

Faculté Polytechnique



Metabolic Modeling of Cellular Culture Processes – Genetic, Microscopic and Macroscopic Scales

Application to Animal Cells, Bacteria and Microalgae

Doctoral Thesis

*In fulfillment of the requirements for the degree of
Docteur en Sciences de l'Ingénieur et Technologie*

Maxime MATON
Graduate Engineer in Mechanical Engineering



Under the supervision of Pr. Dr. Ir. Alain VANDE WOUWER and Pr. Dr. Ir. Philippe BOGAERTS

June 2025

Jury Members

Pr. Dr. Ir. Alain VANDE WOUWER
Advisor
Université de Mons

Pr. Dr. Ir. Philippe BOGAERTS
Co-Advisor
Université Libre de Bruxelles

Pr. Dr. Baptiste LEROY
President of the Jury
Université de Mons

Dr. Ir. Laurent DEWASME
Secretary of the Jury
Université de Mons

Pr. Dr. Ir. Arnaud VANDAELE
Université de Mons

Dr. Julio R. BANGA
CSIC (Spain)

Dr. Olivier BERNARD
INRIA-BIOCORE (France)

*À ma grand-mère, Huguette,
dont le coeur m'a élevé plus haut que les livres.*

Acknowledgments

This thesis would not have been possible without the guidance, support, and inspiration of many people, to whom I am deeply grateful.

First and foremost, my sincere and profound thanks go to my advisor, **Alain Vande Wouwer**. It is difficult to fully express the extent of my gratitude. His unwavering confidence in my potential, even when I doubted myself, has been a cornerstone of this journey. His trust, patience, and encouragement have made a lasting difference, not only in the work presented here but also in the researcher I have become. For his mentorship, for the freedom he gave me and for his constant push to aim higher: thank you.

I am also very grateful to my co-advisor, **Philippe Bogaerts**, whose involvement in this work was nothing short of remarkable. His precision, insightful feedback, and dedication greatly improved the quality of this doctoral thesis. Working under his guidance has been an immense privilege and a formative experience for which I will always be thankful. Our exchange often sparked the most valuable insights and clarified the story behind the data. I am especially grateful for his professionalism and his commitment to quality, which set a standard that I have tried to uphold throughout.

My sincere thanks go to the members of my doctoral committee for their time, constructive feedback, ongoing involvement, and thoughtful input provided at every key stage. Furthermore, I would like to thank the members of the jury for agreeing to be part of this important milestone and who agreed to review this final work. Their presence and willingness to engage with my work already mean a lot to me, and I am grateful that they are part of this final chapter of the thesis. It is a privilege to be evaluated by esteemed peers.

A special mention goes to **Baptiste Leroy**, whose knowledge of bacterial metabolism significantly enriched my understanding. Our conversations on photophosphorylation and proton motive force were not only enlightening but also essential to one of the core studies in this thesis. Thank you for your generosity and clarity.

I would also like to express my gratitude to **Jesús Picó** for welcoming me into his Department at the Universitat Politècnica de València. My time there was scientifically stimulating and personally rewarding. Working on gene expression modeling and synthetic biology during that stay broadened my horizons in meaningful ways. I thank him and his team for their hospitality and openness.

To the anonymous reviewers who helped improve the quality of my published work through their careful reading and suggestions: thank you. The same goes for all the researchers I met during conferences, whose questions and conversations brought fresh perspectives and new ideas.

I am also thankful to the entire **SECO** Department of the Faculty of Engineering. To the colleagues who came and went during these years, the professors who joined us for short stays, and the community that took shape within those walls. Thank you for the science, but just as much for the human connection. Doctoral research is often viewed as a solitary pursuit, but this experience was also made of laughter, shared meals, hallway chats, and moments that reminded me that we were all in this together.

To my family, my friends, and all those who supported, encouraged, and carried me, your help may have come in quiet gestures or simple words, but it made a world of difference. Whether in small or significant ways, your presence and support made a lasting impact on my academic and personal growth. I will not name names, but you will know who you are. Thank you for being here.

The financial support provided by the FNRS, the University of Mons, and the SECO Department made it possible for me to attend conferences and undertake an enriching research stay abroad. I gratefully acknowledge their role in enabling these important experiences.

And now, a final and personal acknowledgment. To myself. This work demanded more than knowledge; it required persistence, resilience, and faith. I will not list the challenges I faced or the toll it sometimes took, but I will say this: *This thesis exists because I refused to disappear. That should be more than enough.* Coming from a mechanical engineering background, diving deep into biology and biochemistry was no small feat. I thank myself for the countless hours spent learning, for embracing the unfamiliar, and for staying the course. A special nod to *Lehninger Principles of Biochemistry* by Nelson and Cox – often jokingly referred to as "my Bible", and rightly so.

I also recognize and appreciate the foundation laid by my Engineering education. It taught me the most valuable lesson of all: how to learn. In many ways, this thesis is a testament to that very ability.

Finally, thank you to each and every person who takes the time to read this work. I hope you will find in it something worth learning – just as I did.

Abstract

Metabolic Modeling of Cellular Culture Processes

This doctoral thesis addresses the mathematical modeling of cellular culture processes through a structured progression across three levels of biological complexity: macroscopic dynamics, intracellular metabolism, and gene-level regulation. Each modeling approach is developed with increasing mechanistic detail, reflecting the core objective of better describing and connecting observable culture behavior with the molecular processes that drive cellular function. This stepwise exploration offers a refined and biologically grounded view of cellular systems with direct relevance to the design and optimization of biotechnological applications.

At the macroscopic level, a novel hybrid modeling approach was developed to describe the dynamic behavior of cellular cultures. This method uses elementary flux mode (EFM) analysis to derive a reduced metabolic reaction network capturing key pathways, and combines this inferred reaction scheme with kinetic modeling that integrates mechanistic rate laws and machine learning (artificial neural networks) for improved predictive accuracy. This enables data-driven refinement of the kinetics, while preserving a biochemical basis for the model structure.

At the microscopic level, a constructive methodology was established for elaborating intermediate-size metabolic networks. This bottom-up approach iteratively integrates fundamental biological knowledge (core pathways, thermodynamic constraints) with constraint-based modeling tools (flux balance, variability, and coupling analyses) to curate a metabolic network of manageable size. The method includes explicit modeling of energy metabolism (e.g., proton motive force) to improve physiological realism. Two case studies on photosynthetic microorganisms (a cyanobacterium and a microalga) illustrate the construction of detailed metabolic models under both steady-state and dynamic (diurnal) conditions.

At the genetic-integration level, a new framework was developed to couple metabolic models with gene expression models in the context of synthetic biology. By incorporating a burden-aware gene expression formulation, the model constrains metabolic fluxes based on enzyme availability and resource allocation within the cell. This integrated approach was applied to *E. coli* engineered with a synthetic pathway, allowing direct simulation of how changes in gene expression (through the strengths of the promoter and the ribosome binding site of introduced genes) affect cellular growth and product formation.

Each modeling level yielded significant findings. The macroscopic hybrid model successfully overcame the combinatorial explosion of EFMs and improved predictive performance in animal cell culture case studies, showing that integrating neural net-

works can capture complex kinetics where traditional models fall short. The microscopic modeling approach led to validated metabolic networks for *Arthrospira* (yielding accurate flux predictions under various nutrient conditions) and *Tisochrysis* (capturing carbon storage dynamics in day/night cycles), demonstrating that mid-sized models can achieve a realistic balance between detail and tractability. The gene-integrated model significantly reduced the solution space of feasible flux distributions by enforcing regulatory constraints, and it identifies optimal gene expression strategies for maximizing the biosynthesis of a metabolite of interest (p-coumaric acid) while maintaining cell viability. Collectively, these results form a coherent multi-scale modeling strategy: the higher-level models benefit from mechanistic grounding, and the lower-level models are informed by broader physiological context, all contributing to a more comprehensive understanding of cellular behavior.

The methodologies developed in this thesis advance the state-of-the-art in metabolic modeling. They provide practical tools for researchers and engineers to better predict and manipulate cell behavior, from improving large-scale bioreactor performance to designing genetically modified strains for enhanced production. The scientific relevance of this research is underscored by its adaptability to various organisms and contexts, and by its validation through experimental comparisons and peer-reviewed publications.

In summary, this thesis establishes a structured approach to modeling cellular cultures at multiple levels, demonstrating that progressively expanding the scope of the model can yield mathematical representations that are predictive and biologically meaningful. These models offer strong potential for advancing biotechnological innovation and deepening our understanding of complex cellular systems.

Résumé

Modélisation métabolique des cultures cellulaires

Cette thèse de doctorat traite de la modélisation mathématique des cultures cellulaires à différentes échelles. La recherche s'articule autour de trois niveaux croissants de complexité biologique : la modélisation macroscopique, l'étude du métabolisme intracellulaire et l'intégration des aspects génétiques. Chaque niveau introduit un degré de détail mécanistique plus poussé, avec pour objectif de relier plus finement les comportements observables des cultures aux mécanismes moléculaires sous-jacents. Cette exploration progressive permet une représentation plus fidèle et biologiquement cohérente des systèmes cellulaires, avec une portée directe pour la conception et l'optimisation d'applications biotechnologiques.

Au niveau macroscopique, une approche hybride originale a été développée pour modéliser le comportement dynamique des cultures cellulaires. Cette méthode repose d'abord sur l'analyse des modes de flux élémentaires (EFMs), utilisée pour extraire un schéma réactionnel réduit représentant les principales voies métaboliques. La cinétique est ensuite modélisée à l'aide de lois mécanistiques ou de techniques d'apprentissage automatique (réseaux de neurones artificiels) afin d'améliorer la précision des prédictions. Cette approche permet de modéliser les cinétiques à partir des données expérimentales tout en conservant une structure biochimique cohérente.

Sur le plan microscopique, une méthodologie itérative a été établie pour construire des réseaux métaboliques de taille intermédiaire. Cette approche combine les connaissances biologiques fondamentales (voies centrales, contraintes thermodynamiques) avec des outils de modélisation à base de contraintes (analyses de bilan de flux, de variabilité et de couplage), afin d'obtenir des modèles à la fois précis et maîtrisés en complexité. La méthode inclut une modélisation explicite du métabolisme énergétique (notamment via la force proton-motrice) afin de renforcer la pertinence biologique. Deux études de cas, menées sur des microorganismes photosynthétiques (une cyanobactérie et une microalgue), illustrent le développement de modèles métaboliques dans des conditions stationnaires et dynamiques (cycles jour/nuit).

Au niveau de l'intégration génétique, une nouvelle approche a été développée pour coupler des modèles métaboliques avec des modèles d'expression génique dans le contexte de la biologie synthétique. En tenant compte de la charge cellulaire liée à l'expression des gènes, le modèle contraint les flux métaboliques en fonction de la disponibilité enzymatique et de l'allocation des ressources intracellulaires. Cette approche intégrée a été appliquée à une souche de *E. coli* modifiée génétiquement avec une voie synthétique, permettant de prédire l'effet de variations dans l'expression génique (via les forces de promoteurs ou des sites de liaison ribosomique) sur la croissance cellulaire et la production du métabolite cible.

Chaque niveau de modélisation a conduit à des résultats significatifs. Le modèle hybride macroscopique a permis de surmonter l'explosion combinatoire du nombre d'EFMs et d'améliorer les performances prédictives dans des cas d'étude de cultures cellulaires animales, montrant que l'intégration de réseaux neuronaux peut capturer des cinétiques complexes là où les modèles classiques échouent. L'approche de modélisation microscopique a abouti à des réseaux métaboliques validés pour *Arthrospira* (avec des prédictions de flux précises sous différentes conditions d'alimentation) et *Tisochrysis* (reproduisant les dynamiques de stockage du carbone en cycle jour/nuit), démontrant que des modèles de taille intermédiaire peuvent atteindre un bon compromis entre réalisme et complexité. Le modèle intégrant la régulation génétique a réduit significativement l'espace des solutions en imposant des contraintes régulatrices, et a permis d'identifier des stratégies d'expression génique optimales pour maximiser la biosynthèse d'un métabolite d'intérêt (l'acide p-coumarique) tout en préservant la viabilité cellulaire. Ensemble, ces résultats constituent une progression cohérente en modélisation : les modèles de plus haut niveau bénéficient d'un ancrage mécanistique, tandis que les modèles de plus bas niveau sont éclairés par un contexte physiologique plus large, contribuant à une compréhension plus complète du comportement cellulaire.

Les méthodologies développées dans cette thèse apportent de nouvelles stratégies de modélisation métabolique à différentes échelles biologiques. Elles offrent des outils concrets pour prédire et manipuler le comportement cellulaire, qu'il s'agisse d'optimiser les performances des bioréacteurs à grande échelle ou de concevoir des souches génétiquement modifiées pour une production accrue. La portée scientifique de ce travail est soulignée par son adaptabilité à divers organismes et contextes d'application, ainsi que par sa validation expérimentale et sa reconnaissance au travers de publications évaluées par les pairs.

En résumé, cette thèse établit une approche structurée de la modélisation des cultures cellulaires à plusieurs niveaux, montrant que l'élargissement progressif du champ de modélisation permet d'obtenir des représentations mathématiques à la fois prédictives et biologiquement pertinentes. Ces modèles offrent un fort potentiel pour faire progresser l'innovation biotechnologique et approfondir notre compréhension des systèmes cellulaires complexes.

Nomenclature

Acronym or symbol	Meaning
AA	Amino Acid
ANN	Artificial Neural Network
ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
B	Functional Biomass
CARB	Carbohydrate
CoA	Coenzyme A
CV	Coefficient of Variation
dFBA (DFBA)	Dynamic Flux Balance Analysis
dFVA (DFVA)	Dynamic Flux Variability Analysis
DMFA	Dynamic Metabolic Flux Analysis
DNA	Deoxyribonucleic Acid
DRUM	Dynamic Reduction of Unbalanced Metabolism
EFM	Elementary Flux Mode
ER	Endoplasmic Reticulum
ETC	Electron Transport Chain
FAD ⁺	Flavin Adenine Dinucleotide – Oxidized form and electron acceptor
FADH ₂	Flavin Adenine Dinucleotide – Reduced form and electron donor
FBA	Flux Balance Analysis
FCA	Flux Coupling Analysis
FVA	Flux Variability Analysis
G3P (GAP)	Glyceraldehyde-3-phosphate
G6P	Glucose-6-phosphate
GEM	Genome-scale Metabolic Model
GTP	Guanosine-5'-triphosphate
H ⁺	Hydrogen ion – Proton
MFA	Metabolic Flux Analysis
MLP	Multilayer Perceptron
MPA	Metabolic Pathway Analysis
MR	Macro-reaction
mRNA	Messenger RNA
NAD ⁺	Nicotinamide Adenine Dinucleotide – Oxidized form and electron acceptor
NADH	Nicotinamide Adenine Dinucleotide – Reduced form and electron donor
NADP ⁺	Nicotinamide Adenine Dinucleotide Phosphate – Oxidized form
NADPH	Nicotinamide Adenine Dinucleotide Phosphate – Reduced form
NH ₃	Ammonia

NH_4^+	Ammonium
NO_2^-	Nitrite
NO_3^-	Nitrate
NN	Neural Network
ODE	Ordinary Differential Equation
P_i	Inorganic phosphate
PA	Phosphatidic Acid
PBR	Photobioreactor
PEP	Phosphoenolpyruvate
PMF	Proton Motive Force
PPP	Pentose Phosphate Pathway
PSI	Photosystem I
PSII	Photosystem II
QSSA	Pseudo-steady-state assumption – Quasi-steady-state assumption
SN	Sub-network
STR	Stirred Tank Reactor
TC	Total Carbon
TCA	Tricarboxylic Acid cycle
THF	Tetrahydrofolate
TIC	Total Inorganic Carbon
TOC	Total Organic Carbon
TN	Total Nitrogen
R5P	Ribose-5-phosphate
RBS	Ribosome Binding Site
RHS	Right-Hand-Side
RNA	Ribonucleic Acid

E	Matrix of the elementary flux modes
K	Stoichiometric matrix of the macroscopic reactions
$\mathbf{m}_e$	Number of extracellular measurements
N	Stoichiometric matrix of the metabolic network
$\mathbf{N}_e$	Stoichiometric matrix of the extracellular measurements
γ^*	Photon
Λ	Second target number for the reduced set of EFMs in the reduction procedure
μ	Specific growth rate of the cell
$\underline{\nu}_m$	Vector of specific reaction rates obtained from experimental data (often referred to as reaction rates in the manuscript for convenience, although strictly speaking they correspond to specific rates) - $\mathcal{N}$ is the matrix formalism
ξ	Vector of metabolite concentrations
Ξ	Performance index of the optimization problem in the reduction procedure
$\underline{\phi}$	Vector of specific metabolic fluxes equivalent to specific reaction rates - Φ is the matrix formalism
$\underline{\phi}_{num}$	Vector of reaction rates obtained numerically from the EFM reduction procedure
Ω	First target number for the reduced set of EFMs in the reduction procedure
$\underline{r}$	Vector of actual metabolic fluxes equivalent to total reaction rates - R is the matrix formalism

Contents

1	Introduction	1
1.1	Motivations	1
1.2	Contributions	4
1.3	Thesis Outline	6
1.4	List of Publications	7
2	Preliminaries	9
2.1	Cellular Foundations	10
2.2	Cellular Metabolism	14
2.2.1	Glycolysis and Pyruvate Fate	15
2.2.2	Pentose Phosphate Pathway	16
2.2.3	Tricarboxylic Acid Cycle	17
2.2.4	Lipid Metabolism	18
2.2.5	Amino Acid Metabolism	18
2.2.6	Nucleotide Metabolism	19
2.2.7	Protein Synthesis	19
2.3	Bioreactors and Operating Modes	19
2.3.1	Batch Mode	20
2.3.2	Fed-Batch Mode	20
2.3.3	Continuous Mode	20
2.3.4	Perfusion Mode	21
2.4	Basics of Metabolic Analysis	21
2.4.1	Theoretical Background	21
2.4.2	Metabolic Flux Analysis	25
2.4.3	Toy Example	26
2.5	Fundamentals of Genetics	32
2.5.1	Gene Expression	32
2.5.2	Genetic Elements	33
2.5.3	Genetic Models	33
2.5.4	Synthetic Biology	34
2.6	Artificial Neural Networks	34
2.6.1	Neural Network Learning Algorithms	35
2.6.2	Neural Network Architectures	36
2.7	Parameter Estimation	38
2.7.1	Linear Parameter Estimation	38
2.7.2	Nonlinear Parameter Estimation	39

3	Macroscopic Modeling	41
3.1	Literature Review	45
3.1.1	Model Derivation	45
3.1.2	Kinetic Modeling	47
3.2	Hybrid Modeling	49
3.2.1	EFM Selection	49
3.2.1.1	Theoretical Background	49
3.2.1.2	EFM Reduction Procedure	50
3.2.1.2.1	Generation of EFMs	51
3.2.1.2.2	Biological Interpretation of the EFMs	52
3.2.1.2.3	Main Reduction of EFMs	52
3.2.1.2.4	Generation of the Final Macroreaction Set	53
3.2.1.3	Dynamic Mass Balance Model	54
3.2.2	Kinetic Identification	55
3.2.2.1	Model Structure	55
3.2.2.1.1	Mechanistic and Empirical Models	56
3.2.2.1.2	Data-Driven Models	57
3.2.2.2	Parameter Estimation	59
3.3	Application to Hybridoma Cell Line HB-58	60
3.3.1	Cell Line and Media	60
3.3.2	Metabolic Network	60
3.3.3	Measurement Configuration	61
3.3.4	Numerical Results	62
3.3.4.1	EFM Reduction Procedure	62
3.3.4.2	Kinetic Modeling	70
3.3.4.2.1	Simple Kinetic Laws	70
3.3.4.2.2	Generalized Kinetic Functions	70
3.3.4.2.3	Multilayer Perceptrons	70
3.3.4.3	Cross-Validation	81
3.3.5	Overflow Metabolism	81
3.4	Application to Cell Line CHO-320	85
3.4.1	Cell Line and Media	85
3.4.2	Metabolic Network	85
3.4.3	Measurement Configuration	86
3.4.4	Numerical Results	86
3.4.4.1	EFM Reduction Procedure	86
3.4.4.2	Kinetic Modeling	94
3.4.4.3	Cross-Validation	97
3.4.4.4	Positivity Constraints	97
3.5	Prospects	100
3.5.1	EFM Selection Process	100
3.5.2	Kinetic Modeling	100
3.5.3	Further Considerations	100
3.6	Conclusions	101
4	Microscopic Modeling	104
4.1	Literature Review	107
4.1.1	Genome-Scale Metabolic Networks	107
4.1.2	Small-Size Metabolic Networks	108
4.1.3	Mid-Size Metabolic Networks	108
4.2	Constructive Modeling Approach	110
4.2.1	Core Metabolic Pathway Selection	111

4.2.2	Reaction Reversibility	112
4.2.3	Mathematical Methods	113
4.2.3.1	Elementary Mass Conservation	114
4.2.3.2	Constraint-Based Methods	115
4.2.3.2.1	Classical Assumptions	115
4.2.3.2.2	Flux Balance Analysis	116
4.2.3.2.3	Flux Variability Analysis	116
4.2.3.2.4	Flux Coupling Analysis	117
4.2.4	Cellular Energy Metabolism	118
4.3	Application to <i>Arthrospira</i> Sp. PCC 8005	122
4.3.1	Elaboration of a Mid-Size Metabolic Network	123
4.3.1.1	Main Metabolic Routes	123
4.3.1.1.1	Carbon Uptake	124
4.3.1.1.2	Embden-Meyerhof-Parnas Pathway	124
4.3.1.1.3	Pentose Phosphate Pathway	124
4.3.1.1.4	Calvin-Benson Cycle	124
4.3.1.1.5	Tricarboxylic Acid Cycle	125
4.3.1.1.6	GS-GOGAT Pathway	125
4.3.1.1.7	Urea Cycle-Like Pathway	126
4.3.1.1.8	Energy Metabolism	127
4.3.1.1.9	Other Metabolic Routes	127
4.3.1.2	Iterative Methodology	128
4.3.1.2.1	First Iteration	128
4.3.1.2.2	Second Iteration	130
4.3.1.2.3	More Iterations	131
4.3.1.2.4	Final Metabolic Network	131
4.3.2	In-Silico Validation	131
4.3.2.1	Day Phase	132
4.3.2.1.1	Carbon-Only Uptake Scenario	133
4.3.2.1.2	Carbon-Nitrogen Uptake Scenario	134
4.3.2.2	Dark Phase	134
4.3.2.2.1	Carbon-Only Uptake Scenario	135
4.3.2.2.2	Carbon-Nitrogen Uptake Scenario	135
4.3.3	Experimental Validation	136
4.3.3.1	Cell Line and Media	136
4.3.3.2	Direct Validation	136
4.3.3.2.1	Experimental Constraints	138
4.3.3.2.2	Fundamental Tests	139
4.3.3.2.3	Dynamic Flux Variability Analysis	140
4.3.3.2.4	Dynamic Flux Balance Analysis	144
4.3.3.2.5	Impact of Nitrogen Sources on Metabolism	145
4.3.3.3	Further Validation	148
4.4	Application to Unicellular Microalgae	148
4.4.1	Elaboration of a Mid-Size Metabolic Network	149
4.4.1.1	Starting Point	149
4.4.1.1.1	Main Metabolic Routes	149
4.4.1.1.2	Eukaryotic vs. Prokaryotic Modeling Paradigms	150
4.4.1.2	Iterative Methodology	152
4.4.1.2.1	First Iteration	153
4.4.1.2.2	Second Iteration	154
4.4.1.2.3	Eukaryotic vs. Prokaryotic Networks	154
4.4.2	The DRUM Framework	155

4.4.2.1	Basic Principle	155
4.4.2.2	No Accumulation of Internal Metabolites	156
4.4.2.3	Internal Metabolite Accumulation Issue	156
4.4.2.4	Accumulating Metabolites vs. Excreted Products	157
4.4.3	Experimental Validation	158
4.4.3.1	Cell Line and Media	158
4.4.3.2	Numerical Results	158
4.4.3.2.1	Comparison of Macroscopic Reactions	159
4.4.3.2.2	Macroscopic Model Derivation	160
4.4.3.2.3	Discussion	163
4.5	Prospects	167
4.5.1	Methodological Improvements	167
4.5.2	Perspectives for <i>Arthrospira Sp. PCC 8005</i>	169
4.5.3	Perspectives for <i>Tisochrysis Lutea</i>	171
4.6	Conclusions	171
5	Microscopic Modeling With Genetic Insights	174
5.1	Literature Review	177
5.1.1	Modeling Frameworks	177
5.1.2	Regulatory and Enzyme-Constrained Models	178
5.1.3	Conciliation Strategies	179
5.2	Synthetic Biology Foundations	180
5.2.1	Regulatory Design Elements	180
5.2.1.1	Promoter Strength	181
5.2.1.2	RBS Strength	181
5.2.2	Endogenous vs. Engineered Systems	181
5.2.2.1	Endogenous Genes	182
5.2.2.2	Exogenous Genes	183
5.2.2.3	Parameter Tuning	184
5.3	Constraint-Based Burden-Aware Modeling	184
5.3.1	Metabolic Modeling	184
5.3.2	Genetic Modeling	185
5.3.2.1	Protein-Coding Gene Dynamics	185
5.3.2.2	Cell Growth Expression	186
5.3.3	Conciliation Strategy	187
5.3.3.1	Model Adaptations and Extensions	188
5.3.3.1.1	Mapping Substrate Availability	189
5.3.3.1.2	Adding Engineered Pathways	189
5.3.3.1.3	Expressing Exogenous Genes	190
5.3.3.2	Algorithms	190
5.3.3.2.1	Conciliation Algorithm	190
5.3.3.2.2	Promoter Strength Optimization	191
5.4	Application to p-Coumaric Acid Biosynthesis	193
5.4.1	Engineering E. coli for p-Coumaric Acid Production	194
5.4.2	Preliminary Setup	194
5.4.2.1	Substrate Mapping	194
5.4.2.2	Engineered Pathway Integration	196
5.4.2.3	Exogenous Gene Expression	197
5.4.3	Numerical Results	197
5.5	Prospects	201
5.5.1	Alternative Gene Expression Models	201
5.5.2	Advancing the Conciliation Strategy	202

5.5.3	Extensions to Endogenous Proteins	203
5.5.4	Multi-Step Synthetic Pathway	203
5.5.5	Further Considerations	204
5.6	Conclusions	205
6	Conclusion and Prospects	208
6.1	EFM-Derived Macroscopic Modeling	208
6.2	Microscopic Metabolic Modeling	209
6.3	Constraint-Based Burden-Aware Modeling	211
6.4	Final Conclusion	212
6.5	Perspectives	213
A	Systematic Methodology for Bioprocess Model Identification Based on Generalized Kinetic Functions	218
A.1	Original Method	218
A.2	Revisited Method	220
B	Metabolic Reactions for the Metabolism of Hybridoma Cell Line HB-58	222
C	Metabolic Reactions for the Metabolism of Chinese Hamster Ovary Line CHO-320	224
D	Experimental Rate Derivation	227
D.1	Derivation From Smoothing Splines and Differentiation Methods	227
D.2	Derivation From Linear Regression	228
D.3	Uncertainty Estimation of Experimental Rates	229
E	Neural Network Parameter Values	231
F	Supplementary Quantitative Analysis	232
F.1	Methods and Metrics	232
F.2	Applications	233
G	Metabolic Network for Arthrospira Sp. PCC 8005	238
H	Focused Metabolic Network for Arthrospira Sp. PCC 8005	245
H.1	Network Reduction	245
H.2	Data Preprocessing	246
H.3	Well-Posedness Analysis	247
H.4	Calculability Analysis	247
I	Metabolic Network for Tisochrysis Lutea	253
I.1	Original Network	253
I.2	Extended Network	256
J	Metabolic Compartmentalization	258
K	Additional Metabolic Reactions for the Metabolism of Purple Non-Sulfur Bacteria in Rhodospirillum Rubrum	260
L	Identifiability Analysis	262

List of Figures

1	Schematic view of the thesis organization. Cellular culture processes are studied at three levels. (i) Macroscopic modeling derives macroscopic models from elementary flux modes (EFMs), assuming prior knowledge of the metabolic network through its stoichiometric matrix N . (ii) Microscopic modeling addresses the construction of the metabolic network itself and the quantification of intracellular fluxes using constraint-based approaches. (iii) Genetic regulation integrates catalytic constraints and gene expression models, further restricting the solution space by accounting for regulatory effects on flux values. This figure is presented at the beginning to guide the reader through the different modeling scales discussed in this doctoral thesis.	xxv
1.1	The different modeling approaches: (i) macroscopic modeling derived from elementary flux modes connecting extracellular substrates (e.g., glucose) to extracellular products (e.g., lactate and alanine), (ii) microscopic modeling emphasizing the study of intracellular fluxes v_i , and (iii) microscopic modeling integrating genetic aspects, as enzymes that catalyze biochemical reactions are products of genetic transcription and translation.	7
2.1	Features of living cells, from [1]. Nucleus or nucleoid contains genetic material-DNA and associated proteins, plasma membrane is a tough, flexible lipid bilayer, and cytoplasm contains an aqueous solution and suspended particles and organelles.	11
2.2	Classification of organisms according to their source of energy (sunlight or oxidizable chemical compounds) and their source of carbon for the synthesis of cellular material, from [1].	12
2.3	Eukaryotic cell structure. Schematic illustrations of two major types of eukaryotic cell: (a) a representative animal cell and (b) a representative plant cell, from [1].	13
2.4	Prokaryotic cell structure. Schematic illustration of bacterial cells, from [1].	14
2.5	Relationships between catabolic and anabolic pathways. Catabolic pathways deliver chemical energy in the form of ATP, NADH, NADPH and $FADH_2$ used in anabolic pathways, from [1].	15
2.6	Catabolic fates of pyruvate formed in glycolysis.	16
2.7	Pentose Phosphate Pathway: oxidative phase in red and chemical reactions related to non-oxidative phase in blue.	17
2.8	Tricarboxylic acid (TCA) cycle and its role in anabolism. Red arrows represent anaplerotic reactions that replenish depleted cycle intermediates, from [1].	18
2.9	Schematic representation of a stirred tank bioreactor.	20

2.10	The different operating modes of bioreactors: batch, fed-batch, continuous, and perfusion.	21
2.11	Bioreactor in perfusion operating mode.	22
2.12	Pointed polyhedral cone $\mathbf{S}$ in the positive orthant, the edges of $\mathbf{S}$ representing the elementary modes e_i , and the polytope $\mathbf{F}$	24
2.13	Metabolic network containing 18 internal metabolites and 24 intracellular reactions. Internal metabolites are represented by white boxes, extracellular substrates are represented by pink boxes, extracellular products are represented by blue boxes, and the terminal biosynthetic product is in the orange box. The intracellular fluxes are v_i and some elementary flux modes are illustrated as e_i	28
2.14	Transcription and translation steps in eukaryotes.	33
2.15	A single layer perceptron with one output neuron: x_i represents the input signals, w_i denotes synaptic coefficients, b is a bias, f is an activation function and y is the output signal.	35
2.16	Sequential decision process of reinforcement learning.	37
3.1	Hybrid modeling using an elementary flux mode reduction procedure and neural networks. N is the stoichiometric matrix, N_e is the stoichiometric matrix of extracellular measurements, $\underline{\nu}_m$ is the vector of specific uptake and excretion rates obtained from experimental data, K describes the stoichiometry of the bioreactions, $\underline{\phi}_{num}$ is the vector of specific metabolic fluxes obtained numerically from the reduction procedure, $\underline{\phi}$ is the vector of specific reaction rates modeled using classical kinetic laws or ANNs and $\underline{\xi}$ is the vector of concentrations of culture species.	49
3.2	The modular EFM selection procedure. N is the stoichiometric matrix, N_e is the stoichiometric matrix of extracellular measurements, $\underline{\nu}_m$ is the vector of specific uptake and excretion rates obtained experimentally, E is the matrix of EFMs and K describes the stoichiometry of the bioreactions. Ω is a first target number for the reduced set of EFMs, close to the number of measured extracellular species m_e , and Λ is a second target number for the reduced set of EFMs, smaller than m_e	51
3.3	Kinetic identification procedure. $\underline{\xi}$ is the concentration vector of measured extracellular species, $\underline{\phi}$ is the vector of reaction rates predicted by the kinetic model, $\underline{\phi}_{num}$ is the vector of reaction rates to model (i.e., the rates obtained from the EFM reduction procedure) and $\underline{\theta}$ is the vector of parameters of the model.	55
3.4	A one hidden-layer MLP with Z neurons, n input signals in the input layer and m output signals in the output layer. The nonlinear activation functions in the hidden layer are sigmoid functions, while activation functions in the output layer are identity functions.	58
3.5	Architecture of a multilayer perceptron before and after the application of pruning algorithms – the synapse pruning technique is shown in green and the neuron pruning technique in yellow.	59
3.6	Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) in datasets 1, 2 and 4 (direct validation) computed using smoothing splines and differentiation methods.	65
3.7	Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) in datasets 1, 2 and 4 (direct validation): numerical results (black curves); $K_e \Phi$ for $n_{EFM} = 214$ (green crosses) – results from the prefiltering step.	66

3.8	Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) in datasets 1, 2 and 4 (direct validation): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = \Omega = 10$ (red crosses); $K_e\Phi$ for $n_{EFM} = \Lambda = 5$ (blue bullets).	67
3.9	Time evolution of the measured concentrations in datasets 1, 2 and 4 (direct validation) – experimental data points (black bullets); concentration profiles on the basis of $n_{EFM} = 214$ (green lines) and reducing to $n_{EFM} = \Omega = 10$ (red dashed lines).	68
3.10	Time evolution of the measured concentrations in datasets 1, 2 and 4 (direct validation) – experimental data points (black bullets); concentration profiles on the basis of $n_{EFM} = \Omega = 10$ (red lines) and selecting $n_{EFM} = \Lambda = 5$ (blue dashed lines).	69
3.11	A one hidden-layer MLP with 3 neurons, 6 input signals (which are the measured concentrations of culture species) in the input layer and 5 output signals (which are the total reaction rates of each bioreaction) in the output layer. The biases are denoted by b_j in the hidden layer and c_k in the output layer. The nonlinear activation functions in the hidden layer are sigmoid functions, while activation functions in the output layer are identity functions.	72
3.12	Time evolution of the total reaction rates r_i (in mM.h ⁻¹) in datasets 1, 2 and 4 (direct validation) – target values, i.e., rates coming from the EFM reduction procedure (black curves); prediction with an MLP with ten neurons (red curves), five neurons (blue curves), four neurons (magenta curves), three neurons (green curves) and two neurons (cyan curves) in the hidden layer.	76
3.13	Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) – results coming from the EFM reduction procedure (black curves); prediction with an MLP with ten neurons (red curves), five neurons (blue curves), four neurons (magenta curves), three neurons (green curves) and two neurons (cyan curves) in the hidden layer.	77
3.14	Time evolution of the total reaction rates r_i (in mM.h ⁻¹) in datasets 1, 2 and 4 (direct validation) – target values, i.e., rates coming from the EFM reduction procedure (black curves); prediction with an MLP with three neurons in the hidden-layer using backpropagation algorithm (red curves), MLP with three neurons and nonlinear re-identification of weights and biases (blue curves) and pruned MLP with three neurons (green curves).	78
3.15	Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) – results coming from the EFM reduction procedure (black curves); prediction with an MLP with three neurons in the hidden-layer using backpropagation algorithm (red curves), MLP with three neurons and nonlinear re-identification of weights and biases (blue curves) and pruned MLP with three neurons (green curves).	79
3.16	Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) – results coming from the EFM reduction procedure (black curves); prediction with a pruned MLP with three neurons using magnitude pruning algorithm (red curves) and prediction with a pruned MLP with three neurons using the pruning strategy based on CV values (blue curves).	80

3.17	Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation – red curves) and dataset 3 (cross-validation – purple curves) – results coming from the EFM reduction procedure considering $\Lambda = 5$ EFMs.	83
3.18	Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation – red curves) and dataset 3 (cross-validation – purple curves) – prediction with the pruned MLP using the magnitude pruning algorithm.	84
3.19	Prediction of the time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = m_e = 6$ (black curves), reducing $n_{EFM} = \Lambda_1 = 5$ (red curves) and $n_{EFM} = \Lambda_2 = 4$ (dashed blue curves).	89
3.20	Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) in dataset 1 (left-hand side) and dataset 2 (right-hand side): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = 173$ (red crosses); $K_e\Phi$ for $n_{EFM} = \Omega = 10$ (blue bullets).	90
3.21	Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = 173$ (red curves), reducing $n_{EFM} = \Omega = 10$ (blue curves).	91
3.22	Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) in dataset 1 (left-hand side) and dataset 2 (right-hand side): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = \Omega = 10$ (red crosses); $K_e\Phi$ for $n_{EFM} = \Lambda_1 = 5$ (blue bullets) and $K_e\Phi$ for $n_{EFM} = \Lambda_2 = 4$ (green squares).	92
3.23	Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = \Omega = 10$ (red curves), reducing $n_{EFM} = \Lambda_1 = 5$ (blue curves) and reducing to $n_{EFM} = \Lambda_2 = 4$ (green curves).	93
3.24	Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points are black bullets and the concentration profiles computed from the ODE system (dynamical simulator) are depicted by the red curves.	96
3.25	Time evolution of the extracellular measured species in dataset 3 (cross-validation) – experimental data points (black bullets), results from the EFM reduction procedure considering $n_{EFM} = \Lambda_2 = 4$ modes (left-hand side); concentration profiles computed using the dynamical simulator (right-hand side).	98
3.26	Time evolution of the index $ K_e\Phi - \nu $ for the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – positivity constraint violation (blue curves) when $n_{EFM} \geq m_e$ ($n_{EFM} = 173$ and $m_e = 6$); model discrepancy (pink curves) when $n_{EFM} < m_e$ ($n_{EFM} = 5$ and $m_e = 6$).	99
4.1	Existing modeling approaches on the right-hand side and the proposed constructive methodology on the left-hand side.	111
4.2	Algorithmic scheme of the mathematical methods. Note that FCA is preferred to FVA for larger networks. The latter methods should give equivalent results (cross-validation of the methods).	114

4.3	Schematic representation of the proton motive force across the mitochondrial membrane (left-hand side), the thylakoid membrane of chloroplasts (center) and the plasma membrane (right-hand side), from [1].	119
4.4	Flow of electrons and protons through the 4 complexes of the respiratory chain depicting the process of oxidative phosphorylation, from [1]. Q represents the cofactor FADH ₂	120
4.5	Illustration of the process of photophosphorylation, from [1]. Photosystem II (PSII) absorbs photons, splits water into O ₂ , H ⁺ protons and electrons while generating a proton gradient for ATP synthesis. Photosystem I (PSI) then uses the excited electrons to reduce NADP ⁺ into NADPH.	121
4.6	The three stages of CO ₂ assimilation in photosynthetic organisms: (1) fixation of CO ₂ into 3-P-glycerate, (2) conversion of 3-P-glycerate to glyceraldehyde-3-P and (3) regeneration of RubisCO from triose phosphates.	125
4.7	Incomplete citric acid cycle and the main intermediates which are precursors in many biosynthetic pathways. Anaplerotic reactions are shown in red.	126
4.8	Overview of the methodological framework used to validate the metabolic network. Preliminary processes are shown in blue, including data pre-processing, a well-posedness check and a calculability analysis. These steps ensure the model is computationally tractable before proceeding to experimental validation.	137
4.9	Experimental growth rate μ_{exp} and specific growth rate μ_{opt} obtained from dFBA for the different feeding strategies: 30mM-N nitrate (top-left, top-right, center-left), 30mM-N nitrate in high salinity medium (center-right), 15mM-N nitrate + 15mM-N nitrite (bottom-left) and 15mM-N nitrate + 15mM-N urea (bottom-right).	140
4.10	Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{27} to v_{36} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the tricarboxylic acid cycle.	141
4.11	Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{108} to v_{111} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the cellular energy metabolism.	142
4.12	Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{116} to v_{123} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the biosynthesis of carbohydrates.	143
4.13	Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{126} to v_{141} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the biosynthesis of lipids.	144
4.14	Simplified flux map (using <i>CellDesigner</i>) at a specific time instant for the 30mM-N nitrate feeding strategy. The flux map depicts the energy cellular metabolism, glycolysis and the pentose phosphate pathway. Active reactions are represented by green arrows.	145
4.15	The fluxes are expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$. Fluxes v_{38} and v_{39} are related to the GS-GOGAT pathway, v_{47} , v_{60} , v_{147} and v_{149} correspond to the production rate of arginine, proline, chlorophyll and cyanophycin, respectively.	147

4.16	Integration of the bottom-up strategy into modeling frameworks. The first case study allows microscopic validation under constant light regimes via the classical MFA tools, while the second concludes with macroscopic validation using the DRUM framework under dynamic light/dark conditions.	155
4.17	Evolution of intracellular carbon quotas for GAP, G6P, PEP, carbohydrates (CARB), PA (storage lipids), and functional biomass B throughout a day/night cycle. Comparison between the three models (original model from [2] in red, the re-calibrated model in purple, and the newly constructed model in green). Light intensity is shown to distinguish day and night phases.	164
4.18	Dynamics of the functional biomass B over time during a day/night cycle. Predictions from the original model (red curves), the re-calibrated model (purple curves), and the new model (dark green curves). Experimental data are represented by black and gray bullets.	165
4.19	Evolution of the total biomass X expressed both as carbon content (mg_C/L) and nitrogen content (mg_N/L) during a day/night cycle. Prediction of the three models against experimental data.	166
4.20	Dynamics of chlorophyll content ($\mu\text{g}/\text{L}$) over time during a day/night cycle. Predictions from the original model (red curve), the re-calibrated model (purple curve), and the new model (dark green curve).	167
4.21	Profiles of the macroscopic reaction rates over time for the photosynthesis, carbon allocation, lipid synthesis, biomass formation, and related processes. Predictions from the original model (red curves), the re-calibrated model (purple curves), and the new model (dark green curves).	168
5.1	Schematic representation of the transcription stage (initiation step, elongation step and termination step).	181
5.2	Schematic representation of the translation stage and synthesis of the protein (sequence of amino acids).	182
5.3	Schematic view of cellular resource allocation between protein types. On the left-side, the partitioning of ribosomal resources for the synthesis of ribosomal, non-ribosomal endogenous proteins, and exogenous proteins (when expressed). On the right-side, pie charts depict how the fractions of cell dry weight vary with growth rate and substrate availability under different promoter/RBS strategies (from [3]).	186
5.4	Input-output characteristics of constraint-based metabolic models and gene expression models. Additional elements highlighted in red indicate the extra considerations required when accounting for the expression of exogenous genes.	188
5.5	Top-left: Growth rate μ versus substrate availability s from the gene expression model (black points: simulation data; red curve: fitted Michaelis-Menten-type law). Bottom-left: Growth rate μ versus glucose uptake rate ν_{Glc} from constraint-based metabolic model with biomass optimization (simulation data). Right-side: Mapping between s and ν_{Glc} by matching growth rates (black bullets: matched data; curve: fitted functional relationship).	195
5.6	Evolution of the mass fractions of the cell, the growth rate, and the concentration of the exogenous enzyme, at steady-state, for different values of the promoter strength controlling the expression of the TAL enzyme.	198

5.7	Evolution of v_{TAL}^{ub} for different values of the promoter strength of the exogenous protein TAL. FVA results from the classical constraint-based modeling framework are depicted in blue and outcomes from the constraint-based burden-aware modeling framework are shown in purple.	200
5.8	Full biosynthetic pathway for Naringenin from L-tyrosine.	204
6.1	Multi-scale nature of cellular culture processes: macroscopic process dynamics, microscopic intracellular metabolism and gene-level regulatory integration.	217
G.1	Complete flux map of the final metabolic network obtained from the constructive procedure (the flux map is constructed using <i>CellDesigner</i>). Online reading is advised to identify the main metabolic routes.	243
G.2	Complete flux map during the day phase for a net assimilation of 100 mol of HCO_3^- . Online reading is advised to identify the main metabolic routes. Active routes are represented in orange.	244
J.1	Left-hand side: Prokaryotic cells (e.g., bacteria such as <i>E. coli</i>) have a simple, two-compartment system. Right-hand side: Eukaryotic organisms offer many specialized organelles, each with conditions distinct from the cytosol, that may provide a better environment for a range of metabolic pathways. From [4]. As explained in Sec. 2.1, glyoxysomes are a specialized form of peroxisomes.	258

List of Tables

2.1	Reactions for the metabolism of CHO-320 cells during growth phase. . .	26
2.2	Stoichiometric matrix N describing the metabolism of CHO-320 cells during growth phase.	29
2.3	Stoichiometric matrix of the extracellular measurements, denoted N_e . .	30
2.4	Matrix of elementary flux modes, denoted $E^{n \times n_{EFM}}$, where $n_{EFM} = 11$. .	31
3.1	Summary of culture conditions for experiments in perfusion.	60
3.2	The stoichiometric coefficients of the reaction network.	64
3.3	Number of NN parameters for 6 input signals and 5 output signals. . .	73
3.4	Coefficients of variation (in %) of the estimated weights and biases – colored boxes are the removed synaptic connections or biases after using pruning algorithms.	74
3.5	Results of the magnitude pruning algorithm – colored boxes are the removed synaptic connections or biases.	74
3.6	Estimated values of the specific reaction rates ν_{ξ_i} (in mM.h ⁻¹ per 10 ⁹ cells) and ν_X (in h ⁻¹) obtained by linear regression (L.R.) and prediction from the optimization problem of the reduction procedure.	87
3.7	The stoichiometric coefficients of the reaction network.	94
3.8	Value of the kinetic parameters.	95
4.1	Information related to metabolic networks from the iterative procedure. The number of metabolites is comprised of internal metabolites and intracellular energetic cofactors, and the number of reactions includes the intracellular reactions and the transport/exchange reactions.	129
4.2	Relative flux error (RFE) for all folds.	148
4.3	Information related to metabolic networks from the iterative procedure. The number of metabolites is comprised of internal metabolites and intracellular energetic cofactors, and the number of reactions includes the intracellular reactions and the transport/exchange reactions.	153
4.4	Macroscopic reactions obtained using the DRUM framework.	159
4.5	Kinetic structure of the macroscopic reactions.	161
4.6	Parameters for the original model of Baroukh et al., the original model re-calibrated using the normalized cost function, and for the newly constructed model.	162
4.7	Additional reactions for the synthesis of phycocyanobilin.	170
4.8	Additional reactions for the synthesis of the polypeptide chains and the formation of phycocyanin.	170
5.1	Parameter values and constants	199
B.1	Metabolic reactions for the metabolism of Hybridoma cells.	222

C.1	Metabolic reactions for the metabolism of Chinese Hamster Ovary cells during the growth phase.	224
E.1	Neural network parameter values after reidentification of the biases and weights – colored boxes indicate synaptic connections or biases that were removed following the application of the magnitude pruning algorithm. .	231
F.1	Results of the pre-filtering step using the cosine-similarity algorithm. The value Ξ^0 represents the best value of the indicator when all the EFMs are considered without any reduction. More information can be found in our original study in [5].	234
F.2	Results of the optimization-based reduction and selection steps.	234
F.3	RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification with ANNs) with $n_{EFM} = \Lambda = 5$	234
F.4	Global RMSE applied to MLP identification of metabolic fluxes.	235
F.5	RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with ANNs.	235
F.6	RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with MLP with 3 neurons in the hidden-layer.	235
F.7	RMSE applied to the prediction of the measured concentrations after identification of fluxes with a pruned MLP with 3 neurons in the hidden-layer.	236
F.8	Results of the pre-filtering step using the cosine-similarity algorithm. . .	236
F.9	Results of the optimization-based reduction and selection steps.	236
F.10	RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification).	236
F.11	RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification) with $n_{EFM} = \Lambda_2 = 4$.	237
F.12	RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with classical kinetic laws.	237
G.1	Metabolic reactions for the metabolism of <i>Arthrospira spirulina platensis</i> PCC 8005 (photoautotrophic regime). Protons H^+ involved in most metabolic reactions correspond to H_n^+	238
H.1	Metabolic reactions for the metabolism of <i>Arthrospira spirulina platensis</i> PCC 8005 (photoautotrophic regime). Protons H^+ involved in most metabolic reactions correspond to H_n^+	248
I.1	Original metabolic network for <i>Tisochrysis lutea</i> , from [2].	253
I.2	Additional reactions for the pathway of carbohydrate synthesis of <i>Tisochrysis lutea</i> (First iteration).	256
I.3	Additional reactions for the pathway of carbohydrate synthesis of <i>Tisochrysis lutea</i> (Second iteration).	257
J.1	Overview of the main compartmentalized metabolic pathways (catabolic and anabolic) across different cell types.	259
K.1	Additional metabolic reactions for the metabolism of purple non-sulfur bacteria in <i>Rhodospirillum rubrum</i>	260

Reading Guide

Guiding Thread: The Modeling of Cellular Cultures Through Increasingly Detailed Representations

- **Chapter 1 – Introduction**
- **Chapter 2 – Preliminaries**
- **Chapter 3 – Macroscopic Modeling**

Two case studies illustrating the modularity of the method:

- *Case 1:* Medium-sized metabolic network → Full enumeration of EFMs and neural-based kinetics;
- *Case 2:* Larger metabolic network → Identification of EFM subsets and mechanistic kinetics.

Application to animal cells (HB-58 and CHO-320 cell lines)

- **Chapter 4 – Microscopic Modeling**

Two case studies applied to photosynthetic microorganisms:

- *Case 1 (Arthrospira sp. PCC 8005):* De novo metabolic network reconstruction + Experimental validation under constant light → Classical MFA tools;
- *Case 2 (Tisochrysis Lutea):* Extension of an existing metabolic network + Experimental validation beyond QSSA applicability → DRUM approach.

- **Chapter 5 – Microscopic Modeling With Genetic Insights**

One case study in synthetic biology applied to engineered *E. coli*.

- **Chapter 6 – Conclusion and Prospects**
- **Appendices**

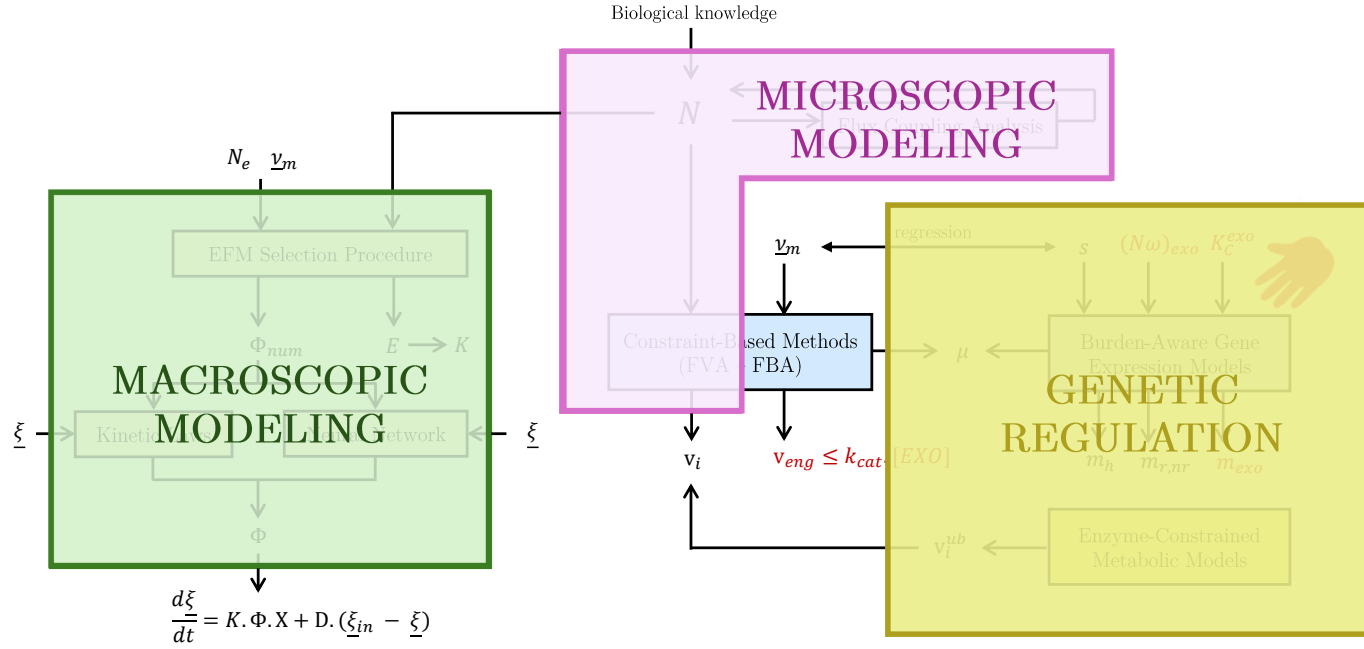


Figure 1: Schematic view of the thesis organization. Cellular culture processes are studied at three levels. (i) Macroscopic modeling derives macroscopic models from elementary flux modes (EFMs), assuming prior knowledge of the metabolic network through its stoichiometric matrix N . (ii) Microscopic modeling addresses the construction of the metabolic network itself and the quantification of intracellular fluxes using constraint-based approaches. (iii) Genetic regulation integrates catalytic constraints and gene expression models, further restricting the solution space by accounting for regulatory effects on flux values. This figure is presented at the beginning to guide the reader through the different modeling scales discussed in this doctoral thesis.

Chapter 1

Introduction

The study of cellular cultures has long stood as a cornerstone in biotechnology, biomedical research, and systems biology. From the production of bio-pharmaceuticals to advancing regenerative medicine, cellular processes support a multitude of critical applications. Yet, the inherent complexity of these systems, comprising multi-scale interactions among molecular, biochemical, and genetic components, necessitates the use of robust analytical tools to unravel and predict cellular behavior. Mathematical modeling of cellular processes provides a powerful framework for this endeavor, enabling researchers to systematically study, simulate and manipulate these systems. This introduction explores the importance of mathematical modeling, focusing on metabolic modeling as a key approach for advancing the understanding of cellular culture processes.

This chapter is structured into distinct sections: (i) **Motivations**, which emphasizes the critical importance of metabolic modeling in unraveling the complexities of cellular systems. This section highlights the multi-scale nature of cellular processes and explains how the different levels of description (ranging from macroscopic models to microscopic models integrating genetic aspects) are essential for a comprehensive understanding; (ii) **Contributions**, where the innovative aspects, key findings, and specific advancements made during this work are articulated; (iii) **Thesis Outline**, providing a clear road-map of the manuscript's organization; and (iv) **List of Publications**, which enumerates the scientific outputs and peer-reviewed publications that have resulted from this research endeavor, showcasing its academic impact.

1.1 Motivations

Mathematical modeling serves as an indispensable tool for bridging the gap between experimental data and theoretical understanding of biological systems. Cellular culture processes are influenced by numerous factors, including nutrient availability, environmental conditions, genetic regulation, and intracellular dynamics. These factors interact in non-linear and often unintuitive ways, making it challenging to predict outcomes without a formalized framework.

In the context of industrial and clinical applications, mathematical models enable the optimization of bioprocesses, such as maximizing the yield of therapeutic proteins and scaling up the production of cellular products. Beyond their utility in engineering, these models are critical for advancing fundamental biological research. They allow for hypothesis generation, identification of key regulatory pathways, and elucidation of mechanisms underlying cellular responses to perturbations. By integrating experimental observations with computational predictions, mathematical models provide a unified

platform for iterative cycles of experimentation and refinement, accelerating the pace of discovery.

Mathematical modeling is also important for the development of control strategies. In industrial bioprocesses, the ability to maintain optimal operating conditions, manage variability, and respond quickly to perturbations is essential. Models enable the design of control algorithms that can dynamically adjust inputs, such as nutrient feeds or environmental parameters, to achieve desired results. For instance, model predictive control systems rely on accurate mathematical representations to anticipate system behavior and implement corrective actions in real time. These capabilities are vital to ensure efficiency, reliability, and scalability in cellular culture processes.

Metabolic modeling is a specific branch of mathematical modeling dedicated to understanding and predicting the behavior of metabolic networks. Cellular metabolism encompasses the network of biochemical reactions that convert nutrients into energy, biomass, and other cellular components essential for growth. By formalizing these reactions in mathematical terms, metabolic models provide a structured approach to study the flow of metabolites through these pathways.

At its core, metabolic modeling involves the use of stoichiometric representations that describe the quantitative relationships between reactants and products in metabolic reactions. These models can be expanded to include kinetic parameters, such as reaction rates and enzyme affinities, to capture the dynamic behavior of the system. Depending on the level of detail required, models may range from simple macro-models to comprehensive kinetic models incorporating gene expression and regulatory mechanisms.

The importance of metabolic modeling lies in its versatility and applicability. In basic research, these models aim to elucidate fundamental principles of cellular function, such as resource allocation and energy efficiency. In an applied context, metabolic modeling is involved in designing optimized production strains, identifying therapeutic targets, and understanding disease mechanisms. For example, metabolic models can simulate how genetic modifications or environmental changes influence cellular output, providing a predictive framework for experimental design.

Among the various modeling approaches, metabolic modeling stands out as a particularly powerful tool for understanding cellular functions. Cellular metabolism, which encompasses the network of biochemical reactions that convert nutrients into energy and biomass, is central to the survival and proliferation of cells. Disruptions in metabolic pathways are implicated in numerous diseases, including cancer, diabetes, and metabolic disorders, underscoring the importance of studying these systems.

Metabolic models leverage stoichiometric and kinetic information about biochemical reactions to describe and predict the flow of metabolites through cellular pathways. At their core, these models capture the interplay between substrates, enzymes, and cofactors, enabling researchers to investigate questions ranging from resource allocation to metabolic flux distributions under varying conditions. Importantly, these models can be tailored to the needs of specific applications, whether focusing on high-level predictions for bioprocess optimization or detailed mechanistic insights for biomedical research.

A fundamental distinction in mathematical modeling is between structured and unstructured models. These two paradigms represent different approaches to capture the complexity of cellular systems.

Unstructured models consider the cell as a black box, focusing primarily on input-output relationships without explicitly accounting for the underlying of intracellular mechanisms. These models are typically macroscopic and describe the main properties such as biomass production, substrate consumption, and product formation. Thanks

to their relative simplicity, these models are particularly suitable for the prediction, the optimization, the state estimation and the control of bioprocesses [6]. However, although their simplicity makes them computationally efficient and practical for certain applications, unstructured models are limited in their ability to capture dynamic and regulatory features of cellular processes.

Structured models, on the other hand, incorporate detailed representations of intracellular mechanisms. These models include descriptions of metabolic pathways, regulatory networks, and even gene expression dynamics. Structured models are capable of capturing the intricate interplay between different cellular components, providing insight into how these interactions influence the overall behavior of the system. Thanks to the important development of "omics" technologies that grant access to very detailed metabolic information, these models can be very complex and contain hundreds of reactions. However, their complexity often requires significant computational resources and detailed experimental data for parametrization.

Both structured and unstructured models have their merits and limitations, and the choice of approach depends on the specific goals of the study. However, recent studies [7–15] have highlighted the strong link between both modeling approaches, allowing the combination of elements of both paradigms to balance computational efficiency with biological realism.

This thesis follows a structured progression toward increasingly detailed modeling, reflecting the multi-scale nature of cellular systems. The different modeling approaches are depicted in Fig. 1.1. First, macroscopic modeling is addressed. Macroscopic modeling, particularly when derived from elementary flux modes (EFMs), exemplifies a hybrid approach that bridges structured and unstructured paradigms. EFMs represent the simplest, non-decomposable pathways within a metabolic network that can operate at steady-state and connect extracellular substrates to products. By identifying these fundamental pathways, it becomes possible to construct macroscopic models that capture essential features of metabolic networks without delving into the full complexity of intracellular dynamics.

In this approach, the reaction schemes used to derive macroscopic models are deduced from EFMs, which effectively summarize the capabilities of the metabolic network. This methodology enables researchers to study the overall behavior of the system while preserving a connection to the underlying biochemical mechanisms. For example, macroscopic models derived from EFMs can be used to predict flux distributions, optimize metabolic performance, and analyze the impact of genetic or environmental changes on cellular function.

While macroscopic models abstract away many intracellular details, their foundation in EFMs ensures that they remain biologically meaningful. This balance makes them particularly useful for applications where computational efficiency and scalability are important, such as bioprocess optimization or high-throughput screening.

To address the limitations of macroscopic models, the next step involves microscopic modeling, which delves into intracellular dynamics and interactions. This level of detail captures the kinetics of individual reactions, the influence of enzyme concentrations, and the feedback mechanisms that regulate metabolic fluxes. By incorporating these factors, microscopic models provide a more nuanced understanding of cellular behavior, revealing how cells dynamically respond to changes in their environment or genetic makeup. Microscopic models are computationally demanding, but offer significant advantages, particularly when studying transient phenomena, such as the cellular response to stress or nutrient depletion. These models also provide insight into the coordination of metabolic pathways, enabling researchers to identify potential bottlenecks or vulnerabilities that

could be targeted for therapeutic or industrial applications.

The final stage of the modeling integrates genetic information, reflecting the ultimate level of detail in understanding cellular processes. Cellular metabolism is intrinsically related to gene expression, as enzymes that catalyze biochemical reactions are products of genetic transcription and translation. By incorporating genetic data, including regulatory networks and transcriptional dynamics, these models bridge the gap between genotype and phenotype.

This integration allows for the study of how genetic variations or mutations influence metabolic function, providing a framework for personalized medicine or synthetic biology applications. Moreover, models that incorporate genetic information are essential for understanding metabolic regulation, such as how cells prioritize certain pathways under specific conditions or adapt to long-term changes in their environment.

1.2 Contributions

This thesis presents several key contributions to the field of metabolic modeling, focusing on innovative methodologies and their applications to address critical challenges in cellular system analysis. These contributions are outlined as follows:

Macroscopic Models Derived from Elementary Flux Modes

1. Handling Large Sets of Elementary Flux Modes (EFMs)

A central challenge in metabolic modeling is the exponential growth in the number of EFMs as the size and complexity of the metabolic network increase. For large networks, enumerating all EFMs becomes computationally infeasible. To address this issue, a novel method has been developed to reduce the number of EFMs and select the most informative ones from a potentially large or incomplete set. This method ensures that the selected EFMs capture the essential metabolic features without requiring a complete enumeration of all possible modes.

2. Hybrid Modeling Approach

This thesis introduces a hybrid modeling framework that integrates EFMs with advanced machine learning techniques. Specifically, reaction schemes are derived from EFMs, while reaction rates are modeled using neural networks. Neural networks provide the flexibility to capture the non-linear and complex relationships inherent in metabolic reaction rates, offering a significant improvement over traditional modeling approaches.

3. General Applicability

The developed methods for EFM selection and reaction rate modeling are generalizable and applicable to various biological systems, including scenarios where the total number of EFMs cannot be fully enumerated. This approach ensures computational efficiency while maintaining biological accuracy.

Microscopic Models of Intracellular Dynamics

1. Constructive Method for Developing Metabolic Networks

A step-by-step method is proposed to elaborate detailed metabolic networks, focusing on intracellular dynamics and interactions. This method provides a more comprehensive representation of cellular behavior.

2. Development of Mid-Sized Metabolic Models

This thesis develops metabolic networks and models of intermediate size, which strike a balance between overly simplistic and excessively detailed models. This level of granularity is important for capturing sufficient biological detail without incurring prohibitive computational costs, making the models both insightful and practical for applications.

3. Modeling of Proton Motive Force and ATP Synthesis

The developed models incorporate the proton motive force (PMF), a critical component of cellular energy metabolism related to ATP synthesis, which is a molecule that serves as the primary energy currency of cells. Most existing studies overlook this aspect, yet it is essential for understanding energy transduction in cells. Including PMF in the models provides a more accurate representation of cellular energy and highlights its role in metabolic regulation and efficiency.

4. Application to Photosynthetic Microorganisms

The utility of microscopic models is demonstrated through their application to photosynthetic microorganisms, particularly beyond the pseudo-steady-state assumption, considering some internal metabolites accumulate over time. These models provide insights into the coordination of metabolic pathways and how cells respond dynamically to environmental changes, offering valuable implications for bioengineering and renewable energy production.

Integration of Genetic Aspects in Microscopic Models

1. Genetic-Level Integration

This thesis advances the integration of genetic information into metabolic modeling by incorporating gene regulatory networks and transcriptional dynamics. This approach bridges the gap between genotype and phenotype, enabling a more detailed understanding of the genetic control of metabolism.

2. Improvement of Flux Predictions

By connecting metabolic and genetic information, the developed method improves the accuracy of flux predictions in metabolic analysis. This integration provides a more holistic understanding of cellular behavior, accounting for both metabolic and regulatory influences.

3. Optimization of Gene Expression Parameters

In the context of synthetic biology, a novel approach is proposed for the optimal selection of gene expression parameters, particularly for exogenous genes. This approach ensures efficient expression of these genes to optimize the production of target products, addressing a key challenge in synthetic biology applications.

4. Application to Synthetic Biology

The methods developed are applied to synthetic biology, focusing on the model organism *E. coli*. This application highlights how genetic-level modeling can inform the design of engineered metabolic systems, advancing the field of synthetic biology.

The contributions of the thesis span multiple levels of modeling, from macroscopic to microscopic, and genetic-level integration. The methodologies developed address key challenges in metabolic modeling, such as handling large sets of EFMs, incorporating complex reaction dynamics, and integrating genetic information. These contributions provide both theoretical advancements and practical tools for the analysis and engineering of cellular systems with applications ranging from bioengineering to synthetic biology and renewable energy.

1.3 Thesis Outline

By progressing from macroscopic to microscopic modeling, and finally to genetic-level integration, this thesis aims to provide a comprehensive framework for understanding and modeling cellular culture processes. In subsequent chapters, each level of modeling will be explored in detail, highlighting their contributions, challenges, and applications. Together, these approaches form a cohesive narrative that underscores the power of mathematical modeling in unraveling the complexities of cellular systems.

The thesis is structured into several chapters to explore the different modeling approaches and their applications. The manuscript is divided as follows:

- **Chapter 2: Preliminaries** - This chapter reviews the fundamentals of cellular metabolism, the basics of metabolic modeling and metabolic flux analysis, and provides reminders of genetics and other foundational concepts such as neural networks and parameter estimation. This chapter provides the foundational concepts and background necessary for the discussions that follow.
- **Chapter 3: Macroscopic Modeling** - Here, macroscopic models derived from elementary flux modes are explored. The chapter focuses on methods for selecting the best EFMs to infer reaction schemes and how to model reaction rates using traditional modeling approaches or neural networks. Several applications of these methods to animal cells are discussed. Chapters 3 through 5 share a unified structure designed to provide clarity and coherence. Each chapter begins with a comprehensive background section that contextualizes the research contributions within existing studies. This is followed by a detailed presentation of the research itself, highlighting novel methods, insights, and applications. Finally, each chapter concludes with a discussion of prospects and potential avenues for improvement, paving the way for future research.
- **Chapter 4: Microscopic Modeling** - This chapter presents a constructive methodology for developing detailed metabolic models, emphasizing the study of intracellular interactions and their dynamics. The originality of the procedure lies in the sequential and iterative use of biological knowledge and mathematical methods to refine the definition of the metabolic network. In this chapter, two primary focuses are addressed: the appropriate sizing of the metabolic network, striking a balance between model simplicity and detail, and the incorporation of energy-related aspects such as the proton motive force. First, the approach is applied to photosynthetic microorganisms under constant light regimes to apply the classical tools of metabolic flux analysis. Then, to further validate the developed methodology, this chapter extends its application beyond the pseudo-steady-state assumption, considering cases where internal metabolites accumulate over time. This approach enhances the ability of the model to capture dynamic cellular behavior and transient metabolic states. Validation is performed in a different biological context, specifically in unicellular microalgae, demonstrating the versatility and robustness of the proposed modeling framework.
- **Chapter 5: Genetic Aspects in Microscopic Modeling** - This chapter explores the integration of genetic information into microscopic models in the context of synthetic biology. The combination of metabolic and genetic models facilitates a significant reduction in the solution space for flux distributions when performing metabolic flux analysis, enhancing the precision of predictions. Furthermore, this integrated approach enables the adjustment and optimization of gene expression parameters, particularly those related to exogenous genes, to maximize the

production rates of desired products. The methodology is applied to the model organism *E. coli*, demonstrating the potential of this framework to advance synthetic biology applications.

- **Chapter 6: Conclusion and Prospects** - This chapter synthesizes the insights gained throughout the thesis and outlines potential future directions for research.
- **Appendices** - Additional materials, detailed derivations, and additional information are included to support the main text.

Together, these chapters form a cohesive framework for advancing the understanding of cellular culture processes through progressively detailed modeling approaches, providing both theoretical insights and practical applications.

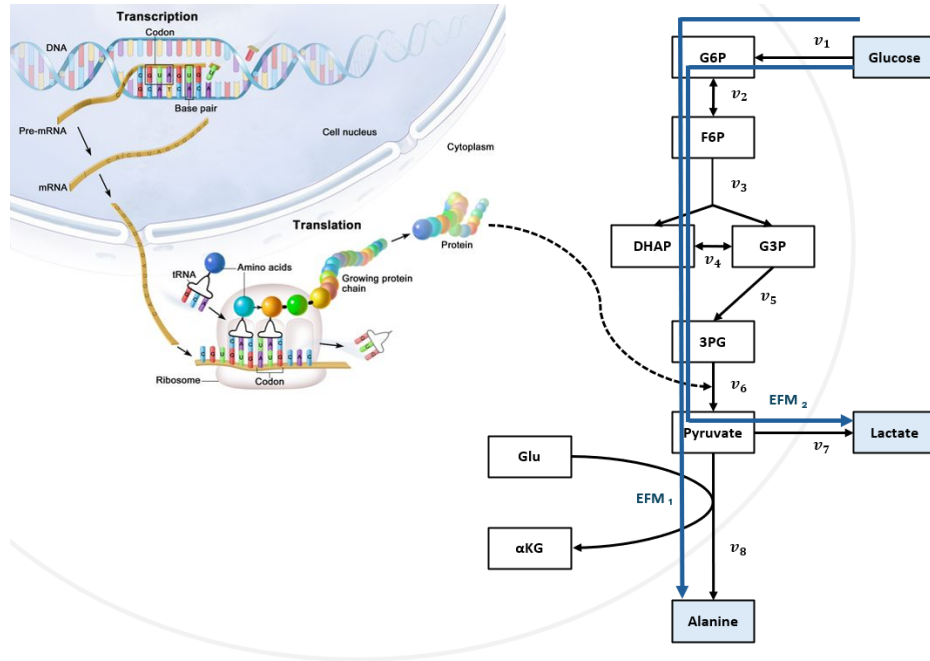


Figure 1.1: The different modeling approaches: (i) macroscopic modeling derived from elementary flux modes connecting extracellular substrates (e.g., glucose) to extracellular products (e.g., lactate and alanine), (ii) microscopic modeling emphasizing the study of intracellular fluxes v_i , and (iii) microscopic modeling integrating genetic aspects, as enzymes that catalyze biochemical reactions are products of genetic transcription and translation.

1.4 List of Publications

Journal Papers in Special Issues

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. A Systematic Elementary Flux Mode Selection Procedure for Deriving Macroscopic Bioreaction Models from Metabolic Networks. *Journal of Process Control* **2022**, *118*, 170 - 184.
- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Hybrid Dynamic Models of Bioprocesses Based on Elementary Flux Modes and Multilayer Perceptrons. *Processes*

2022, 10, 2084.

Journal Paper

- Maton, M.; Leroy, B.; Vande Wouwer, A. A Constructive Bottom-Up Approach for the Elaboration of Metabolic Networks: Case Study of Photosynthetic Cyanobacteria *Arthrospira Spirulina Platensis* PCC 8005. *Biochemical Engineering Journal* **2025**, 221, 109770.

Conference Paper as Keynote Paper

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Selection of a Minimal Suboptimal Set of EFMs for Dynamic Metabolic Modelling. *IFAC-PapersOnLine*. 11th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2021: Venice, Italy, 13-16 June 2021. **2021**, 54, 667 - 672.

Conference Papers as Regular Papers

- Maton, M.; Santos-Navarro, F.N.; Pico, J.; Bogaerts, Ph.; Vande Wouwer, A. Towards Constraint-Based Burden-Aware Models for Metabolic Engineering. *IFAC-PapersOnLine*. 12th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2024: Toronto, Canada, 14-17 July 2024. **2024**, 58 (14), 247 - 252.
- Maton, M.; Vande Wouwer, A. Mathematical Modeling of Photosynthetic Eukaryotic Microorganisms Using Metabolic Networks. *IFAC-PapersOnLine*. 11th Vienna International Conference on Mathematical Modelling MATHMOD 2025: Vienna, Austria, 19-21 February 2025. **2025**, 59 (1), 397 - 402.
- Maton, M.; Leroy, B. ; Vande Wouwer, A. Metabolic Modeling of *Arthrospira* Sp. PCC 8005 - Network Definition and Experimental Validation. *IFAC-PapersOnLine*. 14th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems DYCOPS 2025: Bratislava, Slovakia, 16-19 June 2025. **2025**, 59 (6), 391 - 396.

Conferences with Abstract

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Selection of Elementary Flux Modes for the Development of Reduced Macroscopic Models. *40th Benelux Meeting on Systems and Control*, **2021**.
- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. A Modular Elementary Flux Mode Reduction Procedure for Dynamic Metabolic Modelling. *41st Benelux Meeting on Systems and Control*, **2022**.

Chapter 2

Preliminaries

The present chapter lays the theoretical foundations necessary for the subsequent discussions in this thesis. Because this work involves cellular metabolism, bioprocess engineering, and computational modeling, a solid understanding of the fundamental principles is essential. This chapter serves as a compendium of essential concepts, providing a structured review of key aspects of cellular metabolism, cell culture processes, metabolic modeling, and computational tools relevant to the study.

To establish a proper framework, an initial section will introduce cellular foundations, covering the structural organization of cells and the classification of different types of organisms. This will include a brief discussion of prokaryotic and eukaryotic cells, their key morphological and functional differences, and their relevance to metabolic studies. Understanding these fundamental aspects is crucial, as cellular architecture and evolutionary lineage play a significant role in shaping metabolic capabilities and regulatory mechanisms.

Following this, a general overview of cellular metabolism will be presented, encompassing the core biochemical pathways shared among most organisms. While these fundamental metabolic routes provide a universal framework, the specific metabolic traits of the strain under investigation will be explored in greater detail in dedicated chapters later. This discussion will elucidate the essential mechanisms underlying energy production, precursor biosynthesis, and cellular maintenance, providing a metabolic context for subsequent analyses.

In parallel, this chapter will outline the fundamental principles of cell culture processes, including bioreactor configurations and operational strategies. The interplay between bioreactor design, environmental control, and cellular physiology is crucial in biotechnological applications, influencing both experimental reproducibility and industrial scalability. A detailed examination of these elements will provide a contextual foundation for understanding metabolic responses in controlled culture environments.

A primer on metabolic modeling and metabolic flux analysis (MFA) will be introduced, bridging the gap between experimental data and quantitative insights into cellular function. The mathematical formalisms employed in these methodologies provide a framework for analyzing metabolic networks, enabling the assessment of intracellular fluxes and the prediction of cellular responses under varying conditions. By establishing the computational base for metabolic analysis, this section will set the scene for later chapters that employ these techniques to elucidate strain-specific metabolic properties.

Furthermore, key principles in genetics will be reviewed as genetic modifications and regulatory mechanisms play a central role in metabolic engineering and synthetic biology. The current chapter will also provide a concise introduction to artificial neural networks and other computational tools that have gained prominence in biological data analysis and predictive modeling, along with key information on parameter estimation techniques, which are fundamental for mathematical modeling.

By consolidating these theoretical elements, this chapter establishes the necessary groundwork for the scientific discussions and methodological developments that follow. The interdisciplinary nature of this thesis requires a comprehensive perspective, integrating biochemical, biotechnological, and computational principles to drive a deeper understanding of metabolic function and optimization.

2.1 Cellular Foundations

The smallest cellular organisms consist of single cells and are microscopic, while the more complex multicellular organisms consist of a variety of cell types, which vary in size, shape, and specialized function. Despite these structural and functional differences, all cells share certain fundamental properties, which can be seen at the biochemical level [1].

Cells exhibit common structural features, as depicted in Fig. 2.1. The plasma membrane defines the cell boundary, separating its internal content from the external environment. This membrane is primarily composed of lipids and proteins, forming a thin, flexible, and hydrophobic barrier that regulates the exchange of substances between the cell and its surroundings. However, transport proteins in the plasma membrane allow the passage of certain ions and molecules; receptor proteins transmit signals into the cell to respond to external stimuli; and membrane-associated enzymes participate in some metabolic pathways.

Within the plasma membrane, the cytoplasm is composed of an aqueous medium known as the cytosol, which contains a variety of suspended particles with specific functions. The cytosol is a highly concentrated solution filled with enzymes and RNA molecules that encode them, as well as the building blocks required for macromolecule synthesis, such as amino acids and nucleotides. In addition, it contains hundreds of small organic molecules called metabolites, i.e., intermediates in biosynthetic and degradative pathways; coenzymes, i.e., compounds essential to many enzyme-catalyzed reactions; ribosomes, i.e., the molecular machines responsible for protein synthesis; and proteasomes, which degrade proteins that are no longer needed by the cell.

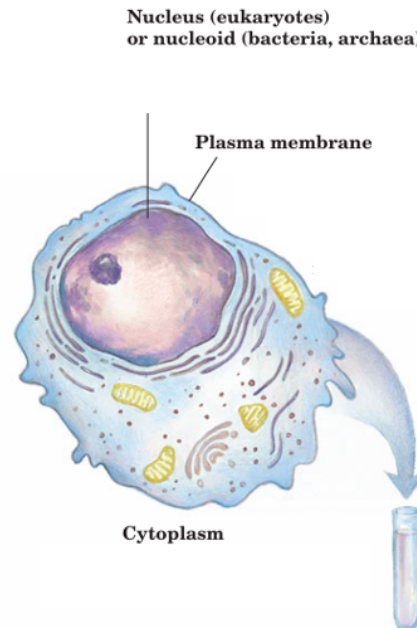


Figure 2.1: Features of living cells, from [1]. Nucleus or nucleoid contains genetic material-DNA and associated proteins, plasma membrane is a tough, flexible lipid bi-layer, and cytoplasm contains an aqueous solution and suspended particles and organelles.

All cells have either a nucleus or a nucleoid, in which the genome, i.e., the complete set of genes, composed of DNA, is stored and replicated. In bacteria and archaea, this genetic material is concentrated in a region called the nucleoid, which is not enclosed by a membrane. In contrast, eukaryotic cells possess a nucleus, where the genetic material is housed within a double membrane known as the nuclear envelope. Organisms with a defined nuclear compartment belong to the large group of *Eukarya*. Microorganisms lacking a nuclear envelope, previously classified as prokaryotes, are now forming two very distinct groups: *Bacteria* and *Archaea*. The latter differ in their cellular membrane composition, cell wall structure, ribosomal RNA sequences, and metabolic pathways. Bacteria inhabit soils, surface waters, and tissues of other living or decaying organisms. Archaea inhabit extreme environments such as salt lakes, hot springs, highly acidic bogs, and ocean depths. The classification of the different types of organisms is given in Fig. 2.2.

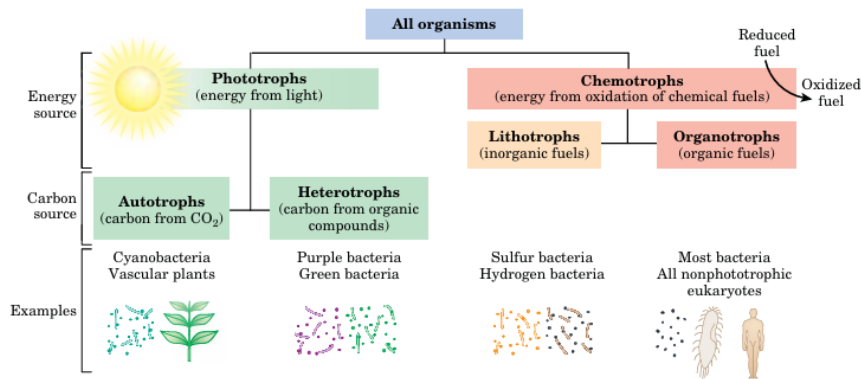


Figure 2.2: Classification of organisms according to their source of energy (sunlight or oxidizable chemical compounds) and their source of carbon for the synthesis of cellular material, from [1].

In order to establish a fundamental understanding of the structure of cells, a concise explanation of key organelles and cellular components will be provided. These structures play essential roles in cellular metabolism, energy production, and biosynthetic processes, forming the basis for more advanced discussions in the subsequent chapters. Fig. 2.3 depicts the eukaryotic cell and key cellular components:

- **Mitochondria:** Double-membrane-bound organelles responsible for ATP synthesis through oxidative phosphorylation. Mitochondria play a central role in cellular respiration, energy metabolism, and apoptotic signaling;
- **Ribosomes:** Macromolecular complexes, composed of ribosomal RNA and proteins, serving as the sites of protein synthesis. They facilitate the translation of messenger RNA (mRNA) into functional polypeptides, playing a crucial role in gene expression;
- **Peroxisomes:** Small membrane-bound organelles containing oxidative enzymes involved in lipid metabolism. They are critical for breaking down long-chain fatty acids via β -oxidation;
- **Glyoxysomes:** Specialized form of peroxisomes found in plant cells involved in the glyoxylate cycle, a metabolic pathway that enables the conversion of stored lipids into carbohydrates during seed germination (crucial for energy production in young plants before they establish photosynthetic capability);
- **Chloroplasts:** Double-membrane-bound organelles responsible for photosynthesis. They contain their own DNA and ribosomes, similar to mitochondria;
- **Thylakoids:** Membrane-bound compartments that house the photosynthetic machinery. Thylakoid membranes contain pigments, such as chlorophyll, which capture light energy to drive the light-dependent reactions of photosynthesis.

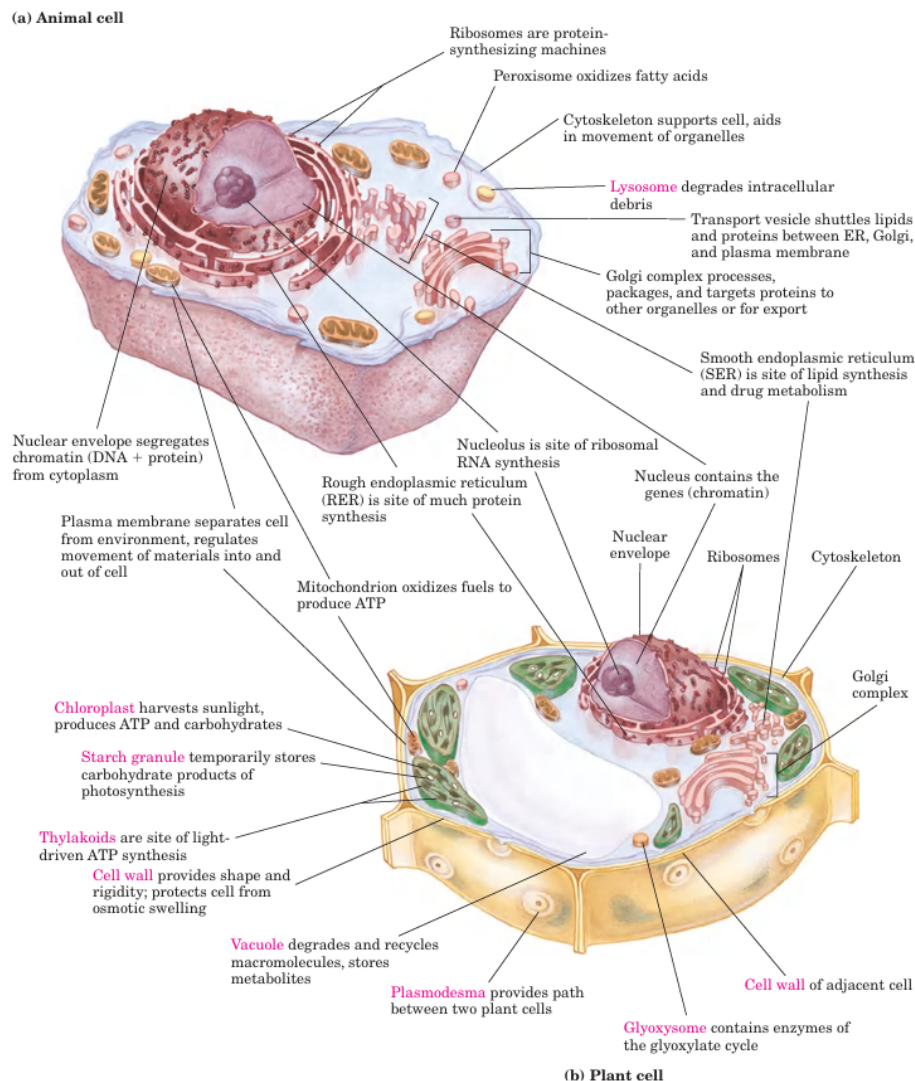


Figure 2.3: Eukaryotic cell structure. Schematic illustrations of two major types of eukaryotic cell: (a) a representative animal cell and (b) a representative plant cell, from [1].

Fig. 2.4 shows the structure of prokaryotic cells and the key cellular components. A prokaryote has a protective outer membrane and an inner plasma membrane that encloses the cytoplasm and the nucleoid. The plasma membrane and the layers outside it constitute the cell envelope. The nucleoid contains a single circular molecule of DNA, and the cytoplasm contains one or more smaller circular segments of DNA called plasmids.

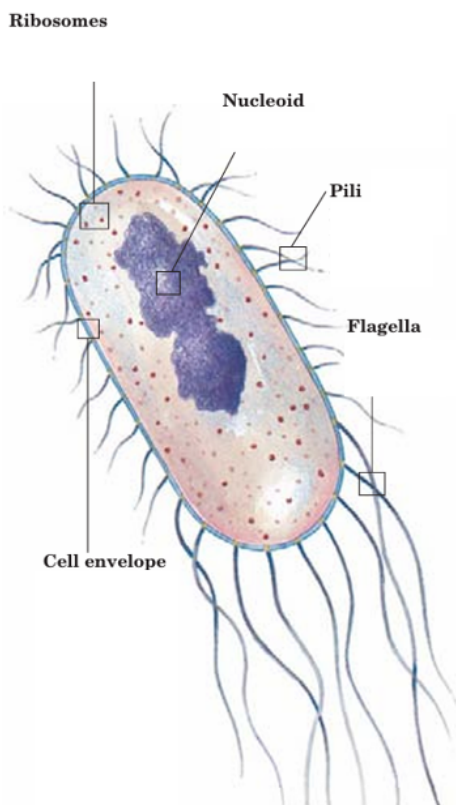


Figure 2.4: Prokaryotic cell structure. Schematic illustration of bacterial cells, from [1].

2.2 Cellular Metabolism

Cell metabolism is defined as the set of all chemical transformations that occur within the cell or an organism, through a series of enzyme-catalyzed reactions. These reactions constitute metabolic pathways, which consist of sequential enzymatic steps that transform substrates into products, facilitating the conversion of energy and the synthesis of essential biomolecules. The small molecules involved in these pathways are known as metabolites and include intermediates, precursors, and end products. Among these transformations, reactions that convert organic nutrient molecules (carbohydrates, fats, and proteins) into smaller molecules (lactic acid, CO_2 , NH_3) form what is called catabolism. In contrast, anabolism, also called biosynthesis, is the set of reactions that synthesize larger and more complex molecules (lipids, polysaccharides, proteins, and nucleic acids) from small and simple precursors [1, 16].

These two metabolic phases are interconnected. As a matter of fact, catabolic pathways release energy in the form of adenosine triphosphate molecules (ATP) and reduced electron carriers (NADH, NADPH, and FADH_2) used in the anabolic pathways to achieve the biosynthesis reactions. The rest of the energy released by catabolic pathways is lost as heat. The interconnection between the metabolic routes is illustrated in Fig. 2.5.

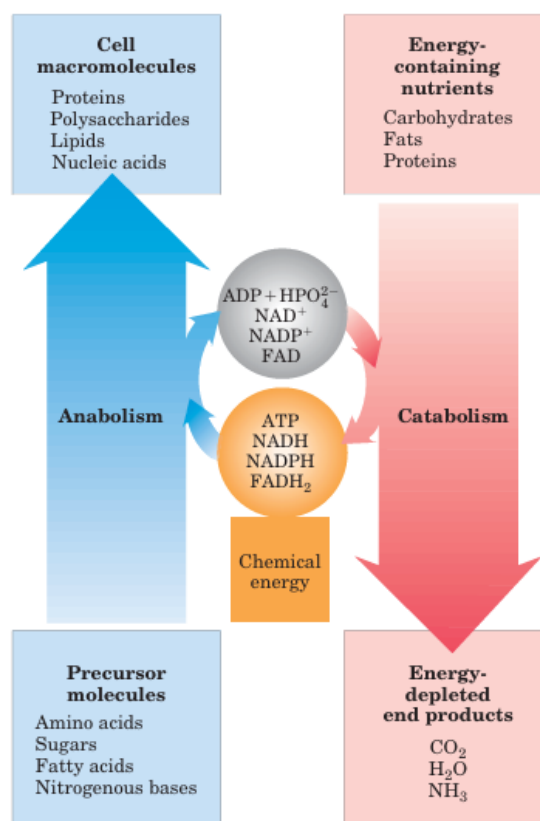


Figure 2.5: Relationships between catabolic and anabolic pathways. Catabolic pathways deliver chemical energy in the form of ATP, NADH, NADPH and FADH₂ used in anabolic pathways, from [1].

Cellular metabolism comprises a highly regulated network of biochemical reactions that enable the conversion of nutrients into energy and essential molecules. Despite the vast diversity of living organisms, certain metabolic pathways are common, reflecting their fundamental roles in cellular function. This section provides an overview of key metabolic pathways, including glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway (PPP), the lipid metabolism and the amino acid (AA) biosynthesis. The other metabolic pathways specific to the strain under study will be explored in detail in later chapters.

2.2.1 Glycolysis and Pyruvate Fate

Glycolysis, also known as the Embden-Meyerhof-Parnas pathway, occupies a central position in the metabolism of most organisms. One glucose molecule is degraded through a series of enzyme-driven reactions, producing two molecules of pyruvate, three-carbon compounds. During glycolysis, some of the free energy released from glucose is captured and stored in the form of ATP and NADH. The ten-step enzymatic sequence takes place in the cytoplasm and serves as a primary energy source under both aerobic and anaerobic conditions.

It is worth noting that the pyruvate formed by glycolysis can be further metabolized

via different catabolic routes depending on the type of organisms, as depicted in Fig. 2.6. Its utilization is primarily determined by oxygen availability and cellular energy demands:

- **Aerobic Respiration:** Under aerobic conditions, pyruvate is transported into mitochondria (or analogous structures in prokaryotes), where it is oxidized to form the acetyl group of acetyl-coenzyme A. Acetyl-CoA then enters the tricarboxylic acid cycle, leading to the production of NADH and FADH_2 , which subsequently donate electrons to the electron transport chain (ETC), driving ATP synthesis via oxidative phosphorylation;
- **Fermentation:** In anaerobic conditions, pyruvate is metabolized through fermentation to regenerate NAD^+ , which is required for glycolysis to continue. In lactic acid fermentation, pyruvate is reduced to lactate. It is a process observed in many animal cells, including muscle cells under low oxygen conditions. In alcoholic fermentation, common in yeast and some bacteria, pyruvate is first converted into acetaldehyde and then reduced to ethanol, releasing CO_2 in the process;
- **Anabolic fates:** The oxidation of pyruvate is an important catabolic process. However, pyruvate has anabolic fates as well. For instance, it can provide the carbon skeleton for the synthesis of several amino acids, including alanine, or for the synthesis of fatty acids.

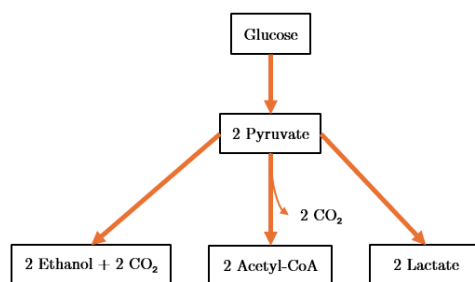


Figure 2.6: Catabolic fates of pyruvate formed in glycolysis.

2.2.2 Pentose Phosphate Pathway

As mentioned previously, the main catabolic pathway of glucose-6-phosphate (G6P) – an intermediate in glycolysis – is its conversion to pyruvate, which is largely oxidized in the citric acid cycle, ultimately contributing to ATP production. However, glucose-6-phosphate also serves other catabolic roles that lead to the formation of specialized cellular products. Notably, it can be oxidized to pentose phosphates via the pentose phosphate pathway (PPP) and forms the ribose-5-phosphate (R5P), essential for synthesizing RNA, DNA, and coenzymes such as ATP, NADH, FADH_2 , and coenzyme A. Another key output of this pathway is NADPH, which is vital for reductive biosynthesis and for combating the harmful effects of oxygen radicals.

The pentose phosphate pathway occurs in the cytoplasm of the cell. During the oxidative phase, NADP^+ is the electron acceptor and reacts with glucose-6-phosphate to produce ribose-5-phosphate, a precursor for nucleotide synthesis, and NADPH, a reductant for biosynthetic reactions. The reactions of this oxidative phase are essentially

irreversible in the cell. During the non-oxidative phase, the pentose phosphates produced during the oxidative phase are recycled into glucose-6-phosphate through a series of rearrangements of the carbon skeletons. The reactions within the non-oxidative segment of the pentose phosphate pathway are easily reversible, allowing the conversion of hexose phosphates back to pentose phosphates. The general scheme of the pentose phosphate pathway is depicted in Fig. 2.7.

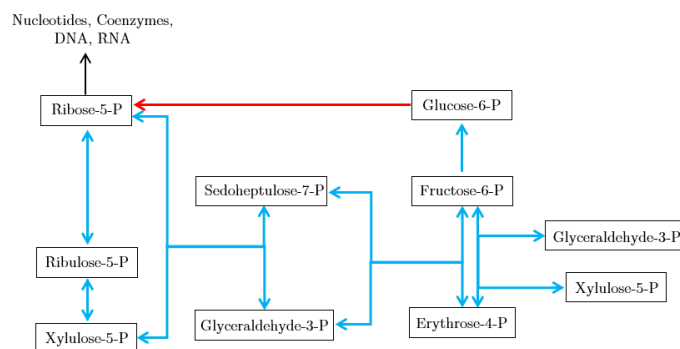


Figure 2.7: Pentose Phosphate Pathway: oxidative phase in red and chemical reactions related to non-oxidative phase in blue.

2.2.3 Tricarboxylic Acid Cycle

Under aerobic conditions, glycolysis is the first stage in the complete oxidation of glucose. Henceforth, the pyruvate produced by glycolysis is further oxidized to H_2O and CO_2 during the so-called process of cellular respiration. It consists of three primary stages. Initially, organic fuel molecules are oxidized in the form of the acetyl group in acetyl-coenzyme A. In the second stage, these acetyl groups are introduced in the citric acid cycle, enabling their oxidation to CO_2 . The energy released during this process is captured in the form of electron carriers NADH and $FADH_2$. Finally, the third stage involves the oxidation of these reduced coenzymes during a process called oxidative phosphorylation, during which they transfer electrons to oxygen through the electron transport chain, driving proton translocation and enabling ATP synthesis.

The tricarboxylic acid (TCA) cycle also called the citric acid cycle or the Krebs cycle, is essential to energy-yielding metabolism, but its role is not limited to energy conservation. Indeed, most intermediates of the cycle serve as precursors for a wide variety of products, especially in the formation of amino acids from α -ketoglutarate [17]. To replace intermediates that are used for various metabolic functions, cells utilize anaplerotic reactions. As a matter of fact, the operation of the citric acid cycle depends on the anaplerotic role of the phosphoenolpyruvate carboxylase (among others), which is essential for supplying the oxaloacetate needed for the synthesis of citrate and α -ketoglutarate. Fig. 2.8 illustrates the main reactions of the tricarboxylic acid cycle, its role in anabolism, and some anaplerotic reactions that replenish depleted cycle intermediates. Also, it is worth noting that the TCA cycle might be slightly different for some bacteria. Reactions of the TCA cycle take place in the mitochondrial matrix in eukaryotic cells or occur in the cytoplasm in prokaryotic cells, which lack mitochondria.

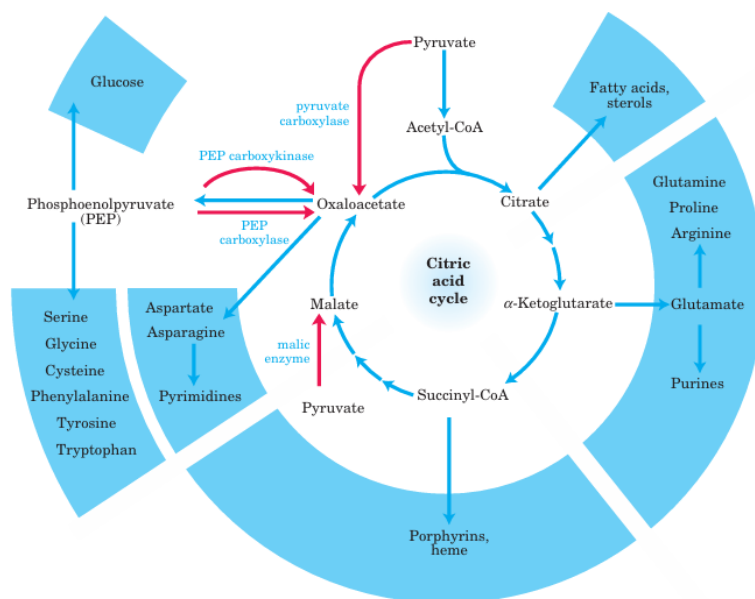


Figure 2.8: Tricarboxylic acid (TCA) cycle and its role in anabolism. Red arrows represent anaplerotic reactions that replenish depleted cycle intermediates, from [1].

2.2.4 Lipid Metabolism

Lipids serve as key energy reserves, structural components of membranes, and signaling molecules. Lipid metabolism encompasses both fatty acid metabolism and the synthesis and degradation of complex lipids such as phospholipids and sterols.

