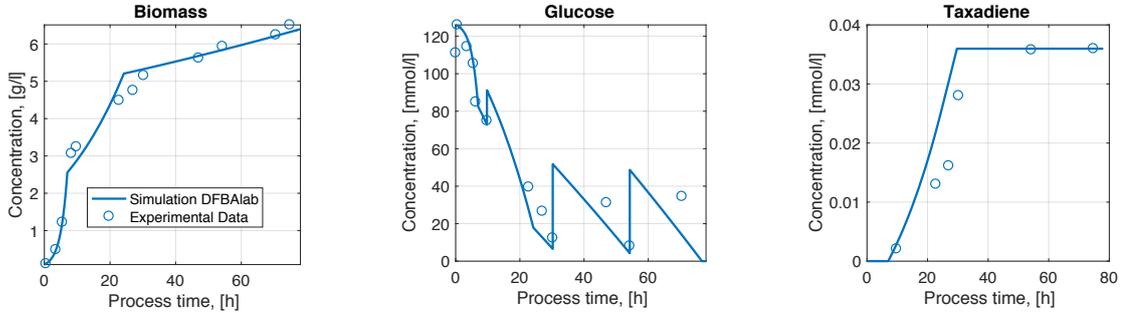


Figure B.5.: **Modeling Time Course Concentrations: Batch 6.** Production conditions corresponds to:  $T_{production} = 27^{\circ}C$ ,  $t_{induction} = 6.25$  h.  $P_{max}$  amounted to  $9.2615 \times 10^{-4}$  mmol taxadiene/1 h

Table B.11.: **Determination Coefficients for All Batches.** For average calculation, only positive values are considered.

Batch ID	Biomass	Glucose	Taxadiene
1	0.982	0.923	0.912
2	0.994	0.934	0.982
3	0.993	0.756	0.629
4	0.878	0.926	-0.634
5	0.991	0.794	0.809
6	0.862	0.770	0.627
<b>Average</b>	<b>0.950</b>	<b>0.850</b>	<b>0.792</b>