In eukaryotes, fatty acid oxidation, also known as β -oxidation, occurs in mitochondria and peroxisomes, breaking down long-chain fatty acids into acetyl-CoA, which enters the TCA cycle for further oxidation. Conversely, fatty acid biosynthesis occurs in the cytoplasm and involves the conversion of acetyl-CoA into long-chain fatty acids, which are utilized for membrane formation and signaling molecule synthesis. In prokaryotes, lipid metabolism occurs in the cytoplasm and in the plasma membrane.

2.2.5 Amino Acid Metabolism

Amino acids (AA) are the building blocks of proteins and serve as precursors for various biomolecules. While some amino acids can be synthesized endogenously through transamination and related pathways, others, called essential amino acids, cannot be produced by the cell and must be obtained exogenously, by supplying nutrients. Amino acid degradation contributes to cellular metabolism by producing intermediates that enter glycolysis, the TCA cycle, or other biosynthetic pathways, enabling their integration into broader metabolic networks.

Although amino acid metabolism occurs entirely in the cytoplasm in prokaryotes, the metabolism of amino acids is compartmentalized between the cytoplasm and mitochondria in eukaryotic cells.

2.2.6 Nucleotide Metabolism

Nucleotides are essential for genetic information storage, energy transfer (e.g., ATP, GTP), and cellular signaling. Two types of pathways lead to nucleotides: the *de novo* pathways and the salvage pathways. De novo synthesis begins with their metabolic precursors: amino acids, ribose-5-phosphate and CO_2 . Salvage pathways recycle the free bases and nucleotides released from nucleic acid breakdown. These processes ensure a continuous supply of nucleotides for DNA replication, RNA transcription, and other cellular functions.

Since prokaryotic cells lack membrane-bound organelles, all reactions related to nucleotide metabolism take place in the cytoplasm. In eukaryotes, nucleotide metabolism is primarily cytoplasmic, but mitochondria contribute to pyrimidine biosynthesis and nucleotide pool regulation.

2.2.7 Protein Synthesis

Cells are composed of a vast array of proteins, which are essential macromolecules formed by chains of amino acids. These proteins have diverse functions, critical to cellular processes. For example, enzymes, i.e., a major class of proteins, act as biological catalysts that accelerate metabolic reactions, thereby regulating cellular function and efficiency. Other proteins play structural roles, facilitate intracellular signaling, or serve as transport proteins, enabling the movement of molecules across cellular membranes.

Despite the functional diversity of the proteins, they are all synthesized from the same set of 20 standard amino acids. The specific function of a protein is determined by its amino acid sequence, which dictates its three-dimensional structure and biochemical properties. In addition, protein synthesis is fundamental to biomass formation. Theoretically, an average protein synthetic reaction should be assessed for each cell strain under specific conditions. Therefore, the average amino acid composition of the studied strain is represented by the stoichiometric coefficients in the protein formation reaction. Such amino acid composition data, which vary across different organisms and cell types, have been documented in various studies, including [18], for multiple cells. This information is also essential for accurately modeling biomass synthesis.

2.3 Bioreactors and Operating Modes

The cultivation of mammalian, bacterial, or yeast cells in controlled environments requires precise control over physicochemical parameters to optimize cell growth and product formation. Bioreactors serve as essential vessels for these processes, supporting cell proliferation and biosynthesis through various operational strategies. The choice of bioreactor configuration significantly influences the efficiency, scalability, and productivity of the process.

Among the various bioreactor designs, the stirred tank bioreactor (STR) is the most widely used in the biotechnology industry. As illustrated in Fig. 2.9, a stirred tank bioreactor consists of a cylindrical vessel equipped with an impeller for mixing, one or more feeding pumps, and sensors to perform measurements. The agitation system ensures uniform distribution of nutrients, oxygen, and pH throughout the culture, facilitating optimal cell growth and product formation.

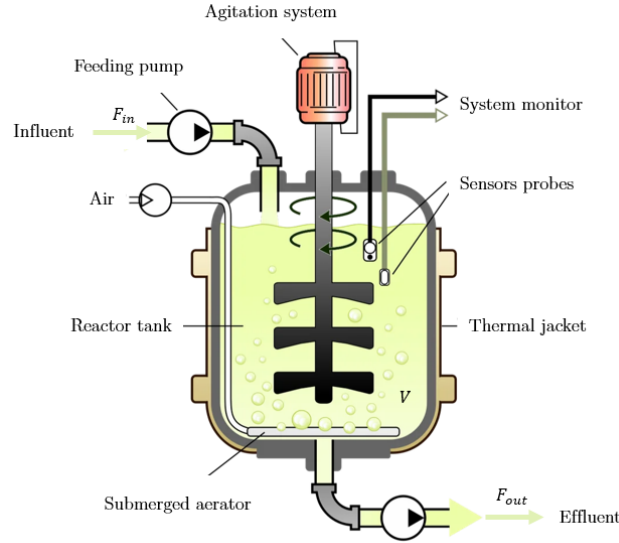


Figure 2.9: Schematic representation of a stirred tank bioreactor.

Bioreactor configurations can be classified on the basis of operational modes, including batch, fed-batch, continuous, and perfusion processes. Each operation mode presents advantages and challenges related to productivity, process control, and scalability. The different operating modes are shown in Fig. 2.10.

2.3.1 Batch Mode

Batch culture is a closed system in which cells grow in a fixed volume of culture medium without additional nutrient supplementation after inoculation. No fresh culture medium is added or taken out (V is constant, $\forall t > t_0$, $F_{in}(t) = F_{out}(t) = 0$). Although simple to implement and monitor, batch culture suffers from limited productivity due to nutrient depletion and accumulation of inhibitory metabolites.

2.3.2 Fed-Batch Mode

In fed-batch mode, fresh medium is added and the culture volume is not constant over time ($\forall t > t_0$, $F_{in}(t) \geq 0$, $F_{out}(t) = 0$ and $V(t) \geq V(t_0)$). This operating mode prolongs the duration of the culture by preventing substrate depletion, thereby improving cell densities and productivity. However, because there is no removal of the medium from the system, inhibitory metabolites accumulate over time, which can significantly affect product quality.

2.3.3 Continuous Mode

Continuous culture maintains cells in a steady-state environment by continuously adding fresh medium while simultaneously removing an equivalent volume of spent medium (V is constant, $\forall t > t_0$, $F_{in}(t) = F_{out}(t) \geq 0$). This operating mode prevents both substrate depletion and inhibitory metabolite accumulation, enabling prolonged operation and enhanced productivity while maintaining cell viability over extended periods.

Continuous cultures are typically operated in a chemostat or turbidostat configuration, where nutrient feed rates or optical density regulate cell growth, respectively.

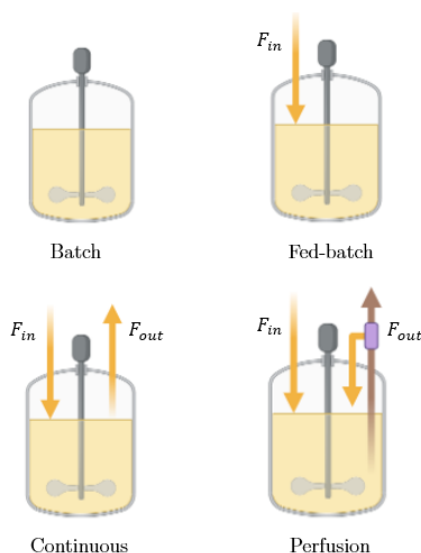


Figure 2.10: The different operating modes of bioreactors: batch, fed-batch, continuous, and perfusion.

2.3.4 Perfusion Mode

As detailed in Fig. 2.11, perfusion culture is a specialized continuous process in which cells are retained within the bioreactor, while the spent medium is removed and replenished with fresh nutrients. Cell retention can be achieved through filtration, centrifugation, or membrane-based separation methods. Perfusion processes enable high cell densities, increased product titers, and improved product quality by maintaining optimal culture conditions.

2.4 Basics of Metabolic Analysis

Metabolic analysis is a critical aspect in understanding cellular function, allowing researchers to study biochemical pathways, predict metabolic capabilities, and optimize biotechnological applications. This section provides the fundamental mathematical and conceptual foundations necessary for metabolic network analysis. More advanced notions will be explored in subsequent chapters, where deeper theoretical and computational aspects will be introduced.

2.4.1 Theoretical Background

Metabolic modeling aims to mathematically model the metabolism of a cell and to understand how biochemical reactions interact to sustain life. Metabolism is a complex network of reactions that transform molecules to generate energy, build cellular structures, and regulate biochemical functions. In that respect, it is important to make a distinction between the types of molecules involved:

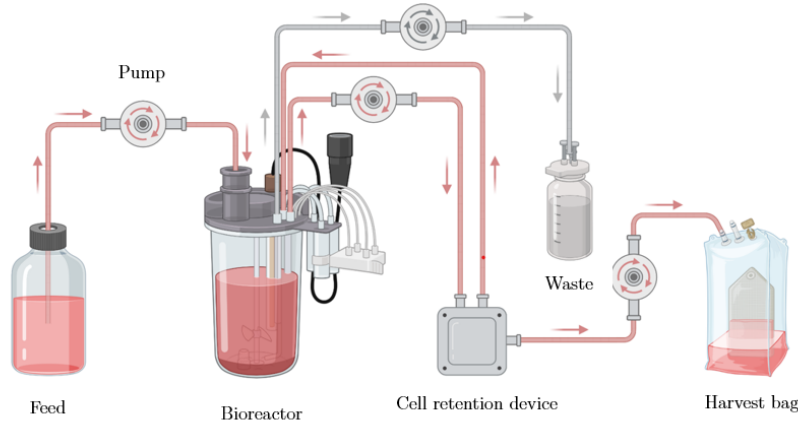


Figure 2.11: Bioreactor in perfusion operating mode.

- **Internal Metabolites:** Metabolites that are involved in intracellular reactions;
- **External Metabolites:** Molecules exchanged with the environment (e.g., glucose uptake, lactate secretion);
- **Cofactors and Enzymes:** Essential molecules facilitating reactions but not consumed in the process.

At the reaction level, a **substrate** refers to a reactant that is transformed by an enzymatic reaction, and a **product** is the compound formed. However, at the cellular level, a substrate often refers to a compound available in the extracellular medium that the cell uptakes and metabolizes, while a product is a metabolite excreted as a result of cellular metabolism. To avoid ambiguity, the latter are called **extracellular substrate**, sometimes called nutrients, and **extracellular product**.

A metabolic network is a structured representation of how metabolites interact through enzymatic reactions. A metabolic network is mathematically represented using stoichiometric matrices. The primary matrix is the stoichiometric matrix of the metabolic network and is denoted N . It describes how internal metabolites participate in metabolic reactions. It is an $m \times n$ matrix where m represents the number of internal metabolites and n is the number of intracellular reactions. In addition, another key matrix is the stoichiometric matrix of extracellular measurements and is denoted N_e . This matrix encodes the stoichiometry of the metabolites exchanged with the external environment, providing a macroscopic view of metabolic inputs and outputs. It is an $m_e \times n$ matrix where m_e stands for the number of extracellular measurements.

In metabolic network analysis, the biological system under study typically encompasses an entire cell or a specific cellular compartment, to which the principle of mass conservation of internal metabolites is applied. Within this system, internal metabolites participate in enzyme-catalyzed reactions to produce other metabolic intermediates (e.g., via intracellular reactions), or they are transported across the system's boundaries through exchange reactions [10]. The general equation of the internal metabolite dynamics is built on the mass balance principle and is expressed as follows:

$$\frac{dC}{dt} = Nv - \mu C \quad (2.1)$$

where $\underline{C} \in \mathbf{R}^m$ is the vector of intracellular metabolite concentrations (normalized to the cellular volume), $\underline{v} \in \mathbf{R}^n$ is the vector of network fluxes and $N \in \mathbf{R}^{m \times n}$ is the stoichiometric matrix of the metabolic network where the element $N_{i,j}$ is related to the stoichiometric coefficient of the metabolite i taking place in reaction j . In Eq. 2.1, the term $\mu \underline{C}$ expresses the dilution that occurs due to cell growth and μ represents the specific growth rate of the cell.

Solving such models requires detailed knowledge of kinetic parameters, enzyme kinetics, and regulatory mechanisms, which are often difficult to measure or estimate accurately. In this regard, some assumptions are made to simplify the study while remaining biologically relevant. Slow/fast dynamics approximation is applied according to the pseudo-steady-state assumption (QSSA) of internal metabolites. On the one hand, intracellular reactions occur much faster than extracellular reactions. Therefore, it is reasonable to assume that internal metabolites reach steady-state instantaneously in comparison to extracellular compounds and that their transient behavior can be neglected [19]. On the other hand, the influence of dilution due to cell growth is negligible in comparison with the internal and exchange reaction rates so that $\mu \underline{C}$ vanishes. Therefore, Eq. 2.1 becomes an homogeneous system of linear equations:

$$N\underline{v} = 0 \quad (2.2)$$

Besides, network fluxes are often constrained to positivity assuming that the direct reactions prevail over their reverse counterparts:

$$\underline{v} \geq 0 \quad (2.3)$$

System in Eq. 2.2 under the positivity constraints imposed by Eq. 2.3 constitutes a problem of non-negative linear algebra or convex analysis whose solution space is a polyhedral cone, denoted $\mathbf{S}$ in the positive orthant [20]. The pointed polyhedral cone is illustrated in Fig. 2.12.

Another key notion in metabolic analysis is the concept of elementary flux modes (EFMs) [21, 22]. First, in metabolic networks, the number of reactions n typically exceeds the number of metabolites m (i.e., $n > m$), leading to an underdetermined system. Indeed, since the number of unknown fluxes is greater than the number of equations enforcing steady-state constraints $N\underline{v} = 0$, the solution $\underline{v}$ is typically not unique. Then, not all the solutions within this space are of equal biological significance.

Therefore, in order to decompose this space into meaningful biological pathways and to extract a finite set of representative solutions, the notion of elementary flux modes is introduced by adding a non-decomposability constraint. Elementary flux modes are defined as the simplest, smallest, non-decomposable pathways that satisfy the steady-state condition, connecting substrates to products without accumulation of internal metabolites. Since it is elementary, it means that it cannot be decomposed into two or more simpler steady-state flux vectors that also satisfy Eq. 2.2 under Eq. 2.3. Geometrically, EFMs form the extreme rays of the polyhedral cone $\mathbf{S}$ (i.e., the edges of $\mathbf{S}$ constructed by the intersection of the hyperplanes defined by the rows of N), as depicted in Fig. 2.12.

Any steady-state flux distribution $\underline{v}$ contained in the pointed polyhedral cone can be expressed as a non-negative linear combination of EFMs, meaning that elementary modes serve as a finite basis for exploring the metabolic capabilities of the system. It follows:

$$\underline{v} = \sum_{k=1}^K \lambda_k \underline{e}^{(k)}, \quad \lambda_k \geq 0 \quad (2.4)$$

where $\underline{e}^{(k)}$ are the elementary flux modes (finite in number) and λ_k are non-negative scalar coefficients, ensuring that any feasible flux distribution remains within the positive orthant.

Mathematically, the computation of EFMs is achieved using the Double Description (DD) method, a fundamental approach for enumerating convex polyhedral cones [23, 24]. The method incrementally constructs a representation of the flux solution space by iteratively incorporating constraints from the null space of the stoichiometric matrix N , identifying the extreme rays corresponding to EFMs. These extreme rays are minimal, meaning that they cannot be decomposed into combinations of other modes. Elementary flux modes can be calculated by using computational tools such that EFMTOOL [25] or METATOOL [26].

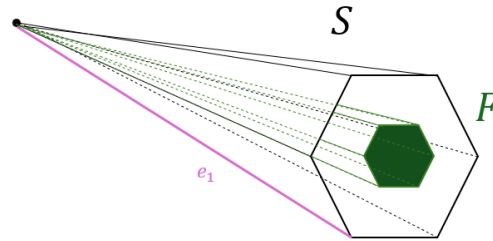


Figure 2.12: Pointed polyhedral cone $\mathbf{S}$ in the positive orthant, the edges of $\mathbf{S}$ representing the elementary modes e_i , and the polytope $\mathbf{F}$.

Extracellular dynamics can also be introduced into the system through cell uptake and excretion/production rates [27]. In that case, the flux distribution is further constrained, as depicted in Fig. 2.12 and the set of solutions is included in the polytope $\mathbf{F}$ ($\mathbf{F} \subset \mathbf{S}$) and only specific combinations of modes provide a solution. By adding a set of linear constraints, making use of experimental uptake and excretion rates $\underline{\nu}_m$, the system in Eq. 2.2 becomes:

$$\begin{pmatrix} N \\ N_e \end{pmatrix} \underline{v} = \begin{pmatrix} 0 \\ \underline{\nu}_m \end{pmatrix} \quad (2.5)$$

The matrix N_e is the stoichiometric matrix of extracellular measurements and the vector $\underline{\nu}_m$ is inferred from the measurement of the time evolution of extracellular concentrations, for example, using regression based on the smoothed derivative of the concentration signals.

Depending on the structure of the stoichiometric matrix N and the availability of external measurements, different approaches are proposed in the literature to solve the system in Eq. 2.2 (or in Eq. 2.5 when experimental rates are taken into account) together with the positivity constraints imposed by Eq. 2.3. These methods include metabolic flux analysis (MFA), metabolic pathway analysis (MPA), flux balance analysis (FBA), and flux variability analysis (FVA).

2.4.2 Metabolic Flux Analysis

Metabolic flux analysis (MFA) is a fundamental computational approach used to quantify the flow of metabolites through a metabolic network. Consider the system in Eq. 2.2. If the vector $\underline{v}$ is partitioned into one vector containing the measured or known fluxes $\underline{v}_k$ and another containing the unknown rates $\underline{v}_u$ such that:

$$\underline{v} = \begin{bmatrix} \underline{v}_k \\ \underline{v}_u \end{bmatrix} \quad (2.6)$$

Then, the stoichiometric matrix N can be partitioned likewise into:

$$N = \begin{bmatrix} N_k & N_u \end{bmatrix} \quad (2.7)$$

where N_k and N_u are the stoichiometric matrices associated with $\underline{v}_k$ and $\underline{v}_u$. Therefore, the equality in Eq. 2.2 is conserved and can be rewritten as:

$$N_u \underline{v}_u + N_k \underline{v}_k = 0 \quad \Longleftrightarrow \quad N_u \underline{v}_u = -N_k \underline{v}_k \quad (2.8)$$

Initially, let us assume that external flux measurements are not taken into account. The vector $\underline{v}_k$ consists of intracellular fluxes that are determined using specific experimental techniques (e.g., isotope labeling, nuclear magnetic resonance, mass spectrometry) or obtained from the literature. In this case, these fluxes are not directly constrained by extracellular uptake or excretion rates. Therefore, the system in Eq. 2.8 is determined when there is a sufficient number of linearly independent constraints such that the solution, i.e., the unmeasured fluxes $\underline{v}_u$, is unique. This is the case where the rank of the matrix N_u is equal to the dimension of $\underline{v}_u$. When this is not the case, the system is either over-determined or under-determined ($\text{rank}(N_u) < \dim(\underline{v}_u)$).

However, in general, it is very difficult to determine intracellular fluxes directly because they require specific experimental techniques or indirect computational approaches. Consequently, it means that $\dim(\underline{v}_k)$ is often equal to zero and that $\dim(\underline{v}_u) = \dim(\underline{v}) = n$. In addition, in metabolic networks, the system is often under-determined because biochemical systems tend to have more reactions n than independent mass balance equations m . It results in under-determined systems with $n - m$ degrees of freedom and there are infinitely many solutions.

A common approach to solve this issue is to apply additional constraints. For instance, additional experimental data (e.g., experimental uptake and excretion rates) can reduce this under-determination. Indeed, let us consider the addition of m_e extracellular fluxes inferred from the measurement of external species. In this case, the degree of freedom is $n - m - m_e$ and the solution space is further constrained. As introduced in Sec. 2.4.1, a set of linear constraints is added:

$$N_e \underline{v} = \underline{v}_m \quad (2.9)$$

and the system in Eq. 2.2 becomes:

$$\begin{pmatrix} N & \underline{0} \\ N_e & -\underline{v}_m \end{pmatrix} \begin{pmatrix} \underline{v} \\ 1 \end{pmatrix} = 0 \quad (2.10)$$

where $\underline{0}$ is a null column vector $\in \mathbf{R}^{m \times 1}$. However, despite the additional constraints from external fluxes, the system in Eq. 2.10 often remains under-determined. This means that classical MFA approaches are insufficient for uniquely determining all the fluxes. Consequently, more advanced computational methods are required. These methods encompass MPA, FBA and FVA and are based on the addition of other constraints.

The latter strategies will be introduced in subsequent chapters.

It is important to note that when the system is over-determined ($n < m$), no exact solution exists unless the equations are consistent (i.e., some constraints are redundant or dependent). If an exact solution does not exist, it is possible to use least-squares approaches to find the best approximate solution.

2.4.3 Toy Example

This section applies the theoretical foundations introduced in Sec. 2.4.1 and 2.4.2 to characterize the metabolism of CHO-320 cells during growth phase. Chinese Hamster Ovary (CHO) are animal cells that are widely used in bio-pharmaceutical production. The choice of a small metabolic network at this stage ensures that fundamental concepts are clearly established before expanding into more complex models in subsequent chapters.

Cells are cultivated in batch mode in stirred flasks in a serum-free medium [28]. During the growth phase, cellular metabolism is given by the metabolic network in Fig. 2.13 and the corresponding metabolic reactions are detailed in Table 2.1 [29]. It focuses on the utilization of two primary energetic nutrients: glucose and glutamine. The metabolism of amino acids is not included. This network encompasses the key pathways such as glycolysis, the pentose phosphate pathway and the tricarboxylic acid cycle. Additionally, a portion of glutamine is assumed to contribute to nucleotide synthesis, with purines and pyrimidines grouped into a single species in equal proportions.

Table 2.1: Reactions for the metabolism of CHO-320 cells during growth phase.

Fluxes	Metabolic reactions
v_1	glucose _{ext} $\rightarrow$ glucose-6-P
v_2	glucose-6-P $\rightarrow$ fructose-6-P
v_3	fructose-6-P $\rightarrow$ glyceraldehyde-3-P + dihydroxyacetone-P
v_4	dihydroxyacetone-P $\rightarrow$ glyceraldehyde-3-P
v_5	glyceraldehyde-3-P $\rightarrow$ pyruvate
v_6	pyruvate $\rightarrow$ lactate _{ext}
v_7	pyruvate + glutamate $\rightarrow$ α -ketoglutarate + alanine _{ext}
v_8	pyruvate $\rightarrow$ acetyl-CoA + CO ₂
v_9	acetyl-CoA + oxaloacetate $\rightarrow$ citrate
v_{10}	citrate $\rightarrow$ α -ketoglutarate + CO ₂
v_{11}	α -ketoglutarate $\rightarrow$ fumarate + CO ₂
v_{12}	fumarate $\rightarrow$ malate
v_{13}	malate $\rightarrow$ oxaloacetate
v_{14}	malate $\rightarrow$ pyruvate + CO ₂
v_{15}	glutamate + oxaloacetate $\rightarrow$ α -ketoglutarate + aspartate
v_{16}	glutamate $\rightarrow$ α -ketoglutarate + NH ₄ ⁺ _{ext}
v_{17}	glutamine _{ext} $\rightarrow$ glutamate + NH ₄ ⁺ _{ext}
v_{18}	3 glutamine _{ext} + 2 aspartate + CO ₂ + 2 ribose-5-P $\rightarrow$ nucleotides + fumarate + 3 glutamate
v_{19}	CO ₂ $\rightarrow$ CO _{2ext}
v_{20}	glucose-6-P $\rightarrow$ ribulose-5-P + CO ₂
v_{21}	ribulose-5-P $\rightarrow$ ribose-5-P
v_{22}	ribulose-5-P $\rightarrow$ xylose-5-P
v_{23}	erythrose-4-P + xylose-5-P $\rightarrow$ fructose-6-P + glyceraldehyde-3-P
v_{24}	ribose-5-P + xylose-5-P $\rightarrow$ fructose-6-P + erythrose-4-P

The metabolic network consists of 24 intracellular reactions ($n = 24$) and 18 internal metabolites ($m = 18$), which are subject to mass balance constraints under steady-state conditions. Additionally, nucleotide synthesis is included in the network as a biosynthetic reaction but is not considered an internal metabolite. In this model, nucleotides represent a terminal biosynthetic product. Instead of detailing the entire biosynthetic pathway, the model assumes that a fraction of glutamine is directly converted into nucleotides, which are then retained inside the cell. Furthermore, as depicted in Fig. 2.13, glucose and glutamine serve as the primary substrates, while lactate, CO_2 , NH_4^+ , and alanine are identified as metabolic byproducts. These compounds correspond to measurable uptake and excretion fluxes, meaning that they are exchanged between the cell and its environment. As a result, they are classified as external metabolites. In this case study, only the uptake rates of glucose and glutamine and the secretion rates of lactate, NH_4^+ and alanine are measured ($m_e = 5$).

Based on the definitions introduced in Sec. 2.4.1, the stoichiometric matrix of the metabolic network $N \in \mathbf{R}^{18 \times 24}$ and the stoichiometric matrix of the extracellular measurements $N_e \in \mathbf{R}^{5 \times 24}$ are given in Table 2.2 and Table 2.3, respectively.

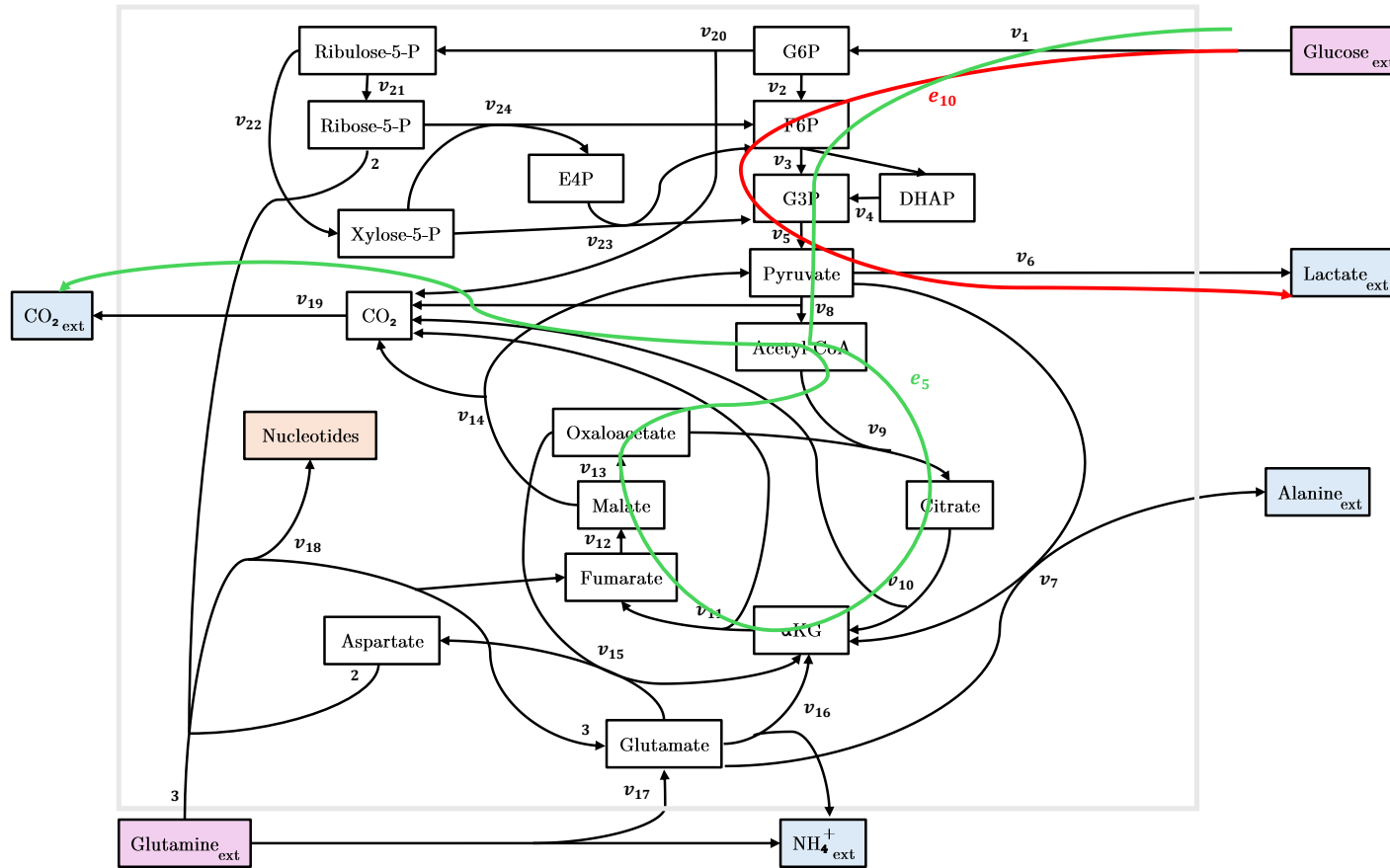


Figure 2.13: Metabolic network containing 18 internal metabolites and 24 intracellular reactions. Internal metabolites are represented by white boxes, extracellular substrates are represented by pink boxes, extracellular products are represented by blue boxes, and the terminal biosynthetic product is in the orange box. The intracellular fluxes are v_i and some elementary flux modes are illustrated as e_i .

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉	V ₁₀	V ₁₁	V ₁₂	V ₁₃	V ₁₄	V ₁₅	V ₁₆	V ₁₇	V ₁₈	V ₁₉	V ₂₀	V ₂₁	V ₂₂	V ₂₃	V ₂₄
G6P	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0
F6P	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
G3P	0	0	1	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
DHAP	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pyruvate	0	0	0	0	1	-1	-1	-1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ru5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0
R5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2	0	0	1	0	0	-1
X5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1
E4P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	1
Acetyl-COA	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Citrate	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
α kG	0	0	0	0	0	0	1	0	0	1	-1	0	0	0	1	1	0	0	0	0	0	0	0	0
Fumarate	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	1	0	0	0	0	0	0
Malate	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0
Oxaloacetate	0	0	0	0	0	0	0	0	-1	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0
Aspartate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	-2	0	0	0	0	0	0
Glutamate	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	-1	-1	1	3	0	0	0	0	0	0
CO ₂	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	-1	-1	1	0	0	0	0

Table 2.2: Stoichiometric matrix N describing the metabolism of CHO-320 cells during growth phase.

	v ₁	v ₂	v ₃	v ₄	v ₅	v ₆	v ₇	v ₈	v ₉	v ₁₀	v ₁₁	v ₁₂	v ₁₃	v ₁₄	v ₁₅	v ₁₆	v ₁₇	v ₁₈	v ₁₉	v ₂₀	v ₂₁	v ₂₂	v ₂₃	v ₂₄
Glucose	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lactate	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Glutamine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-3	0	0	0	0	0	0
NH ₄ ⁺	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Alanine	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2.3: Stoichiometric matrix of the extracellular measurements, denoted N_e .

From the stoichiometric matrix N , the elementary flux modes can be computed using the Double Description (DD) method. They are calculated using EFMTOOL and are gathered into the matrix E of dimension $n \times n_{EFM}$, given in Table 2.4. For the metabolic network under study, $n_{EFM} = 11$ EFMs are obtained. They represent the simplest, non-decomposable pathways connecting substrates to products without accumulation of internal metabolites. The elementary vectors e_5 and e_{10} are represented by green and red arrows, respectively, in Fig. 2.13.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}
v_1	2	2	0	3	1	0	2	0	3	1	2
v_2	0	0	0	0	1	0	0	0	0	1	0
v_3	0	0	0	2	1	0	0	0	2	1	0
v_4	0	0	0	2	1	0	0	0	2	1	0
v_5	0	0	0	5	2	0	0	0	5	2	0
v_6	0	0	1	5	0	0	1	0	0	2	2
v_7	1	0	0	0	0	1	1	0	0	0	0
v_8	1	2	0	0	2	0	0	1	5	0	0
v_9	1	2	0	0	2	0	0	1	5	0	0
v_{10}	1	2	0	0	2	0	0	1	5	0	0
v_{11}	4	5	1	0	2	1	3	2	5	0	3
v_{12}	5	6	1	0	2	1	4	2	5	0	4
v_{13}	3	4	0	0	2	0	2	1	5	0	2
v_{14}	2	2	1	0	0	1	2	1	0	0	2
v_{15}	2	2	0	0	0	0	2	0	0	0	2
v_{16}	0	1	1	0	0	0	0	1	0	0	1
v_{17}	0	0	1	0	0	1	0	1	0	0	0
v_{18}	1	1	0	0	0	0	1	0	0	0	1
v_{19}	9	12	2	3	6	2	6	5	18	0	6
v_{20}	2	2	0	3	0	0	2	0	3	0	2
v_{21}	2	2	0	1	0	0	2	0	1	0	2
v_{22}	0	0	0	2	0	0	0	0	2	0	0
v_{23}	0	0	0	1	0	0	0	0	1	0	0
v_{24}	0	0	0	1	0	0	0	0	1	0	0

Table 2.4: Matrix of elementary flux modes, denoted $E^{n \times n_{EFM}}$, where $n_{EFM} = 11$.

Finally, the primary objective in metabolic flux analysis is to determine a feasible solution for the system described in Eq. 2.2 under the positivity constraints given in Eq. 2.3. Unfortunately, the current system is under-determined ($n > m$), meaning there is no unique solution for $\underline{v}$. However, the integration of experimental data provides additional constraints that help restrict the solution space. In this case study, five extracellular fluxes ($m_e = 5$) are experimentally measured, thereby partially constraining the system. Despite this reduction in degrees of freedom, the system remains under-determined, implying that Eq. 2.10 cannot be solved using the classical MFA approach. Therefore, further constraints must be considered, and methodologies such as MPA, FBA or FVA must be used.

2.5 Fundamentals of Genetics

Genetics is the cornerstone of modern molecular biology and biotechnology, providing the framework for understanding gene expression, gene regulation, and inheritance. Fundamental genetic principles allow researchers to manipulate biological systems for applications ranging from synthetic biology to medical therapeutics. This section provides an overview of key genetic processes and elements, as well as an introduction to genetic modeling. These concepts will serve as the basis for subsequent discussions on gene circuit design, gene expression optimization, and synthetic biology approaches.

2.5.1 Gene Expression

Gene expression is the process by which genetic information encoded in DNA is converted into functional proteins. It involves two main steps: **transcription** and **translation**. These steps are represented in Fig. 2.14.

Transcription is the process in which an RNA molecule is synthesized from a DNA template. This process occurs in three stages:

- **Initiation** - RNA polymerase binds to a specific DNA sequence known as the promoter, facilitating the unwinding of the DNA strands;
- **Elongation** - RNA polymerase moves along the DNA strand, synthesizing a complementary RNA transcript;
- **Termination** - Transcription stops when RNA polymerase encounters a termination signal, and the RNA transcript is released.

In prokaryotes, transcription occurs in the cytoplasm, and the RNA transcript is directly a messenger RNA (mRNA) that can be immediately used for translation. In eukaryotes, transcription occurs mainly in the nucleus, and the initial transcript is called pre-mRNA, which undergoes several post-transcriptional modifications, including capping, splicing, and polyadenylation. After these modifications, the mature mRNA is exported from the nucleus to the cytoplasm for translation [30, 31].

Translation is the process by which ribosomes synthesize proteins using the mRNA transcript as a template. It consists of three key steps:

- **Initiation** - The ribosome assembles around the start codon of the mRNA, guided by the ribosome binding site (RBS) in prokaryotes or the Kozak sequence in eukaryotes;
- **Elongation** - Transfer RNA (tRNA) molecules deliver amino acids to the ribosome, where they are joined together to form a polypeptide chain;
- **Termination** - Translation ends when the ribosome encounters a stop codon, releasing a newly synthesized protein.

In prokaryotes, translation occurs in the cytoplasm, often occurring simultaneously with transcription. In eukaryotes, translation takes place in the cytoplasm, usually on ribosomes either freely floating in the cytosol or attached to the endoplasmic reticulum (ER) [32].

The efficiency of gene expression is modulated at multiple levels, including transcriptional and translational control. This regulation ensures that genes are expressed at the right time, location, and quantity to meet cellular needs [33, 34].

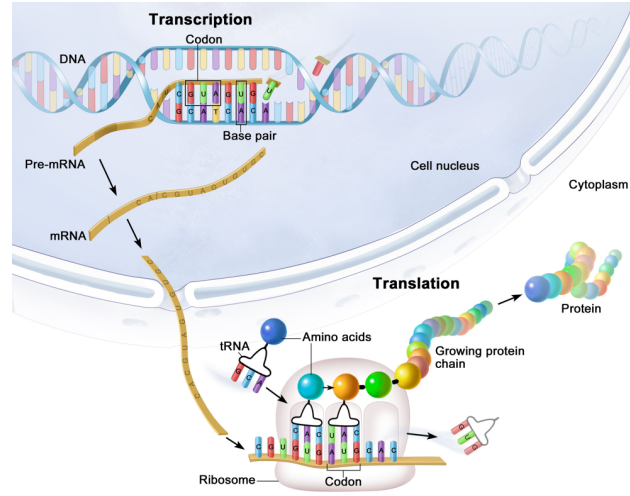


Figure 2.14: Transcription and translation steps in eukaryotes.

2.5.2 Genetic Elements

A **promoter** is a DNA sequence that initiates the transcription of a gene by providing a binding site for RNA polymerase and associated transcription factors. The promoters then play a crucial role in regulating gene expression and can be classified based on their strength, specificity, and inducibility [31].

The **promoter strength** refers to the efficiency with which a promoter can initiate transcription, which is influenced by sequence composition, transcription factor availability and chromatin structure (for eukaryotic cells). Strong promoters drive high levels of gene expression, whereas weak promoters yield lower transcription rates. The field of synthetic biology often employs engineered promoters to fine-tune gene expression for specific applications [3, 35, 36].

The **ribosome binding site (RBS)** is a short nucleotide sequence upstream of the start codon that facilitates the recruitment of ribosomes for translation initiation [32].

The **RBS strength** determines the efficiency of translation initiation and, consequently, the protein synthesis rate. Strong RBS sequences enhance ribosome recruitment, leading to higher protein production, while weaker RBS sequences reduce translation efficiency. The strength of a ribosome binding site can be quantitatively predicted and optimized using computational tools such as the RBS CALCULATOR [37], which considers RNA secondary structure, ribosome binding affinity, and translation initiation rates [38].

2.5.3 Genetic Models

Genetic models are mathematical or computational frameworks used to describe and predict gene expression and regulatory interactions within biological systems. These models are based on experimental data, biochemical kinetics, and system biology approaches [39]. Models can be deterministic (i.e., based on ordinary differential equations) or stochastic (i.e., accounting for intrinsic noise in gene expression).

Genetic models play an important role in synthetic biology, metabolic engineering, and evolutionary biology. They enable:

- The **prediction of gene expression dynamics** based on promoter strength and RBS strength;
- The **optimization of synthetic gene circuits** for desired expression levels and regulatory behaviors;
- The **analysis of gene regulatory networks** to understand natural and synthetic biological systems.

Furthermore, recent advances in machine learning and computational biology have further enhanced the predictive power of genetic models, allowing for the rational design of gene expression systems [40].

2.5.4 Synthetic Biology

Synthetic biology is an interdisciplinary field that applies engineering principles to biological systems, enabling the design and construction of new biological functions and systems. It involves the rational design of genetic circuits, metabolic pathways, and entire organisms for applications in medicine, bioenergy, and industrial biotechnology [41, 42].

In this context, synthetic gene circuits are engineered genetic systems that function similarly to electronic circuits, controlling gene expression in a predictable and programmable manner. These circuits often comprise genetic elements such as promoters, ribosome binding sites, transcription factors, and regulatory sequences, which work together to achieve specific biological functions. More information can be found in [39, 43].

The introduction of synthetic genetic circuits into cells can impose metabolic and regulatory burdens, affecting cellular growth, stability, and native gene expression. These effects include:

- **Resource Competition** - The demand for ribosomes, RNA polymerases, and other cellular machinery can limit the efficiency of both native and synthetic gene expression;
- **Metabolic Load** - The expression of additional genes may divert energy and precursor molecules from essential cellular processes, impacting cell viability;
- **Unintended Crosstalk** - Synthetic circuits may interact unpredictably with endogenous genetic networks, leading to undesired regulatory effects.

These factors are critical in the design and optimization of synthetic circuits and must be accounted for in genetic models. While the present section introduces these considerations, a more in-depth exploration of their impact and mitigation strategies will be provided and discussed in Sec. 5.2.

2.6 Artificial Neural Networks

Machine learning is a subfield of artificial intelligence (AI) that focuses on developing algorithms capable of learning patterns from data. Today, machine learning techniques have become central to a variety of applications, including natural language processing, computer vision, and biomedical data analysis [44]. The primary goal of machine

learning is to develop models that generalize well to unseen data, which is achieved through various learning paradigms, including supervised, unsupervised, and reinforcement learning [45].

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks and form a subset of machine learning methods. While machine learning encompasses a broad range of techniques, artificial neural networks specifically refer to structures composed of interconnected artificial neurons designed to approximate complex functions [46]. The recent resurgence of ANNs, particularly deep learning, is attributed to increased computational power, large datasets, and advances in optimization techniques [44].

The fundamental unit of a neural network for supervised learning is called a single layer perceptron, as depicted in Fig. 2.15. The perceptron, first introduced in [47], is one of the first artificial neural network models. It consists of an input layer that holds input values and does not perform computations, and one output layer that contains one or possibly more neurons, each of them applying a weighted sum followed by an activation function, typically a step function. However, the perceptron can only model linearly separable problems, which limits its applicability.

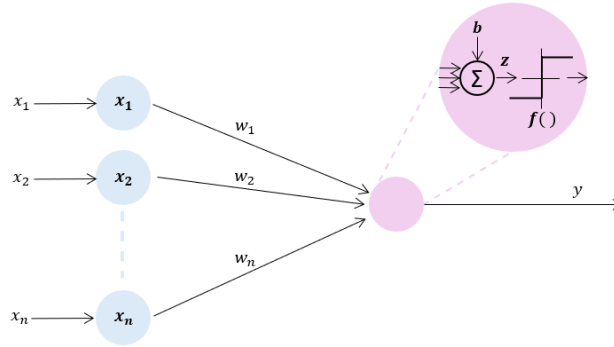


Figure 2.15: A single layer perceptron with one output neuron: x_i represents the input signals, w_i denotes synaptic coefficients, b is a bias, f is an activation function and y is the output signal.

Mathematically, the perceptron illustrated in Fig. 2.15 involves a set of input signals x_i weighted by synaptic coefficients w_i , plus a bias b :

$$z = b + \sum_{i=1}^n w_i x_i \quad (2.11)$$

This weighted sum enters an activation function f to predict the output signal $y = f(z)$.

2.6.1 Neural Network Learning Algorithms

In the context of machine learning, learning refers to the process by which a model adjusts its parameters based on input data to improve its performance on a given task. Learning (or training) enables the network to generalize from examples and make predictions or decisions. The objective is to minimize errors, optimize performance, and adapt to new data. Learning in ANNs can be categorized into different strategies depending on the availability of labeled data, the presence of feedback signals, and the

type of problem to be solved [48].

Supervised learning is the most widely used training paradigm in ANNs, where the model learns from labeled data, which means that both input and output data are provided. The network processes the input signals, predicts the output signals, and compares the predicted outputs against the actual outputs of the labeled data. The objective is to minimize the error by adjusting the parameters of the network (i.e., weights and biases) through an optimization process.

The training data are composed of input-output pairs (X, Y) , where X represents the input features and Y denotes the corresponding target values. The ANN makes predictions using a function $f(X, \Theta)$, where Θ represents the network parameters (weights and biases). A cost function quantifies the error between the predicted and actual outputs, and the network parameters are optimized using algorithms such as Stochastic Gradient Descent (SGD) or Adam. The learning process continues until the network converges to an optimal set of parameters that generalizes well to unseen data [49].

Supervised learning is commonly used for image classification (e.g., Convolutional Neural Networks for object recognition), speech recognition (e.g., Recurrent Neural Networks for speech-to-text systems) and medical diagnosis (e.g., detecting diseases from medical imaging).

Unsupervised learning is another type of training in which networks learn from input data without explicit output labels. Unlike supervised learning, there are no target values to guide the training [48]. Instead, the model identifies the underlying patterns, structures, or relationships in the data. This approach is particularly useful for exploratory analysis and feature learning.

For example, some common unsupervised learning techniques are clustering algorithms (e.g., k-means clustering). This type of training is used for customer segmentation in marketing, anomaly detection, and feature extraction and dimensionality reduction (e.g., principal component analysis).

Reinforcement learning is a learning paradigm in which an agent learns by interacting with an environment. The term *reinforcement* describes the principle of learning by means of a *reward function*. This function assigns *penalties* or *rewards* to the network output, depending on how the output relates to the learning goal [48].

As depicted in Fig. 2.16, the reinforcement learning consists of an agent, an environment, actions, a reward signal, and a policy. The agent takes an action a_t at the time step t , based on its policy $\pi(a_t|s_t)$, which maps states s_t to actions. The environment responds by transitioning to a new state s_{t+1} and providing a reward r_t based on the action taken. The objective of the agent is to learn a policy that maximizes the cumulative reward over time, typically optimized using algorithms such as Q-learning and Policy Gradient methods.

2.6.2 Neural Network Architectures

At a later stage of the work, artificial neural networks (ANNs) will be used as part of the methodology. Given the availability of labeled data, the use of ANNs trained through supervised learning constitutes a suitable choice. The most commonly used neural network architectures in supervised learning include **Feedforward Neural Networks** (FNNs), **Convolutional Neural Networks** (CNNs) and **Recurrent Neural Networks** (RNNs).

Feedforward Neural Networks (FNNs) represent the simplest class of arti-

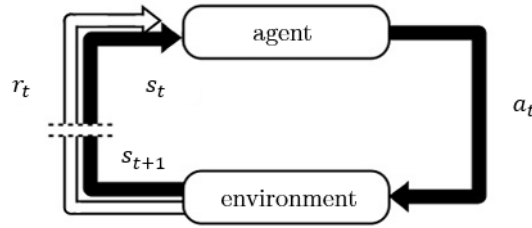


Figure 2.16: Sequential decision process of reinforcement learning.

cial neural networks where information flows in one direction, from the input layer to the output layer, passing through one or more hidden layers (without cycles or loops). The backpropagation algorithm is the foundation of supervised learning in FNNs where errors between predicted and actual outputs are propagated back through the system, causing the system to adjust the parameters.

FNNs are typically fully connected networks, which means that every neuron in a given layer is connected to every neuron in the next layer. This leads to a large number of parameters, making them computationally expensive compared to more specialized architectures such as CNNs and RNNs.

For example, multilayer perceptrons (MLPs) are classified as FNNs. To overcome the limitations of the perceptron, MLPs incorporate multiple layers and non-linear activation functions, such as the sigmoid, hyperbolic tangent ($\tanh$), and the rectified linear unit (ReLU) [50]. This allows the network to model complex and non-linear relationships [51]. As a matter of fact, the universal approximation theorem states that a multilayer perceptron with at least one hidden layer and a non-linear activation function can approximate any continuous function given sufficient neurons [52, 53].

Convolutionary Neural Networks (CNNs) are designed for object recognition and are used in image processing tasks. CNNs use convolutional layers to extract spatial features from input data that is an image. These layers apply learnable filters, also called kernels, which are small matrices that move across small regions of the image and perform convolution operation at each step. This sliding process creates a feature map (a new matrix) that represents the detected patterns. Once the features are extracted, pooling layers reduce the spatial dimensions of the feature maps by down-sampling them, reducing computational complexity while retaining important information. The final extracted features are then passed through fully connected layers, where they are used for classification or regression tasks [54]. Training a CNN follows a process similar to that of FNNs.

Recurrent Neural Networks (RNNs) are designed for sequential data, making them particularly suitable for applications such as speed recognition and language modeling. Unlike FNNs and CNNs, RNNs have recurrent connections that allow information to persist across time steps. Regarding the learning strategy of RNNs, the backpropagation algorithm does not rely on the classical *feed-forward* layout but on a feed-forward layout with *feedbacks*. As a matter of fact, training a recurrent neural network involves a variant of backpropagation called backpropagation through time (BPTT). More information can be found in [55].

2.7 Parameter Estimation

Parameter estimation is a fundamental aspect of mathematical modeling, as it ensures that models accurately represent real-world systems. Developing any mathematical model requires defining an appropriate model structure that captures the essential dynamics of the system. However, for the model to be meaningful and applicable, the value of its parameters must be determined based on empirical data. Parameters define key characteristics of the system, such as growth rates in biological models or reaction coefficients in chemical processes. Since these parameters are often unknown or uncertain, estimation techniques, such as statistical inference, optimization methods, or machine learning, are used to determine their values from experimental observations.

Accurate parameter estimation improves model reliability, enhances predictive power, and allows for better decision-making in fields like engineering. Without proper estimation, even well-structured models can yield incorrect or misleading results, reducing their effectiveness in practical applications. This section introduces the basics of both linear and non-linear parameter estimation. These concepts will be further explored and applied in the subsequent chapters.

2.7.1 Linear Parameter Estimation

Linear parameter estimation is applicable when the model can be expressed in a linear form with respect to its parameters. This approach often relies on least-squares regression techniques (e.g., ordinary least squares, weighted least squares) to estimate the parameters efficiently. Linear methods are computationally less expensive and generally provide unique solutions when the model structure allows for linearization.

Ordinary least squares (OLS) is the most common estimator. Let consider the following linear regression model (scalar case where i is the time index):

$$y(i) = y_m(i) + \epsilon(i) = \underline{x}^T(i)\underline{\theta} + \epsilon(i) \quad i = 1 \rightarrow M \quad (2.12)$$

where $\underline{y} = [y(1) \ y(2) \ \dots \ y(M)]^T$ is the vector of measurements (stochastic variables), $\underline{y}_m$ is the vector of outputs given by the model, $\underline{\epsilon} = [\epsilon(1) \ \epsilon(2) \ \dots \ \epsilon(M)]^T$ is the vector of measurement errors (stochastic variables), $\underline{x}(i) = [x_1(i) \ x_2(i) \ \dots \ x_p(i)]^T$ is the vector of explicative variables, supposed perfectly known, whose the matrix notation is $X = [\underline{x}(1) \ \underline{x}(2) \ \dots \ \underline{x}(M)]^T$, and $\underline{\theta} = [\theta_1 \ \theta_2 \ \dots \ \theta_p]^T$. M is the number of experimental points (that is, the number of observations) and p is the number of parameters.

The objective of OLS is to estimate $\underline{\theta}$ such that the residual sum of squares is minimized:

$$J(\underline{\theta}) = \sum_{i=1}^M \epsilon^2(i) = \sum_{i=1}^M \left(y(i) - \underline{x}^T(i)\underline{\theta} \right)^2 \quad (2.13)$$

The OLS estimator of $\underline{\theta}$ is denoted $\hat{\underline{\theta}}$ and can be obtained as follows:

$$\hat{\underline{\theta}} = \arg \min_{\underline{\theta}} J(\underline{\theta}) = (X^T X)^{-1} X^T \underline{y} \quad (2.14)$$

Furthermore, understanding the estimation error and its covariance is essential for constructing confidence intervals and assessing model reliability. The estimation error is given as follows:

$$\tilde{\underline{\theta}} = (X^T X)^{-1} X^T \underline{\epsilon} \quad (2.15)$$

The covariance matrix of the estimation error describes the variability of the estimated parameters $\hat{\underline{\theta}}$ and serves to compute the confidence intervals (CIs). The covariance matrix of the estimation error is given as follows:

$$E[\tilde{\underline{\theta}} \tilde{\underline{\theta}}^T] = \sigma^2 (X^T X)^{-1} \quad (2.16)$$

where σ^2 represents the variance of the residuals.

Moreover, a correct estimate for the error variance can be obtained from the residuals as follows:

$$\hat{\sigma}^2 = \frac{J(\hat{\underline{\theta}})}{M - p} \quad (2.17)$$

Therefore, for a 95% confidence interval, considering a normal distribution, it comes:

$$\hat{\underline{\theta}} - 1.96 \sqrt{\text{diag}\left(E[\tilde{\underline{\theta}} \tilde{\underline{\theta}}^T]\right)} \leq \hat{\underline{\theta}} \leq \hat{\underline{\theta}} + 1.96 \sqrt{\text{diag}\left(E[\tilde{\underline{\theta}} \tilde{\underline{\theta}}^T]\right)} \quad (2.18)$$

2.7.2 Nonlinear Parameter Estimation

Nonlinear parameter estimation is required for models where the parameters appear non-linearly. These problems are typically solved using iterative optimization algorithms such as Levenberg-Marquardt, genetic algorithms, or Bayesian approaches. Nonlinear estimation can be more computationally intensive and sensitive to initial guesses but provides greater flexibility for capturing complex behaviors.

Nonlinear least squares (NLS) is a generalization of the OLS method and is used when the relationship between dependent and independent variables is non-linear in the parameters. In this case, Eq. 2.12 becomes:

$$y(i) = y_m(i) + \epsilon(i) = y_m(\underline{x}(i), \underline{\theta}) + \epsilon(i) \quad i = 1 \rightarrow M \quad (2.19)$$

where $y_m(\underline{x}(i), \underline{\theta})$ represents a nonlinear function of $\underline{x}(i)$ and parameters $\underline{\theta}$.

As before, the objective of NLS is to estimate the parameters $\underline{\theta}$ by minimizing the sum of squared residuals:

$$J(\underline{\theta}) = \sum_{i=1}^M \epsilon^2(i) = \sum_{i=1}^M \left(y(i) - y_m(\underline{x}(i), \underline{\theta}) \right)^2 \quad (2.20)$$

And the parameter estimates can be obtained as follows:

$$\hat{\underline{\theta}} = \arg \min_{\underline{\theta}} J(\underline{\theta}) \quad (2.21)$$

In this case, because NLS involves nonlinear functions, there is no closed-form solution for Eq. 2.21 and a numerical optimization procedure is required such as the Gauss-Newton method, the Levenberg-Marquardt algorithm (e.g., *lsqnonlin* in MATLAB) or the use of Nelder Mead (e.g., *fminsearch* in MATLAB).

For nonlinear least-squares, the covariance matrix of the estimation error cannot be simplified as in Eq. 2.16 and requires to compute explicitly the Fisher Information Matrix F :

$$E[\tilde{\underline{\theta}} \tilde{\underline{\theta}}^T] \approx F^{-1} = \sigma^2 P^{-1} \quad (2.22)$$

where σ^2 represents the measurement variance and a correct estimate is still given by Eq. 2.17. The matrix P is defined as follows:

$$P = \sum_{i=1}^M \left(\underline{y}_{\underline{\theta}}(\underline{x}(i), \underline{\theta}) \cdot \underline{y}_{\underline{\theta}}^T(\underline{x}(i), \underline{\theta}) \right) \quad (2.23)$$

where $\underline{y}_{\underline{\theta}}(\underline{x}(i), \underline{\theta})$ contains the parametric sensitivities and is defined as:

$$\underline{y}_{\underline{\theta}}(\underline{x}(i), \underline{\theta}) = \left. \frac{\partial y_m(\underline{x}(i), \underline{\theta})}{\partial \underline{\theta}} \right|_{\underline{\theta}=\hat{\underline{\theta}}} \quad (2.24)$$

When the covariance matrix of the estimation error is computed, confidence intervals can be calculated as before (see Eq. 2.18).

Considering the extension to the vector case:

$$\underline{y}(i) = \underline{y}_m(\underline{x}(i), \underline{\theta}) + \underline{\epsilon}(i) \quad i = 1 \rightarrow M \quad (2.25)$$

The expression of the cost function in Eq. 2.20 becomes:

$$J(\underline{\theta}) = \sum_{i=1}^M \left(\underline{y}(i) - \underline{y}_m(\underline{x}(i), \underline{\theta}) \right)^T \cdot Q^{-1}(i) \cdot \left(\underline{y}(i) - \underline{y}_m(\underline{x}(i), \underline{\theta}) \right) \quad (2.26)$$

For weighted least-squares, Q is a positive semi-definite matrix and for the maximum likelihood, Q is the covariance matrix of the measurement noise.

Chapter 3

Macroscopic Modeling

Mathematical modeling has long been instrumental in deciphering the complexities of cellular culture processes, providing a framework to predict and optimize cellular behavior in various biotechnological applications. These models can be broadly categorized into microscopic and macroscopic approaches. While microscopic models investigate intricate intracellular mechanisms, macroscopic models simplify the system to focus on observable extracellular phenomena, offering a more streamline perspective for large-scale applications.

Macroscopic modeling emphasizes the relationships between extracellular inputs and outputs, such as nutrient uptake and product formation rates, without detailing the underlying intracellular pathways. This approach simplifies the complexity inherent in cellular systems, making it particularly valuable for industrial bioprocess optimization [56]. In [57], a comprehensive review is proposed in which the authors outline the main strategies in macroscopic modeling, highlighting the identification of stoichiometric relationships that connect extracellular substrates to products, followed by the development of kinetic models to simulate cell growth and metabolism.

Within the realm of macroscopic modeling, macroscopic metabolic models have gained prominence for their ability to represent cellular metabolism by linking extracellular measurements to metabolic activity. A systematic methodology for designing such models involves the use of metabolic networks to derive macroscopic bioreactions that connect extracellular substrates and products [9, 58, 59]. This approach employs elementary flux modes (EFMs) to establish a direct connection between intracellular metabolic pathways and observable macroscopic behavior [10, 22].

In this context, EFMs are defined as the simplest sets of enzyme-catalyzed reactions that can operate at steady-state, converting substrates into products without any net accumulation of internal metabolites. As such, these elementary modes provide a comprehensive representation of all possible metabolic pathways within a network, and researchers can then infer macroscopic reactions that capture the essential features of complex intracellular processes. Therefore, this method enables the development of macroscopic models that maintain a connection to the underlying cellular metabolism, offering a balance between simplicity and biological relevance.

However, the derivation of macroscopic models from elementary flux modes (EFMs) presents significant challenges, particularly when dealing with large metabolic networks. Indeed, as the level of detail increases, so does the size of the corresponding stoichiomet-

ric matrix, leading to a combinatorial explosion in the number of EFMs [60, 61]. This phenomenon poses a fundamental challenge, as the enumeration of all EFMs rapidly becomes computationally intractable for complex networks. As a consequence, addressing this issue requires efficient methodologies for selecting a representative subset of EFMs, which will serve as the basis for constructing an appropriate reaction scheme for macroscopic modeling.

Beyond the combinatorial challenge in EFM enumeration, another critical difficulty in metabolic modeling lies in the description of reaction kinetics. As a matter of fact, once a set of EFMs is selected to represent the macroscopic reactions, the next step involves the modeling of the reaction rates. Traditionally, kinetic models rely on classical rate laws such as Michaelis-Menten or Monod kinetics [62, 63], which assume predefined mechanistic interactions. However, in many cases, particularly for complex and highly interconnected metabolic pathways, the derivation of such kinetic laws is highly challenging, due to incomplete knowledge of enzyme activities, regulatory effects, and metabolite concentrations [64].

To overcome these challenges, this chapter proposes a hybrid modeling approach that integrates EFM-based macroscopic reaction derivation with machine learning techniques for kinetic modeling, if necessary. First, methods are developed to reduce the number of elementary modes and select a subset of EFMs that best represent the system’s macroscopic behavior, even when the complete set of modes is not completely enumerated. Then, the problem of kinetic modeling is addressed by comparing classical approaches with artificial neural networks (ANNs). ANNs provide a data-driven alternative to traditional kinetics, making use of experimental or simulation data to learn complex flux patterns that are otherwise difficult to capture using conventional mechanistic models [65, 66].

The applications and case studies presented in this chapter focus on the modeling of animal cell cultures, a field of significant interest in biotechnology and biopharmaceutical production. Animal cells are widely used for the production of recombinant proteins, monoclonal antibodies, and viral vectors, requiring efficient bioprocess optimization [67]. Metabolic models play a crucial role in improving cell culture performance by predicting cellular responses to environmental and genetic perturbations, ultimately enabling more efficient process control and optimization [68, 69].

Modeling the metabolism of animal cells is particularly challenging due to their complex regulatory mechanisms, non-trivial metabolic interactions, and dependency on extracellular factors [70]. In this context, the development of accurate macroscopic models is highly valuable for optimizing bioprocess parameters. The proposed hybrid modeling approach addresses these challenges by providing a structured methodology for deriving macroscopic reaction networks from EFMs, and capturing kinetic behaviors, possibly using ANNs.

Chapter Outline

The remainder of this chapter is structured as follows:

- **Literature Review** - A review of existing methodologies for deriving macroscopic models from EFMs is presented, with a focus on EFM enumeration strategies, reduction techniques, and selection criteria. Additionally, existing approaches for kinetic modeling, including mechanistic and data-driven methods, are discussed;

- **Theoretical Background and Methods** - This section introduces the proposed hybrid modeling framework. The first part describes how a representative and informative set of EFMs can be selected to infer the macroscopic reaction scheme. The second part details different approaches for modeling the kinetics of these reactions, emphasizing the advantages of using ANNs in complex cases;
- **Case Studies** - Two case studies are presented to illustrate the applicability of the methodology and highlight the modular nature of the modeling procedure. In some cases, a complete enumeration of EFMs is feasible, whereas in others, only a subset of elementary modes can be identified due to computational constraints. Similarly, for kinetic modeling, classical kinetic laws can be applied when the flux patterns exhibit well-defined relationships with extracellular concentrations, but when the system displays highly intricate or nonlinear flux behaviors, artificial neural networks (ANNs) provide a more suitable alternative;
- **Future Perspectives** - Potential improvements to the proposed modeling approach are discussed, including alternative selection strategies for EFMs, integration of additional biological constraints, and refinements in ANNs-based kinetic modeling;
- **Conclusions** - A summary of key findings is provided, emphasizing the advantages and limitations of the hybrid approach.

Contributions

The primary contributions of this chapter are as follows:

- A systematic approach for selecting a representative set of EFMs to deduce a macroscopic reaction scheme, even in cases where the full set of EFMs is not enumerated;
- A hybrid modeling framework that integrates mechanistic and machine-learning-based approaches for kinetic modeling;
- A demonstration of the applicability of the proposed methodology through case studies in animal cell cultures, highlighting scenarios where the use of traditional kinetic laws is possible and where ANNs offer superior predictive performance.

By addressing both the combinatorial explosion of EFMs and the challenge of kinetic modeling, this work provides a practical solution to derive macroscopic models from EFMs with direct applications in the biotechnological and biomedical fields. See more information in Sec. 1.2.

List of Publications

Journal Papers in Special Issues

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. A Systematic Elementary Flux Mode Selection Procedure for Deriving Macroscopic Bioreaction Models from Metabolic Networks. *Journal of Process Control* **2022**, *118*, 170 - 184.
- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Hybrid Dynamic Models of Bioprocesses Based on Elementary Flux Modes and Multilayer Perceptrons. *Processes* **2022**, *10*, 2084.

Conference Paper as Keynote Paper

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Selection of a Minimal Suboptimal Set of EFMs for Dynamic Metabolic Modelling. *IFAC-PapersOnLine*. 11th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2021: Venice, Italy, 13-16 June 2021. **2021**, *54*, 667 - 672.

Conferences with Abstract

- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. Selection of Elementary Flux Modes for the Development of Reduced Macroscopic Models. *40th Benelux Meeting on Systems and Control*, **2021**.
- Maton, M.; Bogaerts, Ph.; Vande Wouwer, A. A Modular Elementary Flux Mode Reduction Procedure for Dynamic Metabolic Modelling. *41st Benelux Meeting on Systems and Control*, **2022**.

3.1 Literature Review

The use of mammalian cell cultures has become a cornerstone of biotechnology, particularly in biopharmaceutical production, where they are essential for manufacturing monoclonal antibodies, viral vaccines, and therapeutic proteins for treating genetic disorders [70–72]. Despite their critical role in industry, large-scale cultivation presents several engineering challenges, including high cell densities, process reproducibility, and optimization of nutrient uptake and product formation [12]. To address these challenges and improve process efficiency, macroscopic metabolic modeling has emerged as a valuable tool, offering insights that support model-based optimization, control strategies, and the development of digital twins for bioprocesses [73].

3.1.1 Model Derivation

The derivation of minimal bioreaction models is of primary importance in the development of monitoring and control strategies for the production of cell or microorganism cultures. As discussed previously, mathematical modeling of bioprocesses can be categorized into structured and unstructured models [8] and both are used in different types of applications. On the one hand, structured models, also called microscopic models, offer detailed intracellular description, explicitly capturing complex interactions within metabolic pathways [74]. Such models attempt to explicitly describe the physiological state of a studied cell strain, its composition or its regulatory adaptation to environmental changes. However, the description of these numerous and complex reactions comes at the cost of a large number of equations that are usually difficult to handle. On the other hand, unstructured models, often referred to as macroscopic models, focus solely on the extracellular environment, describing the evolution of nutrient consumption and product formation, without considering intracellular dynamics [75]. They are known to describe the evolution of extracellular variables, such as biomass, extracellular substrates and products by a smaller set of macroscopic reactions. Their lower complexity level eases the parametric identification procedures in view of process optimization.

Over the years, connections have been established between the two approaches, and intra- and extracellular information can be combined to develop reduced-order macroscopic models, allowing simplified but biologically relevant representations of cellular metabolism [8, 15, 73, 76]. A central concept in the development of minimal bioreaction models is the notion of elementary flux modes (EFMs) [21] or extreme pathways [74], which can be seen as the simplest metabolic pathways linking extracellular substrates to products. The EFMs form a convex basis of the flux space and allow the derivation of macroscopic bioreaction models that retain essential metabolic characteristics. However, one of the primary challenges of EFM-based modeling is the combinatorial explosion in the number of modes as the size of the metabolic network increases. Therefore, this computational burden requires efficient selection and reduction techniques to extract only the most relevant EFMs for model construction.

A first set of methods are **network reduction methods** where simplified metabolic networks can be obtained on the basis of detailed networks through metabolic flux analysis [75], for instance, by discarding insignificant fluxes [77]. In this way, the dimension of the network is reduced, and the elementary vectors can all be computed and enumerated using specific software tools such as METATOOL [78] and EFMTOL [25].

Alternatively, when the number of EFMs is prohibitively large [79], methods allowing only the **selection of EFM subsets** are required. To this extent, several techniques

have emerged in the last decades, including the computation of the shortest elementary flux modes [80], subsystem analysis [81], decomposition of the flux distribution into a minimal number of elementary flux modes [82], random sampling where random subsets of new combinations of modes are selected on the basis of a given probability function [83, 84], or MFA-based optimization [85]. Also, [86, 87] develop an optimization criterion that compromises error, efficiency of the modes, and model size, and the selection of the elementary flux vectors relies on ranking and controlled random search. In the same context, for the purpose of reducing the dimensionality of the solution space and building dynamic models, most of the previous methods incorporate additional biological constraints, such as cell specific uptake and secretion rates, to further reduce the EFM solution space, as specifically studied in [7, 13] and illustrated in Fig. 2.12. Methods using dynamic metabolic flux analysis (DMFA), flux balance analysis (FBA) and dynamic flux balance analysis (DFBA) are also proposed in [15]. Recently, [85, 88] have applied column generation techniques to determine subsets of elementary flux modes in metabolic networks and [89, 90] introduce the poly-pathway model approach to account for the metabolic behavior of multiple experimental conditions.

Another strategy consists in the **reduction of the number of elementary vectors** from a larger initial pool to keep only a small set of representative modes. In connection with this, a study conducted in [91] demonstrates that the number of EFMs can be reduced by projection of the modes into yield space and [92] suggests how to reduce the number of EFMs by grouping them into clusters and computing an average EFM for each one. Hebing et al. [14] use an EFM reduction procedure based on a geometrical collinearity criterion, followed by the definition of a multi-objective genetic algorithm to find a suitable set of EFMs. More recently, our research group has developed several procedures for EFM reduction, i.e., Abbate et al. [93] introduce a method to select the best EFM candidates based on the formulation of a linear optimization problem and Maton et al. [5] develop a reduction methodology based on a combination of several criteria based on collinearity and a series of constrained least-squares problems. The latter work represents the first significant scientific contribution of this thesis and constitutes the initial version of the final reduction algorithm.

In this chapter, this earlier work serves as a foundation for further developments. The methodology is extended by introducing several key improvements to the original algorithm, which now proceeds in four steps. It addresses challenges such as the enumeration of the initial set of elementary flux modes, the differentiate consideration of positivity constraints on the fluxes (see Sec. 3.4.4.4), and the prediction error of reduced macroscopic reaction sets (reduced below the number of measured components). These advances build directly on the initial contribution in [5], while significantly enhancing its scope and applicability.

It is worth emphasizing the distinction between the present reduction methodology and the approach developed by Abbate et al. [93]. In the latter, the number of retained modes was determined through a principal component analysis (PCA), which selects a subset of modes that explain most of the variance in the experimental data. In contrast, the methodology proposed here does not rely on statistical variance reduction. Instead, the number of modes is directly reduced to a desired target value Λ , through an optimization framework that simultaneously computes the time-varying numerical flux signals $\underline{\phi}(t)$. These signals are specifically adjusted to best fit the experimental uptake and excretion rates under positivity constraints. This integrated reduction-identification strategy ensures not only a compact reaction scheme but also consistency with the experimental data at the kinetic level, providing a stronger biological and predictive basis

than a variance-only criterion.

In contrast to recent algorithms such as [85, 88, 90], it is difficult to claim that our algorithm would presumably be more efficient, and we certainly do not want to make this claim. In fact, benchmark comparisons are delicate to achieve in view of the limited access to software packages that implement the various approaches. However, two comments can be made:

- The approaches are different. For instance, the column generation algorithm aims at generating a set of basis vectors, whereas our approach aims at reducing a large initial set of EFMs in order to extract the most relevant vectors;
- The current approach is modular, fast and simple as it combines several pre-existing tools, making the methodology and the code easy to understand and apply for non-expert users.

Furthermore, it is worth noticing that advancements have been made since the development of the hybrid method presented in this chapter. In particular, significant progress has been achieved in the enumeration of elementary flux modes (EFMs), addressing one of the major computational bottlenecks in metabolic network analysis. Notably, the EFM-Ta method, introduced in [94], proposes an efficient approach to extract EFMs by leveraging information from the final tableau of the simplex method. By applying additional elementary steps and eliminating redundant calculations, this method accelerates the enumeration process while maintaining accuracy. More recently, Mores et al. [95] have introduced a method that enables accelerated enumeration of extreme rays through a positive-definite elementary test, which refines the selection of extreme pathways. This technique allows for a more computationally efficient identification of relevant EFMs, further reducing the complexity of large-scale metabolic models. Additionally, [96] introduces a method that uniquely determines EFM weights in unimolecular metabolic networks by applying a Markovian constraint. This approach reformulates the flux decomposition problem into a steady-state analysis of a discrete-time Markov chain, ensuring a unique and biophysically plausible solution. Their algorithm addresses the ambiguity inherent in traditional optimization-based methods, providing a robust framework for EFM weight determination in specific network classes.

A promising perspective for future work would be to explore the potential integration of these advanced algorithms into our hybrid method to enhance its scalability and efficiency. A more detailed discussion of this topic will be provided in the dedicated section of this chapter (see Sec. 3.5).

3.1.2 Kinetic Modeling

What makes the use of elementary flux modes attractive in bioprocess modeling is the possibility to derive minimal bioreaction schemes. Henceforth, it remains to model the kinetics using specific functions (such as Monod or Haldane laws or more complex non-linear functions of the extracellular species) and parameter identification to establish a functional model from sets of experimental data. In the last decades, many studies have addressed the issue of kinetic identification. However, the determination of kinetic structures is often based on arbitrary choices. Indeed, specific kinetic phenomena such as activation, saturation, and inhibition can sometimes be described by different mathematical expressions with similar evolutions with respect to the culture species. Attempts have been made and achieved to propose general kinetic formalisms, as in [97, 98], where power-law equations were used to represent overall reaction rates of complex biological systems [99–104]. Nevertheless, although these methods are effective and have the merit

of being systematic, they do not capture a double component effect. More recently, generalized Monod equations were proposed in [7, 11, 89, 105]. In addition, a systematic procedure was developed in [106] to select the most likely kinetic structures using decision graphs, nested models, and likelihood ratio tests. Recently, another systematic procedure for modeling the kinetics using generalized-kinetic functions and a three-step parameter identification procedure has been proposed and detailed in [107, 108].

All previous approaches are based on kinetic functions that incorporate factors representing various phenomena of activation, inhibition, and saturation, providing some biological interpretation. However, when reaction rates are difficult to describe as they include influences from several culture species, an alternative approach is to resort to purely data-driven techniques such as neural networks [109–111]. Although these models present a high flexibility to capture non-linear kinetic phenomena, they lack biological interpretation.

As will be detailed in the subsequent sections, the application of our EFM reduction algorithm [5, 112] yields, on the one hand, a candidate stoichiometric matrix and, on the other hand, the time evolution of metabolic fluxes corresponding to the selected EFMs. These fluxes often exhibit complex dynamics, as they result from the interplay of multiple phenomena triggered by different culture species. In such cases, a purely mechanistic approach for kinetic modeling becomes challenging. Therefore, the objective of the present study is also to explore the potential of artificial neural networks, and particularly multilayer perceptrons (MLPs), to describe these fluxes. The motivation for using this type of neural network lies in their flexibility, their ability to capture nonlinear dependencies, and the availability of efficient computational toolboxes such as the Deep Learning Toolbox in MATLAB. However, a known drawback of MLPs is their fully connected structure, leading to a rapid increase in the number of parameters as the number of neurons in the hidden layers grows. To mitigate this issue, this study will also explore pruning algorithms to maintain a minimal but effective model structure. Furthermore, the original contribution here lies in the fact that the neural networks are not applied directly to raw input-output data, as in existing macroscopic modeling approaches, but are specifically trained to approximate the time-series fluxes that are computed numerically by solving the optimization problem underlying the EFM reduction method.

The current chapter proposes a modular modeling procedure in which minimal bioreaction models are constructed by selecting a basis of elementary flux modes (EFMs). This process starts with a relatively large set of EFMs, which is then progressively reduced based on geometric criteria and least-squares residual minimization. Since EFM-associated reaction rates are often influenced by multiple activation, inhibition, and saturation effects from different culture species, artificial neural networks provide a powerful framework to model these complex dynamics, resulting in a hybrid dynamic model. However, in certain cases where reaction rates exhibit well-defined functional relationships with extracellular species, classical kinetic laws remain applicable, and parameter estimation can be performed using standard optimization techniques. This flexibility allows the approach to adapt to different levels of complexity, ensuring a balance between mechanistic interpretability and predictive accuracy.

Note that the concept of hybrid modeling, where mechanistic descriptions are combined with data-driven components, was first formalized by [113] in the context of fed-batch bioreactor. Their work demonstrated that coupling neural networks with first-principles models can improve predictive accuracy and extrapolation capabilities. More

recently, recent reviews [114, 115] have summarized the main categories of hybrid models, including serial and parallel combinations of mechanistic and machine learning models, residual (delta) correction strategies, and feature-augmented formulations. These works highlight the current shift toward modular hybrid workflows, underlining the relevance of the methodology developed in this thesis.

3.2 Hybrid Modeling

This section introduces the hybrid modeling methodology and emphasizes the main concepts related to metabolic network analysis, kinetic identification, and artificial neural networks. The hybrid modeling methodology is sketched in Fig. 3.1. Starting from the definition of a metabolic network with the stoichiometric matrix N , the measurement configuration N_e , and the experimentally determined specific uptake and excretion rates $\underline{\nu}_m$, this method first uses the mechanistic information based on the concept of elementary flux modes. For this purpose, an EFM reduction procedure is applied to infer a minimal bioreaction model and its corresponding stoichiometric matrix K . Second, the procedure takes advantage of the time evolution of the specific metabolic fluxes $\underline{\phi}_{num}$, which are provided by the EFM reduction algorithm, to model the kinetics. When flux patterns are straightforward to model, classical kinetic laws are used; otherwise, a data-driven modeling approach is applied, using MLPs with the concentration of extra-cellular measurements $\underline{\xi}$ as input. The model components K and $\underline{\phi}$ can then be used in a predictive model. The complete methodology is described in detail in the sequel.

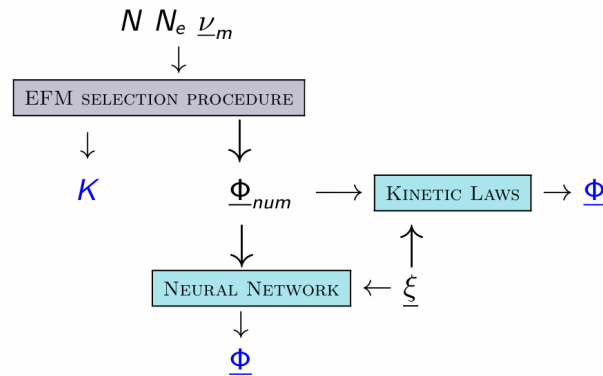


Figure 3.1: Hybrid modeling using an elementary flux mode reduction procedure and neural networks. N is the stoichiometric matrix, N_e is the stoichiometric matrix of extracellular measurements, $\underline{\nu}_m$ is the vector of specific uptake and excretion rates obtained from experimental data, K describes the stoichiometry of the bioreactions, $\underline{\phi}_{num}$ is the vector of specific metabolic fluxes obtained numerically from the reduction procedure, $\underline{\phi}$ is the vector of specific reaction rates modeled using classical kinetic laws or ANNs and $\underline{\xi}$ is the vector of concentrations of culture species.

3.2.1 EFM Selection

3.2.1.1 Theoretical Background

The fundamental concepts underlying the metabolic network analysis have been introduced in Sec. 2.4. As a reminder, a metabolic network is mathematically represented by

its stoichiometric matrix N , which encodes the relationships between internal metabolites and intracellular reactions. Furthermore, another key concept in metabolic network analysis is the notion of elementary flux modes (EFMs), which characterize the minimal, steady-state flux distributions that sustain the metabolic network. These elementary flux modes are assembled in the matrix denoted E .

As illustrated in Fig. 3.1, the EFM reduction procedure yields the matrix K . This matrix is derived from the matrix of elementary flux modes E and is known as the stoichiometric matrix of macroreactions (or the stoichiometric matrix of bioreactions). This matrix facilitates the derivation of a reaction scheme, which is particularly useful for constructing macroscopic models, as discussed in Sec. 3.2.1.3. According to [7], this matrix is obtained as follows:

$$K = N_e E \quad (3.1)$$

Where $N_e^{m_e \times n}$ is the stoichiometric matrix of extracellular measurements and $E^{n \times n_{EFM}}$ is the matrix of EFMs. Consequently, K is an $m_e \times n_{EFM}$ matrix. Notably, the matrix K retains the essential stoichiometric and thermodynamic properties of the original metabolic network.

3.2.1.2 EFM Reduction Procedure

This section focuses on the reduction algorithm, which starts from a usually large number of EFMs to end up into a small set of macroreactions which can be the basis for a macroscopic bioreaction model. The target is therefore a bioreaction scheme with fewer reactions than components, i.e., measured species. Biomass is included in the measured species to ensure that the reduced macroreaction model can be used for process simulation. The complete methodology is depicted in Fig. 3.2.

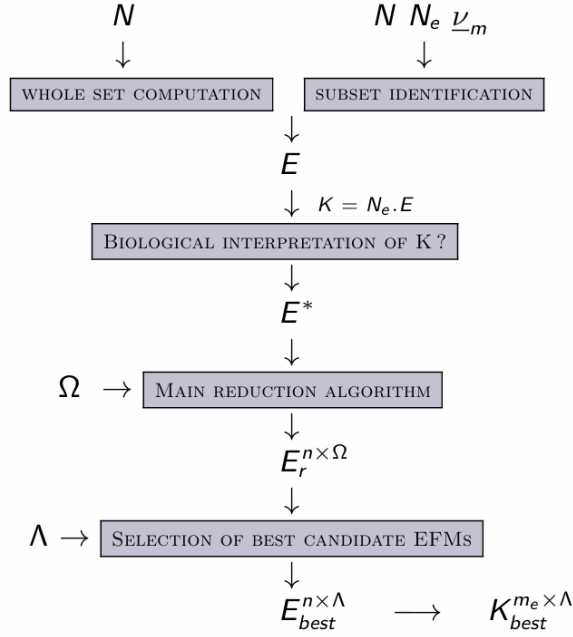


Figure 3.2: The modular EFM selection procedure. N is the stoichiometric matrix, N_e is the stoichiometric matrix of extracellular measurements, ν_m is the vector of specific uptake and excretion rates obtained experimentally, E is the matrix of EFMs and K describes the stoichiometry of the bioreactions. Ω is a first target number for the reduced set of EFMs, close to the number of measured extracellular species m_e , and Λ is a second target number for the reduced set of EFMs, smaller than m_e .

3.2.1.2.1 Generation of EFMs

The first step consists of the generation of the elementary flux modes on the basis of a metabolic network. As mentioned previously, depending on the size of the metabolic network, two cases can be distinguished: on the one hand, when the enumeration of all possible pathways is achievable (i.e., for metabolic networks of modest sizes) and on the other hand, the situation where only subsets of modes can be identified (i.e., when the EFM enumeration becomes computationally prohibitive).

3.2.1.2.1.1 Enumeration When enumeration of all the modes is possible, a matrix of elementary flux modes, denoted E , can be obtained using software tools such as EFMTOOL. The matrix of EFMs is an $n \times n_{EFM}$ matrix where n_{EFM} represents the number of modes. From a geometrical point of view, elementary flux modes represent the edges of a polyhedral cone $\mathbf{S}$ constructed by the intersection of the hyperplanes defined by the rows of the stoichiometric matrix N , as illustrated in Fig. 2.12. For the enumeration of all extreme rays of the polyhedral cone, software's such as EFMTOOL make use of the double description method [25] and requires only the definition of the matrix N as input.

3.2.1.2.1.2 Generation of Subsets When the enumeration of all EFMs becomes impossible or their number difficult to handle, the use of alternative methods for computing reduced sets, such as the ones reviewed in the literature review in Sec.

3.1.1, is required. In our algorithm, we have selected the method proposed in [82], where a fast algorithm is proposed to randomly compute a decomposition of admissible flux vectors in a minimal number of elementary flux modes. In fact, it can be shown that the decomposition of any vector v , inside the polytope $\mathbf{F}$, in the convex basis formed by elementary flux vectors is not unique. The objective of the fast algorithm is to determine one of the minimal decompositions, and to this end, it requires the definition of the stoichiometric matrix N , the matrix of extracellular measurements N_e and the vector of specific excretion and uptake rates $\underline{\nu}_m$. If the algorithm is used only once, a minimal set of EFMs (with a number of EFMs equal to the number of measurements) is directly obtained, which can be further reduced in the following steps of our algorithm to get a number of macroreactions lower than the number of measured species. However, it might be interesting to generate a larger pool of candidate EFMs before proceeding to the reduction, and in this case, the generation algorithm is executed several times until the desired pool is built.

3.2.1.2.2 Biological Interpretation of the EFMs

Although the reduction algorithm is applied to the matrix of elementary flux modes E obtained in the previous step, the stoichiometric matrix of macroreactions K is given by Eq. 3.1. It is an $m_e \times n_{EFM}$ matrix, whose columns provide the stoichiometry of each macroreaction. A biological interpretation of this matrix, e.g., detecting macroreactions with no reactants or no products, might be useful to reduce the matrix to a matrix K^* , corresponding to a reduced matrix of EFMs, denoted E^* .

3.2.1.2.3 Main Reduction of EFMs

The next step consists of drastically reducing the number of modes, with a target number Ω , which is generally set to a number close (maybe slightly greater than) the number of extracellular measurements. This reduction is an iterative procedure based on the following optimization problem:

$$\Xi = \sum_{k=1}^M \left(K_e \underline{\phi}_k - \underline{\nu}_{m,k} \right) W^{-1} \left(K_e \underline{\phi}_k - \underline{\nu}_{m,k} \right)^T$$

$$\min_{\underline{\phi}_k} \Xi \quad s.t. \quad \underline{\phi}_k \geq 0 \quad (3.2)$$

where $\underline{\nu}_{m,k}$ is the vector of specific uptake and excretion rates at every time step k , $\underline{\phi}_k$ corresponds to a time-varying decomposition of the flux $\underline{\nu}_{m,k}$ into a reduced set of vectors stored in K_e , and W is a weighting diagonal matrix whose diagonal elements are given by $\max_i \nu_{m,k,i}^2$ where $\nu_{m,k,i}$ is the i^{th} element of the vector $\underline{\nu}_{m,k}$ ($i \in [1, n_{EFM}]$).

This criterion represents how well the positivity constraints can be satisfied while reproducing the evolution of the extracellular fluxes. Note that $K_e = N_e E_e$, where E_e denotes a reduced matrix of EFMs and it is important to emphasize that $\underline{\nu}_m$ represents the vector of specific rates. As noted in [116, 117], EFMs are often associated with specific reaction rates, as they correspond to independent metabolic pathways in the network.

The index Ξ should be equal to zero when: (i) the number of modes is greater than or equal to the number of extracellular measurements ($n_{EFM} \geq m_e$), (ii) K_e is full row rank, and (iii) there are no constraints on $\underline{\phi}(t)$. However, the positivity constraints can be active in some time periods depending on the compatibility of the measurements

(which are always affected by experimental errors) with the network. The index Ξ therefore indicates how well the positivity constraints are satisfied, and the minimization of the index selects the flux values that achieve the best constraint satisfaction. More information is provided in [112] and an illustration is given in Fig. 3.26 (see Sec. 3.4.4.4).

The reduction of EFMs can be split into two steps: (i) the first is a pre-filtering of the set of modes relying on the concept of cosine-similarity introduced in [14] and (ii) the second step is based on the optimization problem in Eq. 3.2 and includes several checks where a user-defined tolerance value guides the selection.

3.2.1.2.3.1 Cosine-Similarity Algorithm This algorithm involves computing the collinearity between the elementary flux vectors and removing the collinear modes. In other words, if the cosine of the angle between two extreme rays is greater than a defined threshold $\chi \in [0,1]$, then the two elementary flux vectors point in the same direction, and one of them can be removed. The reduction is a compromise between the value of the indicator Ξ and the number of EFMs. Finally, the remaining modes are stored in a reduced matrix E_e and the optimization problem in Eq. 3.2 can be formulated.

In our original work, a numerical method is developed to find the best compromise between the value of the indicator Ξ and the number of EFMs. To maintain clarity and conciseness, the detailed derivation of this method is not reproduced here. Interested readers are referred to our original publication for a comprehensive exposition of the methodology [5].

3.2.1.2.3.2 Optimization-Based Reduction Although the previous prefiltering step allows for a significant cut in the number of modes, especially when starting from the complete set of elementary flux vectors, the cosine criterion alone does not allow for a reduction to the target number Ω . For this purpose, an optimization-based reduction is proposed, where a randomly selected vector could be removed if the corresponding reduced elementary flux vector collection in E_e satisfies the following condition:

$$|\Xi^* - \Xi| < tol \quad (3.3)$$

Otherwise, the selected EFM is kept. In Eq. 3.3, Ξ^* is the performance index for the candidate elimination, whereas Ξ is the prior value of the performance index (prior to the random reduction), and tol is a tolerance value defined by the user.

Since this reduction is based on a random selection of the modes, the reduced set with Ω EFMs is not unique and may not be the best over all combinations, but it achieves nevertheless an acceptable satisfaction of the positivity constraints.

3.2.1.2.4 Generation of the Final Macroreaction Set

The final step of the algorithm involves reducing the number of EFMs below the number of extracellular measurements from the previous set containing Ω modes. To achieve this objective, the choice of a target number Λ is necessary ($\Lambda < m_e$). This reduction is still based on the value of the indicator Ξ (i.e., based on the optimization problem in Eq. 3.2), which however has now another interpretation and can be seen as a sum of squared residuals measuring the distance to the experimental data. Indeed, when the number of EFMs is smaller than m_e , the non-zero value of the indicator corresponds to the combined impact of the positivity constraints and of the fitting of the experimental measurements with a model which now has a reduced number of degrees of freedom. As a matter of fact, with a set of EFMs smaller than the number of measured species, it is

no longer possible to explain all the variance in the data, but a satisfactory model can be obtained in the classical least-squares sense. More information is given in [112].

As mentioned in the previous section, the set of retained modes is not unique, leading to equivalent models. Indeed, as explained in [118], even when the number of EFMs is limited (either initially or after reduction), the resulting set of bioreactions is usually redundant for the design of dynamic models that fully explain the available experimental data. There exists a set of equivalent bioreaction models because they all provide exactly the same values of observed uptake and production rates. Our reduction procedure aims at finding a minimal set of bioreactions, but it is important to keep in mind that this set is not unique, both regarding the number of bioreactions (a number of EFMs equal to the number of measured species allows an exact representation of the fluxes, and a smaller number allows an approximate representation) and regarding the equivalent selections of EFMs for a given number of them. In terms of input-output behavior, these equivalent models will, therefore, lead to equivalent control options. Some of the models might, however, be preferable in terms of biological interpretability and exploitation, as discussed in Sec. 3.2.1.2.2.

For the sake of clarity, it is worth noting that the optimization-based reduction in Sec. 3.2.1.2.3.2 and the selection of the final set of macroreactions in Sec. 3.2.1.2.4 are different. While the first is a reduction based on several random successive eliminations of one mode so as to satisfy Eq. 3.3, the other selects the best combination of λ EFMs among the reduced set of Ω modes according to the smallest value of the performance index. This last step no longer relies on random eliminations of modes, but on the selection of the best subset of EFMs. Therefore, the final set obtained in this way could differ from directly choosing Ω equal to Λ in the previous step.

Furthermore, both steps are necessary as the enumeration of all EFM combinations may become computationally prohibitive, their number being given by:

$$n_{comb} = \frac{\Omega!}{\Lambda! (\Omega - \Lambda)!} \quad (3.4)$$

Hence the importance of reducing first the number of modes to a small value Ω before going through the final step.

3.2.1.3 Dynamic Mass Balance Model

From the reduction procedure presented in Sec. 3.2.1.2, a stoichiometric matrix of macroreactions K is obtained, as well as the time evolution of the specific metabolic fluxes (equivalent to the specific reaction rates) $\Phi(t)$ (see Fig. 3.1). The product of these two quantities defines the biological uptake and excretion rates of the culture species $\underline{\xi}$ in the following mass balance equation system:

$$\frac{d\underline{\xi}}{dt} = K \Phi X + D(\underline{\xi}_{in} - \underline{\xi}) \quad (3.5)$$

In the previous equation, $\underline{\xi}$ is the vector of measured extracellular species concentrations (normalized to volume of bioreactor), D represents a diagonal dilution matrix and $\underline{\xi}_{in}$ denotes the vector of inflow concentrations. X represents the biomass concentration. Note that in Eq. 3.5, the model is expressed in terms of the specific reaction rates Φ , consistent with the fluxes derived from the reduction procedure. Nevertheless, it is worth noting that there exists an alternative formalism that directly models the total reaction rates. This alternative representation may be preferred in some contexts and

will be further discussed in Sec. 3.3.4.2.

To develop a predictive model, it is now necessary to represent the flux vector Φ using a parametric model describing the influence of the extracellular culture species. Depending on the complexity of the flux dynamics and their nonlinear behavior, classical kinetic laws or artificial neural networks can be used.

3.2.2 Kinetic Identification

Kinetic identification is a critical step in the development of mathematical models for bioprocesses. The accuracy of kinetic modeling directly impacts the predictive capability of the model and consequently the optimization and control of bioprocesses. Two main approaches are commonly employed for modeling reaction rates: (i) mechanistic approaches based on classical kinetic laws, such as Monod, Michaelis-Menten, and Haldane models, and (ii) data-driven approaches, such as artificial neural networks (ANNs), which provide greater flexibility when dealing with highly complex systems. Regardless of the approach, estimating the kinetic parameters is a fundamental step in the identification process, which can be performed using linear or nonlinear techniques, as introduced in Sec. 2.7.

Kinetic identification involves both model structure selection and parameter estimation, and these two steps are inherently interconnected, as depicted in Fig. 3.3. The process is often iterative, meaning that initial assumptions about the kinetic model may need to be refined based on the results of parameter estimation. As a matter of fact, parameter estimation is not a standalone process but is tied to the chosen model framework.

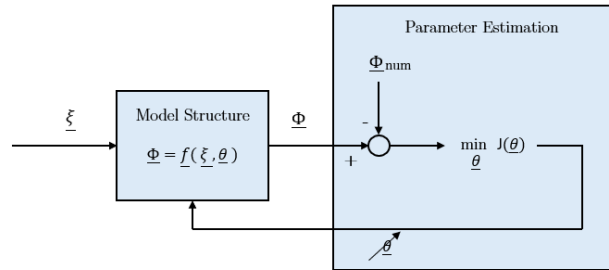


Figure 3.3: Kinetic identification procedure. ξ is the concentration vector of measured extracellular species, ϕ is the vector of reaction rates predicted by the kinetic model, ϕ_{num} is the vector of reaction rates to model (i.e., the rates obtained from the EFM reduction procedure) and θ is the vector of parameters of the model.

3.2.2.1 Model Structure

As mentioned above, the first step in kinetic identification, prior to parameter estimation, is to define the structure of the model that describes reaction rates. In macroscopic bioprocess modeling, reaction rate expressions are typically formulated as functions of measurable extracellular culture components, as these provide practical and experimentally accessible data for model calibration and validation. These chemical components generally fall into three main categories: **substrates**, which serve as primary sources of nutrients and energy for microbial and enzymatic activity; **products**, which include desired metabolites or by-products of the bioprocess; and **biomass**, which represents

the growth and proliferation of microbial or cellular populations. As explained in [119], bioprocess engineering models utilize biomass, substrate, and product concentrations to predict the behavior of the system at the reactor scale.

Selecting an appropriate model is crucial because it dictates the number and type of parameters that need to be estimated. A poorly chosen model structure can lead to inaccurate predictions, even if parameter estimation is performed optimally. In addition, the complexity of the model should be balanced. Indeed, overly simple models may fail to capture key system dynamics, while overly complex models may introduce excessive uncertainty in parameter estimation. The selection of a model structure is often an iterative process in which different kinetic formulations are tested, and their performance is evaluated based on goodness-of-fit metrics and predictive capability.

3.2.2.1.1 Mechanistic and Empirical Models

These models are derived from fundamental biochemical and enzymatic principles. Common kinetic formulations include the Monod, Michaelis-Menten, and Haldane equations, which describe microbial growth, enzyme kinetics, and substrate inhibition, respectively. Mechanistic and empirical models are advantageous when prior knowledge of the system's behavior is available, as they provide interpretable parameters with clear physical meanings.

3.2.2.1.1.1 Simple Kinetic Laws The **Michaelis-Menten** kinetics [120] is a fundamental model for enzyme kinetics, describing the rate of enzymatic reactions as a function of substrate concentration:

$$\phi = \phi_{max} \frac{S}{S + K_m} \quad (3.6)$$

where ϕ is the reaction rate, ϕ_{max} is the maximum reaction rate, S is the substrate concentration, and K_m is the Michaelis constant.

The **Monod** model [121] describes microbial growth as a function of substrate concentration, analogous to the Michaelis-Menten equation but applied to the biomass growth:

$$\mu = \mu_{max} \frac{S}{S + K_s} \quad (3.7)$$

where μ is the specific growth rate, μ_{max} is the maximum specific growth rate, S is the substrate concentration, and K_s is the half-saturation constant.

The **Haldane** model [122] extends Michaelis-Menten kinetics to account for substrate inhibition effects:

$$\phi = \phi_{max} \frac{S}{S + K_m + \frac{S^2}{K_i}} \quad (3.8)$$

where K_i is the inhibition constant.

3.2.2.1.1.2 Generalized Kinetic Functions The systematic methodology developed in [108] provides a structured framework for the selection and estimation of kinetic models in bioprocess engineering. A detailed explanation of the method is provided in the Appendix A.1, where its use in the context of this study is also discussed.

3.2.2.1.2 Data-Driven Models

When the system is highly complex, or when explicit kinetic mechanisms are unknown or difficult to derive, data-driven models can be employed. These models rely on experimental data to infer relationships between input and output variables without requiring explicit mechanistic assumptions. They encompass a broad spectrum of methods beyond artificial neural networks (ANNs). For instance, sparse identification techniques have been successfully applied to elucidate macroscopic stoichiometry and bioprocess kinetics [123, 124], while implicit sparse identification approaches have addressed the determination of reaction numbers and stoichiometry. Other strategies include matrix factorization methods, such as data-driven linear predictors based on maximum-likelihood nonnegative matrix decomposition for batch cultures of hybridoma cells [125], or low-rank matrix approximation approaches for the inference of bioprocess models [126]. These works, many of which have been developed in our department, illustrate the diversity and growing relevance of data-driven methodologies in bioprocess modeling.

Although flexible and capable of capturing nonlinear behaviors, data-driven approaches require large and high-quality datasets to ensure robustness and avoid overfitting. In the present study, we focus specifically on artificial neural networks, and in particular multilayer perceptrons (MLPs), due to their ability to approximate highly nonlinear mappings and capture complex kinetic behaviors that were not satisfactorily represented by classical kinetic laws or generalized Monod-type expressions.

3.2.2.1.2.1 Multilayer Perceptrons As discussed in Sec. 2.6, the single-layer perceptron (SLP) is a fundamental building block of artificial neural networks (ANNs). Although SLPs can approximate linearly separable functions, they are inherently limited in their ability to model nonlinear relationships, which are often characteristic of reaction kinetics. To overcome this limitation, multilayer perceptrons (MLPs) have been developed as an extension of the SLP architecture. Multilayer perceptrons are organized as follows:

- **One input layer**, which distributes the input values to the first hidden layer. This layer represents the independent variables (e.g., species concentrations in our case study);
- **One or several hidden layers**, composed of neurons that apply weighted transformations followed by nonlinear activation functions;
- **One output layer**, which recovers the output of each neuron of the last hidden layer. For example, it produces the reaction rates predicted by the neural structure in our case study.

Mathematically, the transformation from the input layer to the hidden layer is given as follows:

$$h_j = f\left(\sum_{i=1}^n w_{ji} x_i + b_j\right), \quad j \in [1 ; Z] \quad (3.9)$$

where h_j is the activation of the j^{th} hidden neuron, w_{ji} represents the weight from input neuron i to hidden neuron j , b_j is the bias term for hidden neuron j and $f(\cdot)$ is the activation function applied to the hidden neurons. Then, the output layer computes:

$$y_k = g\left(\sum_{j=1}^Z v_{kj} h_j + c_k\right), \quad k \in [1 ; m] \quad (3.10)$$

where y_k is the output, v_{kj} represents the weight from hidden neuron j to output neuron k , c_k is the bias term for output neuron k and $g(\cdot)$ is the activation function for the output layer. In the previous equations, the activation function $f(\cdot)$ is typically nonlinear (e.g., ReLU, sigmoid, or tanh), allowing the network to capture complex input-output relationships. The output activation function $g(\cdot)$ depends on the nature of the problem (e.g., linear for regression or softmax for classification).

The general structure of a multilayer perceptron with one hidden layer is depicted in Fig. 3.4. It is worth noting that the number of hidden layers constitutes the true computational engine of the MLP, and the more hidden layers, the more powerful the artificial neural network, but at the expense of a large number of parameters and the risk of overparameterization. However, as highlighted in [127], even one hidden-layer MLP can approximate the mapping of any continuous function.

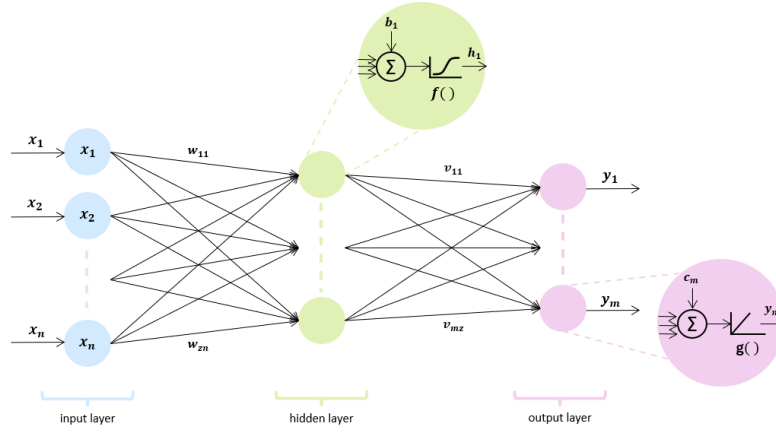


Figure 3.4: A one hidden-layer MLP with Z neurons, n input signals in the input layer and m output signals in the output layer. The nonlinear activation functions in the hidden layer are sigmoid functions, while activation functions in the output layer are identity functions.

As explained in Sec. 2.6, the multilayer perceptron is known as a feedforward network, as data flow in the forward direction from the input layer to the output layer. Neurons can be trained with a backpropagation learning algorithm that computes the gradient of a loss function. Basically, perceptron learning involves adjusting the weights and biases of the network to minimize the deviations between the outputs predicted by the network and the expected ones. For MLP training, backpropagation is an efficient algorithm for estimating gradients, which can be exploited in various methods, such as gradient descent or stochastic gradient descent, updating the parameters of the structure to minimize the loss function. More information on learning algorithms can be found in [128].

3.2.2.1.2.2 Pre- and Post-Processing Steps Before training the neural network, it is common practice to perform pre- and post-processing steps on the input and output of the network in order to make the training process more efficient. Depending on the magnitude of the input values, the sigmoid activation functions may become saturated, leading to small gradients and slow convergence of the training procedure. To tackle this issue, a normalization step is applied to input signals in the data set during

a pre-processing step, and the output signals can be reversely transformed back into the original range in a post-processing step.

3.2.2.1.2.3 Pruning Algorithms MLPs are fully interconnected and the number of parameters rapidly increases with the number of hidden layers and neurons in these layers. In this regard, pruning algorithms remove redundant and unnecessary connections [129–133]. By removing synaptic connections with little influence, a sparsely connected network with multiple zero weights can be obtained. Methods for pruning MLPs are proposed in [134]. Fig. 3.5 illustrates the architecture of a network before and after different pruning techniques.

In this study, a simplified version of the magnitude pruning algorithm proposed in [131] is developed. The idea is to assign a score to each parameter of the network (weights and biases) corresponding to its absolute value. Due to the pre- and post-processing steps, the score of each parameter is assumed to represent its relative importance to the accuracy of the trained network. Hence, a synaptic connection or bias may be removed if its score is smaller than a threshold value defined by the user.

3.2.2.2 Parameter Estimation

Parameter estimation is essential for kinetic identification, as it allows for the determination of unknown kinetic coefficients from experimental data. When the model structure is established, the next step is to estimate the parameters that best fit the experimental data. This process involves defining a cost function, typically based on minimizing the difference between actual and predicted reaction rates, and applying optimization techniques to obtain parameter values (see Sec. 2.7).

Parameter estimation is relevant to both mechanistic and data-driven kinetic models. For classical kinetic laws, parameters such as reaction rate constants and affinity coefficients are fitted using experimental data. In data-driven models, including ANNs, parameter estimation involves optimizing the network weights and biases to minimize error in prediction. In both cases, robust estimation techniques, adequate experimental design, and sensitivity analysis play crucial roles in ensuring reliable parameter identification [135].

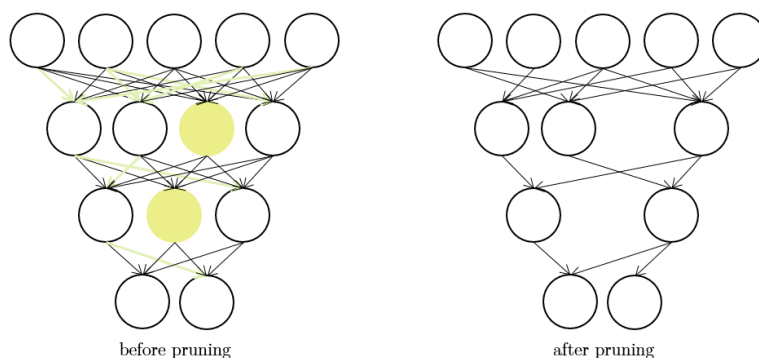


Figure 3.5: Architecture of a multilayer perceptron before and after the application of pruning algorithms – the synapse pruning technique is shown in green and the neuron pruning technique in yellow.

3.3 Application to Hybridoma Cell Line HB-58

3.3.1 Cell Line and Media

Hybridoma cell line HB-58 is a specialized eukaryotic cell line derived from animal cells, specifically used in biotechnological and biomedical applications. Hybridoma cells are a fusion of B-lymphocytes (antibody-producing immune cells) and myeloma (cancer) cells, allowing them to proliferate indefinitely while maintaining the ability to produce monoclonal antibodies. In this application, data come from the Hybridoma cell line HB-58 enabling the production of IgG1 monoclonal antibodies, specific for the mouse kappa light chain. Serum-free medium – prepared from DMEM/F12 (1:1) and completed with glutamine, 500 μmol of ethanolamine, 10 mg of bovine insulin, cholesterol, Pluronic F-68, HEPES and other additives – was used. The cells were kept as suspension cultures in the medium in shake flasks at 37 °C in a 5% CO₂ incubator. Experiments were conducted at the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology (ECUST), in Shanghai [12].

Perfusion cultures were performed in a 2 L stirred bioreactor and were settled in a working volume of 1.8 L during the whole duration of the culture. Cultures were carried out in a controlled environment (36.8 °C, 40% DO, pH = 7.0 $\pm$ 0.2). The perfusion phase started after 44–56 h of batch culture with a constant dilution rate of 0.0197 h⁻¹. Cells were retained by a spin-filter (20 μm) and the stirring speed was fixed at 200 rpm. Data acquisition and process control were performed using the supervisory software MFCS/Win 3.0. Table 3.1 summarizes the operating conditions of the cultures.

Table 3.1: Summary of culture conditions for experiments in perfusion.

Experiment	X_0 (10 ⁹ cells/L)	t_{perf} (h)	Feed Stream (mM)			
			Glc_{in}	Gln_{in}	Ala_{in}	...
Cult. 1	0.19	54	11	5	0.5	
Cult. 2	0.23	56	15	11.5	0.1	
Cult. 3	0.36	48	28	4	0.3	
Cult. 4	0.36	44	28	9.5	0.05	

3.3.2 Metabolic Network

Hybridomas have a complex metabolism. Depending on the level of description required, networks of different sizes might be considered. In this application, a metabolic network of hybridoma cells with 70 biochemical reactions and 44 internal metabolites is considered. The details of this metabolic network can be found in [136] and the network is depicted in the Appendix B. It includes the major reactions of central metabolism, such as the pathways of glycolysis, the tricarboxylic acid cycle, amino acid metabolism, and the synthesis of biomass and antibodies. The pentose phosphate pathway is not taken into account in the network definition because, in most tissues, 80 to 90% of glucose oxidation is achieved by glycolysis. The stoichiometric coefficients for biomass and antibody synthesis are taken from the literature [12]. Note that considering a metabolic reaction for the formation of biomass allows for its prediction when deriving dynamical models.

3.3.3 Measurement Configuration

In addition to the definition of the metabolic network and its corresponding stoichiometric matrix N , the methodology described in Sec. 3.2 and depicted in Fig. 3.1 requires the definition of the stoichiometric matrix of extracellular measurements N_e . Although various measurement scenarios can be considered, this study focuses on a set of six measured components – namely, glucose, glutamine, ammonia, lactate, alanine and biomass – for the sake of simplicity. These concentration measurements are then collected in vector $\underline{\xi}$. While only the first five are extracellular metabolites, biomass is also included in the measurement configuration. The rationale is methodological: by incorporating biomass into the matrix N_e , the EFM reduction procedure yields a macroscopic reaction scheme in which biomass formation appears explicitly. This is essential for building a predictive model of both metabolite dynamics and cell growth. Thus, the term *extracellular measurements* should here be understood more broadly as the set of measured components available for model calibration, encompassing both extracellular metabolites and biomass concentration.

The selection of glucose, glutamine, ammonia, lactate, alanine, and biomass as primary measurements is based on their pivotal roles in central carbon and nitrogen metabolism, as well as their significance in bioprocess monitoring. These metabolites provide essential information on cellular physiology, substrate utilization, and metabolic by-products, which are fundamental for accurate macroscopic modeling.

- **Glucose** - As the principal carbon source in many bioprocesses, glucose uptake is a key metabolite of cell growth. Monitoring glucose concentration provides insights into glycolytic activity, energy generation, and the balance between catabolic and anabolic pathways [19];
- **Glutamine** - This amino acid is a critical nitrogen donor and anaplerotic substrate, feeding into the TCA cycle through glutamate and α -ketoglutarate. Its consumption rate is related to cell growth and the biosynthesis of nucleotides and proteins [137];
- **Ammonia** - As a major byproduct of glutamine metabolism, ammonia accumulation can lead to metabolic inhibition and altered cellular physiology. Monitoring ammonia allows for the assessment of nitrogen metabolism efficiency and potential toxic effects, which are critical for optimizing cell culture conditions;
- **Lactate** - A key fermentative byproduct, lactate accumulation is indicative of metabolic overflow, particularly under conditions of oxygen limitation or excess glucose supply, as discussed in Sec. 3.3.5;
- **Alanine** - This amino acid is a key metabolite in transamination reactions and is an indicator of the interplay between glycolysis and amino acid metabolism. Alanine production is often related to pyruvate metabolism and nitrogen shuttling [138];
- **Biomass** - The overall biomass concentration is a fundamental measure of cellular growth and viability and serves as a crucial output variable in macroscopic models.

By restricting the measurement scenario to these six extracellular metabolites, the model remains manageable and easy to interpret. Although additional measurements, such as the full spectrum of extracellular amino acids, could provide a more detailed and precise model, they significantly increase model complexity and parameterization. The

effectiveness of the EFM reduction procedure presented in Sec. 3.2.1.2 has been thoroughly demonstrated through its application to extended measurement scenarios, including cases with more than 20 extracellular amino acids. These case studies, which show that the method yields significant reductions, are detailed in our published papers [5, 112, 139].

Therefore, given the measurement scenario, the stoichiometric matrix of extracellular measurements N_e can be defined accordingly. Finally, the specific uptake and excretion rates $\underline{\nu}_m(t)$ can be calculated using smoothing splines and differentiation methods from experimental concentration measurements. Appendix D.1 provides the complete mathematical development underlying the computation of the specific reaction rates from the experimental data. In addition, Appendix D.3 discusses how confidence intervals on these estimated rates could be computed, for instance by bootstrap resampling of the spline fits.

Fig. 3.6 shows the time evolution of the experimental uptake and excretion rates for datasets 1, 2 and 4 (used for direct validation). Dataset 3 is not considered at this stage as it is reserved for cross-validation. Cross-validation ensures that the model is evaluated on unseen data, preventing overfitting and assessing its generalization ability. This allows us to validate the robustness of the modeling approach in later stages of the study.

3.3.4 Numerical Results

3.3.4.1 EFM Reduction Procedure

The first step of the reduction procedure involves computing an initial set of elementary flux vectors. As stated in Sec. 3.2.1.2, different approaches exist depending on the size of the metabolic network. In this case, using EFMTOOL as the EFM generator leads to an initial set of 22,563 modes. Alternative methods can be used, such as the one proposed in [82] which identifies only representative subsets of EFMs and allows obtaining a few hundred modes as a starting point. However, in order to prove the effectiveness of the reduction procedure, the whole set of elementary vectors is considered in the present work, since their number remains manageable. Note that these elementary modes represent the edges of the pointed polyhedral cone $\mathbf{S}$ depicted in Fig. 2.12. Then, to ensure a biological interpretation of the reaction scheme, any vector leading to a macroreaction with no biological meaning is discarded and the reduced matrix of modes E^* is made up of 21,874 vectors. The power of the methodology lies in the main reduction algorithm. With a collinearity indicator of 99% between the vectors, a drastic reduction is achieved to 214 modes. Next, an optimization-based reduction is performed with a target number Ω equal to 10. The last step is the selection of a few modes from the previous set to derive models with fewer reactions than components. In this study, only five elementary flux modes are finally selected ($\Lambda = 5$). Note that the target Λ could have been set to an even smaller value. However, because overflow metabolism is noticed in hybridoma cells, a further reduction may cause the loss of essential biological information.

A first direct validation of the reduced model can be achieved by comparing the experimental uptake and excretion rates $\underline{\nu}_m(t)$ to the product $K_e\Phi(t)$ computed in the optimization problem (with K_e being the corresponding stoichiometric matrix of macroreactions after reduction). The results of the prefiltering step and the optimization-based reduction are illustrated in Fig. 3.7 and Fig. 3.8, respectively. A quantitative assessment of these steps is provided in the Appendix F (in Table F.1 and Table F.2), where the corresponding performance index Ξ is reported. This validation can be pursued by

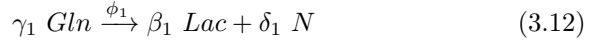
analyzing the prediction of the measured concentrations by integration of the computed fluxes and identification of the most likely initial conditions of the measured species. The predicted measured concentrations are shown in Fig. 3.9 and Fig. 3.10. Almost no loss of information is noticed (especially until the reduction up to Ω modes) and the prediction of concentrations is very satisfactory, highlighting the merits of the modular EFM selection procedure. A quantitative summary of this concentration-based validation is provided in Table F.3, which lists the root mean square errors (RMSEs) for each dataset and globally.

A further improvement in the graphical presentation of the results would be to include confidence intervals on the experimental data points. This requires prior knowledge of the measurement uncertainty, typically quantified by the standard deviation of the experimental assays. Such a representation would provide a more rigorous visual assessment of the agreement between the model predictions and the experimental data.

Besides the previous validations, it might be interesting to examine the final set of macroreactions obtained from the reduction algorithm:

$$K = \begin{bmatrix} 0 & -\alpha_2 & 0 & -\alpha_4 & -\alpha_5 \\ \beta_1 & 0 & 0 & 0 & \beta_5 \\ -\gamma_1 & -\gamma_2 & -\gamma_3 & -\gamma_4 & 0 \\ \delta_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \epsilon_3 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 \end{bmatrix} \quad (3.11)$$

where α_i , β_i , γ_i , δ_i , ϵ_i and σ_i are stoichiometric coefficients of the bioreactions, listed in Table 3.2. K is an $m_e \times \Lambda$ matrix with $m_e = 6$ and $\Lambda = 5$. Equivalently, a macroreaction scheme can be drawn:



This reaction scheme makes sense and is validated by the study conducted in [140]. Indeed, glucose (Glc) and glutamine (Gln) are consumed with biomass growth and metabolite production, defining the so-called respiratory metabolism. Moreover, the phenomenon of overflow metabolism is pointed out with lactate (Lac) production from glucose excess, while excess glutamine drives deamination leading to the production of ammonia (N), with alanine (Ala) and lactate as associated byproducts. This point will be further discussed in Sec. 3.3.5.

It is worth noting that, under perfusion, the feed stream also supplies amino acids beyond glutamine. Although not explicitly represented in this reduced macroscopic model, these species contribute to additional amino-nitrogen that can be routed into biomass nitrogen and, under C/N imbalance, into ammonium NH_4^+ via deamination.

Table 3.2: The stoichiometric coefficients of the reaction network.

Parameter	Value	Parameter	Value	Parameter	Value
α_2	0.0150	β_5	0.1459	γ_4	0.0194
α_4	0.0258	γ_1	0.2425	δ_1	0.4851
α_5	0.0729	γ_2	0.0092	ϵ_3	0.2408
β_1	0.2425	γ_3	0.1204	σ_4	0.7925

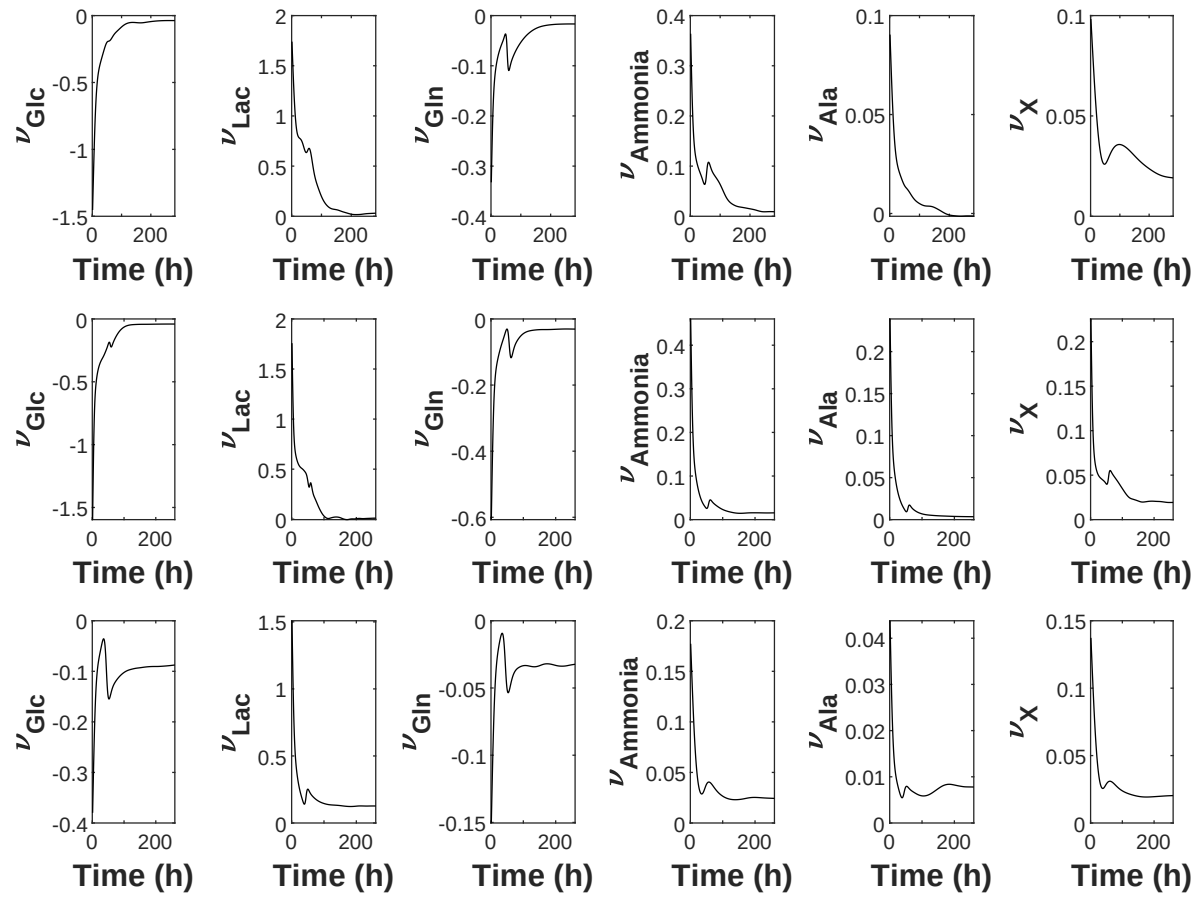


Figure 3.6: Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in $\text{mM}\cdot\text{h}^{-1}$ per 10^9 cells) and ν_X (in h^{-1}) in datasets 1, 2 and 4 (direct validation) computed using smoothing splines and differentiation methods.

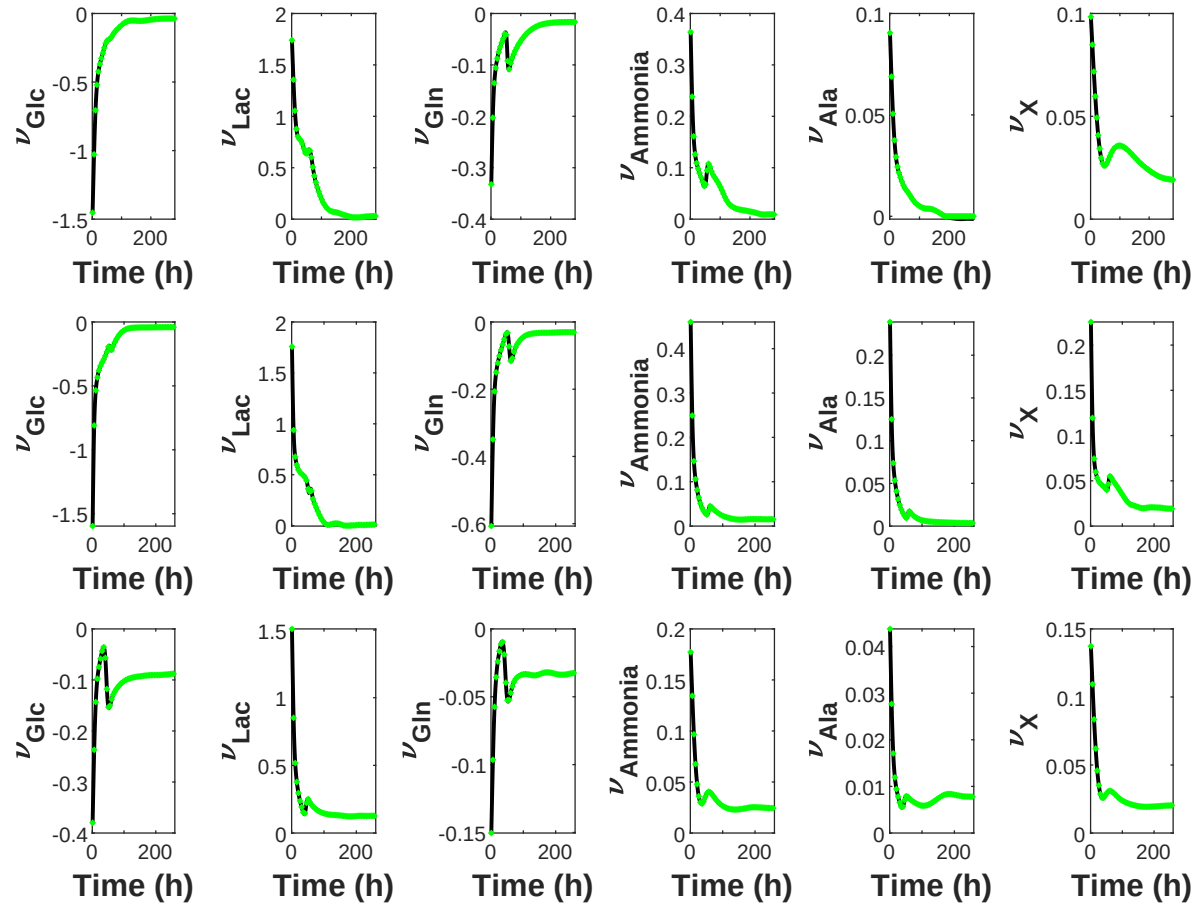


Figure 3.7: Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in $\text{mM}\cdot\text{h}^{-1}$ per 10^9 cells) and ν_X (in h^{-1}) in datasets 1, 2 and 4 (direct validation): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = 214$ (green crosses) – results from the prefiltering step.

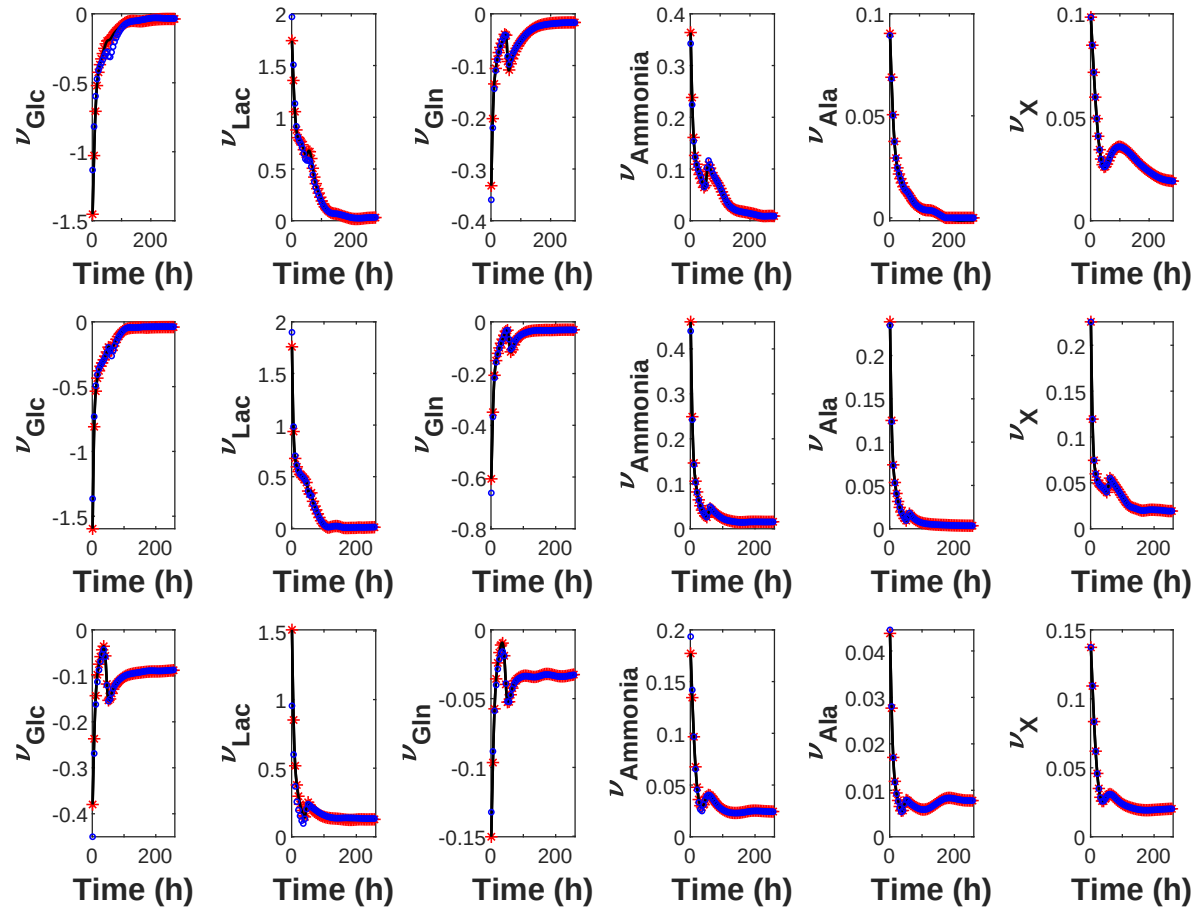


Figure 3.8: Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in $\text{mM} \cdot \text{h}^{-1}$ per 10^9 cells) and ν_X (in h^{-1}) in datasets 1, 2 and 4 (direct validation): numerical results (black curves); $K_e \Phi$ for $n_{EFM} = \Omega = 10$ (red crosses); $K_e \Phi$ for $n_{EFM} = \Lambda = 5$ (blue bullets).

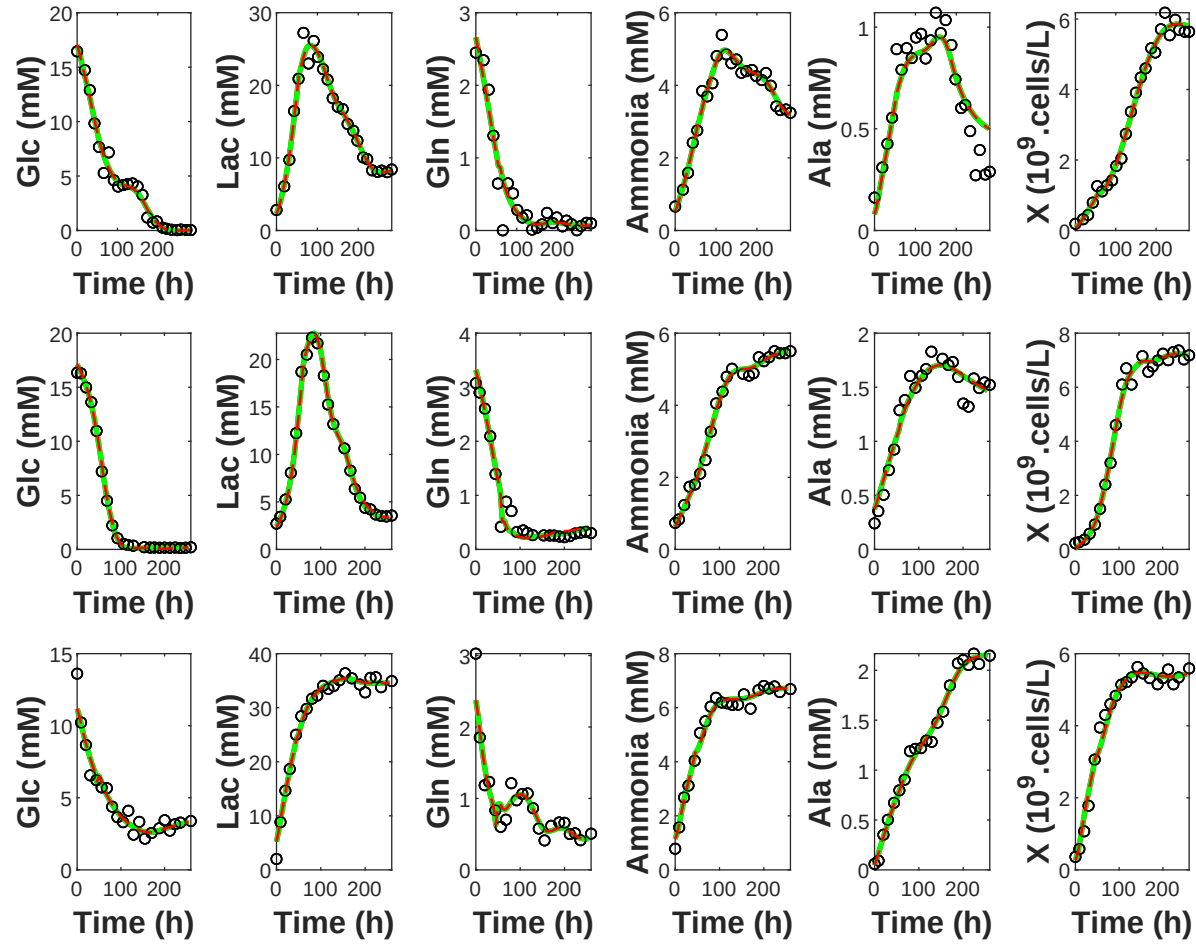


Figure 3.9: Time evolution of the measured concentrations in datasets 1, 2 and 4 (direct validation) – experimental data points (black bullets); concentration profiles on the basis of $n_{EFM} = 214$ (green lines) and reducing to $n_{EFM} = \Omega = 10$ (red dashed lines).

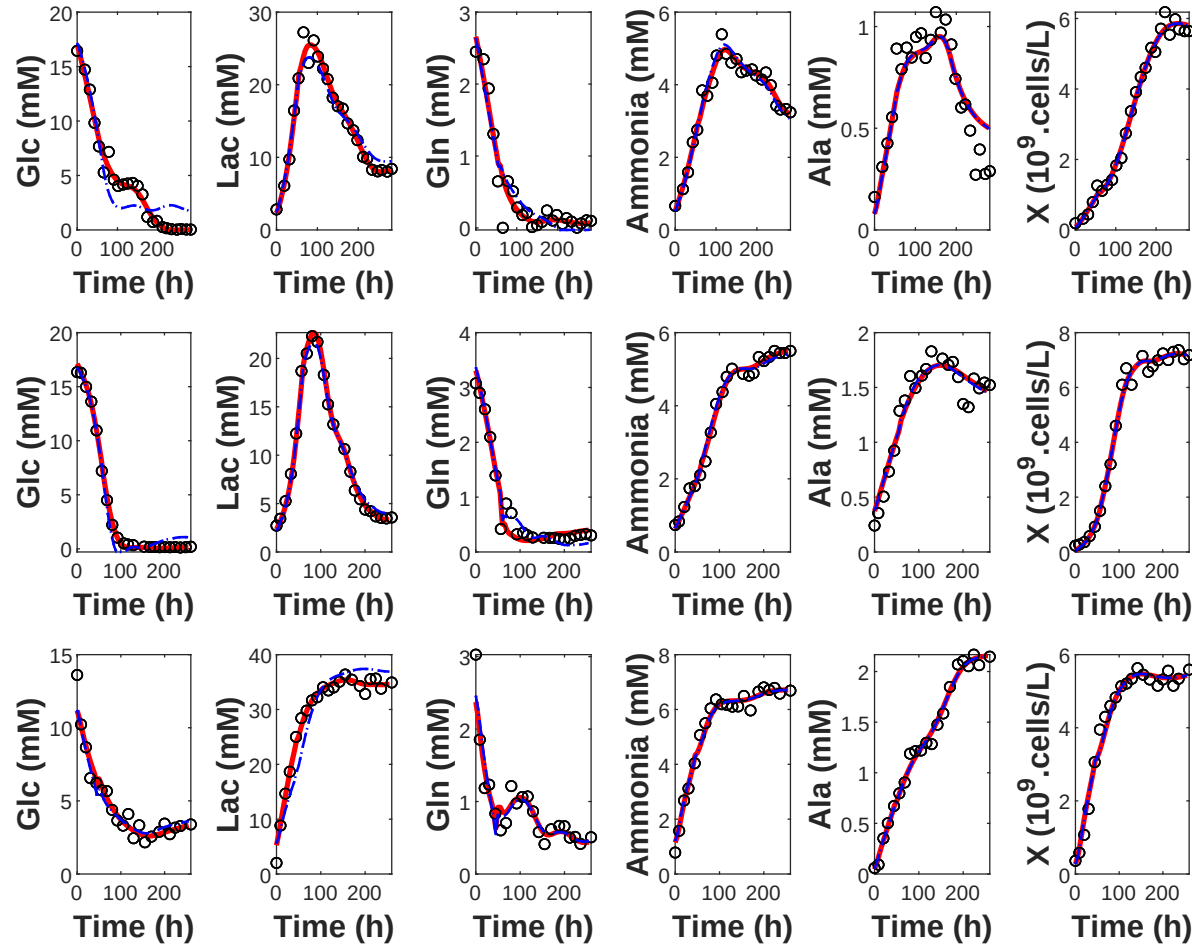


Figure 3.10: Time evolution of the measured concentrations in datasets 1, 2 and 4 (direct validation) – experimental data points (black bullets); concentration profiles on the basis of $n_{EFM} = \Omega = 10$ (red lines) and selecting $n_{EFM} = \Lambda = 5$ (blue dashed lines).

3.3.4.2 Kinetic Modeling

For the derivation of a dynamic macroscopic model, it remains to describe the numerical signals Φ_{num} coming from the EFM reduction algorithm using classical kinetic laws or neural networks.

3.3.4.2.1 Simple Kinetic Laws

As a first attempt to model the kinetics, the use of classical kinetic laws is explored (i.e., Michaelis-Menten, Monod or Haldane kinetics). These classical formulations assume predefined relationships between substrate concentrations and reaction rates, as defined by Eq. 3.6, Eq. 3.7 and Eq. 3.8. However, in our case, these approaches failed to adequately describe the macroscopic reaction kinetics, as they do not sufficiently capture the system's nonlinear behavior and regulatory interactions.

3.3.4.2.2 Generalized Kinetic Functions

Given the limitations of the classical kinetic laws, we then applied the generalized kinetic function identification method developed in [108]. This method is based on the Monod formalism and systematically identifies generalized kinetic rate expressions by optimizing model parameters to fit experimental data. Like classical kinetic laws, this generalized modeling approach also seeks to identify specific reaction rates Φ . Despite its theoretical advantages, this generalized modeling approach also failed to provide an adequate representation of the reaction kinetics. Therefore, the inability to obtain satisfactory results suggests that the kinetic relationships governing the macroscopic reactions obtained from the EFM reduction procedure may be more complex than what can be captured through generalized Monod-like expressions.

3.3.4.2.3 Multilayer Perceptrons

Given the challenges encountered with both classical and generalized kinetic models, a data-driven approach is proposed by using artificial neural networks (ANNs) to model reaction kinetics. The flexibility of ANNs, especially MLPs, allows them to approximate highly nonlinear relationships, making them well-suited for capturing complex metabolic behaviors.

Originally, the EFM reduction procedure is an optimization-based method to derive a minimal reaction scheme through the matrix K and to numerically compute the time evolution of specific reaction rates Φ to match the specific uptake and excretion rates $\underline{\nu}_m$ obtained from the experimental data. The latter rates $\underline{\nu}_m$ are computed by Eq. D.2. Although biomass concentration is interpolated (from experimental data) using cubic or B-smoothing splines, the division by the interpolated biomass could lead to abrupt variations in the rate profiles. Since neural networks prefer smooth and continuous functions, modeling total reaction rates instead of specific rates could be a better choice.

As outlined in Sec. 3.2.1.3, the reaction kinetics can be expressed in two different forms: either using the **specific** reaction rates or the **total** reaction rates. As introduced in Eq. 3.5, Φ represents the specific reaction rates. However, the reaction term can also be expressed in terms of total reaction rates by rearranging the equation as follows:

$$\frac{d\xi}{dt} = K R + D(\xi_{in} - \xi) \quad (3.17)$$

where $R = \Phi \cdot X$ denotes the total reaction rates (back-transformation to the total rates by multiplying the specific rates by the biomass concentration). The key distinction is

that Φ is normalized by biomass, while R represents the actual metabolic fluxes at a given time. Both formulations are mathematically equivalent, and the choice between one formalism or the other depends on the modeling context or whether the kinetics are to be expressed per unit of biomass or as total reaction rates.

Also, it is worth emphasizing that:

1. First, the multiplication by the biomass concentration (interpolated from experimental data) can act as a smoothing filter (if proper and smooth interpolation of biomass);
2. Second, errors in biomass estimation from the ODE system in Eq. 3.5 impact the total reaction rates $\Phi.X$. Therefore, directly modeling the total reaction rates R in Eq. 3.17 avoids unnecessary error propagation;
3. Third, inputs of neural networks are the extracellular concentrations ξ . In this case study, biomass X is measured (i.e., $X \in \xi$), meaning that the network can directly learn the dynamics of the total reaction rates.

To identify the most suitable formulation for ANN training, the smoothness of Φ and R was evaluated using time-series data. Several methods such as variance analysis, Sobel filtering, and Fourier transform analysis were applied. These analyses confirm that the total reaction rates R are generally smoother than the specific reaction rates Φ , thus supporting the decision to model the total rates using MLPs in this context. However, it should be noted that the ANN was trained to predict both the specific reaction rates Φ and the total reaction rates R . On the basis of the results and observations, it was noticed that modeling the total reaction rates yields better predictive performance and stability, likely due to reduced error propagation and smoother target functions.

Furthermore, both the specific reaction rates Φ and the total reaction rates R rely on the interpolated biomass. Therefore, it means that reaction rates remain strongly influenced by how the biomass is interpolated, and any errors or biases in biomass estimation will propagate through the entire identification process. An alternative could be to directly work on the total reaction rates in the reduction procedure, as will be discussed in Sec. 3.5. Nevertheless, since the biomass concentration measurements were interpolated using cubic splines or B-smoothing splines, it is highly likely that this will not significantly affect the results. Besides, after modeling the kinetics with neural networks, a global nonlinear re-identification step is added, which involves simulating the full ODE system and optimizing the parameters to better fit experimental data. This additional step could mitigate error propagation to some extent.

Fig. 3.11 illustrates the multilayer perceptron considered in the context of the present study. The multilayer perceptron uses as input signals the concentration of extracellular species ξ and as target values the total reaction rates. The latter are obtained by multiplying the specific reaction rates Φ computed numerically in the EFM reduction procedure by the biomass concentration interpolated from experimental data. Therefore, there are six input signals, namely, the concentrations ξ_i ($i \in [1, 6]$) of the measured species and five output signals r_k ($k \in [1, 5]$), which are the total rates of each reaction. Only one hidden layer is considered to limit the number of parameters (weights and biases), and the nonlinear activation functions in the hidden layer are hyperbolic tangent sigmoid functions, while the activation functions in the output layer are identity functions. In summary, the following equations can be written as follows:

$$h_j = f\left(\sum_{i=1}^6 w_{ji} \xi_i + b_j\right), \quad j \in [1; Z] \quad (3.18)$$

$$r_k = g\left(\sum_{j=1}^Z v_{kj} h_j + c_k\right), \quad k \in [1; 5] \quad (3.19)$$

with

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (3.20)$$

The function in Eq. 3.20 has the same shape as the classical sigmoid function, but horizontal asymptotes are in -1 and 1 so that the output values are squeezed into $[-1, 1]$. It can be seen as a rescaled version of the logistic sigmoid that improves numerical stability and accelerates convergence during training. In the previous equations, Z is the number of neurons in the hidden layer. Hence, the number of parameters is $5 + 12 \cdot Z$. Depending on the complexity of the problem, the number of neurons in the hidden layer can be increased at the expense of a significant increase in the number of parameters to identify. Table 3.3 summarizes the number of parameters of the neural network according to the number of neurons in the hidden layer. The rapid increase in the number of parameters with the number of neurons justifies the use of pruning algorithms.

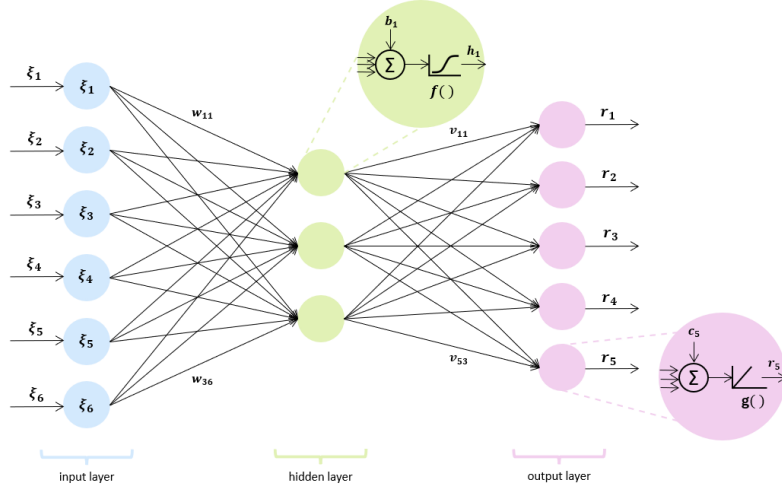


Figure 3.11: A one hidden-layer MLP with 3 neurons, 6 input signals (which are the measured concentrations of culture species) in the input layer and 5 output signals (which are the total reaction rates of each bioreaction) in the output layer. The biases are denoted by b_j in the hidden layer and c_k in the output layer. The nonlinear activation functions in the hidden layer are sigmoid functions, while activation functions in the output layer are identity functions.

Multilayer perceptron training was carried out using the Deep Learning Toolbox in MATLAB, which provides efficient implementations of backpropagation and network optimization routines. In our framework, the input data correspond to the smoothed concentration profiles, while the target outputs are the time-series fluxes computed numerically by solving the optimization problem of the EFM reduction method. Because these trajectories are obtained numerically, they can be evaluated at finer time intervals, making it possible to generate a substantial number of input–output pairs for training.

Table 3.3: Number of NN parameters for 6 input signals and 5 output signals.

# Neurons	# Parameters
10	125
5	65
4	53
3	41
2	29
1	17

In practice, the procedure yielded on the order of 250-300 data points per dataset, providing a sufficient basis to estimate the parameters of the MLPs while limiting the risk of overfitting.

As a direct validation, Fig. 3.12 shows the time evolution of the reaction rates obtained numerically through the reduction procedure (i.e., Φ_{num} signals multiplied by the biomass concentration and used as target values) and the prediction with MLPs with different numbers of neurons in the hidden layer. As expected, the more neurons, the better the fitting to numerical results. The global RMSE values for these identifications are summarized in Table F.4 in the Appendix F, which confirms the monotonic decrease of the error as the number of neurons in the hidden-layer increases. Fig. 3.13 shows the corresponding concentration profiles. Although significant deviations appear in the reaction rates, they have a limited impact on the prediction of the time evolution of the concentrations. A quantitative assessment is provided in Table F.5. Hence, an MLP with only three neurons in the hidden layer is an acceptable compromise between the fitting to the experimental data and the number of network parameters. Note that the experimental data has been partitioned to avoid overfitting, i.e., 70% for training, 15% for validation and 15% for testing. The **training** set is used to optimize model parameters by learning patterns from data; the **validation** set helps tuning hyper-parameters (learning rate, stopping criteria, regularization) and prevent overfitting by assessing model performance during training; and the **testing** set is used only after training is complete to evaluate the model generalization to unseen data, ensuring unbiased performance assessment.

In order to improve the MLP prediction, in addition to backpropagation, a global nonlinear identification of the weights and biases of the network can be performed using the Nelder–Mead simplex optimization algorithm in order to minimize a weighted nonlinear least-squares criterion:

$$\min_{\hat{\underline{\Theta}}} J(\hat{\underline{\Theta}}) = \sum_{j=1}^n \sum_{i=1}^{N_j} \left(R_{ij}(\hat{\underline{\Theta}}) - R_{ij}^{num} \right) \Sigma^{-1} \left(R_{ij}(\hat{\underline{\Theta}}) - R_{ij}^{num} \right)^T \quad (3.21)$$

where $\hat{\underline{\Theta}}$ is the vector of parameters to be re-identified containing the weights and biases of the neural network, R_{ij} gives the network prediction at the i -th time instant in the j -th experiment, R_{ij}^{num} is the vector of the corresponding numerical signals (obtained by multiplying the specific reaction rates Φ_{ij}^{num} , derived from the EFM reduction procedure, by the interpolated biomass concentration X_{ij} from experimental data) and Σ is a normalization matrix in which diagonal elements are chosen as the squares of the

maximum rate of each reaction.

To improve convergence, the optimization algorithm can be run several times using the parameters found in the previous round as initialization. The confidence intervals and variation coefficients of the estimated parameters can be obtained from the inverse Fisher information matrix [141]. The variation coefficients of the estimated parameters are given in Table 3.4.

Next, the magnitude pruning algorithm is exploited to reduce the number of parameters. The results of the pruning algorithm are shown in Table 3.5. W is a matrix that collects the synaptic connections between the input layer and the hidden layer, b represents the biases of the neurons in the hidden layer, V contains the weights between the hidden layer and the output layer, and c are the biases of the neurons in the output layer. In this way, using the magnitude pruning algorithm, the number of parameters is reduced by 17% (41 $\rightarrow$ 34 parameters).

W						b	V			c
5.8	15.5	33.5	20	7.1	77	6.6	47.1	26.5	105.3	58.7
7.1	4.5	10.4	14.5	8.6	8.3	125.7	1	1	1.6	13.4
7.3	31.7	40.4	11.5	5.8	9.5	4.7	73.3	38.2	41.1	710.7
							32.8	50.9	82	169.6
							3.5	2.1	7.7	11.3

Table 3.4: Coefficients of variation (in %) of the estimated weights and biases – colored boxes are the removed synaptic connections or biases after using pruning algorithms.

W						b	V			c

Table 3.5: Results of the magnitude pruning algorithm – colored boxes are the removed synaptic connections or biases.

The time evolution of the reaction rates and the prediction of the measured concentrations using a pruned multilayer perceptron are illustrated in Fig. 3.14 and Fig. 3.15, respectively. Fig. 3.14 shows the added value of a global nonlinear re-identification of the parameters. Small deviations may be observed when the pruned MLP is used for kinetic modeling but the results remain very satisfactory, highlighting the merit of the simplified pruning method. Fig. 3.15 validates the whole procedure and provides good results for the prediction of the measured concentrations. A quantitative summary is given in Table F.6.

Furthermore, a pruning of the neural network can also be achieved on the basis of the coefficients of variation of the estimated parameters. Indeed, weights and biases identified with a coefficient of variation greater than 30% can be discarded without affecting the fitting of the experimental data. However, the elimination of the weights in

V must be exercised with more care and only the parameters identified with a coefficient of variation greater than $\cong 50\%$ are neglected (see Table 3.4). Fig. 3.16 shows the prediction of the time evolution of the extracellular measured species using this pruning strategy. It yields fewer parameters in the neural structure ($41 \rightarrow 30$ parameters) at the expense of larger deviations from experimental data. However, the results remain acceptable for deriving a macroscopic model for control and optimization purposes. Furthermore, the pruning strategy using the value of variation coefficients covers up most of the parameters defined as non-essential when using the magnitude pruning algorithm, validating the two pruning approaches. Finally, it is worth noting that a nonlinear global re-identification of weights and biases of the pruned MLP could still improve the results. Appendix E provides the parameters of the neural network architecture.

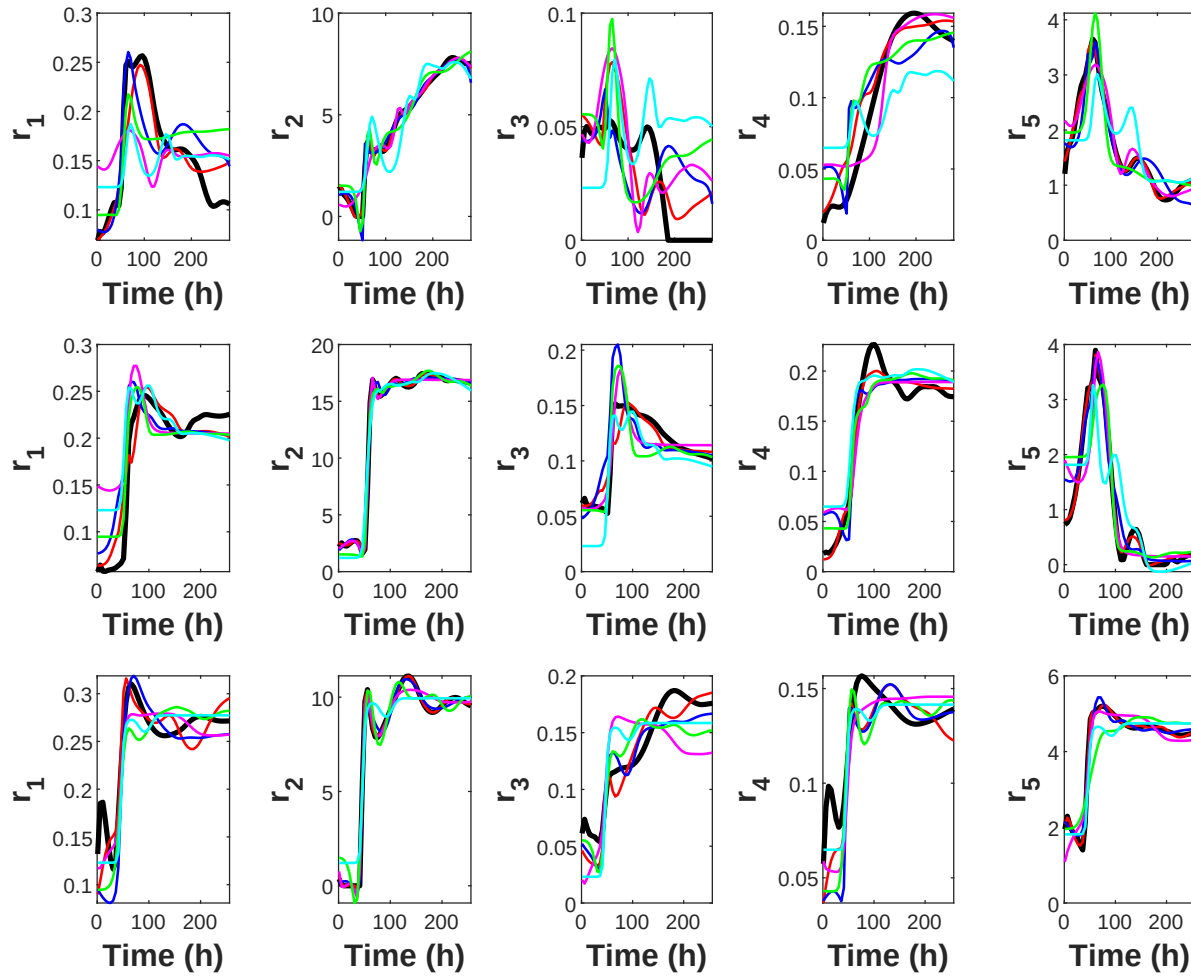


Figure 3.12: Time evolution of the total reaction rates r_i (in mM.h^{-1}) in datasets 1, 2 and 4 (direct validation) – target values, i.e., rates coming from the EFM reduction procedure (black curves); prediction with an MLP with ten neurons (red curves), five neurons (blue curves), four neurons (magenta curves), three neurons (green curves) and two neurons (cyan curves) in the hidden layer.

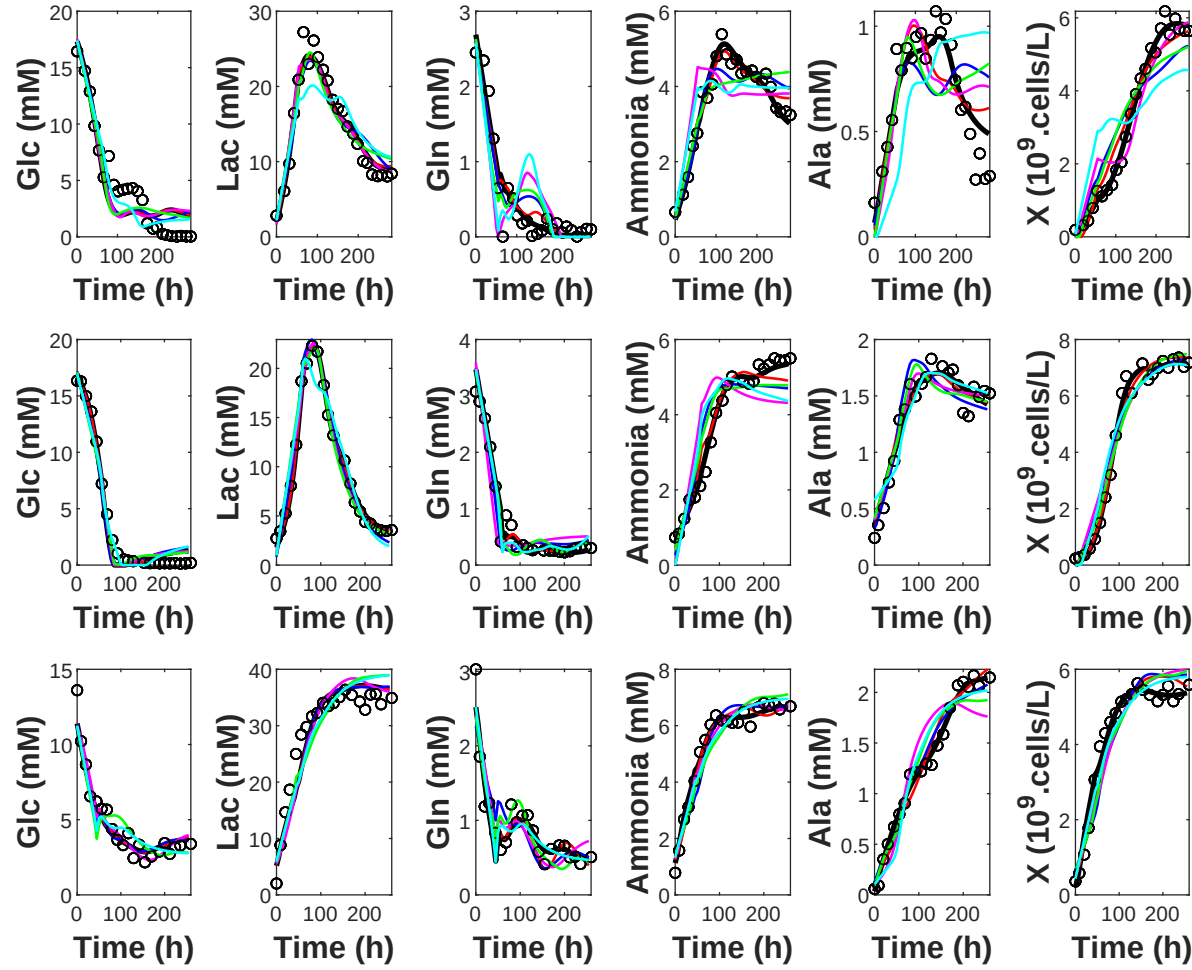


Figure 3.13: Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) — results coming from the EFM reduction procedure (black curves); prediction with an MLP with ten neurons (red curves), five neurons (blue curves), four neurons (magenta curves), three neurons (green curves) and two neurons (cyan curves) in the hidden layer.

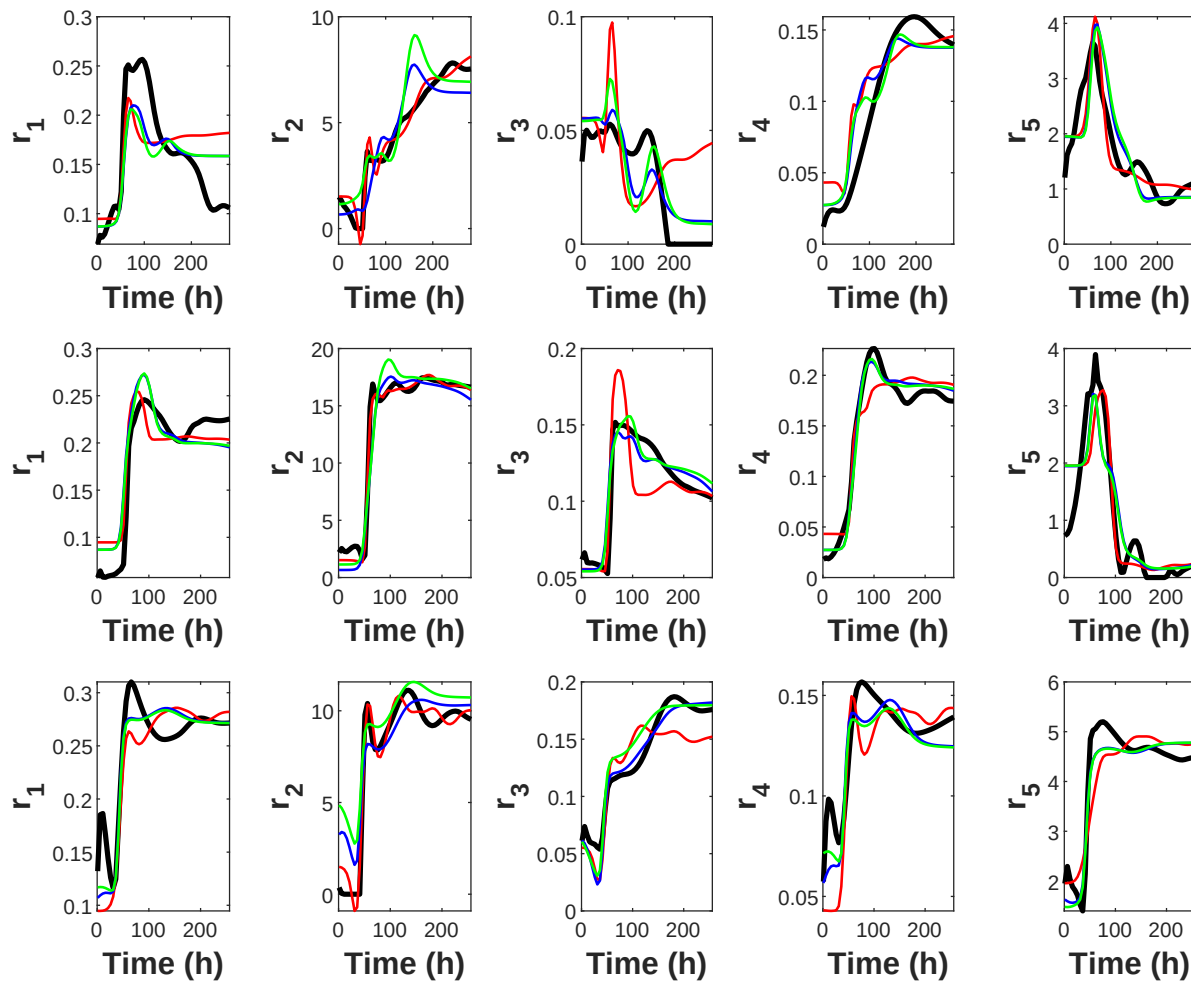


Figure 3.14: Time evolution of the total reaction rates r_i (in mM.h^{-1}) in datasets 1, 2 and 4 (direct validation) – target values, i.e., rates coming from the EFM reduction procedure (black curves); prediction with an MLP with three neurons in the hidden-layer using backpropagation algorithm (red curves), MLP with three neurons and nonlinear re-identification of weights and biases (blue curves) and pruned MLP with three neurons (green curves).

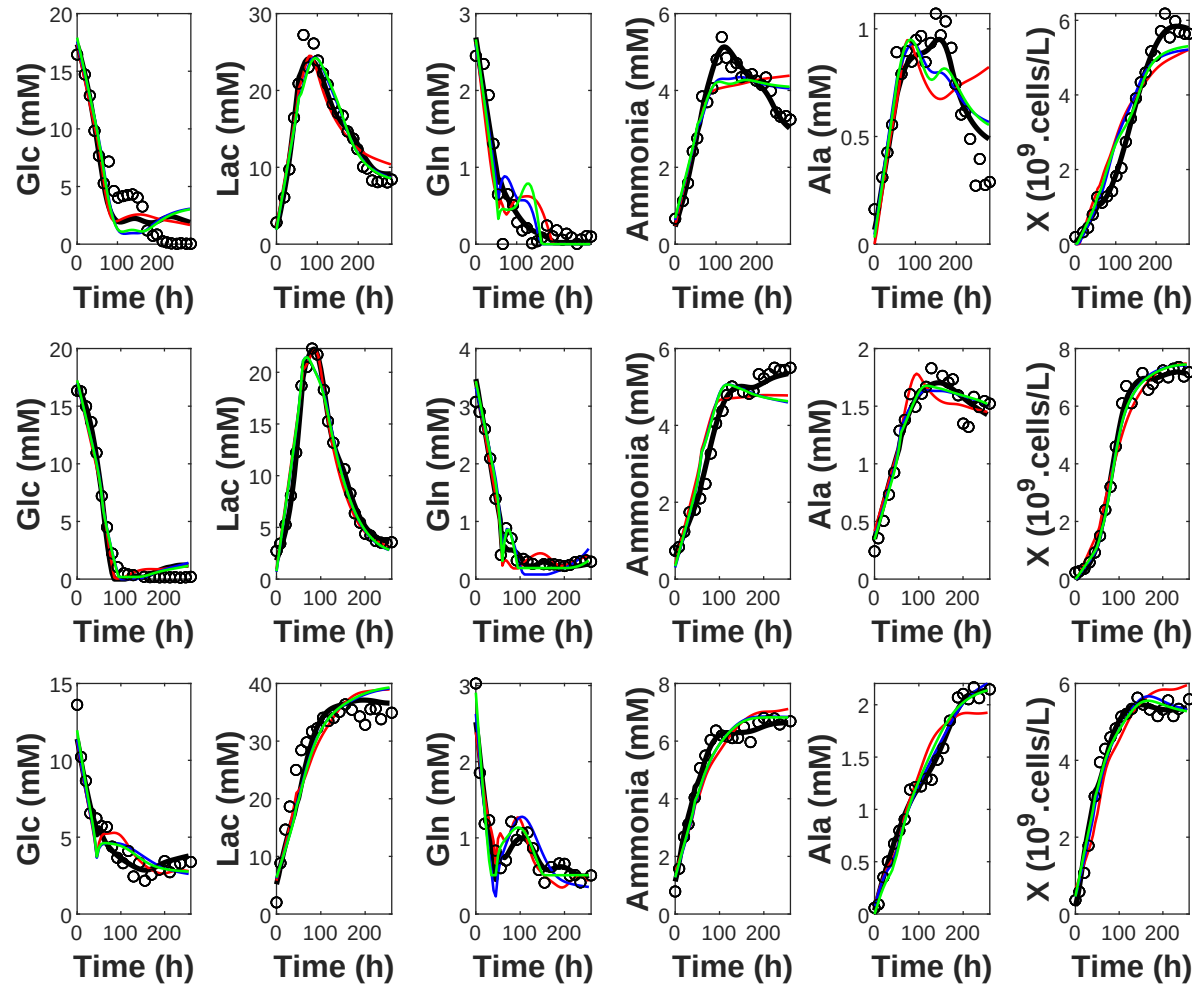


Figure 3.15: Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) — results coming from the EFM reduction procedure (black curves); prediction with an MLP with three neurons in the hidden-layer using backpropagation algorithm (red curves), MLP with three neurons and nonlinear re-identification of weights and biases (blue curves) and pruned MLP with three neurons (green curves).

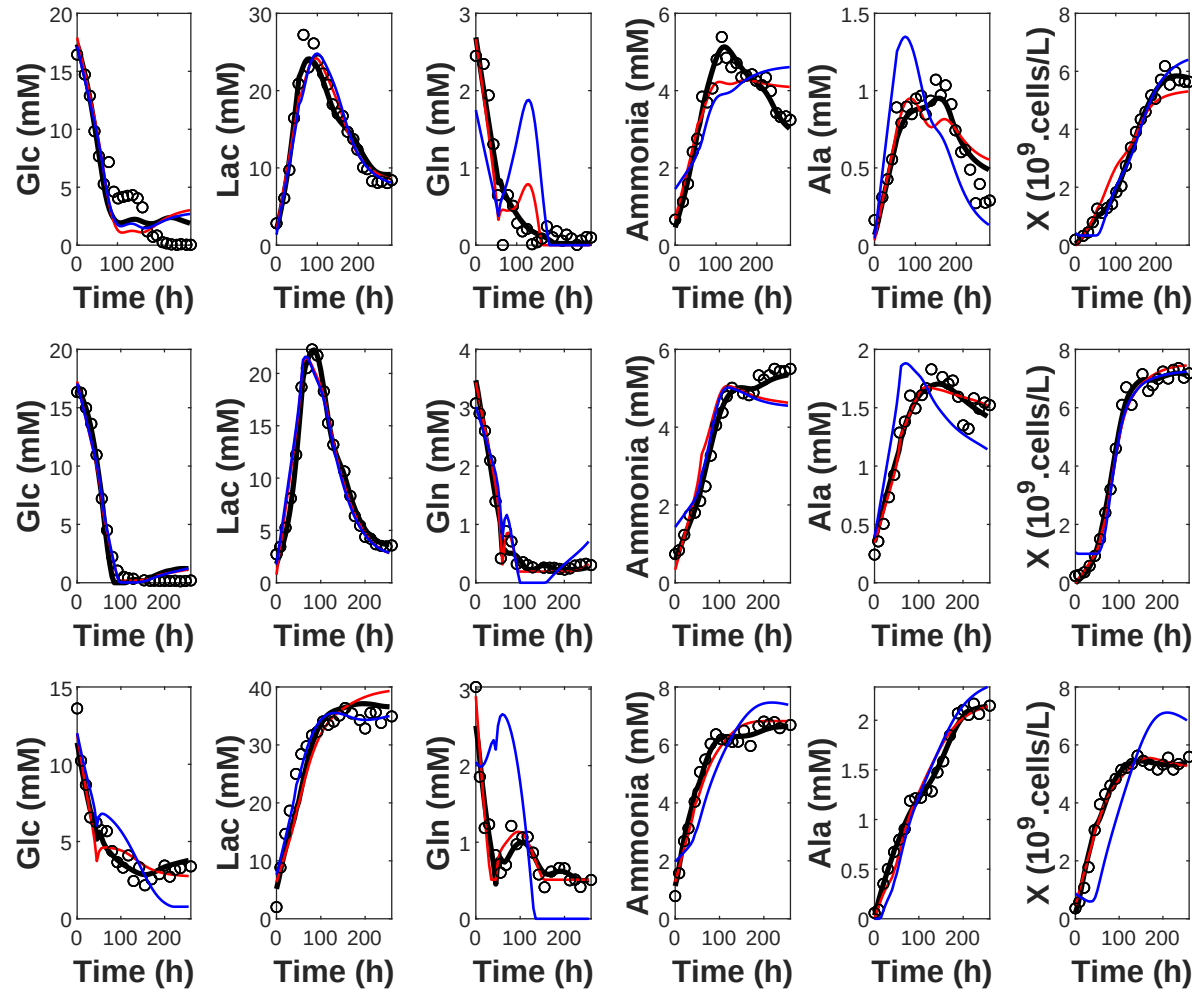


Figure 3.16: Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation) — results coming from the EFM reduction procedure (black curves); prediction with a pruned MLP with three neurons using magnitude pruning algorithm (red curves) and prediction with a pruned MLP with three neurons using the pruning strategy based on CV values (blue curves).

3.3.4.3 Cross-Validation

Based on the previous results, we select for kinetic modeling the multilayer perceptron with three neurons in the hidden-layer, pruned specifically using the magnitude pruning algorithm. To validate the whole methodology and assess the robustness of the proposed modeling framework, a cross-validation is performed using dataset 3 (unseen data), which had not been used in any prior training or calibration steps. This cross-validation procedure serves as a critical step in evaluating both the **EFM-based reduction methodology** and the **kinetic modeling** using ANNs (MLP-based approach). The corresponding RMSEs are summarized in the Appendix F, in Table F.3 for the reduced EFM model before ANN identification and in Table F.7 after the identification of the fluxes with the pruned MLP. The results are depicted in Fig. 3.17 and Fig. 3.18 and show that the reduced network and the final kinetic model remain predictive even for unseen data, with only moderate increases in RMSE during cross-validation.

On the one hand, cross-validation of the EFM reduction procedure exhibits great performance, strongly validating the EFM selection process. This highlights the robustness of the proposed reaction scheme, which effectively captures the essential dynamics of the system. On the other hand, cross-validation of the kinetic identification process remains satisfactory, demonstrating that the neural network-based kinetic model successfully generalizes to new conditions while maintaining reliable predictive capability.

However, a notable discrepancy is observed in the prediction of the time evolution of alanine within dataset 3. This deviation can be attributed to two main factors. First, experimental measurement errors in dataset 3 (from [12]) are well documented, particularly affecting alanine quantification, which introduces uncertainty in both model validation and performance assessment. Second, a fundamental difference in the kinetic behavior of alanine between the training datasets (datasets 1, 2 and 4) and dataset 3 further explains the observed limitations. In datasets 1, 2 and 4, alanine undergoes an extended period of accumulation before being consumed, a pattern effectively learned by the neural network. In contrast, alanine is rapidly consumed after an abrupt production phase in dataset 3, a kinetic regime that is absent from the training data. Therefore, given that MLP-based models are inherently reliant on training data distributions, it is expected that such an unusual kinetic profile is not optimally captured. This highlights a fundamental limitation of data-driven approaches: their ability to generalize is constrained by the diversity of kinetic behaviors represented in the training datasets.

Overall, the results of the cross-validation confirm the strong predictive power of the proposed modeling pipeline. The findings reinforce the effectiveness of the EFM-based reduction methodology and demonstrate that the neural network approach can satisfactorily model the kinetics of reaction rates.

3.3.5 Overflow Metabolism

Beyond numerical validation, it is also valuable to examine the biological relevance of the identified reaction scheme, briefly discussed in Sec. 3.3.4.1. In particular, the structure of the reactions suggests the presence of overflow metabolism, a well-documented phenomenon in mammalian cell cultures.

Overflow metabolism is a phenomenon in which cells achieve incomplete oxidation of an abundantly supplied energy source, commonly glucose and glutamine, despite aerobic conditions. It results in the excretion of organic end-products that are often inhibitory. In this case study, an overflow of glucose is observed, leading to the formation of lactate.

In addition, the production of ammonium reflects a carbon-nitrogen imbalance driven by a surplus amino-nitrogen from the perfusion feed, primarily via deamination of glutamine and other amino acids.

This phenomenon is essentially described by the reaction scheme deduced from the modular EFM selection procedure in Sec. 3.3.4.1. Indeed, in mammalian cell cultures, two main metabolic states can be pointed out: (i) respiratory metabolism at low substrate uptake rates and (ii) overflow metabolism at high substrate uptake rates. During respiratory metabolism, the consumption of glucose and glutamine leads to the growth of biomass and the production of metabolites such as lactate and ammonia. Respiratory growth is captured by reaction in Eq. 3.15 for the substrate consumption and biomass growth, together with Eq. 3.12, Eq. 3.14 and Eq. 3.16 for the production of metabolites. The overflow of glucose leads to the production of lactate, as denoted by the sole reaction in Eq. 3.16 (i.e., not coupled with the reaction in Eq. 3.15), and the excess of glutamine ends in the production of ammonia and lactate but also of other amino acids such as alanine, among others, as described by reactions in Eq. 3.12 and Eq. 3.14. Lastly, the reaction in Eq. 3.13 translates the consumption of glucose and glutamine for cell maintenance. Glucose and glutamine are two energy sources for the production of ATP and reduced pyridine nucleotides, which are essential for cell life. In that respect, all the chemical reactions deduced from the reduced matrix of EFMs have a biological meaning, as expected, consolidating the usefulness of the proposed reduction procedure.

The phenomenon of overflow metabolism can be observed in the values of r_i for $i \in [1, 5]$ in Fig. 3.12 (especially for dataset 2). Indeed, r_5 exhibits large values until about $t = 100$ h, before a significant decrease. This highlights a potential overflow metabolism on glucose during that first period of time, which vanishes at $t = 100$ h when glucose is almost completely depleted. This can be seen in the glucose concentration-time profile in Fig. 3.13.

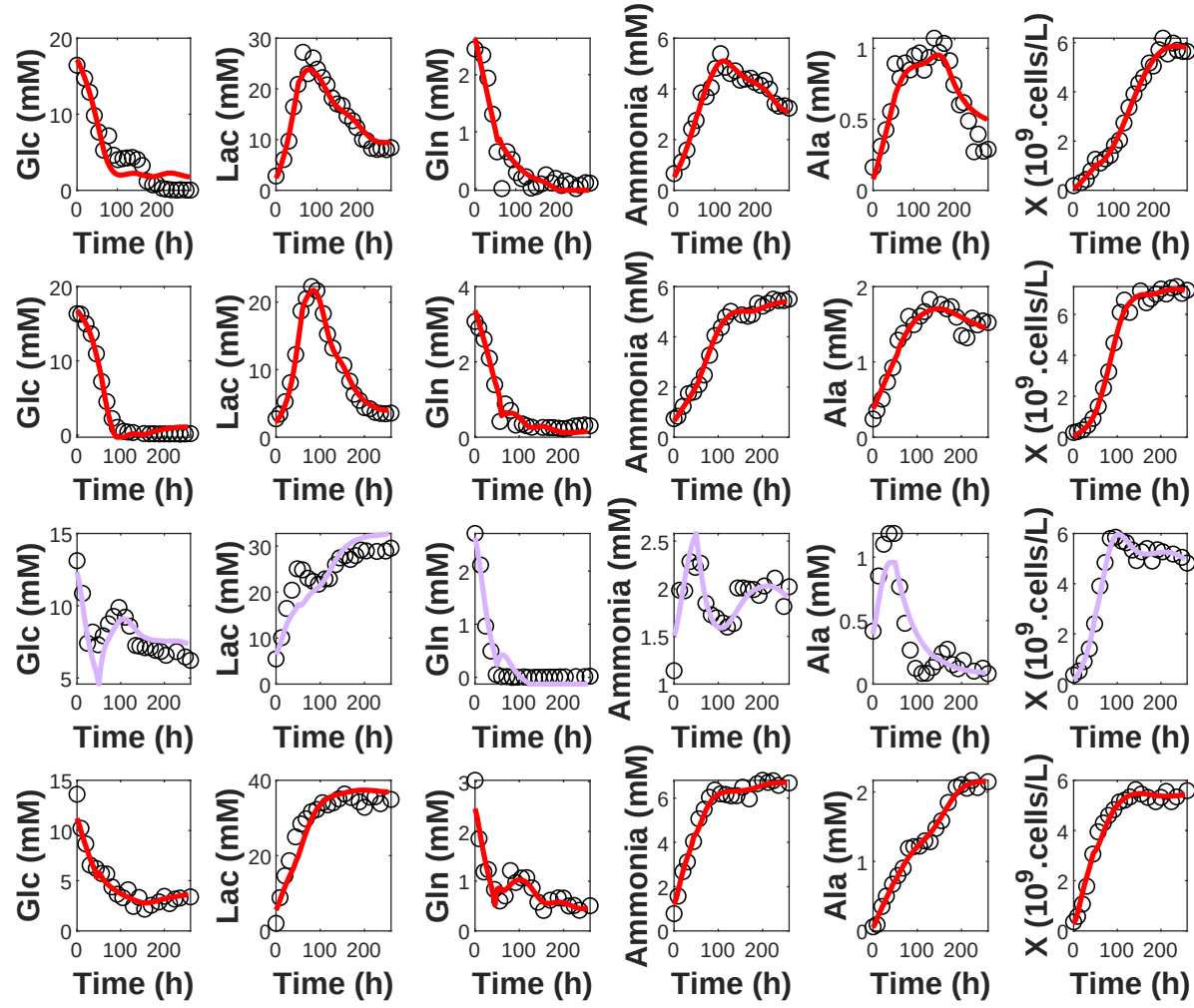


Figure 3.17: Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation – red curves) and dataset 3 (cross-validation – purple curves) – results coming from the EFM reduction procedure considering $\Lambda = 5$ EFMs.

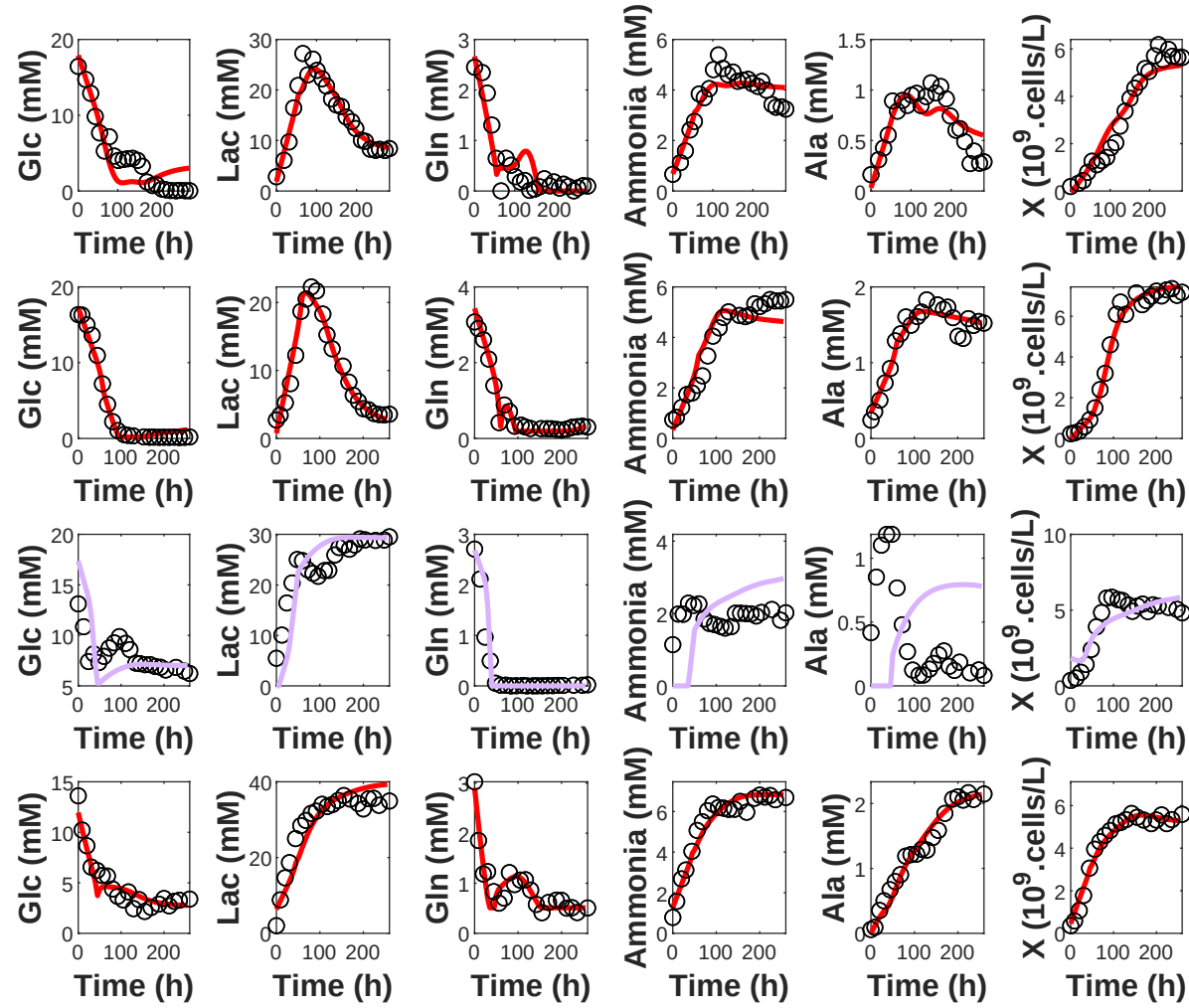


Figure 3.18: Prediction of the time evolution of the extracellular measured species in datasets 1, 2 and 4 (direct validation – red curves) and dataset 3 (cross-validation – purple curves) – prediction with the pruned MLP using the magnitude pruning algorithm.

3.4 Application to Cell Line CHO-320

To further emphasize the modularity of the proposed modeling framework, another application is considered in which the metabolism of the CHO-320 cell line is represented by a significantly larger metabolic network. The increased network complexity leads to a combinatorial explosion in the number of elementary flux modes (EFMs), making their complete enumeration computationally difficult. Consequently, rather than performing a full enumeration of EFMs, a targeted identification of relevant subsets is preferred, leveraging computational efficiency while preserving biological interpretability.

In addition, this case study highlights the crucial advantage of integrating biological knowledge into the modeling process. In some cases, prior biochemical understanding enables more structured and biologically meaningful parameterization of reaction rates, thus facilitating the modeling process. By incorporating biological insights, it is possible to reduce dependency to purely data-driven parameter estimation and improve the predictive accuracy of the model.

3.4.1 Cell Line and Media

Chinese Hamster Ovary (CHO) cell line is one of the most widely used mammalian cell lines for biopharmaceutical production, particularly in the manufacturing of recombinant proteins and monoclonal antibodies. Among its variants, CHO-320 represents a specific line that has been adapted for optimized growth and protein expression under controlled culture conditions. CHO cells are preferred because of their ability to perform complex post-translational modifications, including glycosylation, which is crucial for the functionality and stability of therapeutic proteins.

The experimental data come from CHO-320 cell cultures used in [7, 59]. The experiments were performed by the Biochemistry Laboratory of the Catholic University of Louvain (UCL), Belgium (courtesy of Prof. Yves-Jacques Schneider). The experimental data correspond to measurements collected from the exponential phase of different batch cultures, carried out in a serum-free medium supplemented with rice protein hydrolysate and glutamine. Cultures were settled in a working volume of 25 mL in shake flasks and incubated at 100 rpm in a CO₂ incubator at 37 °C in an atmosphere of 5% CO₂ in air. The cultures were inoculated to reach an initial concentration of $0.3 \cdot 10^9$ cells/L. Cultures were kept for 190 hours, with an exponential growth phase of approximately 80 hours. These experiments have been designed to have an initial concentration of 16 mM of glucose and 6 mM of glutamine.

3.4.2 Metabolic Network

For this application, it is proposed to derive a macroscopic model for the exponential growth phase of CHO cells, during which cells divide and proliferate. In the context of this study, a metabolic network containing 100 biochemical reactions and 72 internal metabolites is considered. This network is specific to the growth phase of the cells and is developed in [13]. The corresponding metabolic reactions can be found in the Appendix C. The network considered in this section is more detailed than the one considered in Sec. 3.3. It is composed of the glycolysis pathways, the TCA cycle, the pentose phosphate pathway (PPP), nucleotide synthesis pathways, and several amino acid synthesis routes. Besides, it is necessary to include a reaction related to biomass formation so as to consider its measurement in the EFM reduction procedure. According

to [13], the cellular composition of the studied CHO cells is the following: 92.26% of proteins, 1.3% of RNA, 0.52% of DNA and 2.97% of lipids.

3.4.3 Measurement Configuration

As in Sec. 3.3, this study focuses on a set of six extracellular measurements, namely glucose, glutamine, ammonia, lactate, alanine and biomass, for the sake of simplicity. Other measurement scenarios can be found in our published articles [5, 112, 139]. Imposing a measurement scenario leads to the definition of the stoichiometric matrix of extracellular measurements N_e . In addition, it is necessary to compute the experimental specific uptake and excretion rates $\underline{\nu}_m(t)$.

In a first approximation, the specific uptake and excretion rates can be assumed constant during the exponential growth phase and are simply evaluated by linear regression using the linear least squares method (see Sec. 2.7.1). The details of the computations are given in the Appendix D.2. In a second evaluation, it is considered that specific extracellular rates can vary over time, as is usually the case during a batch or a fed-batch culture. In that case, the rates are evaluated by applying smoothing splines to the measured concentration profiles and performing numerical differentiation using Eq. D.1 and Eq. D.2 (considering $D = 0$ – batch cultures). It is worth noting that the data is limited for this application (only the measurements of two batch cultures are considered for direct validation).

3.4.4 Numerical Results

3.4.4.1 EFM Reduction Procedure

Before going into technical details related to the two situations, a first insight involves computing the entire set of elementary vectors using the stoichiometric matrix N of the metabolic network. This can be done independently of the measurement scenario using EFMTOOL as an EFM generator, which provides 993,203 elementary modes. The latter represent the edges of the pointed polyhedral cone $\mathbf{S}$ depicted in Fig. 2.12. This very large number of EFMs is quite difficult to manipulate. As such, a first approach would be to partition this initial set into several smaller subsets and apply the reduction procedure on each subset. This approach has been tested and validated and gives good results in terms of reduction, but at the expense of a high computation time.

An alternative approach is to generate a smaller initial set of EFMs using a subset generation algorithm. As mentioned in Sec. 3.2.1.2.1.2, the fast algorithm introduced in [82] can be employed. This algorithm can be executed an arbitrary number of times so as to generate the desired initial number of EFMs. When the specific reaction rates are assumed constant and are computed by linear interpolation, the fast generation algorithm allows obtaining $n_{EFM} = m_e = 6$ EFMs at once. Note that the elementary vectors obtained using this algorithm are now related to the polytope $\mathbf{F}$ depicted in Fig. 2.12. When starting from this small initial set, the intermediate steps of the reduction procedure (i.e., the application of the cosine-similarity algorithm and the optimization-based reduction) can be bypassed. The final stage of the methodology can be applied immediately to reduce the number of EFMs below m_e . Two target numbers are considered in what follows: (i) $\Lambda_1 = 5$ and (ii) $\Lambda_2 = 4$.

A direct validation of the reduction procedure can be achieved by comparing the constant reaction rates $\underline{\nu}_m$ obtained by linear regression with the product $K_e \cdot \Phi$ deduced

from the optimization problem, K_e corresponding to the reduced matrix of modes. The rates being constant, this validation can be summarized in Table 3.6.

Table 3.6: Estimated values of the specific reaction rates ν_{ξ_i} (in mM.h⁻¹ per 10⁹ cells) and ν_X (in h⁻¹) obtained by linear regression (L.R.) and prediction from the optimization problem of the reduction procedure.

	Cult. 1			Cult. 2		
	L.R.	$\Lambda_1 = 5$	$\Lambda_2 = 4$	L.R.	$\Lambda_1 = 5$	$\Lambda_2 = 4$
ν_m						
ν_{Glc}	-0.2072	-0.2054	-0.2075	-0.1393	-0.1308	-0.1345
ν_{Lac}	0.4132	0.4168	0.4127	0.2582	0.2728	0.2664
ν_{Gln}	-0.0511	-0.0510	-0.0464	-0.0481	-0.0476	-0.0387
$\nu_{Ammonia}$	0.0449	0.0449	0.0485	0.0359	0.0362	0.0411
ν_{Ala}	0.0117	0.0117	0.0117	0.0097	0.0097	0.0097
ν_X	0.0278	0.0278	0.0277	0.0319	0.0319	0.0316

This validation can be pursued by the analysis of the concentration profiles, obtained by integration of the computed fluxes and identification of the most likely initial conditions of the measured species. This is illustrated in Fig. 3.19. The results are quite satisfactory, despite the simplifying assumption of constant rates, and even when the number of modes is reduced to $\Lambda = 4$.

In a second evaluation, we now consider that the specific extracellular rates can vary over time and are computed by applying smoothing splines to the measured concentration profiles and numerical differentiation. Based on these time-varying rates, the execution of the fast generation algorithm produces an initial set of 206 elementary flux vectors. The complete reduction procedure can therefore now be applied. First, the test relative to the interpretability of the stoichiometric matrix K yields a reduced set of 173 modes. Second, the collinearity test could be used to further reduce this set, but as the number of EFMs is quite modest, it is preferable to perform directly the optimization-based reduction. Usually, the collinearity test is reserved for situations where the initial set is much larger and a preliminary, relatively drastic (but unsupervised) reduction is required, as discussed in [5]. The reduction is therefore continued using random elimination based on an optimization criterion ensuring the positivity constraints, with a target value $\Omega = 10$ (this latter number can of course be chosen differently, but we recommend a target slightly above m_e). Finally, the last step involves finding the best combination of $\Lambda_1 = 5$ or $\Lambda_2 = 4$ modes among these 10 EFMs, minimizing the least squares deviation from the measured data.

Fig. 3.20 and Fig. 3.21 show the time evolution of the reaction rates and the corresponding concentration profiles, respectively, predicted on the basis of 173 elementary flux modes or $\Omega = 10$ EFMs. Fig. 3.22 and Fig. 3.23 extend the illustration to the final sets of $\Lambda_1 = 5$ or $\Lambda_2 = 4$ modes. For $\Lambda_1 = 5$, the results are very satisfactory. With $\Lambda_2 = 4$, significant deviations appear in reaction rates, though they have a limited impact on the prediction of concentration profiles over time. Mostly the prediction of ammonia is adversely affected, but the results remain quite acceptable in view of establishing simple dynamic models for bioprocess control purposes. A quantitative assessment of these successive reduction steps is provided in the Appendix F in Tables F.8 to F.11, confirming that the chosen reductions preserve a high level of predictive accuracy.

Furthermore, the reconstruction of the time evolution of the concentrations in Fig. 3.23 is better than in Fig. 3.19, which highlights the benefit of computing the reaction rates from smoothing splines and differentiation methods instead of making use of the simplifying assumption of constant rates.

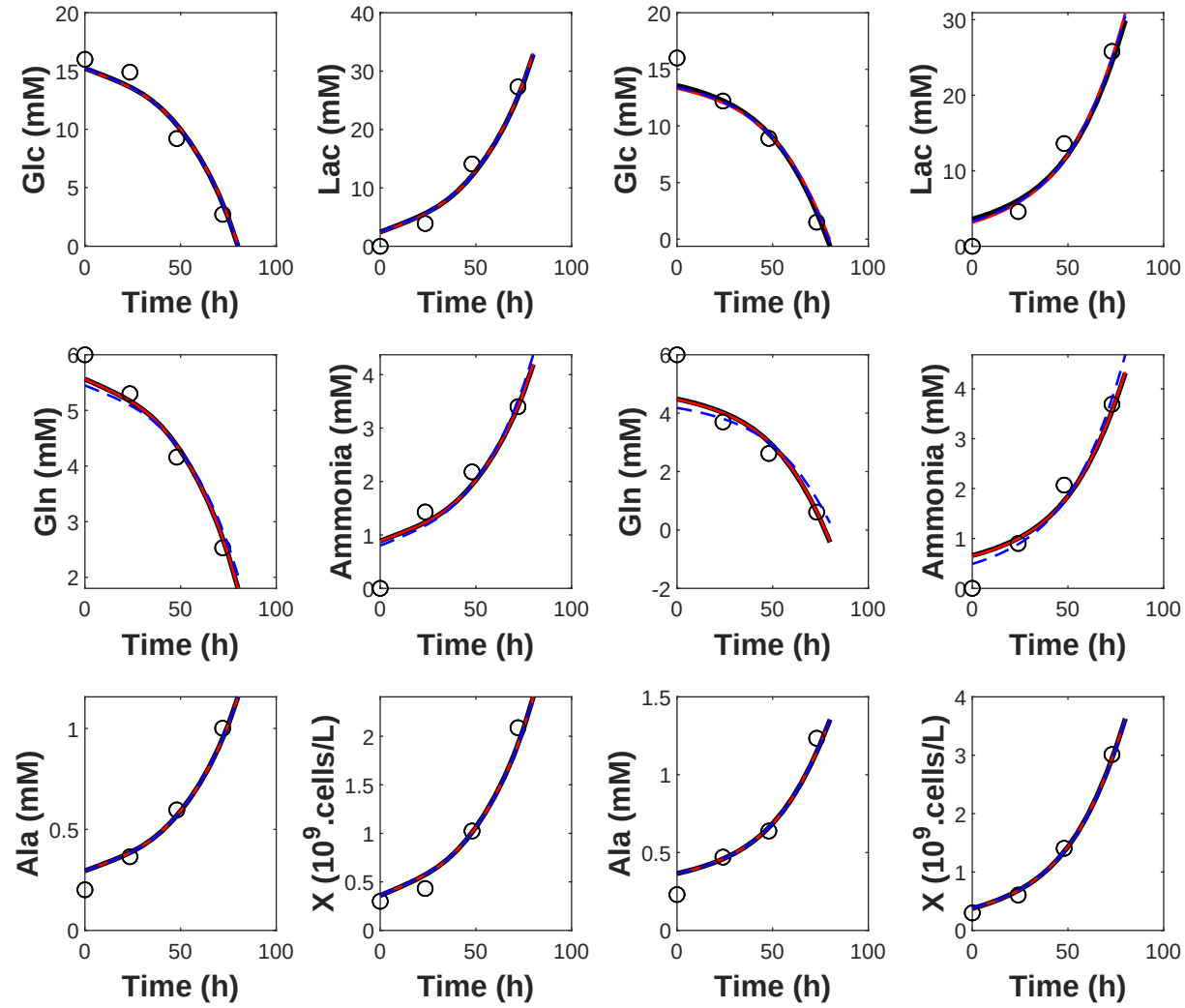


Figure 3.19: Prediction of the time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = m_e = 6$ (black curves), reducing $n_{EFM} = \Lambda_1 = 5$ (red curves) and $n_{EFM} = \Lambda_2 = 4$ (dashed blue curves).

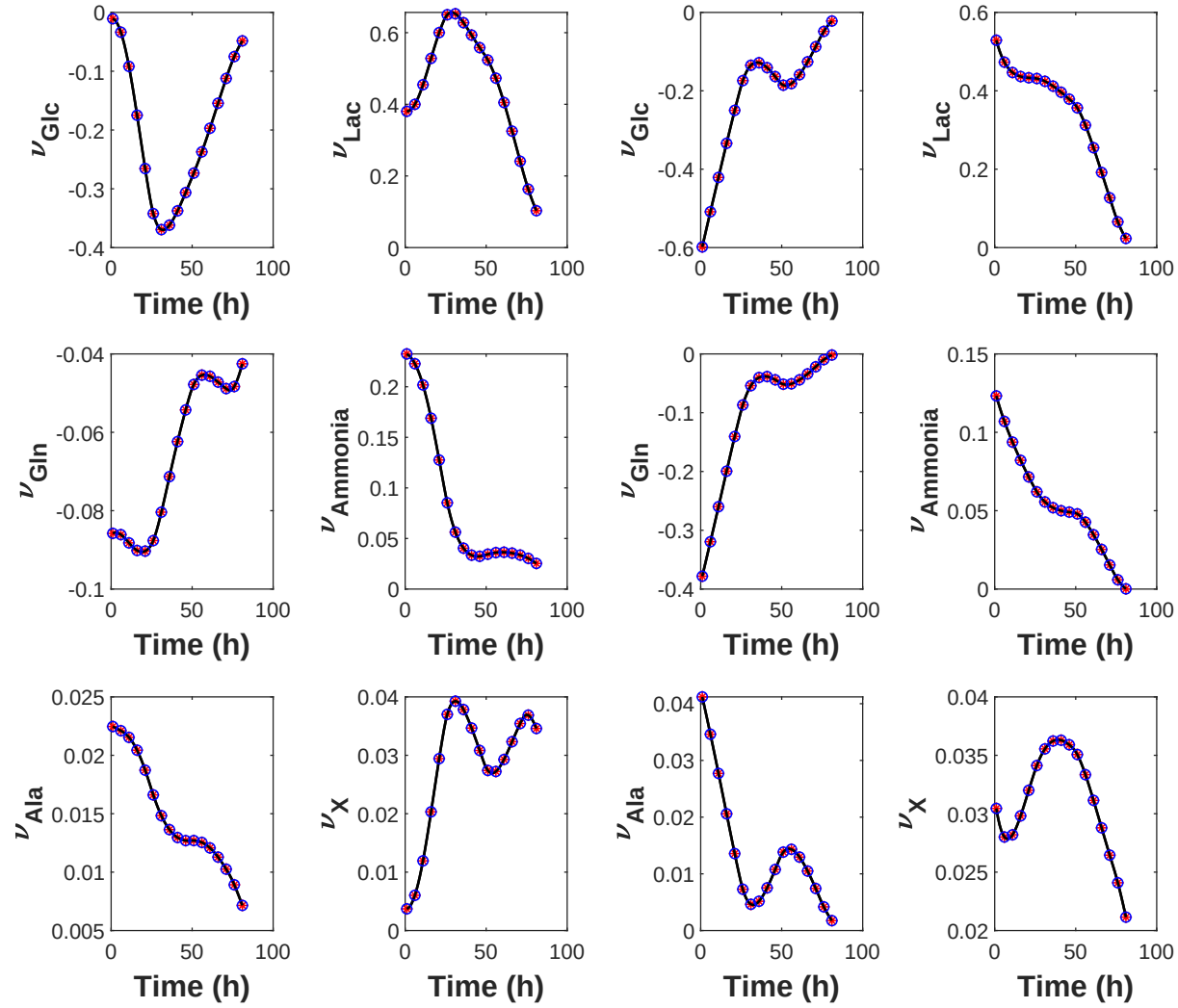


Figure 3.20: Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in $\text{mM}\cdot\text{h}^{-1}$ per 10^9 cells) and ν_X (in h^{-1}) in dataset 1 (left-hand side) and dataset 2 (right-hand side): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = 173$ (red crosses); $K_e\Phi$ for $n_{EFM} = \Omega = 10$ (blue bullets).

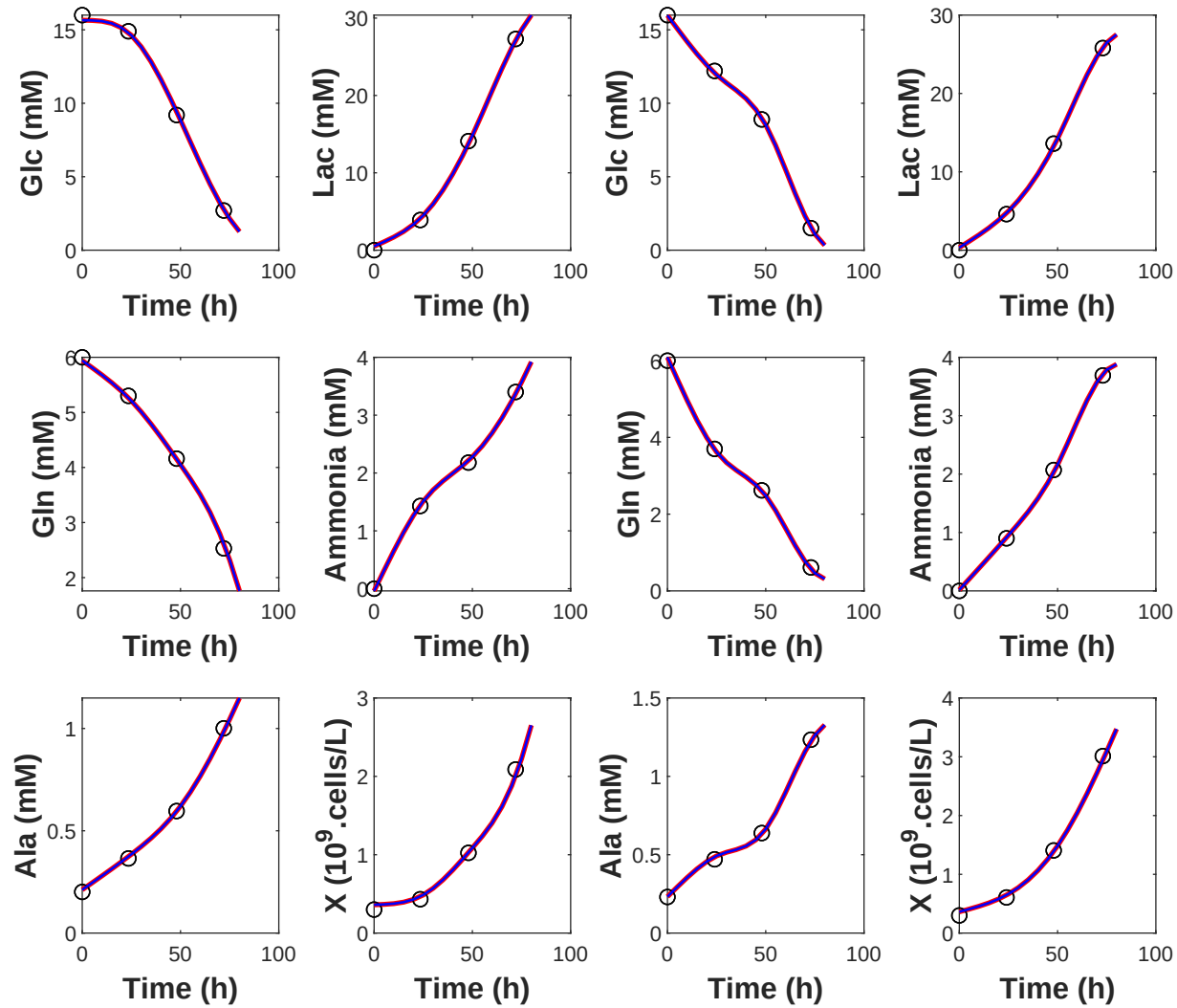


Figure 3.21: Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = 173$ (red curves), reducing $n_{EFM} = \Omega = 10$ (blue curves).

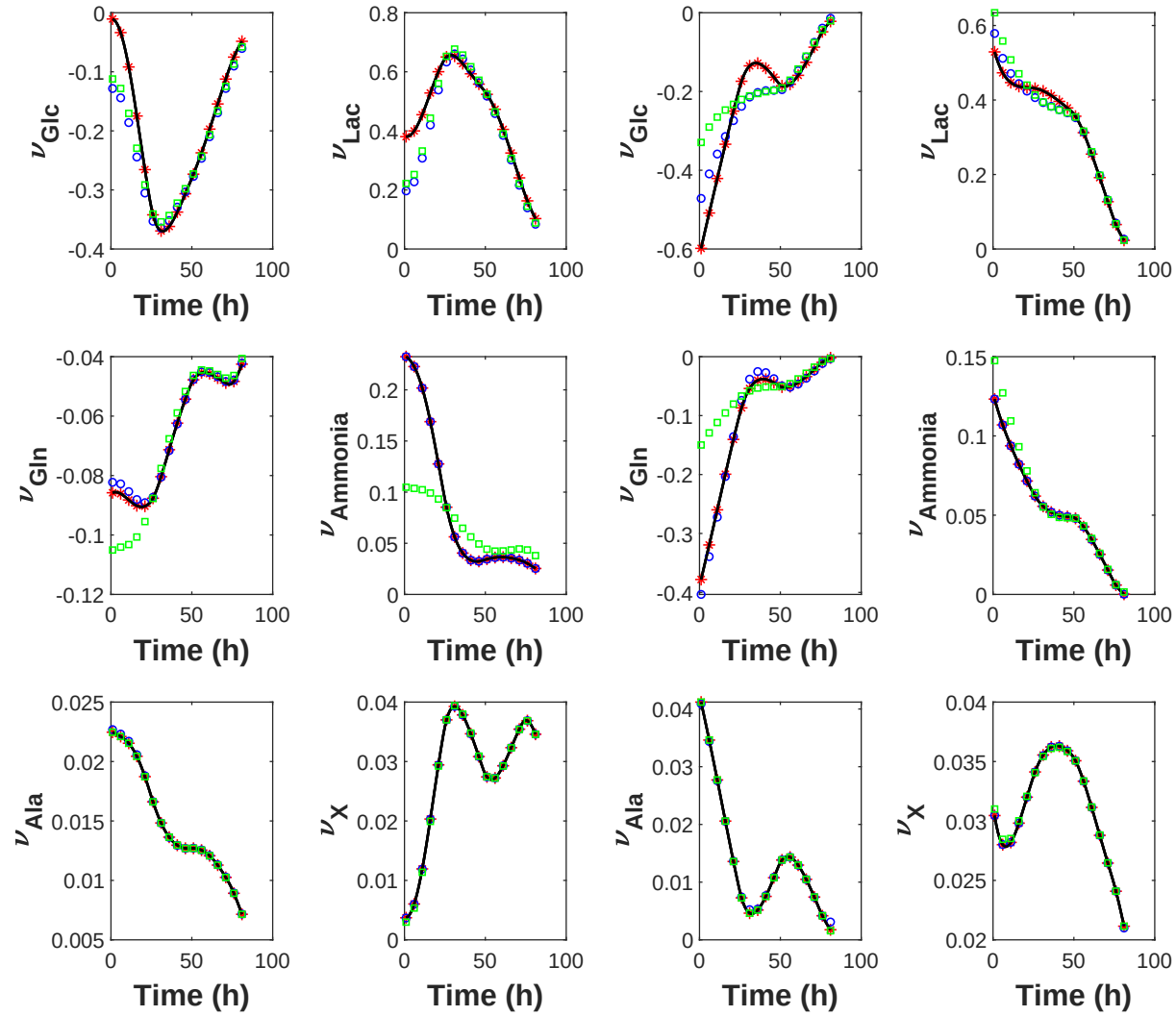


Figure 3.22: Time evolution of the specific uptake and excretion rates ν_{ξ_i} (in mM.h^{-1} per 10^9 cells) and ν_X (in h^{-1}) in dataset 1 (left-hand side) and dataset 2 (right-hand side): numerical results (black curves); $K_e\Phi$ for $n_{EFM} = \Omega = 10$ (red crosses); $K_e\Phi$ for $n_{EFM} = \Lambda_1 = 5$ (blue bullets) and $K_e\Phi$ for $n_{EFM} = \Lambda_2 = 4$ (green squares).

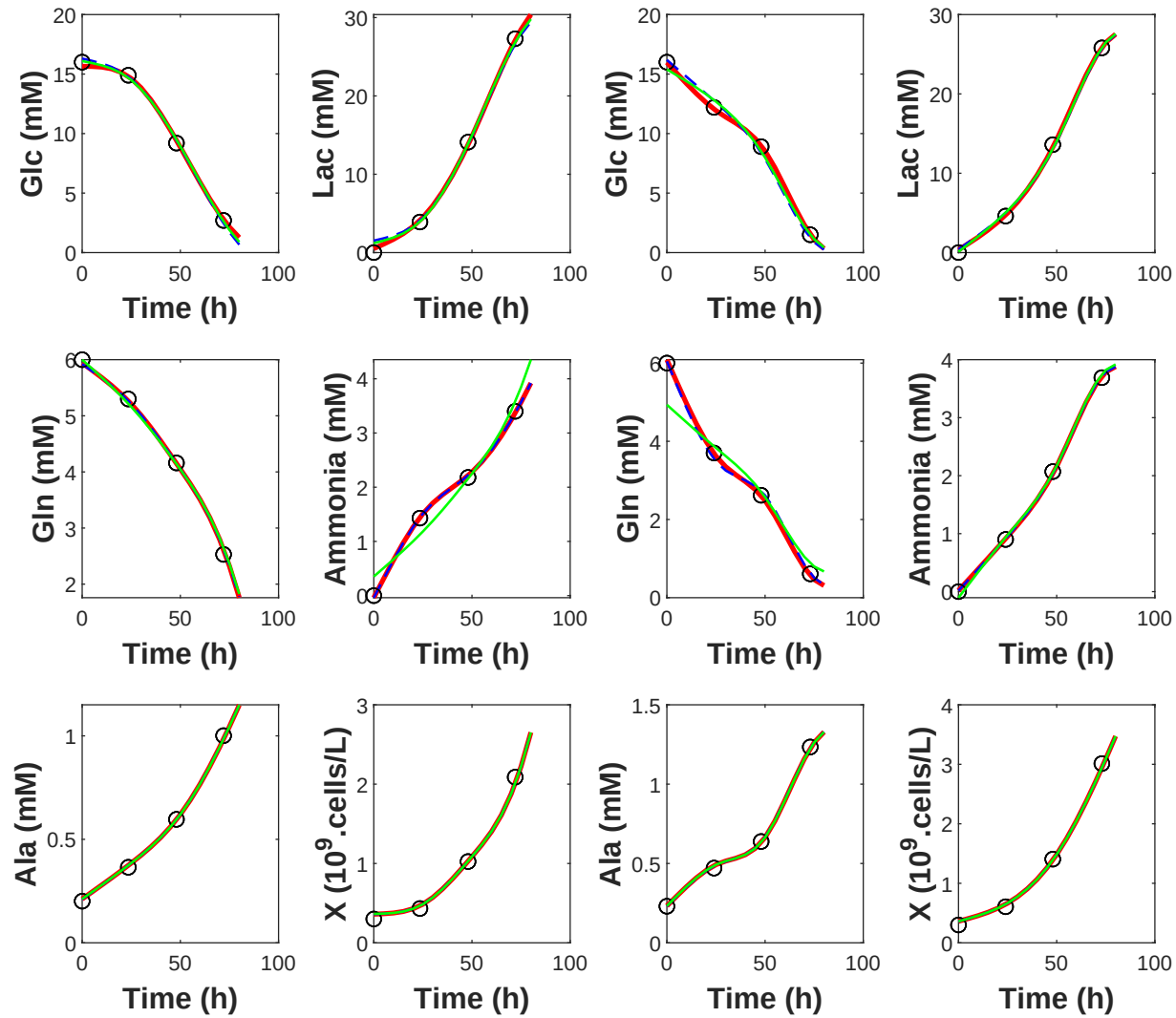


Figure 3.23: Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points (black bullets), concentration profiles on the basis of $n_{EFM} = \Omega = 10$ (red curves), reducing $n_{EFM} = \Lambda_1 = 5$ (blue curves) and reducing to $n_{EFM} = \Lambda_2 = 4$ (green curves).

In order to assess the merits of the reduction procedure, in addition to the previous validations, it might be interesting to examine the set of macro-reactions given by the reduced stoichiometric matrix K :

$$K = \begin{bmatrix} -\alpha_1 & 0 & -\alpha_3 & 0 \\ 0 & 0 & \beta_3 & 0 \\ -\gamma_1 & -\gamma_2 & 0 & -\gamma_4 \\ 0 & \delta_2 & 0 & \delta_4 \\ 0 & 0 & 0 & \epsilon_4 \\ \sigma_1 & 0 & 0 & 0 \end{bmatrix} \quad (3.22)$$

where α_i , β_i , γ_i , δ_i , ϵ_i and σ_i are the stoichiometric coefficients. The value of these coefficients is given in Table 3.7. K has dimension $m_e \times \Lambda$ with $m_e = 6$ and $\Lambda = 4$. Equivalently, a macro-reaction scheme can be drawn:



This reaction scheme makes sense, as glucose and glutamine are consumed during the growth phase, while lactate, ammonia, alanine, and biomass are produced, as confirmed by the time evolution of these species. The first reaction translates the growth of biomass by consumption of glucose and glutamine, the second reaction indicates that the consumption of glutamine leads to the production of ammonia, the third reaction shows that lactate is produced from glucose, and the last reaction involves the consumption of glutamine to produce ammonia and alanine.

Table 3.7: The stoichiometric coefficients of the reaction network.

Parameter	Value	Parameter	Value	Parameter	Value
α_1	0.1781	γ_2	0.5000	ϵ_4	0.2673
α_3	0.2500	γ_4	0.2673	σ_1	0.4517
β_3	0.5000	δ_2	0.5000		
γ_1	0.0347	δ_4	0.2673		

3.4.4.2 Kinetic Modeling

In order to derive a macroscopic model, it remains to model the reaction rates Φ using classical kinetic laws or artificial neural networks, knowing that the time-series data are obtained numerically from the reduction method. As a reminder, the mass-balance dynamic model is given as follows:

$$\frac{d\xi}{dt} = K \Phi X + D(\xi_{in} - \xi)$$

In this equation, ξ is the vector of concentrations of the extracellular species in the culture medium, K represents the stoichiometry of the reaction network built on the basis of the reduced set of EFMs, Φ is the vector of specific reaction rates, X represents the biomass concentration, D represents the diagonal dilution matrix and ξ_{in} denotes the

vector of inflow concentrations. In this application, the cultures operate in batch mode and the second term in the RHS vanishes ($D = 0$).

Without diving into the problem of the formulation of kinetic laws and the identification of their parameters, a basic dynamic simulator is proposed based on the assumption that the macro-reactions proceed at their maximal rate during the exponential growth phase. The reaction rates are modeled using Michaelis-Menten kinetics:

$$\phi_1 = \phi_{max_1} \frac{Glc}{Glc + \kappa} \frac{Gln}{Gln + \kappa} \quad (3.27)$$

$$\phi_2 = \phi_{max_2} \frac{Gln}{Gln + \kappa} \quad (3.28)$$

$$\phi_3 = \phi_{max_3} \frac{Glc}{Glc + \kappa} \quad (3.29)$$

$$\phi_4 = \phi_{max_4} \frac{Gln}{Gln + \kappa} \quad (3.30)$$

On the one hand, the parameters ϕ_{max_i} can be obtained by selecting the maximum entries in the vector $\Phi_{num}(t)$. On the other hand, the half-saturation constants κ (which would not be identifiable based on the limited data at disposal) are taken as small numbers to ensure the positivity of the model, but not too small so as to avoid excessive stiffness of the system of ordinary differential equations (ODEs). The value of the parameters is given in Table 3.8.

Table 3.8: Value of the kinetic parameters.

	Unit	Value
ϕ_{max_1}	h^{-1}	0.06
ϕ_{max_2}	h^{-1}	0.1
ϕ_{max_3}	h^{-1}	0.71
ϕ_{max_4}	h^{-1}	0.05
κ	mM	0.01

The simulation results are shown in Fig. 3.24 and are very satisfactory, as confirmed by the low RMSE reported in Table F.12. By relying on biological insight to guide parameter selection, rather than performing a full parameter identification, we are able to construct a functional and predictive model. Although further refinement through classical parameter estimation could improve accuracy, it is not required in this case to capture the system dynamics adequately. Thus, we have demonstrated that the EFM reduction procedure opens the door to the fast development of dynamic macroscopic models.

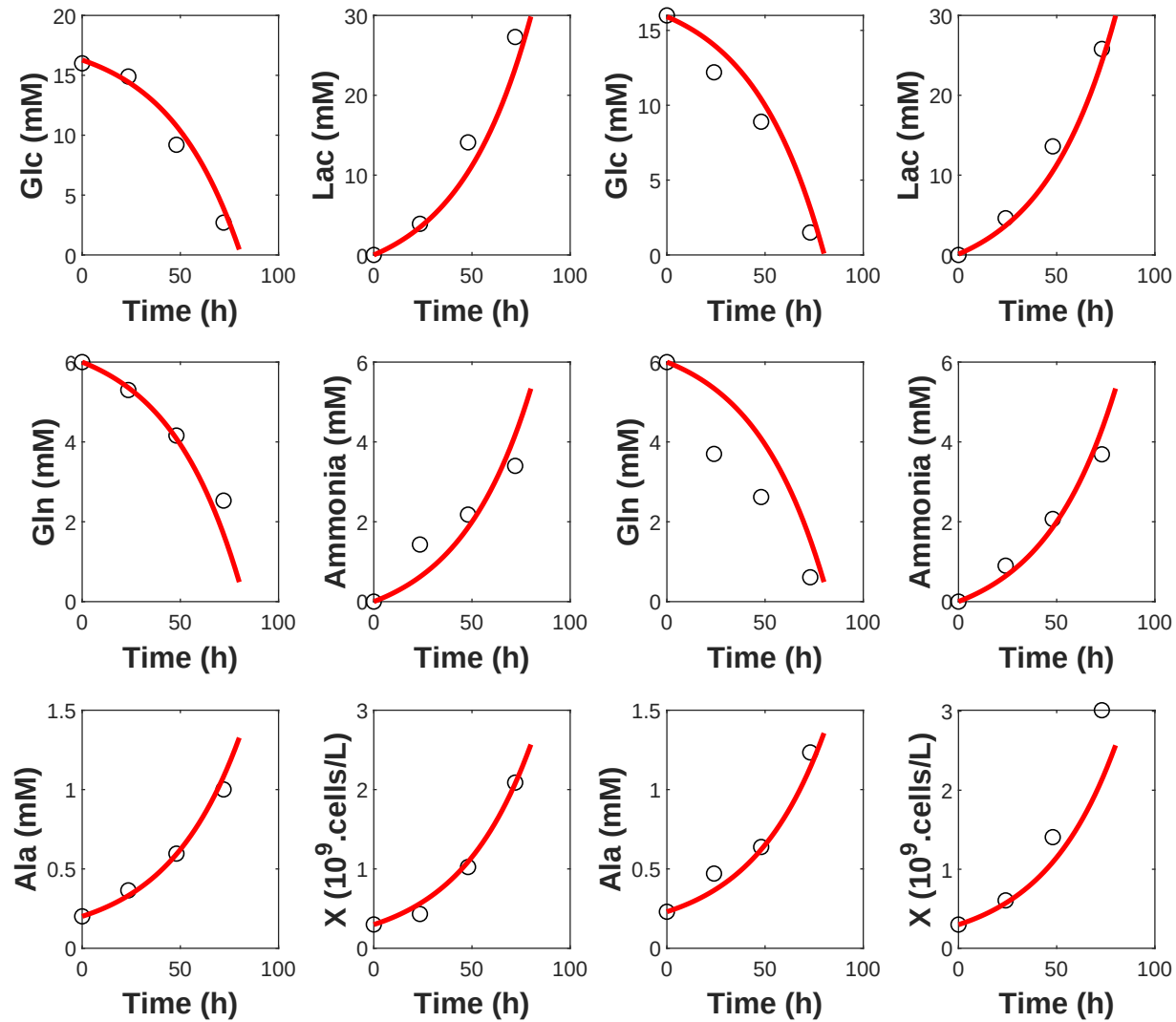


Figure 3.24: Time evolution of the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – experimental data points are black bullets and the concentration profiles computed from the ODE system (dynamical simulator) are depicted by the red curves.

3.4.4.3 Cross-Validation

As before, to validate the whole methodology and assess the robustness of the proposed modeling framework, a cross-validation is performed using dataset 3 (unseen data), which had not been used in any prior training or calibration steps. This cross-validation procedure serves as a critical step in evaluating both the **EFM-based reduction methodology** and the **kinetic modeling** based on simple kinetic laws. The results are depicted in Fig. 3.25, and the corresponding RMSEs are summarized in Table F.11 and Table F.12, before and after kinetic identification, respectively.

On the one hand, cross-validation of the EFM reduction procedure exhibits great performance, strongly validating the EFM selection process. This highlights the robustness of the proposed reaction scheme, which effectively captures the essential dynamics of the system. On the other hand, the cross-validation of the kinetic modeling process remains satisfactory, despite the simplified assumptions we have made. Note that better results could have been obtained if larger quantities of informative data were available.

3.4.4.4 Positivity Constraints

Finally, to provide mathematical rigor and support the interpretation of the results, we introduce this section, which focuses on the role of positivity constraints in the flux estimation process. Beyond predictive performance, the method also aims to characterize how these constraints influence flux estimation. In particular, we investigate how the index Ξ responds when the number of selected EFMs is varied with respect to the number of extracellular measurements, especially in cases where the number of modes is reduced below m_e , limiting the ability to capture the full variance present in the data.

As mentioned in Sec. 3.2.1.2.3, the index Ξ should be equal to zero when: (i) the number of modes is greater than or equal to the number of extracellular measurements ($n_{EFM} \geq m_e$), (ii) K_e is full row rank and (iii) there are no constraints on $\phi(t)$. However, the positivity constraints can be active in some time periods depending on the compatibility of the measurements (which are always affected by experimental errors) with the network. The index Ξ therefore indicates how well the positivity constraints are satisfied, and the minimization of the index selects the flux values that achieve the best constraint satisfaction. This is illustrated in Fig. 3.26. The blue curves represent the situation where $n_{EFM} \geq m_e$. Therefore, it is apparent that each time the index is non-zero, a positivity constraint is active. In contrast, the pink curves represent the situation where $n_{EFM} < m_e$. In this case, the non-zero value of the indicator corresponds to the combined impact of both the positivity constraints and the fitting of the experimental measurements with a model which has now a reduced number of degrees of freedom.

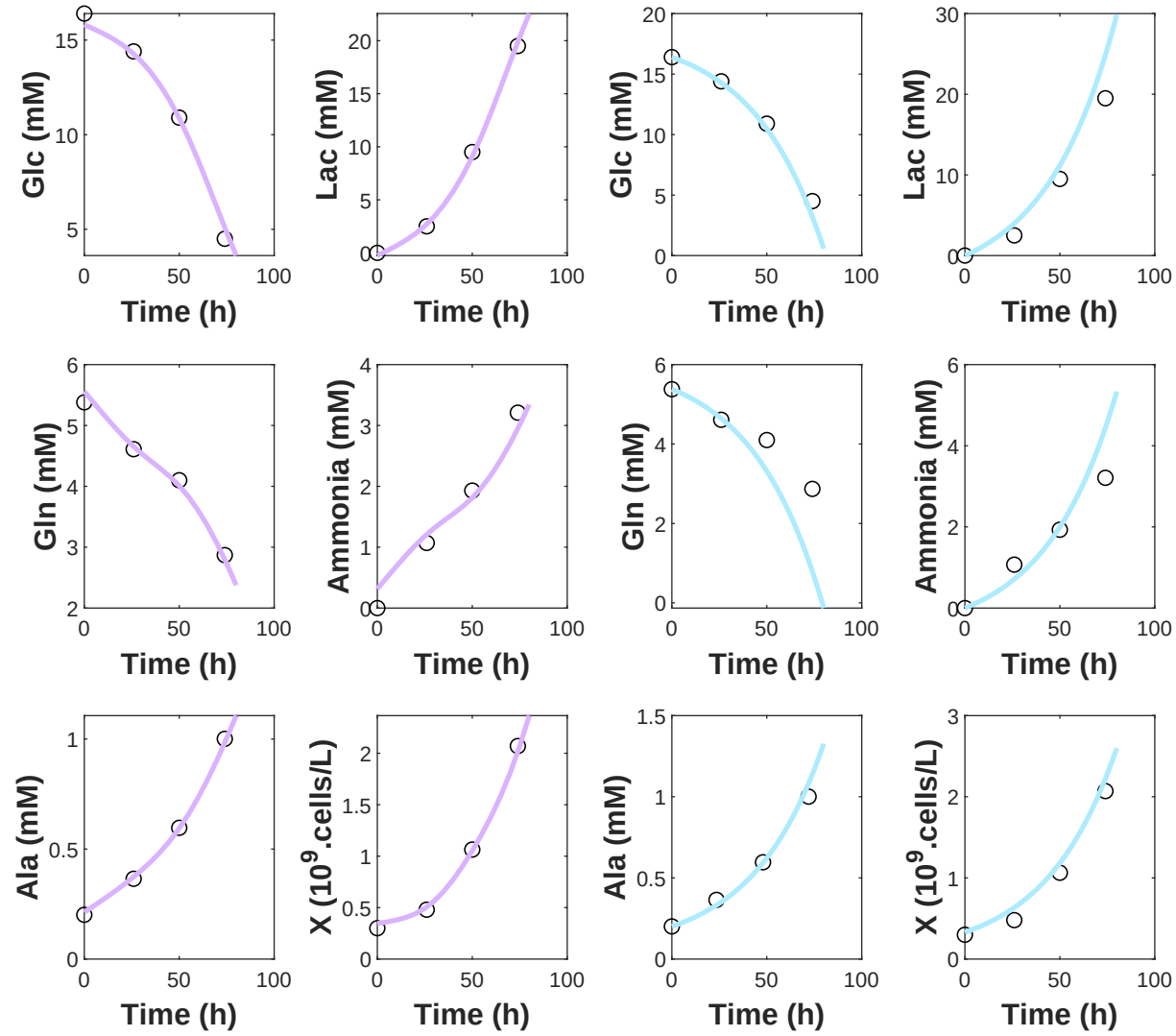


Figure 3.25: Time evolution of the extracellular measured species in dataset 3 (cross-validation) – experimental data points (black bullets), results from the EFM reduction procedure considering $n_{EFM} = \Lambda_2 = 4$ modes (left-hand side); concentration profiles computed using the dynamical simulator (right-hand side).

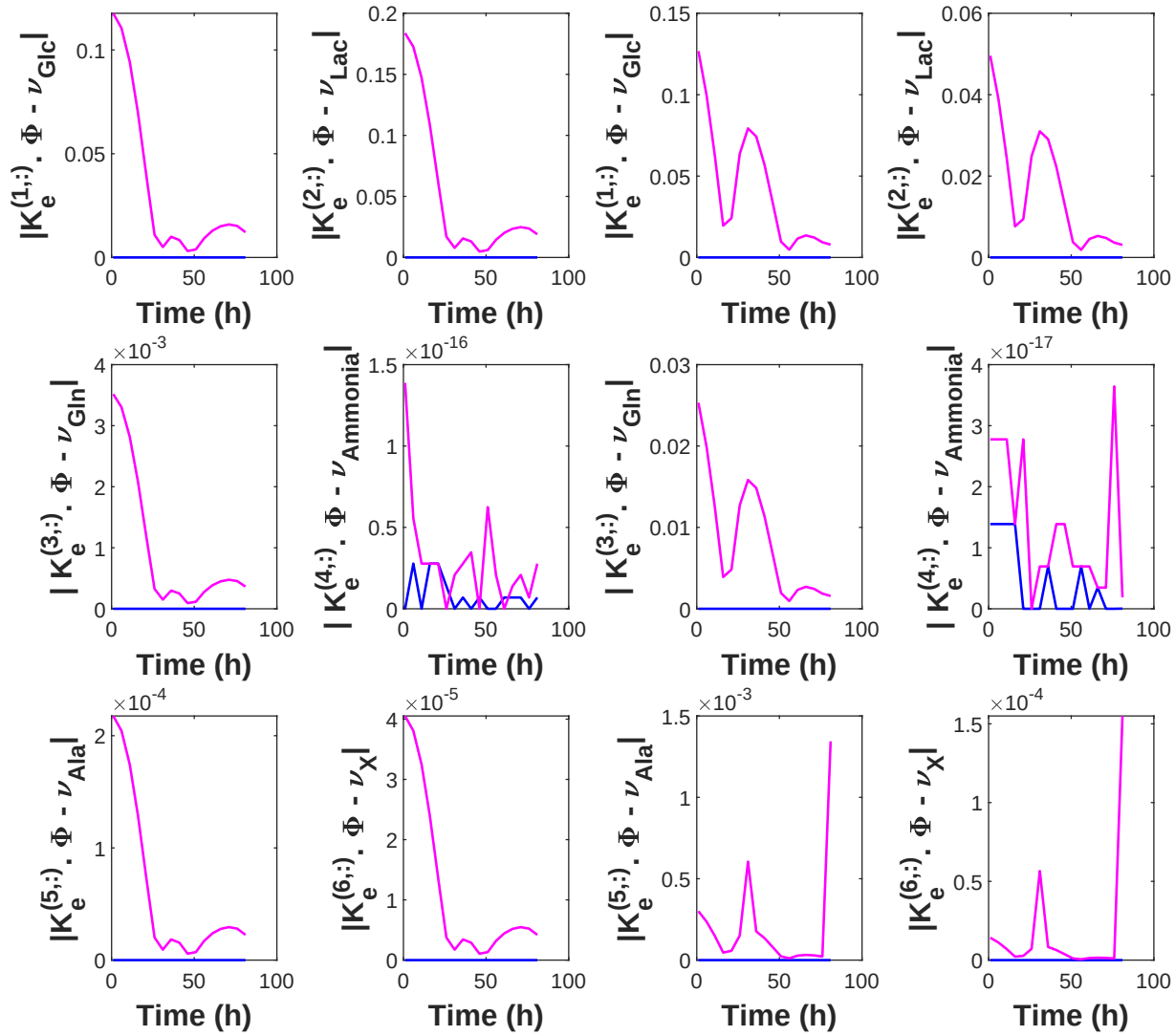


Figure 3.26: Time evolution of the index $|K_e \Phi - \nu|$ for the extracellular measured species in dataset 1 (left-hand side) and in dataset 2 (right-hand side) – positivity constraint violation (blue curves) when $n_{EFM} \geq m_e$ ($n_{EFM} = 173$ and $m_e = 6$); model discrepancy (pink curves) when $n_{EFM} < m_e$ ($n_{EFM} = 5$ and $m_e = 6$).

3.5 Prospects

Although the macroscopic modeling framework developed in this study has demonstrated its effectiveness in deriving meaningful metabolic models for animal cell cultures, several avenues for future research and improvement remain open. These perspectives primarily concern enhancements in the selection of elementary flux modes (EFMs), refinements in kinetic modeling approaches, and the exploration of alternative computational strategies.

3.5.1 EFM Selection Process

First, the EFM reduction procedure, although effective, could benefit from the integration of alternative selection strategies. The current approach relies on cosine similarity and optimization-based reductions, followed by a selection step optimizing the performance index Ξ . Future work could investigate more sophisticated techniques, such as clustering-based reductions or probabilistic approaches, to identify the most relevant elementary flux modes.

Another promising direction involves improving the generation and identification of subsets of EFMs when full enumeration becomes computationally prohibitive. Beyond the current approach of generating random subsets of modes (using the algorithm developed in [82]), advanced algorithms could be explored to extract the most informative EFMs. For example, methods such as the ones introduced in Sec. 3.1.1 can be exploited, particularly the method developed in [95] that leads to an accelerated enumeration of extreme pathways through a positive-definite elementary set. This technique allows for a more computationally efficient identification of relevant EFMs, further reducing the complexity of large-scale metabolic models.

3.5.2 Kinetic Modeling

Regarding kinetic modeling, while multilayer perceptrons (MLPs) have been used successfully for reaction rate identification, alternative neural architectures could provide additional benefits. Recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, for example, could better capture temporal dependencies in metabolic fluxes, particularly for dynamic bioprocesses [44]. In addition, physics-informed neural networks (PINNs), which integrate mechanistic constraints into deep learning models, present an exciting avenue to combine mechanistic and data-driven approaches in a biologically consistent manner [142].

Further improvements could also be made in the pruning techniques used for ANN-based modeling. The current pruning strategy, based on the magnitude of the weight values, is relatively simple and could be replaced or complemented by more advanced methods such as Hessian-based pruning or structured sparsity techniques [143]. These structures could lead to more compact and computationally efficient networks without compromising predictive accuracy.

3.5.3 Further Considerations

An interesting research direction would be the integration of additional biological constraints into the macroscopic modeling framework. This could include regulatory constraints inferred from transcriptomic or proteomic data, metabolic thermodynamics, or dynamic constraints based on metabolite turnover rates. Such extensions would enhance

the biological relevance and predictive power of the models, ultimately improving their applicability for bioprocess optimization and control.

The aforementioned perspectives constitute only a subset of potential improvements. Several other avenues remain open, depending on experimental constraints and research objectives. One alternative worth exploring, though without a guaranteed improvement in results, would be to directly compute total reaction rates instead of specific reaction rates from the EFM reduction procedure. This approach would eliminate the dependence on interpolated biomass values, potentially reducing error propagation and biases introduced by biomass estimation in the identification process. Following the formalism of the mass balance equation system in Eq. 3.17, Eq. D.2 and Eq. D.3 become:

$$\underline{\kappa}_m(t_k) = \left(\frac{d\underline{\xi}(t_k)}{dt} - D(t_k)(\underline{\xi}_{in}(t_k) - \underline{\xi}(t_k)) \right) \quad (3.31)$$

$$\underline{\kappa}_m(t_k) = K \underline{r}(t_k) \quad (3.32)$$

And the performance index in Eq. 3.2 in the EFM reduction procedure becomes:

$$\begin{aligned} \Xi &= \sum_{k=1}^M \left(K_e \underline{r}_k - \underline{\kappa}_{m,k} \right) W^{-1} \left(K_e \underline{r}_k - \underline{\kappa}_{m,k} \right)^T \\ \min_{\underline{r}_k} \Xi \quad &s.t. \quad \underline{r}_k \geq 0 \end{aligned} \quad (3.33)$$

Nevertheless, it is important to note that specific reaction rates are more biologically interpretable. Metabolic models based on Flux Balance Analysis (FBA) and Elementary Flux Mode (EFM) analysis primarily work with specific reaction rates, and many kinetic laws are often formulated in terms of specific rates rather than total rates.

3.6 Conclusions

In this work, a novel macroscopic modeling framework for the metabolic modeling of bioprocesses was developed and applied to animal cell cultures, namely, Hybridoma Cell Line HB-58 and Chinese Hamster Ovary Cell Line CHO-320. The proposed methodology follows a **two**-step approach: **(i)** the derivation of a macroscopic reaction scheme based on a reduced set of elementary flux modes (EFMs), and **(ii)** the modeling of the reaction rates using either classical kinetic laws or artificial neural networks (ANNs), depending on the complexity of flux patterns. The methodology ensures that the resulting models remain both interpretable and predictive, offering valuable insights into the metabolic behavior of these cell cultures.

For the derivation of the macroscopic reaction scheme, the study introduces a systematic EFM reduction procedure to overcome the combinatorial explosion associated with the full EFM enumeration. This approach, combining cosine similarity filtering and optimization-based reductions, allows the selection of a sub-optimal yet informative set of EFMs. More precisely, this process consists of **four** main steps: **(1)** the generation of EFMs which involves either a full enumeration for moderate-sized networks or the selection of representative subsets when the full enumeration is computationally prohibitive or impossible; **(2)** a biological interpretation check, ensuring that EFMs correspond to biologically meaningful macroreactions; **(3)** the main reduction step, which involves cosine-similarity filtering followed by optimization-based reductions to retain

informative subsets of EFMs; and (4) the final selection step, where the best subset of EFMs is chosen based on a performance index. This structured methodology provides a reduced stoichiometric matrix of the macroreactions and the corresponding time evolution of the reaction rates computed numerically from the optimization problem. For reaction rate modeling, classical kinetic laws are used when well-defined relationships with extracellular measurements exist, whereas MLP-based neural networks are applied in cases of complex flux dependencies and highly nonlinear behaviors.

The overall framework is validated through **two** case studies: **(A)** perfusion cultures of Hybridoma HB-58 cells, and **(B)** batch cultures of CHO-320 cells during the growth phase. In the first case study, a metabolic network comprising 70 biochemical reactions and 44 internal metabolites is considered, and a scenario of six extracellular measurements is handled. Since the complete EFM enumeration is feasible, the complete set of modes is identified before applying the reduction part of the procedure. Reaction rates are modeled using artificial neural networks (MLPs) due to the complexity of flux patterns. In the second case study, a larger metabolic network is examined, with 100 biochemical reactions and 72 internal metabolites. Given the prohibitive number of EFMs (approximately one billion), only subsets of modes are generated for further reduction. In this case, reaction rates are modeled using classical kinetic laws due to biological insights and well-defined relationships between flux patterns and extracellular measurements.

Despite the success of the developed methodology, limitations remain. Although the EFM reduction procedure has proven to be highly effective in obtaining a small yet biologically meaningful reaction scheme, the use of advanced algorithms to generate subsets of EFMs might be investigated. Furthermore, the application of advanced pruning techniques and alternative neural architectures, such as PINNs, could enhance the efficiency and interpretability of ANN-based models.

Beyond the case studies presented in this manuscript, the EFM reduction procedure has been applied to various other metabolic networks of different sizes, alternative measurement configurations, and different cell lines. These additional investigations provide further validation of the approach and highlight its versatility. However, for the sake of conciseness, these results are not included in detail in the present document. These case studies are mentioned below:

- Hybridoma Cell Line HB-58 in perfusion mode;
 - Metabolic network with 24 reactions and 17 internal metabolites ($N^{17 \times 24}$) with 5 extracellular measurements ($m_e = 5$) (see [5]);
 - Metabolic network with 70 reactions and 44 internal metabolites ($N^{44 \times 70}$) with 5 and 22 extracellular measurements ($m_e = 5$ or 22) (see [5]);
 - Metabolic network with 126 reactions and 86 internal metabolites ($N^{86 \times 126}$) with 22 extracellular measurements ($m_e = 22$).
- Chinese Hamster Ovary Cell Line CHO-320 in batch cultures (growth phase);
 - Metabolic network with 100 reactions and 72 internal metabolites ($N^{72 \times 100}$) with 20 extracellular measurements ($m_e = 20$) (see [112]).
- Animal cell cultures in fed-batch mode operated with multi-input feeding of glucose, glutamine, and amino-acid media.

- Metabolic network with 24 reactions and 17 internal metabolites ($N^{17 \times 24}$) with 5 extracellular measurements ($m_e = 5$);
- Metabolic network with 70 reactions and 44 internal metabolites ($N^{44 \times 70}$) with 22 extracellular measurements ($m_e = 22$).

Overall, the macroscopic modeling approach developed in this chapter represents a significant advance in the field of metabolic modeling for bioprocesses. By bridging mechanistic and data-driven modeling strategies, it offers a flexible yet biologically meaningful methodology to capture the metabolic behavior of complex cell cultures. The insights gained from this work pave the way for future enhancements in bioprocess modeling, optimization, and control, contributing to the broader goal of improving industrial bioproduction efficiency.

Chapter 4

Microscopic Modeling

Metabolism is the foundation of all cellular processes, enabling cells to convert nutrients into energy and biomass while maintaining homeostasis. Therefore, the ability to model cellular metabolism with precision is essential for advancing the understanding of biological systems and optimizing biotechnological applications. As outlined in Chap. 1, this thesis follows a structured approach to metabolic modeling, progressively moving from macroscopic to microscopic levels. While the previous chapter focused on macroscopic metabolic modeling, in which cellular metabolism is described in terms of extracellular inputs and outputs, this chapter delves into the microscopic level, where intracellular fluxes and reactions are explicitly quantified.

Microscopic metabolic modeling seeks to describe the complex network of intracellular biochemical reactions that sustain cellular function. A key component of this process is the definition of a metabolic network, represented mathematically by its stoichiometric matrix. This matrix encodes all intracellular reactions that occur in the cell, allowing a rigorous and quantitative analysis of metabolic flux distributions. Afterward, once the metabolic network is defined, flux balance analysis (FBA), flux variability analysis (FVA), and other metabolic flux analysis (MFA) techniques can be applied to estimate intracellular reaction rates. Unlike macroscopic approaches, which primarily focus on overall nutrient consumption and product formation, microscopic modeling provides insights into how cells allocate resources and regulate their metabolic pathways.

The development of a metabolic network is not an easy task and goes beyond the process of collecting biochemical reactions. A significant challenge in the elaboration of metabolic networks lies in ensuring biological accuracy while maintaining computational feasibility. It requires ensuring that the entire set of reactions is biochemically and mathematically consistent, which involves careful consideration of reaction reversibility, ensuring that there are no gaps in the network reconstruction where metabolites would otherwise accumulate or disappear without biochemical justification, and verifying mass and charge balance across all reactions.

For this purpose, this chapter presents a step-by-step and iterative methodology that combines biological knowledge with mathematical techniques to develop metabolic networks and validate the network structure. Special attention is given to the size of the network, ensuring that it is neither too large (which would increase computational complexity, particularly for elementary flux mode (EFM) analysis) nor too small (which could lead to an oversimplified representation of cellular metabolism). Another focus of this study is the modeling of the proton motive force (PMF), a key element in cellular

energy metabolism that drives ATP synthesis and other essential bioenergetic processes.

The applications and case studies addressed in this chapter are related to the modeling of photosynthetic microorganisms due to their metabolic complexity and ecological significance. Understanding the metabolism of these organisms is crucial not only for fundamental biological research but also for biotechnological applications, including biofuel production, carbon sequestration, and bioproduct synthesis. However, modeling such systems poses additional challenges, particularly because some internal metabolites can accumulate over time. Therefore, the pseudo-steady-state assumption that underpins many classical MFA techniques does no longer hold (see Sec. 2.4).

To address this issue, two case studies are explored. The first involves modeling the metabolism of a strain of *Arthrospira* sp. under constant light conditions, where the pseudo-steady-state assumption remains valid. The second extends the definition of an existing metabolic network for unicellular microalgae and takes into account the accumulation of internal metabolites when applying MFA techniques. This second case study serves a dual purpose: (i) validate the definition of the metabolic network, and (ii) demonstrate how metabolic modeling methodologies can be effectively applied when classical assumptions no longer hold. In this study, the method proposed by [2] is applied to account for internal metabolite accumulation, and a macroscopic model is derived based on the definition of the metabolic network and the concept of EFMs. This case study thus bridges microscopic and macroscopic modeling approaches, reinforcing the multi-scale nature of cellular systems and, more importantly, underscoring the critical role of metabolic network development in both microscopic flux analysis and macroscopic reaction scheme inference based on EFMs, as discussed in Chap. 3.

Chapter Outline

The remainder of this chapter is structured as follows:

- **Literature Review** - This section provides an overview of existing methodologies for the elaboration of metabolic networks. It reviews approaches ranging from genome-scale reconstructions to reduced metabolic models and discusses different techniques for ensuring network consistency. Furthermore, this section positions the proposed methodology within the landscape of existing approaches;
- **Theoretical Background and Methods** - This section presents the constructive methodology for the development of metabolic networks introducing the iterative approach that integrates biological knowledge and mathematical methods. In addition, it discusses the incorporation of the proton motive force (PMF) into the metabolic network to improve the representation of energy metabolism;
- **Case Studies** - Two case studies are presented to illustrate the application of the methodology. The first case study considers the development of a metabolic network for a strain of *Arthrospira* sp. under constant light regimes, where the pseudo-steady-state assumption is valid. The second case study extends an existing network for unicellular microalgae, addressing internal metabolite accumulation and demonstrating how classical metabolic flux analysis (MFA) techniques can be used under non-balanced growth. The second case study also serves as a validation of the proposed methodology by linking metabolic network definition to elementary flux mode analysis;

- **Future Perspectives** - This part explores potential extensions of the methodology. It also discusses the interest of modeling phycobiliproteins in the metabolic network of *Arthrospira* sp. (high economical value);
- **Conclusions** - This section summarizes the key contributions of the constructive methodology, emphasizing the importance of a structured and validated approach to metabolic network construction. The implications of the findings are highlighted for both theoretical research and practical applications.

Contributions

The primary contributions of this chapter are as follows:

- A constructive methodology for the development of mid-size metabolic networks that combines both biological knowledge and mathematical methods. In addition, this bottom-up approach minimizes reliance on genome-scale reconstructions while striking a balance between excessive computational complexity and oversimplified representation of cellular metabolism;
- The developed metabolic networks incorporate energy aspects such as the proton motive force to enhance the representation of cellular energy metabolism;
- The methodology is applied to photosynthetic microorganisms, providing insights into their complex metabolic regulation. In addition, challenges related to the accumulation of internal metabolites are addressed.

By addressing these key aspects, this work provides a step-by-step methodology to construct a consistent metabolic network for any organism and derive its corresponding stoichiometric matrix, which will serve as the basis for the application of other metabolic flux analysis techniques. See more information in Sec. 1.2.

List of Publications

Journal Paper

- Maton, M.; Leroy, B.; Vande Wouwer, A. A Constructive Bottom-Up Approach for the Elaboration of Metabolic Networks: Case Study of Photosynthetic Cyanobacteria *Arthrospira Spirulina Platensis* PCC 8005. *Biochemical Engineering Journal* **2025**, *221*, 109770.

Conference Papers as Regular Papers

- Maton, M.; Vande Wouwer, A. Mathematical Modeling of Photosynthetic Eukaryotic Microorganisms Using Metabolic Networks. *IFAC-PapersOnLine*. 11th Vienna International Conference on Mathematical Modelling MATHMOD 2025: Vienna, Austria, 19-21 February 2025. **2025**, *59 (1)*, 397 - 402.
- Maton, M.; Leroy, B. ; Vande Wouwer, A. Metabolic Modeling of *Arthrospira* Sp. PCC 8005 - Network Definition and Experimental Validation. *IFAC-PapersOnLine*. 14th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems DYCOPS 2025: Bratislava, Slovakia, 16-19 June 2025. **2025**, *59 (6)*, 391 - 396.

4.1 Literature Review

Mathematical modeling of microbial metabolism is a crucial aspect of biotechnology, system biology, and ecological research. Metabolic modeling is becoming a powerful tool for understanding, predicting, manipulating and optimizing the complex metabolic processes of microorganisms [19]. In particular, this framework allows for studying the metabolism of a living cell and acquiring closer knowledge of the physiological and biochemical properties of the cell under study. In that context, the use of metabolic networks is particularly interesting despite the difficulties underlying their reconstruction. In this way, metabolic networks are used in various applications in biotechnology, medicine, and environmental science and enable scientists to gain insights into the operation of organisms and processes of energy production, nutrient utilization, and waste elimination, among others. However, constructing a metabolic network is not an easy task. Indeed, it requires integrating high-quality, reliable and available data from multiple sources, including genomics, transcriptomics, proteomics, and metabolomics. In addition, metabolic networks are highly interconnected, making their reconstruction time-consuming and their analyses computationally intensive. Furthermore, to ensure their accuracy and predictive capability, an iterative process of model refinement is needed, comparing model predictions with experimental observations and leading to updating the network structure when necessary.

4.1.1 Genome-Scale Metabolic Networks

With the advances in sequencing techniques and genome annotation methods, complete genome sequences are available for many microorganisms, enabling the generation of genome-scale metabolic reconstructions and leading to genome-scale metabolic models [144–146]. However, to ensure the consistency of such reconstructions and comprehend the organizational principles of metabolic interactions within cellular networks, examination of the structural and topological properties of metabolic networks is important. In the past, many studies [26, 147–149] have paved the way and have covered this issue, e.g., the identification of groups of reactions unable to carry a flux due to the stoichiometry of the network under steady-state conditions or the establishment of enzyme subsets, i.e., groups of reactions operating together in fixed flux proportions at steady-state. Such analyses enable identifying potentially missing reactions as well as reactions under coordinated regulation and allow a continuous refinement of metabolic reconstructions through an iterative model-building process [150–153].

Similarly, frameworks based on convex analysis have been developed to examine structural and topological network properties. Those methods are mainly based on a series of constraints that govern the operation of the metabolic network at steady-state, including stoichiometric and thermodynamic constraints, which limit the range of possible network behaviors corresponding to different metabolic phenotypes. For instance, flux balance analysis is a mathematical approach for analyzing the flow of metabolites through the network and identifies an optimal flux distribution characterizing a particular phenotype [117]. Other techniques rely on the concept of extreme pathways [154] and elementary modes [21, 155] and are often used as a quantitative measure of network robustness. However, even if constraint-based modeling and analysis have attracted huge interest in the last decades, most algorithms for the identification of elementary modes and extreme pathways do not scale well for genome-scale models of complex microorganisms due to the combinatorial explosion of the identified modes. In that context, several mathematical techniques and algorithms have recently emerged to reduce the number of modes and identify the most informative ones [88, 112]. In addition to

these popular approaches, flux coupling analysis (FCA) is concerned with describing dependencies between metabolic reactions and finding coupled reaction sets and blocked reactions in genome-scale metabolic systems [156, 157]. The latter method is essentially used for exploring biological questions such as network evolution, gene essentiality, and gene regulation and requires solving a series of linear programs.

4.1.2 Small-Size Metabolic Networks

The above methods are efficient and take part in the model-building process. However, they require the computation of null-space matrices or involve the solution of a huge sequence of linear programs that can be computationally costly. Another challenge with genome-scale networks is the complexity involved in constructing the stoichiometric matrix, which serves as the foundation for many computational analyses. Therefore, building an accurate and balanced stoichiometric matrix can be time-consuming and prone to errors. Additionally, genome-scale networks depend on the availability of high-quality genomic data and comprehensive gene annotations, which are not always accessible or reliable, particularly for lesser-studied or newly isolated strains.

Furthermore, although the models deduced from genome-scale networks provide a complete skeleton of the metabolic reactions taking place in an organism, these models are difficult to analyze, interpret biologically and use for control and optimization strategies because of their complexity and their dimension [158]. Henceforth, to facilitate the analyses and the application of control strategies, small-size metabolic networks have been deduced from larger networks by focusing only on the main metabolic functions of the microorganisms [7, 13, 29, 159–161]. Such networks often encompass essential pathways such as glycolysis, the citric acid cycle, and other specific biosynthetic pathways and lead to simplified models that are used to understand the fundamental aspects of cellular metabolism and the principles of enzymatic activity and regulatory mechanisms. However, even though they are easier to manage than large-scale networks and require reduced computational resources, small-size metabolic networks are often oversimplified. Hence, important interactions that occur in the full network are omitted, and insights gained from the simplified networks might not be directly generalized to more complex systems. In addition, the behavior and effectiveness of the network can vary depending on specific environmental conditions and the physiological state of the organism, which are key aspects that are not represented in small-size networks.

4.1.3 Mid-Size Metabolic Networks

A middle ground is then the development of metabolic networks of intermediate size, which are detailed enough to capture the complexity of the cellular metabolism while remaining straightforward to analyze at the process level. These networks can be constructed using different approaches in metabolic modeling. On the one hand, top-down approaches start with a comprehensive genome-scale or a large model and aim to simplify or reduce the network by applying reduction algorithms and techniques like transcriptomics or proteomics-based filtering for further refinement [162–165]. These methods typically rely on the structure of the genome-scale model. For this purpose, powerful algorithms have emerged in recent years, such as *NetworkReducer* [158], *MinNM* or *MetNet* [166] among others. Nevertheless, these methods inherit many of the same limitations as genome-scale networks themselves (i.e., availability of genomic data, errors or gaps in the original genomic reconstruction, the complexity and the size of the original genome-scale model as well as computational challenges associated with handling large matrices). On the other hand, bottom-up approaches consist of constructing metabolic

networks by carefully curating reactions based on biological insight, biochemical and physiological data, and experimental findings. These methods often simplify the networks in a context-specific manner. For instance, the elaboration of the network can start from condition-specific metabolome data [162] or mass spectroscopy data [167]. Then, metabolic networks are constructed using Gaussian Graphical Models, statistical techniques for dimensionality reduction [167], or optimization-based methods to infer the active networks [162].

More recently, studies have been conducted to leverage both top-down and bottom-up strategies, leading to a flexible methodology for constructing functional and testable models. This hybrid approach enables enhancing model accuracy and biological relevance and is based on graph-based transformation and stoichiometric and retro-synthetic analyses [168, 169]. Nevertheless, it is worth noting that, in many cases, methods used for metabolic network construction (both top-down or bottom-up approaches) utilize genome-scale metabolic models (GEMs) as foundational components, though their usage can differ depending on the approach [164]. As a matter of fact, top-down approaches use GEMs as a starting point for simplification, bottom-up approaches begin with experimental data to construct smaller networks but still sometimes make use of GEMs to ensure a broader metabolic context (like *redGEM* and *lumpGEM* in [164]) and hybrid methods leverage both approaches using GEMs and experimental data to build and refine the model.

The methodology introduced in this chapter offers an alternative to existing methods and can be considered as a constructive approach to derive a metabolic network. In contrast to current methods, it is a bottom-up approach with minimal reliance on genomic reconstructions. Therefore, it proves particularly valuable in cases where the strain under study is more specific and genomic reconstructions are not readily accessible in the literature. The method is basic and simple, yet it is modular. Specifically, it enables the development of metabolic networks either from scratch or by incorporating experimental data, genomic reconstructions, and biochemical databases if available. As discussed previously, the inferred networks typically comprise a couple of hundred reactions, which are sufficiently detailed to accurately reflect the complexity of the cellular metabolism but remain simple to analyze at the process level. Furthermore, this hands-on strategy provides a deep understanding of the metabolic layout and highlights the connections between key pathways, which might be lost in networks simplified from genome-scale models.

As illustrated in Fig. 4.1, the method is, in essence, iterative and can also make use of genomic and genetic studies, as background material, if available. When applicable, genomic studies are exploited only to identify key metabolites and detect specific reactions proper to the strain under study (in contrast to top-down approaches, where genome-scale models are simplified using reduction algorithms). In that case, the modeling procedure also requires the use of lumping techniques [170–172] to overlook some intermediate metabolites. Finally, the approach uses constraint-based methods [173–176] to ensure network consistency, as illustrated in Fig. 4.2 (more details will be given in the sequel of the chapter). Furthermore, it uses in-silico and experimental data to quantify the intracellular mechanisms and refine the metabolic network definition. The contributions of the present work are then twofold: (i) propose a constructive approach to elaborate a metabolic network of intermediate size and (ii) establish a comprehensive network for photosynthetic microorganisms, as will be described in the case studies.

The original aspect of this constructive approach lies (i) in the use of the above-

mentioned techniques sequentially and iteratively while ensuring network consistency, (ii) in the attempt of modeling the proton motive force, which is an electrochemical gradient driving the synthesis of ATP, and (iii) in the elaboration of a fully generalized network, in contrast to the context-specific designs typical of most bottom-up approaches. Note that this general framework, especially with respect to reaction reversibility, enables broader adaptability and applicability across varied metabolic contexts. The sequential and iterative exploitation of the mathematical methods is illustrated in Fig. 4.2 and will be detailed in the dedicated section of this chapter.

4.2 Constructive Modeling Approach

As introduced previously, the elaboration of a metabolic network is not an easy task. In that context, the present study proposes a constructive methodology to develop an intermediate-size network that is sufficiently detailed to capture the whole metabolism of the microorganism under study but simple enough for control and analysis purposes. Indeed, while GEMs and large networks are comprehensive, their size and redundancy make them computationally expensive, difficult to interpret and to parametrize, and overly complex for dynamic analysis and control. In contrast, small networks may oversimplify the system dynamics and fail to capture essential metabolic interactions. In that respect, mid-size networks retain only key pathways and metabolites relevant to the biological or engineering objective, providing sufficient detail for meaningful analysis while remaining tractable for dynamic studies and control. Moreover, reducing network size also mitigates issues such as excessive multiplicity in constrained optimization solutions, which can arise in GEMs from redundant or rarely active pathways. Although some multiplicity reflects realistic biological scenarios, intermediate-size networks enable reducing irrelevant alternative solutions and allow highlighting the most critical metabolic strategies. Additionally, mid-size networks offer practical advantages in control strategies beyond reducing the number of dynamic states through pseudo-steady-state assumptions. They simplify the structure of the model, making it more suitable for feedback design, linearization, and robustness analysis (supporting dynamic optimization and control).

Nonetheless, it is worth noting that the notion of *intermediate-size* or *mid-size* networks is subjective and depends on the organism, the research context and the objectives. The choice of this scale is often influenced by available experimental data, dynamic behaviors under investigation and computational considerations. As such, intermediate-size metabolic networks are typically developed through iterative refinement, including only the necessary pathways for meaningful insight, while ensuring that the model remains manageable for dynamic simulation and control purposes. All of this ensures flexibility, tailoring the network size and complexity to meet specific research needs.

The modeling procedure is illustrated in Fig. 4.1 and Fig. 4.2. Fig. 4.1 positions the approach within the landscape of existing methods, and Fig. 4.2 depicts the algorithmic scheme of the mathematical methods. The first step of the methodology involves collecting metabolic reactions that define the metabolism of the microorganism from the literature. When genome sequences of the strain are available, specific reactions proper to the strain under study might be added. In that case, lumping techniques are applied to overlook nonessential intermediate metabolites and identify only the main reactions and key metabolites. Furthermore, because the reversibility of chemical reactions is essential to ensure the consistency of the network, genetic studies and biochemical databases (such as KeGG) are used to refine the development of the network and to develop a

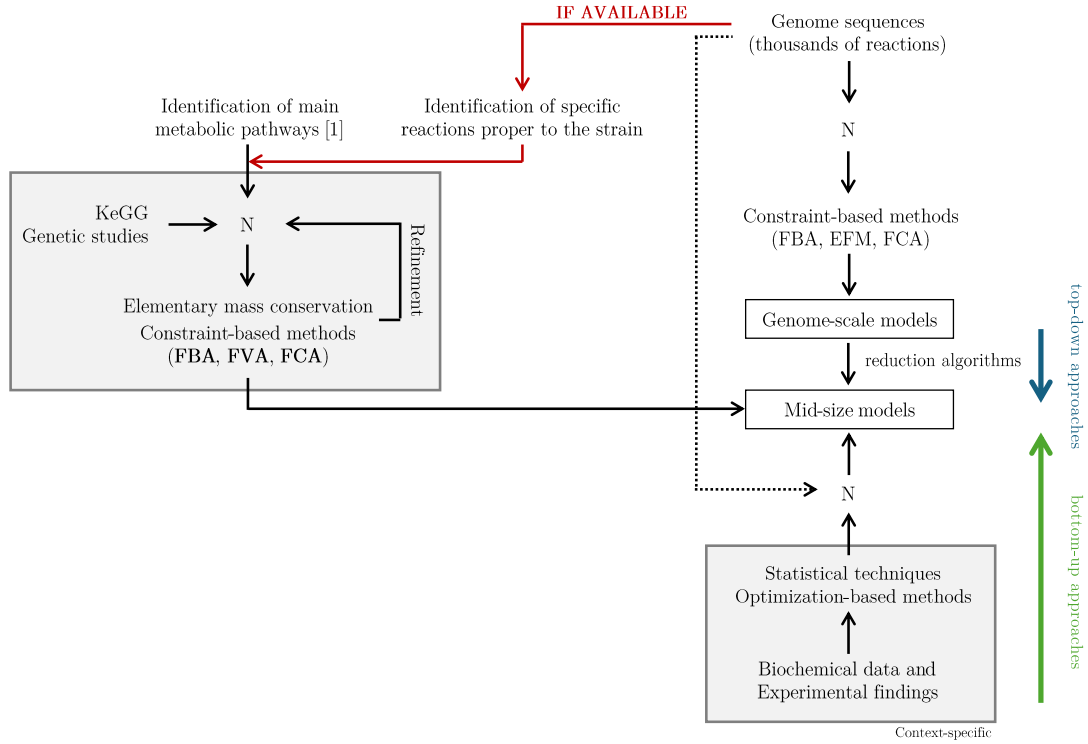


Figure 4.1: Existing modeling approaches on the right-hand side and the proposed constructive methodology on the left-hand side.

fully generalized network. At this stage, it is possible to build the stoichiometric matrix N that characterizes the metabolic network and finally apply constraint-based methods to validate the network or update the network structure when necessary. As mentioned previously, the proposed modeling strategy stands out from existing methods (classical top-down and bottom-up approaches) due to its minimal reliance on genome-scale reconstruction.

4.2.1 Core Metabolic Pathway Selection

The constructive procedure to elaborate a metabolic network starts with identifying the main metabolic pathways that describe the metabolism of the cell under study. Even though it is presumably the most arduous and time-consuming step of the methodology, this stage is essential to understand the operation of the organism and obtain a consistent network. There is no magic recipe for building a coherent network. First, it is crucial to determine the kind of organism under study. Indeed, although most metabolic routes are similar for most organisms, other pathways are very specific and can affect the whole metabolism. Essentially, organisms are broadly classified into different categories based on various characteristics, such as cell structure, mode of nutrition, reproduction, and evolutionary history (see Sec. 2.1). For instance, regarding the case study presented in Sec. 4.3, *Arthrospira* sp. PCC 8005 are prokaryotes, more specifically, photosynthetic cyanobacteria. Then, after investigating the kind of organism, consulting the literature and surveys is the most relevant way to gather information and identify the main

metabolic routes that form the network. For this purpose, biochemistry textbooks, on-line biochemical databases, and genome sequences (when available) can be used.

Because genome sequences gather thousands of biochemical reactions, lumping techniques can be applied to reduce the dimensionality and complexity of the identified network. In this study, lumping involves combining similar chemical species or reactions into a single entry to simplify the mathematical modeling. To illustrate the lumping operation, the following reactions are considered, representative of a simplified biochemical process:



The metabolite A is converted to B through enzyme E_1 , B is converted to D via enzyme E_2 , and C_1 and C_2 are intermediate complexes. To simplify the network, intermediate states C_1 and C_2 , and the transient binding of enzymes are disregarded to focus only on the net conversion of the metabolite A to the metabolite D . The lumped reaction is then:



In lumping techniques, the intermediate metabolites that are eliminated from the model are typically those that do not serve as precursors for the synthesis of other essential compounds within the network. These metabolites serve as transient intermediates in specific pathways without contributing to broader connectivity or downstream processes. By removing these non-essential intermediates, the complexity of the network is reduced while retaining its core functional and mechanistic properties.

4.2.2 Reaction Reversibility

After identifying the main metabolic pathways that define the metabolism of the microorganism of interest, it is crucial to study whether a reaction is reversible. Indeed, the reversibility of reactions in metabolic networks is a critical aspect that influences the efficiency, flexibility, and regulation of metabolic pathways simultaneously. Reversible reactions allow cells to adapt to environmental changes, maintain metabolic balance, and optimize resource utilization. For *Arthrospira* sp. PCC 8005, this aspect is particularly important to take into account because cyanobacteria behave differently depending on light conditions. For example, experimental observation demonstrates a large accumulation of lipids and carbohydrates during the day and their consumption at night. In mathematical modeling, considering reaction reversibility is a standard approach in network reconstruction and can be addressed in multiple ways. As mentioned in many studies [177–181], the most consistent approach is to define concentration ranges for metabolites and to evaluate the thermodynamic feasibility of flux solutions. As will be shown later on, the reversibility of reactions can have a massive impact on the outcomes of metabolic flux analyses and constraint-based methods.

Determining whether a reaction is reversible poses a significant challenge, as it involves several considerations, such as thermodynamics, kinetics, and physiological context. A reaction tends to be spontaneous in the forward direction if the *standard* Gibbs free energy change ΔG° is negative. Henceforth, the reaction is said to be irreversible if this quantity is large and negative, and the reaction is said to be reversible if the quantity is close to zero. This value is a constant characteristic of each specific reaction and

is computed by K_{eq} , the equilibrium constant being experimentally measurable under standard conditions (i.e., 25°C and 101.3 kPa). Nevertheless, in a biological system, the actual concentrations of reactants and products can influence the direction of the reaction. Consequently, the quantity ΔG is preferred and represents the *actual* free energy change. It is computed on the basis of the standard free energy change, but is also a function of reactant and product concentrations and the temperature prevailing during the reaction. In that respect, when ΔG is large and negative, the reaction tends to proceed in the forward direction, and when ΔG is large and positive, the reaction tends to go in the reverse direction.

Besides these thermodynamic aspects, kinetic properties (rate constants, activation energies) and physiological conditions, including enzyme regulation, must be considered to determine whether a reaction is reversible. In that context, biochemical databases and the literature can provide additional information to assess the reversibility of metabolic reactions accurately. The databases BRENDA [182] and KEGG [183] are relevant and helpful resources. In this study, the reversibility of the reactions is stated specifically based on the standard Gibbs free energy change and the enzyme regulation, i.e., the enzyme's capability to catalyze a metabolic reaction in both directions depending on cellular conditions. In addition, reaction reversibility can be tuned and cross-validated by using a constraint-based method to identify blocked reactions in the network. See Sec. 4.2.3.2.4 for more information.

4.2.3 Mathematical Methods

After a representative set of biochemical reactions is collected and the reaction reversibility is determined, it is possible to infer the corresponding stoichiometric matrix of the network. In metabolic engineering, a metabolic network is defined by an $m \times n$ stoichiometric matrix N where m represents the number of internal metabolites and n corresponds to the number of biochemical reactions (see Sec. 2.4.1). Based on the stoichiometric matrix N , several mathematical methods can be applied to ensure network consistency.

Furthermore, when defining the stoichiometric matrix N , it is crucial to differentiate between different metabolites that interact within the cellular metabolism. In this way, the internal balanced metabolites are distinguished from the unbalanced species comprising intracellular energetic cofactors and extracellular metabolites. Moreover, a distinction can be made between internal reactions and exchange or transport reactions to be as rigorous as possible.

This section is devoted to mathematical tools and constraint-based methods for checking/debugging/validating the metabolic network. Fig. 4.2 suggests an algorithmic scheme regarding the sequential and iterative exploitation of the mathematical methods. First, to ensure network consistency, mass conservation is applied to the biochemical reactions of the network. Then, constraint-based methods are exploited in an iterative way. The subsequent sections are dedicated to the description of the algorithms. On the one hand, elementary mass conservation is outlined. On the other hand, the main concepts relative to metabolic network analysis are recalled, the classical assumptions are discussed, and the most popular constraint-based methods are introduced. Finally, an enhanced version of flux coupling analysis is presented, and its interest in metabolic network elaboration is pointed out.

As introduced in Sec. 2.4.1, the general equation of internal metabolite dynamics

can be expressed based on the mass balance principle:

$$\frac{d\mathbf{C}}{dt} = \mathbf{N}\mathbf{v} - \mu\mathbf{C} \quad (4.6)$$

where $\mathbf{C} \in \mathbf{R}^m$ denotes the vector of metabolite concentrations, $\mathbf{v} \in \mathbf{R}^n$ is the vector of network fluxes and $\mu\mathbf{C}$ represents the dilution occurring due to the cell growth. As stated above, $\mathbf{N} \in \mathbf{R}^{m \times n}$ is the stoichiometric matrix where the element N_{ij} is related to the metabolite i participating in the reaction j .

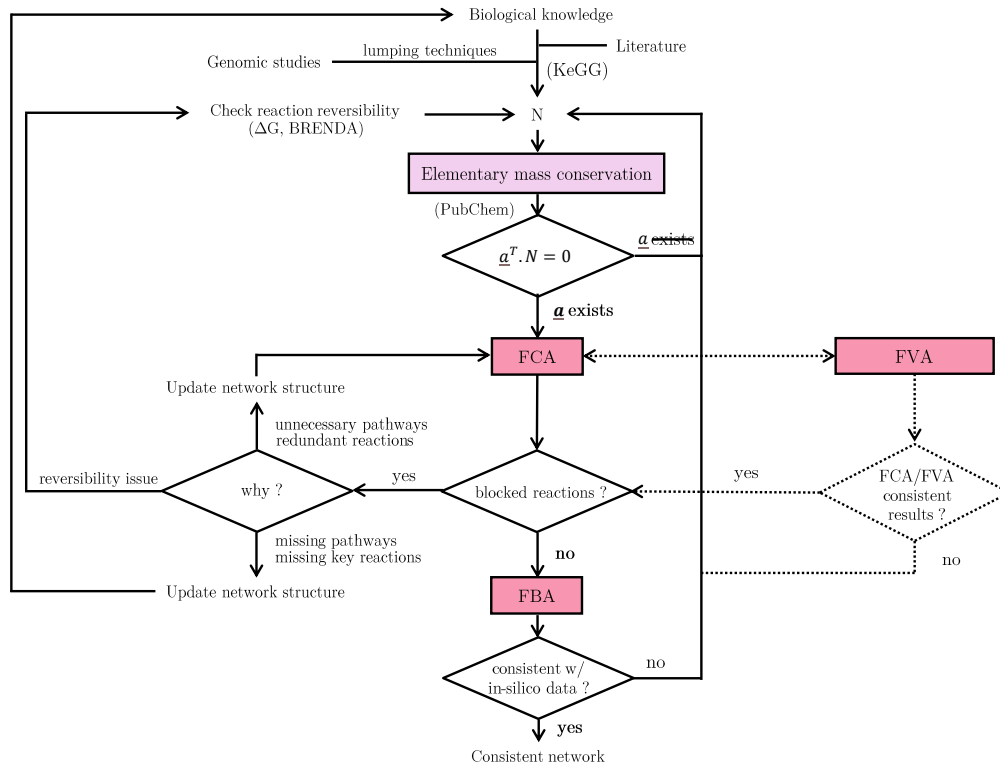


Figure 4.2: Algorithmic scheme of the mathematical methods. Note that FCA is preferred to FVA for larger networks. The latter methods should give equivalent results (cross-validation of the methods).

4.2.3.1 Elementary Mass Conservation

To ensure mass conservation, a variable z can be defined. It represents the total number of relevant atoms involved in metabolic reactions that participate in cellular metabolism. Consider the general equation of internal metabolite dynamics in Eq. 4.6 where the dilution term is neglected because it does not affect the fundamental mass balance:

$$\frac{d\mathbf{C}}{dt} = \mathbf{N}\mathbf{v}(\mathbf{C}) \quad (4.7)$$

where $\mathbf{N}$ is the stoichiometric matrix, $\mathbf{C} \in \mathbf{R}^m$ denotes the vector of metabolite concentrations, $\mathbf{v} \in \mathbf{R}^n$ is the vector of reaction rates. A non-negative vector $\mathbf{a}$ is introduced,

which contains information about the number of atoms (of carbon, nitrogen, or phosphorus) in each metabolite:

$$z = \underline{a}^T \underline{C} \quad (4.8)$$

By differentiating z with respect to time, it follows:

$$\frac{dz}{dt} = \underline{a}^T \frac{d\underline{C}}{dt} = \underline{a}^T N \underline{v}(\underline{C}) \quad (4.9)$$

Henceforth, if mass conservation applies, then z must be constant, so its time derivative is null:

$$\underline{a}^T N \underline{v}(\underline{C}) = 0, \forall \underline{v}(\underline{C}) \rightarrow \underline{a}^T N = 0 \quad (4.10)$$

Therefore, if the vector $\underline{a}$ exists, then there is mass conservation when constructing the reactions of the network. It is important to note that for the mass balance of biomass and macromolecules (proteins, nucleic acids, lipids), an average atomic composition must be considered in agreement with the monomeric composition used to state their biosynthetic reactions. In addition, the elemental composition of biomass is highly variable over time due to factors such as species-specific differences, growth phase, and environmental conditions. This variability is critical and must be considered in metabolic analyses, as it directly influences the stoichiometry of biochemical reactions. However, an average elemental composition is often assumed as a practical simplification to ensure consistency and facilitate the verification of elementary mass conservation in such analyses.

4.2.3.2 Constraint-Based Methods

4.2.3.2.1 Classical Assumptions

4.2.3.2.1.1 Pseudo-Steady-State Assumption This first assumption (QSSA) postulates that internal metabolites reach steady-state instantaneously compared to extracellular compounds so that their transient behavior can be neglected ($\frac{d\underline{C}}{dt} \rightarrow 0$). Indeed, when intracellular metabolites are consumed or produced at rates much higher than the scale of observation, their concentrations reach a dynamic equilibrium rapidly. However, for metabolites with slow turnover, such as cofactors, secondary metabolites, or macromolecule precursors, the steady-state assumption no longer holds because these metabolites may exhibit slow dynamics over the relevant time scale. This point will be discussed further in Sec. 4.4.2.

4.2.3.2.1.2 Effects of Dilution The second assumption states that the influence of dilution due to cell growth is negligible compared to the internal and exchange reaction rates so that $\mu \underline{C}$ vanishes. This assumption is valid for central metabolites where their production and consumption rates exceed the effect of dilution as outlined in [184]. In addition, in mass-conserving metabolic networks, where key chemical elements (e.g., carbon, nitrogen, phosphorus) are balanced, the dilution term is inherently accounted for within flux dynamics [185]. However, dilution must be explicitly considered when it significantly affects metabolite dynamics. For example, in models that include ribosome allocation, enzyme kinetics, or proteome constraints, dilution affects protein/enzyme concentrations and thus must be accounted for [186–188]. This is particularly relevant in mechanistic models like ME-models, where enzyme synthesis, degradation, and dilution are explicitly modeled and directly constrain metabolic capacity. In enzyme-constrained models such as GECKO, enzyme levels are typically derived from proteomics or scaled with growth, and the effects of dilution are incorporated implicitly rather than through dynamic mass balances. More information in Sec. 5.5.5.

In most metabolic network analyses and constrained-based methods, these two assumptions are widely used [117, 155, 189]. In classical MFA techniques such as Flux Balance Analysis (FBA) and Elementary Flux Mode (EFM) analysis, the dilution term is assumed to be negligible by default. In fact, in central metabolism, intracellular metabolites generally turn over much faster than the cell grows. Moreover, because the metabolite pools are small and tightly regulated, their dilution due to growth is negligible compared to reaction rates (metabolite homeostasis).

Therefore, under QSSA, the metabolic steady-state condition simplifies to:

$$N \underline{v} = 0 \quad (4.11)$$

In addition, network fluxes are often subject to positivity constraints assuming that direct reactions prevail over their reverse counterparts:

$$\underline{v} \geq 0 \quad (4.12)$$

System in Eq. 4.11 under the positivity constraints imposed by Eq. 4.12 constitutes a problem of non-negative linear algebra or convex analysis. In any realistic large-scale metabolic model, there are more biological reactions than compounds, leading to under-determined systems for which no unique solution exists (as discussed in Sec. 2.4). Constraint-based methods can alleviate this issue by applying constraints to limit possible solutions and explore the solution space.

4.2.3.2.2 Flux Balance Analysis

Flux balance analysis is undoubtedly the most popular constraint-based method that consists of computing an optimal flux distribution $\underline{v}$ which maximizes or minimizes an objective function $Z = \underline{c}^T \underline{v}$ such that:

$$\underline{v}^{opt} = \arg \max_{\underline{v}} (Z) \text{ s.t. } \{N\underline{v} = 0 ; \underline{v} \geq 0 ; \underline{v}_l \leq \underline{v} \leq \underline{v}_u\} \quad (4.13)$$

where $\underline{c}$ is a vector of weights translating how much each reaction contributes to the objective function, and $\underline{v}_l$, and $\underline{v}_u$ are vectors of lower and upper bounds, respectively. This method assumes that cells regulate their fluxes in order to optimize some relevant biological objectives, e.g., maximization of biomass growth or the ATP production rate [117]. Despite the efficiency of FBA in finding optimal solutions, the latter may not correspond to the actual flux distribution, which, moreover, might not be unique.

4.2.3.2.3 Flux Variability Analysis

Flux variability analysis is another constraint-based method that computes the maximal and minimal values of each reaction flux, and therefore the intervals bounding internal fluxes [190]. In this way, it ensures that the actual flux distribution belongs to the interval. The extreme values of the set of admissible fluxes can be obtained using the FVA approach defined as a series of 2 n linear programs. This means solving two optimization problems for each flux i :

$$\begin{cases} \underline{v}_i^{min} = \min_{\underline{v}} (\underline{v}_i) \text{ s.t. } \{N\underline{v} = 0 ; \underline{v} \geq 0 ; \underline{v}_l \leq \underline{v} \leq \underline{v}_u\}, \\ \underline{v}_i^{max} = \max_{\underline{v}} (\underline{v}_i) \text{ s.t. } \{N\underline{v} = 0 ; \underline{v} \geq 0 ; \underline{v}_l \leq \underline{v} \leq \underline{v}_u\} \end{cases} \quad (4.14)$$

4.2.3.2.4 Flux Coupling Analysis

Flux coupling analysis is another framework for studying the topological and flux connectivity features of genome-scale or large metabolic networks. This method requires the solution of a sequence of linear programs and enables reducing the dimensionality of the network by the identification of pairs of metabolic fluxes. Three types of coupling are distinguished:

- Directional coupling ($v_i \rightarrow v_j$) if a non-zero flux for v_i implies a non-zero flux for v_j but not necessarily the inverse;
- Partial coupling ($v_i \rightleftharpoons v_j$) if a non-zero flux for v_i implies a non-zero flux for v_j and vice versa;
- Full coupling ($v_i \Leftrightarrow v_j$) if a non-zero flux for v_i implies not only a non-zero but a fixed flux for v_j and vice versa.

Pairs of metabolic fluxes that do not belong to one of these categories are considered uncoupled, defining blocked reactions. In view of reducing the number of linear programs to solve, it is suggested to develop an algorithm that identifies only blocked reactions, defined as reactions incapable of carrying a flux under steady-state conditions for a particular uptake scenario. In the context of this study, blocked reactions are used to identify incomplete pathways, to pinpoint errors or omissions in metabolic reconstructions, and also to detect potential problems with reaction reversibility. This last feature appears particularly interesting because, as mentioned in Sec. 4.2.2, the actual Gibbs free energy change ΔG depends on the actual concentrations of reactants and products, which can affect the directionality of the reaction. Furthermore, depending on the uptake scenario, it is feasible to determine inconsistencies in the network and redundant reactions.

Mathematically, the set of blocked reactions is identified by solving the following linear programming problem once for every flux k :

$$\text{maximize} \quad v_k \quad (4.15)$$

$$\text{subject to} \quad \sum_{j=1}^N N_{ij} v_j = 0, \quad \forall i \in M \quad (4.16)$$

$$v_j^{\text{uptake}} \leq v_j^{\text{uptake}^{\text{max}}}, \quad \forall j \in N_{\text{transport}} \quad (4.17)$$

$$v_j \geq 0, \quad \forall j \in N \quad (4.18)$$

It consists in the maximization of a particular flux v_j for a metabolic network comprised of a set of $M = \{1, \dots, M\}$ metabolites and a set of $N = \{1, \dots, N\}$ metabolic reactions. This formulation requires reversible reactions to be expressed as two irreversible reactions in opposite directions, constraining all fluxes to positive values. Therefore, if the maximum value of the flux is zero, then the reaction is said to be unusable or blocked. In addition, it is possible to limit the uptake of resources to the network using the constraint in Eq. 4.17, and the maximum uptake rate of any metabolite absent from the external medium is set to zero. Also, it is important to note that the solution of the linear program depends on the steady-state assumption, the imposed uptake/secretion scenarios, growth requirements, and energy production requirements. Therefore, a blocked reaction does not necessarily mean that the reaction is incorrect, inconsistent, or irrelevant in describing the metabolism of the cell; it could only mean that the reaction is not activated with respect to the actual uptake/secretion scenario.

To evaluate the merits and effective operation of the flux coupling analysis, the algorithm has been performed on existing networks for the modeling of *Tisochrysis lutea* [2], *Chlorella sorokiniana* [191] and the mixotrophic growth of *Chlorella sp.* [192]. The results of the algorithm are consistent, demonstrating its reliability. Therefore, it can be used with confidence for the current network under study. It is important to note that the present analysis identifies *context-specific blocked reactions*, which depend on the imposed uptake and secretion scenarios. This contrasts with *structurally blocked reactions*, which are inherently inactive under all possible conditions due to the network's topology and constraints [157]. As such, structural blockage is more restrictive and context-independent, while context-specific blockage may vary depending on environmental assumptions.

The latter constraint-based techniques are closely related but emphasize different aspects of the feasible flux space. FVA determines the minimum and maximum possible value for each reaction flux under the imposed constraints, requiring the solution of two linear programs per reaction (i.e., $2n$ problems in total). FCA, in contrast, focuses on qualitative dependencies between reactions by assessing whether the flux of one reaction implies the flux of another. This involves solving on the order of n optimization problems and yields a coupling matrix that classifies reaction pairs as fully coupled, directionally coupled, or uncoupled. Both methods can identify blocked reactions, though the interpretation differs: in FVA, they are blocked due to constraint satisfaction, whereas in FCA, they are blocked because of structural interdependencies. For the present study, using both approaches provided complementary insights and allowed cross-validation of the results.

Finally, as mainly discussed in Sec. 4.1, it should be noted that all the methods used to elaborate the **structure** of the metabolic network rely on the quasi-steady-state assumption (QSSA). The network is refined using classical constraint-based tools, such as FBA and FCA, which assume that intracellular metabolite concentrations remain approximately constant over time. This assumption is standard and valid for reconstructing metabolic networks, as it enables the identification of feasible flux distributions and the exploration of topological and functional properties. However, during the experimental **validation** phase (particularly under dynamic environmental conditions such as day/night cycles), this assumption no longer holds – see Sec. 4.4. Therefore, the steady-state assumption is appropriate for defining the **structure** of the metabolic network (i.e., which reactions are possible and how they interact), but not necessarily for modeling the **dynamics** of the metabolism under changing conditions. Using complementary approaches at different stages ensures both robust network reconstruction and realistic dynamic validation.

4.2.4 Cellular Energy Metabolism

Another focus of the current work, in addition to the elaboration of consistent mid-size metabolic networks, is the modeling of the proton motive force (PMF), which is an electrochemical gradient that drives the synthesis of adenosine triphosphate (ATP), a molecule that serves as the primary energy currency of cells.

Essentially, organisms can synthesize ATP through various mechanisms depending on the availability of oxygen and the presence of light. First of all, as it does in all living organisms, the microorganisms under study generate ATP through a process called **substrate-level phosphorylation**, where ATP is generated directly during metabolic pathways such as glycolysis and the citric acid cycle, as mentioned in Sec. 2.2.1. This

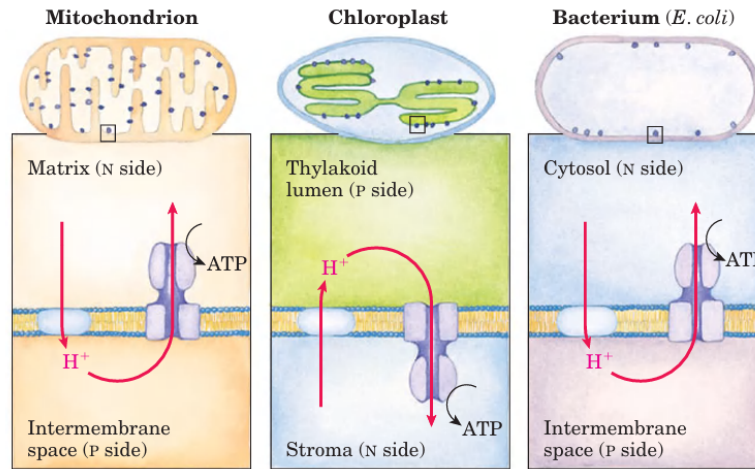


Figure 4.3: Schematic representation of the proton motive force across the mitochondrial membrane (left-hand side), the thylakoid membrane of chloroplasts (center) and the plasma membrane (right-hand side), from [1].

metabolic process generates ATP directly through enzymatic reactions without the involvement of an electron transport chain or a proton gradient or membrane-bound ATP synthase complexes. Therefore, it provides a means to generate ATP quickly in the cytoplasm, particularly under anaerobic conditions or when light is not available for photosynthesis.

In aerobic organisms, ATP is also produced through a process called **oxidative phosphorylation**, where cells use oxygen as the final electron acceptor in the electron transport chain (ETC) to produce ATP. Furthermore, in photosynthetic cells, ATP can be generated under light conditions through a process called **photophosphorylation**, involving the oxidation of H_2O into O_2 , with $NADP^+$ as the ultimate electron acceptor. In these cases, the synthesis of ATP is driven by the proton motive force, denoted PMF, which is a critical component of energy generation and is involved in various cellular processes. This electrochemical gradient is generated across a particular membrane and is based on the chemiosmotic theory that states that the movement of electrons through the electron transport chain creates a proton gradient across a membrane, which drives the synthesis of ATP [1]. Fig. 4.3 depicts the proton motive force in oxidative phosphorylation and photophosphorylation. It can be shown that, as electrons move through the electron transport chain, protons (i.e., H^+ ions) are pumped across the membrane. The pumping of protons creates a higher concentration of H^+ on one side of the membrane, producing a proton gradient that will serve as the driving force for ATP synthesis from ADP and P_i , catalyzed by a membrane-bound ATP synthase complex.

For the process of oxidative phosphorylation, the membrane is the inner mitochondrial membrane in eukaryotic cells or the plasma membrane in prokaryotic cells. For the process of photophosphorylation, the membrane is the thylakoid membrane of chloroplasts in eukaryotic cells or the thylakoid membrane of thylakoid-like membrane structures inside the prokaryotic cells that serve as sites for light-dependent reactions.

Even though the mechanism of ATP synthesis is well documented, the corresponding

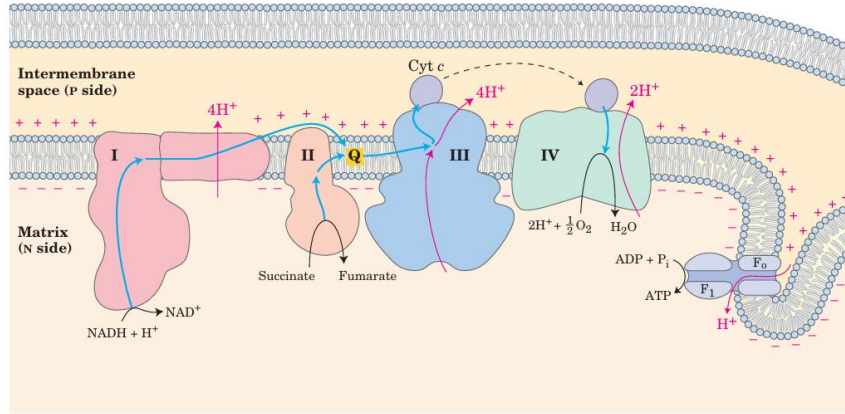
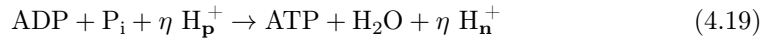


Figure 4.4: Flow of electrons and protons through the 4 complexes of the respiratory chain depicting the process of oxidative phosphorylation, from [1]. Q represents the cofactor FADH_2 .

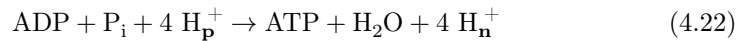
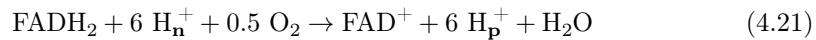
chemical reactions differ according to studies or biochemistry textbooks, which leads to confusion, particularly in mathematical modeling. In that respect, this study proposes to differentiate the cytoplasmic protons H_n^+ (referred to as H^+) and the periplasmic protons H_p^+ , to attempt to model the proton gradient. The synthesis of ATP from ADP and P_i catalyzed by a membrane-bound ATP synthase complex is thus given as follows:



Where η represents the number of protons required to produce one molecule of ATP and varies depending on the specific process and pathway. In photosynthetic microorganisms, four protons are required to form one molecule of ATP ($\rightarrow \eta = 4$).

In cellular respiration, the synthesis of ATP also involves other cofactors (NADH and FADH_2) that contribute to the generation of the gradient proton. The summary of the flow of electrons and protons through the four complexes of the respiratory chain is given in Fig. 4.4, and the detailed chemical equations can be found in [1]. The oxidation of NADH induces the movement of 10 protons (through complexes I, III, and IV) across the membrane, while the oxidation of FADH_2 results in the translocation of 6 protons (through complexes III and IV) across the membrane.

Thus, the simplified equations are written as follows:



Eq. 4.20 represents the oxidation of NADH, Eq. 4.21 depicts the oxidation of FADH_2 and Eq. 4.22 is the reaction of ATP synthesis in photosynthetic cells. Therefore, it follows that 2.5 molecules of ATP are produced per molecule of NADH, and 1.5 molecules of ATP are produced per molecule of FADH_2 .

Photophosphorylation is a light-driven process in which chlorophyll and accessory pigments capture solar energy, driving electron transport through photosystems II and

I. This generates a proton gradient across the thylakoid membrane, which powers ATP synthesis. In oxygenic photosynthesis (as in plants and cyanobacteria), water is split at photosystem II, releasing O_2 , while the ATP and NADPH produced are subsequently used in the Calvin-Benson cycle to fix CO_2 into carbohydrates. The underlying mechanisms of this light-driven energy conversion can be described as follows: during this process, electrons flow through a series of membrane-bound carriers, including cytochromes, quinones, and iron-sulfur proteins embedded in the thylakoid membrane, as depicted in Fig. 4.5. Then, as electrons move through the electron transport chain, protons are pumped across the membrane from the stroma into the thylakoid lumen. The pumping of protons creates a higher concentration of H^+ in the thylakoid lumen compared to the stroma, producing a proton gradient that will serve as the driving force for ATP synthesis from ADP and P_i , catalyzed by a membrane-bound ATP synthase complex. Finally, ATP and NADPH can be used to reduce CO_2 to form triose phosphates, starch, sucrose, and other products (see Sec. 4.3.1.1.4).

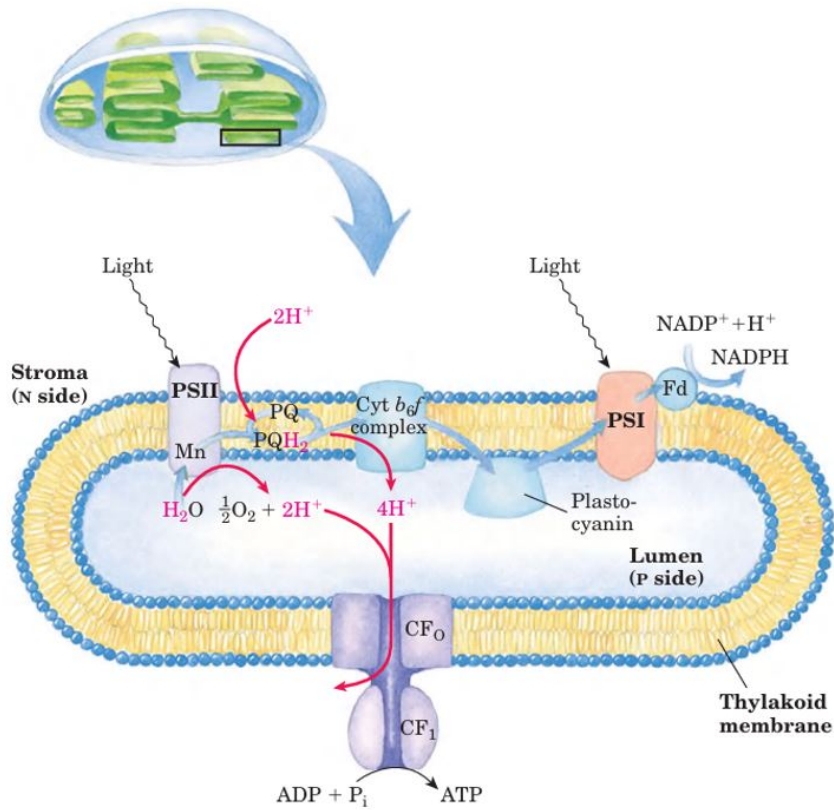
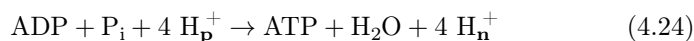
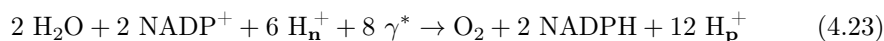


Figure 4.5: Illustration of the process of photophosphorylation, from [1]. Photosystem II (PSII) absorbs photons, splits water into O_2 , H^+ protons and electrons while generating a proton gradient for ATP synthesis. Photosystem I (PSI) then uses the excited electrons to reduce $NADP^+$ into NADPH.

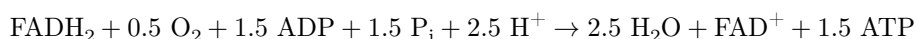
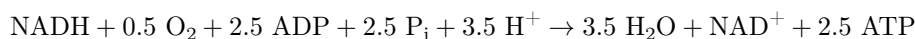
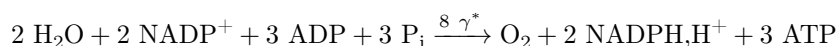
At least 8 photons (denoted γ^*) must be absorbed to drive four electrons from H_2O to $NADP^+$, leading to the movement of 12 protons across the thylakoid membrane. In

this way, the reactions relative to photophosphorylation are obtained:



It is important to note that Eq. 4.23 is not the classical overall redox equation, but expresses the transmembrane proton transfer explicitly. These movements of protons reflect contributions from water oxidation, the Q-cycle in the cytochrome b_6f complex, and NADP^+ reduction (6 protons are removed from the stroma – 4 via Q-cycle, 2 via NADP^+ reduction – and 12 protons accumulate in the lumen – 4 from water oxidation, 8 from PQH_2 release and Q-cycle). Coupled with Eq. 4.24, these H^+ movements generate the proton gradient that drives ATP synthesis. Therefore, this proton gradient can produce approximately 3 molecules of ATP, which is in agreement with the approximate stoichiometry of photophosphorylation established by Peter Mitchell in [1] stating that experimental measurements yield values of about 3 ATP molecules per molecule of O_2 produced.

The previous chemical equations can be compared with the global reactions related to photophosphorylation and oxidative phosphorylation, commonly found in literature:



The equations are in agreement. However, while the latter are more straightforward, they do not express explicitly the proton motive force.

4.3 Application to *Arthrospira* Sp. PCC 8005

Arthrospira *Spirulina platensis* (also referred to as *Limnospira indica*) is a filamentous non- N_2 -fixing cyanobacterium, widely recognized for its nutritional benefits. This species of blue-green algae has become an important source of commercially produced nutraceutical compounds and uses sunlight and CO_2 to produce chemical compounds that are essential for life. For centuries, this cyanobacterium has been consumed by indigenous people for its highly nutritious profile [193]. It has a high protein content (about 60 to 70% of its dry weight) and contains large amounts of healthy nutritional molecules such as beta-carotene, phycocyanin, vitamins, minerals, and essential fatty acids. In addition, this microorganism has considerable therapeutic benefits such as improving the immune system and helping prevent heart disease, cancer, and diabetes [194] and is one of the algae capable of producing bioenergy and renewable energy, which could potentially help reduce the effects of global warming [195]. In addition, they are photosynthetic prokaryotes of crucial importance to the Earth's nitrogen and carbon cycles and in the conversion of solar light into chemical energy and play a significant role in the primary production of biomass and oxygen [196]. With such impressive advantages, especially as a nutraceutical, this cyanobacterium has become one of the preferred choices for a sustainable photobiological cell factory.

The choice for this organism is not trivial. First, *Arthrospira* sp. PCC 8005 is a photosynthetic prokaryote that uses light as an energy source, water as an electron donor, and CO_2 as a carbon source, which makes its metabolism interesting to model. Beyond its terrestrial applications as a source of food supplements, pigments, and environmental bioremediation, this strain has also been selected as the best nutritional resource

for long-haul space exploration missions due to its ability to convert CO₂ into edible biomass and oxygen and its ability to remove certain compounds from contaminated waters [197–199]. Therefore, whether for Earth-based biotechnological processes or for bioregenerative life-support systems in space, a deeper understanding of the metabolic capabilities of this organism is essential to fully harness its biotechnological potential. Besides, studies on similar strains have been conducted in recent years that allow a deep examination and assessment of the results of metabolic analyses, using existing studies as a benchmark. So far, analysis of this specific strain has been scarce. In [200], a basic metabolic flux model containing 22 reactions has been developed to predict rate-limiting enzymes for the production of γ -linolenic acid. Subsequently, a more extensive metabolic network for *S. platensis* has been identified by [201] in a context-specific manner. However, in addition to inconsistencies regarding reaction reversibility, these models did not provide the microorganisms' whole-cell characteristics and metabolic capabilities. Recently, available data and genome sequences of *S. platensis* (different strains) have enabled the development of genome-scale metabolic models [202–204]. While the first models lack details, the others are difficult to comprehend. In this context, the suggested procedure provides a bridge between these two extremes while facilitating the learning and understanding of the metabolism of the studied strain cell, particularly for non-expert users (through a step-by-step constructive approach). After careful refinement of the network structure, the identified metabolic network comprises 198 reactions and 171 metabolites (see the Appendix G) and includes energy aspects such as the proton motive force, barely outlined in existing networks [205].

4.3.1 Elaboration of a Mid-Size Metabolic Network

This section is devoted to applying the iterative modeling procedure depicted in Fig. 4.1 and Fig. 4.2 and described in Sec. 4.2 to the cyanobacteria *Arthrospira* sp. PCC 8005. First, the current section presents the main metabolic pathways of the organism under study. Then, the methodology and the iterative process are described. Finally, the results of metabolic analyses are presented in Sec. 4.3.2 and Sec. 4.3.3 as a validation of the network structure.

4.3.1.1 Main Metabolic Routes

As part of the first step of the constructive methodology, this section presents a simplified overview of the main metabolic routes of *Arthrospira* sp. under photoautotrophic conditions. The metabolic pathways and information related to the strain are mainly derived from the reference textbook on biochemistry [1]. Unlike eukaryotic plants and algae that possess chloroplasts as distinct organelles within their cells, *Arthrospira* sp. PCC 8005 is a prokaryotic photosynthetic organism that has specialized structures that are the sites for light-dependent reactions. These thylakoid-like membrane structures are spread throughout the cell's cytoplasm and contain chlorophyll and other pigments necessary for photosynthesis. Therefore, their metabolism differs from the metabolism of eukaryotic unicellular microalgae studied in [2, 191].

In addition to providing a brief overview of the key pathways, this section aims to point out the basic principles of these metabolic pathways to interpret the results of metabolic analyses later. As depicted in Fig. 4.1, collecting relevant information about the organism under study constitutes the first step of the methodology. The pathways presented in the following subsections are the main routes identified after completing the whole procedure after a couple of iterations. See Table 4.1 for more information related to the iterative procedure.

4.3.1.1.1 Carbon Uptake

Bicarbonate ions serve as the main carbon source for the microorganism under study. These ions enter the cell through active transport. Within the cell, carbonic anhydrase dehydrates these ions into CO_2 , which is then integrated into the Calvin cycle by ribulose-1,5-bisphosphate carboxylase (commonly known as RubisCO). The carbon assimilated is concurrently utilized in producing all major macromolecules in cells, including proteins, carbohydrates, lipids, and nucleic acids, using 3-phosphoglycerate as a key intermediate metabolite. The metabolic reaction corresponding to carbon uptake is denoted v_1 in Table G.1 in the Appendix G.

4.3.1.1.2 Embden-Meyerhof-Parnas Pathway

The Embden-Meyerhof-Parnas pathway, also known as glycolysis, occupies a central position in the metabolism of many microorganisms (see Sec. 2.2.1). Photosynthetic organisms produce glucose by initially reducing atmospheric CO_2 to triose sugars and then converting these trioses into glucose (via the Calvin-Benson cycle and gluconeogenesis). One glucose molecule is degraded through a series of enzyme-driven reactions, producing two molecules of pyruvate, three-carbon compounds. Metabolic reactions corresponding to the Embden-Meyerhof-Parnas pathway correspond to the fluxes v_6 to v_{13} in Table G.1.

4.3.1.1.3 Pentose Phosphate Pathway

The pentose phosphate pathway is depicted in Fig. 2.7 and is described in Sec. 2.2.2. Metabolic reactions related to the pentose phosphate pathway correspond to the fluxes v_{18} to v_{26} in Table G.1.

4.3.1.1.4 Calvin-Benson Cycle

In contrast to animal cells, plants and photosynthetic microorganisms can synthesize carbohydrates directly from CO_2 and water. This process involves the reduction of CO_2 using the energy and reducing power provided by ATP and NADPH produced during light-dependent reactions of photosynthesis. The conversion of CO_2 to simple reduced organic compounds is a process called CO_2 assimilation or the Calvin-Benson cycle and occurs in three stages, as shown in Fig. 4.6.

The assimilation of CO_2 into biomolecules begins with the carbon-fixation reaction, where CO_2 combines with ribulose-1,5-bisphosphate, resulting in the formation of two molecules of 3-phosphoglycerate. In the subsequent stage, these 3-phosphoglycerate molecules are reduced to triose phosphates. Collectively, three molecules of CO_2 react with three molecules of ribulose-1,5-bisphosphate to form six molecules of glyceraldehyde-3-phosphate in equilibrium with dihydroxyacetone phosphate. In the third stage of this cycle, five of the six molecules of triose phosphates are utilized to regenerate three molecules of ribulose-1,5-bisphosphate, returning to the cycle's starting point. The metabolic reactions corresponding to the three stages of the Calvin cycle correspond to the fluxes v_2 to v_5 in Table G.1. The remaining molecule of triose phosphate, which is the net product of photosynthesis, serves as a precursor for the synthesis of hexoses used as energy and structural materials, sucrose for transport to non-photosynthetic tissues, or glycogen for storage, as depicted in Fig. 4.6.

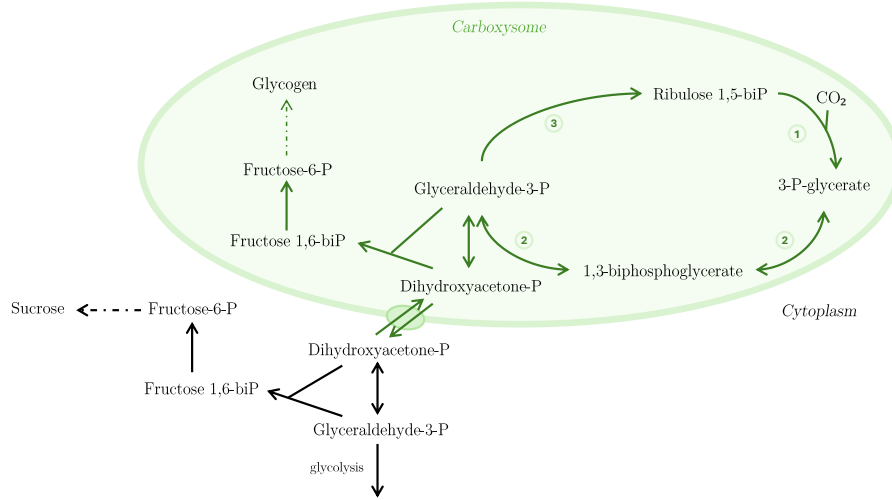


Figure 4.6: The three stages of CO₂ assimilation in photosynthetic organisms: (1) fixation of CO₂ into 3-P-glycerate, (2) conversion of 3-P-glycerate to glyceraldehyde-3-P and (3) regeneration of RubisCO from triose phosphates.

4.3.1.1.5 Tricarboxylic Acid Cycle

The citric acid cycle, also called the tricarboxylic acid cycle or the Krebs cycle, is explained in Sec. 2.2.3. The metabolic reactions related to the citric acid cycle are represented by the fluxes v_{27} to v_{36} in Table G.1 and the anaplerotic reaction is expressed by the flux v_{37} . In addition, it is important to notice that α -ketoglutarate dehydrogenase activity is lacking in cyanobacteria [206]. Therefore, the citric acid cycle reactions are slightly different from those found in most bacteria. In the past, it was believed that the citric acid cycle in cyanobacteria did not function as a complete cycle but rather as two independent branches due to the absence of α -ketoglutarate dehydrogenase, as illustrated schematically in Fig. 4.7. However, more recent studies have revealed the existence of an alternative pathway involving α -ketoglutarate decarboxylase and succinyl semialdehyde dehydrogenase, which bypasses the missing α -ketoglutarate dehydrogenase step. This is described by fluxes v_{31} and v_{32} .

4.3.1.1.6 GS-GOGAT Pathway

The GS-GOGAT pathway is a crucial biochemical pathway for nitrogen assimilation in many microorganisms. Assimilated through photosynthetic processes, cyanobacteria preferentially use nitrate as a nitrogen source, which is then reduced in the form of NH_4^+ . However, ammonium can be toxic to cells at high concentrations. Therefore, the GS-GOGAT pathway allows the incorporation of ammonium into organic molecules, specifically amino acids and other nitrogen-containing biomolecules. The first step of the pathway is catalyzed by glutamine synthetase (GS). Glutamate reacts with NH_4^+ to yield glutamine, which carries an additional nitrogen atom, making it a key molecule for transporting nitrogen within cells. The second step of the pathway is catalyzed by glutamate synthase (GOGAT). This enzyme transfers the amide nitrogen from glutamine to α -ketoglutarate, an intermediate of the citric acid cycle, to regenerate two glutamate molecules. In addition, the GS-GOGAT pathway is not only essential for nitrogen assimilation but also plays a key role in regulating nitrogen metabolism in response to

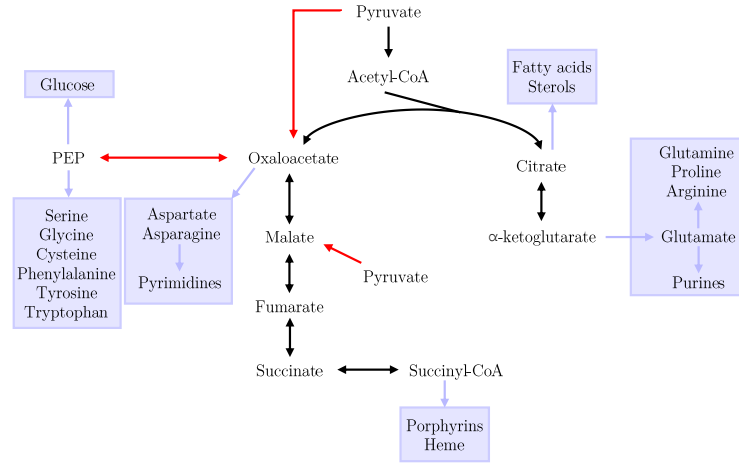


Figure 4.7: Incomplete citric acid cycle and the main intermediates which are precursors in many biosynthetic pathways. Anaplerotic reactions are shown in red.

changing environmental conditions. The GS-GOGAT pathway and the reactions related to nitrogen assimilation are represented by the fluxes v_{38} to v_{43} in Table G.1.

4.3.1.1.7 Urea Cycle-Like Pathway

Another way to regulate ammonium levels is through the urea cycle-like pathway (also referred to as the Ornithine-Ammonia cycle), also called the ornithine cycle. While photosynthetic organisms primarily manage nitrogen through the GS-GOGAT pathway, certain cyanobacteria, including *Arthrospira* sp. PCC 8005, possess the enzymatic machinery to synthesize urea under specific conditions. In these organisms, urea may serve as a temporary storage form of nitrogen, which can later be broken down by urease to release ammonium for assimilation into amino acids.

Genome sequence analysis of a similar strain, i.e., *A. platensis* NIES-39 [203], confirms the presence of all the enzymatic steps required for this pathway, including carbamoyl phosphate synthase (for v_{45}), ornithine carbamoyltransferase (for v_{46}), argininosuccinate synthase (for v_{47}) and argininosuccinate lyase (for v_{48}). Proteomic studies further support its activity, showing that these enzymes are upregulated under nitrogen-depleted conditions [198]. This suggests that, rather than acting as an excretory pathway, as in ureotelic organisms, the urea cycle-like pathway in cyanobacteria probably plays a role in nitrogen recycling and metabolic flexibility.

Moreover, the inclusion of reactions v_{44} and v_{49} within the metabolic network is justified by experimental observations. In fact, reaction v_{49} , representing the formation of urea via arginase, is supported by continuous culture experiments using nitrate as the sole nitrogen source (30mM-N NO_3^-), where residual urea was detected in the medium. This suggests that *Arthrospira* sp. PCC 8005 is capable of producing urea (in small quantity) under certain growth conditions. Additionally, in experiments in which the feeding medium includes a mixture of nitrate and urea (15mM-N NO_3^- + 15mM-N urea), the ability to assimilate urea becomes essential, thereby justifying the inclusion of v_{44} in the network, which reflects the activity of urease. These aspects will be discussed in more detail in Sec. 4.3.3.

In this pathway, ammonium combines with carbon dioxide and ATP to form carbamoyl phosphate, which enters the urea cycle-like pathway. First, carbamoyl phosphate transfers its carbamoyl group to ornithine, resulting in the formation of citrulline and the release of inorganic phosphate. Then, citrulline is combined with aspartate, which supplies an additional nitrogen, to form argininosuccinate. The latter is cleaved to form arginine and fumarate. Finally, arginine is hydrolyzed to produce urea and regenerate ornithine to repeat the cycle, while fumarate joins the pool of citric acid cycle intermediates. Although some low-level excretion of urea has been detected experimentally, its primary function in cyanobacteria is likely nitrogen retention and redistribution rather than elimination.

The urea cycle-like pathway is biochemically interconnected with the tricarboxylic acid cycle through fumarate, forming what is known as the "Krebs bicycle" [1]. However, each cycle can function independently, and their interaction relies on the exchange of crucial intermediates within the cytosol. The urea cycle-like pathway is depicted by fluxes v_{44} to v_{49} in Table G.1.

4.3.1.1.8 Energy Metabolism

As mentioned in Sec. 4.2.4, *Arthrospira* sp. PCC 8005 can synthesize ATP through different mechanisms depending on oxygen availability and the presence of light. First, ATP is synthesized through substrate-level phosphorylation, allowing a fast production of ATP in the cytoplasm. Then, in the presence of sufficient light, ATP is generated through the process of photophosphorylation occurring in the thylakoid membranes of prokaryotic thylakoids. Finally, in the absence of sufficient light for photosynthesis or during the night, cyanobacteria can also synthesize ATP through a process called oxidative phosphorylation. This mechanism occurs in the plasma membrane for prokaryotic cells. The chemical reactions describing ATP synthesis through photophosphorylation and oxidative phosphorylation are given by the fluxes v_{120} to v_{123} in Table G.1.

4.3.1.1.9 Other Metabolic Routes

Most of the previous pathways (other than the Calvin-Benson cycle) are catabolic, meaning that the major metabolic fuels are degraded to enter the citric acid cycle and yield their electrons to the respiratory chain, enabling, inter alia, the synthesis of ATP. Nevertheless, anabolic pathways are also fundamental in describing the whole metabolism of an organism. Those pathways use chemical energy in the form of ATP, NADH or NADPH to synthesize cellular components from simple precursor molecules. The most important anabolic pathways are the biosynthesis of carbohydrates, the biosynthesis of lipids, and the biosynthesis of amino acids and nucleotides. Finally, information pathways (proteins, DNA, and RNA metabolism) must be included in the metabolic network to form biomass. Most anabolic pathways can be found in [1], especially metabolic reactions characterizing amino acid metabolism (fluxes v_{70} to v_{101} in Table G.1) and the nucleotide salvage pathway (fluxes v_{102} to v_{118}).

As highlighted in Sec. 4.2, a benefit of the methodology presented in this chapter is its minimal reliance on genomic reconstructions (not always accessible or reliable for lesser-studied or newly-isolated strains). As a consequence, in the first instance, and for the sole purpose of demonstrating feasibility (i.e., handling situations where genome sequences are not available), simplified chemical reactions are considered for some anabolic pathways. Then, in the second instance, because genome sequences of similar strains are available, the latter will be exploited to better estimate the composition of

the main components of the cell material and to identify specific metabolic routes.

(A) When genome sequences are not available, it can be assumed that the synthesis of carbohydrates begins with glucose-6-phosphate (G6P), an intermediate in glycolysis, which is converted into glucose-1-phosphate (G1P) by the enzyme phosphoglucomutase. Glucose-1-phosphate is then activated by nucleotides such as UTP to form UDP-glucose, a building block for synthesizing polysaccharides such as glycogen, which serve as carbohydrate reserves in cells. Regarding the synthesis of lipids, it can be assumed that it starts with glycerol-3-phosphate (G3P), which is derived from glycolysis through the reduction of dihydroxyacetone phosphate. Then, glycerol-3-phosphate is acylated with two molecules of acyl-CoA, forming phosphatidic acid via acyltransferases. The acyl-CoA molecules are derived from fatty acids, which are themselves synthesized from acetyl-CoA, a product of central carbon metabolism. Phosphatidic acid is a precursor of various lipids. The latter can form triglycerides or alternatively be modified to produce phospholipids, key components of cell membranes. Finally, the elemental composition of proteins and biomass can be estimated using data from similar organisms. Experimental methods, such as protein hydrolysis for amino acid profiles or solvent extraction for lipids, can also be useful to determine key components without genomic data. Therefore, if desired, all major metabolic pathways could be expressed without relying on genomic data.

(B) However, in this particular case, because genome sequences are available for *Arthrospira* sp. (different strains), the composition of the main components of cell material (proteins, carbohydrates, lipids, and nucleic acids) and other macromolecules can be provided by genomic studies [202–204], as well as the average formula of biomass. Doing so, by using the genome sequence of *S. platensis* C1 PCC 9438 in [202] and applying lumping techniques, the reactions describing the synthesis of carbohydrates are given by fluxes v_{133} to v_{146} , the reactions for the synthesis of lipids are represented by fluxes v_{147} to v_{166} and the elemental composition of the macromolecules and the biomass is given by fluxes v_{173} to v_{178} in Table G.1 (note that the stoichiometry will be adjusted later using experimental data relative to the specific strain). More specific pathways relative to the strain under study can also be derived from the latter genome sequence and will be detailed in the sequel. For example, the synthesis of chlorophyll, especially important for photosynthetic organisms, is depicted by fluxes v_{167} to v_{171} .

4.3.1.2 Iterative Methodology

As mentioned previously, the method is iterative and model refinement is possible by comparing model predictions with experimental observations, which may include available experimental data or well-established findings from the literature or surveys. This section outlines a couple of iterations before obtaining a final draft for the metabolic network, while the next sections will present the results of metabolic analyses as validation. Table 4.1 summarizes the iterations.

4.3.1.2.1 First Iteration

For the first iteration of the procedure, the main metabolic pathways that are considered are the carbon uptake, the Calvin-Benson cycle, the glycolysis, the pentose phosphate pathway, the citric acid cycle, the urea cycle-like pathway and the GS-GOGAT pathway, the pathway of the amino acid metabolism, the nucleotide salvage pathway, the pathways of biosynthesis of carbohydrates and lipids, information pathways and the routes for ATP synthesis by oxidative phosphorylation and photophosphorylation. Essentially, the first iteration takes into account the metabolic reactions introduced in Sec. 4.3.1.1.

Table 4.1: Information related to metabolic networks from the iterative procedure. The number of metabolites is comprised of internal metabolites and intracellular energetic cofactors, and the number of reactions includes the intracellular reactions and the transport/exchange reactions.

Iteration	# METABOLITES	# REACTIONS	# BLOCKED REACTIONS
1	157	167	39
2	170	191	16
3	161	182	7
4	161	179	7
5	154	172	0
6	155	182	0

The reversibility of the reactions is accounted for by considering the standard Gibbs free energy change, and elementary-mass conservation is verified. This leads to a metabolic network that contains 157 metabolites and 167 reactions.

In order to check network consistency, as well as structural and topological network properties, constraint-based methods can be applied. For this purpose, a particular uptake scenario is considered, for which light, a carbon source (HCO_3^-), and a nitrogen source (NO_3^-) are available. This uptake scenario makes sense for photosynthetic organisms, since light is used as an energy source, bicarbonate is a common source of inorganic carbon, and nitrate is the preferential nitrogen source for cyanobacteria. To avoid infeasible solutions, a set of transport/exchange reactions is added to the network (fluxes v_{179} to v_{198}), as well as a metabolic reaction including ATP maintenance requirements (flux v_{124}) to simulate non-growth-associated energy demands. In addition, the objective function of the flux balance analysis should represent a biologically relevant process. Accordingly, a flux balance analysis with biomass maximization is suggested, and a flux coupling analysis is performed to prevent large computation times. After careful analysis, the network provides inconsistent results. In fact, the biomass production rate obtained from the FBA is null, which is not expected considering the uptake scenario, and the flux coupling analysis provides several blocked reactions, meaning that there are gaps in the network structure.

However, these inconclusive results can be explained. First, in the presence of light, it is common knowledge that cyanobacteria reduce atmospheric CO_2 to triose sugars before converting them into glucose using the Calvin-Benson cycle and gluconeogenesis. The latter pathway is missing from the network and needs to be considered (addition of fluxes v_{14} to v_{17} — found in [1] — and adjustment of the reversibility of several reactions of the Embden-Meyerhof-Parnas pathway). Then, from the first iteration, flux coupling analysis revealed 39 blocked reactions, which are reactions unable to carry a flux under the defined uptake scenario. These reactions often indicate gaps in the metabolic network, where some metabolites are consumed without being explicitly produced by upstream reactions. By carefully inspecting the blocked reactions, it is possible to identify the missing pathways or transport steps necessary for the synthesis or utilization of intermediate metabolites. Indeed, addressing these gaps improves the accuracy and completeness of the network reconstruction, ensuring that all essential metabolic connections are captured. In this case, it appears that several reactions enabling the metabolism of amino acids and reactions belonging to the nucleotide salvage pathways are inactive,

which explains why biomass production is impossible when performing FBA. The main reasons are the following:

- The Brønsted-Lowry acid-base equilibrium v_{40}

Initially, the reaction v_{40} was set as irreversible, assuming that ammonia NH_3 primarily acts as a weak base and reacts with a proton H^+ to form ammonium NH_4^+ . However, as mentioned in Sec. 4.2.2, acid-base equilibria are inherently reversible, with the direction of the equilibrium shifting based on factors such as the concentration of reactants and products, the pH of the medium, the temperature or the presence of other ions or compounds. Although the equilibrium generally favors ammonium formation under physiological conditions, FCA revealed blocked reactions under the specific uptake scenario. Therefore, to ensure a more generalized and functional metabolic network, the reaction was modeled as reversible.

- The missing pathways of tetrahydrofolate metabolism and S fixation

The formation of several amino acids requires H_2S and tetrahydrofolate (THF) or its byproducts (CHO-THF; CH_2 -THF; CH_3 -THF; DHF) as reactants. In addition, some of them also intervene in the nucleotide salvage pathway. Because the synthesis of such intermediate metabolites is not considered in the network, it leads to several inactive reactions that are essential in the formation of biomass, which explains the result of the FBA. Therefore, they need to be synthesized to avoid incomplete pathways or gaps in the network reconstruction, leading to infeasible solutions. In that regard, the pathways of tetrahydrofolate metabolism and S fixation are added to the network using genome sequences and lumping techniques. The corresponding metabolic reactions are represented by fluxes v_{50} to v_{69} in Table G.1.

4.3.1.2.2 Second Iteration

After making the reaction v_{40} reversible and adding pathways of neoglucogenesis, tetrahydrofolate metabolism and S fixation, the metabolic network consists of 170 metabolites and 191 reactions. Elementary-mass conservation is still verified. For the same uptake scenario as the one considered in Sec. 4.3.1.2.1, FBA provides plausible results (e.g., the specific growth rate is non-zero), which highlights the benefits of the iterative procedure. More details on the outcomes of the metabolic analysis are given in Sec. 4.3.2 and Sec. 4.3.3. However, FCA still identifies unusable/blocked reactions.

In this case, the unusable reactions are due to an excessively detailed representation of the metabolism of tetrahydrofolate. In this way, the reactions detailing the formation of dihydrofolate (DHF) from glutamate, para-aminobenzoate, and GTP (fluxes v_{53} to v_{61} , which were obtained from the genome sequence in [202] and using lumping techniques) are not strictly necessary and can be removed from the network because these reactions involve intermediate metabolites that are not currently used in the synthesis of other essential components. However, it is worth noting that it depends on the scope and focus of the model. If more specific pathways are added later on, these metabolites might become relevant and necessary.

Henceforth, it is worth noticing that flux coupling analysis can be used not only as a method to identify blocked reactions (i.e., inconsistencies in the network structure), but also as a method to simplify over-detailed networks.

4.3.1.2.3 More Iterations

The third iteration involves removing unnecessary reactions from the pathway of tetrahydrofolate metabolism. Besides, more iterations are needed to improve the structure of the metabolic network. In this regard, the fourth iteration involves removing redundant reactions due to lumping techniques or the conflicting nomenclature of chemical compounds from the literature. The fifth iteration allows for discarding reactions involving specific metabolites that do not take place in the biosynthesis of carbohydrates, lipids, proteins, or other macromolecules. The sixth iteration allows the addition of cofactor metabolism reactions (fluxes v_{119} and v_{125} to v_{132} – found in [1]) and the reaction of formation of cyanophycin, a key biopolymer that stores excess nitrogen during periods of nitrogen abundance (flux v_{172}). More iterations can be made, but they are not detailed in this work for the sake of clarity.

4.3.1.2.4 Final Metabolic Network

After several iterations and a compromise between the level of detail and the size of the network, a final draft of the metabolic network can be obtained. It comprises 198 reactions (including intracellular reactions and transport reactions) and 171 metabolites (including internal metabolites and intracellular energetic cofactors). For the sake of completeness and to enable the development of enriched networks in further research, the final network also considers the synthesis of specific precursors involved in the synthesis of important biomolecules and vitamins. The metabolic reactions of the network are given in Table G.1. This network models the metabolism of the cyanobacterium *Arthrospira* sp. PCC 8005 under a photoautotrophic regime. Recall that this strain is particularly interesting due to its use in bioregenerative life-support systems whose main objectives are to supply a crew with food, water, and oxygen while eliminating waste.

4.3.2 In-Silico Validation

Validating the metabolic network using data is a crucial step to ensure its reliability and predictive power. In this section, in-silico data are used to assess the network consistency and functionality.

As a first validation or when experimental data are not available, a common strategy for validating a metabolic network is to leverage in-silico data, i.e., data generated through computational models and simulations rather than obtained from experiments, and explore the behavior of the network under defined conditions. The idea is then to analyze whether the model predictions align with known biological behavior under imposed scenarios, such as nutrient uptake or environmental conditions, in order to assess the network validity. This approach enables identifying inconsistencies or unrealistic outputs, ensuring that the network reflects biological reality and not just mathematical consistency.

As mentioned previously, this first application focuses on the mathematical modeling of *Arthrospira* sp. PCC 8005 under constant light regimes (i.e., under continuous illumination), where the pseudo-steady-state assumption is valid, enabling the use of classical metabolic analysis tools. In such conditions, the metabolic fluxes reach a dynamic equilibrium, where the production and consumption of internal metabolites balance out over time. Unlike fluctuating light conditions, where rapid transitions between photosynthetic activity and carbon storage mobilization induce significant accumulation or depletion of metabolites, continuous illumination provides a stable supply of energy

(ATP, NADPH) and fixed carbon (via CO_2 assimilation). It ensures that carbon fluxes remain balanced and that storage compounds (such as glycogen) do not accumulate significantly, as cells continuously metabolize their resources rather than alternating between storage and consumption phases. As a result, the metabolic network maintains approximately constant flux distributions, which makes it possible to apply classical steady-state analysis methods such as FBA and FVA. Although minor fluctuations in enzyme activities and metabolite pools may still occur due to regulatory mechanisms, they do not significantly disrupt the overall metabolic balance. Therefore, under well-controlled conditions such as chemostats or turbidostats with continuous light exposure (see Sec. 4.3.3.1), the assumption that internal metabolite concentrations remain approximately constant holds, justifying the use of the classical MFA tools.

The metabolism adaptation under dynamic conditions, such as permanent fluctuating light, will be considered and discussed in the second case study, in Sec. 4.4. Nevertheless, even though the present study assumes a steady-state framework (and a static regime for in-silico simulations), it offers insights into the photoautotrophic metabolism of cyanobacteria under the day-night cycle. During the day, when light levels are sufficient to meet carbon and energy demands, the autotrophic regime occurs. At night, metabolism changes to a heterotrophic regime, breaking down carbon storage molecules into precursor metabolites and energy for growth and maintenance. Otherwise, when light is not intense enough to meet the carbon and energy growth demands, the regime is said to be mixotrophic and is characterized by the consumption of carbon storage molecules. In this section, a static in-silico analysis is conducted and diverse uptake scenarios are explored to assess the robustness of the constructed network and quantify intracellular metabolic mechanisms.

Because the proposed network is of modest size, a graphical representation of the latter is adequate to analyze the metabolic fluxes more easily. The use of *CellDesigner* is suggested [207]. It is a structured diagram editor for drawing gene-regulatory and biochemical networks using the Systems Biology Markup Language (SBML). Although implementing a network is time-consuming, CellDesigner is a powerful tool for modeling biochemical networks and offers a good balance between usability and functionality. The flux map is represented in Fig. G.1 in the Appendix G.

4.3.2.1 Day Phase

The uptake scenarios considered in this section correspond to an autotrophic regime in which cyanobacteria use sunlight, CO_2 and water to produce organic compounds, among others. To ensure the metabolic consistency of our network, we first perform simulations under a strict carbon-only uptake scenario, where only light and a source of carbon are provided and no nitrogen source is available. In this case, as expected, biomass production is not feasible because nitrogen is essential for the synthesis of amino acids and nucleotides. However, this scenario allows us to validate our metabolic model by ensuring that under nitrogen limitation, only the core metabolic pathways, such as glycolysis, the tricarboxylic acid cycle and the pentose phosphate pathway, are activated.

In the study conducted by [208], a similar uptake scenario is considered in a metabolic model of unicellular microalgae. However, while their simulations result in biomass formation under light and CO_2 uptake alone, our results suggest that nitrogen must be explicitly available for biomass production. This difference may arise from implicit nitrogen uptake or relaxed nitrogen constraints in their model, whereas our formulation strictly enforces nitrogen balance.

Then, to further explore nitrogen metabolism, in-silico simulations are performed considering a nitrogen source in addition to light and a carbon source, enabling the activation of amino acid biosynthesis and macromolecule formation, which are required for biomass growth.

Furthermore, to analyze how the network behaves during the day phase, it is necessary to impose a light uptake, through a specific photon uptake rate through the corresponding transport reaction v_{181} . This is achieved by computing a specific amount of photon moles, ensuring that light availability is properly integrated into the metabolic simulations. However, light is commonly expressed in watts per square meter (W/m^2). For that reason, finding a relationship between the energy and the number of photons is essential. Basically, it requires calculating the energy of a single photon using the following equation:

$$E_{\gamma^*} = \frac{h c}{\lambda} \quad (4.25)$$

Where h is the Planck constant (expressed in joule seconds: J.s), c is the speed of light (in meters per second), and λ is the wavelength of light in meters. The energy obtained using Eq. 4.25 is then expressed in joules. Next, to obtain the photon flux density, i.e., the number of photons per second per square meter, it remains to divide the power density (in W/m^2) by the energy of a single photon. Finally, the photon flux density expressed as the number of moles of photons per second per square meter is obtained using the Avogadro number (see [1] for more details).

As an example, if a light with an average wavelength of 600 nm and a power density of 150 W/m^2 is considered, the photon flux density is equal to $7.5185 \cdot 10^{-4}$ moles of photons per square meter per second (also expressed in $\text{einstein} \cdot \text{m}^{-2} \cdot \text{s}^{-1}$).

4.3.2.1.1 Carbon-Only Uptake Scenario

As mentioned previously, the use of in-silico data requires the definition of an uptake scenario to mimic what might happen biologically. It can be done by specifying the availability of nutrients, environmental conditions, or flux bounds for exchange reactions. Metabolic analyses are then performed and the results of the simulations are compared with known biological insights. In this section, light energy and a carbon source are imposed to mimic an autotrophic regime. Flux variability analysis and flux balance analysis are performed to examine intracellular mechanisms and predict the metabolic distribution of the fluxes under constant environmental conditions.

In order to perform the flux balance analysis, it is required to select a biological objective. Although biomass formation is infeasible under a carbon-only uptake scenario due to the absence of a nitrogen source, performing FBA with biomass optimization remains informative. This approach ensures that the model does not allocate flux toward biomass formation, confirming that elemental balance is properly enforced. In addition, it allows us to analyze how the core metabolic pathways operate under nitrogen starvation. By maintaining a consistent optimization objective across the different uptake scenarios, we can also assess how metabolic fluxes shift when nitrogen becomes available, providing insights into the network's response to nutrient limitations.

In that respect, it is necessary to set the fluxes of the corresponding transport reactions. In this case, simulations are performed for the net assimilation of 100 mol of bicarbonate ion HCO_3^- through reaction v_{179} . Because bicarbonate is the only carbon source considered and there is no nitrogen source, reactions related to the transport of

glucose, nitrate, nitrite, and ammonia are set to zero. Since cells can accumulate carbohydrates and lipids during the day and use them during the night, the corresponding transport reactions are possibly reversible. However, since simulations are representative of the day phase in the present scenario, the lower bounds of v_{195} and v_{196} can be set to zero. With regard to light energy, the associated transport reaction v_{181} is constrained such that sufficient photon flux is available to support growth under the given simulation conditions. The corresponding flux map is depicted in Fig. G.2.

As illustrated in the flux map, autotrophy is characterized by high fluxes in the photosynthetic pathways and by the activation of the Calvin-Benson cycle. Upper glycolysis operates in the glyconeogenic direction to produce carbohydrates and sugar precursor metabolites such as phosphoenolpyruvate, glucose-6-phosphate and ribose-5-phosphate, essential for growth. In addition, the pentose phosphate pathway is in the reductive mode, and the citric acid cycle enables the formation of metabolite precursors for biomass growth. These results are in concordance with [208]. As detailed in Sec. 4.2.4, ATP is generated through glycolysis and the citric acid cycle, on the one hand. On the other hand, ATP is driven by the proton motive force (depicted by the electrochemical gradient H_p^+/H_n^+) and is synthesized via photophosphorylation and oxidative phosphorylation. In addition, it can be shown that the influence of light on the qualitative flux distribution is minimal as long as light remains the limiting factor, as discussed in [208]. Otherwise, the excessive light energy is dissipated by means of photophosphorylation, and the rest of the metabolism remains relatively unchanged in terms of flux distribution. In this way, it enhances the robustness of photosynthetic pathways against light variation, maintaining a consistent energy supply to the metabolism. As a reminder, it is important to notice that no biomass is formed (meaning that the specific growth rate is equal to zero) because no nitrogen source is provided to the organism.

4.3.2.1.2 Carbon-Nitrogen Uptake Scenario

In this section, the day phase is still addressed, but a nitrogen source is provided to the organism, in addition to light and a carbon source. In this case, the net assimilation of 10 mol of nitrate NO_3^- is imposed using the exchange reaction v_{186} , known as the preferential nitrogen source for cyanobacteria. The fluxes of the other transport reactions remain the same as in Sec. 4.3.2.1.1. The same conclusions as before can be drawn regarding the Calvin-Benson cycle, the Embden-Meyerhof-Parnas pathway, PPP, and the routes of TCA. In addition, it is interesting to note that glyceraldehyde-3-phosphate (G3P) is produced mainly via the Calvin-Benson cycle. The reductive PPP contributes a little, but glycolysis (via the breakdown of fructose-1,6-bisphosphate) is not a major pathway for G3P production, particularly when the organism is in a photosynthetic state. Furthermore, for this uptake scenario, the nitrogen source allows the biosynthesis of amino acids and the activation of the nucleotide salvage pathway. Besides, the GS-GOGAT pathway and the urea cycle-like pathway are involved in nitrogen assimilation. Finally, biomass is formed.

4.3.2.2 Dark Phase

The uptake scenarios analyzed in the following sections correspond to a heterotrophic regime, meaning that carbon storage molecules are degraded into precursor metabolites and energy, essentially for growth and maintenance. To ensure the validity of the metabolic network and facilitate comparisons with existing studies, the heterotrophic regime (equivalent to a chemotrophic regime in this case) is simulated by imposing a net assimilation of 100 mol of glucose, one of the organic carbon sources represented in the network. As in previous analyses (see Sec. 4.3.2.1), simulations are first conducted

with glucose as the sole energy to examine the behavior of the network under carbon-only conditions and verify that the core metabolic pathways function as expected under heterotrophic conditions. Subsequently, a source of nitrogen is introduced to enable biosynthetic processes and biomass formation.

In heterotrophic conditions during the dark phase, optimizing biomass formation remains a relevant objective, as many cyanobacteria are capable of growth when both carbon and nitrogen sources are available. However, depending on the physiological state of the organism, alternative metabolic objectives may be more biologically meaningful. For instance, if only a carbon source is available, cells are unlikely to prioritize growth and may focus on maximizing ATP production to maintain vital cellular functions. Another plausible objective could be the accumulation of storage compounds, such as glycogen, which would later serve as an energy reserve for the next light phase.

4.3.2.2.1 Carbon-Only Uptake Scenario

Flux variability analysis and flux balance analysis are performed by setting the glucose transport reaction through the exchange reaction v_{192} . Only a carbon source is assumed, meaning that the transport reactions related to nitrogen sources are set to zero, and no light energy is considered because of the dark phase. The corresponding flux map is not represented for the sake of clarity, but the operation of the main pathways is resumed thereafter. The absence of light inhibits the activation of the Calvin-Benson cycle, which normally produces glucose from CO_2 during the day phase. When glucose is used as the carbon source, upper glycolysis is in the downward direction and the pentose phosphate pathway is in the oxidative mode. In addition, primary carbon flux is generated through the tricarboxylic acid cycle, which generates precursor metabolites for growth and energy via oxidative phosphorylation. During the night, ATP is not synthesized by photophosphorylation. Furthermore, when proceeding in the downward direction, glycolysis also contributes to energy demands, and it can be established that approximately 40% of the carbon is lost through respiration [151]. These latter conclusions are supported by existing studies that cover the modeling of photosynthetic bacteria.

4.3.2.2.2 Carbon-Nitrogen Uptake Scenario

This uptake scenario deals with the dark phase, considering glucose as a carbon source and nitrate as a nitrogen source. The same conclusions can be drawn as in Sec. 4.3.2.1.2, except the Calvin-Benson cycle is not activated due to the absence of light. However, an interesting observation can be made, which is in agreement with the study conducted in [208]. As mentioned in the previous section, PPP is activated in the oxidative mode in heterotrophic regimes, mainly to fulfill the NADPH requirements, enabling the synthesis of macromolecules such as amino acids, nucleotides, and lipids. However, the flux distribution shows that the presence of NADPH/NAD⁺ transhydrogenase alpha subunit (EC number: 1.6.1.2) that converts NADH to NADPH through v_{126} is usually preferred to produce NADPH, implying a nearly null flux into the pentose phosphate pathway through v_{18} . This result contradicts experimental findings proving that PPP remains the primary route for NADPH synthesis. This allows highlighting the limitation of in-silico data and reminds of the importance of experimental data. This discrepancy can arise from missing regulatory mechanisms or oversimplified assumptions in the model. Furthermore, the more data, the better the flux predictions provided by the metabolic analyses. Recall that the identified metabolic network is underdetermined. Therefore, a calculability/observability analysis could be conducted to determine the number of fluxes that can be uniquely calculated [209]. In the present study, the inaccuracy pointed out can be corrected by limiting the flux of NADPH/NAD⁺ transhydrogenase complex in

the FBA approach, leading to results consistent with experimental findings.

Essentially, autotrophic and heterotrophic fluxes differ in the arrangement of the core carbon network, including the Calvin-Benson cycle, the Embden-Meyerhof-Parnas pathway, the pentose phosphate pathway, and the citric acid cycle. In contrast, the remaining pathways involved in the biosynthesis of amino acids and macromolecules and the information pathways show relatively consistent flux patterns regardless of growth conditions. In that respect, it suggests that anabolic processes operate independently of growth conditions, which is explained by the classical bow-tie structure of microorganisms.

4.3.3 Experimental Validation

Using in-silico data is a powerful first step in evaluating the structure of a metabolic network. By simulating specific uptake scenarios, we can assess whether the network behaves in a biologically plausible manner, ensuring that it reflects expected metabolic features and not just mathematical consistency. This initial validation helps identify potential gaps or errors in the identified network, improving its structure before further analysis. However, in-silico data have limitations. Therefore, evaluating the network against experimental data is crucial to confirm that it accurately represents actual metabolic behavior. Some methods for experimental validation are shown in Fig. 4.8.

4.3.3.1 Cell Line and Media

Data were provided by the Department of Proteomics and Microbiology of the University of Mons. More information on the cell line, the medium, the bioreactor operation mode, and the analysis methods can be found in [199].

The photo-bioreactor (PBR) started in batch mode and continuous feeding started after 7 days (only data in continuous mode are provided). For the continuous PBR experiments, cyanobacteria were incubated in a radially illuminated 2 L cylindrical double jacket reactor under a constant light power density of 125 W/m^2 with the agitation settled at 150 rpm. Note that this constant light regime is achieved using artificial lighting systems that provide uniform light conditions, meaning that the quasi-steady-state assumption (QSSA) holds. Also, light is the limiting factor in the experiments. The pH was automatically maintained at 8.5 with HCl (0.5 M) whereas the dilution rate was kept around 0.2 per day with fresh Zarrouk medium (30mM-N NaNO_3 or a mixture of nitrogen sources). Experimental data in continuous mode are provided for different feeding strategies (i.e., 30mM-N nitrate; 15mM-N nitrate + 15mM-N nitrite; 15mM-N nitrate + 15mM-N urea and 30mM-N nitrate in high salinity medium).

Experimental data include biomass productivity and oxygen productivity (expressed in $\text{g.L}^{-1}.\text{d}^{-1}$); residual nitrogen concentration (i.e., NO_2^- , NO_3^- , NH_4^+ and urea in mM); pigment concentration (i.e., chlorophyll_a and phycobiliproteins in g.L^{-1}); TC, TOC, TIC and TN in biomass and supernatant (in mg.L^{-1}) and protein, lipid and carbohydrate content in biomass. O_2 is monitored in the exhaust gas (purging with air to prevent an increase in the oxygen saturation of the medium). The temperature is around 36°C and the carbon source is a mixture of 10.5 g.L^{-1} of NaHCO_3 and 7.6 g.L^{-1} of Na_2CO_3 .

4.3.3.2 Direct Validation

Experimental validation of a metabolic network is quite challenging because all available data (uptake and secretion rates, metabolite concentrations, and flux measurements, if

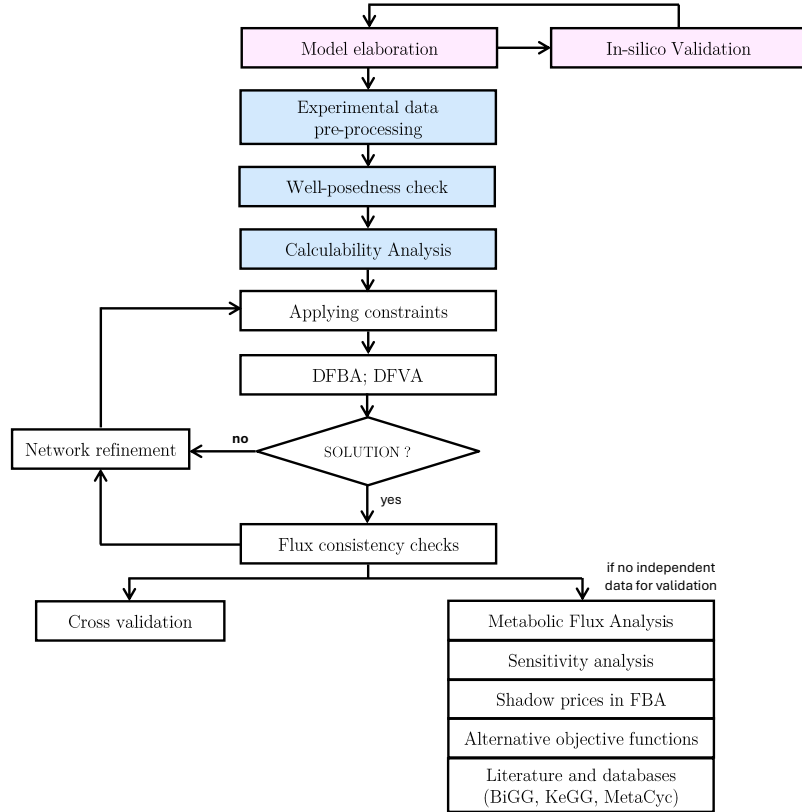


Figure 4.8: Overview of the methodological framework used to validate the metabolic network. Preliminary processes are shown in blue, including data pre-processing, a well-posedness check and a calculability analysis. These steps ensure the model is computationally tractable before proceeding to experimental validation.

applicable) are used to constrain the system. As mentioned previously, this study applies the quasi-steady-state assumption under constant light regimes, using the full spectrum of classical metabolic analysis tools. In this case, dynamic metabolic frameworks, such as dynamic flux balance analysis (dFBA) and dynamic flux variability analysis (dFVA), are used because experimental uptake and secretion rates are computed from concentration profiles over time. These methods extend traditional FBA and FVA by integrating time-dependent metabolic constraints. In this section, several tests are performed to validate the network based on the set of methods suggested in Fig. 4.8.

As discussed in Sec. 4.3.1.2.4, the final iteration of the iterative procedure resulted in a generalized and enriched metabolic network that offers a comprehensive representation of the metabolic capabilities of the system. It includes reactions that generate intermediate metabolites, some of which may be involved in the biosynthesis of essential biomolecules such as cofactors or vitamins, thus facilitating potential future model extensions. However, because of the limited availability of experimental measurements and the underdetermined nature of the metabolic network, it was necessary to refine the scope of the analysis. For this purpose, a network reduction is applied, focusing on essential pathways and resulting in a focused metabolic network. It results in a metabolic network comprising 175 biochemical reactions and 152 metabolites, as shown in Table

H.1 in the Appendix H. More information is given in Sec. H.1.

This refined network is the result of several foundational steps that were required to ensure coherence, consistency, and interpretability of the model. These include data pre-processing and mathematical verification of the well-posedness of the network. In addition, a calculability analysis was carried out to identify fluxes that can be uniquely determined. This enhances the interpretability of the resulting flux distributions and helps avoid misleading conclusions from non-unique solutions. Full details of these preliminary steps are provided in the Appendix H.

4.3.3.2.1 Experimental Constraints

Experimental constraints derived from the cultivation data are integrated into the metabolic flux analysis framework to ensure that the simulations reflect biologically and physiologically realistic conditions. The following fluxes are either imposed directly or derived from measured concentrations and rates:

- The photon uptake rate, based on light intensity and reactor geometry;
- The bicarbonate (HCO_3^-) uptake rate;
- The nitrogen source (NO_2^- , NO_3^- , NH_4^+ and urea) uptake or secretion rates, depending on the feeding strategy;
- The chlorophyll secretion rate;
- The oxygen production rate;
- The specific growth rate μ .

In addition to flux constraints, the macromolecular composition of biomass (including carbohydrate, lipid, and protein content) was monitored over time. These measurements were used to adjust the stoichiometric coefficients in the biomass reaction (flux v_{155} in Table H.1) to reflect the actual composition of cells under different conditions.

Although several fluxes could be accurately estimated from experimental data, certain parameters are subject to uncertainty because of assumptions or measurement limitations. In particular, the photon uptake rate depends on: (i) the average wavelength of the incident light (assumed to be 600 nm), (ii) the illuminated surface area of the photobioreactor (estimated from reactor geometry), and (iii) an assumed light absorption efficiency of the culture (see Sec. H.2). Indeed, since not all incident light is absorbed by the cells, a light efficiency parameter is introduced to estimate the fraction of usable photons. To account for uncertainty in these assumptions, the photon uptake rate is not fixed but instead bounded between a lower and upper limit, defined on the basis of a sensitivity analysis.

Similarly, the oxygen production rate, derived from experimental productivity data, is treated with caution. Due to potential inaccuracies in oxygen sensor measurements, the corresponding flux is also defined using upper and lower bounds rather than a fixed value. These bounds reflect realistic uncertainty in the measured values and prevent the model from being overconstrained by potentially imprecise data.

4.3.3.2.2 Fundamental Tests

A series of fundamental tests is conducted to verify the coherence of the network. For instance, flux variability analysis (FVA and/or dFVA) is achieved by imposing only experimental uptake rates and ensuring that experimental secretion rates belong to the predicted intervals. Moreover, flux balance analysis (FBA and/or dFBA) with biomass optimization is performed by imposing both experimental uptake and secretion rates and ensuring that the predicted optimal value for the growth rate is higher than the measured growth rate. Doing these analyses is a prudent approach to enhance confidence in the analyses and to verify that the model accurately reflects the organism's growth potential and efficiency.

Fig. 4.9 shows the results of the dynamic flux balance analysis (dFBA) with biomass optimization under different nitrogen feeding strategies. The experimental growth rates ($\mu_{exp} \approx D = 0.083 \text{ h}^{-1}$) are compared with the specific growth rates predicted by the model (μ_{opt}), providing an overview of the predictive capabilities of the model.

Across all conditions, the model-predicted growth rates are higher than the experimental values, which is expected given that in vivo systems do not necessarily operate at maximal metabolic efficiency. Under the nitrate-only (30mM-N) and nitrate-nitrite (15mM-N each) conditions, the predicted growth rates closely match the experimental ones, suggesting that cells are operating near their physiological limits. This result is consistent with the relatively high energetic cost of nitrate and nitrite assimilation, which requires one or two reduction steps, respectively, prior to incorporation into biomass. The strong agreement observed in the nitrate-nitrite condition can be attributed to the fact that nitrite requires only a single reduction step (v_{42} in Table H.1), while nitrate must be reduced in two steps (v_{41} and v_{42}), making it more energy intensive. The energetic burden of assimilating these oxidized nitrogen forms likely limits the cell's ability to grow faster, which is consistent with the simulated optimal growth rates close to the dilution rate.

The nitrate condition under high salinity medium follows a similar trend, with the model predicting a growth rate very close to the experimental value. This behavior is biologically plausible, as elevated salinity imposes osmotic stress that increases maintenance energy demands and reduces the resources available for biomass synthesis. The fact that the model reproduces this growth-limiting effect suggests that it is sufficiently sensitive to environmental factors beyond the nutrient supply.

In contrast, the model predicts a significantly higher growth rate under the nitrate-urea condition. Urea is energetically more favorable because it bypasses the need for reduction, resulting in lower ATP and reductant requirements for nitrogen assimilation. Although the experimental growth rate remains constrained by the fixed dilution rate of the photobioreactor, the higher μ_{opt} is consistent with this metabolic advantage. This highlights the difference between metabolic potential, as predicted by the model, and actual performance under controlled conditions. This is in line with proteomic evidence showing that urea supplementation leads to specific upregulation of urea assimilation in *Arthrospira*, reflecting a metabolic adjustment to the more energetically favorable nitrogen source [198].

Discrepancies between optimized and experimental growth rates are common in constraint-based metabolic modeling, particularly in underdetermined networks. It is well established that optimal and observed growth can differ by a factor of 1.5 to 3 depending on environmental context, regulatory constraints, and model assumptions [117].

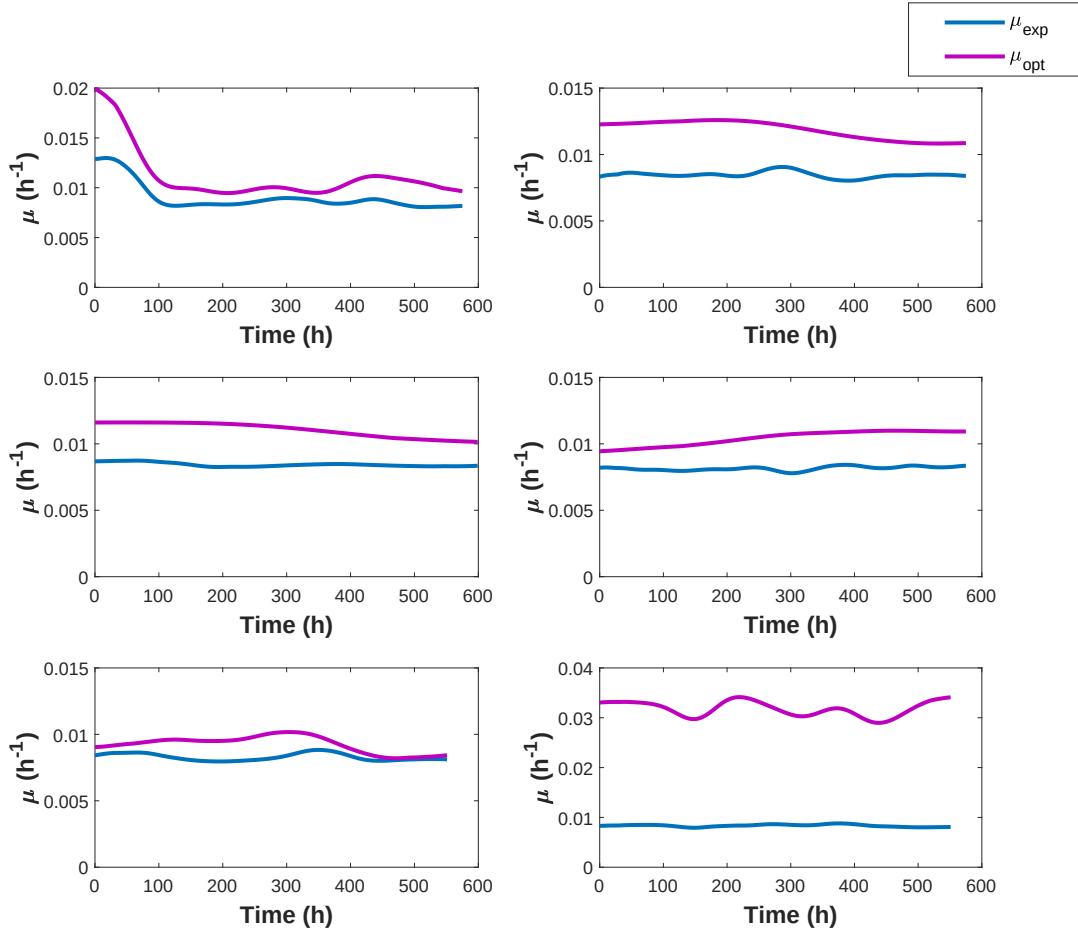


Figure 4.9: Experimental growth rate μ_{exp} and specific growth rate μ_{opt} obtained from dFBA for the different feeding strategies: 30mM-N nitrate (top-left, top-right, center-left), 30mM-N nitrate in high salinity medium (center-right), 15mM-N nitrate + 15mM-N nitrite (bottom-left) and 15mM-N nitrate + 15mM-N urea (bottom-right).

These results, particularly the strong agreement between the predicted and measured growth rates under nitrate-only, nitrate-nitrite, and high-salinity nitrate conditions, underscore the reliability and robustness of the model in capturing the metabolic behavior of *Arthrospira* across diverse nitrogen sources.

4.3.3.2.3 Dynamic Flux Variability Analysis

Subsequently, all experimental constraints are applied and dFVA is performed to explore the range of possible fluxes through each reaction in the network. Fig. 4.10 to Fig. 4.13 present the dFVA results for some metabolic reactions under the 30mM-N nitrate feeding condition. Specifically, reactions v_{27} to v_{36} are related to the tricarboxylic acid cycle, reactions v_{108} to v_{111} are related to cellular energy metabolism, reactions v_{116} to v_{123} represent the biosynthesis of carbohydrates, and reactions v_{126} to v_{141} represent the biosynthesis of lipids. Narrow flux intervals indicate higher prediction confidence, while broader intervals suggest uncertainty or metabolic flexibility and imply that the model could benefit from additional constraints. In particular, when a flux interval narrows to

a single value, it signifies that the corresponding flux is calculable, as defined through calculability analysis, meaning that it is uniquely determined by the imposed constraints.

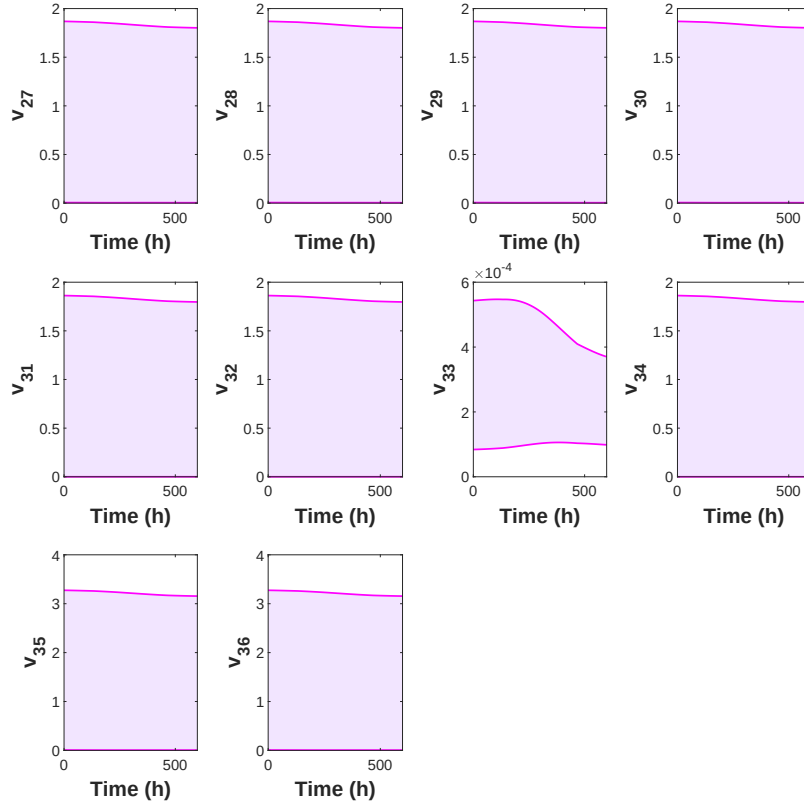


Figure 4.10: Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{27} to v_{36} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the tricarboxylic acid cycle.

As expected, reactions involved in the biosynthesis of biomass precursors (e.g., lipids, amino acids, nucleotides, and carbohydrates) exhibit minimal variability (see Fig. 4.12 and Fig. 4.13). This is directly due to their direct coupling with the fixed specific growth rate, which constrains the amounts of biomass components that must be produced. As a result, the lower bounds of the fluxes in these anabolic pathways are necessarily non-zero, as the required precursors must be synthesized at a rate proportional to the enforced growth rate. This tight coupling reinforces the biological plausibility of these results and confirms that the network structure correctly reflects biomass demands.

In contrast, reactions from central carbon metabolism (glycolysis, tricarboxylic acid cycle, and pentose phosphate pathway) show broader flux intervals, as illustrated in Fig. 4.10. These pathways provide intrinsic metabolic flexibility in carbon routing and co-factors balancing, offering multiple alternative routes for metabolite interconversion and redox maintenance. Similarly, reactions associated with cellular energy metabolism (e.g., ATP, NADH and NADPH generation and utilization) display considerable variability, as shown in Fig. 4.11, consistent with their multiple sources of production throughout the network.

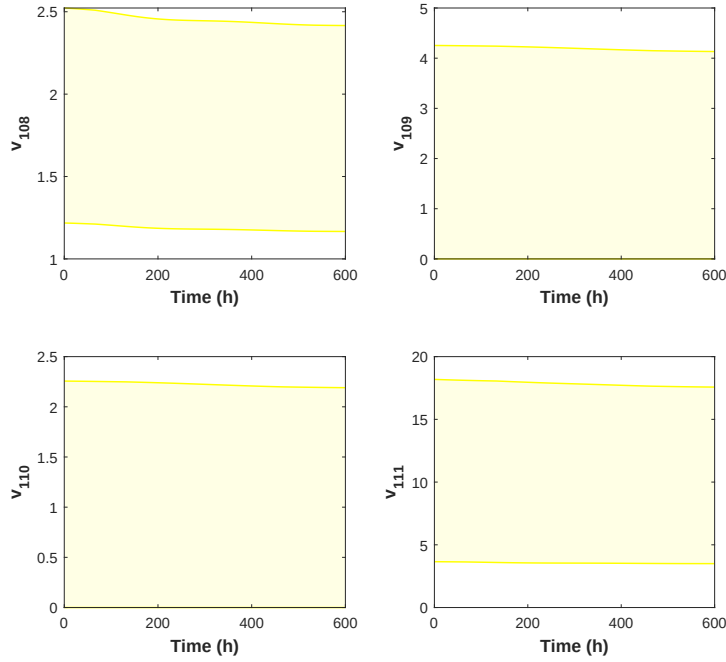


Figure 4.11: Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{108} to v_{111} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the cellular energy metabolism.

In particular, the flux range for the ATP synthesis reaction (via v_{111}) remains large, as ATP can be produced via substrate-level phosphorylation, oxidative phosphorylation, and photophosphorylation. Since all of these routes are present and not fully constrained in the network, the model retains the freedom to mix and redistribute flux across these pathways as needed to satisfy energy demands, which is a property that reflects biological flexibility, but also introduces modeling ambiguity.

This flexibility is especially pronounced in the absence of constraint on non-growth associated maintenance energy (commonly referred to as ATP maintenance, via the reaction v_{112} in Table H.1). According to the literature, such as the study conducted in [210], maintenance energy requirements for microorganisms typically range between 0.1 and 0.4 $\text{mmol ATP.g}_{DW}^{-1}.\text{h}^{-1}$, though these values are organism-specific. In the present study, no ATP maintenance constraint was imposed due to the lack of experimentally determined values for *Arthrospira* sp. PCC 8005. However, incorporating a species-specific maintenance demand in future refinements may help to narrow the flux intervals related to ATP turnover and reduce some of the flexibility observed in energy-related pathways.

Additionally, while the photon uptake rate was bounded on the basis of physical and geometric parameters (e.g., light intensity, illuminated surface, light efficiency), these constraints only influence the photophosphorylation route of ATP synthesis (via v_{108}). The model remains free to exploit other ATP-generating routes, such as oxidative phosphorylation, if doing so satisfies the objective function without violating any constraints. This behavior is a well-known limitation of flux balance analysis. The latter optimizes the objective function under a set of stoichiometric and constraint-based assumptions but does not inherently select biologically preferred or realistic flux distributions unless

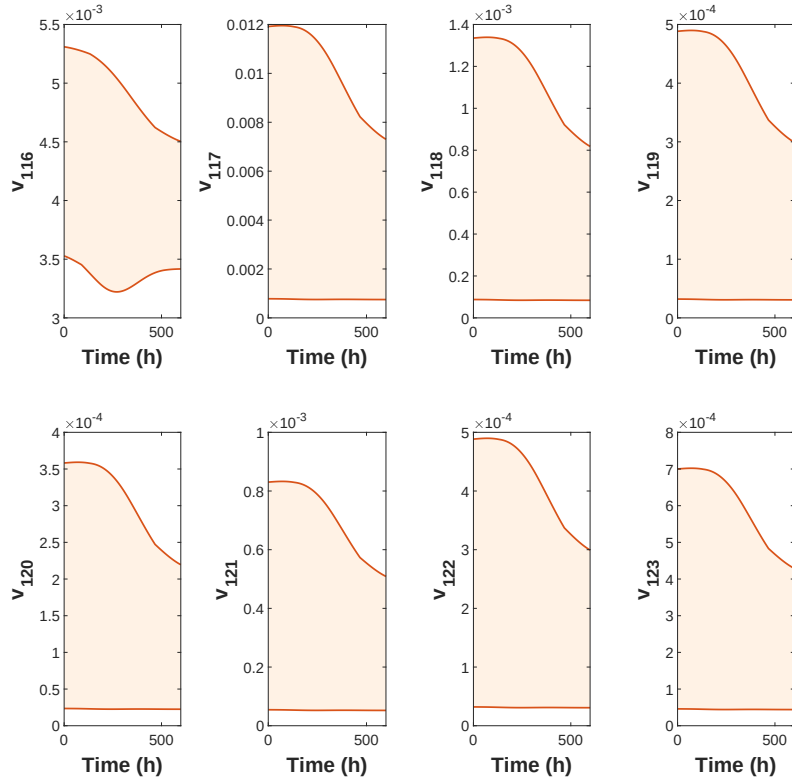


Figure 4.12: Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{116} to v_{123} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the biosynthesis of carbohydrates.

additional constraints or objectives are introduced. For instance, when energy is not limiting, the model may satisfy ATP requirements entirely through photophosphorylation and ignore other pathways, even if they are active in vivo.

Several modeling strategies can be adopted to address this issue:

- Introducing ATP balancing constraints to ensure the participation of multiple energy-generation routes;
- Using multi-objective optimization (e.g., maximizing biomass while minimizing photon use or balancing cofactor production);
- Modeling explicitly ATP from different sources as separate pseudo-reactions, each with distinct costs or limits.

Despite the inherent flexibility of the FVA framework, the analysis captures biologically meaningful trends and supports the structural validity of the network. Moreover, these results are consistent with the findings of [202], who performed a static FVA analysis on a genome-scale metabolic network (875 reactions), and also reported similar FVA patterns where anabolic pathways were tightly constrained and central carbon metabolism retained greater flexibility. Overall, the flux variability analysis supports the validity of the network and highlights important areas for future refinement.

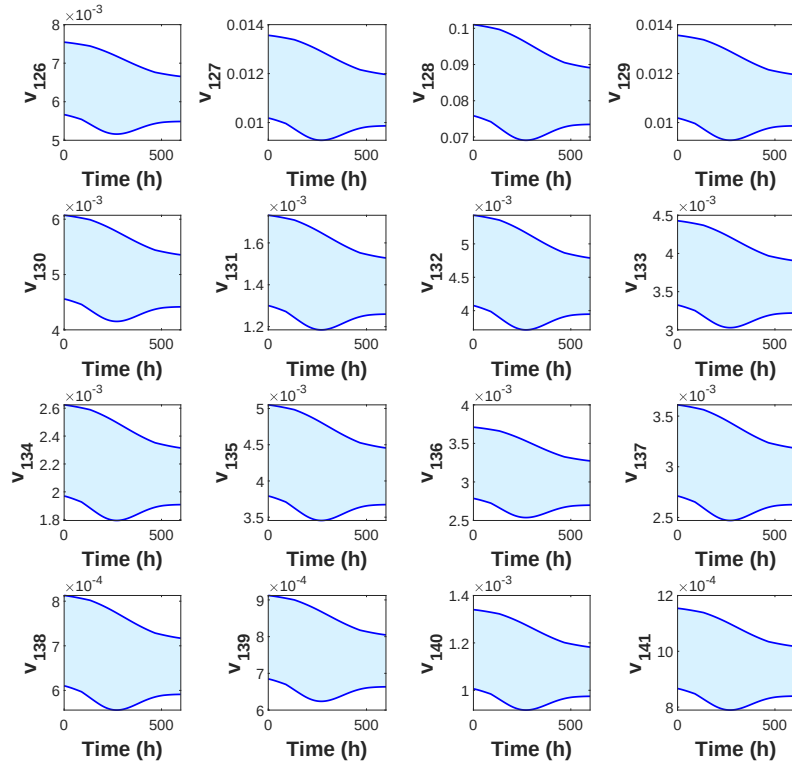
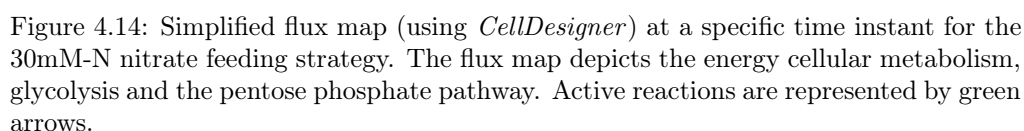


Figure 4.13: Results from the dynamic flux variability analysis under the 30mM-N nitrate feeding condition. Fluxes v_{126} to v_{141} (expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$) are related to the biosynthesis of lipids.

4.3.3.2.4 Dynamic Flux Balance Analysis

Flux balance analysis is proposed to analyze the flux distribution of the model. This is performed by imposing the experimental uptake and secretion rates, as well as the specific growth rate, while maximizing ATP production. After the analysis of the flux map, the results are in agreement with similar studies [208]. The simplified flux map is given in Fig. 4.14 for a specific time instant for the 30mM-N nitrate feeding condition. It is observed that the regime is autotrophic and is characterized by high fluxes in the photosynthetic pathways and the activation of the Calvin-Benson cycle. Upper glycolysis operates in the glyconeogenic direction to produce carbohydrates and sugar precursor metabolites (PEP, G6P and R5P), essential for growth. The pentose phosphate pathway (PPP) is in the reductive mode, and ATP synthesis is driven by the PMF and is synthesized via photophosphorylation and oxidative phosphorylation (in addition to substrate-level phosphorylation). Also, as already discussed in Sec. 4.3.2.1.2, it is interesting to note that glyceraldehyde-3-phosphate (G3P) is produced mainly via the Calvin-Benson cycle. Reductive PPP contributes a little, but glycolysis (that is, the breakdown of fructose-1,6-bisphosphate) is not a major pathway for G3P production, particularly when the organism is in a photosynthetic state. In addition, analysis of the complete map shows the activation of the citric acid cycle and that the GS-GOGAT pathway and the urea cycle-like pathway are involved in nitrogen assimilation.



In addition to growth simulations, analyzing how the organism responds to different nitrogen sources provides deeper insights into the predictive power of the metabolic model. For this purpose, it is proposed to perform flux balance analysis (FBA and/or dFBA) simulations under the four experimental feeding strategies: 30mM-N nitrate; 15mM-N nitrate + 15mM-N nitrite; 15mM-N nitrate + 15mM-N urea; and 30mM-N nitrate under high salinity conditions. These simulations are constrained by measured uptake and secretion rates and a fixed specific growth rate, reflecting the controlled conditions of a continuous photobioreactor. The fluxes obtained are then compared with established physiological knowledge and experimental observations to validate the structure of the network.

145

growth rate is fixed by experimental data, the internal redistribution of fluxes remains highly sensitive to assumptions about cofactor balancing, energy demands, and maintenance reactions. For example, ATP synthesis via photophosphorylation (reactions v_{108} and v_{111} in Table H.1) can produce excessive ATP if unconstrained, especially under energy-favorable nitrogen sources (e.g., urea) that require less reductive assimilation effort. Meanwhile, the representation of non-growth associated maintenance energy (e.g., ATP maintenance reaction through flux v_{112}) remains difficult to calibrate due to limited organism-specific data. Furthermore, even when photon uptake is constrained, the model may not fully capture how light energy is partitioned in vivo between linear and cyclic electron flow, oxygen evolution, and photoprotective processes. These uncertainties particularly affect the predictions of energy metabolism, making it difficult to compare flux values across different nitrogen-feeding conditions. However, when interpreted with care, the model provides consistent and biologically meaningful insights into how the cell reallocates its internal resources.

As illustrated in Fig. 4.15, several fluxes reflect known physiological adaptations. When urea is used as a nitrogen source, higher fluxes are observed through reactions v_{38} and v_{39} of the GS-GOGAT pathway, indicating a more efficient capture of ammonium released from urea hydrolysis. This facilitates faster glutamine and glutamate synthesis, which supports increased production of storage and N-rich compounds such as cyanophycin (via reaction v_{149}) and arginine (via reaction v_{47}). In addition, chlorophyll production (through reaction v_{147}) is enhanced under urea, possibly due to internal release of CO_2 during hydrolysis of urea, which can be recycled for photosynthesis.

In contrast, nitrate- and nitrite-fed cultures exhibit lower fluxes through the GS-GOGAT and pigment synthesis pathways, likely due to the greater energy and reductant demands associated with the two-step reduction of nitrate/nitrite. These conditions also increase metabolic pressure on glycolysis, the tricarboxylic acid cycle, and the pentose phosphate pathway to supply ATP, NADH, and NADPH for nitrogen assimilation. These simulation results are consistent with the experimental findings in *Arthrospira* sp. PCC 8005 under similar conditions as described by [199].

Under high salinity, additional energy and redox equivalents are required to support osmotic balance, ion homeostasis, and stress responses. As shown in Fig. 4.15, fluxes through proline synthesis (using reaction v_{60}) and arginine biosynthesis (using reaction v_{47}) increase, reflecting the production of osmoprotective solutes (e.g., proline, glycine) and the involvement of the ornithine cycle in managing internal ammonium levels. In addition, the increased flux through ATP synthesis supports the increased maintenance demands imposed by salt stress. Furthermore, under nitrate feeding in high salinity conditions, the GS-GOGAT pathway (through fluxes v_{38} and v_{39}) exhibits increased fluxes compared to standard nitrate feeding. This likely reflects a cellular response to osmotic stress, where glutamate (the product of GS-GOGAT) serves not only as a key metabolite in nitrogen assimilation but also as a precursor for compatible solutes and other stress-related compounds. Therefore, the increased flux values indicate a physiological adaptation to maintain nitrogen homeostasis and support osmoprotection under salinity stress.

Although metabolic fluxes are generally stable across most nitrogen feeding strategies in continuous steady-state cultivation, an exception is observed in the condition involving 30mM-N supplementation under high salinity (notably for fluxes v_{38} and v_{39} in Fig. 4.15). In this case, intracellular fluxes associated with the GS-GOGAT cycle display a noticeable time-dependent increase, diverging from the expected steady-state behavior.

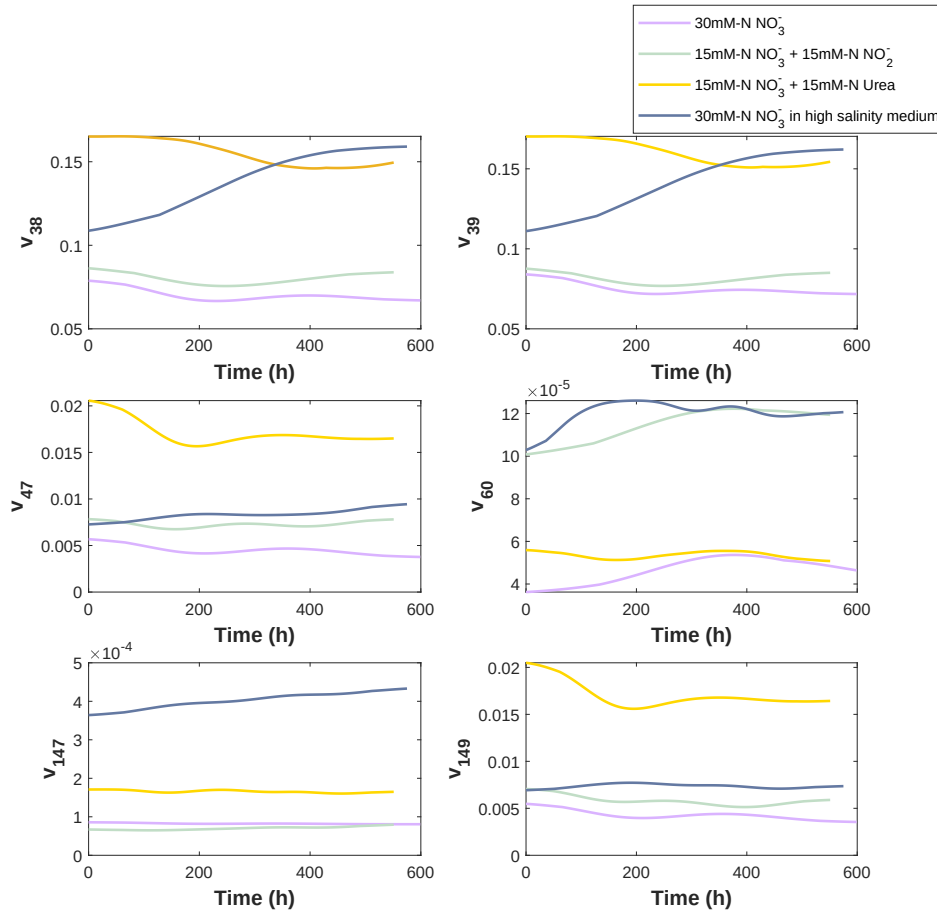


Figure 4.15: The fluxes are expressed in $\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$. Fluxes v_{38} and v_{39} are related to the GS-GOGAT pathway, v_{47} , v_{60} , v_{147} and v_{149} correspond to the production rate of arginine, proline, chlorophyll and cyanophycin, respectively.

Upon closer examination of the experimental data, it was found that the residual ammonium concentration in the culture medium increased significantly over the course of the experiment. This substantial increase is inconsistent with steady-state assumptions and suggests either a deviation in the culture dynamics (e.g., nitrogen overaccumulation or altered uptake capacity due to salt stress) or a potential experimental disturbance.

Consequently, it is important to note that the comparison of the metabolic impact of different nitrogen sources presents inherent challenges because of the high sensitivity of metabolic flux analysis to experimental measurement errors. Even under controlled continuous cultivation, small inaccuracies in extracellular metabolite concentrations, biomass estimates, or gas exchange rates can significantly affect the calculated flux distributions. Although every effort was made to pre-process and smooth the data consistently across conditions, the interpretation of flux differences should be approached with caution, particularly when variations fall within the margin of experimental error.

Overall, despite the model simplifications, the predicted intracellular fluxes align well with the expected physiological responses to different nitrogen regimes. The trends observed – such as higher pigment and nitrogen reserve synthesis under urea, elevated osmolyte production under high salt, and increased energetic demands under nitrate-only feeding – underscore the model ability to reproduce condition-specific metabolic

adjustments. These findings reinforce the validity of the metabolic network and its utility in exploring metabolic trade-offs in *Arthrospira* sp. under varying environmental conditions.

4.3.3.3 Further Validation

Cross-validation is quite complex when applied to metabolic networks because data are typically used to constrain the model. Most cross-validation strategies in MFA focus on comparing predicted fluxes with experimental fluxes when ^{13}C -labeled flux data are available as ground truth. Such data are not available for the present study. However, three experiments have been conducted using the same feeding strategy (i.e., 30mM-N nitrate feeding). Consequently, a further validation is proposed by computing mean exchange fluxes from two datasets (training phase) to predict intracellular fluxes via FBA, which will then be compared against the results from the third dataset (test phase) using flux similarity metrics. The idea is to test how well the general trends in flux distributions generalize when faced with a completely new set of constraints. The results are summarized in Table 4.2 for all folds, and the relative flux error (RFE) is used as a metric (see [211] for more information).

The metric in Eq. 4.26 measures the relative difference between two flux distributions and is normalized to range between 0 (indicating perfect similarity) and 1 (indicating complete dissimilarity). The proposed indicator is defined as follows:

$$RFE = \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{J} \sum_{j=1}^J \frac{|f_{train,i,j} - f_{test,i,j}|}{\max(|f_{train,i,j}|, |f_{test,i,j}|, \epsilon)} \right) \quad (4.26)$$

where $f_{train,i,j}$ and $f_{test,i,j}$ are the i^{th} flux at time instant j from the training set or the test set, respectively. ϵ is a small constant to avoid division by zero (typically, $\epsilon = 10^{-6}$). The low RFE values across all folds indicate a high degree of similarity between predicted and observed fluxes, suggesting that the model captures generalizable flux trends. Therefore, combined with previous validation efforts, these results reinforce the validity and robustness of the underlying network structure. The results of this further validation further support the constructive methodology presented in this chapter for developing consistent metabolic networks.

Table 4.2: Relative flux error (RFE) for all folds.

	FOLD (1-2) VS 3	FOLD (1-3) VS 2	FOLD (2-3) VS 1
RFE	0.0518	0.0538	0.0581

4.4 Application to Unicellular Microalgae

Unicellular microalgae are a diverse group of photosynthetic eukaryotic microorganisms found in aquatic environments. These organisms have gained increasing scientific attention due to their metabolic plasticity, ecological relevance, and potential for biotechnological exploitation in sectors such as biofuel production, nutraceuticals, and carbon sequestration. As primary producers, microalgae contribute significantly to global carbon cycling and represent model organisms for studying cellular responses to fluctuating environmental conditions [212, 213]. Their relatively simple cellular organization, coupled with a high degree of metabolic complexity, makes them particularly attractive for

systems biology approaches aimed at elucidating the regulatory and functional architecture of metabolism under various growth conditions.

Among the available modeling frameworks for understanding and predicting microalgal metabolic behavior, the use of constraint-based models and elementary flux mode (EFM) analysis has been shown to be particularly effective in capturing the structure and capabilities of metabolic networks [74, 155]. One notable contribution in this field is the **DRUM** framework developed by Baroukh et al. [2, 15, 191, 208], which enables metabolic modeling under non-balanced growth conditions by incorporating a macroscopic, EFM-derived structure tailored to the carbon metabolism of microalgae. This approach provides a valuable foundation for exploring intracellular flux distributions, especially in systems characterized by dynamic accumulation of internal metabolites.

The present case study builds on the existing DRUM approach to further develop and validate the constructive bottom-up methodology proposed in Sec. 4.2 for elaborating mid-scale metabolic networks. The objectives of this study are twofold: (i) to demonstrate the validity and robustness of the proposed methodology by reconstructing an extended version of the metabolic network originally proposed by [2], and comparing both the network topology and the model outputs, and (ii) to address the challenge of internal metabolite accumulation, which is a phenomenon that can disrupt the assumption of metabolic steady-state and complicate the application of classical metabolic flux analysis (MFA) techniques. In doing so, this case study also showcases how the constructive modeling framework can be adapted to reconcile EFM-based modeling (covered in Chapter 3) with non-balanced growth conditions, thereby extending the applicability of MFA to more physiologically realistic scenarios.

4.4.1 Elaboration of a Mid-Size Metabolic Network

In this section, the constructive bottom-up modeling strategy is applied for the development of a mid-size metabolic network tailored to the carbon metabolism of unicellular microalgae. Unlike the first case study, where the network was assembled *de novo* from elementary metabolic functions, the present study benefits from the availability of an existing and validated network proposed in [2]. First, the current section presents the existing network and the main metabolic pathways. Then, an extension of the network is proposed by applying the methodology and the iterative process described in Fig. 4.2.

4.4.1.1 Starting Point

This section outlines the main metabolic pathways that are included in the existing network, detailing key modeling assumptions and organism-specific modifications introduced during its construction. In addition, particular attention is paid to the distinguishing features of eukaryotic microalgal metabolism (relevant to the present case) in contrast to the prokaryotic system investigated in the previous case study.

4.4.1.1.1 Main Metabolic Routes

The initial metabolic network used in this study is derived from the work developed by [214], which serves as the foundational scaffold for the model proposed by [2]. This original network offers a compact and generic representation of eukaryotic microalgal metabolism, comprising 152 metabolites and 160 reactions. It captures the essential core pathways common to most microalgae, including photosynthesis, glycolysis, the pentose phosphate pathway, the tricarboxylic acid (TCA) cycle, oxidative phosphorylation, and the biosynthetic pathways for macromolecules such as carbohydrates, lipids,

proteins, nucleic acids, and chlorophyll. The choice of this model is justified by its alignment with other reconstructions of eukaryotic microalgal metabolism, providing a well-established basis for further elaboration.

Several organism-specific modifications were introduced. Transport reactions, particularly those associated with chloroplast exchanges, were refined. Lipid biosynthesis was explicitly represented through the formation of phosphatidic acids (PA), modeled using detailed fatty acid profiles obtained from a similar strain. The protein, DNA and RNA synthesis reactions were derived from the average amino acid composition of *Chlamydomonas reinhardtii*, leveraging the assumption that the macromolecular composition is relatively conserved in eukaryotic microalgae. These biosynthetic pathways were complemented by data-driven modifications informed by the experimental study of [215].

A central conceptual adjustment in the DRUM framework lies in the definition of biomass. Traditional biomass formulations assume fixed stoichiometric contributions of macromolecules such as carbohydrates and lipids. However, under dynamic conditions such as diel cycles, these components may undergo significant accumulation or depletion. To account for this, the network distinguishes between **functional biomass** (comprising proteins, RNA, DNA, and chlorophyll) and a separate pool of accumulating metabolites (e.g., lipids and carbohydrates). This separation enables the modeling of non-steady-state phenomena without artificially constraining the dynamics of key storage compounds. Consequently, the biomass equation was reformulated to include a stoichiometric balance of only functional components, while accumulating pools were treated as independent and dynamically regulated metabolites. Adjustments to the functional biomass composition, including the addition of PA, were made to match observed elemental ratios (e.g., C/N), thus ensuring consistency with empirical data.

Further refinements were applied to improve physiological realism. Glycerol synthesis and excretion were removed. Starch synthesis was lumped into the broader carbohydrate pathway. The reaction artificially converting NADPH to NADH was excluded to prevent metabolic shortcuts that bypass the pentose phosphate pathway – an issue commonly observed in flux balance analysis, as discussed in Sec. 4.3.2.2.2. In addition, photorespiration was omitted, given its minimal activity in the experimental observations. The quantum requirement for photosynthesis was revised from 8 to 10 photons per CO₂ molecule fixed, reflecting typical values found in the literature and accounting for photon losses. Finally, several minor adjustments were made to reaction reversibility and chloroplast export routes. Most notably, glyceraldehyde-3-phosphate (GAP) was modeled as the exported metabolite instead of dihydroxyacetone-phosphate (DHAP), based on more recent experimental evidence. The metabolic network proposed by [2] is given in Table I.1 in the Appendix I and contains 157 metabolites and 162 reactions.

4.4.1.1.2 Eukaryotic vs. Prokaryotic Modeling Paradigms

This section provides a comparative perspective on the metabolic modeling paradigms applied to eukaryotic and prokaryotic organisms. It serves three primary goals: (i) to emphasize fundamental biological differences between eukaryotic and prokaryotic cellular organization, (ii) to analyze in detail the network proposed by [2], and (iii) to compare this network structure with the metabolic network developed in Sec. 4.3 for a prokaryotic microorganism. In doing so, our objective was to assess the structural validity of our earlier prokaryotic network and establish points of convergence between the modeling strategies despite the underlying biological divergence.

4.4.1.1.2.1 Biological and Organizational Differences Fundamentally, eukaryotic cells are structurally more complex than their prokaryotic counterparts, featuring membrane-bound organelles such as chloroplasts, mitochondria, nuclei and, in some cases, glyoxysomes (see Fig. 2.3). This compartmentalization not only spatially segregates metabolic activities, but also introduces additional layers of regulation and transport dynamics. In contrast, prokaryotic cells lack internal membrane-bound organelles, resulting in a more homogeneous intracellular environment where metabolic processes occur in a shared cytoplasmic space. This organizational difference directly impacts the construction of the metabolic network and its interpretation. In eukaryotic models, metabolic reactions must often be localized to specific compartments, with associated transport reactions ensuring connectivity between them. This leads to increased network complexity and a richer set of constraints, especially when accounting for energetic coupling between compartments. In prokaryotes, while intracellular micro-compartmentalization (e.g., carboxysomes) can exist, the metabolic architecture is generally more streamlined. A schematic comparison of metabolic compartmentalization is provided in the Appendix J.

4.4.1.1.2.2 Comparative Network Architecture The metabolic network elaborated by Baroukh et al. [2] is designed for *Tisochrysis lutea*, an eukaryotic microalga, and incorporates this structural complexity by explicitly modeling chloroplast-associated reactions. The light-dependent reactions and the Calvin-Benson-Bassham cycle are located within the chloroplast compartment and involve the reactions R_1 through R_{14} in Table I.1. These reactions are associated with metabolites prefixed with *c* (e.g., cGAP, cRuBP), which denote their chloroplastic localization. This compartmentalized representation aligns with the known intracellular organization of photosynthetic eukaryotes, where light harvesting and carbon fixation occur within the chloroplast.

Central carbon metabolism is largely conserved and encompasses glycolysis and gluconeogenesis (R_{15} to R_{25}), as well as the pentose phosphate pathway (R_{37} to R_{43}). These subsystems are functionally analogous to those in the prokaryotic network constructed in our previous case study. The TCA cycle (R_{26} to R_{36}) is modeled in its canonical form here, in contrast to the prokaryotic network, which included a bypass through succinate semialdehyde to circumvent the step of α -ketoglutarate dehydrogenase, which is a known adaptation in some prokaryotes for redox balancing or under stress conditions (see Sec. 4.3.1.1.5).

Nitrogen assimilation also shows differences. In the present network, the glutamine synthase/glutamate synthase (GS-GOGAT) pathway is represented by reactions R_{56} and R_{57} . In contrast, this eukaryotic network does not include the urea cycle-like pathway, which was present in the prokaryotic network. This absence may reflect either a lack of the corresponding enzymatic machinery in microalgae or simply the fact that urea was not considered in the experimental data. In practice, microalgae often recycle nitrogen through alternative routes such as amino acid interconversion or direct assimilation of ammonium or nitrate.

The biosynthesis of amino acids is covered in the reactions R_{55} to R_{93} and largely mirrors the set of anabolic functions present in the prokaryotic network, although there are some differences in pathway structure and individual reactions, reflecting organism-specific enzymatic pathways. Importantly, the eukaryotic network includes reactions for the synthesis of metabolites such as anthranilate, hydroxyproline, and glyoxylate, which were not explicitly modeled in the prokaryotic counterpart.

Lipid metabolism is represented through the reactions R_{103} to R_{123} , where lipids are collectively represented as phosphatidic acid (PA). Although this abstraction is slightly less granular than the detailed acyl-chain representation used in our prokaryotic model, it remains biologically sound and sufficient for macroscopic flux modeling. Similarly, THF metabolism (R_{94} to R_{102}), nucleic acid synthesis (R_{124} to R_{141}), and chlorophyll biosynthesis (R_{142} to R_{147}) are modeled in a conceptually equivalent manner to our own network.

A key distinction arises in carbohydrate metabolism. In the network proposed in [2], carbohydrate synthesis is represented by a single lumped reaction R_{148} , while our prokaryotic network offers a more detailed description of the enzymatic steps involved. Furthermore, a significant divergence is observed in the modeling of energy metabolism. The present network considers photophosphorylation (R_1) and oxidative phosphorylation (R_{50} and R_{51}) using the simplified and classical stoichiometries that capture net ATP and reductant yields. In contrast, our previous model includes an explicit representation of the proton motive force (PMF), including proton translocation and membrane energetics. Although mechanistically more detailed, both approaches ultimately describe equivalent biological processes and result in comparable ATP generation from redox reactions and light energy, particularly when used in macroscopic simulations.

This comparative analysis demonstrates that, despite differences in structural complexity, cellular compartmentalization, and modeling resolution, the essential architecture of metabolism is largely conserved across our previous prokaryotic network and the present eukaryotic system. Therefore, this convergence across models supports the structural validity of the network developed in Sec. 4.3 for prokaryotic cells and supports the methodological soundness of the constructive bottom-up strategy.

4.4.1.2 Iterative Methodology

The previous sections introduced the metabolic network developed by [2] for *Tisochrysis lutea*, summarized its main pathways and outlined the structural and biological differences between the eukaryotic and prokaryotic metabolic systems. By analyzing the organization of this eukaryotic network and comparing it with our previously developed prokaryotic model, we have ensured both the biological and mathematical relevance of the reference network. This analysis also reinforced the consistency and generality of our constructive bottom-up approach.

Building upon this validated network, we now proceed to elaborate a more detailed version, using our iterative and modular strategy, presented in Sec. 4.2. It is important to emphasize that the existing network is already comprehensive in most aspects of central metabolism. Therefore, our efforts focus on refining a specific functional module, namely the **biosynthesis of carbohydrates**, which is represented by a single lumped reaction in the original formulation. By constructing a more biochemically explicit pathway for carbohydrate synthesis, we aim to explore whether increased structural granularity leads to improved predictive accuracy and closer alignment with experimental observations.

As a reminder, this case study thus serves two main objectives. First, it offers a complementary perspective of the **DRUM** framework by demonstrating how metabolic accumulation – though not explicitly addressed via novel mechanism – can be better contextualized within a richer metabolic structure. Second, and more importantly, it provides an additional validation of our methodology for identifying and elaborating mid-

scale metabolic networks, showing that even localized refinements based on bottom-up principles can enhance model fidelity without compromising parsimony.

4.4.1.2.1 First Iteration

To improve biological fidelity and validate the effectiveness of the constructive bottom-up modeling strategy, a more detailed pathway for carbohydrate biosynthesis in *Ti-sochrysis lutea* is proposed. As depicted in Fig. 4.2, this is achieved by identifying relevant intermediary steps through the literature and genomic data (if available) from *T. lutea* and closely related haptophyte microalgae. The revised network incorporates explicit formation and utilization of activated sugar donors such as UDP-glucose, UDP-N-acetylglucosamine, GDP-mannose, and mannitol-1-phosphate. The additional reactions are given in Table I.2 and are deduced from [216–219].

For the first iteration, all metabolic reactions in Table I.2 are added. The newly constructed pathway distinguishes between three major biosynthetic routes contributing to total carbohydrate biosynthesis: (i) **chrysolaminarin** (β -1,3-glucan), the main storage polysaccharide, (ii) **mannitol**, a soluble polyol and secondary storage molecule, and (iii) **glycan precursors** (UDP-N-acetylglucosamine and GDP-mannose), which include nucleotide-activated sugars involved in glycoprotein biosynthesis and cell wall components. These products are integrated into a composite metabolite pool denoted as $carbohydrate_{av}$ via R_{148}^* .

Based on the findings on haptophyte carbohydrate metabolism [220] and in the absence of species-specific quantification, we estimated the relative contributions of these components to the total carbohydrate pool as follows: 70% chrysolaminarin, 20% mannitol and 10% glycan precursors. This breakdown enables defining the reaction R_{148}^* that aggregates the various biosynthetic routes while preserving key factors. Then, reactions R_{163} to R_{173} are added and reaction R_{148}^* replaces the original oversimplified lumped reaction R_{148} and allows for finer resolution of carbon flux toward various carbohydrate forms.

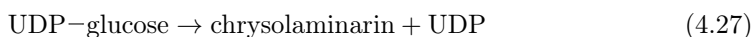
In order to check the consistency of the network as well as structural and topological network properties, flux coupling analysis is performed. For this purpose, a particular uptake scenario is considered for which light, a carbon source, and a nitrogen source are available. Table 4.3 summarizes some iterations. For the first iteration, FCA shows blocked reactions, i.e., reactions incapable of carrying a flux under steady-state conditions. Therefore, the structure of the network needs to be revised.

Table 4.3: Information related to metabolic networks from the iterative procedure. The number of metabolites is comprised of internal metabolites and intracellular energetic cofactors, and the number of reactions includes the intracellular reactions and the transport/exchange reactions.

Iteration	# METABOLITES	# REACTIONS	# BLOCKED REACTIONS
0	157	162	0
1	169	173	14
2	162	167	0

4.4.1.2.2 Second Iteration

The blocked reactions identified in the previous iteration can be rationalized on the basis of the network structure. First, the compound $(\beta\text{-1,3-glucan})_n$ involved in reaction R_{164} is not explicitly represented as an intermediate metabolite in the current network. To address this, it is appropriate to model the polymer as a terminal product, e.g., either as a global sink or as an accumulating metabolite, and to define a lumped biosynthetic reaction such as:



This chemical equation replaces R_{164} and represents the $\beta\text{-1,3-glucan}$ polymerization, even if each glucosyl unit is not modeled individually.

Similarly, the reaction R_{173} , which describes the conversion of mannitol-1-P to free mannitol, would be relevant only in a model in which mannitol is tracked as a distinct and accumulating metabolite within a soluble carbohydrate pool. However, in the present case, mannitol-1-P is incorporated directly into the composite *carbohydrate_{av}* pool, and is not further used or exchanged as a free intermediate.

Finally, the reactions R_{165} to R_{168} , as well as the reactions R_{169} to R_{171} , can be further simplified by applying lumping techniques, as described in Sec. 4.2.1. On this basis, a final draft of the metabolic network can now be established, comprising 162 metabolites and 167 reactions.

4.4.1.2.3 Eukaryotic vs. Prokaryotic Networks

An important point of divergence between metabolic networks across different domains of life lies in the biosynthesis and composition of carbohydrates. The metabolism of carbohydrates is not universally conserved in its structural or functional detail, and instead reflects the evolutionary adaptations and physiological roles of carbohydrates in different organisms. In the prokaryotic system examined in the first case study (specifically, *Arthrospira* sp.), genomic analysis revealed the presence of pathways leading to the synthesis of glycogen (as the main storage polymer), peptidoglycan (a major component of the cell wall), cyclitol derivatives and dTDP-rhamnose, each of which plays a structural or protective role in bacterial physiology. In contrast, the eukaryotic system investigated in the second case study (*Tisochrysis lutea*, a unicellular haptophyte alga) showed a distinctly different carbohydrate profile. Here, the main carbohydrate products include chrysolaminarin (a $\beta\text{-1,3-glucan}$ storage polymer), mannitol-1-phosphate (a soluble carbon reserve), UDP-glucose and UDP-N-acetylglucosamine, reflecting the absence of peptidoglycan and the reliance of different storage and glycosylation strategies typical of microalgae.

These differences underscore the fact that carbohydrate metabolism must be tailored to the biological context of each organism. Although prokaryotes frequently use carbohydrates for structural purposes (e.g., cell wall biosynthesis), many eukaryotes, particularly photosynthetic microalgae, use them primarily for carbon storage and osmotic regulation. Even among eukaryotes, significant variability exists: green algae such as *Chlamydomonas reinhardtii* synthesize starch, while diatoms may produce chrysolaminarin along with other unique polysaccharides. These observations highlight the importance of accounting for organism-specific metabolic features when reconstructing carbohydrate biosynthetic pathways.

4.4.2 The DRUM Framework

In this section, a brief overview of the DRUM framework is provided, emphasizing its suitability for modeling metabolism under non-steady-state conditions, such as those imposed by diurnal cycles.

Although this chapter focuses on the microscopic modeling of cellular cultures through the development of a constructive bottom-up approach to elaborate mid-scale metabolic networks (see Sec. 4.2), this second case study concludes with a macroscopic validation phase. To evaluate the predictive capability of the extended network and address the challenges posed by the accumulation of internal metabolites under dynamic growth conditions, we adopt the DRUM (Dynamic Reduction of Unbalanced Metabolism) framework proposed by Baroukh et al. [2]. While this framework is not a methodological contribution of this work, its relevance lies in its ability to translate elementary flux mode (EFM)-based networks into dynamic, macroscopic models capable of capturing non-steady behavior. Readers interested in the full mathematical formulation and theoretical underpinnings are referred to the original publication.

This brief detour into macroscopic modeling is thus intended solely as a means of validating the network constructed using our bottom-up methodology, illustrating its integration into dynamic simulation frameworks, and establishing a conceptual bridge with the previous chapter, which focuses on EFM-derived macroscopic modeling. This guiding thread is illustrated in Fig. 4.16.

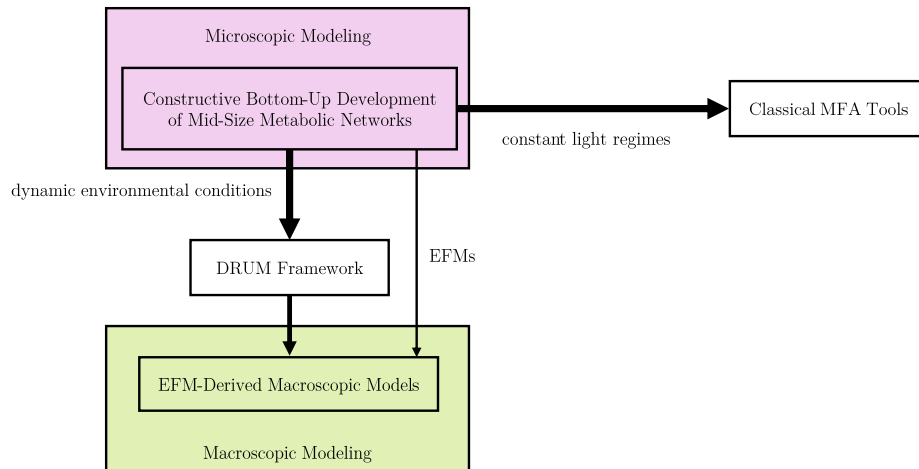


Figure 4.16: Integration of the bottom-up strategy into modeling frameworks. The first case study allows microscopic validation under constant light regimes via the classical MFA tools, while the second concludes with macroscopic validation using the DRUM framework under dynamic light/dark conditions.

4.4.2.1 Basic Principle

The core idea of the DRUM approach is to partition the metabolic network into sub-networks, each grouping reactions that are assumed to operate under balanced growth conditions. The remaining metabolites, which interconnect these sub-networks, are

treated dynamically. Each sub-network is then reduced to a set of macroscopic reactions, resulting in an overall system of differential equations (ODEs) that describes substrate consumption, biomass production, product excretion, and the accumulation of some internal metabolites. Depending on the size of the original network, the number and complexity of sub-networks, and the dimensionality of their associated elementary flux mode (EFM) spaces, the number of EFMs may become prohibitively large. In such cases, the approach can be combined with the reduction algorithm presented in Sec. 3.2.1.2.

4.4.2.2 No Accumulation of Internal Metabolites

To introduce the DRUM framework, we begin with the classical mass balance formulation in Eq. 3.5. For clarity, we consider a continuous bioprocess scenario in which microorganisms grow in a perfectly mixed, constant volume stirred-tank reactor, operating at a dilution rate D , and are supplied with a set of incoming substrates S_{in} . Under the quasi-steady-state assumption (QSSA) – that is, assuming no net accumulation of internal metabolites – the system can be described by the following set of ordinary differential equations:

$$\frac{d}{dt} \begin{pmatrix} S \\ P \\ X \end{pmatrix} = \begin{bmatrix} K_S \\ K_P \\ K_X \end{bmatrix} \Phi X - D \begin{pmatrix} S \\ P \\ X \end{pmatrix} + D \begin{pmatrix} S_{in} \\ 0 \\ 0 \end{pmatrix} \quad (4.28)$$

$$\frac{dC}{dt} = 0 \quad (4.29)$$

where S and P represent respectively the extracellular substrates and excreted products, X denotes the total biomass and C corresponds to intracellular metabolites governed by the QSSA. These equations are structurally equivalent to Eq. 3.5, but explicitly distinguish between substrates, products, biomass, and internal metabolites to facilitate the introduction of unbalanced growth dynamics in later steps of the DRUM formalism. Recall that the global matrix K represents the stoichiometric matrix of macroscopic reactions and is an $m_e \times n_{EFM}$ matrix, where m_e is the number of external species and n_{EFM} is the number of selected elementary flux modes. Accordingly, the vector Φ contains n_{EFM} macroscopic fluxes – one associated with each EFM.

4.4.2.3 Internal Metabolite Accumulation Issue

When the accumulation of internal metabolites cannot be neglected, the metabolic network is partitioned into sub-networks in which the quasi-steady-state assumption (QSSA) remains applicable. These sub-networks are typically defined based on metabolic functions, co-regulated reaction groups, or compartmental organization within the cell. The remaining metabolites, denoted by A , serve as connectors between the sub-networks and are explicitly allowed to accumulate, thereby introducing dynamics into the system.

Taking into account this accumulation of internal metabolites, the dynamic system is described by the following set of ordinary differential equations:

$$\frac{d}{dt} \begin{pmatrix} S \\ P \\ A \\ B \end{pmatrix} = \begin{bmatrix} K_S^{SN_1} & \dots & K_S^{SN_k} \\ K_P^{SN_1} & \dots & K_P^{SN_k} \\ K_A^{SN_1} & \dots & K_A^{SN_k} \\ K_B^{SN_1} & \dots & K_B^{SN_k} \end{bmatrix} \begin{bmatrix} \Phi^{SN_1} \\ \vdots \\ \Phi^{SN_k} \end{bmatrix} B - D \begin{pmatrix} S \\ P \\ A \\ B \end{pmatrix} + D \begin{pmatrix} S_{in} \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (4.30)$$

$$\frac{dC_j}{dt} = 0 \quad \forall C_j \in C \setminus A \quad (4.31)$$

where S , P , and B represent the extracellular substrates, excreted products, and functional biomass, respectively. The metabolites $A \subset C$ denote the accumulating internal metabolites, while the remaining metabolites $C \setminus A$ are assumed to be at steady-state. Moreover, this formulation is based on the decomposition of the global stoichiometric matrix N into sub-matrices N_{SN_i} , corresponding to each sub-network:

$$N = (N_{SN_1} \dots N_{SN_k}) \quad (4.32)$$

For each sub-network SN_i , the stoichiometric matrix of the macroreactions K^{SN_i} is derived from its stoichiometric matrix N_{SN_i} , the set of relevant external metabolites, and the corresponding matrix of elementary modes computed using EFMTOOL as EFM generator. Specifically, $K^{SN_i} = N_e^{SN_i} E^{SN_i}$, where E^{SN_i} is the matrix of EFMs calculated on the basis of N^{SN_i} . The matrix $N_e^{SN_i}$ represents the stoichiometric matrix of external metabolites.

Furthermore, it is worth noting that the variable B denotes the functional biomass, which differs from conventional total biomass X . Under QSSA, any chemical element from substrate S ends up in either biomass or excreted products P . However, when there is accumulation of internal metabolites, chemical elements from substrate S can also be retained in the accumulating metabolites A . As a result, the total biomass X can be reconstructed by applying an elemental mass balance:

$$X_Z(t) = Z_B B(t) + \sum_A Z_A A(t) \quad (4.33)$$

where $Z \in \{C ; N ; O ; H ; P ; S ; \dots\}$ represents a chemical element, Z_A and Z_B are the number of atoms of the chemical element Z per mole of accumulating metabolites A and functional biomass B , respectively, and $A(t)$ and $B(t)$ stand for the concentration of A and B at time t . The resulting value $X_Z(t)$ corresponds to the concentration of the chemical element Z in total biomass X at time t .

Finally, to fully characterize the model, only the kinetics of the macroscopic fluxes Φ^{SN_i} in Eq. 4.30 must be defined. These fluxes can be modeled using various kinetic formalisms, including mass-action, power-law, Michaelis-Menten, Hill-type, or cybernetic models, depending on the biological context and available data.

4.4.2.4 Accumulating Metabolites vs. Excreted Products

Within the DRUM framework, metabolites classified as A (accumulating) and P (excreted products) serve fundamentally distinct roles in cellular carbon metabolism under non-balanced growth conditions. Accumulating metabolites A , such as storage lipids or carbohydrates, are intracellular compounds that may build up over time. They are explicitly modeled with accumulation terms in the mass balance equations, allowing their intracellular concentrations to vary dynamically. Experimentally, a metabolite is identified as belonging to the A category if it shows a net increase in the intracellular pool (e.g., via targeted metabolomics or GS-MS/HPLC analyses) without immediate export from the cell. In contrast, excreted products P represent compounds synthesized by the cell and directly released into the extracellular medium, such as CO_2 or other metabolic byproducts. These metabolites are modeled without accumulation terms and their fluxes are strictly outward. Detection is typically based on their appearance in the extracellular environment over time.

By distinguishing between internal accumulation and external release, the DRUM framework enables a clear separation between storage metabolites and true excretion

products, offering an accurate representation of carbon partitioning under dynamic growth conditions.

4.4.3 Experimental Validation

Photosynthetic microorganisms such as *Tisochrysis lutea* present unique challenges for metabolic modeling, particularly under dynamic environmental conditions. Unlike heterotrophic or continuously illuminated systems (see Sec. 4.3), phototrophic organisms exposed to diurnal light-dark cycles exhibit transient storage and mobilization of internal metabolites, which are processes that inherently violate the steady-state assumptions (QSSA) underlying classical metabolic flux analysis (MFA). In the previous case study involving *Arthrospira* sp., the assumption of constant light availability allowed us to approximate the system using steady-state conditions and classical MFA techniques. However, in the present case study, experimental data were acquired under alternating day and night conditions, where photosynthetic activity and carbon allocation shift dramatically over time. This dynamic context necessitates a modeling framework capable of accounting for internal metabolite accumulation and turnover, motivating the use of the **DRUM** approach in conjunction with the refined metabolic network developed in Sec. 4.4.1.2.

In this section, we briefly describe the experimental conditions under which the data were collected. Then, practical implementation and parameterization of the DRUM framework are also discussed. Finally, we present simulation results and compare model predictions with experimental measurements, assessing both the consistency of the model with observed dynamics and the added value of our more detailed network construction. It is important to note that in-silico validation was not performed, as the original model had already been experimentally validated in the study conducted in [2].

4.4.3.1 Cell Line and Media

The experimental data used in this case study originate from a continuous culture of *Isochrysis affinis galabana* (clone T-iso, CCAP927/14) [215] cultivated under a day/night cycle, as described in [2]. This microbial strain, recently renamed *Tisochrysis lutea*, is known for its ability to accumulate high levels of lipids. Cultures were grown in duplicate in 5 L cylindrical vessels maintained at a constant temperature of 22°C and a pH of 8.2, regulated through automated CO₂ injection. The following variables were measured throughout the experiments: nitrate concentration, particulate carbon and nitrogen, chlorophyll content, total carbohydrates, and neutral lipid concentrations.

4.4.3.2 Numerical Results

To assess the merits of the extended metabolic network developed through the constructive bottom-up strategy, it is essential to compare its performance against the reference network proposed in [2]. This comparison is structured in three stages. First, a comparison of the macroscopic reactions derived from both networks is performed, using the same partitioning into sub-networks as in the original study. Particular attention is paid to the differences that arise from the more detailed representation of carbohydrate biosynthesis in the extended network. Second, on the basis of the derived macroscopic models, the reaction kinetics are modeled and the kinetic parameters are identified. This enables a quantitative comparison between the two models, both through numerical metrics such as the cost function obtained during parameter estimation, and through graphical assessment of data reproducibility. Finally, a discussion of the results is provided, highlighting the biological relevance of the findings and proposing possible

interpretations regarding carbon partitioning and intracellular storage behavior. This structured comparison aims to validate the extended network while emphasizing the contributions of a more detailed network construction to dynamic metabolic modeling.

4.4.3.2.1 Comparison of Macroscopic Reactions

As part of the DRUM-based macroscopic model construction, the network is partitioned into sub-networks following the same scheme as in the original study conducted in [2]. For consistency, the same subdivision is applied here, with the exception of the sub-network associated with carbohydrate biosynthesis, which required a more detailed treatment due to the elaboration of the underlying pathways. Specifically, the carbohydrate biosynthesis sub-network incorporates the synthesis of necessary precursors, including activated sugars, reflecting the metabolic complexity of *Tisochrysis lutea*. Following the DRUM methodology, elementary flux modes (EFMs) are computed for each sub-network, normalized by unit production of the macromolecule of interest, and principal component analysis is used to verify that a single dominant metabolic behavior is sufficient to represent the flux space. Macroscopic reactions are shown in Table 4.4. As in the original study, the set of accumulating metabolites A contains $\{GAP; G6P; PEP; CARB; PA\}$ (see [2] for more information).

Table 4.4: Macroscopic reactions obtained using the DRUM framework.

Sub-network SN	Macroscopic Reaction
Photosynthesis	$30 \text{ Light} + 3 \text{ CO}_2 + 2 \text{ H}_2\text{O} + \text{P}_i \rightarrow \text{GAP} + 3 \text{ O}_2$
Upper glycolysis	$\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{P}_i + \text{H}^+$ $2 \text{ GAP} + \text{H}_2\text{O} \rightarrow \text{G6P} + \text{P}_i$ $\text{G6P} + \text{ATP} \rightarrow \text{H}^+ + \text{ADP} + 2 \text{ GAP}$
Lower glycolysis	$\text{GAP} + \text{ADP} + \text{P}_i + \text{NAD}^+ \longleftrightarrow \text{PEP} + \text{ATP} + \text{NADH} + \text{H}_2\text{O} + \text{H}^+$
Carbohydrate synthesis	$\text{G6P} \longleftrightarrow \text{CARB} + \text{P}_i$
Carbohydrate synthesis	$0.05 \text{ PEP} + 0.26 \text{ O}_2 + 0.3 \text{ H}^+ + 0.15 \text{ NO}_3^- + 1.17 \text{ G6P} + 1.12 \text{ P}_i \longleftrightarrow \text{CARB} + 0.61 \text{ CO}_2 + 0.5 \text{ H}_2\text{O}$
Lipid synthesis	$\text{GAP} + 16.61 \text{ PEP} + 2 \text{ ADP} + 13.46 \text{ NAD}^+ + 29.3 \text{ NADPH} + 34.48$ $\text{H}^+ + 2.15 \text{ O}_2 \longleftrightarrow \text{PA} + 14.61 \text{ P}_i + 2 \text{ ATP} + 13.46 \text{ NADH} + 29.3$ $\text{NADP}^+ + 4.31 \text{ H}_2\text{O} + 16.61 \text{ CO}_2$
Biomass synthesis	$3.13 \text{ PEP} + 7.37 \text{ O}_2 + 4.46 \text{ H}^+ + 1.31 \text{ NO}_3^- + 1.14 \text{ G6P} + 0.11 \text{ PA} +$ $0.03 \text{ SO}_4^{2-} + 0.0025 \text{ Mg}^{2+} \rightarrow \text{B} + 11.67 \text{ CO}_2 + 4.23 \text{ P}_i + 6 \text{ H}_2\text{O}$

The macroscopic reaction defined in [2] for carbohydrate production is represented in blue in Table 4.4 and is a highly simplified lumped reaction. In contrast, the more detailed network reconstruction led to the macroreaction depicted in yellow. This reaction is biologically relevant for several reasons. First, the minor involvement of phosphoenolpyruvate (PEP) suggests the utilization of high-energy intermediates to support carbohydrate polymerization and precursor synthesis. Second, the consumption of molecular oxygen (O_2) and the release of carbon dioxide (CO_2) reflect partial oxidative metabolism, which is consistent with the known respiratory activity of *Tisochrysis lutea*, especially under day/night cycling conditions. Third, the presence of a small nitrate flux indicates limited nitrogen incorporation, which is consistent with the biosynthesis of nitrogen-containing carbohydrate derivatives such as UDP-N-acetylglucosamine. Furthermore, the production of water (H_2O) is a typical byproduct of polymerization reactions involving the condensation of sugar units. Therefore, the stoichiometry and the nature of the involved metabolites closely align with known biochemical processes

and cellular behaviors of *Tisochrysis lutea*, supporting the biological realism and validity of the reconstructed macroscopic reaction.

4.4.3.2.2 Macroscopic Model Derivation

As mentioned previously, eight macroscopic reactions are obtained on the basis of the six defined sub-networks. To derive the macroscopic model, it is now necessary to define the kinetics associated with these macroscopic reactions. For this purpose, simple proportional kinetics are assumed. According to the formalism introduced in Sec. 4.4.2, Eq. 4.30 can be simplified as follows:

$$\frac{dM}{dt} = \frac{d}{dt} \begin{pmatrix} S \\ A \\ B \end{pmatrix} = K \Phi B - D \begin{pmatrix} S \\ A \\ B \end{pmatrix} + D \begin{pmatrix} S_{in} \\ 0 \\ 0 \end{pmatrix} \quad (4.34)$$

In Eq. 4.34, M is the vector of kept metabolites composed of substrates S (i.e., metabolites transported in the cell: light, CO_2 , O_2 , P_i , H_2O , H^+ , NO_3^- , SO_4^{2-} , Mg^{2+}), metabolites authorized to accumulate A (either at branching points of metabolic pathways such as GAP, G6P, PEP or end products of metabolic pathways – PA and CARB – or energy metabolites like ATP, ADP, NADH, NAD^+ , NADPH, NADP^+), and functional biomass B . In the previous dynamic model, the kinetic vector Φ (of dimension 8×1) represents the specific rates of the macroscopic reactions and must be modeled using appropriate kinetic laws.

As mentioned in the previous chapter in Sec. 3.2.2, the calibration of a macroscopic model requires two successive steps: first, the selection of a kinetic model structure and second, the estimation of the model parameters by minimizing the discrepancy between model predictions and experimental measurements, typically through the minimization of a cost function.

4.4.3.2.2.1 Kinetic Structure Selection As a first validation step, we propose to adopt the same simple kinetic laws as those used in the original study. Each macroscopic reaction rate is modeled using proportional or first-order kinetics, following the original assumptions. The complete set of kinetic expressions is detailed in Table 4.5. This choice ensures consistency and facilitates a direct comparison with previous results. Moreover, given the relatively limited experimental dataset available, the use of simple kinetic structures is appropriate to avoid overfitting. Although more sophisticated kinetic formulations, such as Michaelis-Menten or cybernetic models, could potentially improve predictive performance, their implementation would require richer experimental data to enable reliable parameter identification.

As discussed in [2], energetic cofactors such as ATP, NADP, NADPH, and FADH_2 are difficult to balance directly within the DRUM framework. Typically, it is achieved by introducing maintenance terms – such as the maintenance reaction (i.e., the second macroscopic reaction in Table 4.4) – calibrated to fit experimental growth rates and substrate consumption data. However, in the present study, critical experimental measurements such as carbon incorporation (e.g., absorbed light per unit of biomass) or dissolved CO_2 concentration were not available. Consequently, estimating maintenance parameters and achieving an accurate energetic balance was not feasible. For these reasons, as in the previous study conducted in [2], we decided not to explicitly model the balance of energetic cofactors and we neglected the description of ATP, ADP, NADPH, NADP^+ , NADH, and NAD^+ dynamics. Practically, this is reflected in the model by setting $\phi_{MR_2} = 0$.

Table 4.5: Kinetic structure of the macroscopic reactions.

Sub-network (SN)	Kinetic Structure
Photosynthesis	$\phi_{MR_1} = k_{MR_1} \cdot I$
Upper glycolysis	$\phi_{MR_2} = 0$ $\phi_{MR_3} = k_{MR_3} \cdot \text{GAP}$ $\phi_{MR_4} = k_{MR_4} \cdot \text{G6P}$
Lower glycolysis	$\phi_{MR_5} = k_{MR_5} \cdot \text{GAP} - k'_{MR_5} \cdot \text{PEP}$
Carbohydrate synthesis	$\phi_{MR_6} = k_{MR_6} \cdot \text{G6P} - k'_{MR_6} \cdot \text{CARB}$
Lipid synthesis	$\phi_{MR_7} = k_{MR_7} \cdot \text{PEP} \cdot \text{GAP} - k'_{MR_7} \cdot \text{PA}$
Biomass synthesis	$\phi_{MR_8} = k_{MR_8} \cdot \text{PEP} \cdot \text{G6P} \cdot \text{NO}_3$

4.4.3.2.2 Model Calibration To estimate the parameters of the model, it is required to minimize a cost function that depends on model predictions (i.e., simulation results) and experimental measurements (see Sec. 2.7). As mentioned in Sec. 4.4.3.1, the available experimental data include light intensity I (expressed in $\mu\text{E.m}^{-2}.\text{s}^{-1}$), particulate carbon and nitrogen concentrations, chlorophyll content, total carbohydrate concentration and neutral lipid concentration. Moreover, the incoming substrate $S_{in} = \text{NO}_3^- = 0.287 \text{ mmol.L}^{-1}$ and the dilution rate $D = 0.040833 \text{ h}^{-1}$. Given that the particulate carbon and nitrogen contents of the biomass were measured experimentally, the model must be able to simulate the total biomass in terms of carbon and nitrogen content. Specifically, the carbon and nitrogen content of the total biomass X are computed as follows:

$$X_C(t) = C_B B(t) + \sum_A C_A A(t) \quad (4.35)$$

$$X_N(t) = N_B B(t) + \sum_A N_A A(t) \quad (4.36)$$

where $A \in \{ \text{GAP} ; \text{G6P} ; \text{PEP} ; \text{CARB} ; \text{PA} \}$ is the set of authorized accumulating metabolites; C_A and C_B represent the number of carbon atoms per molecule of metabolite A and functional biomass B, respectively; N_A and N_B represent the number of nitrogen atoms per molecule of metabolite A and functional biomass B, respectively; $A(t)$ and $B(t)$ denote the concentrations of accumulating metabolites and functional biomass at time t ; and $X_C(t)$ and $X_N(t)$ denote the total carbon and nitrogen concentration in the total biomass at time t .

It is important to note that energetic cofactors are not considered in the balance equations, as their contribution to the total carbon and nitrogen pools is assumed to be negligible compared to the contributions from functional biomass and the accumulating molecules (CARB, PA, PEP, G6P, GAP).

Finally, while the original study defined the cost function as the sum of squared errors between simulated and experimental measurements (using the average of experimental duplicates), we propose here to use a cost function based on squared relative errors. The cost function is expressed as follows:

$$J = \sum_{\xi} \sum_k \left(\frac{\xi_{meas_k} - \xi_{sim_k}}{\xi_{meas_k}} \right)^2 \quad ; \quad \xi \in \{ \text{CARB} ; \text{PA} ; X_C ; X_N \} \quad (4.37)$$

where ξ_{meas_k} is the measured variable at the k -th time instant and ξ_{sim_k} is the corresponding simulated value. This formulation corresponds to the assumption of constant relative measurement error and, therefore, gives equal weight to deviations expressed as a percentage of the measured value. Although other normalizations are also possible (e.g., division by maximum values, standard deviations, or experimental variances), the approach chosen here provides a simple and effective way to mitigate scale effects in calibration. To minimize the cost function, the Nelder-Mead algorithm (*fminsearch* in MATLAB) was used and several optimizations were performed with random initial parameters set to prevent the risk of finding local minima (multistart strategy). In addition, since very few data were available, all data were used to estimate the model parameters. To ensure consistency with the original work, the initial conditions and the structure of the macroscopic model were aligned with those described in the existing study. Although the simulation framework was originally available under SCILAB, the model was implemented in MATLAB for this study. All parameter identification and model calibration procedures were developed independently. For a detailed description of the initialization and structure of the original model, the reader is referred to [2]. The values of the parameters for the different models are shown in Table 4.6.

Table 4.6: Parameters for the original model of Baroukh et al., the original model recalibrated using the normalized cost function, and for the newly constructed model.

	Unit	Original model	Recalibration	New model
k_{MR_1}	$\mu E^{-1} . m^2 . s . m M . h^{-1} . m M B^{-1}$	0.0011	0.0013	0.0013
k_{MR_3}	$h^{-1} . m M B^{-1}$	223.53	200.95	213.12
k_{MR_4}	$h^{-1} . m M B^{-1}$	10.30	13.05	14.37
k_{MR_5}	$h^{-1} . m M B^{-1}$	436.95	385.20	351.72
k'_{MR_5}	$h^{-1} . m M B^{-1}$	5.00	6.40	5.95
k_{MR_6}	$h^{-1} . m M B^{-1}$	70.00	56.08	50.26
k'_{MR_6}	$h^{-1} . m M B^{-1}$	6.50	4.12	4.10
k_{MR_7}	$m M^{-1} . h^{-1} . m M B^{-1}$	$4.50 \cdot 10^3$	$2.59 \cdot 10^3$	$2.63 \cdot 10^3$
k'_{MR_7}	$h^{-1} . m M B^{-1}$	0.60	0.52	0.49
k_{MR_8}	$m M^{-2} . h^{-1} . m M B^{-1}$	$2.18 \cdot 10^4$	$1.95 \cdot 10^4$	$1.89 \cdot 10^4$
J		72.5117	47.2158	44.3625

In the following, three models will be compared systematically: (i) the original macroscopic model of Baroukh et al. calibrated using the classical (non-normalized) cost function, (ii) the same original model re-calibrated using the normalized cost function (representing a first level of methodological improvement) and (iii) the newly constructed macroscopic model, calibrated using the normalized cost function, reflecting both the extended network and the updated calibration approach.

First, the models are compared on the basis of the cost function value. The cost function value of the original model using the classical definition of the cost function (non-normalized) was estimated based on the first iteration of the recalibration process, where the initial guess corresponded exactly to the set of parameters originally defined in [2]. Although this provides only an approximate evaluation, it ensures that the comparison remains meaningful and consistent with the methodology applied throughout this work. As shown in Table 4.6, the results show a clear progression. The cost function associated with the original macroscopic model is the highest ($J \approx 73$), followed by the re-calibrated version of the same model using the normalized cost function (J

≈ 47), and finally the newly constructed macroscopic model, which achieves the lowest cost function value ($J \approx 44$). These results highlight the benefits of using a properly normalized cost function during calibration and demonstrate the merits of extending the metabolic network to improve both the model structure and the data reproducibility.

Beyond the numerical comparison of cost function values, it is equally important to visually assess the ability of each model to reproduce the experimental dynamics of key biological variables. The simulation results are given in Fig. 4.17 to Fig. 4.21 and it can be noticed that all three models capture the general dynamics of *T. lutea* under a day/night cycle, but with clear differences in accuracy and biological realism.

4.4.3.2.3 Discussion

To comprehensively evaluate the modeling approaches, a multi-level comparison is conducted. First, a global comparison between the three models is performed to assess the overall performance. Second, a focused comparison between the re-calibrated original model and the newly constructed model is proposed to highlight the added value of the bottom-up network reconstruction.

4.4.3.2.3.1 Global Comparison At a global level, the three models – the original model (MODEL 1), the re-calibrated original model (MODEL 2, with a normalized cost function), and the new model (MODEL 3, obtained from the extended metabolic network) – all succeed in reproducing the main dynamic trends of *Tisochrysis lutea* under day/night cycling. However, notable differences emerge regarding their predictive accuracy and biological plausibility. The re-calibrated original model significantly improves the numerical fit to experimental data compared to the original model, highlighting the importance of a correct cost function definition. The new model, while yielding only a slightly better cost function value than the re-calibrated model, demonstrates a more accurate simulation of carbon storage dynamics. In particular, MODEL 1 underestimates the accumulation of carbohydrates compared to experimental observations and tends to overpredict the formation of lipids (PA). Consequently, the predicted carbohydrate quota increases less than measured, while the lipid quota is overestimated, as illustrated in Fig. 4.17. MODEL 2, after recalibration, corrects part of this deviation, improving the simulation of carbohydrate accumulation and limiting lipid overproduction. Finally, the newly constructed MODEL 3 offers a further improvement: it closely matches the observed carbohydrate buildup and predicts a more moderate accumulation of lipids, thus providing a more accurate representation of the cellular storage dynamics.

4.4.3.2.3.2 Focused Comparison Focusing on the comparison between the re-calibrated original model and the new model isolates the specific impact of updating the macroscopic reaction for carbohydrate synthesis. Although both models achieve comparable fits to biomass and storage pools (see Fig. 4.17 to Fig. 4.20), the new model explicitly integrates energetic and redox costs associated with carbohydrate polymerization, resulting in a more realistic representation of carbon allocation. As shown in Table 4.4, carbohydrate production in MODEL 3 is coupled with the consumption of a small amount of phosphoenolpyruvate (PEP) and oxygen, and the release of carbon dioxide and water, introducing an energetic cost associated with carbon storage. This stoichiometric modification influences the intracellular dynamics during the day phase. In particular, as shown in Fig. 4.17, MODEL 3 shows a higher accumulation of the intermediate metabolite G6P during early light exposure compared to MODEL 2, reflecting the slower, energetically restricted pathway toward carbohydrate storage. Meanwhile, PEP levels in MODEL 3 are slightly lower than in MODEL 2, since PEP is now

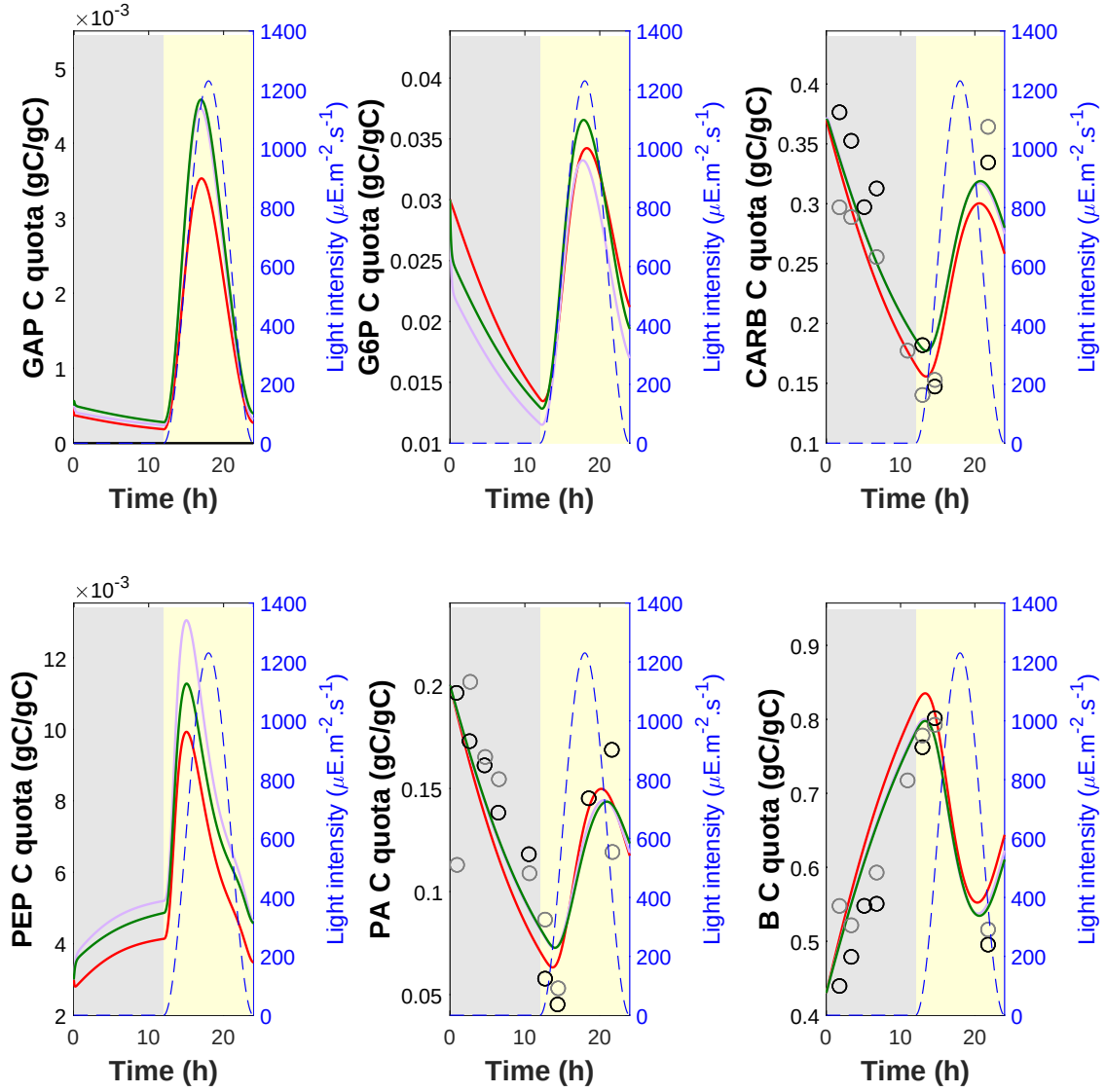


Figure 4.17: Evolution of intracellular carbon quotas for GAP, G6P, PEP, carbohydrates (CARB), PA (storage lipids), and functional biomass B throughout a day/night cycle. Comparison between the three models (original model from [2] in red, the re-calibrated model in purple, and the newly constructed model in green). Light intensity is shown to distinguish day and night phases.

consumed as a substrate to produce carbohydrates. Despite these internal differences, both models ultimately reproduce similar trends in total carbohydrate accumulation. The main improvements of MODEL 3 are thus visible mainly at the level of intermediate metabolite dynamics (G6P, PEP) rather than in terms of the final storage pool. Moreover, by coupling carbohydrate synthesis to oxygen consumption and CO_2 release, the new model introduces a built-in regulation mechanism for carbon overflow, providing biologically meaningful insights into how *Tisochrysis lutea* manages excess carbon under nutrient limitation.

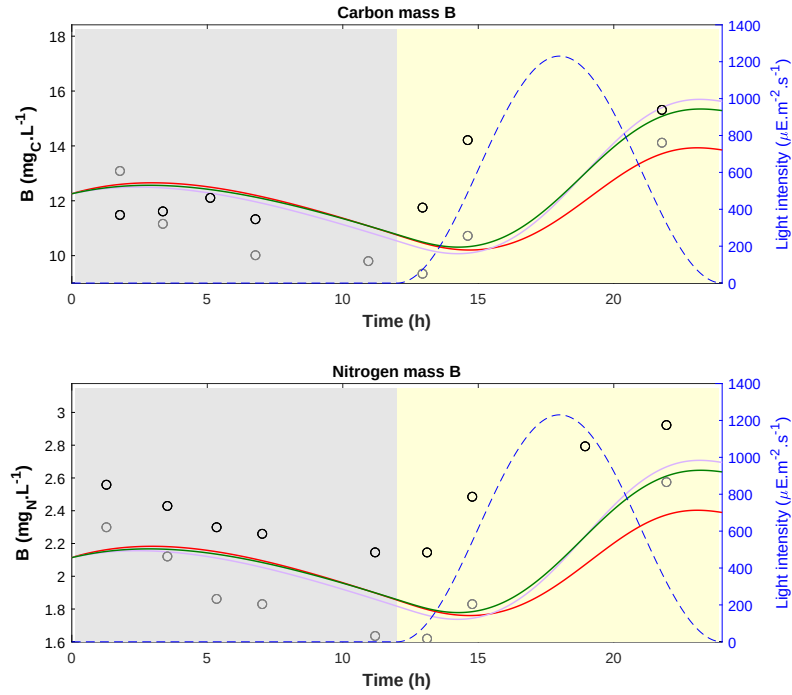


Figure 4.18: Dynamics of the functional biomass B over time during a day/night cycle. Predictions from the original model (red curves), the re-calibrated model (purple curves), and the new model (dark green curves). Experimental data are represented by black and gray bullets.

Under fluctuating environmental conditions such as light/dark cycles, microalgae often prioritize the accumulation of storage reserves (such as carbohydrates and lipids) over structural biomass production. This physiological behavior, known as carbon overflow or carbon partitioning toward storage, enables cells to buffer against periods of unfavorable growth conditions by stocking excess energy. In MODEL 3, this strategy is reflected by a slightly lower proportion of functional biomass related to total biomass, compared to the original model structure in MODEL 2 (see Fig. 4.17 to Fig. 4.19). The improved allocation toward energy reserves (CARB and PA) is thus consistent with known metabolic strategies in phototrophic microorganisms.

Finally, when examining the dynamic rates of the key macroscopic reactions in Fig. 4.21, both the re-calibrated original model (MODEL 2) and the new model (MODEL 3) preserve similar overall trends for primary processes such as photosynthesis and biomass synthesis. However, the rate of carbohydrate accumulation ϕ_{MR_6} shows a small distinction: MODEL 3 exhibits a more gradual and attenuated accumulation compared to MODEL 2. This behavior is consistent with the newly introduced stoichiometric coupling between carbohydrate synthesis, oxygen consumption, and CO_2 release, which imposes an energetic cost on carbon storage. As a result, carbohydrate storage dynamics in MODEL 3 is more progressive, reflecting a biologically plausible regulation of reserve synthesis rather than immediate and unrestricted accumulation.

In summary, the new model represents an improvement over the original formulation, both in terms of reproducing experimental observations and aligning with the mecha-

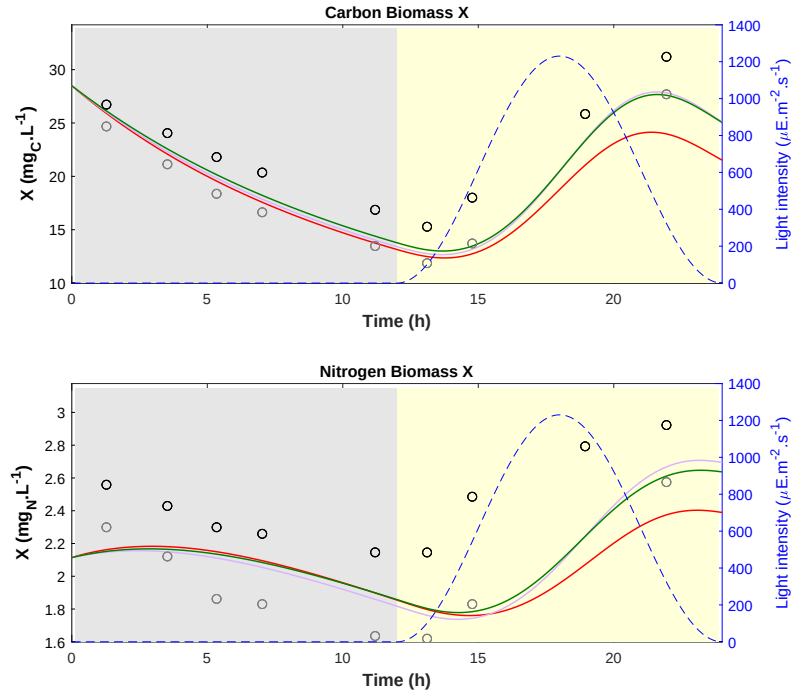


Figure 4.19: Evolution of the total biomass X expressed both as carbon content (mg_C/L) and nitrogen content (mg_N/L) during a day/night cycle. Prediction of the three models against experimental data.

nistic understanding of microalgal carbon metabolism. It addresses prior limitations by normalizing the calibration process — ensuring that all state variables are fitted without bias — and by refining the storage reaction to account for the real metabolic costs associated with carbohydrate biosynthesis. These modifications yield a model that not only better captures growth and reserve accumulation but also provides a framework more consistent with how microalgae dynamically balance carbon and energy under fluctuating environmental conditions.

Importantly, recent studies [221, 222] have emphasized carbon overflow mechanisms (through partial respiration or excretion) as key elements of microalgal acclimatization to environmental stresses, particularly nitrogen limitation, and the new model incorporates this concept more explicitly than the original version. Although the model proposed in [2] constitutes a pioneering foundation for DRUM-based modeling under non-balanced growth, the present work builds upon it by (slightly) enhancing both numerical fidelity and biological plausibility.

Finally, this second case study serves as a validation of the constructive bottom-up network reconstruction approach proposed in this thesis. Starting from a basic network architecture and carefully incorporating key pathways, the approach led to improved model structure, parameterization, and predictive power. This demonstrates that expansion of metabolic networks (even at intermediate scales) can enhance the dynamic modeling of complex biological systems such as photosynthetic microorganisms.

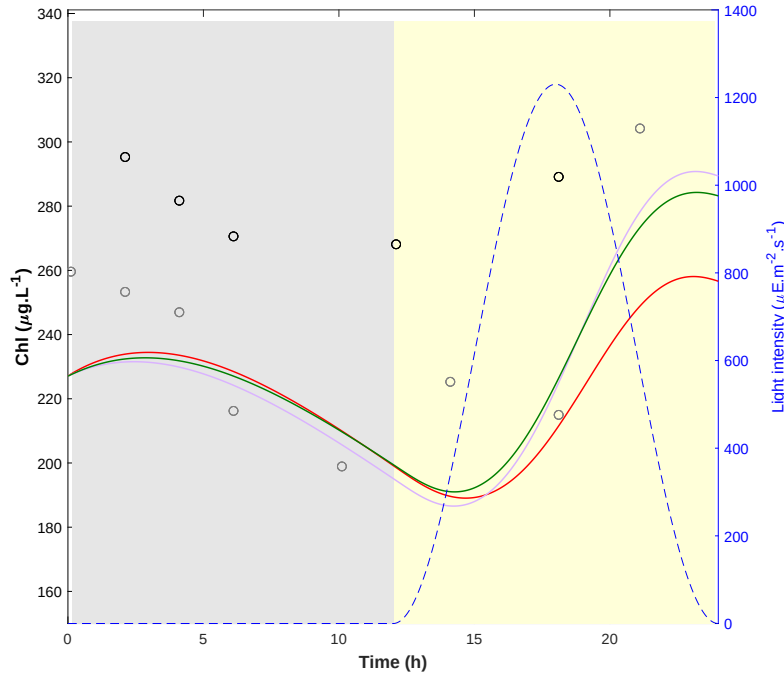


Figure 4.20: Dynamics of chlorophyll content ($\mu\text{g/L}$) over time during a day/night cycle. Predictions from the original model (red curve), the re-calibrated model (purple curve), and the new model (dark green curve).

4.5 Prospects

The present work has explored and validated the bottom-up strategy for constructing mid-sized metabolic networks through two complementary case studies. First, a metabolic network was developed from scratch for *Arthrospira* sp. PCC 8005, followed by an experimental validation under conditions where the pseudo-steady-state assumption (QSSA) remained applicable. Second, an existing network for unicellular microalgae was extended and dynamically validated under non-steady-state conditions by integrating the model into the DRUM framework. These two applications highlight the potential of the constructive modeling approach based on basic biological understanding, minimal genomic data, and constraint-based methods to generate biologically relevant and efficient models. Building on these findings, several perspectives for improvement and further developments can be proposed.

4.5.1 Methodological Improvements

The bottom-up strategy proposed in this thesis was designed to construct mid-sized metabolic networks through a stepwise and biologically intuitive approach. A key strength of the methodology is its minimal reliance on genomic data. Using basic biological knowledge, generic metabolic principles, and constraint-based modeling, it offers a viable pathway to model organisms even in the absence of complete genome-scale models. Furthermore, the constructive process inherently fosters a deeper understanding of the organism's metabolism, which is particularly valuable for engineers and non-biologists

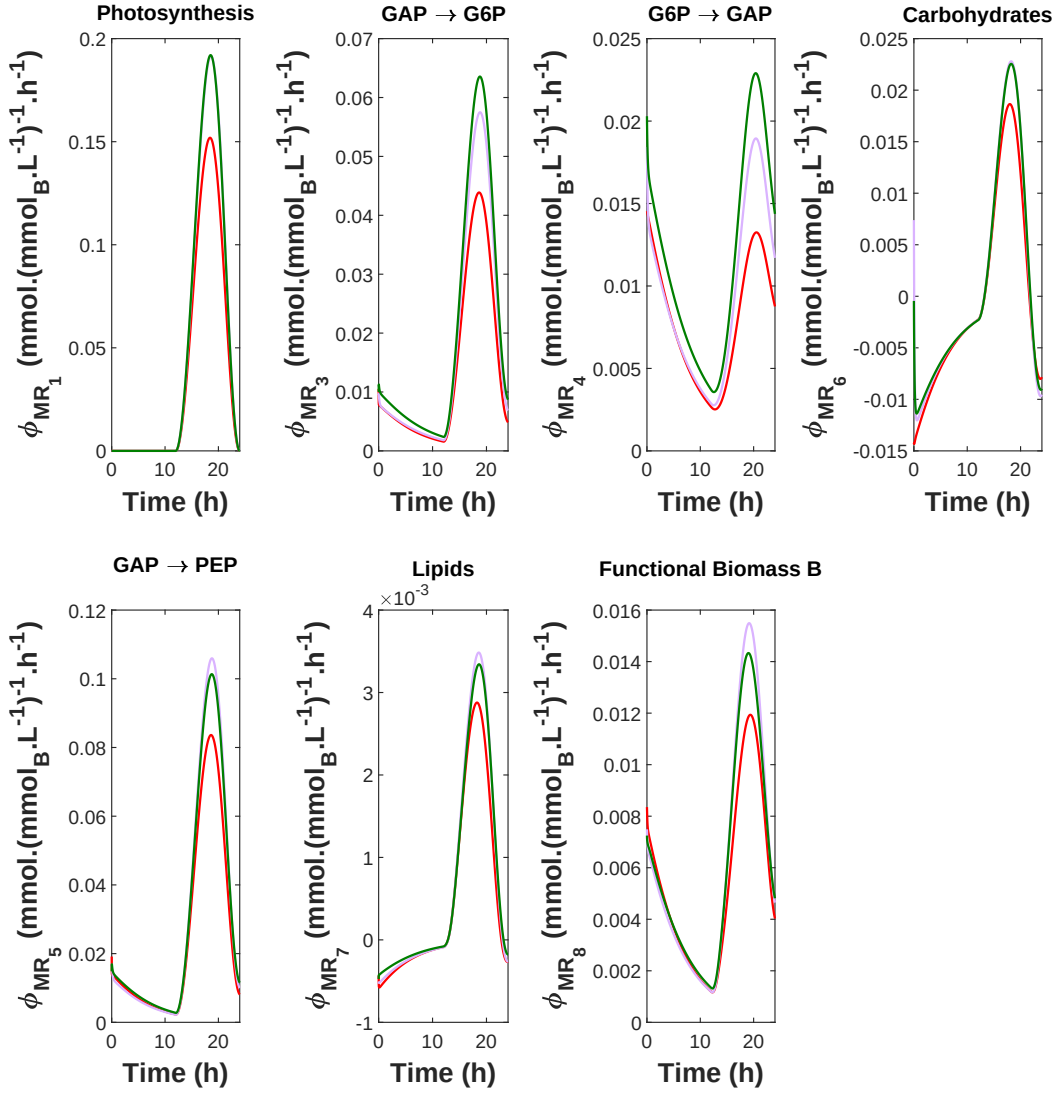


Figure 4.21: Profiles of the macroscopic reaction rates over time for the photosynthesis, carbon allocation, lipid synthesis, biomass formation, and related processes. Predictions from the original model (red curves), the re-calibrated model (purple curves), and the new model (dark green curves).

approaching biological systems modeling. However, several avenues for methodological improvements can be identified.

First, although the approach is well suited for cases where little omics information is available, it would be valuable to benchmark its performance against existing reconstruction pipelines that typically rely on genome-scale information. In particular, for well-documented model organisms, it would be insightful to compare networks obtained using the current bottom-up strategy to those derived from semi-automated genome-based methods (e.g., CARVEME, MODELSEED, METADRAFT) [150, 153, 223]. Such comparisons could highlight the respective strengths, weaknesses, and applicability domains of each approach.

Second, the iterative, trial-and-error nature of the current methodology, while educational, could be complemented by systematic refinement procedures. For instance, constraint-based gap-filling algorithms, which are usually employed to improve genome-scale models, could be adapted to mid-sized networks developed manually. This would help identify missing links in the metabolic network and systematically suggest plausible reaction additions.

Third, a key limitation remains the availability of experimental data to constrain and validate the models. Although measuring extracellular concentrations is common, more detailed experimental datasets, such as ^{13}C -labeled flux measurements (fluxomics), would provide critical insights into intracellular metabolic fluxes. Incorporating such data would help to reduce the under-determination inherent to metabolic networks and improve the predictive power and robustness of the resulting models.

Finally, particular attention should be paid to the modeling of cellular energy metabolism. In the present work, an alternative representation based on the proton motive force (PMF) was proposed instead of classical photophosphorylation and oxidative phosphorylation formulations. However, energy metabolism remains challenging to model accurately, primarily because of the scarcity of direct experimental data. Typically, maintenance terms are empirically tuned to match observed growth rates and substrate consumption. Yet, in the present studies, crucial experimental measurements such as absorbed light per biomass unit and dissolved CO_2 concentration were not always available, making an accurate calibration of maintenance energetics difficult. A major improvement would thus be the integration of direct measurements related to energy fluxes, enabling a more mechanistic and biologically grounded modeling of energy production and consumption processes.

In summary, while the proposed strategy is particularly useful for organisms with limited omics data and offers an accessible framework for model construction and biological understanding, its future refinement could involve integration with systematic gap-filling methods, exploitation of richer experimental datasets, and a deeper mechanistic representation of cellular energetics.

4.5.2 Perspectives for *Arthrospira Sp. PCC 8005*

In addition to the general methodological improvements discussed in Sec. 4.5.1, specific developments can be considered for the case of *Arthrospira* sp. PCC 8005. In particular, extending the current metabolic network to explicitly account for the biosynthesis of **phycobiliproteins**, and especially phycocyanin, represents a promising direction.

Phycobiliproteins are essential water-soluble pigments that act as light-harvesting antennas in cyanobacteria, transferring absorbed energy to chlorophyll during photosynthesis. Beyond their biological function, phycocyanins have gained significant industrial interest due to their strong fluorescence, antioxidant properties, and vivid coloration. They are widely used as natural colorants in the food and cosmetic industries, fluorescent markers in biomedical applications, and potential therapeutic agents [224]. Modeling their production would thus not only refine the biological accuracy of the metabolic network but also support the biotechnological valorization of *Arthrospira*. Moreover, experimental measurements of phycobiliproteins were available in the datasets analyzed in this study (see Sec. 4.3.3.1), making this integration both feasible and scientifically meaningful.

Despite their importance, reactions associated with phycocyanin biosynthesis are typically absent even from existing genome-scale reconstructions of diverse strains of *Arthrospira*. To address this gap, a target strategy was developed. Phycocyanin was considered to comprise two distinct components: (i) the **prosthetic group** (phycocyanobilin) and (ii) the **protein backbone** (polypeptide chains). The synthesis of the prosthetic group could be modeled based on existing biochemical knowledge. Indeed, it is well known that phycocyanobilin is formed from protoporphyrin-IX, an intermediate metabolite already present in the current network (via reaction v_{167} in Table G.1 or v_{144} in Table H.1). The relevant biosynthetic steps can be identified from published biochemical studies. The corresponding chemical reactions for the synthesis of the prosthetic group are derived from the study conducted in [202] and are summarized in Table 4.7 after the application of lumping techniques.

Table 4.7: Additional reactions for the synthesis of phycocyanobilin.

Flux	Metabolic reaction
v_1^*	$\text{Fe}^{2+} + \text{protoporphyrin} \rightarrow 2 \text{H}^+ + \text{protoheme}$
v_2^*	$\text{protoheme} + 3 \text{NADH} + 3 \text{O}_2 \rightarrow \text{biliverdin-}\alpha + \text{Fe}^{2+} + \text{CO} + 3 \text{NAD}^+ + 3 \text{H}_2\text{O}$
v_3^*	$\text{biliverdin-}\alpha + \text{Fe}_{red} \rightarrow \text{phycocyanobilin} + \text{Fe}_{ox}$

For the polypeptide chains, a deep analysis of the amino acid sequence was performed using the **UNIPROT** database, allowing the determination of the amino acid composition of the phycocyanin α - and β -subunits. To account for the energetic cost of polymerization, a classical biochemical assumption was made: approximately 2 ATP molecules are consumed per peptide bond formation during protein synthesis. This led to a full stoichiometric description of the assembly of phycocyanin from its metabolic precursors. The corresponding metabolic reactions are given in Table 4.8.

Table 4.8: Additional reactions for the synthesis of the polypeptide chains and the formation of phycocyanin.

Flux	Metabolic reaction
v_4^*	$24 \text{ alanine} + 2 \text{ cysteine} + 9 \text{ aspartate} + 8 \text{ glutamate} + 5 \text{ phenylalanine} + 13 \text{ glycine} + 1 \text{ histidine} + 11 \text{ isoleucine} + 9 \text{ lysine} + 13 \text{ leucine} + 4 \text{ methionine} + 6 \text{ asparagine} + 5 \text{ proline} + 12 \text{ serine} + 10 \text{ threonine} + 4 \text{ valine} + 11 \text{ tyrosine} + 7 \text{ arginine} + 1 \text{ tryptophan} + 7 \text{ glutamine} + 324 \text{ ATP} + 162 \text{ GTP} \rightarrow \text{CpcA-cys-84}$
v_5^*	$29 \text{ alanine} + 4 \text{ cysteine} + 10 \text{ aspartate} + 8 \text{ glutamate} + 5 \text{ phenylalanine} + 11 \text{ glycine} + 9 \text{ isoleucine} + 4 \text{ lysine} + 14 \text{ leucine} + 6 \text{ methionine} + 7 \text{ asparagine} + 5 \text{ proline} + 16 \text{ serine} + 10 \text{ threonine} + 12 \text{ valine} + 5 \text{ tyrosine} + 11 \text{ arginine} + 1 \text{ tryptophan} + 5 \text{ glutamine} + 344 \text{ ATP} + 172 \text{ GTP} \rightarrow \text{CpcB-cys-155}$
v_6^*	$\text{CpcA-cys-84} + \text{phycocyanobilin} \rightarrow \text{CpcA-PCB}$
v_7^*	$\text{CpcB-cys-155} + \text{phycocyanobilin} \rightarrow \text{CpcB-PCB}$
v_8^*	$\text{CpcA-PCB} + \text{CpcB-PCB} \rightarrow (\alpha\beta)$
v_9^*	$6 (\alpha\beta) \rightarrow \text{phycocyanin}$

With these elements now characterized, it remains to integrate the corresponding reactions into the existing metabolic network, calibrate the model using the available experimental measurements of phycobiliproteins (expressed in g.L^{-1}), and evaluate the predictive performance. If successful, this extension could significantly enhance the biological realism of the model and support future applications focused on optimizing

phycocyanin production.

This reconstruction approach demonstrates the flexibility and robustness of the bottom-up strategy proposed in this work, allowing the incorporation of complex biosynthetic routes even when standard genome-scale resources are unavailable. By carefully combining known metabolic intermediates, biochemical knowledge, experimental amino acid compositions, and realistic energy requirements for peptide bond formation, the developed extension preserves both biological accuracy and modeling tractability.

4.5.3 Perspectives for *Tisochrysis Lutea*

In addition to the general methodological improvements discussed previously, several specific avenues can be envisioned to further refine the modeling of *Tisochrysis lutea*'s carbon metabolism. First, exploring alternative kinetic laws for macroscopic reactions could enhance model flexibility and biological realism, although this will critically depend on the availability of richer experimental datasets.

Beyond carbohydrate storage, extending the description of lipid metabolism is a promising direction. In the current model, lipid reserves are represented by a phosphatic acid (PA) pool. However, microalgal lipid metabolism is more complex, encompassing neutral lipids such as triacylglycerols (TAGs) – the main energy storage form – as well as membrane lipids such as glycolipids and phospholipids [225]. A more detailed partition of lipid species, supported by experimental data, if available, could better capture physiological shifts between membrane synthesis and energy storage under nutrient stress. In particular, explicitly modeling TAG accumulation would align the model with established microalgal responses to environmental fluctuations and could improve predictions of carbon allocation dynamics under varying growth conditions.

Another perspective concerns the application of the present network extension to more recent models. In particular, the approach developed here to expand the carbohydrate synthesis module in the study of [2] could be applied in the same way to the updated network proposed in [222], where additional data are available. Since the metabolic network proposed in [222] still represents carbohydrate biosynthesis by a simplified lumped reaction, the extension procedure developed in Sec. 4.4.1.2 can be applied directly.

4.6 Conclusions

This work proposes a constructive methodology for the identification and development of metabolic networks of intermediate size. It combines basic biological knowledge with a series of constraint-based methods in an iterative procedure, allowing for continuous refinement of the network definition. Two case studies, supported by experimental validation, were conducted to test the methodology, and the results obtained were both consistent and biologically relevant.

To the best of our knowledge, there is currently no comprehensive step-by-step methodology to guide the construction of metabolic networks from scratch. On the one hand, genome-scale reconstructions are often available, but remain extremely large, difficult to analyze, and heavily reliant on genome annotations. On the other hand, small-scale networks may lack sufficient biological realism. A middle ground is therefore the development of mid-sized networks sufficiently detailed to capture the complexity of cellular metabolism while remaining tractable for dynamic process-level modeling and

analysis.

The approach developed in this study follows a bottom-up and constructive logic. First, it identifies the key core metabolic pathways, i.e., glycolysis, the TCA cycle, pentose phosphate pathway, etc., that are common across organisms. Particular attention is paid to the directionality of the reactions, based on thermodynamic information (e.g., $\Delta G'^{\circ}$ values) and enzymatic properties gathered from resources such as KeGG and genome studies when available. The resulting set of reactions is used to build the stoichiometric matrix N of the metabolic network, allowing the application of constraint-based tools such as Flux Balance Analysis (FBA), Flux Variability Analysis (FVA), and Flux Coupling Analysis (FCA). These tools help ensure network consistency, identify blocked reactions, and iteratively refine the network. If genome sequences are available, specific pathways or organism-specific features can be integrated into the network, while applying lumping strategies when necessary to manage complexity. Importantly, this methodology also emphasizes the modeling of cellular energy metabolism (here through the proton motive force), providing a more mechanistic description of energy transduction. An important strength of the approach is its minimal reliance on genomic information: genomic data can enhance the model if available but are not strictly required.

Two experimental case studies were conducted to illustrate the methodology. **First**, a network for *Arthrospira* sp. PCC 8005 was developed entirely from scratch. The model was validated both in-silico under steady-state assumptions and experimentally under constant light regimes, where the pseudo-steady-state assumption (QSSA) remains valid, allowing the application of classical tools of metabolic flux analysis. Experimental validations also investigated the impact of different nitrogen sources on metabolism. **Second**, the unicellular microalga *Tisochrysis lutea* was modeled by extending an existing metabolic network. The network was extended to better describe carbohydrate storage pathways and was experimentally validated under day/night cycling, a condition where QSSA no longer holds. Here, the DRUM (Dynamic Reduction of Unbalanced Metabolism) framework was used to accommodate non-steady-state growth phases. In both case studies, the proposed methodology showed its capacity to deliver biologically plausible results and dynamically validated metabolic models, supporting its general applicability.

Despite its strengths, the methodology does have limitations. First, the iterative nature of the process is partly dependent on the expertise and intuition of the user. Moreover, the availability of experimental data remains a major bottleneck: while external metabolite fluxes are often measurable, internal fluxes (e.g., through ^{13}C -labeling experiments) are rare and would greatly improve model accuracy and validation. Another challenge lies in properly modeling cellular energy metabolism (here approached via the PMF formalism) where critical experimental data such as light absorption per biomass or dissolved CO_2 concentrations are required.

Beyond the case studies treated here, the methodology has already been successfully extended to other organisms, although experimental validations are still pending. These case studies are mentioned below:

- The modeling of photosynthetic eukaryotic microorganisms;

In this case, specific care must be taken regarding the existence of intracellular compartments, impacting the definition and modeling of the proton motive force and associated transport reactions. Furthermore, anabolic pathways, especially for carbohydrate and lipid biosynthesis, can differ significantly depending on the

type of organism and must be specifically considered (see [226]).

- The modeling of purple non-sulfur bacteria in *Rhodospirillum rubrum*.

The latter are photosynthetic prokaryotes for which diverse organic carbon sources (acetate, propionate, butyrate, valerate, and hexanoate) have been considered to produce polyhydroxyalkanoates and biohydrogen. The metabolism of these photosynthetic organisms is comparable to the one discussed in Sec. 4.3 but additional pathways need to be taken into account, such as ethylmalonyl-CoA and methylbutanoyl-CoA pathways, degradation pathways, β -oxidation pathways and the synthesis of fatty acids, among others. These pathways involve 59 other metabolites and 67 additional reactions in the metabolic network. The additional metabolic reactions specific to the metabolism of purple non-sulfur bacteria are detailed in the Appendix K. This application is particularly noteworthy, as it highlights the strength of the proposed methodology, which remains effective even in the absence of genomic data.

In summary, this work demonstrates that a structured, biologically informed bottom-up approach can efficiently provide mid-sized metabolic networks, even in the absence of full genome-scale information. The methodology is flexible, adaptable to a wide range of organisms, and capable of integrating dynamic experimental validation. The two case studies support the idea that such networks, i.e., detailed enough to capture key metabolic characteristics, yet simple enough for dynamic simulation and analysis, represent a promising direction for microbial and bioprocess modeling. The constructive approach developed here provides a practical and scalable solution that bridges the gap between small oversimplified models and massive genome-scale reconstructions, offering a powerful tool for systems biology and metabolic engineering.

Chapter 5

Microscopic Modeling With Genetic Insights

Understanding and optimizing cell culture processes is a fundamental objective in biotechnology, especially for applications ranging from biopharmaceutical production to metabolic engineering. As previously elaborated in this thesis, modeling these complex systems can be approached through a multi-scale strategy: from a **macroscopic** level, where global process variables are considered, to a **microscopic** level, where intracellular metabolism is resolved with greater biochemical detail. In the previous chapters, we established two core frameworks: (i) a method for deriving macroscopic models based on elementary flux modes (EFMs), and (ii) a bottom-up approach for constructing mid-sized mechanistically informed microscopic metabolic networks. These foundational models offer robust insights into intracellular flux distributions under various conditions.

However, such models often overlook one critical dimension of cellular behavior: the genetic regulation of enzymatic activity, which fundamentally determines the capacity of the cell to maintain specific metabolic functions. Although classical constraint-based metabolic models, such as flux balance analysis (FBA) [117], assume a quasi-steady-state for metabolic fluxes constrained by stoichiometry and possibly thermodynamics, they rarely incorporate detailed molecular-level constraints such as enzyme availability, a feature inherently tied to gene expression. Bridging this gap between metabolism and gene regulation is essential for capturing the full cellular response, particularly when synthetic biology interventions introduce non-native elements that perturb the natural regulatory hierarchy.

In this chapter, we propose to extend the microscopic modeling framework by incorporating genetic insights, specifically by considering the gene-protein-reaction relationship that connects gene expression levels to enzyme abundance and consequently to reaction fluxes. The underlying rationale is to impose constraints on intracellular fluxes not only from stoichiometric and thermodynamic principles but also from proteome allocation and gene expression dynamics. This approach aligns with recent advances in enzyme-constrained models (ecModels) [188], where enzyme kinetics and proteomic limitations are introduced to better capture the biophysical reality of cellular systems.

The context in which this integration becomes particularly impactful is that of **synthetic biology**, a field that seeks to design and construct novel biological parts and systems. Synthetic biology often involves the expression of exogenous genes, which impose a metabolic burden on host cells, typically manifesting as a reduction in growth

rate or altered flux distribution [227]. In such scenarios, modeling the coupling between metabolic fluxes and gene expression becomes indispensable for rational design and optimization.

This chapter proposes a reconciliation framework to integrate gene expression models (particularly host-aware genetic models) with microscopic metabolic networks. The core strategy is to constrain the maximum allowed flux through each metabolic reaction by the availability of its catalyzing enzyme, whose concentration is, in turn, dictated by the transcriptional and translational machinery. This method enables a more coherent coupling between metabolic capacity and proteomic cost, accommodating the dual impact of synthetic constructs on metabolism and growth.

To illustrate this integrative approach, we focus on a simple case study involving *Escherichia coli* engineered for the biosynthesis of p-Coumaric acid, a key intermediate in the production of flavonoids, pharmaceuticals and biodegradable plastics [228]. The synthesis of this metabolite involves the heterologous expression of genes from non-native pathways, thereby serving as an ideal benchmark for evaluating the impact of gene expression burden on metabolic performance.

This project was initiated as part of a collaborative effort with the Synthetic Biology and Biosystems Control Lab of the **Universitat Politècnica de Valencia** (UPV), in Spain, after a four-week research stay abroad. This short-term mobility provided an invaluable opportunity to exchange ideas on synthetic biology modeling and laid the groundwork for the development of the integrative framework proposed in this chapter. Interdisciplinary discussions and complementary expertise across institutions significantly enriched the methodological design and practical orientation of the study.

Chapter Outline

The remainder of this chapter is structured as follows:

- **Literature Review** - This section surveys the state-of-the-art in genetic modeling frameworks, including deterministic and stochastic models of gene expression, as well as methods that integrate metabolic and genetic layers;
- **Theoretical Background and Methods** - This section introduces key theoretical principles in synthetic biology, including design principles for genetic circuits and the trade-offs involved in resource allocation. We also formalize the modeling frameworks used in this chapter: (i) classical constraint-based models for metabolism, and (ii) burden-aware models for gene expression. The methodology section details our reconciliation strategy to couple both frameworks via enzyme constraints;
- **Case Study** - This part describes the implementation of the proposed framework on a realistic synthetic biology application. We detail the engineered metabolic pathway, gene constructs used, and simulation results showing how gene expression impacts metabolic output and cellular growth;
- **Future Perspectives** - This section includes a critical discussion of the current limitations and potential extensions, including integration with dynamic models, hybrid stochastic-deterministic approaches, and multi-omics data incorporation;

- **Conclusions** - This part provides a summary of findings and discusses the added value of integrating genetic constraints into microscopic metabolic models, particularly in the context of synthetic systems.

Contributions

The primary contributions of this chapter are as follows:

- The development of a novel integrative modeling framework that bridges constraint-based metabolic models with burden-aware gene expression models by explicitly constraining intracellular fluxes based on enzyme availability;
- Demonstration of the framework relevance in a synthetic biology context, particularly for analyzing how exogenous gene expression influences host metabolism and growth;
- The improvement of intracellular flux estimation at the metabolic level and the optimization of pathway enzyme expression strength at the genetic level;
- The methodology is applied to a case study on *E. coli* engineered for p-Coumaric acid production, showcasing the practical utility of the model in guiding rational strain design.

By addressing these critical challenges, this work introduces a framework for integrating metabolic and gene expression models, enabling improved intracellular flux prediction and enzyme allocation strategies. This approach serves as a foundation for rational design and optimization in synthetic biology applications. See more information in Sec. 1.2.

List of Publications

Conference Paper as Regular Paper

- Maton, M.; Santos-Navarro, F.N.; Pico, J.; Bogaerts, Ph.; Vande Wouwer, A. Towards Constraint-Based Burden-Aware Models for Metabolic Engineering. *IFAC-PapersOnLine*. 12th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2024: Toronto, Canada, 14-17 July 2024. **2024**, 58 (14), 247 - 252.

5.1 Literature Review

This literature review presents a structured overview of the modeling approaches relevant to this work. First, we review the main modeling frameworks independently: constraint-based models at the metabolic level and gene expression models at the genetic level. Then, we examine extensions of these frameworks, focusing on models that incorporate regulatory mechanisms or enzyme capacity constraints, the latter generally doing so without explicitly modeling the underlying gene expression dynamics. Finally, we review recent strategies that explicitly connect gene expression dynamics and metabolic modeling, particularly in the context of synthetic biology.

5.1.1 Modeling Frameworks

Microbial cell factory development through metabolic engineering seeks to obtain high levels of products of interest through genetic modification of microorganisms. Although microbial cell factories are a promising solution to achieve this objective, traditional approaches in metabolic engineering based on extensive combinatorial trial-and-error experiments are expensive, time-consuming, and undirected. The design of complex biological systems with novel functions, done rationally and systematically, aims to eliminate this bottleneck by introducing principles and tools elucidated by engineering, including mathematical modeling and computational optimization.

With the advances in sequencing techniques and genome annotation methods, several mathematical modeling approaches have been developed to analyze the integrated behavior of microbial cells on a genome scale [229, 230]. In such integrative analysis, biological data from genome to metabolome are combined and the prediction of cellular responses to different kinds of genetic and environmental perturbations is evaluated.

At the metabolic level, stoichiometric modeling, also known as **constraint-based modeling**, is now routinely applied. This class of methods relies on mass balances over intracellular metabolites and the assumptions of pseudo-steady-state conditions to estimate intracellular metabolic fluxes. Among the realm of possible methods, flux balance analysis (FBA) and flux variability analysis (FVA) are very popular approaches. While the first method reduces the solution space considering that the cell behaves to satisfy a biological objective [117], flux variability analysis determines the range of possible reaction fluxes that still satisfies the original FBA problem within some optimality factor [231]. Constraint-based analyses are implemented in well-established software such as the COBRA Toolbox [232].

Gene expression models have been widely used in systems biology to describe the regulation and dynamics of mRNA and protein production. These models, often based on ODEs, provide mechanistic insight into gene activation, repression, and feedback control. They are used both in systems biology to model gene regulatory networks and in synthetic biology to design predictable genetic circuits. In the past, many studies have been conducted for the design of gene circuits with the desired behavior [35] and to model and optimize the regulation of engineered metabolic pathways [233, 234].

In synthetic biology applications, gene expression models are often extended to include the expression of exogenous genes, i.e., genes that are not native to the host organism and that are artificially introduced into the host genome to encode novel functions. Unlike endogenous genes, the expression of exogenous genes contributes to the cellular mass but not directly to the host's growth. Importantly, expressing these genes

draws on the same limited cellular machinery (e.g., ribosomes, polymerases), which can interfere with the host's own physiological processes – a phenomenon commonly referred to as gene expression burden. Recently, host-aware gene expression models, also called burden-aware gene expression models, have been developed to address cell burden in the context of designing increasingly complex synthetic genetic circuits in synthetic biology [3, 235]. These models represent an important step toward integrating metabolism and gene expression by capturing how exogenous gene expression perturbs the allocation of cellular resources. They formalize the notion of gene expression burden (also known as cell burden) by describing how the synthesis of proteins encoded by exogenous genes competes with endogenous processes for limited translational resources, ultimately affecting cellular growth and metabolic performance.

In parallel, resource allocation models consider the global budgeting of cellular resources such as the total proteome or ribosome pool. Frameworks like Resource Balance Analysis (RBA) and proteome partition models allocate a finite proteome between cellular functions (e.g., metabolism, translation, housekeeping) and predict growth rate-dependent enzyme expression profiles consistent with empirical bacterial growth laws [236].

5.1.2 Regulatory and Enzyme-Constrained Models

Although constraint-based models provide a powerful framework for understanding cellular metabolism, they lack regulatory detail. As a matter of fact, stoichiometric models do not include information on metabolic regulation such as gene expression and gene regulation via translation or enzyme activity. Henceforth, the integration of transcriptome data into stoichiometric metabolic models has been considered to obtain improved flux predictions and constrain the range of possible flux distributions. In this context, [237] extends the stoichiometric modeling framework with a transcriptional regulatory network using a Boolean logic formalism and [238] uses extreme pathways analysis and reduces the solution space by removing a pathway when it becomes inconsistent with the imposed regulatory constraints. Later, [239] extends the approach by constraining a metabolic flux to zero in the absence of any essential component (sub-unit, assembly factor or translational activator). Yet, many regulatory phenomena cannot be accurately described by Boolean logic, and the previous methods are limited by the available knowledge of regulatory processes.

To tackle this issue, dynamic methods have been developed to account for the feedback of FBA on the regulatory network by considering the concentrations of the metabolites. [240] integrates a set of ODEs into their previous approach (at the expense of the kinetic parameters to identify) to predict the phenotype of diauxic growth of *E. coli*. [241] proposes a data-driven method to limit a reaction to a percentage of its maximum flux value, corresponding to the probability of activation of the gene related to the reaction. [242] develops the first metabolic flux analysis tool that integrates regulatory networks directly to constrain the FBA problem. Nevertheless, regulatory networks are supposed to be known and must be provided by the user. To guarantee the optimality of the predicted phenotype, [243] suggests a model transformation resulting in a stoichiometric representation of gene-protein-reaction associations that can be directly integrated into the stoichiometric matrix. Then, a two-step variant of FBA determines the flux distribution that satisfies an optimal objective while also ensuring the minimization of enzyme usage.

Although regulatory models can capture gene activity and expression states, enzyme-

constrained models extend constraint-based modeling by introducing flux bounds that reflect enzyme catalytic capacity. These models account for proteome allocation trade-offs by constraining reaction rates on the availability of enzymes, where enzyme levels may come from proteomic measurements or assumed allocations. In this context, the method developed in [188] improves steady-state flux predictions by integrating enzymatic capacities derived from proteomic data and kinetic data. However, although powerful for static resource allocation analysis, this framework does not simulate gene expression or dynamic adaptation over time.

A key development in this area is the introduction of the dynamic enzyme-cost flux balance analysis (deFBA) proposed in [244], which extends the enzyme-constrained framework to simulate both enzyme and metabolite levels over time. Subsequent work by [245] adapts deFBA to genome-scale simulations, while [246] proposes a short-term variant (sdeFBA) to improve computational efficiency. Later, the study in [247] introduces a robust formulation (rdeFBA) to handle environmental uncertainty. Despite these advances, the core limitation of enzyme-constrained models remains: enzyme levels are treated as parameters or optimization variables, not as an output of simulated gene expression. As a result, while these methods are powerful in systems biology, especially for predicting resource allocation strategies under nutrient limitations or dynamic shifts, they remain poorly suited for applications in synthetic biology, where dynamic control of gene expression and burden from exogenous proteins are central concerns.

In summary, regulatory models and enzyme-constrained models each extend constraint-based metabolic modeling in complementary ways: one by describing control over gene activity and the other by accounting for catalytic limits. However, neither alone fully describes how gene expression affects metabolic flux in engineered systems.

5.1.3 Conciliation Strategies

The previous methods deal with enzyme capacities as parameters to optimize with no consequences for the host cell but limiting metabolic fluxes. These approaches have proven useful in modeling microbial physiology, but do not address the challenges posed by synthetic biology contexts, where exogenous genes are introduced deliberately into the host genome.

In fact, in the context of synthetic biology and the design of gene circuits, the enzymes that catalyze engineered metabolic pathways consist of exogenous proteins that must be expressed. Although genetic models account for this expression at the regulatory level, its implications for metabolic modeling require further attention. On the one hand, protein expression consumes cellular substrates, a factor already incorporated into constraint-based models through appropriate stoichiometric representations. On the other hand, exogenous gene expression competes with endogenous genes for limited cellular resources such as ribosomes, creating a burden on the host. This burden is not captured in classical constraint-based models, despite its significant impact on host physiology, cell growth, and ultimately on the performance of the designed pathway and the production rate of the target compound. As such, accounting for this burden is essential not only for accurate metabolic modeling but also for guiding the optimal selection of gene expression levels in pathway design.

Bridging metabolism and gene expression explicitly is an area of active research. In systems biology, metabolism and expression models (ME Models) [187] represent a comprehensive approach that includes transcription and translation reactions with a

metabolic model. While mechanistically rich, these models require substantial data and are computationally intensive.

To the best of our knowledge, none of the above-mentioned methods considers the problem of conciliation of constraint-based metabolic and host-aware gene expression models when exogenous enzymes are expressed so as to optimize the flux through a de novo metabolic pathway. This chapter addresses this challenge by proposing a simple and generalizable method to link metabolic flux constraints with gene expression dynamics (specifically through enzyme availability) targeted toward synthetic gene circuits in *E. coli*. As will be discussed in the subsequent sections, the benefits of the proposed method are: (i) constraining the solution space of the flux distribution taking into account enzyme activity on the one hand and the cell burden caused by the expression of exogenous enzymes on the other, (ii) having a better prediction of the production rate of the metabolite of interest, and (iii) assessing/tuning essential enzyme gene expression parameters: transcription and translation strengths. This approach combines several preexisting tools, making the methodology easy to understand and apply by non-expert users.

Furthermore, it is worth noting that advancements have been made since the development of our current conciliation strategy. Recently, more modular and tractable approaches have been proposed. Enhanced Flux Potential Analysis (eFPA) [248] uses enzyme abundance data to modulate flux potentials in a condition-specific manner, while dynamic Constrained Allocation FBA (dCAFBA) [249] couples FBA with proteome allocation to simulate resource reallocation during growth transitions.

5.2 Synthetic Biology Foundations

Although this section begins with a brief overview of gene expression and endogenous gene organization, its purpose is to lay the groundwork for synthetic biology. Core regulatory elements, such as promoters and ribosome binding sites, are not only central to native gene expression but are also repurposed to express exogenous genes in engineered systems.

5.2.1 Regulatory Design Elements

As discussed in Sec. 2.5, transcription and translation are the core processes by which genetic information is expressed as functional protein. During transcription, RNA polymerase binds to a DNA sequence called a **promoter**, located at the start of a gene, and synthesizes a messenger RNA (mRNA) copy of that gene. In bacteria, the promoter region is recognized by the RNA polymerase σ -factor, and its DNA sequence determines how tightly the polymerase binds and how often transcription initiates [250]. A schematic representation of the transcription stage is given in Fig. 5.1.

During translation, the ribosomes assemble on the mRNA at a specific site known as the **ribosome binding site** (RBS). The RBS is a short mRNA sequence that base-pairs with the ribosomal RNA to position the ribosome for start codon recognition. Once a ribosome is bound, it proceeds to synthesize by reading the mRNA codons and assembling the amino acids in sequence. A schematic representation of the translation stage is shown in Fig. 5.2.

Promoters and RBS elements are key regulatory sequences that control gene expression: the promoter governs transcription (mRNA production rate), and the RBS

governs translation initiation (how efficiently ribosomes launch protein synthesis on the mRNA). However, not all promoters and RBSs are created equal, they vary in *strength*, which refers to their effectiveness in driving gene expression.

5.2.1.1 Promoter Strength

Promoter strength denotes the frequency of transcription initiation events from that promoter. It is determined by the DNA sequence elements and how optimally they bind to RNA polymerase and any transcription factors. A *strong promoter* has a DNA sequence close to the consensus recognized by RNA polymerase, leading to frequent initiation of mRNA transcripts, whereas a *weak promoter* has less optimal sequence and initiates transcription less frequently [250].

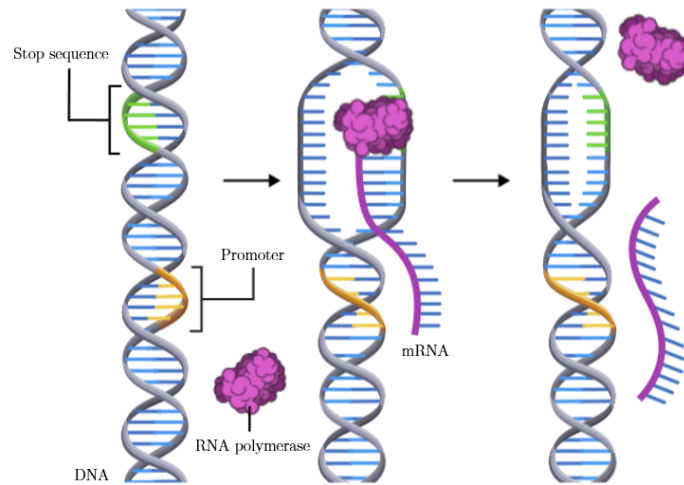


Figure 5.1: Schematic representation of the transcription stage (initiation step, elongation step and termination step).

5.2.1.2 RBS Strength

RBS strength refers to the efficiency with which a ribosome can initiate translation at that site on the mRNA. A strong RBS has a nucleotide sequence (and structure) that robustly recruits the ribosome, resulting in frequent translation initiation. The translation initiation rate of an RBS is largely encoded in its sequence, and synthetic biologists can predict and tune the RBS strength by sequence design [251]. A *strong RBS* yields more ribosome-loaded mRNAs and thus more proteins, while a *weak RBS* leads to fewer initiation events and lower protein production.

It is worth noting that beyond initiation, other factors such as mRNA secondary structure and codon usage also affect translation, but the RBS is a primary determinant of how hard each mRNA is pushed into producing protein.

5.2.2 Endogenous vs. Engineered Systems

In both systems and synthetic biology, the characterization of gene expression often depends on the nature and origin of the genes involved. Broadly, we distinguish be-

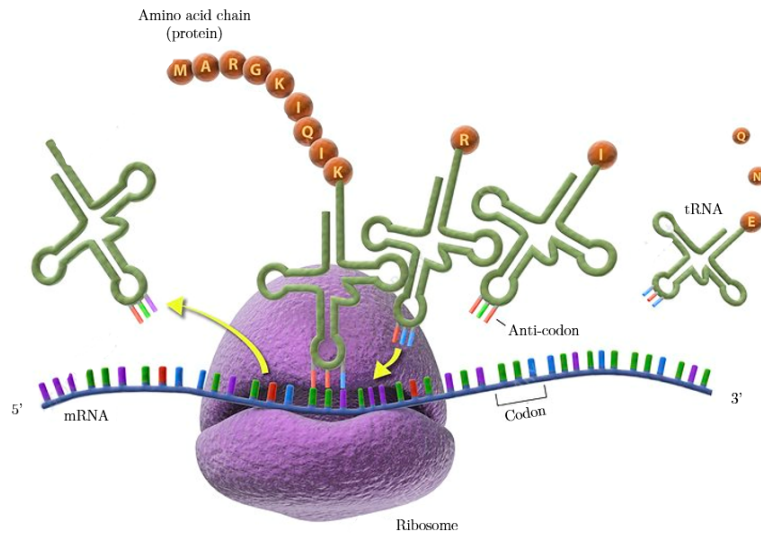


Figure 5.2: Schematic representation of the translation stage and synthesis of the protein (sequence of amino acids).

tween endogenous genes, which are naturally present in the host genome, and exogenous genes, which are introduced through genetic engineering. While endogenous genes govern the cell's native physiological functions, exogenous genes are typically associated with synthetic gene circuits or engineered metabolic pathways, introduced to confer new capabilities or reroute cellular functions. Their expression can lead to cellular burden, a phenomenon in which competition for shared resources perturbs both host physiology and circuit performance.

5.2.2.1 Endogenous Genes

In systems biology, we often distinguish between **ribosomal genes**, i.e., those encoding components of the protein synthesis machinery, and **non-ribosomal genes**, which include all other genes that serve various cellular function ¹. **Ribosomal genes** primarily consist of those encoding the ribosomal RNAs (rRNAs) and the ribosomal proteins (E. coli's ribosome contains 55 different proteins along with 3 rRNAs) as well as several translation factors. These genes are highly expressed, especially when the cell is growing rapidly, because the cell must produce many ribosomes to support high protein synthesis demands.

In contrast, **non-ribosomal genes** encompass all other genes, i.e., enzymes, regulatory proteins or structural proteins, which collectively use the remaining portion of gene expression capacity. Because ribosomal components take up such a large share of the biosynthetic budget, there is an inherent competition between the production of ribosomes and the production of other proteins. The cell balances this allocation according to growth conditions. At steady-state, faster growth rates demand a larger fraction of resources devoted to ribosomal gene expression to produce enough ribosomes, leaving a smaller portion for other proteins [252]. Conversely, in slower growth or nutrient-poor conditions, relatively fewer ribosomes are needed, and more resources can go toward

¹Here, **non-ribosomal** simply means any gene not involved in the ribosome; this usage is distinct from **non-ribosomal peptides**, which refers to a different biosynthetic process outside of ribosomes.

non-ribosomal functions (stress responders, catabolic enzymes).

The important takeaway is that protein capacity is a finite resource. Translational resources (ribosomes, tRNAs, elongation factors) and transcriptional resources (RNA polymerases, nucleotides) are shared between making ribosomal proteins/RNA and making all other proteins. This trade-off underpins many cellular behaviors and is a key consideration when we introduce additional genes in synthetic biology, as it is explained in Sec. 5.2.2.2. To understand and predict gene expression dynamics, researchers often build mathematical models of the transcription and translation processes. These **gene expression models** are differential equations that simulate how mRNA and protein concentrations change over time.

5.2.2.2 Exogenous Genes

A central theme in synthetic biology is the introduction of **exogenous genes**, i.e., genes that are not originally part of the host's genome, to endow cells with new capabilities. These genes could encode fluorescent reporters, regulators, or entire metabolic enzymes for new pathways (e.g., **engineered metabolic pathways**). When such a gene is inserted (for example, on a plasmid or via genome integration) and expressed in a host cell, it is essentially diverting a fraction of the host's transcriptional and translational machinery. The host's RNA polymerases will transcribe the foreign gene and the host's ribosomes will translate the mRNA into an exogenous protein. As a result, the expression of externally introduced genes inevitably draws on the finite pool of cellular resources for the host's own gene expression and growth.

The use of RNA polymerases, ribosomes, tRNAs, amino acids, ATP/GTP, by the foreign genes can impose a **cellular burden** on the host [253]. In practical terms, this burden often manifests as slower growth rate, lower endogenous protein expression, or other physiological stresses, because some portion of the host's biosynthetic energy and resources is being siphoned off to produce non-native proteins.

In the context of **engineered metabolic pathways**, synthetic biologists can insert a suite of exogenous genes encoding enzymes that catalyze a series of reactions to produce a desired chemical (for instance, a biofuel, pharmaceutical precursor, or novel metabolite). Each of those enzymes must be expressed, and together they can constitute a substantial new drain on the cell's resources. Moreover, beyond the gene expression burden, an active metabolic pathway can divert substrates, ATP, and reducing power away from native metabolic processes, compounding the stress on the cell (metabolic burden). The cell must not only continue its normal metabolic functions and growth, but also supply the materials and energy to run the introduced pathway. If not carefully engineered, this can lead to reduced growth and productivity. This is a well-recognized challenge: any exogenous gene or pathway expression perturbs the host's resource allocation. In that context, **host-aware** or **burden-aware gene expression models** have been developed to represent gene expression dynamics taking into account cellular burden and competition for host resources.

Synthetic biology also creates complex **gene circuits** that are networks of interacting genes and regulatory elements designed to implement functions such as logic gates, oscillators, or feedback controllers [234, 254]. Classic examples include toggle switches and oscillators engineered in *E. coli*. These circuits often involve multiple exogenous regulatory proteins (repressors, activators) that control each other's expression. Like any other heterologous genes, the components of a synthetic gene circuit draw on the

host’s finite pool of RNA polymerases and ribosomes.

5.2.2.3 Parameter Tuning

The strength of a promoter or ribosome binding site (RBS) determines the rate of transcription and translation, respectively, and is a key lever for modulating gene expression. In both endogenous and exogenous contexts, tuning these regulatory elements allows for precise control over protein production, whether to optimize metabolic fluxes, balance pathway expression, or manage cellular burden (see Sec. 5.3.3.2.2).

For exogenous genes, tuning is typically achieved during construct design by selecting or modifying promoter and RBS sequences to achieve desired expression levels. Libraries of promoters and RBSs with graded strengths, or computational tools, can assist in this process. For endogenous genes, expression can be modulated by replacing their native regulatory elements or using targeted regulation strategies, though care is needed to avoid disrupting essential cellular functions.

This ability to adjust gene expression is fundamental to optimizing engineered pathways, minimizing cellular burden, and ensuring predictable behavior in synthetic systems. A detailed overview of regulatory design principles, including promoter and RBS tuning strategies, can be found in the foundational work by Alon [39].

5.3 Constraint-Based Burden-Aware Modeling

In this section, we present the modeling frameworks used to describe the cellular behavior at two distinct but complementary levels: metabolic models to characterize metabolism, and burden-aware gene expression models to capture the dynamics of resource allocation at the genetic level. After introducing the main features of each framework, we describe the proposed strategy to reconcile the two levels in a unified modeling approach.

The conciliation method suggested in this work is deliberately simple and conceptual in nature. Rather than developing a highly intricate or data-heavy integration, the objective is to propose a first methodological step toward linking metabolic flux predictions with gene expression constraints, especially in the context of synthetic biology when exogenous proteins are synthesized. While the approach offers a practical and accessible way to bridge the two modeling scales, it also leaves considerable room for improvement. More detailed or mechanistically rich connections could be envisioned in future work, and further research is needed to deepen the integration between metabolic and genetic modeling frameworks, a challenge that remains both complex and highly promising.

5.3.1 Metabolic Modeling

As a reminder, cellular metabolism is defined by a set of chemical reactions, possibly catalyzed by enzymes, taking place within the cell and forming metabolic pathways. The structure of cellular metabolism can be described by a stoichiometric matrix N representing the network of biochemical reactions and known as the **metabolic network**. The major concepts underlying the analysis of such networks have been largely covered in the previous chapter, specifically in Sec. 4.2.3. Here, we briefly recall that intracellular flux distributions can be evaluated using **constraint-based methods**, which rely on applying physico-chemical and biological constraints, such as mass balances under

pseudo-steady-state assumptions (QSSA), bounds on reaction rates, and cellular objectives to the metabolic network. Among these methods, flux balance analysis (FBA) and flux variability analysis (FVA) are widely employed to predict feasible flux distributions without requiring detailed kinetic parameters. Throughout this chapter, we will use these constraint-based frameworks to characterize metabolic activity in connection with gene expression considerations.

5.3.2 Genetic Modeling

Models of gene expression accounting for host-circuit interactions are particularly interesting for understanding how the cell growth and the expression of cell endogenous genes evolve when exogenous genes are introduced into the host cell. The current work relies on the host-aware gene expression model proposed in [3]. The latter accounts for the dynamics of the expression of the cell host endogenous protein-coding genes (contributing to cell mass and cell growth) on the one hand and the expression of protein-coding exogenous genes (contributing to cell mass only) on the other hand. For this purpose, Fig. 5.3 illustrates a schematic representation of how cellular resources are allocated to the synthesis of ribosomal and non-ribosomal proteins, and exogenous proteins when expressed.

5.3.2.1 Protein-Coding Gene Dynamics

The gene expression dynamics for a generic k -type-protein-coding gene in prokaryote cells, under the assumption that transcription is faster than translation and that ribosomes are the limiting shared resource required for protein expression, is given by the following equation:

$$\dot{p}_k = \frac{\nu_t(s)}{l_{pk}} J_k(\mu, r) r - (d_k + \mu) p_k \quad (5.1)$$

where p_k is the number of copies of the protein of class k , l_{pk} represents the length of the protein (expressed as equivalent number of amino acids), d_k is the protein degradation rate, μ stands for the specific growth rate of the cell, r is the number of free ribosomes and $\nu_t(s)$ symbolizes the substrate-dependent effective peptide elongation rate. As a first approximation, the latter is expressed using a Michaelis-Menten expression that is considered to be organism- and substrate-dependent but does not depend on the nucleotide sequence.

The key functional coefficient $J_k(\mu, r)$ is defined as the resource recruitment strength and allows explaining the distribution of resources (free ribosomes) between the host cell and the genes of interest k (either endogenous ribosomal genes, endogenous non-ribosomal genes or exogenous genes). It takes into account lab-accessible parameters such as promoter and ribosome binding site (RBS) strengths characterizing the steps of genetic transcription and translation, respectively. The resource recruitment strength is defined as follows:

$$J_k(\mu, r) \triangleq E_{mk}(l_{pk}, l_e) \omega_k \left(\frac{d_{mk}}{K_C^k(s)} + \mu r \right)^{-1} \quad (5.2)$$

where μr is the flux of free ribosomes, E_{mk} represents the ribosome density-related term and can be approximated to $0.62 l_{pk} (l_e)^{-1}$ with $(l_e)^{-1}$ being the specific ribosome density, ω_k represents the mRNA production rate (i.e., a parameter related to the promoter strength), d_{mk} is the mRNA degradation rate and $K_C^k(s)$ symbolizes the effective

RBS strength and depends on the availability of the substrate s . The latter parameter is expressed as follows:

$$K_C^k(s) \triangleq \frac{K_b^k}{K_u^k + K_e(s)} \quad (5.3)$$

where K_b^k and K_u^k represent, respectively, the association and dissociation rate between a free ribosome and the RBS and $K_e(s)$ is the translation initiation rate.

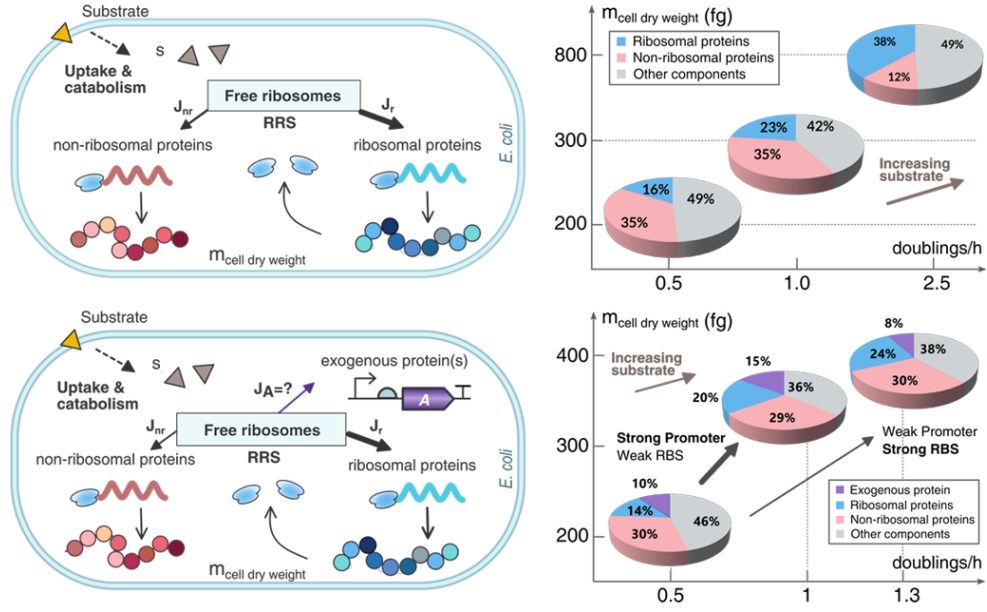


Figure 5.3: Schematic view of cellular resource allocation between protein types. On the left-side, the partitioning of ribosomal resources for the synthesis of ribosomal, non-ribosomal endogenous proteins, and exogenous proteins (when expressed). On the right-side, pie charts depict how the fractions of cell dry weight vary with growth rate and substrate availability under different promoter/RBS strategies (from [3]).

5.3.2.2 Cell Growth Expression

Cell growth can be defined as the variation of the protein fraction of the total cell mass. However, not all protein mass contributes to cell growth. Some proteins may be actively degraded, while others, such as exogenous genes, do not play an active role in promoting cell growth. Therefore, only endogenous ribosomal and non-ribosomal proteins are accounted for when calculating the cell-specific growth rate. Under the assumption that the protein mass for each cell rapidly reaches the steady-state, the cell-specific growth rate is obtained as follows:

$$\mu(s) = \frac{m_{aa}}{m_h} \nu_t(s) \Phi_t^h \Phi_m r_T \quad (5.4)$$

where m_{aa} is the average amino acid mass, m_h is the mass of the native host cell, $\Phi_m r_T$ represents the number of available mature ribosomes and Φ_t^h can be defined as a modulation function that is an image of the cellular burden and represents the fraction of

ribosomes elongating along endogenous ribosomal and non-ribosomal transcripts relative to the total number of available mature ribosomes. This modulation function can be expressed as follows:

$$\Phi_t^h = \frac{N_r J_r(\mu, r) + N_{nr} J_{nr}(\mu, r)}{1 + \Psi(\mu, r)} \quad (5.5)$$

with the function $\Psi(\mu, r)$ defined as:

$$\Psi(\mu, r) = N_r J_r \left(1 + \frac{1}{E_{mr}}\right) + N_{nr} J_{nr} \left(1 + \frac{1}{E_{mnr}}\right) + \sum_{k=1}^{N_{exo}} J_k \left(1 + \frac{1}{E_{mk}}\right) \quad (5.6)$$

where N represents the number of protein-coding genes. The indexes r and nr stand for ribosomal and non-ribosomal protein-coding endogenous genes and N_{exo} allows the existence of exogenous genes.

To maintain clarity, only a subset of the equations of the burden-aware model proposed by [3] is presented here. The original model not only describes the dynamics of ribosome allocation and protein expression but also explicitly tracks the evolution of key variables, including the total mass of ribosomes, the total mass of non-ribosomal endogenous proteins, the mass of exogenous proteins, and the overall host cell mass. It also predicts important outputs such as the cell's specific growth rate and the productivity of target proteins over time. Although the full model includes a rich set of variables and equations, it remains experimentally tractable, as many of the biological parameters involved can be obtained or approximated from empirical data.

In practice, most parameters associated with endogenous gene expression, such as ribosome production rates, basal translation efficiencies, and protein degradation rates, can be inferred from standard experimental techniques such as proteomics, ribosome profiling, or growth assays. These experimental platforms provide quantitative estimates of protein abundance, ribosome occupancy, and cellular growth dynamics under various conditions. Parameters related to exogenous genes, such as promoter strength, ribosome binding site strength, and expression burden coefficients, typically require characterization or fitting in the engineered system. More information and details can be found in the original paper.

5.3.3 Conciliation Strategy

The objective of this work is to propose a simple and conceptual method to reconcile the two modeling frameworks introduced previously, i.e., constraint-based metabolic models and burden-aware gene expression models, in the context of synthetic biology where exogenous enzymes associated with engineered metabolic pathways are expressed.

In classical enzyme-constrained models, such as the one introduced in [188], intracellular fluxes are limited by catalytic constraints of the form:

$$v_E^{ub} = k_{cat,E} [E] \quad (5.7)$$

where v_E^{ub} is the upper bound of the flux related to the reaction catalyzed by enzyme E , $k_{cat,E}$ is the catalytic rate constant and $[E]$ is the concentration of the corresponding enzyme. This class of approaches has proven effective for improving metabolic flux predictions by incorporating enzyme capacity limitations. However, they typically rely on static enzyme abundance estimates and do not account for the dynamic effects of gene

expression, resource competition, or cellular burden that are factors that become particularly important in synthetic biology applications involving exogenous gene expression.

Here, we propose taking advantage of the burden-aware gene expression model to dynamically predict enzyme concentrations and cell growth rates as functions of substrate availability and promoter and RBS strengths. This information is then used to constrain the corresponding reaction fluxes in the constraint-based metabolic model. In particular, for each enzyme involved in the engineered pathway, the upper bound of the flux that it catalyzes is determined on the basis of the predicted enzyme concentration and its catalytic constant, following the relationship in Eq. 5.7.

Thus, given a fixed amount of limiting substrate, the primary objective is to reconcile the constraint-based metabolic model and the burden-aware gene expression model in a coherent framework, where the predicted enzyme concentrations constrain the intracellular fluxes. The secondary objective is to estimate the promoter and RBS strengths required to express the pathway enzymes in a way that maximizes the flux of the target metabolite while accounting for the effects of exogenous expression burden on the host cell physiology. Fig. 5.4 illustrates the general input-output characteristics of the two modeling frameworks.

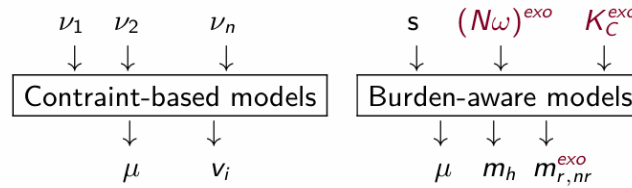


Figure 5.4: Input-output characteristics of constraint-based metabolic models and gene expression models. Additional elements highlighted in red indicate the extra considerations required when accounting for the expression of exogenous genes.

As illustrated in Fig. 5.4, the constraint-based metabolic model and the gene expression model differ substantially from an input-output perspective. Constraint-based models require experimental measurements of uptake and excretion rates (e.g., glucose uptake, metabolite secretion) as inputs, and predict intracellular metabolic flux distributions and the associated cell growth rate as outputs. In contrast, the burden-aware gene expression model considers as input an abstract quantity s , representing the availability of the limiting substrate, and predicts as outputs the host cell's growth rate together with the mass of the host cell and the masses of ribosomal and non-ribosomal endogenous proteins and exogenous proteins. The corresponding concentrations can be subsequently calculated by combining the predicted protein masses with their molar masses, the total host cell mass and the Avogadro's number.

5.3.3.1 Model Adaptations and Extensions

This part describes the modifications required in both metabolic and gene expression models to enable their conciliation. First, we establish how substrate availability is mapped between the two frameworks. Then, we describe how engineered metabolic pathways are incorporated into the stoichiometric model, followed by the necessary adjustments to the burden-aware gene expression model to account for exogenous protein expression.

5.3.3.1.1 Mapping Substrate Availability

To reconcile the two modeling frameworks, it is necessary to align their respective descriptions of substrate limitation. In the gene expression model, substrate limitation is abstracted as an availability parameter s , representing the accessible quantity of the limiting nutrient from a resource allocation perspective. In contrast, the constraint-based metabolic model represents substrate limitation through specific uptake rates (e.g., glucose uptake rate ν_{Glc}) in the stoichiometric network. Therefore, a mapping between s and ν_{Glc} must be established to ensure consistent growth rate predictions between both models.

This mapping is constructed based on the wild-type system, i.e., the native metabolic network without engineered pathways and the gene expression model without exogenous gene expression. To this end, the growth rate μ is first characterized as a function of substrate availability s using the gene expression model. The resulting $\mu(s)$ relationship exhibits a Monod-type behavior, as expected under substrate-limited conditions.

In parallel, a series of simulations are performed using the constraint-based metabolic model by prescribing different specific glucose uptake rates ν_{Glc} and recording the corresponding growth rates obtained under biomass maximization. By matching the growth rates predicted by the two frameworks for the wild-type configuration, a correspondence between ν_{Glc} and s is established. This resulting functional mapping $\nu_{Glc} = f(s)$ is obtained through regression and is used to ensure consistent integration of the two modeling layers (see Sec. 5.4.2.1).

Note that this procedure becomes more complex when multiple limiting substrates (e.g., glucose and oxygen) are simultaneously considered.

5.3.3.1.2 Adding Engineered Pathways

To model the introduction of a synthetic metabolic pathway, one or more reactions are incorporated into the stoichiometric network. Specifically, reactions that represent the formation of the targeted metabolite are added to the stoichiometric matrix N . To maintain mass balance and avoid the introduction of artificial gaps, the synthetic reaction must originate from intermediate metabolites that are already synthesized by the host cell. This ensures that the engineered pathway is biologically feasible and is properly integrated into the host's metabolic capabilities.

The newly added reaction is assumed to be catalyzed by an exogenous enzyme, i.e., a protein encoded by a gene that is not part of the endogenous genome and must be artificially expressed. Consequently, the corresponding intracellular flux will be constrained by the catalytic capacity of the exogenous enzyme. If the flux of the engineered metabolic reaction is denoted v_A and the exogenous enzyme is noted A , then the following relationship is introduced into the metabolic flux analysis formulation:

$$v_A^{ub} = k_{cat,A} [A] \quad (5.8)$$

This formulation directly links intracellular flux limits to gene expression dynamics and resource allocation, ensuring that the capacity of the engineered pathway remains consistent with the burden imposed on the host cell. In doing so, it extends classical constraint-based modeling by making selected flux bounds dynamically dependent on the outcomes of a burden-aware gene expression model.

5.3.3.1.3 Expressing Exogenous Genes

As depicted in Fig. 5.4, the burden-aware gene expression model developed in [3] predicts several key variables: the specific growth rate μ , the masses of ribosomal (m_r) and non-ribosomal (m_{nr}) endogenous proteins, the mass of exogenous proteins if present (m_{exo}), and the overall host cell mass (m_h). Furthermore, as discussed in Sec. 5.3.2, proteins are modeled in terms of copy numbers (p_k) and associated masses, aggregated into functional pools rather than tracking individual gene products.

Endogenous gene expression is divided into two major classes: ribosomal and non-ribosomal proteins. For each class, average transcription, translation, and degradation parameters were estimated using experimental datasets such as proteomics and ribosome profiling. The model does not resolve individual endogenous genes but instead assumes representative behaviors for each class.

To introduce an exogenous gene, the genetic model is extended by adding a new pool associated with the expression of the exogenous protein. This requires specifying genetic parameters, including:

- The coding sequence length (expressed in number of amino acids), directly influencing the translation elongation time;
- The promoter strength (controlling transcription initiation frequency);
- The RBS strength (controlling ribosome binding and translation initiation rate);
- The degradation rate (if known or assumed similar to endogenous non-ribosomal proteins).

Typically, values for coding sequence length can be extracted from sequence databases (e.g., UniProt), while promoter and RBS strengths are either selected from characterized parts or inferred from standard synthetic biology libraries. In the absence of detailed characterization, degradation rates and elongation parameters can be reasonably assumed to match the endogenous non-ribosomal pool as a first approximation.

The exogenous gene imposes an additional load on the cellular translation machinery, directly impacting the allocation of ribosomes and thus indirectly reducing the resources available for endogenous gene expression. Consequently, the predicted growth rate μ will reflect not only the metabolic state of the cell but also the burden associated with exogenous protein production. This adaptation allows for dynamic, burden-aware prediction of enzyme concentrations for synthetic pathways, providing the necessary bridge to integrate gene expression dynamics into constraint-based metabolic modeling.

5.3.3.2 Algorithms

5.3.3.2.1 Conciliation Algorithm

A stepwise conciliation procedure is proposed in **Algorithm 1** for the case of a single exogenous enzyme A , which catalyzes the flux v_A of the engineered metabolic pathway, and considering a limiting substrate s . In this initial formulation, the promoter and ribosome binding site (RBS) strengths controlling the expression of the exogenous gene are considered fixed parameters rather than optimization variables. The objective is to determine the intracellular flux distribution of the metabolic network when an exogenous gene is expressed, accounting for the constraints imposed by enzyme availability and the burden on cellular resources. This conciliation strategy aims to bridge

the constraint-based metabolic model and the burden-aware gene expression model for systems involving synthetic metabolic pathways catalyzed by exogenous enzymes.

It is important to note that the expression machinery associated with endogenous genes (both ribosomal and non-ribosomal) has been previously parametrized and validated in [3], and is assumed to be already configured in this study.

In this context, we introduce an additional exogenous gene encoding the enzyme A responsible for catalyzing the flux through the synthetic pathway of interest. Therefore, the burden-aware model is accordingly extended to account for the expression of this exogenous gene, requiring the specification of certain genetic parameters – such as the coding sequence length, the promoter strength and the strength of the ribosome binding site (RBS) – which can be obtained from the literature or experimental characterization.

The conciliation process begins by setting a value for the substrate availability parameter s , mapped to the corresponding glucose uptake rate ν_{Glc} via the calibration step procedure described earlier. Fixed values for the promoter and RBS strengths associated with the exogenous gene are assigned, ensuring biological realism. Using the burden-aware model, the cell growth rate μ , the masses of endogenous ribosomal, non-ribosomal and exogenous protein pools, and the host cell mass are predicted. Then, the concentration of the exogenous enzyme is derived from mass ratios and molar mass information.

An upper bound v_A^{ub} on the flux catalyzed by the exogenous enzyme is then imposed based on the catalytic relationship in Eq. 5.8. Subsequently, a flux variability analysis (FVA) is performed under the following constraints:

- The mass balance ($N\underline{v} = 0$);
- The reaction directionality ($\underline{v} \geq 0$);
- The catalytic constraint ($0 \leq v_A \leq v_A^{ub}$);
- A fixed cell growth rate consistent with that predicted by the burden-aware model.

If a feasible flux distribution is found, the corresponding fluxes are recorded. Otherwise, the promoter and RBS strengths must be adjusted. Failure to find a feasible solution typically indicates that initial expression parameters are incompatible with maintaining the necessary balance between enzyme availability, cellular burden, and host growth, possibly due to an excessive burden slowing cell doubling, or insufficient expression failing to meet metabolic demands.

5.3.3.2.2 Promoter Strength Optimization

To further refine the reconciliation between the metabolic and gene expression models, an extended algorithm is proposed to optimize the promoter strength controlling the expression of the exogenous gene. In this strategy, the RBS strength is fixed a priori to a biologically realistic value based on synthetic biology libraries, and only the promoter strength is iteratively adjusted. Although the procedure focuses on optimizing the promoter strength, it could equivalently be adapted to optimize the RBS strength if desired. This conciliation strategy is given in **Algorithm 2**.

The optimization is motivated by biological constraints on cellular resource allocation. Experimental studies have shown that, to maintain proper cellular function and growth, the mass of all exogenous proteins should not exceed approximately 40% of

Algorithm 1: Conciliation Strategy

Input: s , parameters related to the exogenous gene
Output: v_i

- 1 Select $s \longleftrightarrow \nu_{Glc}$;
- 2 Fix promoter/RBS strengths (related to the expression of the exogenous gene);
- 3 Derive μ , m_A and m_h from the gene expression model;
- 4 Calculate $[A]$ using m_A , m_h and M_A ;
- 5 Compute v_A^{ub} using Eq. 5.8;
- 6 Perform FVA s.t. $\{N\underline{v} = 0 ; \underline{v} \geq 0 ; 0 \leq v_A \leq v_A^{ub} ; \mu \text{ and } \nu_{Glc} \text{ fixed}\}$;
- 7 **if** *feasible solution* **then**
 - 8 | Record v_i from metabolic analysis;
- 9 **else**
 - 10 | Adjust promoter/RBS strengths to ensure feasibility;
 - 11 | Back to **3**;
- 12 **end**

Algorithm 2: Promoter Strength Optimization

Input: s , parameters related to the exogenous gene
Output: v_i , promoter strength ^{opt}

- 1 Select $s \longleftrightarrow \nu_{Glc}$;
- 2 Initial guess for promoter strength (RBS strength fixed);
- 3 Derive m_A and m_h from the gene expression model;
- 4 **while** $m_A < 0.4 m_h$ **do**
 - 5 | Increase promoter strength;
 - 6 | Get m_A , m_h and μ from the gene expression model;
- 7 **end**
- 8 Impose $\mu_{min} = \mu$;
- 9 Calculate $[A]$ using m_A , m_h and M_A ;
- 10 Compute v_A^{ub} using Eq. 5.8;
- 11 Perform FVA s.t. $\{N\underline{v} = 0 ; \underline{v} \geq 0 ; 0 \leq v_A \leq v_A^{ub} ; \mu_{min} \text{ and } \nu_{Glc} \text{ fixed}\}$;
- 12 **if** *feasible solution* **then**
 - 13 | Record v_i from metabolic analysis;
 - 14 | Record promoter strength ^{opt} ;
- 15 **else**
 - 16 | Decrease promoter strength or change RBS strength;
 - 17 | Back to **3**;
- 18 **end**

the total host cell mass [227, 255]. Exceeding this threshold would imply an excessive burden on the host’s translational machinery, impairing its ability to grow and sustain physiological functions. Therefore, the promoter strength is tuned to reach a level of exogenous enzyme expression that remains within this acceptable burden range. However, it is worth noting that this threshold is not absolute and may depend on the microbial strain, growth conditions, and specific burden characteristics. The 40% value is used here as a practical guideline, but the procedure itself remains general and could accommodate different thresholds if needed, as long as metabolic feasibility is preserved.

The goal of the optimization is to determine the promoter strength that maximizes the flux v_A through the engineered metabolic pathway, while accounting for enzyme availability constraints, the impact of exogenous burden on host growth, and metabolic network feasibility. The procedure iteratively adjusts the promoter strength associated with the exogenous gene expression until the optimal balance between burden, enzyme production, and metabolic performance is achieved, with a fixed substrate availability.

Practically, the optimization procedure proceeds in two stages. First, a preliminary iteration is performed using the burden-aware gene expression model, incrementally increasing the promoter strength until the mass of the exogenous protein pool approaches, but not exceed, the predefined burden threshold (e.g., 40% of the total host cell mass). Once a suitable promoter strength and associated growth rate are identified, the concentration of the exogenous enzyme is computed from its predicted mass, molar mass, and the host cell mass. Then, this concentration is used to fix an upper bound on the corresponding flux v_A through the catalytic constraint expressed in Eq. 5.8. Subsequently, a flux variability analysis (FVA) is performed under the same constraints as the ones described in **Algorithm 1** while possibly restricting the value of other fluxes if experimental data are available.

If a feasible solution is found, the corresponding intracellular fluxes and the optimal promoter strength are recorded. Otherwise, it indicates that the current promoter strength was excessive relative to the host’s resource allocation capacity under the given conditions. In such cases, the promoter strength is reduced (or alternatively, the fixed RBS strength can be modified) and the procedure is repeated until a feasible solution is obtained.

5.4 Application to p-Coumaric Acid Biosynthesis

This section presents a case study that illustrates the application of the proposed conciliation strategy that connects constraint-based metabolic models and burden-aware gene expression models. The goal is to model and optimize the biosynthesis of p-Coumaric acid, a valuable aromatic metabolite of industrial interest, in a synthetically engineered *E. coli* strain.

After introducing the biological and synthetic context, we detail the preliminary setup steps required to align the two modeling frameworks, including substrate mapping, pathway integration, and adaptation of the gene expression model. Then, the conciliation strategy is applied and the results are analyzed in terms of metabolic flux predictions, resource allocation, and pathway productivity.

5.4.1 Engineering *E. coli* for p-Coumaric Acid Production

In this case study, the production of p-Coumaric acid (p-hydroxycinnamic acid) is considered a representative application of synthetic biology. p-Coumaric acid is an aromatic organic compound known for its antioxidant and antimicrobial properties and plays an essential role in human health. It has been shown to decrease low-density lipoprotein peroxidation and may contribute to reducing the risk of stomach cancer by limiting the formation of carcinogenic nitrosamines. Beyond its direct biological activities, p-Coumaric acid serves as a critical precursor for the biosynthesis of phenolic acids, flavonoids, lignin, and a wide range of other secondary metabolites of industrial and pharmaceutical interest [254].

Regarding its biosynthesis, p-Coumaric acid can be produced from cinnamic acid through hydroxylation by the action of the enzyme 4-cinnamic acid hydroxylase, or more directly from L-tyrosine via the deamination reaction catalyzed by tyrosine ammonia lyase (TAL). In microbial hosts such as *E. coli*, which naturally secrete L-tyrosine under certain conditions, the TAL-catalyzed route provides a particularly attractive engineering target for the production of p-Coumaric acid.

Recent studies [254, 256] have demonstrated that the production of p-Coumaric acid in engineered *E. coli* strains forms an essential step toward the biosynthesis of more complex molecules, such as naringenin, a key precursor for flavonoid production. Efficient expression of TAL enzyme and balancing the metabolic flux toward p-Coumaric acid are thus critical for optimizing the overall pathway performance in synthetic biology applications.

Accordingly, this study focuses on modeling and optimizing the synthetic production of p-Coumaric acid from L-tyrosine in a genetically modified *E. coli* host, integrating the metabolic and genetic modeling frameworks described previously.

5.4.2 Preliminary Setup

Before applying the conciliation strategy, several preparatory steps must be completed to ensure a consistent and biologically meaningful simulation framework. This preliminary setup focuses on three key aspects: (i) aligning substrate descriptions between the two modeling frameworks, (ii) integrating the engineered metabolic pathway responsible for p-Coumaric acid production, and (iii) adapting the gene expression model to account for the expression of the exogenous enzyme catalyzing the synthetic reaction.

5.4.2.1 Substrate Mapping

In the burden-aware gene expression model developed in [3], the quantity s represents an abstract measure of substrate availability, introduced to modulate key biological processes such as translation elongation rates $\nu_t(s)$ and ribosome binding efficiencies $K_C^k(s)$. Rather than representing a physical quantity, s abstracts the general nutrient richness of the environment as perceived by the cell. In that context, it is important to emphasize that s is a dimensionless model parameter, not a directly measurable experimental variable. In the original study, the burden-aware model was calibrated against global physiological data, particularly from [257, 258], which provide relationships between cellular macromolecular composition and growth rate under different conditions. However, no experimental data directly connect s to physical quantities such as extracellular substrate concentrations or uptake rates.

In the absence of direct experimental measurements of s and recognizing its role as an internal model abstraction rather than a physical substrate concentration, a mapping procedure based solely on simulations was developed (see Sec. 5.3.3.1.1). This calibration strategy ensures a consistent coupling between gene expression and metabolic models by relating the substrate availability parameter s to the specific glucose uptake rate ν_{Glc} by matching the predicted growth rates. The numerical results are depicted in Fig. 5.5.

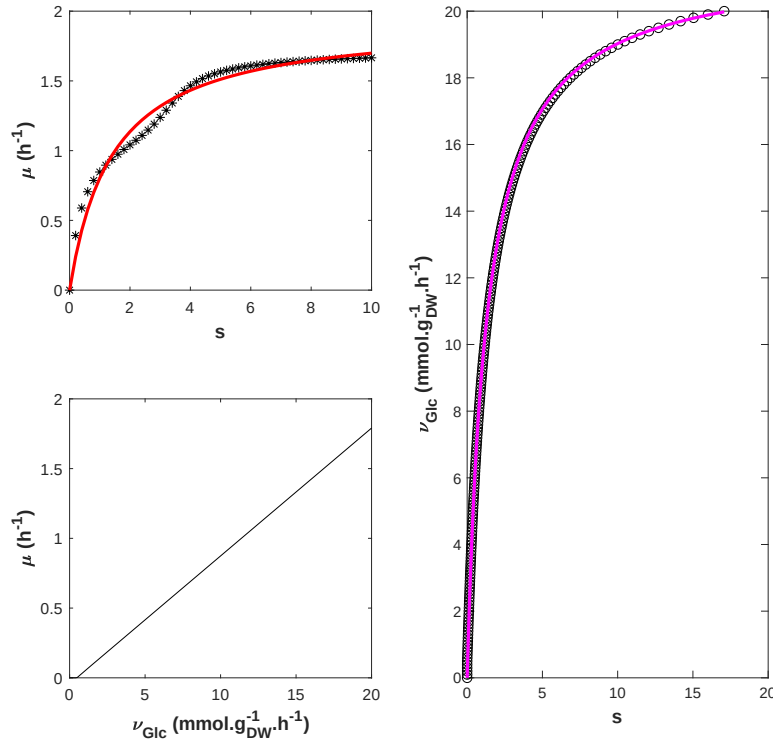


Figure 5.5: Top-left: Growth rate μ versus substrate availability s from the gene expression model (black points: simulation data; red curve: fitted Michaelis-Menten-type law). Bottom-left: Growth rate μ versus glucose uptake rate ν_{Glc} from constraint-based metabolic model with biomass optimization (simulation data). Right-side: Mapping between s and ν_{Glc} by matching growth rates (black bullets: matched data; curve: fitted functional relationship).

As expected from the structure of the burden-aware model, the relationship between substrate availability s and the predicted growth rate μ exhibits a Michaelis-Menten-like saturation behavior. The growth rate increases sharply at low s and gradually saturates at high s , reflecting the biological coupling between nutrient availability, translational efficiency and cell growth as discussed in Sec. 5.3.2. Conversely, the constraint-based metabolic model, under biomass maximization, predicts an approximately linear relationship between glucose uptake rate and growth rate across the physiological range of uptake values typically observed for *E. coli* (i.e., -5 to -20 mmol.g_{DW}⁻¹.h⁻¹). Only under unrealistically high uptake rates would saturation effects be expected.

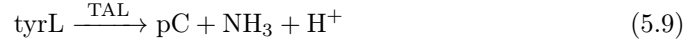
Because of this difference in the input-output relationships between the two modeling frameworks, the mapping between s and ν_{Glc} cannot be determined analytically. Instead, a numerical correspondence is constructed by matching the predicted growth rates. The resulting mapping curve exhibits a sublinear trend, where glucose uptake increases rapidly with s at low values and progressively flattens at higher s , reflecting the biological saturation of the translation capacity under nutrient-rich conditions.

Through this mapping procedure, a substrate availability value of $s = 3.6$, commonly used as a reference point in [3], was found to correspond to a glucose uptake rate of $-15.8 \text{ mmol.g}_{DW}^{-1}.\text{h}^{-1}$. Although this value does not derive from direct experimental calibration, it was pragmatically selected because it leads to growth rates consistent with physiological observations reported by [257] for *E. coli* growing on glucose in minimal media. Thus, this reference ensures that the model predictions remain anchored to biologically relevant operating conditions.

5.4.2.2 Engineered Pathway Integration

From the metabolic point of view, considering an engineered metabolic pathway to produce p-Coumaric acid requires extending the metabolic network of the host cell. In this study, the *E. coli* core metabolic model from [259] is used, which consists of 1805 internal metabolites and 2583 metabolic reactions, including the metabolism of L-tyrosine.

To enable p-Coumaric acid production, the network is expanded by adding the following synthetic reaction, representing the deamination of L-tyrosine into p-Coumaric acid, ammonia and a proton:



where tyrL represents the metabolite L-tyrosine, pC is the p-Coumaric acid and NH_4^+ is the ammonium cation. Additionally, the corresponding exchange reactions for the secretion of p-Coumaric acid are introduced to maintain mass balance and ensure proper integration with the existing network. This network modification is achieved using the COBRA Toolbox in MATLAB [178].

Because this synthetic reaction is catalyzed by an exogenous enzyme, i.e., tyrosine ammonia lyase TAL, the corresponding intracellular flux must be constrained by the enzyme's catalytic capacity. Specifically, the flux through the engineered reaction is limited by:

$$v_{TAL}^{ub} = k_{cat,TAL} [TAL] \quad (5.10)$$

where $k_{cat,TAL}$ is the turnover number of TAL, and $[TAL]$ is the intracellular concentration of the enzyme. The enzyme concentration is dynamically predicted by the burden-aware gene expression model as a function of substrate availability s , as well as the promoter strength and ribosome binding site (RBS) strength controlling the expression of the exogenous gene.

By integrating the synthetic pathway in this way, the constraint-based metabolic model is expanded to capture the new production capability while maintaining a mechanistic coupling between metabolic fluxes, enzyme availability, and gene expression burden. This coherent framework allows for realistic simulation of p-Coumaric acid biosynthesis while accounting for the physiological impact of exogenous enzyme expression.

5.4.2.3 Exogenous Gene Expression

From the genetic point of view, the integration of an engineered metabolic pathway not only modifies the metabolic network structure but also necessitates the addition of the corresponding exogenous genes encoding the required enzymes. In this study, the exogenous pathway involves a single enzyme TAL, which catalyzes the conversion of L-tyrosine into p-Coumaric acid. The burden induced by the expression of this enzyme on the host's physiology must be carefully accounted for.

For this purpose, the open-source modeling tool **ONEMODEL** [260] is used. The latter is a modular, text-based framework designed to facilitate the construction and extension of SBML-compatible biological models, focusing on accessibility, simplicity, and modularity. It provides a flexible platform to build models incrementally, making it particularly suitable for expanding existing host-aware models by introducing exogenous gene expression dynamics.

The original host-aware gene expression model is implemented with OneModel and represents the unmodified wild-type strain, with all endogenous gene parameters calibrated against experimental data. In this work, this baseline is extended by introducing the dynamics of the exogenous TAL enzyme. Specifically, the model is augmented by adding new ordinary differential equations (ODEs) that describe the synthesis, translation, and dilution of the TAL mRNA and protein.

To define the exogenous gene in OneModel, several parameters must be specified:

- The protein length $l_{p_{TAL}}$;
- The mRNA production rate ω_{TAL} linked to promoter strength;
- The copy number of protein-coding exogenous genes N_{TAL} ;
- The ribosome binding kinetics parameters K_b^{TAL} and K_u^{TAL} , characterizing the RBS strength.

The parameters for endogenous gene expression are inherited from the original burden-aware model without modification, ensuring consistency across simulations. Meanwhile, the specific properties of the TAL enzyme are identified either from the literature or biological databases such as BRENDA. By capturing the dynamics and resource demands of exogenous protein production, the model ensures a faithful representation of the interplay between synthetic pathway expression and host cell growth.

5.4.3 Numerical Results

Following the **Algorithm 2** presented in Sec. 5.3.3.2, the first step consists of selecting a value for the substrate availability parameter s . In this study, the quantity s is chosen equal to 3.6, a reference value for which the burden-aware model developed in [3] best aligns with experimental physiological data. This value corresponds approximately to typical *E. coli* growth conditions on glucose minimal medium.

Once the substrate availability is fixed, it is necessary to define the genetic parameters controlling the transcription and the translation of the exogenous enzyme. Fig. 5.6 illustrates the evolution of key cellular variables at steady-state — namely, the mass fractions of the cellular components, the growth rate, and the concentration of the exogenous protein — as a function of the promoter strength.

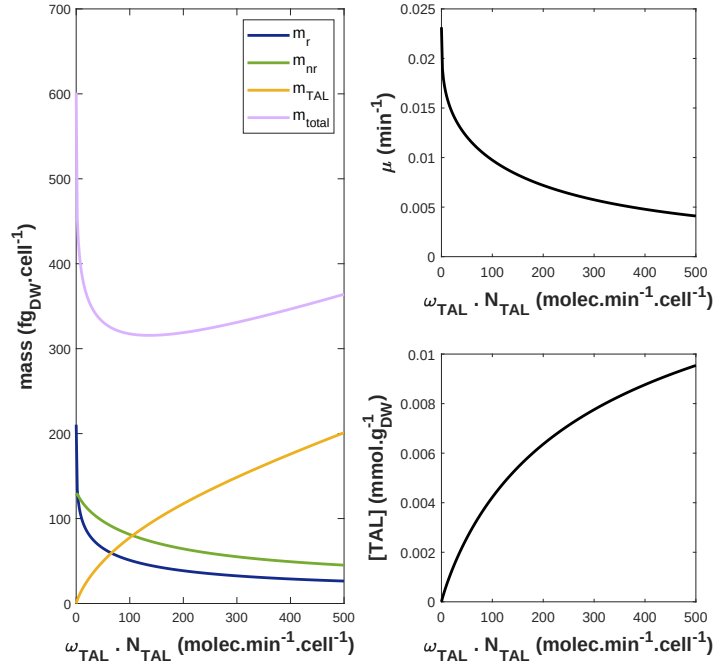


Figure 5.6: Evolution of the mass fractions of the cell, the growth rate, and the concentration of the exogenous enzyme, at steady-state, for different values of the promoter strength controlling the expression of the TAL enzyme.

As expected, increasing the promoter strength, which controls the transcription rate of the exogenous gene, results in a higher steady-state concentration of the TAL enzyme but induces a progressive reduction in cell growth rate due to the expression burden. This trade-off arises because stronger promoter activity sequesters a larger fraction of the available ribosomes for the synthesis of the exogenous proteins, thereby limiting the resources available for the production of endogenous ribosomal and non-ribosomal proteins. Therefore, as promoter strength increases, cellular resource allocation shifts toward the exogenous pathway, ultimately slowing down cellular proliferation.

A practical strategy is to select, as an initial guess, a promoter strength such that the doubling time of the bacteria remains acceptable despite the expression burden. Subsequently, a more refined optimization can be performed: an optimal promoter strength is sought to ensure that the mass of the exogenous protein pool does not exceed approximately 40% of the total host cell mass in steady-state, a threshold commonly cited to preserve cell viability and function [227, 255]. In this study, only the promoter strength is optimized, while the RBS strength is fixed at a biologically realistic value.

Once the promoter strength is obtained, the concentration of the exogenous enzyme can be computed using its predicted mass fraction, its molar mass, and the total host cell mass provided by the burden-aware model. This enzyme concentration is then used to calculate the upper bound of the flux through the engineered pathway, according to Eq. 5.10. Table 5.1 summarizes the main genetic parameters used for optimal promoter strength, including essential characteristics such as the enzyme length l_{pTAL} and its molar mass M_{TAL} .

Table 5.1: Parameter values and constants

	Value	Unit
l_{pTAL}	531	aa
M_{TAL}	57'895	Da
$(\omega_{TAL} N_{TAL})^{opt}$	234	molec.min ⁻¹ .cell ⁻¹
μ_{min}	0.0066	min ⁻¹
m_h	322.3596	fg _{DW} .cell ⁻¹
$[TAL]^{opt}$	0.0069	mmol.g _{DW} ⁻¹
$k_{cat,TAL}$	1.2	min ⁻¹
ν_{Glc}	-15.8	mmol.g _{DW} ⁻¹ .h ⁻¹
K_b^{TAL}	4.7627	cell.min ⁻¹ .molec ⁻¹
K_u^{TAL}	119.7956	min ⁻¹

At the optimal expression level determined, the minimal growth rate is 0.0066 min⁻¹, corresponding to a bacterial doubling time of roughly 1.75 hours. This moderate slow-down relative to wild-type growth rates remains within the expected range for engineered *E. coli* strains subject to a significant but tolerable burden. The steady-state concentration of the TAL enzyme is predicted to be 0.0069 mmol.g_{DW}⁻¹, a level consistent with the expression of key metabolic enzymes or strongly expressed recombinant proteins in synthetic biology applications. Previous studies have shown that engineered *E. coli* systems can achieve recombinant protein expression levels corresponding to several hundred thousand up to millions of molecules per cell, especially when using strong promoters and optimized expression systems [261, 262]. The optimized transcriptional activity, quantified by $(\omega \cdot N)_{TAL}$, reaches 234 mRNA molecules produced per minute per cell, indicating a moderately strong promoter that efficiently drives expression without critically overloading the host's transcriptional machinery. These results confirm that the selected promoter and RBS strengths achieve a biologically coherent trade-off between sufficient enzyme availability for catalytic demands and acceptable host cell viability and growth performance. Notably, the relatively moderate RBS strength chosen (via K_b and K_u) likely contributes to controlling ribosome allocation, preventing extreme competition for translational resources despite the high-level expression of the exogenous pathway.

Having established a biologically coherent expression setting, the corresponding enzyme concentration can now be used to constrain the synthetic flux, enabling an analysis that reflects the true catalytic potential of the engineered pathway. To perform flux variability analysis (FVA), the specific glucose uptake rate ν_{Glc} is obtained from the substrate value s using the regression mapping established previously. The growth rate μ is provided by the burden-aware gene expression model, while the upper bound of the synthetic flux is computed as 0.4965 mmol.g_{DW}⁻¹.h⁻¹ using Eq. 5.10.

To assess the merits of the conciliation strategy, a comparative FVA is performed without imposing any enzyme-based upper bound on the synthetic flux. In this case, the feasible flux interval for the TAL-catalyzed reaction is found to be [0 ; 4.8483] mmol.g_{DW}⁻¹.h⁻¹, indicating a larger degree of uncertainty. In contrast, incorporating enzyme availability constraints dramatically narrows the feasible range to [0 ; 0.4965]

$\text{mmol.g}_{DW}^{-1}.\text{h}^{-1}$, reflecting the realistic catalytic capacity of the expressed TAL enzyme. This significant reduction in the solution space, illustrated in Fig. 5.7, highlights the importance of coupling gene expression burden with constraint-based metabolic modeling.

It is worth emphasizing that although the results are presented at steady-state, the underlying genetic model incorporates dynamic formulations for gene expression. Based on the assumptions made – particularly rapid mRNA and protein turnover relative to the timescale of metabolic adaptation – the system quickly converges to steady-state conditions. Therefore, presenting steady-state results is fully justified and reflects the biological timescales of interest.

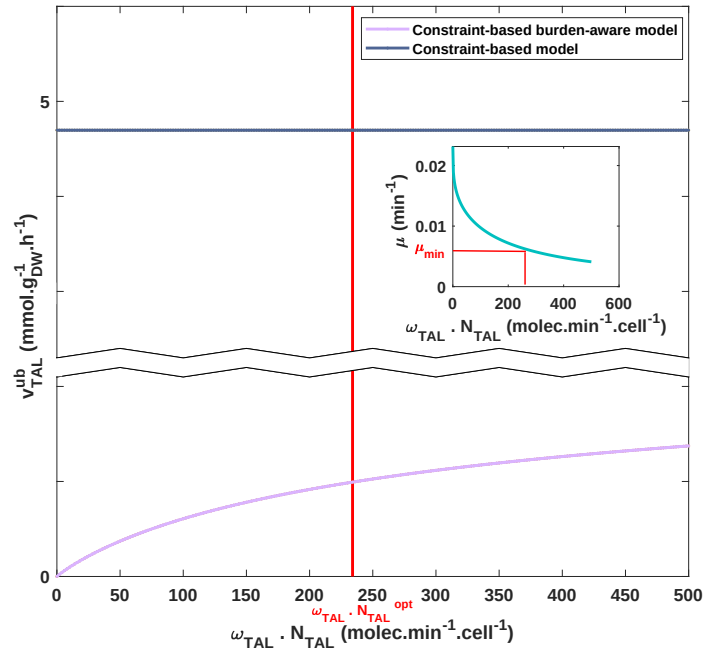


Figure 5.7: Evolution of v_{TAL}^{ub} for different values of the promoter strength of the exogenous protein TAL. FVA results from the classical constraint-based modeling framework are depicted in blue and outcomes from the constraint-based burden-aware modeling framework are shown in purple.

For the sake of clarity, the intervals of possible values for other metabolic fluxes are not presented in this work. Because the conciliation procedure constrains only the flux through the engineered synthetic pathway, the feasible flux intervals for endogenous reactions remain comparable to those obtained by standard constraint-based analysis. However, Fig. 5.7 highlights that combining both constraint-based metabolic modeling with burden-aware gene expression modeling leads to a significantly improved prediction for the production rate of the metabolite of interest, by considerably narrowing the feasible range for the synthetic flux.

To further reduce the solution space for the entire intracellular flux distribution, the existing methods discussed in Sec. 5.1.2, such as the integration of regulatory constraints or thermodynamic feasibility, could be employed. Alternatively, the present procedure could also be extended to model enzyme availability constraints associated with the expression of non-ribosomal endogenous proteins. Although substantial effort would be required to characterize the corresponding genetic parameters, it is most likely that the feasible solution space would be substantially reduced. Addressing this extension constitutes an important direction for future research.

This case study illustrates the practical implementation and benefits of the proposed conciliation strategy between the metabolic and genetic modeling frameworks. By explicitly linking intracellular flux predictions to gene expression dynamics through enzyme availability, the procedure addresses two key challenges in metabolic engineering and synthetic biology: improving the realism of flux predictions at the metabolic level and guiding the selection of the promoter strength at the genetic level to balance productivity and cellular burden. Although the analysis focused on a single engineered reaction, the results demonstrate that integrating constraint-based and burden-aware models enables a more coherent and biologically grounded understanding of synthetic pathway behavior. Then, this integrative framework lays the groundwork for more comprehensive approaches.

5.5 Prospects

This work presented a novel framework to reconcile constraint-based metabolic models with burden-aware gene expression models in the context of synthetic biology. While effective, this approach remains conceptual and offers several avenues for future development, both methodologically and in terms of biological realism.

5.5.1 Alternative Gene Expression Models

The burden-aware gene expression model used in this study was developed by Santos et al. [3] and was applied during a research stay at the Universitat Politècnica de València in collaboration with the team behind its development. The model captures key physiological trade-offs in resource allocation by partitioning the total proteome into ribosomal endogenous, non-ribosomal endogenous, and exogenous protein pools. However, the formulation makes several simplifying assumptions: notably, that transcription is faster than translation (relevant assumption) and that ribosomes are the principal limiting factor in protein expression. In practice, other elements such as RNA polymerases, amino acid availability, tRNAs, chaperones, or cellular energy (ATP, GTP) may also impose significant constraints on gene expression.

Moreover, the model incorporates substrate availability through an empirical and dimensionless parameter s , which must be assigned to real experimental variables through indirect calibration. While this abstraction offers flexibility, it also limits the interpretability and direct experimental validation. Future work could then consider alternative models that use extracellular substrate concentrations or uptake rates as direct inputs, or that include additional layers of regulation (e.g., post-transcriptional effects, stress responses). The integration of models that predict absolute protein abundances or dynamic gene regulatory effects (e.g., through stochastic modeling or hybrid ODE/Bayesian frameworks) could also enhance the mechanistic realism and predictive utility of the framework.

In addition, the genetic stability of engineered constructs is an important aspect that must be addressed. In synthetic biology, long-term maintenance of engineered DNA constructs is a critical design constraint. Cells naturally tend to revert to the wild type because large or highly expressed genetic additions slow growth. Reversion can occur through DNA rearrangements, deletions, or point mutations that inactivate or remove the synthetic pathway. This instability is accentuated when the construct imposes a high metabolic cost, not only through protein synthesis but also through the extra effort of DNA replication itself. To mitigate these effects, stabilizing elements – such as antibiotic-resistance markers with continuous selection or toxin-antitoxin systems – are frequently incorporated. These safeguards add additional metabolic load and must be taken into account in predictive models of gene expression and growth.

Furthermore, the current framework is tailored to prokaryotic systems, where burden-aware gene expression models can explicitly quantify the dynamics of resource allocation, often using mechanistic descriptions of RBS-mediated translation initiation. In contrast, translation initiation in eukaryotic systems is governed by the Kozak sequence, as explained in Sec. 2.5, and involves a scanning mechanism by the ribosome, in addition to more intricate layers of transcriptional and post-transcriptional regulation. These differences present challenges in extending the current formulation to eukaryotic systems. Although burden-aware models for eukaryotes are less developed, some recent efforts have begun to incorporate elements such as promoter variability, splicing, and translation efficiency based on Kozak sequence features [263]. Future work could explore the integration of such models into the proposed framework to expand its applicability to eukaryotic systems in synthetic biology.

5.5.2 Advancing the Conciliation Strategy

In this study, glucose was assumed to be the sole limiting nutrient, allowing a one-to-one mapping between s and the specific glucose uptake rate ν_{Glc} . Although this assumption is reasonable for controlled minimal media, it may not hold under more complex environmental conditions. In real applications, multiple substrates such as oxygen, nitrogen, or phosphate can also limit growth and influence proteome allocation. Extending the mapping strategy to include multiple substrate dimensions would involve defining a multi-input function $s = f(\nu_{Glc}, \nu_{O_2}, \dots)$ and calibrating against multi-omic datasets or bioreactor measurements. Although more complex, this direction would allow the model to better reflect physiological adaptation to diverse conditions.

Furthermore, the current reconciliation strategy relies on aligning growth rate predictions as the common output of both models. Another way to connect these models could be to introduce a coupling constraint that links gene expression and metabolic flux predictions within an iterative optimization framework. The implementation of this strategy could benefit from formal mathematical methods such as dynamic optimization, parameter estimation, or model reduction. Sophisticated integration strategies could also involve combining transcriptomic or proteomic datasets with fluxomic constraints to define a shared state-space across models.

The proposed conciliation strategy is deliberately simple and relies on a fixed mapping and sequential execution of both models. While this ensures accessibility and interpretability, future research could explore more robust and biologically grounded methods to integrate genetic and metabolic models. For example, recent developments have introduced frameworks that estimate fluxes and gene expression levels simultaneously, often by incorporating transcriptomic or proteomic data directly into the constraint system.

Such strategies are exemplified in ME-models (metabolism and expression models) [187] or hybrid formulations that combine FBA with RNA-seq-derived constraints [264, 265]. Another promising line of work involves enhanced flux potential analysis (eFPA) [248] or dynamic constrained allocation FBA (dCAFBA) [249], which use enzyme abundance data or proteome partitioning to simulate dynamic resource allocation under changing growth conditions.

These methods could be adapted or extended to include host-aware or burden-aware concepts, especially in synthetic systems where exogenous gene expression causes strong perturbations. Incorporating feedback from expression burden into FBA objective function, or embedding gene expression models as differential constraints within the FBA problem, would increase integration depth and realism. These extensions may require more computational effort or experimental data, but would substantially improve the biological fidelity of the model.

5.5.3 Extensions to Endogenous Proteins

In this work, enzyme capacity constraints were applied only to the TAL enzyme involved in the engineered p-Coumaric acid pathway. While this already improves flux prediction and narrows uncertainty, a more comprehensive approach would extend catalytic constraints to selected endogenous non-ribosomal enzymes. However, this is not trivial. The current gene expression model aggregates all non-ribosomal endogenous proteins into a single pool, making it difficult to resolve individual enzyme concentrations. Separating this pool into multiple explicit protein classes would require detailed kinetic and expression data that may not always be available or experimentally accessible.

A practical alternative would be to combine the proposed framework with classical enzyme-constrained metabolic models (ecModels) [188], which use fixed enzyme concentrations derived from the literature or proteomics to constrain fluxes. In such a hybrid strategy, dynamically predicted enzyme levels from the burden-aware model could be used for the exogenous pathways, while static capacities would be used for endogenous reactions. This would preserve biological interpretability while avoiding intractable parameter estimation for every gene in the host genome.

5.5.4 Multi-Step Synthetic Pathway

An exciting direction for future work is to expand the current case study beyond p-Coumaric acid to encompass the entire biosynthetic pathway of naringenin, a precursor of flavonoids of industrial and pharmaceutical relevance. As depicted in Fig. 5.8, the extended pathway involves several exogenous enzymes and offers a tractable challenge for evaluating the scaling behavior of the conciliation framework. As additional exogenous proteins are introduced, the model must account for cumulative burden effects, more complex trade-offs in expression tuning, and potential bottlenecks in metabolic flux. Moreover, each added gene would require its own expression parameters and flux constraints, thus testing the robustness of the integrative approach in multi-gene circuit design.

Furthermore, it is worth highlighting the different expression requirements for protein production versus metabolite synthesis. Synthetic designs aimed at overproducing a single protein, such as an antibody or industrial enzyme, differ fundamentally from those intended to generate a small-molecule metabolite. In the first case, very high transcription and translation of one gene product are desirable, so the main burden is

translational. In contrast, engineering a metabolic pathway requires coordinated expression of multiple enzymes to ensure sufficient flux through the pathway, and limitations often arise from cofactor balance or intermediate accumulation, rather than from the abundance of any single protein. Recognizing this distinction is important for setting appropriate expression targets and for evaluating metabolic load in complex synthetic constructs.

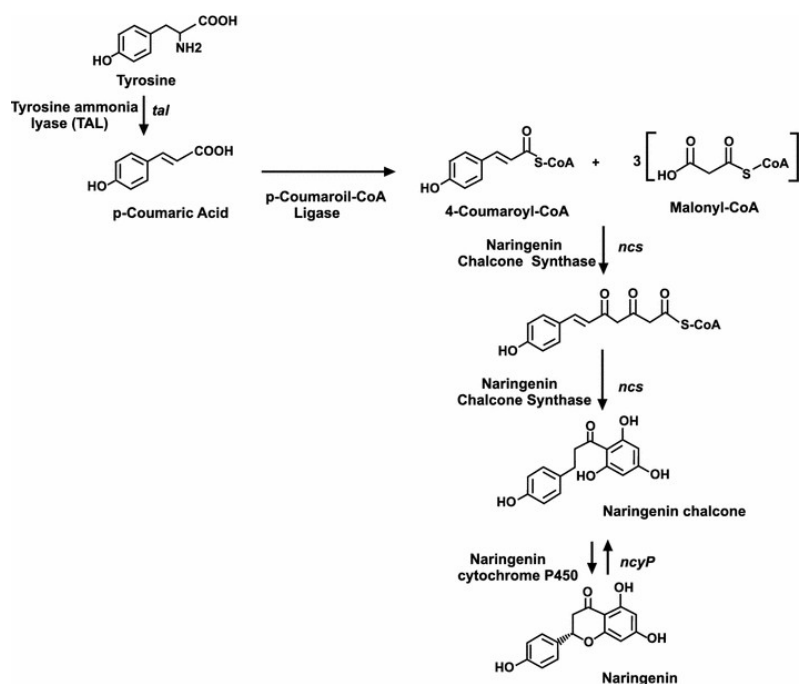


Figure 5.8: Full biosynthetic pathway for Naringenin from L-tyrosine.

5.5.5 Further Considerations

As discussed in Sec. 4.2.3.2.1, models that incorporate ribosome allocation, enzyme kinetics, or proteome constraints must account for the dilution of macromolecules due to growth, particularly when these components are treated as dynamic variables. During exponential growth, enzyme and protein concentrations would decrease unless they are continuously replenished. Therefore, maintaining steady-state concentrations requires synthesis at a rate proportional to the growth rate (i.e., $\mu \cdot [E]$). In such models, dilution becomes a critical factor in determining cellular resource allocation and metabolic capacity.

In mechanistic models such as ME-models [187], this effect is directly incorporated through growth-coupled constraints on macromolecule synthesis. These models require that protein production compensates for dilution explicitly, even though they do not model concentrations via differential equations. In contrast, enzyme-constrained models, such as GECKO [188], use fixed or proteomics-derived enzyme levels and do not explicitly model dilution. Instead, the impact of dilution is typically embedded implicitly in the parameter values or in empirical scaling.

To highlight the difference with the classical MFA framework, which assumes that internal metabolite concentrations are constant over time, we can consider the more general dynamic mass balance:

$$\frac{d\mathcal{C}}{dt} = N\mathcal{v} - \mu\mathcal{C} \quad (5.11)$$

This equation applies to any intracellular compound whose concentration changes due to both reaction fluxes and growth-induced dilution. Depending on the modeling framework, $\mathcal{C}$ may represent metabolite concentrations, macromolecular components such as enzymes and ribosomes, or both. In enzyme- or proteome-constrained models, the dilution of macromolecules is critical and must be balanced through synthesis. For metabolites, especially fast-turnover metabolites, dilution is often negligible, and the quasi-steady-state assumption holds. However, when simulating metabolite concentrations dynamically or modeling slow-turnover species such as cofactors, precursors or signaling intermediates, the dilution term becomes relevant and must be retained.

This formulation in Eq. 5.11 is relevant in dynamic modeling frameworks, such as dynamic enzyme-cost FBA [244], where metabolite concentrations evolve over time and growth-driven dilution must be explicitly balanced. In contrast, steady-state models such as resource balance analysis (RBA) [266] do not simulate internal metabolite dynamics and apply dilution only to macromolecules.

In the present work, dilution is already handled upstream through the gene expression model. This model computes the concentrations of exogenous enzymes dynamically, accounting for both synthesis and growth-dependent dilution. Consequently, when the predicted enzyme levels and growth rate are passed into the FBA module to constrain the fluxes, the effect of dilution is implicitly included. As a result, using the classical steady-state assumption $N\mathcal{v} = 0$ within FBA/FVA remains consistent and justified in the modeling framework proposed in Sec. 5.3.

The dilution term would need to be explicitly included in the mass balance in scenarios where: **(i)** endogenous enzyme concentrations are treated as dynamic variables, **(ii)** the growth rate is not externally provided but derived from the metabolic model itself, or **(iii)** intracellular metabolite concentrations are modeled over time. Since none of these apply to the current framework, the simplification is valid and aligns with common practice in synthetic biology modeling.

5.6 Conclusions

The ability to predict intracellular flux distributions with high fidelity is essential for the rational design of microbial cell factories in metabolic engineering and synthetic biology. While classical constraint-based metabolic models have demonstrated their utility in simulating metabolism at the genome scale, they generally operate under the assumption of enzyme sufficiency, ignoring the underlying genetic constraints that govern protein expression. Conversely, gene expression models capture regulatory and translational dynamics but are typically disconnected from the biochemical demands of metabolism. Bridging these two domains, i.e., metabolic and genetic modeling, is thus a critical step toward more realistic, resource-aware simulation frameworks that can account for the burden imposed by synthetic constructs.

This issue is particularly important in synthetic biology, where the deliberate introduction of exogenous genes rewires host metabolism and places additional demands on

limited cellular resources. The over-expression of non-native enzymes not only consumes amino acids and energy but also sequesters transcriptional and translational machinery, perturbing the host's physiology in ways that are not reflected in standard flux balance analysis. Therefore, integrating metabolic models with gene expression dynamics is not only methodologically meaningful but also biologically indispensable for designing efficient and viable synthetic circuits.

In this chapter, we proposed a simple yet effective conciliation strategy to connect a burden-aware gene expression model with a constraint-based metabolic model. The integration was achieved by mapping an abstract substrate availability parameter s , used in the gene expression framework, to the specific glucose uptake rate ν_{Glc} used in the metabolic model. The growth rates were matched under identical substrate conditions, ensuring physiological consistency. Enzyme concentrations were then dynamically estimated from the gene expression model and used to constrain the corresponding fluxes in the metabolic model via catalytic capacity constraints. This formulation enabled intracellular flux predictions to reflect both biochemical feasibility and gene expression burden.

Beyond basic integration, the proposed method was extended to guide the optimization of genetic control parameters – specifically, the promoter and ribosome binding site (RBS) strengths associated with exogenous gene expression. By imposing biological constraints (e.g., maintaining the exogenous protein mass fraction below 40% of total cell mass), we were able to determine optimal expression levels that maximize pathway productivity while preserving cellular growth. This extended strategy enables the model to not only predict fluxes, but also inform genetic circuit design.

The framework was applied to a case study involving the biosynthesis of p-Coumaric acid in *Escherichia coli*, catalyzed by the exogenous enzyme TAL. The results confirmed the consistency and utility of the approach: the predicted upper bound of the synthetic flux was reduced from $4.8483 \text{ mmol.g}_{DW}^{-1}.\text{h}^{-1}$ (without any catalytic constraint) to $0.4965 \text{ mmol.g}_{DW}^{-1}.\text{h}^{-1}$ when enzyme availability was taken into account via the burden-aware model, thus improving the predictive power of the model by nearly an order of magnitude. Furthermore, the corresponding genetic parameters (e.g., transcriptional activity, growth rate, enzyme concentration) were found to be biologically plausible in agreement with empirical ranges observed in literature for engineered *E. coli* strains.

Despite the promising results, several directions for improvement remain. One particularly important extension would be to integrate the current burden-aware strategy with enzyme-constrained models (ecModels) [188] to impose catalytic limits not only on exogenous reactions, but also on selected endogenous enzymes. While this would require additional parameterization efforts and possibly proteomics data, such hybrid models would better capture the trade-offs between global proteome allocation and metabolic function.

In summary, the modeling approach presented in this chapter provides a flexible and biologically grounded framework to reconcile metabolic and genetic layers in synthetic systems. It offers dual benefits: **(i)** at the metabolic level, it improves the prediction of intracellular fluxes by incorporating enzyme availability and reducing the solution space of feasible flux distributions; **(ii)** at the genetic level, it provides an original way to tune the promoter and ribosome binding site strengths, which are crucial parameters for maximizing the flux through an exogenous pathway and thus enhancing the production rate of the target metabolite.

Taking into account the gene expression burden and enzyme allocation constraints, this work contributes to a pragmatic and modular strategy for connecting two traditionally disconnected modeling frameworks. The proposed method blends into the current landscape of metabolic and systems biology while paving the way for further research into integrated, resource-aware models. It illustrates the importance and feasibility of coupling metabolic performance with genetic design, reinforcing the idea that predictive synthetic biology must treat metabolism and gene expression not as isolated domains but as interdependent systems.

Chapter 6

Conclusion and Prospects

The modeling of cellular culture processes undertaken in this thesis spans three levels of description (i.e., macroscopic, microscopic, and genetic-integrative) reflecting the multi-scale nature of cellular systems, as depicted in Fig. 6.1. By progressively increasing the level of detail, from extracellular behavior to intracellular metabolic pathways and finally to genetic regulatory effects, we have built an increasingly detailed representation of the cell. This concluding chapter summarizes the motivations, objectives, methods, and key results at each modeling level and then provides an integrative synthesis across these axes. Sec. 6.5 outlines future research directions across scales, highlighting methodological innovations, integration challenges, and potential application domains. Finally, we stress the significance of mathematical and metabolic modeling in biotechnology and systems biology, noting that the contributions of this work have been validated by several peer-reviewed publications.

6.1 EFM-Derived Macroscopic Modeling

Macroscopic models characterize cell cultures by focusing on measurable extracellular variables such as biomass growth, substrate consumption, and product formation. Such models are indispensable in bioprocess engineering for predictive control and optimization, offering simplicity and computational efficiency for applications such as scale-up and real-time control. However, traditional unstructured models lack a direct connection to intracellular metabolism, limiting their interpretability. The goal of our work was to bridge structured and unstructured paradigms by deriving macroscopic kinetic models that remain grounded in metabolic knowledge. In particular, our goal was to address two challenges: **(i)** the combinatorial explosion of elementary flux modes (EFMs) when connecting macro dynamics to metabolic networks, and **(ii)** the difficulty of kinetic rate law identification.

We developed a systematic workflow to derive macroscopic bioreaction models from metabolic networks. First, a novel EFM selection/reduction procedure was introduced to drastically reduce the number of elementary vectors and identify a representative subset of metabolic pathways without enumerating the entire set of EFMs (which can be prohibitively large). This procedure used cosine similarity metrics and optimization-based criteria to retain the most informative EFMs, effectively deducing a reduced reaction scheme for cell culture. Second, we implemented a modeling framework for reaction kinetics, integrating mechanistic equations with data-driven machine learning components. In practice, this meant using classical rate laws (e.g., Monod-type or mass-action kinetics) when applicable and employing artificial neural networks (ANNs) – specifically

multilayer perceptrons – to model rate relationships that were too complex or poorly understood for first-principles modeling. This hybrid approach combined mechanistic modeling with data-driven components to capture both well-characterized and complex system behaviors, leading to macroscopic models that were adaptable and biologically consistent.

The macroscopic modeling strategy was applied to mammalian cell cultures representative of animal cells widely used in biotechnology. These case studies demonstrated typical bioproduction scenarios (e.g., a fed-batch culture producing a therapeutic protein) and provided a realistic context for evaluating the hybrid modeling approach. By focusing on animal cells, which have complex metabolism and regulation, we ensured that our method could handle challenging systems of industrial relevance.

Our macroscopic modeling framework successfully produced reduced metabolic models with high predictive power for the chosen case studies. The EFM selection algorithm drastically reduced the network complexity while preserving key pathways, thus overcoming the combinatorial explosion problem. The resulting macroscopic reaction schemes captured essential metabolic conversions between substrates and products. For kinetic modeling, mechanistic models performed well in conditions where classical kinetics could be applicable, while ANN components provided significantly better predictions in more complex scenarios. This adaptive strategy proved its value: in cases where traditional rate laws failed to capture the system’s behavior, the neural models accurately reproduced the data, leading to improved overall predictive performance. In summary, we demonstrated a viable path to derive macroscopic models from EFMs, validated on animal cell cultures. The hybrid EFM-ANN models achieved a reliable simulation of growth and production dynamics, illustrating the practical applicability of the method in the optimization and control of bioprocesses. These findings demonstrated the benefit of incorporating mechanistic insights into data-driven macroscopic models and showed that considering metabolic information improved model fidelity, even at a simplified scale.

6.2 Microscopic Metabolic Modeling

While macroscopic models are valued for their simplicity, they abstract away intracellular details. Microscopic modeling delves into the cell’s internal metabolic network, offering a detailed, mechanistic portrayal of intracellular fluxes, metabolite concentrations, and regulatory interactions. Such detailed models are crucial for understanding the behavior of the system under changing conditions and for investigating the impact of genetic or environmental perturbations at the pathway level. However, genome-scale metabolic models (GEMs) are often too large and unwieldy to integrate into dynamic simulations or optimization routines, whereas small models may omit critical biological detail. Our work therefore set out to develop intermediate-scale metabolic models, often referred to as mid-sized networks, that strike a balance between biological realism and tractability. The objective was to establish a constructive, step-by-step methodology for building such metabolic networks from scratch (or from minimal prior information), guided by fundamental biochemical knowledge, and iteratively refined using computational tools.

We introduced a bottom-up network construction framework for metabolic modeling. The process begins by identifying the core metabolic pathways that are ubiquitous and essential (glycolysis, the TCA cycle, pentose phosphate pathway, etc.), ensuring

that the backbone of the network covers central carbon metabolism. In addition, we paid special attention to reaction directionality and thermodynamics, using data such as standard Gibbs free energy and known enzyme reversibility to set physiologically realistic directions for each reaction. From this foundation, additional pathways relevant to the organism's specific physiology are incorporated. Throughout the construction, we used constraint-based analysis techniques – including flux balance analysis (FBA), flux variability analysis (FVA), and flux coupling analysis (FCA) – to verify network consistency, detect blocked reactions or stoichiometric inconsistencies, and iteratively refine the network structure. Importantly, our methodology integrated cellular energy metabolism by explicitly modeling energy cofactors and the proton motive force (PMF). Incorporating the PMF allowed us to account for ATP generation, ion balances, and other energy transduction processes, thus improving the mechanistic representation of growth yields and maintenance costs. The outcome of this construction process was a stoichiometrically consistent metabolic network of intermediate size, which could then be parameterized and used in dynamic simulations (for instance, by formulating ordinary differential equations or applying dynamic FBA).

We validated the above methodology with two distinct case studies of photosynthetic microorganisms, chosen for their scientific and application interest. The first was the cyanobacterium *Arthrospira* sp. PCC 8005 (a photosynthetic prokaryote), for which we built a metabolic network entirely from scratch. The second was the microalga *Tisochrysis lutea* (a photosynthetic eukaryote), where we started from an existing partial network. These organisms allowed us to test the method in both a scenario with minimal prior information (requiring full de novo network construction) and one with a starting network (testing the method's ability to refine and expand an initial model). Both were relevant to biotechnology: *Arthrospira* is used in bioproduct formation and as a model organism for phototrophic culture, while *Tisochrysis* is studied for biofuel precursors and aquaculture feed. Furthermore, these case studies allowed the exploration of both steady-state and dynamic conditions: *Arthrospira* was modeled under constant illumination (approximate steady-state), whereas *Tisochrysis* was modeled under day/night light cycles that forced dynamic metabolic adjustments.

The constructive modeling strategy proved successful in generating biologically faithful and experimentally validated metabolic models. For *Arthrospira*, the assembled network included all major pathways needed for autotrophic growth. We validated the model by comparing its steady-state flux predictions with experimental measurements under constant light, finding good agreement for growth rates and uptake/secretion fluxes. Moreover, by simulating different nitrogen sources (e.g., nitrate versus ammonium), the model correctly captured shifts in metabolic flux distributions, demonstrating its ability to reflect physiological changes. For *Tisochrysis*, we expanded a prior network to include detailed carbohydrate metabolism pathways. We then embedded this network into the DRUM (Dynamic Reduction of Unbalanced Metabolism) framework to simulate diurnal cycles where the quasi-steady state assumption (QSSA) does not hold. The model successfully reproduced the observed oscillations in biomass composition (such as carbohydrate/lipid accumulation during the day and consumption at night), indicating that it can handle transient metabolite pooling. Across both case studies, the intermediate-scale models achieved a level of detail to yield insight into intracellular regulatory behavior, but remained manageable enough for rigorous validation and dynamic simulation. This work was, to our knowledge, the first comprehensive step-by-step methodology for elaborating metabolic networks of intermediate size with minimal reliance on genomic reconstructions. It bridged the gap between small, oversimplified models and large genome-scale models. The resulting networks captured key

metabolic features of the organisms while remaining tractable for analysis and control. In doing so, we demonstrated that combining fundamental biological knowledge with constraint-based computational refinements can efficiently lead to models that are predictive and mechanistically interpretable. The approach is generalizable to other cell types. In fact, preliminary applications to other phototrophs and bacteria have been initiated, underscoring its broad utility in systems biology and metabolic engineering.

6.3 Constraint-Based Burden-Aware Modeling

A purely metabolic model, even with detailed features, typically operates under the assumption that the enzymatic machinery is available to support the computed fluxes. In reality, metabolic capabilities are closely related to gene expression, enzyme levels, and cellular resource allocation. In synthetic biology and metabolic engineering, when foreign pathways are introduced, the burden of expressing additional genes can significantly alter host cell metabolism. Classical flux-based models (e.g., obtained from FBA and FVA) often predict a wide range of feasible flux distributions. However, many of these distributions are unlikely to occur *in vivo* due to regulatory constraints that are not captured by the model. To address these issues, our third level of modeling integrates genetic information and regulatory constraints into the metabolic model. The objective was to develop an integrative framework that couples metabolism with gene expression, thereby constraining the solution space of metabolic fluxes based on enzyme availability and enabling direct exploration of how genetic perturbations (like over-expressing a pathway) impact cell physiology. This approach aimed to provide a more integrated representation of cellular function to link genotype to phenotype and to improve the descriptive and predictive power of the model for scenarios involving genetic modifications.

We achieved this integration by combining a constraint-based metabolic model with a burden-aware gene expression model. The metabolic component was a stoichiometric model of *E. coli* central metabolism (including the introduced synthetic pathway of interest) operating under the usual mass balance constraints and considering the optimization of a cellular objective such as maximizing growth. The gene expression component, based on the model introduced in [3], explicitly allocates cellular proteomic resources into different pools: endogenous ribosomal and non-ribosomal proteins, and exogenous proteins encoded by the introduced synthetic genes. This resource allocation model imposes additional constraints on the metabolic model. For example, a limited total enzyme budget means that increasing the flux through a pathway requires dedicating more enzyme to it, which in turn can reduce the capacity for other pathways. In practice, this integration was implemented by aligning the specific growth rate in both models to ensure consistency. Through this coupling, intracellular fluxes are directly constrained by gene expression levels, creating a unified model of metabolism and gene regulation. In addition, we proposed an algorithm to adjust gene expression parameters (such as promoter and ribosome binding site (RBS) strengths for synthetic genes) within the model, allowing exploration of how genetic engineering choices affect metabolic performance.

This constraint-based burden-aware modeling approach was demonstrated in the bacterium *Escherichia coli*, a workhorse of synthetic biology. As a case study, we focused on a strain engineered to produce p-Coumaric acid, a valuable aromatic compound. This scenario involved a synthetic metabolic pathway inserted into *E. coli*'s native network. It provided an ideal test for our framework because the synthetic genes imposed a burden (they draw on the transcription/translation resources and cellular energy) and there

was a clear trade-off between growth and production that the model should capture. All necessary model parameters (e.g., maximum ribosome content, translation rates, etc.) were taken from the literature or estimated from experimental data, ensuring that the integrated model remained grounded in known physiology.

The coupled metabolic-genetic model provided significant improvements in the predictive precision of flux distribution outcomes for the engineered *E. coli* system. By explicitly representing the protein resource limitation, the model narrowed the feasible solution space for metabolic fluxes compared to a standard metabolic flux analysis. Indeed, many flux combinations that would be allowable in a purely metabolic model are eliminated when the gene expression capacity is considered. This result aligned with our expectations that integrating genetic constraints imposes biologically relevant limitations on metabolic behavior, thereby addressing the indeterminacy common in unconstrained MFA. In the p-Coumaric acid case study, the model successfully captured the growth-production trade-off. Increasing the expression of the enzymes involved in the synthetic pathway beyond a certain threshold caused a predicted slowdown in growth due to resource allocation, which coincided with experimentally observed burden effects. Furthermore, the framework enabled *in silico* optimization of gene expression levels. By considering promoter strengths (or other expression factors) as adjustable parameters, we identified an optimal allocation that maximizes p-Coumaric acid productivity while maintaining viable growth. This kind of analysis demonstrated the practical utility of the integrative model for strain design because it can guide genetic engineers on how strongly to express each enzyme in a pathway to achieve a desired production target without significantly compromising the performance of the host cell. More broadly, this work bridged metabolic modeling with gene regulation, marking a step toward comprehensive whole-cell modeling. It showcased the potential of multi-scale models to predict phenotypes emerging from the interplay of metabolism and gene expression, which is particularly relevant for advanced metabolic engineering strategies and synthetic biology circuits.

6.4 Final Conclusion

The three levels of modeling described above form a continuum of refinement in the representation of cellular systems. Each layer builds upon the foundations of the previous one, adding new dimensions of insight. Macroscopic models offer a top-level view suitable for bioprocess-scale analyses. They abstract away intracellular detail but are directly informed by the cell's metabolic capabilities through the use of EFMs. Microscopic metabolic models provide a more detailed representation of the cell, explicitly modeling the network of biochemical reactions and capturing dynamic intracellular phenomena. They require more data and careful curation, but yield deeper mechanistic understanding. Genetic-integrative models further extend microscopic models by incorporating gene expression and regulation, thus completing the picture by linking gene-level changes to metabolic and growth outcomes. Conceptually, these approaches are connected by a common thread: the need to balance model complexity with interpretability and predictive power. As we progressed from macro to micro to genetic-level, we confronted classic trade-off between simplicity and detail. Our strategy was to start simple and gradually incorporate complexity in a controlled, knowledge-driven manner. Each level addresses the limitations of the previous one. For example, the microscopic model overcomes the black-box nature of the macroscopic model by providing an explicit representation of intracellular metabolic processes, and the genetic-integrative model addresses a limitation of the purely metabolic model by adding regulatory realism.

Methodologically, there is a strong transversal synergy among the three axes. All models employ metabolic stoichiometry and flux analysis as a cornerstone. The macroscopic models connect to stoichiometric pathways via EFMs, the microscopic models use stoichiometric matrices for constraint-based analysis, and the gene-integrated models add another layer of constraints on those same stoichiometric fluxes. This means that improvements in one area can benefit others. For instance, a more accurate metabolic network (from the microscopic level) could be used to derive better EFMs for macroscopic modeling, or knowledge of gene expression and regulation (from the genetic level) could inform which pathways are active at the microscopic level. This, in turn, can simplify assumptions in macroscopic models, for example, by filtering out EFMs that involve reactions whose genes are not expressed or are repressed under specific regulatory conditions, thereby identifying context-specific EFMs. In essence, data and knowledge integration is pivotal. As we incorporate additional data types (e.g., extracellular flux measurements for macroscopic models, intracellular metabolite concentrations and flux data for microscopic models, and transcriptomic/proteomic data for genetic models), our models become more descriptive and constrained, improving their predictive accuracy. In particular, integrating genetic information significantly improved the model fidelity, as seen by the reduction of the flux solution space and alignment with observed phenotypes. This underscores that the integration of genetic and metabolic data is crucial for capturing cellular behavior. By integrating these layers, a multi-scale model gains the capacity to more reliably predict the outcomes of genetic perturbations and environmental variability.

From a broader scientific perspective, the multi-level modeling approach employed in this thesis reaffirms the importance of mathematical and computational modeling in modern biosciences. By bridging scales from the bioreactor to the gene, we provide a framework to translate molecular knowledge into macroscopic predictions and vice versa. This is essential for fields like systems biology, where understanding how genotype influences phenotype requires linking regulatory networks to metabolic function, and for bioprocess engineering, where cell-level models inform reactor-level optimization. Moreover, our work demonstrates how combining disciplines (e.g., control theory, machine learning, metabolic engineering, and molecular biology) can yield powerful tools for deciphering complex biological systems. Each level of modeling not only provides specific insights and tools, but also contributes to a cohesive multi-scale understanding of cellular cultures. In summary, this thesis establishes a comprehensive framework for modeling cellular cultures, demonstrating that multi-scale metabolic models can be developed incrementally and rigorously to address both the extracellular behavior of cultures and the internal mechanisms that govern metabolism and regulation.

6.5 Perspectives

Although the progress achieved in this work is significant, there remain numerous avenues for future research to further enhance and integrate multi-scale models of cellular cultures. In this section, we outline key perspectives that span across modeling scales, focusing on methodological innovations, integration challenges, and expanded applications.

One of the main challenges ahead is to integrate the different modeling layers into a coherent multi-scale framework. This means connecting large-scale models, such as those that describe nutrient dynamics and population growth in bioreactors, with de-

tailed models of intracellular metabolism and regulation. Building simulation platforms where these models interact could improve predictions under dynamic conditions. Such an approach would also make it possible to account for variability between cells or spatial differences within a reactor. However, this kind of integration will require addressing differences in timescales between processes and may involve hybrid modeling strategies to manage complexity. Possible strategies include multi-rate solvers that couple fast and slow processes efficiently, surrogate or reduced-order models that replace detailed sub-modules in large simulations, and modular frameworks where subsystems interact through clearly defined interfaces.

Building on the hybrid and machine learning approaches introduced here, future work can explore richer machine learning architectures and physics-informed modeling. For instance, recurrent neural networks or long short-term memory networks could be employed at the macroscopic level to capture temporal correlations in process data, and physics-informed neural networks (PINNs) that integrate metabolic laws into neural architectures might offer a way to enforce mechanistic consistency in data-driven models. However, interpretability remains a key consideration. Although hybrid models incorporate data-driven components, they need not become black boxes. Future work could further strengthen interpretability through sparse or pruned architecture, explainable AI techniques, and the enforcement of mechanistic constraints in data-driven components. In addition, improved algorithms for EFM selection and network reduction are an open field: clustering techniques, probabilistic selection methods, or more efficient extreme pathway solvers (such as the positive-definite approaches in recent studies) could further streamline the derivation of reduced models. On the gene-metabolism integration front, more sophisticated coupling strategies can be developed. Our current approach aligned growth rates between models as a coupling constraint. Future approaches might use multi-objective or bi-level optimization to formally integrate regulatory and metabolic objectives, or incorporate dynamic gene expression models that operate in tandem with metabolic ODE models. Furthermore, exploring alternative gene expression models that account for additional layers of regulation (e.g., transcriptional feedback loops, global stress responses) is another promising avenue, as different formulations (beyond the one covered in [3]) could improve accuracy under varying conditions.

As models become more detailed, the demand for high-quality data grows. Integrating multi-omics data (genomics, transcriptomics, proteomics, metabolomics) into modeling pipelines is a crucial next step. Future research should focus on methods to incorporate such data for model parameterization and validation. For example, using transcriptomic data to conditionally turn reactions on or off in a metabolic model (context-specific network reconstructions), using proteomics to set enzyme capacity constraints, or using metabolomics to validate predicted accumulation profiles. The challenge is to handle data heterogeneity and uncertainty. In addition, advancing high-throughput validation techniques (including ^{13}C -metabolic flux analysis for intracellular fluxes, or single-cell sequencing for coupling to gene expression models) will be important to refine and trust the next generation of models. The payoff is significant. Indeed, a deeply integrated model could, for instance, predict how a change in gene regulatory network (due to a mutation or environmental signal) will ripple through to metabolism and ultimately affect a bioprocess performance metric.

Another important challenge concerns parameter identifiability. As models integrate more layers of complexity (from kinetic laws to ANN-based kinetics and gene expression constraints), the risk of over-parameterization increases. Although this thesis has addressed the issue partially through model reduction, pruning of neural networks, and the

use of sensitivity-based measures (e.g., Fisher information-derived coefficients of variation), a systematic identifiability analysis was beyond its scope. Future work should incorporate established tools from systems identification, such as structural identifiability analysis (e.g., GenSSI, STRIKE-GOLDD) and practical identifiability methods (profile likelihoods, Fisher information), in order to rigorously assess which parameters are recoverable under realistic data conditions. A short overview of these approaches is provided in the Appendix L, which outlines how such tools could be integrated into the present workflow.

Another promising research direction is the systematic treatment of uncertainty. Although this thesis used local sensitivity analysis as a first step, future work should explicitly address the impact of experimental noise, limited datasets, and prior assumptions on model reliability. Approaches such as Bayesian inference, bootstrapping, and global sensitivity analysis could be employed to propagate data uncertainty into parameter distributions and predictive confidence intervals, thereby providing a more robust assessment of model reliability.

The modeling approaches and findings in this thesis open the doors to various application domains. In industrial biotechnology, the macroscopic and multi-scale models can be used to design better feed profiles and control strategies (e.g., model predictive control) that account for metabolic shifts, improving yields and reducing trial-and-error experimentation. In metabolic engineering and synthetic biology, the integrated metabolic-gene models are directly applicable to rational strain design. By identifying bottlenecks and trade-offs, they guide genetic modifications for enhanced production of bio-based chemicals, pharmaceuticals, or biofuels. Additionally, the constructive modeling approach for building intermediate-sized networks can be applied to less-characterized organisms, providing a path toward building metabolic models for new microbial systems in biotechnology. Medical and environmental research could also benefit from these methodologies. For instance, understanding tumor metabolism (where genetic and metabolic networks interplay) or optimizing microbial communities for waste treatment could benefit from multi-scale models similar to those developed in this work. Ultimately, the continued refinement of multi-scale modeling will establish this framework as a standard tool in systems biology, with broad impacts ranging from fundamental understanding of cellular decision-making to the engineering of cells and processes for human benefit.

In conclusion, this thesis has demonstrated the power of combining mathematical modeling with biological knowledge to tackle the complexity of cellular cultures at multiple levels. By developing and validating models from the extracellular scale down to the genetic regulatory scale, we have shown that it is possible to obtain predictive, mechanistically grounded models for complex biological systems. The significance of such multi-scale metabolic models lies in its dual capacity: it advances fundamental insights into cellular function and also delivers practical solutions for biotechnological innovation.

The scientific contributions of this work (covering new modeling frameworks, methodological developments, and validated case studies) have led to multiple publications in peer-reviewed international journals and conferences, underscoring the rigor and impact of the research. These results confirm the scientific robustness of the work and provide a validated and peer-reviewed foundation for future explorations.

As the complexity of biological systems demands more structured and integrative modeling approaches, the strategies developed in this thesis provide a rigorous frame-

work for addressing that challenge through multi-scale modeling. By reinforcing the connection between genotype, metabolism, and phenotype, this work provides a basis for more efficient bioprocesses, more rational strain engineering, and a deeper understanding of life's organization across scales, reaffirming the essential role of mathematical and metabolic modeling in the future of bioscience and bioengineering.

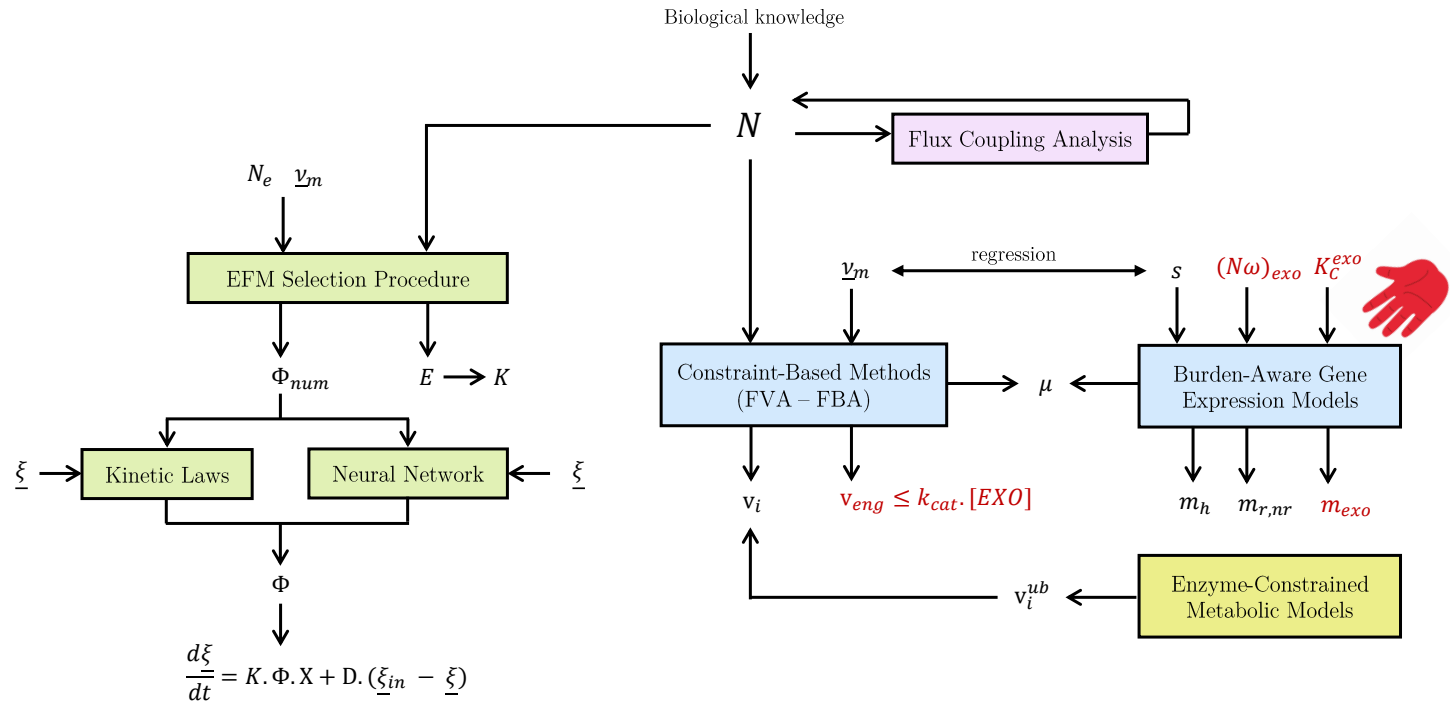


Figure 6.1: Multi-scale nature of cellular culture processes: macroscopic process dynamics, microscopic intracellular metabolism and gene-level regulatory integration.

Appendix A

Systematic Methodology for Bioprocess Model Identification Based on Generalized Kinetic Functions

A.1 Original Method

Based on the definition of a reaction scheme, the general dynamical model of the bioprocess is defined as the system of mass balances of each of the m_e components ξ_i , which can be written as in Eq. 3.5. The systematic methodology developed in [108] provides a structured framework for the selection and estimation of kinetic models in bioprocess engineering. Several kinetic functions can be used to describe the influence of the component ξ_i , which could be substrates (S), products (P), or biomass (X), on the reaction rates $\phi_r(\xi)$. By using the extended version of the Monod law, the specific reaction rates can be expressed as follows:

$$\phi_{r_{Monod}}(\xi) = \phi_{max,r} \prod_{h \in R_r} \frac{\xi_h}{\xi_h + K_{\xi_h,r}} \prod_{l \in P_r} \frac{K_{i_{\xi_l,r}}}{\xi_l + K_{i_{\xi_l,r}}} \quad (\text{A.1})$$

where $\phi_{max,r}$ is the maximum specific rate of the reaction r , $K_{\xi_h,r}$ is the saturation constant associated to ξ_h in reaction r , $K_{i_{\xi_l,r}}$ is the inhibition constant associated to ξ_l in reaction r and R_r and P_r are, respectively, the sets of indices of the components which activate and inhibit the reaction r .

The systematic methodology introduced in [108] is a three-step procedure. The first step involves identifying the activators and inhibitors for each reaction rate ϕ_r . For that, a generalized expression for the reaction rates is considered:

$$\phi_{r_{gen}}(\xi) = \alpha_k \prod_{h \in R_r} \xi_h^{\gamma_{h,r}} \prod_{l \in P_r} e^{-\beta_{l,r} \xi_l} \quad (\text{A.2})$$

where $\alpha_k > 0$ is a kinetic constant, $\gamma_{h,r} \geq 0$ is the activation coefficient of the component ξ_h (activators, e.g., substrates) in reaction r and $\beta_{l,r} \geq 0$ is the inhibition coefficient of the component ξ_l (inhibitors, e.g., products) in reaction r . This structure can be

linearized with respect to its parameters, thanks to a logarithmic transformation:

$$\ln \phi_{r_{gen}}(\xi) = \ln \alpha_k + \sum_{h \in R_r} \gamma_{h,r} \ln \xi_h - \sum_{l \in P_r} \beta_{l,r} \xi_l \quad (\text{A.3})$$

In doing so, it is possible to determine unique initial estimates of the kinetic parameters for the model identification with a simple linear regression.

The **first step** is then a linear least-squares estimation (see Sec. 2.7.1) based on the linearization of the generalized kinetic structure. Since it is a linear regression, the solution is unique and independent of any initial guess. In our case, to model the signals coming from the EFM reduction procedure:

$$\hat{\underline{\theta}}^{(r)} = \min_{\underline{\theta}^{(r)}} \sum_{e=1}^E \sum_{t_s=1}^{t_f} \left(y_{e,t_s}^{(r)} - \underline{x}_{e,t_s}^{(r)T} \underline{\theta}^{(r)} \right)^2 \quad (\text{A.4})$$

where:

$$y_{e,t_s}^{(r)} = \ln \left(\phi_{num,e,t_s}^{(r)} \right) \quad (\text{A.5})$$

$$\underline{x}_{e,t_s}^{(r)T} = [1 \quad \ln \xi_{h_{e,t_s}}^{(r)} \quad \dots \quad -\xi_{l_{e,t_s}}^{(r)} \quad \dots] \quad (\text{A.6})$$

$$\underline{\theta}^{(r)} = [\ln \alpha_k \quad \gamma_{h,r} \quad \dots \quad \beta_{l,r} \quad \dots] \quad (\text{A.7})$$

and under the constraints:

$$[\gamma_{h,r} \quad \dots \quad \beta_{l,r} \quad \dots] \geq 0 \quad (\text{A.8})$$

In the previous equations, $\phi_{num,e,t_s}^{(r)}$ corresponds to the specific reaction rate r obtained numerically from the EFM reduction procedure at a priori defined sampling time t_s of experiment e and $\xi_{i_{e,t_s}}$ corresponds to the measurement of the extracellular component ξ_i at a priori defined sampling time t_s of experiment e (obtained from smoothing splines). Note that, in the context of the work presented in [108], the reaction rates were assumed linear with respect to the biomass concentration to be in concordance with the mathematical formalism associated with the equivalent extended Monod kinetic structure to be identified.

The results of linear least-squares estimation in Eq. A.4 allows for the identification of all potential components ξ_i that could be an influencing factor (activator and/or inhibitor) of reaction r and the estimation of their associated parameter values. This first step is performed in an iterative way by eliminating the regressors in Eq. A.6 associated with factors that do not influence the reaction rate.

With the knowledge of influencing factors and their associated parameter values identified in the first step, the general reaction rate $\phi_{r_{gen}}(\xi)$ is expressed as the product of the influence of each identified factor $\phi_{r_{gen}}^{(\xi_i)}$ such that:

$$\phi_{r_{gen}}(\xi) = a_k \prod_{i \in m_e} \phi_{r_{gen}}^{(\xi_i)} \quad (\text{A.9})$$

To this end, the evolution of each part $\phi_{r_{gen}}^{(\xi_i)}$ is simulated in function of the corresponding influence component ξ_i in its concentration range which has been measured experimentally:

$$\phi_{r_{gen}}^{(\xi_i)} = \xi_i^{\gamma_{i,r}} e^{-\beta_{i,r} \xi_i} \quad (\text{A.10})$$

The equivalent Monod formalism of $\phi_{r_{gen}}^{(\xi_i)}$ can be written as:

$$\phi_{r_{Monod}}^{(\xi_i)} = \delta_{i,r} \frac{\xi_i}{\xi_i + K_{\xi_i,r}} \frac{K_{i_{\xi_i,r}}}{\xi_i + K_{i_{\xi_i,r}}} \quad (\text{A.11})$$

The **second step** consists in the identification of parameter values ($\delta_{i,r}$, $K_{\xi_i,r}$, and $K_{i_{\xi_i,r}}$) associated to the determined Monod equivalent and is performed with a least-squares estimation criterion:

$$\hat{\theta}_{r_{Monod}}^{(\xi_i)} = \min_{\theta_{r_{Monod}}^{(\xi_i)}} \sum_{s=1}^S \left(\phi_{r_{gen},s}^{(\xi_i)} - \phi_{r_{Monod},s}^{(\xi_i)} \right)^2 \quad (\text{A.12})$$

where s is the number of simulated points used for estimation. As outlined in [108], this identification of each Monod equivalent factor can be based on an arbitrarily high number of simulated points using Eq. A.10 and does not involve any measurement. Finally, the global equivalent Monod law of $\phi_{r_{gen}}(\xi)$ can be written as Eq. A.1 with:

$$\phi_{max,r} = a_k \sum_{i \in m_e} \delta_{i,r} \quad (\text{A.13})$$

For this step, since each Monod equivalent factor is a nonlinear function, a multistart strategy – randomly generating initial parameter estimates – is employed to initialize a multidimensional unconstrained nonlinear optimization algorithm (the Nelder-Mead simplex method, using *fminsearch* in MATLAB) in order to cover the broadest range of parameter values as possible.

The **third step** consists in the global identification of the model (identification of the kinetic coefficients together with the initial conditions). This step relies on the complete nonlinear simulation model in Eq. 3.5 and Eq. A.1 and a least-squares estimator using the parameter values found in the second step as initial values for the *fminsearch* optimization algorithm. Thanks to these initial values found in the second step, the risk that the optimization algorithm converges to a local minimum is significantly decreased.

A.2 Revisited Method

In the original method, the reaction rates that we aim to model – whether obtained experimentally or computed numerically – are used only in the first step of the model identification process. These known reaction rates serve as input data to determine only which species act as activators or inhibitors for each reaction. Indeed, the second step is achieved by assuming a generalized kinetic function simulated numerically over a predefined range of species concentrations, while the third step considers the ODE system and a global nonlinear parameter estimation to optimize the kinetic parameters.

In the revisited method that we propose, the known reaction rates are used not only to determine activators and inhibitors for each reaction but also to estimate the parameters of the equivalent Monod structure. Therefore, the **first step** of the methodology remains the same. Then, based on the identified activators and inhibitors, an equivalent Monod structure is directly determined for each reaction and the parameters of this Monod structure are estimated using *lsqnonlin*, a nonlinear least-squares optimization method and a multistart strategy. The **second step** then becomes:

$$\hat{\theta}_{r_{Monod}} = \min_{\theta_{r_{Monod}}} \sum_{e=1}^E \sum_{t_s=1}^{t_f} \left(\phi_{num,e,t_s}^{(r)} - \phi_{Monod,e,t_s}^{(r)} \right) \quad (\text{A.14})$$

Where:

$$\phi_{Monod,e,t_s}^{(r)} = \phi_{max,r} \prod_{h \in R_r} \frac{\xi_{h,e,t_s}^{(r)}}{\xi_{h,e,t_s}^{(r)} + K_{\xi_h,r}} \prod_{l \in P_r} \frac{K_{i_{\xi_l,r}}}{\xi_{l,e,t_s}^{(r)} + K_{i_{\xi_l,r}}} \quad (\text{A.15})$$

In the previous equations, $\phi_{num,e,t_s}^{(r)}$ corresponds to the specific reaction rate r obtained numerically from the EFM reduction procedure at a priori defined sampling time t_s of experiment e and $\xi_{h,e,t_s}^{(r)}$ and $\xi_{l,e,t_s}^{(r)}$ correspond respectively to the measurement of the extracellular component ξ_h and ξ_l (activator or inhibitor) in reaction r at a priori defined sampling time t_s of experiment e (obtained from smoothing splines).

Finally, the **third step**, which involves a global nonlinear parameter estimation using the full ODE system, remains unchanged. This modification allows for a more direct integration of the known reaction rates into the kinetic model while still maintaining the systematic approach of the original methodology.

To validate the revisited methodology, it was first applied to a toy example using in-silico data. The results were highly encouraging, likely due to the underlying kinetic structure following a Monod-like formalism, which aligned well with the modeling assumptions. However, when applied to a real case study using available experimental data, the performance was less conclusive. This discrepancy is attributed to the greater complexity of the actual kinetic structures, which deviate from the Monod-like form. As a result, more flexible modeling approaches, such as neural network architectures, were adopted for the experimental application.

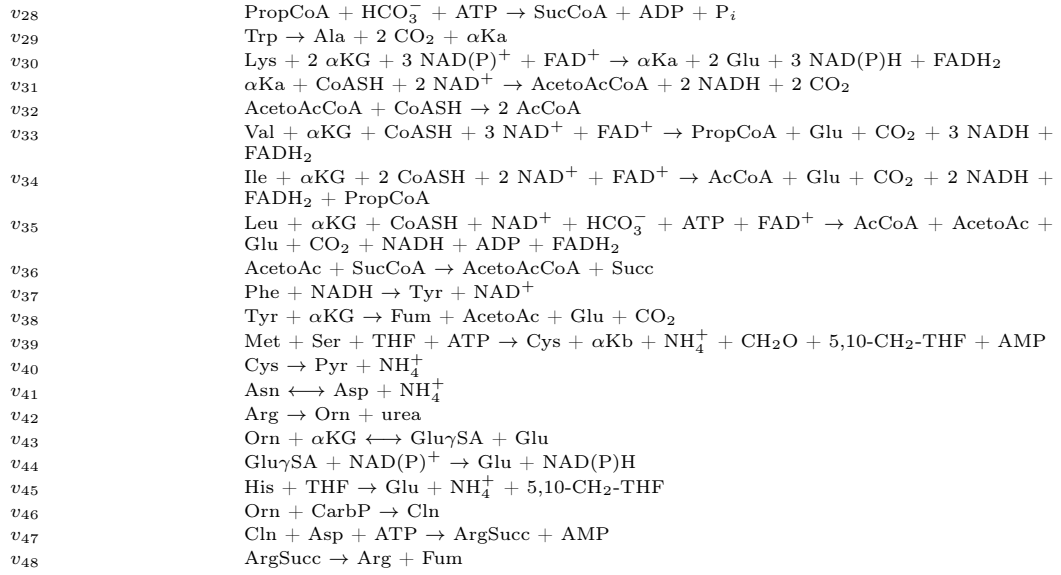
Appendix B

Metabolic Reactions for the Metabolism of Hybridoma Cell Line HB-58

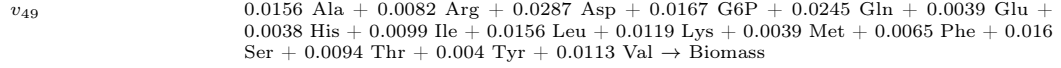
Table B.1: Metabolic reactions for the metabolism of Hybridoma cells.

Flux	Metabolic reaction
Glycolysis	
v_1	$\text{Glc}_{ext} + \text{ATP} \rightarrow \text{G6P} + \text{ADP}$
v_2	$\text{G6P} \longleftrightarrow \text{F6P}$
v_3	$\text{F6P} + \text{ATP} \rightarrow \text{DHAP} + \text{G3P} + \text{ADP}$
v_4	$\text{DHAP} \longleftrightarrow \text{G3P}$
v_5	$\text{G3P} + \text{NAD}^+ + \text{ADP} \longleftrightarrow \text{3PG} + \text{NADH} + \text{ATP}$
v_6	$\text{3PG} + \text{ADP} \rightarrow \text{Pyr} + \text{ATP}$
TCA Cycle	
v_7	$\text{Pyr} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{AcCoA} + \text{CO}_2 + \text{NADH}$
v_8	$\text{AcCoA} + \text{Oxal} + \text{H}_2\text{O} \rightarrow \text{Cit} + \text{CoASH}$
v_9	$\text{Cit} + \text{NAD(P)}^+ \rightarrow \alpha\text{KG} + \text{NAD(P)H} + \text{CO}_2$
v_{10}	$\alpha\text{KG} + \text{CoASH} + \text{NAD}^+ \rightarrow \text{SucCoA} + \text{CO}_2 + \text{NADH}$
v_{11}	$\text{SucCoA} + \text{GDP} + \text{P}_i \longleftrightarrow \text{Succ} + \text{GTP} + \text{CoASH}$
v_{12}	$\text{Succ} + \text{FAD}^+ \longleftrightarrow \text{Fum} + \text{FADH}_2$
v_{13}	$\text{Fum} \longleftrightarrow \text{Mal}$
v_{14}	$\text{Mal} + \text{NAD(P)}^+ \longleftrightarrow \text{Oxal} + \text{NAD(P)H}$
Pyruvate Fates	
v_{15}	$\text{Pyr} + \text{NADH} \longleftrightarrow \text{Lac}_{ext} + \text{NAD}^+$
v_{16}	$\text{Pyr} + \text{Glu} \longleftrightarrow \text{Ala} + \alpha\text{KG}$
Anaplerotic Reaction	
v_{17}	$\text{Mal} + \text{NAD(P)}^+ \longleftrightarrow \text{Pyr} + \text{CO}_2 + \text{NAD(P)H}$
AA Metabolism	
v_{18}	$\text{Glu} + \text{NAD(P)}^+ \longleftrightarrow \alpha\text{KG} + \text{NH}_4^+ + \text{NAD(P)H}$
v_{19}	$\text{Oxal} + \text{Glu} \longleftrightarrow \text{Asp} + \alpha\text{KG}$
v_{20}	$\text{Gln} \rightarrow \text{Glu} + \text{NH}_4^+$
v_{21}	$\text{Thr} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{Gly} + \text{NADH} + \text{AcCoA}$
v_{22}	$\text{Ser} + \text{THF} \longleftrightarrow \text{Gly} + \text{CH}_2\text{O} + 5,10\text{-CH}_2\text{-THF}$
v_{23}	$3\text{PG} + \text{Glu} + \text{NAD}^+ \rightarrow \text{Ser} + \alpha\text{KG} + \text{NADH}$
v_{24}	$\text{Gly} + \text{THF} + \text{NAD}^+ \rightarrow \text{CO}_2 + \text{NH}_4^+ + 5,10\text{-CH}_2\text{-THF} + \text{NADH}$
v_{25}	$\text{Ser} \rightarrow \text{Pyr} + \text{NH}_4^+$
v_{26}	$\text{Thr} \rightarrow \alpha\text{Kb} + \text{NH}_4^+$
v_{27}	$\alpha\text{Kb} + \text{CoASH} + \text{NAD}^+ \rightarrow \text{PropCoA} + \text{NADH} + \text{CO}_2$

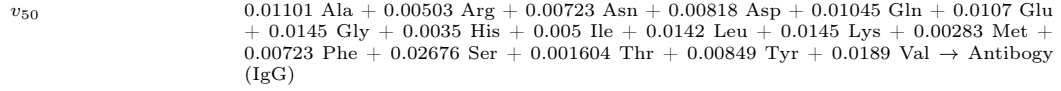
Appendix B – Metabolic Reactions for the Metabolism of Hybridoma Cell Line HB-58



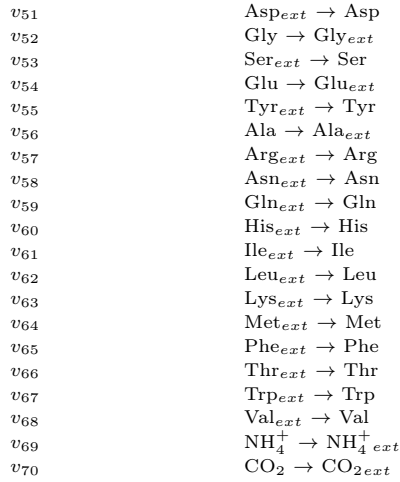
Biomass Synthesis



Antibody Synthesis



Transport Reactions



Appendix C

Metabolic Reactions for the Metabolism of Chinese Hamster Ovary Line CHO-320

Table C.1: Metabolic reactions for the metabolism of Chinese Hamster Ovary cells during the growth phase.

Flux	Metabolic reaction
Glycolysis	
v_1	$\text{Glc}_{ext} + \text{ATP} \rightarrow \text{G6P} + \text{ADP}$
v_2	$\text{G6P} \longleftrightarrow \text{F6P}$
v_3	$\text{F6P} + \text{ATP} \rightarrow \text{DHAP} + \text{G3P} + \text{ADP}$
v_4	$\text{DHAP} \longleftrightarrow \text{G3P}$
v_5	$\text{G3P} + \text{NAD}^+ + \text{ADP} \longleftrightarrow 3\text{PG} + \text{NADH} + \text{ATP}$
v_6	$3\text{PG} + \text{ADP} \rightarrow \text{Pyr} + \text{ATP}$
TCA Cycle	
v_7	$\text{Pyr} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{AcCoA} + \text{CO}_2 + \text{NADH}$
v_8	$\text{AcCoA} + \text{Oxal} + \text{H}_2\text{O} \rightarrow \text{Cit} + \text{CoASH}$
v_9	$\text{Cit} + \text{NAD(P)}^+ \rightarrow \alpha\text{KG} + \text{NAD(P)H} + \text{CO}_2$
v_{10}	$\alpha\text{KG} + \text{CoASH} + \text{NAD}^+ \rightarrow \text{SucCoA} + \text{CO}_2 + \text{NADH}$
v_{11}	$\text{SucCoA} + \text{GDP} + \text{P}_i \longleftrightarrow \text{Succ} + \text{GTP} + \text{CoASH}$
v_{12}	$\text{Succ} + \text{FAD}^+ \longleftrightarrow \text{Fum} + \text{FADH}_2$
v_{13}	$\text{Fum} \longleftrightarrow \text{Mal}$
v_{14}	$\text{Mal} + \text{NAD(P)}^+ \longleftrightarrow \text{Oxal} + \text{NAD(P)H}$
Pyruvate Fates	
v_{15}	$\text{Pyr} + \text{NADH} \longleftrightarrow \text{Lac}_{ext} + \text{NAD}^+$
v_{16}	$\text{Pyr} + \text{Glu} \longleftrightarrow \text{Ala} + \alpha\text{KG}$
PPP	
v_{17}	$\text{G6P} + 2 \text{NADP}^+ + \text{H}_2\text{O} \rightarrow \text{Rbl5P} + 2 \text{NADPH} + \text{CO}_2$
v_{18}	$\text{Rbl5P} \longleftrightarrow \text{R5P}$
v_{19}	$\text{Rbl5P} \longleftrightarrow \text{X5P}$
v_{20}	$\text{X5P} + \text{R5P} \longleftrightarrow \text{F6P} + \text{E4P}$
v_{21}	$\text{X5P} + \text{E4P} \longleftrightarrow \text{G3P} + \text{F6P}$
Anaplerotic Reaction	
v_{22}	$\text{Mal} + \text{NAD(P)}^+ \longleftrightarrow \text{Pyr} + \text{HCO}_3^- + \text{NAD(P)H}$
AA Metabolism	
v_{23}	$\text{Glu} + \text{NAD(P)}^+ \longleftrightarrow \alpha\text{KG} + \text{NH}_4^+ + \text{NAD(P)H}$

Appendix C – Metabolic Reactions for the Metabolism of Chinese Hamster Ovary Line CHO-320

v_{24}	$\text{Oxal} + \text{Glu} \longleftrightarrow \text{Asp} + \alpha\text{KG}$
v_{25}	$\text{Gln} \rightarrow \text{Glu} + \text{NH}_4^+$
v_{26}	$\text{Thr} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{Gly} + \text{NADH} + \text{AcCoA}$
v_{27}	$\text{Ser} \longleftrightarrow \text{Gly}$
v_{28}	$3\text{PG} + \text{Glu} + \text{NAD}^+ \rightarrow \text{Ser} + \alpha\text{KG} + \text{NADH}$
v_{29}	$\text{Gly} + \text{NAD}^+ \rightarrow \text{CO}_2 + \text{NH}_4^+ + \text{NADH}$
v_{30}	$\text{Ser} \rightarrow \text{Pyr} + \text{NH}_4^+$
v_{31}	$\text{Thr} \rightarrow \alpha\text{Kb} + \text{NH}_4^+$
v_{32}	$\alpha\text{Kb} + \text{CoASH} + \text{NAD}^+ \rightarrow \text{PropCoA} + \text{NADH} + \text{CO}_2$
v_{33}	$\text{PropCoA} + \text{HCO}_3^- + \text{ATP} \rightarrow \text{SucCoA} + \text{ADP} + \text{P}_i$
v_{34}	$\text{Trp} \rightarrow \text{Ala} + 2 \text{CO}_2 + \alpha\text{Ka}$
v_{35}	$\text{Lys} + 2 \alpha\text{KG} + 3 \text{NAD(P)}^+ + \text{FAD}^+ \rightarrow \alpha\text{Ka} + 2 \text{Glu} + 3 \text{NAD(P)H} + \text{FADH}_2$
v_{36}	$\alpha\text{Ka} + \text{CoASH} + 2 \text{NAD}^+ \rightarrow \text{AcetoAcCoA} + 2 \text{NADH} + 2 \text{CO}_2$
v_{37}	$\text{AcetoAcCoA} + \text{CoASH} \rightarrow 2 \text{AcCoA}$
v_{38}	$\text{Val} + \alpha\text{KG} + \text{CoASH} + 3 \text{NAD}^+ + \text{FAD}^+ \rightarrow \text{PropCoA} + \text{Glu} + \text{CO}_2 + 3 \text{NADH} + \text{FADH}_2$
v_{39}	$\text{Ile} + \alpha\text{KG} + 2 \text{CoASH} + 2 \text{NAD}^+ + \text{FAD}^+ \rightarrow \text{AcCoA} + \text{Glu} + \text{CO}_2 + 2 \text{NADH} + \text{FADH}_2 + \text{PropCoA}$
v_{40}	$\text{Leu} + \alpha\text{KG} + \text{CoASH} + \text{NAD}^+ + \text{HCO}_3^- + \text{ATP} + \text{FAD}^+ \rightarrow \text{AcCoA} + \text{AcetoAc} + \text{Glu} + \text{CO}_2 + \text{NADH} + \text{ADP} + \text{FADH}_2$
v_{41}	$\text{AcetoAc} + \text{SucCoA} \rightarrow \text{AcetoAcCoA} + \text{Succ}$
v_{42}	$\text{Phe} + \text{NADH} \rightarrow \text{Tyr} + \text{NAD}^+$
v_{43}	$\text{Tyr} + \alpha\text{KG} \rightarrow \text{Fum} + \text{AcetoAc} + \text{Glu} + \text{CO}_2$
v_{44}	$\text{Met} + \text{Ser} + \text{ATP} \rightarrow \text{Cys} + \alpha\text{Kb} + \text{NH}_4^+ + \text{AMP}$
v_{45}	$\text{Cys} \rightarrow \text{Pyr} + \text{NH}_4^+$
v_{46}	$\text{Asn} \longleftrightarrow \text{Asp} + \text{NH}_4^+$
v_{47}	$\text{Arg} \rightarrow \text{Orn} + \text{urea}$
v_{48}	$\text{Orn} + \alpha\text{KG} \longleftrightarrow \text{Glu}\gamma\text{SA} + \text{Glu}$
v_{49}	$\text{Pro} \rightarrow \text{Glu}\gamma\text{SA}$
v_{50}	$\text{Glu}\gamma\text{SA} + \text{NAD(P)}^+ \rightarrow \text{Glu} + \text{NAD(P)H}$
v_{51}	$\text{His} \rightarrow \text{Glu} + \text{NH}_4^+$

Urea Cycle

v_{52}	$\text{Orn} + \text{CarbP} \rightarrow \text{Cln}$
v_{53}	$\text{Cln} + \text{Asp} + \text{ATP} \rightarrow \text{ArgSucc} + \text{AMP}$
v_{54}	$\text{ArgSucc} \rightarrow \text{Arg} + \text{Fum}$

Protein Synthesis

v_{55}	$0.023 \text{ His} + 0.053 \text{ Ile} + 0.091 \text{ Leu} + 0.059 \text{ Lys} + 0.023 \text{ Met} + 0.039 \text{ Phe} + 0.059 \text{ Thr} + 0.014 \text{ Trp} + 0.066 \text{ Val} + 0.051 \text{ Arg} + 0.072 \text{ Gly} + 0.052 \text{ Pro} + 0.032 \text{ Tyr} + 0.078 \text{ Ala} + 0.043 \text{ Asn} + 0.053 \text{ Asp} + 0.019 \text{ Cys} + 0.042 \text{ Gln} + 0.063 \text{ Glu} + 0.068 \text{ Ser} + \text{ATP} + 3 \text{ GTP} \rightarrow \text{Protein} + \text{AMP} + \text{PP}_i + 3 \text{ GDP} + 3 \text{ P}_i$

Nucleotide Synthesis

v_{56}	$\text{R5P} + \text{ATP} \rightarrow \text{PRPP} + \text{AMP}$
v_{57}	$\text{PRPP} + 2 \text{ Gln} + \text{Gly} + \text{Asp} + 4 \text{ ATP} + \text{CO}_2 \rightarrow \text{IMP} + 2 \text{ Glu} + \text{Fum} + 4 \text{ ADP} + 2 \text{ H}_2\text{O}$
v_{58}	$\text{IMP} + \text{Asp} + 2 \text{ ATP} + \text{GTP} \rightarrow \text{ATP}_{RN} + \text{Fum} + 2 \text{ ADP} + \text{GDP}$
v_{59}	$\text{IMP} + \text{Gln} + 3 \text{ ATP} + \text{NAD}^+ + 2 \text{ H}_2\text{O} \rightarrow \text{GTP}_{RN} + \text{Glu} + 2 \text{ ADP} + \text{AMP} + \text{NADH}$
v_{60}	$\text{HCO}_3^- + \text{NH}_4^+ + \text{Asp} + 2 \text{ ATP} + \text{NAD}^+ \rightarrow \text{Orotate} + \text{ADP} + \text{NADH}$
v_{61}	$\text{Orotate} + \text{PRPP} + \text{ATP} \rightarrow \text{UTP}_{RN} + \text{CO}_2 + 2 \text{ ADP}$
v_{62}	$\text{UTP}_{RN} + \text{Gln} + \text{ATP} \rightarrow \text{CTP}_{RN} + \text{Glu} + \text{ADP}$
v_{63}	$0.285 \text{ ATP}_{RN} + 0.285 \text{ UTP}_{RN} + 0.215 \text{ GTP}_{RN} + 0.215 \text{ CTP}_{RN} \rightarrow \text{RNA}$
v_{64}	$\text{ATP}_{RN} \rightarrow \text{dATP}$
v_{65}	$\text{GTP}_{RN} \rightarrow \text{dGTP}$
v_{66}	$\text{CTP}_{RN} \rightarrow \text{dCTP}$
v_{67}	$\text{UTP}_{RN} \rightarrow \text{dTTP}$
v_{68}	$0.285 \text{ dATP} + 0.285 \text{ dTTP} + 0.215 \text{ dGTP} + 0.215 \text{ dCTP} \rightarrow \text{DNA}$

Lipid Synthesis

v_{69}	$\text{DHAP} + \text{NADH} \rightarrow \text{Glyc3P} + \text{NAD}^+$
v_{70}	$\text{Choline} + 18 \text{ AcCoA} + \text{Glyc3P} + 23 \text{ ATP} + 33 \text{ NADH} \rightarrow \text{PC} + 17 \text{ ADP} + 6 \text{ AMP} + 33 \text{ NAD}^+$
v_{71}	$\text{Ethanolamine} + 18 \text{ AcCoA} + \text{Glyc3P} + 23 \text{ ATP} + 33 \text{ NADH} \rightarrow \text{PE} + 17 \text{ ADP} + 6 \text{ AMP} + 33 \text{ NAD}^+$
v_{72}	$\text{PE} + \text{Ser} \rightarrow \text{PS} + \text{Ethanolamine}$
v_{73}	$16 \text{ AcCoA} + \text{Ser} + \text{Choline} + 16 \text{ ATP} + 29 \text{ NADPH} \rightarrow \text{SM} + 2 \text{ CO}_2 + 14 \text{ ADP} + 2 \text{ AMP} + 29 \text{ NADP}^+$
v_{74}	$18 \text{ AcCoA} + 18 \text{ ATP} + 14 \text{ NADPH} \rightarrow \text{Cholesterol} + 6 \text{ CO}_2 + 18 \text{ ADP} + 14 \text{ NADP}^+$
v_{75}	$0.5 \text{ PC} + 0.2 \text{ PE} + 0.075 \text{ PS} + 0.075 \text{ SM} + 0.15 \text{ Cholesterol} \rightarrow \text{Membrane Lipid}$

Appendix C – Metabolic Reactions for the Metabolism of Chinese Hamster Ovary Line CHO-320

Biomass Formation

v_{76} $0.9226 \text{ Protein} + 0.013 \text{ RNA} + 0.0052 \text{ DNA} + 0.0297 \text{ Membrane Lipid} \rightarrow \text{Biomass}$

Transport Reactions

v_{77} $\text{Asp}_{ext} \rightarrow \text{Asp}$
 v_{78} $\text{Cys}_{ext} \rightarrow \text{Cys}$
 v_{79} $\text{Gly} \rightarrow \text{Gly}_{ext}$
 v_{80} $\text{Ser}_{ext} \rightarrow \text{Ser}$
 v_{81} $\text{Glu} \rightarrow \text{Glu}_{ext}$
 v_{82} $\text{Tyr}_{ext} \rightarrow \text{Tyr}$
 v_{83} $\text{Ala} \rightarrow \text{Ala}_{ext}$
 v_{84} $\text{Arg}_{ext} \rightarrow \text{Arg}$
 v_{85} $\text{Asn}_{ext} \rightarrow \text{Asn}$
 v_{86} $\text{Gln}_{ext} \rightarrow \text{Gln}$
 v_{87} $\text{His}_{ext} \rightarrow \text{His}$
 v_{88} $\text{Ile}_{ext} \rightarrow \text{Ile}$
 v_{89} $\text{Leu}_{ext} \rightarrow \text{Leu}$
 v_{90} $\text{Lys}_{ext} \rightarrow \text{Lys}$
 v_{91} $\text{Met}_{ext} \rightarrow \text{Met}$
 v_{92} $\text{Phe}_{ext} \rightarrow \text{Phe}$
 v_{93} $\text{Pro}_{ext} \rightarrow \text{Pro}$
 v_{94} $\text{Thr}_{ext} \rightarrow \text{Thr}$
 v_{95} $\text{Trp}_{ext} \rightarrow \text{Trp}$
 v_{96} $\text{Val}_{ext} \rightarrow \text{Val}$
 v_{97} $\text{Ethanolamine}_{ext} \rightarrow \text{Ethanolamine}$
 v_{98} $\text{Choline}_{ext} \rightarrow \text{Choline}$
 v_{99} $\text{NH}_4^+ \rightarrow \text{NH}_4^+_{ext}$
 v_{100} $\text{CO}_2 \rightarrow \text{CO}_{2_{ext}}$

Appendix D

Experimental Rate Derivation

D.1 Derivation From Smoothing Splines and Differentiation Methods

The general dynamic mass-balance equation in a well-mixed bioreactor under perfusion conditions is expressed as follows:

$$\frac{d\underline{\xi}}{dt} = K \Phi X + D(\underline{\xi}_{in} - \underline{\xi}) \quad (\text{D.1})$$

At each measurement time t_k , the vector of measurable biological uptake and excretion rates of each species is defined as:

$$\underline{\nu}_m(t_k) = \left(\frac{d\underline{\xi}(t_k)}{dt} - D(t_k)(\underline{\xi}_{in}(t_k) - \underline{\xi}(t_k)) \right) \frac{1}{X(t_k)} \quad (\text{D.2})$$

The computation of the vector $\underline{\nu}_m(t_k)$ is obtained by applying smoothing B-splines to the noisy concentration measurements (function *spaps* in MATLAB) and differentiation methods (function *fnder* in MATLAB). Then, it leads to:

$$\underline{\nu}_m(t_k) = K \underline{\phi}(t_k) \quad (\text{D.3})$$

This equation can be written in a matrix form as follows:

$$\mathcal{N} = K \Phi \quad (\text{D.4})$$

where $\mathcal{N}^{m_e \times M}$ is made of the vectors $\underline{\nu}_m(t_k)$ at M measurement times $t_1, t_2, \dots, t_M$.

As shown in Eq. D.2, the computation of the experimental uptake and excretion rates requires the estimation of the first derivative, which is typically obtained by numerical differentiation of noisy data. However, direct numerical differentiation of raw concentration data amplifies noise, leading to unreliable rate estimates. To mitigate this issue, B-smoothing splines are applied to raw concentration measurements using the function *spaps* in MATLAB.

Given a discrete set of noisy measurements (t_k, ξ_i^k) , the smoothing spline $\widehat{\xi}_i(t)$ is obtained by minimizing the penalized least squares criterion:

$$\min_{\widehat{\xi}_i} \left(\sum_k \left(\xi_i^k - \widehat{\xi}_i(t_k) \right)^2 + \lambda \int \left(\frac{d^2 \widehat{\xi}_i}{dt^2} \right)^2 dt \right) \quad (\text{D.5})$$

where λ is the smoothing parameter, controlling the trade-off between data fidelity and smoothness. A larger λ results in a smoother curve, while a smaller λ follows the data more closely but may retain noise.

Then, once the smoothed concentration profile is obtained, the time derivative is computed using the finite difference approximation when time intervals are sufficiently small or are computed analytically from the B-smoothing spline using the function *fnder* in MATLAB. At time t_k , the experimental uptake and excretion rates are given by Eq. D.6 or by Eq. D.7 depending on the time step:

$$\nu_m(t_k) = \left(\widehat{\xi}(t_{k+1}) - \widehat{\xi}(t_k) - D(t_k)(\xi_{in}(t_k) - \xi(t_k)) \right) \frac{1}{X(t_k)} \quad (\text{D.6})$$

$$\nu_m(t_k) = \left(\left. \frac{d\widehat{\xi}}{dt} \right|_{t_k} - D(t_k)(\xi_{in}(t_k) - \xi(t_k)) \right) \frac{1}{X(t_k)} \quad (\text{D.7})$$

In the case study in Sec. 3.3, inflow concentrations are constant over time and are given in Table 3.1. As specified in Sec. 3.3.1, the perfusion phase starts after a batch phase. This means that the dilution rate is null when $t < t_{perf}$ (batch phase), and then the dilution rate is constant over time during the perfusion phase, meaning that $D(t_k)$ is equal to 0.0197 h^{-1} when $t \geq t_{perf}$.

D.2 Derivation From Linear Regression

In a batch culture operating under exponential growth with a constant specific growth rate, the dynamics of cell growth, substrate uptake, and product excretion are described by the following differential equations:

$$\frac{dX}{dt} = \mu X \quad (\text{D.8})$$

$$\frac{dS}{dt} = -\nu_S X \quad (\text{D.9})$$

$$\frac{dP}{dt} = \nu_P X \quad (\text{D.10})$$

Where X represents the biomass concentration, S the substrate concentration and P the product concentration. The parameter μ denotes the specific growth rate, while ν_S and ν_P are the specific uptake and excretion rates, respectively.

The specific growth rate can be estimated by fitting an exponential model to the biomass concentration over time:

$$X(t) = X_0 e^{\mu t} \quad (\text{D.11})$$

To apply ordinary least square (OLS) regression, a logarithmic transformation is applied to linearize the equation. The following equation is obtained:

$$\ln(X(t)) = \ln(X_0) + \mu t \quad (\text{D.12})$$

This allows the specific growth rate μ to be estimated as the slope of a linear regression of $\ln X(t)$ versus time. Using the standard OLS formulation in Eq. 2.12 and defining Γ as the matrix of explicative variables, we can write:

$$\begin{aligned} \underline{y} &= \ln(X(t)) \\ \Gamma &= [\underline{1} \quad t] \end{aligned}$$

Where $\underline{y}$ is the vector of the M measurements (stochastic variables) and $\Gamma \in \mathbf{R}^{M \times 2}$ contains the explicative variables, which are supposed to be perfectly known. The parameters are then estimated as:

$$\hat{\underline{\theta}} = (\Gamma^T \Gamma)^{-1} \Gamma^T \underline{y} \quad (\text{D.13})$$

Yielding:

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} \ln(X_0) \\ \mu \end{bmatrix} \quad (\text{D.14})$$

Similarly, the specific rates of the extracellular species (substrates and products) can be determined from experimental measurements. These rates are assumed to be constant and are calculated by linear regression with respect to biomass. By integrating Eq. D.9 and Eq. D.10, the following linear equations are obtained:

$$S = \frac{-\nu_S}{\mu} X + \frac{\nu_S}{\mu} X_0 + S_0 \quad (\text{D.15})$$

$$P = \frac{+\nu_P}{\mu} X - \frac{\nu_P}{\mu} X_0 + P_0 \quad (\text{D.16})$$

The specific uptake rate (or similarly the specific excretion rate) can then be estimated using linear regression with the following structure:

$$\begin{aligned} \underline{y} &= S(t) \\ \Gamma &= [X(t) \quad \underline{1}] \end{aligned}$$

The corresponding parameter estimates are:

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} \frac{-\nu_S}{\mu} \\ \frac{\nu_S}{\mu} X_0 + S_0 \end{bmatrix} \quad (\text{D.17})$$

The specific uptake rate is then calculated by multiplying $\hat{\theta}_1$ by the specific growth rate estimated previously.

D.3 Uncertainty Estimation of Experimental Rates

The uptake and excretion rates calculated in Eq. D.2 are estimated quantities derived from concentration measurements using smoothing splines and numerical differentiation. As such, they are subject to uncertainty that originates from experimental noise in the concentration data and from the approximation introduced by the spline fitting. Representing this uncertainty is important because it allows confidence intervals to be associated with the rate profiles, thereby providing a more rigorous basis for model calibration.

In practice, confidence intervals on the rates can be obtained by resampling or perturbing the original concentration data, refitting the smoothing splines, and re-estimating the derivatives. A straightforward way to achieve this is through **bootstrap methods**, in which the residuals from the spline fit are resampled repeatedly to generate alternative plausible datasets. For each resampled dataset, the smoothing spline is refitted using the function *spaps* in MATLAB and the corresponding derivatives are recomputed with the function *fnder*. The variability across these bootstrap estimates then provides empirical confidence intervals at each time point. MATLAB facilitates this approach, as the functions *spaps* and *fnder* can be directly embedded in a bootstrap

loop, while resampling can be performed with the built-in function *bootstrap* (Statistics and Machine Learning Toolbox).

Such intervals would also propagate to the model predictions of the measured concentrations, thereby strengthening the statistical interpretation of the results. Although not implemented in the present work, incorporating confidence intervals on uptake and excretion rates represents a valuable extension for future studies and is expected to improve both the transparency and robustness of model-based analyses.

Appendix E

Neural Network Parameter Values

W						b
-4.7461	-0.5569	-0.3064	-0.6936	-2.0738	-0.1662	-2.0046
7.7564	5.4466	-5.0309	-2.0285	2.8096	3.7003	0.0389
5.0274	-0.2005	0.4542	-1.9328	2.9659	2.6954	3.3013

V			c
0.4707	0.7271	0.1896	0.2424
0.9958	0.5455	0.6716	-0.0556
0.4100	0.6717	0.6593	0.0110
0.7971	0.4468	0.2548	0.0999
-0.3652	0.5450	-0.1530	0.0843

Table E.1: Neural network parameter values after reidentification of the biases and weights — colored boxes indicate synaptic connections or biases that were removed following the application of the magnitude pruning algorithm.

Appendix F

Supplementary Quantitative Analysis

While the main text presents a qualitative interpretation of the model results, a quantitative evaluation provides an essential complement. It allows an objective assessment of how closely the model reproduces the experimental data and enables the rigorous comparison of alternative parameter sets or model structures. By coupling qualitative insight with numerical evidence, this Appendix strengthens the overall evaluation of model performance and provides a transparent basis for future model improvements. The performance of the EFM reduction procedure is evaluated by the index Ξ introduced in Eq. 3.2, while the accuracy of kinetic identification and concentration predictions is assessed using standard statistical indicators, namely the root mean square error (RMSE) and the coefficient of determination R^2 .

F.1 Methods and Metrics

Performance Index Ξ

To evaluate the performance of the EFM reduction method, it is proposed to use the indicator Ξ defined by:

$$\Xi = \sum_{k=1}^M \left(K_e \underline{\phi}_k - \underline{\nu}_{m,k} \right) W^{-1} \left(K_e \underline{\phi}_k - \underline{\nu}_{m,k} \right)^T$$

$$\min_{\underline{\phi}_k} \Xi \quad s.t. \quad \underline{\phi}_k \geq 0 \quad (F.1)$$

Recall that the reduction algorithm seeks a small set of EFMs while satisfying positivity constraints and reproducing the experimental uptake and excretion rates $\underline{\nu}_m(t)$. In fact, the indicator measures the weighted sum of squared residuals between the experimental rates and the fluxes predicted by the reduced set of modes. A low value of Ξ indicates that the chosen subset of EFMs reproduces the measured flux data while respecting non-negativity.

- **Model sufficiency:** $\Xi \approx 0$ implies that the reduced macroreaction network is fully consistent with the experimental measurements;
- **Model parsimony:** by monitoring the variation of Ξ as the number of EFMs decreases, it is possible to identify the threshold beyond which further reductions cause unacceptable loss of information.

Root Mean Square Error (RMSE)

The root mean square error (RMSE) measures the average magnitude of prediction errors, expressed in the same physical units as the variable of interest:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (\text{F.2})$$

Where y_i is the experimental value observed at time t_i , $\hat{y}_i$ is the corresponding model prediction, and N is the total number of observations.

Coefficient of Determination R^2

The coefficient of determination R^2 quantifies the proportion of the variance in the experimental data explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (\text{F.3})$$

Where $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ is the mean of the experimental observations, $\sum (y_i - \hat{y}_i)^2$ is the residual sum of squares, and $\sum (y_i - \bar{y})^2$ is the total sum of squares.

A coefficient R^2 of 1 indicates perfect agreement between model and data, whereas values near 0 or negative indicate poor predictive power.

Optional Extensions

Depending on the application, it can be useful to report:

- **Normalized RMSE (NRMSE)**, obtained by dividing the RMSE by the range $\max(y_i) - \min(y_i)$ or by the mean $\bar{y}$, to facilitate comparisons between variables of different scales;
- **Weighted RMSE**, where each residual is multiplied by a weight w_i (e.g., $1/\sigma_i^2$ if measurement uncertainties σ_i are known) to account for varying data confidence or non-uniform time coverage;
- **Global RMSE**, obtained by pooling all residuals across every flux and dataset and applying the standard RMSE formula with the total number of observations N .

Here, the RMSE criterion is applied both to the neural network identification of metabolic fluxes and to the prediction of time-evolution of measured concentrations obtained by integrating the macroscopic model.

F.2 Applications

Application to Hybridoma Cell Line HB-58

Table F.1 and Table F.2 present a quantitative assessment of the pre-filtering and optimization-based reduction steps, respectively. The performance index Ξ measures how well the reduced network – i.e., the retained EFMs together with their numerically computed macroscopic fluxes – reproduces the experimental uptake and excretion rates.

A value of Ξ close to zero indicates excellent consistency. When $\chi = 1$, the complete set of EFMs is retained, giving the best possible reference value Ξ^0 (more details can be found in [5]). Remarkably, the pre-filtering step (using the cosine-similarity algorithm) reduces the set from 21,874 to only 214 EFMs while keeping Ξ unchanged, i.e., still at the best-case level.

Table F.1: Results of the pre-filtering step using the cosine-similarity algorithm. The value Ξ^0 represents the best value of the indicator when all the EFMs are considered without any reduction. More information can be found in our original study in [5].

χ	# EFMs	Ξ	
1.00	21,874	0.0033	Ξ^0
0.99	214	0.0033	

During the optimization-based reduction, the indicator remains essentially at the best-case level down to about 10 EFMs, meaning that the selected EFMs still capture the experimental behavior almost perfectly. Below this point, the index increases reflecting a loss of information as the number of EFMs decreases. Balancing model compactness and fidelity, a final set of modes containing five EFMs is chosen, which provides an acceptable compromise between simplicity and accuracy.

Table F.2: Results of the optimization-based reduction and selection steps.

	# EFMs	Ξ
Optimization-based Reduction	$\Omega = 10$	0.0033
Selection Step	$m_e = 6$	0.0111
	$\Lambda = 5$	0.5056
	4	1.9650

Table F.3 reports the RMSE between measured concentrations and those predicted by integrating the computed macroscopic fluxes, prior to any identification with classical laws or data-driven techniques. The global RMSE of 0.8589 for the direct-validation datasets indicates excellent overall agreement. Cross-validation gives an RMSE of 1.4735, which remains within acceptable limits and supports the predictive ability of the reduced model.

Table F.3: RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification with ANNs) with $n_{EFM} = \Lambda = 5$.

Direct Validation			Cross-Validation
Dataset 1	Dataset 2	Dataset 4	Dataset 3
0.8168	0.4032	1.1791	
0.8589			1.4735

Table F.4 and Table F.5 provide the quantitative assessment of the MLP-based identification of the macroscopic fluxes and their effect on the concentration predictions. Table F.4 shows that the global RMSE on the fluxes decreases steadily from 0.4629 (for two neurons) to 0.0608 (for ten neurons), illustrating the expected gain in fitting capability as network complexity increases.

Table F.4: Global RMSE applied to MLP identification of metabolic fluxes.

# neurons	RMSE
10	0.0608
5	0.1885
4	0.2266
3	0.3267
2	0.4629

Table F.5 presents the RMSE of the corresponding concentration profiles. Although flux deviations are larger when fewer neurons are used, the impact on concentration prediction remains moderate. The global RMSE varies only slightly between three and ten neurons, indicating that increasing the number of neurons beyond three yields limited practical benefit. Considering the trade-off between model parsimony and predictive accuracy, the final MLP architecture was chosen with only three neurons in the hidden-layer.

Table F.5: RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with ANNs.

# neurons	Dataset 1	Dataset 2	Dataset 4	Global RMSE
10	0.8346	0.4119	1.0961	0.8303
5	1.0369	0.5751	1.1295	0.9484
4	0.9318	0.6484	1.3435	1.0130
3	1.0370	0.5869	1.4893	1.0993
2	1.4621	0.7440	1.4039	1.2539

Table F.6 compares three strategies – backpropagation, re-identification, and magnitude pruning – for the final 3-neuron MLP used to identify metabolic fluxes. The global RMSE values are very close, confirming that pruning the network (to some extent) does not degrade predictive accuracy and even slightly improves the fit for some datasets. This robustness further validates the overall identification and reduction procedure.

Table F.6: RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with MLP with 3 neurons in the hidden-layer.

	Dataset 1	Dataset 2	Dataset 4	Global RMSE
Backpropagation	1.0370	0.5869	1.4893	1.0993
Re-identification	1.0734	0.5392	1.4488	1.0861
Magnitude pruning	1.0731	0.5287	1.5028	1.1079

Table F.7 reports the RMSEs obtained when predicting measured concentrations after flux identification with the pruned 3-neuron MLP. The global RMSE increases from 1.1079 (direct validation) to 2.1925 (cross-validation), indicating some expected loss of accuracy on unseen data but overall confirming the robustness of the complete modeling framework.

Table F.7: RMSE applied to the prediction of the measured concentrations after identification of fluxes with a pruned MLP with 3 neurons in the hidden-layer.

Direct Validation			Cross-Validation
Dataset 1	Dataset 2	Dataset 4	Dataset 3
1.0731	0.5287	1.5028	
1.1079			2.1925

Application to Cell Line CHO-320

Tables F.8 to F.11 quantify the effect of each reduction stage on both the performance index Ξ and the RMSE of the concentration predictions. The best-case value Ξ^0 is maintained until the set of EFMs is reduced to $\Omega = 10$, after which it increases as the number of modes decreases. The RMSE analysis shows that even the most compact configuration with $\Lambda_2 = 4$ EFMs retains acceptable predictive capability.

Table F.8: Results of the pre-filtering step using the cosine-similarity algorithm.

χ	# EFMs	Ξ	
1.00	173	$\mathcal{O}(10^{-31})$	Ξ^0
0.995	36	$\mathcal{O}(10^{-31})$	
0.990	34	$\mathcal{O}(10^{-31})$	

Table F.9: Results of the optimization-based reduction and selection steps.

	# EFMs	Ξ
Optimization-based Reduction	$\Omega = 10$	$\mathcal{O}(10^{-31})$
Selection Step	$m_e = 6$	0.0038
	$\Lambda_1 = 5$	0.7518
	$\Lambda_2 = 4$	2.9851

Table F.10: RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification).

# EFMs	Dataset 1	Dataset 2	Global RMSE
173	0.1614	0.1399	0.1510
$\Omega = 10$	0.1614	0.1399	0.1510
$\Lambda_2 = 4$	0.2404	0.2673	0.2542

Table F.11 and Table F.12 provide the quantitative cross-validation of the EFM-based reduction and the subsequent kinetic identification with classical laws. Table F.11 shows that, even for the most compact model with $\Lambda_2 = 4$ EFMs, the global RMSE remains low (0.2542 in direct validation and 0.2078 in cross-validation) confirming the accuracy of the reduced macroscopic model and flux reconstruction.

Table F.11: RMSE applied to the prediction of the measured concentrations by integration of the computed fluxes (before identification) with $n_{EFM} = \Lambda_2 = 4$.

Direct Validation		Cross-Validation
Dataset 1	Dataset 2	Dataset 3
0.2404	0.2673	
0.2542		0.2078

Table F.12 shows that, after identification with classical kinetic laws, the global RMSE increases to approximately 1.31, with only a very small increase to 1.32 under cross-validation. These results indicate that the kinetic model maintains a high predictive ability for unseen data while preserving the simplicity required for bioprocess control applications.

Table F.12: RMSE applied to the prediction of the measured concentrations after identification of the metabolic fluxes with classical kinetic laws.

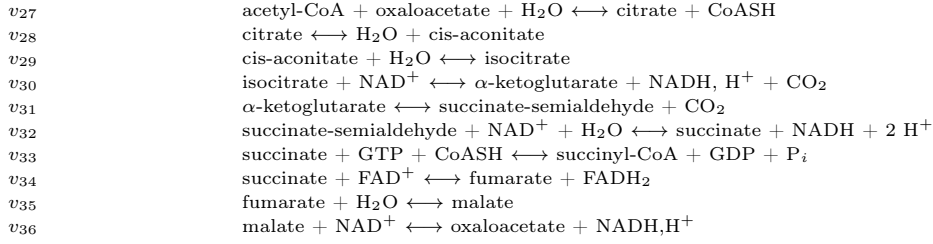
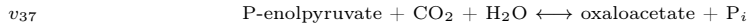
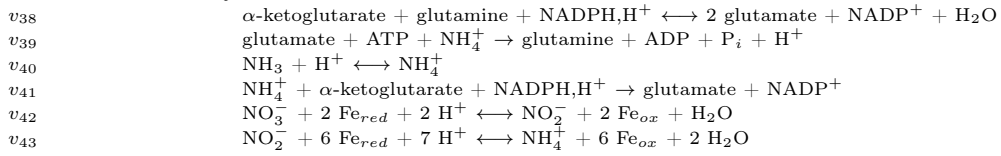
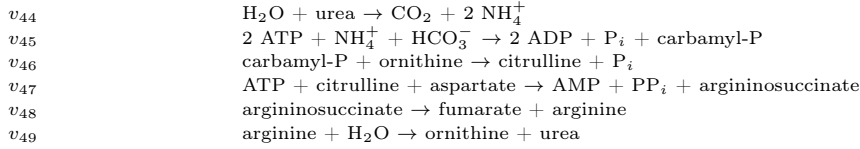
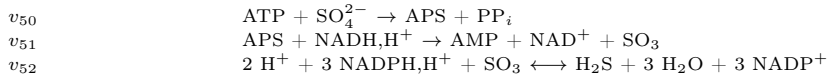
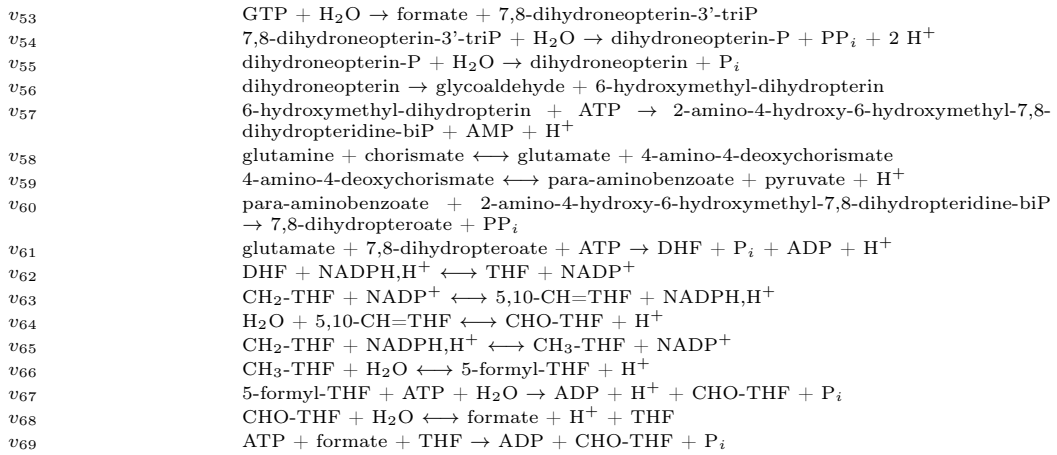
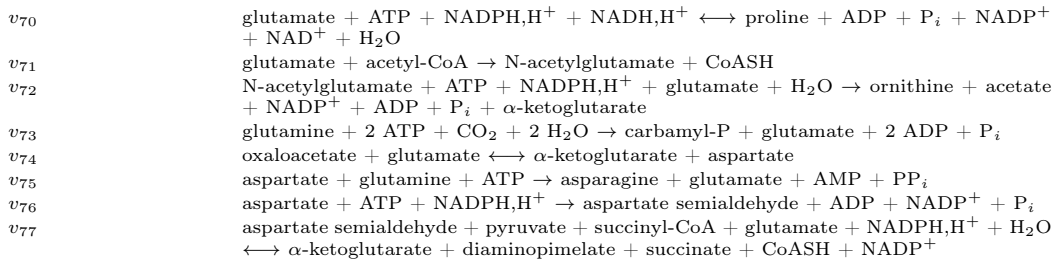
Direct Validation		Cross-Validation
Dataset 1	Dataset 2	Dataset 3
1.3097	1.3088	
1.3093		1.3195

Appendix G

Metabolic Network for Arthrospira Sp. PCC 8005

Table G.1: Metabolic reactions for the metabolism of *Arthrospira spirulina platensis* PCC 8005 (photoautotrophic regime). Protons H^+ involved in most metabolic reactions correspond to H_n^+ .

Flux	Metabolic reaction
Carbon Entry	
v_1	$HCO_3^- + H^+ \rightarrow CO_2 + H_2O$
Calvin Cycle	
v_2	$3 \text{ ribulose-1,5-biP} + 3 CO_2 + H_2O \rightarrow 6 \text{ 3-P-glycerate}$
v_3	$6 \text{ 3-P-glycerate} + 6 ATP \leftrightarrow 6 \text{ 1,3-biphosphoglycerate} + 6 ADP$
v_4	$6 \text{ 1,3-biphosphoglycerate} + 6 NADPH, H^+ \leftrightarrow 6 \text{ glyceraldehyde-3-P} + 6 NADP^+ + 6 P_i$
v_5	$5 \text{ glyceraldehyde-3-P} \rightarrow 3 \text{ ribose-5-P} + 2 P_i$
Glycolysis	
v_6	$\text{glucose} + ATP \rightarrow \text{glucose-6-P} + ADP$
v_7	$\text{glucose-6-P} \leftrightarrow \text{fructose-6-P}$
v_8	$\text{fructose-6-P} + ATP \rightarrow \text{fructose-1,6-biP} + ADP$
v_9	$\text{fructose-1,6-biP} \leftrightarrow \text{glyceraldehyde-3-P} + \text{dihydroxyacetone-P}$
v_{10}	$\text{glyceraldehyde-3-P} \leftrightarrow \text{dihydroxyacetone-P}$
v_{11}	$\text{glyceraldehyde-3-P} + NAD^+ + P_i + ADP \leftrightarrow \text{3-P-glycerate} + NADH, H^+ + ATP$
v_{12}	$\text{3-P-glycerate} \leftrightarrow \text{P-enolpyruvate} + H_2O$
v_{13}	$\text{P-enolpyruvate} + ADP \rightarrow \text{pyruvate} + ATP$
Neoglucogenesis	
v_{14}	$\text{glucose-6-P} + H_2O \rightarrow \text{glucose} + P_i$
v_{15}	$\text{fructose-1,6-biP} + H_2O \rightarrow \text{fructose-6-P} + P_i$
v_{16}	$\text{oxaloacetate} + GTP \rightarrow \text{P-enolpyruvate} + GDP + CO_2$
v_{17}	$\text{pyruvate} + HCO_3^- + ATP \rightarrow \text{oxaloacetate} + ADP + P_i$
PPP	
v_{18}	$\text{glucose-6-P} + 2 NADP^+ + H_2O \rightarrow \text{ribose-5-P} + 2 NADPH, H^+ + CO_2$
v_{19}	$\text{ribose-5-P} \leftrightarrow \text{ribulose-5-P}$
v_{20}	$\text{ribose-5-P} + ATP \rightarrow \text{ribulose-1,5-biP} + ADP$
v_{21}	$\text{xylulose-5-P} \leftrightarrow \text{ribulose-5-P}$
v_{22}	$\text{ribose-5-P} + \text{xylulose-5-P} \leftrightarrow \text{glyceraldehyde-3-P} + \text{sedoheptulose-7-P}$
v_{23}	$\text{glyceraldehyde-3-P} + \text{sedoheptulose-7-P} \leftrightarrow \text{fructose-6-P} + \text{erythrose-4-P}$
v_{24}	$\text{erythrose-4-P} + \text{xylulose-5-P} \leftrightarrow \text{fructose-6-P} + \text{glyceraldehyde-3-P}$
v_{25}	$\text{dihydroxyacetone-P} + \text{erythrose-4-P} \leftrightarrow \text{sedoheptulose-1,7-biP}$
v_{26}	$\text{sedoheptulose-1,7-biP} \rightarrow \text{sedoheptulose-7-P} + P_i$

TCA Cycle

Anaplerotic Reaction

GS-GOGAT Pathway

Urea Cycle-like Pathway

S Fixation

THF Metabolism

AA Metabolism


<i>v</i> ₇₈	diaminopimelate + H ⁺ $\longleftrightarrow$ lysine + CO ₂
<i>v</i> ₇₉	aspartate semialdehyde + NADPH,H ⁺ $\longleftrightarrow$ homoserine + NADP ⁺
<i>v</i> ₈₀	homoserine + ATP + H ₂ O $\rightarrow$ threonine + ADP + P _{<i>i</i>}
<i>v</i> ₈₁	succinyl-CoA + homoserine + cysteine $\longleftrightarrow$ homocysteine + pyruvate + NH ₃ + succinate + CoASH
<i>v</i> ₈₂	homocysteine + CH ₃ -THF $\longleftrightarrow$ methionine + THF
<i>v</i> ₈₃	serine + acetyl-CoA + H ₂ S + H ⁺ $\longleftrightarrow$ cysteine + CoASH + acetate
<i>v</i> ₈₄	serine $\rightarrow$ pyruvate + NH ₄ ⁺
<i>v</i> ₈₅	threonine $\longleftrightarrow$ α-ketobutyrate + NH ₄ ⁺ + H ₂ O
<i>v</i> ₈₆	threonine + NAD ⁺ + CoASH $\rightarrow$ glycine + NADH,H ⁺ + acetyl-CoA
<i>v</i> ₈₇	pyruvate + glutamate + α-ketobutyrate + NADPH,H ⁺ $\longleftrightarrow$ isoleucine + α-ketoglutarate + NADP ⁺ + CO ₂ + H ₂ O
<i>v</i> ₈₈	pyruvate + glutamate $\rightarrow$ alanine + α-ketoglutarate
<i>v</i> ₈₉	2 pyruvate + NADPH,H ⁺ $\longleftrightarrow$ oxoisovalerate + NADP ⁺ + CO ₂ + H ₂ O
<i>v</i> ₉₀	glutamate + oxoisovalerate $\longleftrightarrow$ valine + α-ketoglutarate
<i>v</i> ₉₁	acetyl-CoA + glutamate + oxoisovalerate + NAD ⁺ + H ₂ O $\longleftrightarrow$ leucine + α-ketoglutarate + CoASH + NADH,H ⁺ + CO ₂
<i>v</i> ₉₂	3-P-glycerate + glutamate + NAD ⁺ + H ₂ O $\longleftrightarrow$ serine + α-ketoglutarate + NADH,H ⁺ + P _{<i>i</i>}
<i>v</i> ₉₃	serine + THF $\longleftrightarrow$ glycine + CH ₂ -THF + H ₂ O
<i>v</i> ₉₄	2 P-enolpyruvate + erythrose-4-P + ATP + NADPH,H ⁺ $\rightarrow$ chorismate + ADP + NADP ⁺ + 4 P _{<i>i</i>}
<i>v</i> ₉₅	ribose-5-P + ATP $\rightarrow$ 5'-P-ribosyl-1PP + AMP
<i>v</i> ₉₆	5'-P-ribosyl-1PP + glutamine + serine + chorismate $\longleftrightarrow$ tryptophan + glyceraldehyde-3-P + pyruvate + glutamate + PP _{<i>i</i>} + CO ₂ + 2 H ₂ O
<i>v</i> ₉₇	glutamate + chorismate $\rightarrow$ phenylalanine + α-ketoglutarate + CO ₂ + H ₂ O
<i>v</i> ₉₈	phenylalanine + O ₂ + NADH,H ⁺ $\rightarrow$ tyrosine + NAD ⁺ + H ₂ O
<i>v</i> ₉₉	glutamate + chorismate + NAD ⁺ $\longleftrightarrow$ tyrosine + α-ketoglutarate + NADH,H ⁺ + CO ₂
<i>v</i> ₁₀₀	5'-P-ribosyl-1PP + glutamine + ATP + 2 NAD ⁺ + 3 H ₂ O $\rightarrow$ histidine + α-ketoglutarate + ACR + 2 NADH,H ⁺ + 2 PP _{<i>i</i>} + P _{<i>i</i>}
<i>v</i> ₁₀₁	glycine + NAD ⁺ + THF $\longleftrightarrow$ NH ₄ ⁺ + CH ₂ -THF + NADH,H ⁺ + CO ₂

Nucleotide Salvage Pathway

<i>v</i> ₁₀₂	5'-P-ribosyl-1PP + carbamyl-P + aspartate + NAD ⁺ $\rightarrow$ UMP + NADH,H ⁺ + PP _{<i>i</i>} + P _{<i>i</i>} + CO ₂ + H ₂ O
<i>v</i> ₁₀₃	UMP + ATP $\rightarrow$ UDP + ADP + H ⁺
<i>v</i> ₁₀₄	UDP + ATP $\longleftrightarrow$ UTP + ADP + H ⁺
<i>v</i> ₁₀₅	UTP + ATP + glutamine $\rightarrow$ CTP + ADP + P _{<i>i</i>} + glutamate
<i>v</i> ₁₀₆	CMP + ATP $\longleftrightarrow$ CDP + ADP + H ⁺
<i>v</i> ₁₀₇	CDP + ATP $\longleftrightarrow$ CTP + ADP + H ⁺
<i>v</i> ₁₀₈	5'-P-ribosyl-1PP + 2 glutamine + aspartate + glycine + 4 ATP + CHO-THF + CO ₂ + 2 H ₂ O $\rightarrow$ ACR + fumarate + 2 glutamate + 4 ADP + THF + PP _{<i>i</i>} + 4 P _{<i>i</i>}
<i>v</i> ₁₀₉	ACR + CHO-THF $\longleftrightarrow$ IMP + THF + H ₂ O
<i>v</i> ₁₁₀	IMP + ATP + NAD ⁺ + glutamine + 2 H ₂ O $\rightarrow$ GMP + glutamate + AMP + NADH,H ⁺ + PP _{<i>i</i>}
<i>v</i> ₁₁₁	GMP + ATP $\rightarrow$ GDP + ADP + H ⁺
<i>v</i> ₁₁₂	GDP + ATP $\longleftrightarrow$ GTP + ADP + H ⁺
<i>v</i> ₁₁₃	aspartate + IMP + GTP $\longleftrightarrow$ AMP + fumarate + GDP + P _{<i>i</i>}
<i>v</i> ₁₁₄	UDP + 2 ATP + CH ₂ -THF + NADPH,H ⁺ $\rightarrow$ 2 ADP + dTTP + DHF + NADP ⁺ + P _{<i>i</i>}
<i>v</i> ₁₁₅	dTDP + ATP $\rightarrow$ dTTP + ADP + H ⁺
<i>v</i> ₁₁₆	CDP + ATP + NADPH,H ⁺ $\rightarrow$ dCTP + ADP + NADP ⁺ + H ₂ O
<i>v</i> ₁₁₇	GDP + ATP + NADPH,H ⁺ $\rightarrow$ dGTP + ADP + NADP ⁺ + H ₂ O
<i>v</i> ₁₁₈	ATP + NADPH,H ⁺ $\longleftrightarrow$ dATP + NADP ⁺ + H ₂ O

ATP Synthesis and Cofactor Metabolism

<i>v</i> ₁₁₉	2 Fe _{red} + 2 H ⁺ + NADP ⁺ $\rightarrow$ 2 Fe _{ox} + NADPH,H ⁺
<i>v</i> ₁₂₀	2 H ₂ O + 2 NADP ⁺ + 8 photons + 6 H _{<i>n</i>} ⁺ $\rightarrow$ O ₂ + 2 NADPH + 12 H _{<i>p</i>} ⁺
<i>v</i> ₁₂₁	NADH + 11 H _{<i>n</i>} ⁺ + 0.5 O ₂ $\rightarrow$ NAD ⁺ + 10 H _{<i>p</i>} ⁺ + H ₂ O
<i>v</i> ₁₂₂	FADH ₂ + 6 H _{<i>n</i>} ⁺ + 0.5 O ₂ $\rightarrow$ FAD ⁺ + 6 H _{<i>p</i>} ⁺ + H ₂ O
<i>v</i> ₁₂₃	ADP + P _{<i>i</i>} + 4 H _{<i>p</i>} ⁺ $\rightarrow$ ATP + H ₂ O + 4 H _{<i>n</i>} ⁺
<i>v</i> ₁₂₄	ATP + H ₂ O $\longleftrightarrow$ ADP + P _{<i>i</i>} + MAINT
<i>v</i> ₁₂₅	ATP + AMP $\longleftrightarrow$ 2 ADP + H ⁺
<i>v</i> ₁₂₆	NADPH + NAD ⁺ $\longleftrightarrow$ NADP ⁺ + NADH
<i>v</i> ₁₂₇	PP _{<i>i</i>} + H ₂ O $\rightarrow$ 2 P _{<i>i</i>}
<i>v</i> ₁₂₈	ATP + 2 H ₂ O $\rightarrow$ AMP + 2 P _{<i>i</i>}
<i>v</i> ₁₂₉	NAD ⁺ + H ⁺ + 2 e ⁻ $\longleftrightarrow$ NADH
<i>v</i> ₁₃₀	NADP ⁺ + H ⁺ + 2 e ⁻ $\longleftrightarrow$ NADPH
<i>v</i> ₁₃₁	FAD ⁺ + 2 H ⁺ + 2 e ⁻ $\longleftrightarrow$ FADH ₂
<i>v</i> ₁₃₂	2 H ₂ O $\longleftrightarrow$ O ₂ + 4 H ⁺ + 4 e ⁻

Carbohydrate Synthesis

v_{133}	$\text{glucose-6-P} + \text{UTP} \rightarrow \text{UDP-glucose} + \text{PP}_i$
v_{134}	$\text{UDP-glucose} \rightarrow \text{UDP-galactose}$
v_{135}	$\text{fructose-6-P} + \text{GTP} \rightarrow \text{GDP-mannose} + \text{PP}_i$
v_{136}	$\text{GDP-mannose} + \text{NADPH}, \text{H}^+ \rightarrow \text{GDP-fucose} + \text{NADP}^+ + \text{H}_2\text{O}$
v_{137}	$\text{glucose-6-P} + \text{dTTP} + \text{NADPH}, \text{H}^+ \rightarrow \text{dTDP-rhamnose} + \text{PP}_i + \text{NADP}^+ + \text{H}_2\text{O}$
v_{138}	$\text{UDP-glucose} + 2 \text{NAD}^+ + \text{H}_2\text{O} \rightarrow \text{UDP-glucuronate} + 2 \text{NADH}, \text{H}^+$
v_{139}	$\text{UDP-glucuronate} \rightleftharpoons \text{UDP-galacturonate}$
v_{140}	$\text{UDP-glucuronate} + \text{H}^+ \rightarrow \text{UDP-xylose} + \text{CO}_2$
v_{141}	$\text{fructose-6-P} + \text{glutamine} + \text{acetyl-CoA} + \text{UTP} \rightarrow \text{UDP-N-acetylglucosamine} + \text{glutamate} + \text{CoASH} + \text{PP}_i$
v_{142}	$\text{UDP-N-acetylglucosamine} + \text{glutamate} + 3 \text{alanine} + \text{P-enolpyruvate} + \text{NADPH}, \text{H}^+ + \text{diaminopimelate} + 4 \text{ATP} \rightarrow \text{UDP-N-acetylpentapeptide} + \text{NADP}^+ + 4 \text{ADP} + 5 \text{P}_i$
v_{143}	$\text{UDP-N-acetylglucosamine} + \text{P-enolpyruvate} + \text{ATP} + \text{CTP} + 3 \text{H}_2\text{O} \rightarrow \text{CMP-N-acetylneuraminate} + \text{UDP} + \text{ADP} + \text{PP}_i + 2 \text{P}_i$
v_{144}	$\text{glucose-6-P} + \text{H}_2\text{O} \rightarrow \text{cyclitol}_{av} + \text{P}_i$
v_{145}	$\text{UDP-N-acetylglucosamine} + \text{UDP-N-acetylpentapeptide} + 5 \text{glycine} \rightarrow \text{peptidoglycan}_{av} + \text{alanine} + \text{UMP} + \text{UDP} + \text{P}_i$
v_{146}	$\text{glucose-6-P} + \text{ATP} \rightarrow \text{ADP} + \text{PP}_i + \text{glycogen}$

Lipid Synthesis

v_{147}	$\text{pyruvate} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{acetyl-CoA} + \text{CO}_2 + \text{NADH}, \text{H}^+$
v_{148}	$\text{acetyl-CoA} + \text{H}_2\text{O} \rightarrow \text{acetate} + \text{CoASH} + \text{H}^+$
v_{149}	$\text{dihydroxyacetone-P} + \text{NADH}, \text{H}^+ \rightleftharpoons \text{glycerol-3-P} + \text{NAD}^+$
v_{150}	$\text{acetyl-CoA} + \text{ACP} \rightarrow \text{acetyl-ACP} + \text{CoASH}$
v_{151}	$\text{acetyl-CoA} + \text{ATP} + \text{CO}_2 + \text{H}_2\text{O} \rightarrow \text{malonyl-CoA} + \text{ADP} + \text{P}_i$
v_{152}	$\text{acetyl-ACP} + 7 \text{malonyl-CoA} + 14 \text{NADPH}, \text{H}^+ \rightarrow \text{palmitic-ACP} + 7 \text{CoASH} + 14 \text{NADP}^+ + 7 \text{CO}_2 + 7 \text{H}_2\text{O}$
v_{153}	$\text{malonyl-CoA} + \text{palmitic-ACP} + 2 \text{NADPH}, \text{H}^+ \rightarrow \text{stearic-ACP} + \text{CoASH} + 2 \text{NADP}^+ + \text{CO}_2 + \text{H}_2\text{O}$
v_{154}	$\text{palmitic-ACP} + \text{NADP}^+ \rightarrow \text{palmitoleic-ACP} + \text{NADPH}, \text{H}^+$
v_{155}	$\text{stearic-ACP} + \text{NADH}, \text{H}^+ \rightarrow \text{oleic-ACP} + \text{NAD}^+$
v_{156}	$\text{oleic-ACP} + \text{NADH}, \text{H}^+ \rightarrow \text{linoleic-ACP} + \text{NAD}^+$
v_{157}	$\text{linoleic-ACP} + \text{NADH}, \text{H}^+ \rightarrow \gamma\text{-linolenic-ACP} + \text{NAD}^+$
v_{158}	$\text{glycerol-3-P} + 0.922 \text{palmitic-ACP} + 0.212 \text{palmitoleic-ACP} + 0.059 \text{stearic-ACP} + 0.101 \text{oleic-ACP} + 0.271 \text{linoleic-ACP} + 0.435 \gamma\text{-linolenic-ACP} \rightarrow \text{phosphatidic acid}_{av} + 2 \text{ACP}$
v_{159}	$\text{phosphatidic acid}_{av} + \text{H}_2\text{O} \rightarrow \text{diacylglycerol}_{av} + \text{P}_i$
v_{160}	$\text{diacylglycerol}_{av} + \text{UDP-glucose} \rightarrow \text{monogalactosyldiacylglycerol} + \text{UDP}$
v_{161}	$2 \text{monogalactosyldiacylglycerol} \rightarrow \text{digalactosyldiacylglycerol} + \text{diacylglycerol}_{av}$
v_{162}	$\text{UDP-glucose} + \text{SO}_3 + \text{H}^+ \rightarrow \text{H}_2\text{O} + \text{UDP-sulfoquinovose}$
v_{163}	$\text{diacylglycerol}_{av} + \text{UDP-sulfoquinovose} \rightarrow \text{sulfoquinovosyldiacylglycerol} + \text{UDP}$
v_{164}	$\text{phosphatidic acid}_{av} + \text{glycerol-3-P} + \text{CTP} \rightarrow \text{glycerol} + \text{CMP} + \text{PP}_i + \text{P}_i$
v_{165}	$\text{glycerol-3-P} + 0.956 \text{palmitic-ACP} + 0.573 \text{palmitoleic-ACP} + 0.299 \text{stearic-ACP} + 0.425 \text{oleic-ACP} + 0.377 \text{linoleic-ACP} + 0.370 \gamma\text{-linolenic-ACP} + \text{H}_2\text{O} \rightarrow \text{triglyceride}_{av} + 3 \text{ACP} + \text{P}_i$
v_{166}	$\text{acetate} + \text{CoASH} + \text{ATP} \rightarrow \text{acetyl-CoA} + \text{AMP} + \text{PP}_i$

Chlorophyll Synthesis

v_{167}	$8 \text{glutamate} + 8 \text{ATP} + 8 \text{NADPH}, \text{H}^+ + 2.5 \text{O}_2 \rightarrow \text{protoporphyrin} + 8 \text{ADP} + 8 \text{P}_i + 21 \text{H}_2\text{O} + 8 \text{NADP}^+ + 6 \text{CO}_2 + 4 \text{NH}_3$
v_{168}	$12 \text{acetyl-CoA} + 11 \text{NADPH}, \text{H}^+ + 12 \text{ATP} + 4 \text{H}_2\text{O} \rightarrow \text{phytyl-PP} + 12 \text{CoASH} + 11 \text{NADP}^+ + 12 \text{ADP} + 4 \text{P}_i + 4 \text{CO}_2 + 3 \text{PP}_i$
v_{169}	$\text{methionine} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{S-ad.methionine} + \text{PP}_i + \text{P}_i$
v_{170}	$\text{protoporphyrin} + \text{Mg}^{2+} + \text{S-ad.methionine} + 5 \text{NADPH}, \text{H}^+ + 3 \text{O}_2 + \text{phytyl-PP} + \text{ATP} \rightarrow \text{chlorophyll} + 3 \text{H}^+ + \text{S-ad.homocysteine} + 5 \text{NADP}^+ + 4 \text{H}_2\text{O} + \text{PP}_i + \text{ADP} + \text{P}_i$
v_{171}	$\text{S-ad.homocysteine} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{homocysteine} + \text{AMP} + \text{ADP}$

Nitrogen Storage

v_{172}	$\text{aspartate} + \text{arginine} \rightarrow \text{cyanophycin} + \text{H}_2\text{O}$

Synthesis of Biomass Macromolecules

v_{173}	$0.732 \text{dTDP-rhamnose} + 0.124 \text{UDP-glucose} + 0.051 \text{cyclitol}_{av} + 0.030 \text{peptidoglycan}_{av} + 0.043 \text{glycogen} + 0.022 \text{CMP-N-acetylneuraminate} + 5 \text{ATP} \rightarrow \text{carbohydrate}_{av} + 5 \text{ADP} + 5 \text{P}_i + 0.732 \text{dTDP} + 0.124 \text{UDP} + 0.022 \text{CMP}$
v_{174}	$0.058 \text{isoleucine} + 0.094 \text{leucine} + 0.036 \text{lysine} + 0.018 \text{methionine} + 0.039 \text{phenylalanine} + 0.054 \text{threonine} + 0.009 \text{tryptophan} + 0.076 \text{valine} + 0.053 \text{arginine} + 0.014 \text{histidine} + 0.101 \text{alanine} + 0.098 \text{aspartate} + 0.132 \text{glutamate} + 0.007 \text{cysteine} + 0.086 \text{glycine} + 0.038 \text{proline} + 0.054 \text{serine} + 0.033 \text{tyrosine} + 2 \text{GTP} + 55 \text{ATP} + 2 \text{H}_2\text{O} \rightarrow \text{protein}_{av} + 2 \text{GDP} + 55 \text{ADP} + \text{PP}_i + 55 \text{P}_i$

v_{175}	$0.320 \text{ monogalactosyldiacylglycerol} + 0.131 \text{ digalactosyldiacylglycerol} + 0.147 \text{ sulfoquinovosyldiacylglycerol} + 0.216 \text{ glycerol} + 0.186 \text{ triglyceride}_{av} + \text{ATP} \rightarrow \text{lipid}_{av} + \text{ADP} + \text{P}_i$
v_{176}	$0.262 \text{ ATP} + 0.322 \text{ GTP} + 0.216 \text{ UTP} + 0.200 \text{ CTP} + 2 \text{ ATP} \rightarrow \text{RNA}_{av} + \text{PP}_i + 2 \text{ ADP} + 2 \text{ P}_i$
v_{177}	$0.279 \text{ dTTP} + 0.222 \text{ dCTP} + 0.222 \text{ dGTP} + 0.279 \text{ dATP} + 5 \text{ ATP} \rightarrow \text{DNA}_{av} + \text{PP}_i + 5 \text{ ADP} + 5 \text{ P}_i$
v_{178}	$0.8404 \text{ protein}_{av} + 0.1230 \text{ carbohydrate}_{av} + 0.0182 \text{ lipid}_{av} + 0.0130 \text{ RNA}_{av} + 0.0039 \text{ DNA}_{av} + 0.0016 \text{ chlorophyll} + 97 \text{ ATP} \rightarrow \text{biomass}_{av} + 97 \text{ ADP} + 97 \text{ P}_i$

Transport Reactions

v_{179}	$\text{HCO}_3^- \text{ ext} \longleftrightarrow \text{HCO}_3^-$
v_{180}	$\text{NO}_2^- \text{ ext} \longleftrightarrow \text{NO}_2^-$
v_{181}	$\text{photon ext} \rightarrow \text{photon}$
v_{182}	$\text{CO}_2 \text{ ext} \longleftrightarrow \text{CO}_2$
v_{183}	$\text{O}_2 \text{ ext} \longleftrightarrow \text{O}_2$
v_{184}	$\text{H}_2\text{O ext} \longleftrightarrow \text{H}_2\text{O}$
v_{185}	$\text{P}_i \text{ ext} \longleftrightarrow \text{P}_i$
v_{186}	$\text{NO}_3^- \text{ ext} \longleftrightarrow \text{NO}_3^-$
v_{187}	$\text{SO}_4^{2-} \text{ ext} \longleftrightarrow \text{SO}_4^{2-}$
v_{188}	$\text{NH}_4^+ \text{ ext} \longleftrightarrow \text{NH}_4^+$
v_{189}	$\text{Mg}^{2+} \text{ ext} \longleftrightarrow \text{Mg}^{2+}$
v_{190}	$\text{H}^+ \text{ ext} \longleftrightarrow \text{H}^+$
v_{191}	$\text{urea ext} \longleftrightarrow \text{urea}$
v_{192}	$\text{glucose ext} \rightarrow \text{glucose}$
v_{193}	$\text{chlorophyll} \rightarrow \text{chlorophyll ext}$
v_{194}	$\text{cyanophycin} \rightarrow \text{cyanophycin ext}$
v_{195}	$\text{carbohydrate} \longleftrightarrow \text{carbohydrate ext}$
v_{196}	$\text{lipid} \longleftrightarrow \text{lipid ext}$
v_{197}	$\text{biomass} \rightarrow \text{biomass ext}$
v_{198}	$\text{MAINT} \rightarrow \text{MAINT ext}$

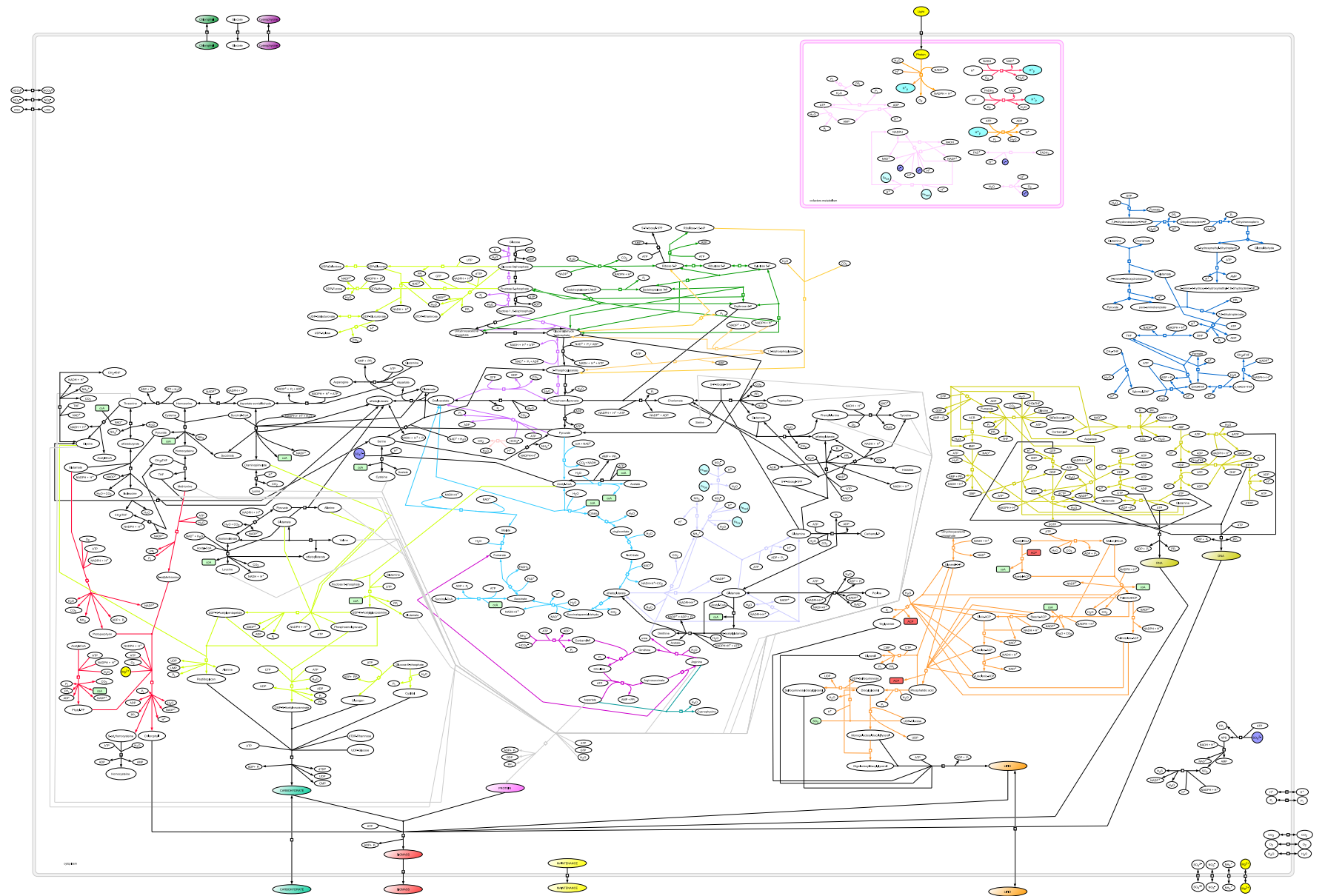


Figure G.1: Complete flux map of the final metabolic network obtained from the constructive procedure (the flux map is constructed using *CellDesigner*). Online reading is advised to identify the main metabolic routes.

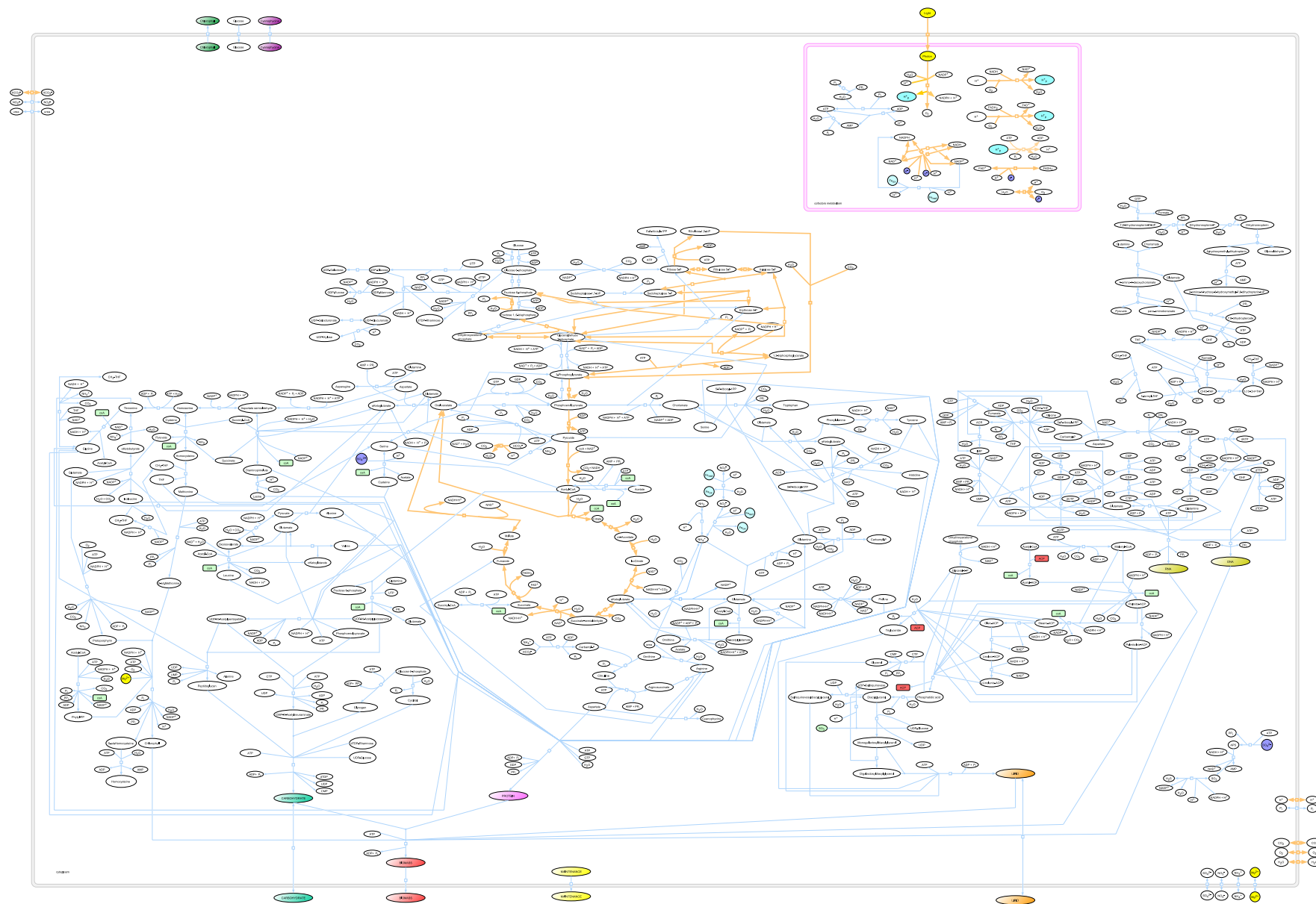


Figure G.2: Complete flux map during the day phase for a net assimilation of 100 mol of HCO_3^- . Online reading is advised to identify the main metabolic routes. Active routes are represented in orange.

Appendix H

Focused Metabolic Network for Arthrospira Sp. PCC 8005

H.1 Network Reduction

In Sec. 4.3.1.2.4, the final iteration of the procedure retains a generalized/enriched metabolic network that provides a comprehensive representation of the metabolic capabilities of the system. In that way, it includes reactions that generate intermediate metabolites, some of which may be involved in the synthesis of essential biomolecules, such as cofactors or vitamins, if future extensions of the model are needed. However, given the limited amount of experimental measurements and the underdetermined nature of the network, it is now necessary to refine the scope of the analysis. At this stage, focusing only on essential pathways (i.e., those directly contributing to central metabolism and biomass precursors) will allow reducing the solution space (see Sec. H.4).

A network reduction ¹ is suggested using FCA. As discussed in [156], flux coupling analysis is also particularly useful to simplify over-detailed networks. While FCA was previously used to identify blocked reactions and pinpoint gaps or errors in the structure of the metabolic network, FCA is now used as a method to preserve only essential pathways. In doing so, the metabolic network is reduced to 175 biochemical reactions and 152 metabolites by simplifying the metabolism of tetrahydrofolate and the metabolism of cofactors, among others. This reduced network is given in Table H.1.

Furthermore, a modeling artifact was detected after a detailed analysis of the intracellular flux distribution. It appears in the metabolic reactions associated with nitrogen assimilation (see reactions v_{41} and v_{42} in Table H.1). These reactions involve the interconversion of ferredoxins Fe_{red} and Fe_{ox} , reflecting the role of iron-sulfur proteins as electron carriers. Because these redox cofactors are internally cycled within the system, their net production and consumption must remain balanced. As a result, the model tightly couples the fluxes of different nitrogen sources through this redox constraint, making it impossible to independently constrain multiple nitrogen assimilation pathways using experimental data. To address this issue, we chose to exclude Fe species from the network representation, which is a common strategy when electron transfer is

¹ This approach does not undermine the value of the generalized network constructed previously; rather, it demonstrates its utility by enabling targeted pathway selection while preserving the option for future extensions as more data become available.

implicit or not the focus of the analysis. An alternative approach would involve modeling Fe explicitly, including transport or buffering reactions to decouple the associated fluxes. This structural limitation emphasizes (i) the importance of experimental validation in addition to in-silico validation, (ii) the necessity of performing flux consistency checks, and (iii) the need for an iterative strategy for both network definition and experimental validation.

H.2 Data Preprocessing

As previously mentioned, the light intensity is commonly expressed in W/m^2 . However, in metabolic flux analyses, light uptake is expressed in millimoles of photons. $\text{g}_{DW}^{-1}.\text{h}^{-1}$. As discussed in Sec. 4.3.2.1, the first step is achieved by assuming an average wavelength $\lambda = 600 \text{ nm}$, expressing the energy of a single photon with Planck's equation and using Avogadro's number. In doing so, considering a constant power density of $125 \text{ W}/\text{m}^2$, a constant illumination of $\pm 6.2654 \cdot 10^{-4} \text{ E}.\text{m}^{-2}.\text{s}^{-1}$ (or moles of photons per square meter per second) is obtained.

Then, to translate this light input into an absolute photon uptake rate for the system, the illuminated surface area of the photobioreactor must be considered. The experiments were carried out in a radially illuminated, cylindrical, double-jacket reactor with a working volume of 2 L. Given the reactor geometry and setup, the illuminated surface was estimated to be approximately 0.08 m^2 . Multiplying the incident photon flux (in $\text{mol photons}.\text{m}^{-2}.\text{s}^{-1}$) by this illuminated area yields the total number of photons incident in the culture per second.

To express this in terms consistent with metabolic flux analysis, the resulting quantity is converted to millimoles per hour. Finally, to obtain the specific photon uptake rate (in $\text{mmol photons}.\text{g}_{DW}^{-1}.\text{h}^{-1}$), this total absorbed photon flux is normalized by biomass concentration ($\text{g}_{DW}.\text{L}^{-1}$). This specific photon uptake rate is then used as a constraint in metabolic simulations.

It is important to note that the above computation relies on simplifying assumptions, in particular, the use of a single average wavelength to represent the incident light. This approach provides an order-of-magnitude estimate of photon flux but does not account for the polychromatic nature of the solar (or artificial) light spectrum. In reality, photons of different wavelengths do not contribute equally to photosynthesis. A more rigorous calculation of the photon flux usable by the cells would require integrating the incident light spectrum with the action spectrum of photosynthesis, which describes the wavelength-dependent efficiency of photon utilization. Such an approach would yield an effective photon flux that better reflects the true energetic input driving the metabolism. More details can be found in [1]. Implementing this method would involve measuring or assuming the spectral distribution of the light source, applying appropriate weighting by the photosynthetic action spectrum, and then converting the result into a photon flux density. Although not implemented here, acknowledging these aspects highlights the limitations of the current simplification and points to a direction for more accurate future analyses.

The biomass concentration X (DCW , $\text{g}.\text{L}^{-1}$) is retrieved from the biomass productivity as mentioned in [267]. In addition, the experimental uptake and secretion rates are expressed in $\text{mmol}.\text{g}_{DW}^{-1}.\text{h}^{-1}$ and are calculated by using smoothing splines and applying differentiation methods to the concentration profiles (see Eq. D.2). Finally, the oxygen secretion rate can be derived from experimental oxygen productivity data.

As mentioned in Sec. 4.3.3.1, the experiments were conducted in a photobioreactor operating in continuous mode, specifically a stirred-tank reactor with constant agitation (150 rpm), continuous illumination, controlled pH, and a fixed dilution rate maintained by the addition of fresh feeding medium. Under these conditions, the system is expected to reach a steady-state continuous mode [119], in which the specific growth rate (μ) should approximately equal the dilution rate (D), and the concentrations of biomass and metabolic products should remain relatively stable over time. Consequently, uptake and secretion rates should also be relatively constant, assuming that the system is truly at steady-state. Nevertheless, it is worth noting that there are exceptions, and these exceptions can be of scientific and practical interest, as discussed in [119].

In this particular case, small fluctuations are observed in the uptake and secretion rate profiles. This can be explained by several factors. First, the system operates initially in batch mode, where cell growth is exponential and nutrient concentrations change dynamically. Upon switching to continuous mode, there is an inevitable transition phase in which biomass concentration, metabolic fluxes, and nutrient uptake rates adjust to the new operating conditions, leading to temporary variations before stabilizing. Second, measurement errors may introduce noise into concentration profiles. Third, the uptake and secretion rates are obtained applying smoothing splines on the noisy concentration profiles and differentiation methods, which could lead to numerical artifacts. Therefore, the fluctuations observed in the uptake and secretion rate profiles should be interpreted with this context in mind.

H.3 Well-Posedness Analysis

Ensuring the formulation of a well-posed problem is essential. Specifically, it involves determining whether certain pathways within the metabolic network connect non-measured inputs to non-measured outputs. If such pathways exist, they can lead to an unbounded set of solutions for the corresponding reaction rates, making the system's behavior difficult to interpret and predict. In that respect, as already discussed in Sec. 2.4, by taking into account experimental measurements, Eq. 4.11 becomes:

$$\begin{pmatrix} N & 0 \\ N_e & -\underline{\nu}_m \end{pmatrix} \cdot \begin{pmatrix} \underline{v} \\ 1 \end{pmatrix} = 0 \quad (\text{H.1})$$

where $\underline{\nu}_m$ is the vector of specific uptake and excretion rates of the measured external species, and N_e is the stoichiometric matrix of extracellular measurements. Note that the previous equation is equivalent to Eq. 2.5. Therefore, checking if the system is well-posed is equivalent to computing the elementary flux modes of the matrix in Eq. H.1. These elementary modes are collected in a matrix E . The system is said to be well-posed if there is no null column in E . In this case, after using METATOOL [26], a matrix containing 303,583 EFMs is obtained. After analysis, no null column is detected, meaning the problem is well-posed.

H.4 Calculability Analysis

Furthermore, a calculability/observability analysis is also relevant to examine under-determined networks, as discussed in [209], and involves identifying fluxes that can be uniquely determined. In this context, it enables an improved interpretation of flux distributions, avoiding misleading conclusions from non-unique solutions, but it can also guide experimental efforts if additional experimental measurements are needed. This analysis is based on Eq. 2.6, Eq. 2.7 and Eq. 2.8 where the vector of network fluxes $\underline{v}$ is

partitioned into: (i) a vector of measured or known fluxes $\underline{v}_k$ and (ii) a vector of unmeasured or unknown fluxes $\underline{v}_u$ and where the matrix N is divided into the corresponding sub-matrices N_k and N_u such that:

$$N_u \cdot \underline{v}_u = -N_k \cdot \underline{v}_k \quad (\text{H.2})$$

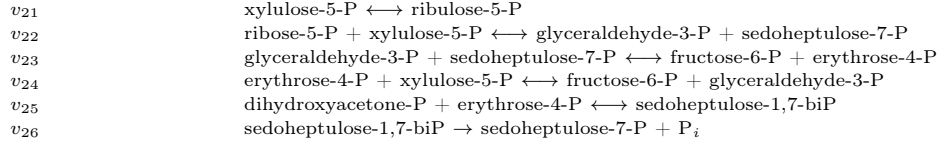
In the general case, the matrix N_u is a non invertible matrix, meaning that $\underline{v}_u$ is obtained using the Moore-Penrose pseudo-inverse operation:

$$\underline{v}_{u,s} = -N_u^\# \cdot N_k \cdot \underline{v}_k + K_u \cdot \underline{a} \quad (\text{H.3})$$

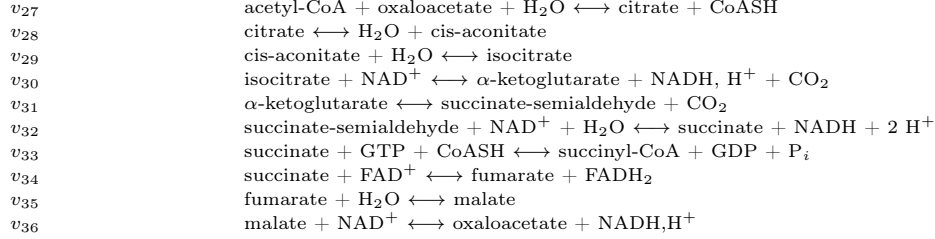
Where K_u is the kernel of the matrix N_u and $\underline{a}$ is an arbitrary vector with $(\dim(\underline{v}_u) - \text{rank}(N_u))$ elements, which reflects the indeterminacy of $\underline{v}_{u,s}$. Consequently, independently of the vector $\underline{a}$, if there are calculable fluxes, all different solutions $\underline{v}_{u,s}$ will give the same value for these calculable fluxes. In this case, after computation, it was determined that only a few dozen fluxes are calculable when the growth rate is considered a well-known flux. Note that in order to further reduce the solution space, it would be necessary to impose additional experimental constraints or make biological assumptions. For example, we could assume that the catabolism of threonine is not a major metabolic pathway in *Arthrospira* sp. compared to its biosynthesis. This is a reasonable assumption, especially in the context of biomass production, protein synthesis, or other biosynthetic goals.

Table H.1: Metabolic reactions for the metabolism of *Arthrospira spirulina platensis* PCC 8005 (photoautotrophic regime). Protons H^+ involved in most metabolic reactions correspond to H_n^+ .

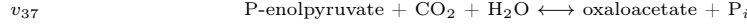
Flux	Metabolic reaction
Carbon Entry	
v_1	$\text{HCO}_3^- + \text{H}^+ \rightarrow \text{CO}_2 + \text{H}_2\text{O}$
Calvin Cycle	
v_2	$3 \text{ ribulose-1,5-biP} + 3 \text{ CO}_2 + \text{H}_2\text{O} \rightarrow 6 \text{ 3-P-glycerate}$
v_3	$6 \text{ 3-P-glycerate} + 6 \text{ ATP} \longleftrightarrow 6 \text{ 1,3-biphosphoglycerate} + 6 \text{ ADP}$
v_4	$6 \text{ 1,3-biphosphoglycerate} + 6 \text{ NADPH,H}^+ \longleftrightarrow 6 \text{ glyceraldehyde-3-P} + 6 \text{ NADP}^+ + 6 \text{ P}_i$
v_5	$5 \text{ glyceraldehyde-3-P} \rightarrow 3 \text{ ribose-5-P} + 2 \text{ P}_i$
Glycolysis	
v_6	$\text{glucose} + \text{ATP} \rightarrow \text{glucose-6-P} + \text{ADP}$
v_7	$\text{glucose-6-P} \longleftrightarrow \text{fructose-6-P}$
v_8	$\text{fructose-6-P} + \text{ATP} \rightarrow \text{fructose-1,6-biP} + \text{ADP}$
v_9	$\text{fructose-1,6-biP} \longleftrightarrow \text{glyceraldehyde-3-P} + \text{dihydroxyacetone-P}$
v_{10}	$\text{glyceraldehyde-3-P} \longleftrightarrow \text{dihydroxyacetone-P}$
v_{11}	$\text{glyceraldehyde-3-P} + \text{NAD}^+ + \text{P}_i + \text{ADP} \longleftrightarrow \text{3-P-glycerate} + \text{NADH,H}^+ + \text{ATP}$
v_{12}	$3\text{-P-glycerate} \longleftrightarrow \text{P-enolpyruvate} + \text{H}_2\text{O}$
v_{13}	$\text{P-enolpyruvate} + \text{ADP} \rightarrow \text{pyruvate} + \text{ATP}$
Neoglucogenesis	
v_{14}	$\text{glucose-6-P} + \text{H}_2\text{O} \rightarrow \text{glucose} + \text{P}_i$
v_{15}	$\text{fructose-1,6-biP} + \text{H}_2\text{O} \rightarrow \text{fructose-6-P} + \text{P}_i$
v_{16}	$\text{oxaloacetate} + \text{GTP} \rightarrow \text{P-enolpyruvate} + \text{GDP} + \text{CO}_2$
v_{17}	$\text{pyruvate} + \text{HCO}_3^- + \text{ATP} \rightarrow \text{oxaloacetate} + \text{ADP} + \text{P}_i$
PPP	
v_{18}	$\text{glucose-6-P} + 2 \text{ NADP}^+ + \text{H}_2\text{O} \rightarrow \text{ribose-5-P} + 2 \text{ NADPH,H}^+ + \text{CO}_2$
v_{19}	$\text{ribose-5-P} \longleftrightarrow \text{ribulose-5-P}$
v_{20}	$\text{ribose-5-P} + \text{ATP} \rightarrow \text{ribulose-1,5-biP} + \text{ADP}$



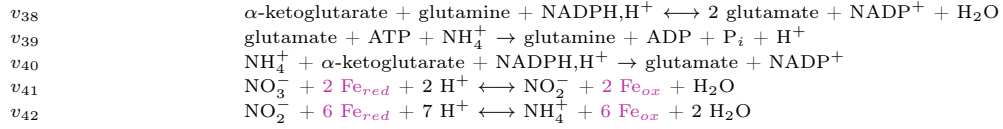
TCA Cycle



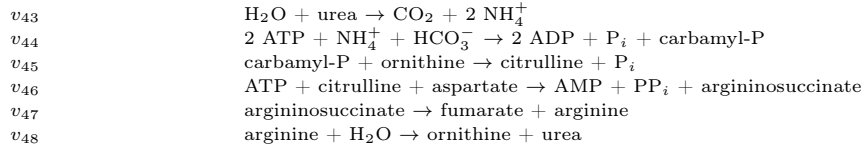
Anaplerotic Reaction



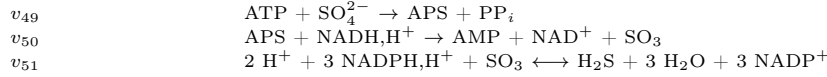
GS-GOGAT Pathway



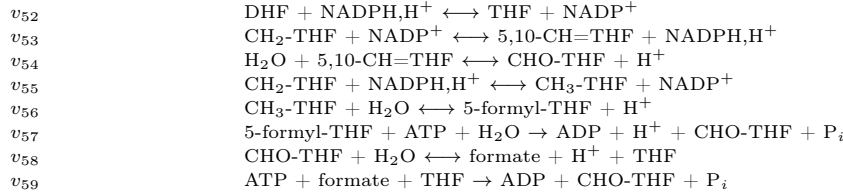
Urea Cycle-like Pathway



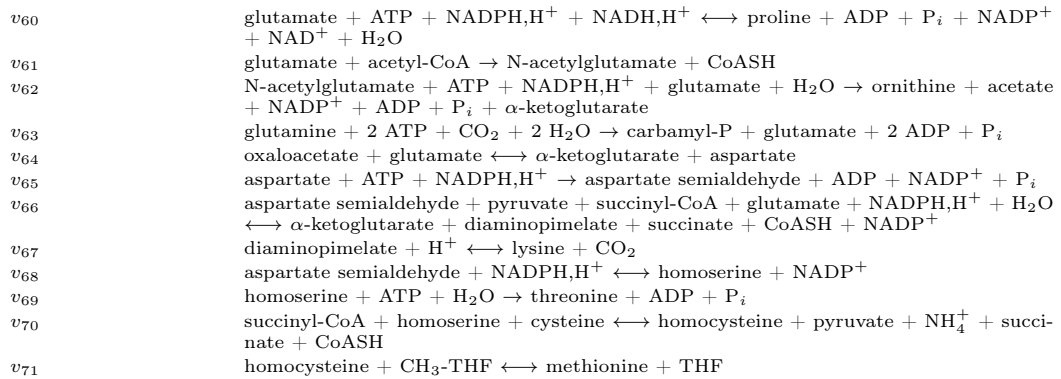
S Fixation

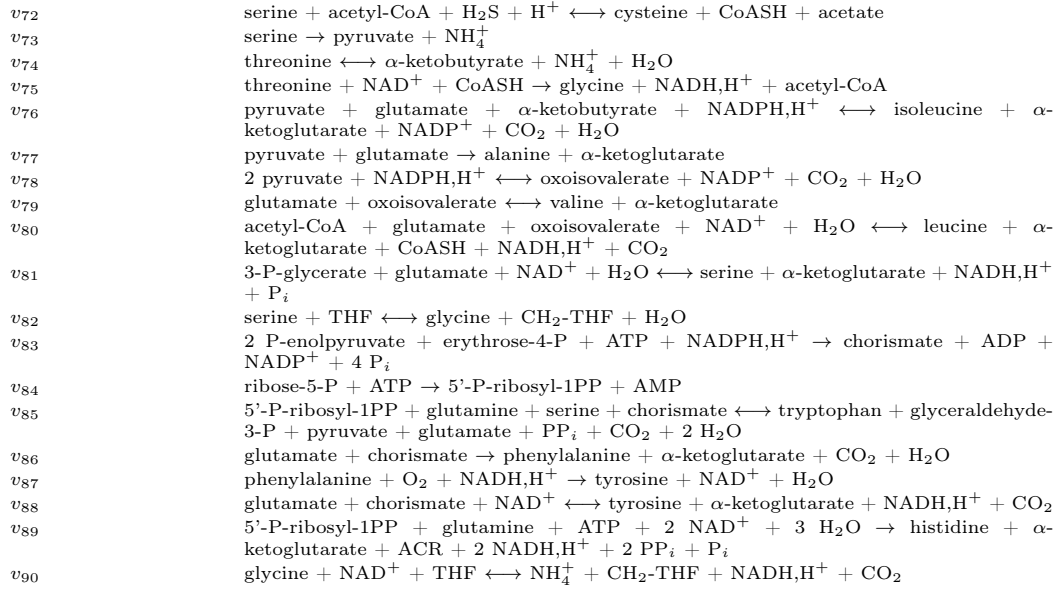


THF Metabolism

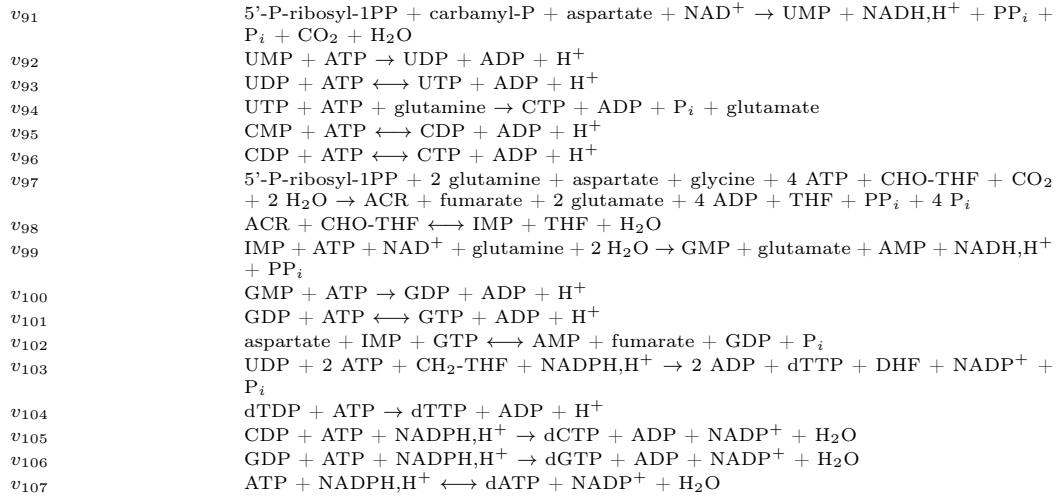


AA Metabolism

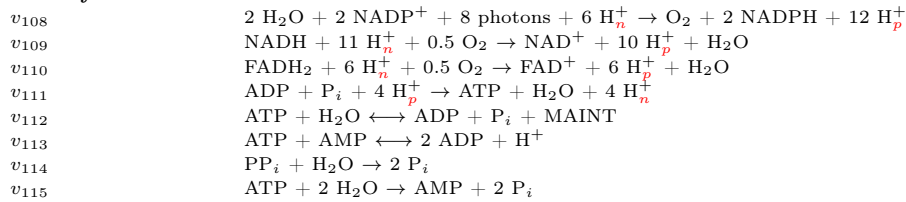




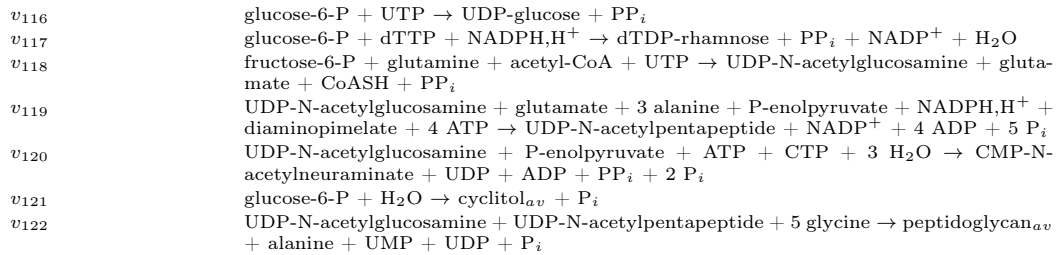
Nucleotide Salvage Pathway



ATP Synthesis and Cofactor Metabolism



Carbohydrate Synthesis



v_{123}	$\text{glucose-6-P} + \text{ATP} \rightarrow \text{ADP} + \text{PP}_i + \text{glycogen}$
Lipid Synthesis	
v_{124}	$\text{pyruvate} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{acetyl-CoA} + \text{CO}_2 + \text{NADH,H}^+$
v_{125}	$\text{acetyl-CoA} + \text{H}_2\text{O} \rightarrow \text{acetate} + \text{CoASH} + \text{H}^+$
v_{126}	$\text{dihydroxyacetone-P} + \text{NADH,H}^+ \rightleftharpoons \text{glycerol-3-P} + \text{NAD}^+$
v_{127}	$\text{acetyl-CoA} + \text{ACP} \rightarrow \text{acetyl-ACP} + \text{CoASH}$
v_{128}	$\text{acetyl-CoA} + \text{ATP} + \text{CO}_2 + \text{H}_2\text{O} \rightarrow \text{malonyl-CoA} + \text{ADP} + \text{P}_i$
v_{129}	$\text{acetyl-ACP} + 7 \text{ malonyl-CoA} + 14 \text{ NADPH,H}^+ \rightarrow \text{palmitic-ACP} + 7 \text{ CoASH} + 14 \text{ NADP}^+ + 7 \text{ CO}_2 + 7 \text{ H}_2\text{O}$
v_{130}	$\text{malonyl-CoA} + \text{palmitic-ACP} + 2 \text{ NADPH,H}^+ \rightarrow \text{stearic-ACP} + \text{CoASH} + 2 \text{ NADP}^+ + \text{CO}_2 + \text{H}_2\text{O}$
v_{131}	$\text{palmitic-ACP} + \text{NADP}^+ \rightarrow \text{palmitoleic-ACP} + \text{NADPH,H}^+$
v_{132}	$\text{stearic-ACP} + \text{NADH,H}^+ \rightarrow \text{oleic-ACP} + \text{NAD}^+$
v_{133}	$\text{oleic-ACP} + \text{NADH,H}^+ \rightarrow \text{linoleic-ACP} + \text{NAD}^+$
v_{134}	$\text{linoleic-ACP} + \text{NADH,H}^+ \rightarrow \gamma\text{-linolenic-ACP} + \text{NAD}^+$
v_{135}	$\text{glycerol-3-P} + 0.922 \text{ palmitic-ACP} + 0.212 \text{ palmitoleic-ACP} + 0.059 \text{ stearic-ACP} + 0.101 \text{ oleic-ACP} + 0.271 \text{ linoleic-ACP} + 0.435 \gamma\text{-linolenic-ACP} \rightarrow \text{phosphatidic acid}_{av} + 2 \text{ ACP}$
v_{136}	$\text{phosphatidic acid}_{av} + \text{H}_2\text{O} \rightarrow \text{diacylglycerol}_{av} + \text{P}_i$
v_{137}	$\text{diacylglycerol}_{av} + \text{UDP-glucose} \rightarrow \text{monogalactosyldiacylglycerol} + \text{UDP}$
v_{138}	$2 \text{ monogalactosyldiacylglycerol} \rightarrow \text{digalactosyldiacylglycerol} + \text{diacylglycerol}_{av}$
v_{139}	$\text{diacylglycerol}_{av} + \text{UDP-sulfoquinovose} \rightarrow \text{sulfoquinovosyldiacylglycerol} + \text{UDP}$
v_{140}	$\text{phosphatidic acid}_{av} + \text{glycerol-3-P} + \text{CTP} \rightarrow \text{glycerol} + \text{CMP} + \text{PP}_i + \text{P}_i$
v_{141}	$\text{glycerol-3-P} + 0.956 \text{ palmitic-ACP} + 0.573 \text{ palmitoleic-ACP} + 0.299 \text{ stearic-ACP} + 0.425 \text{ oleic-ACP} + 0.377 \text{ linoleic-ACP} + 0.370 \gamma\text{-linolenic-ACP} + \text{H}_2\text{O} \rightarrow \text{triglyceride}_{av} + 3 \text{ ACP} + \text{P}_i$
v_{142}	$\text{acetate} + \text{CoASH} + \text{ATP} \rightarrow \text{acetyl-CoA} + \text{AMP} + \text{PP}_i$
v_{143}	$\text{UDP-glucose} + \text{SO}_3 + \text{H}^+ \rightarrow \text{H}_2\text{O} + \text{UDP-sulfoquinovose}$
Chlorophyll Synthesis	
v_{144}	$8 \text{ glutamate} + 8 \text{ ATP} + 8 \text{ NADPH,H}^+ + 2.5 \text{ O}_2 \rightarrow \text{protoporphyrin} + 8 \text{ ADP} + 8 \text{ P}_i + 21 \text{ H}_2\text{O} + 8 \text{ NADP}^+ + 6 \text{ CO}_2 + 4 \text{ NH}_4^+$
v_{145}	$12 \text{ acetyl-CoA} + 11 \text{ NADPH,H}^+ + 12 \text{ ATP} + 4 \text{ H}_2\text{O} \rightarrow \text{phytyl-PP} + 12 \text{ CoASH} + 11 \text{ NADP}^+ + 12 \text{ ADP} + 4 \text{ P}_i + 4 \text{ CO}_2 + 3 \text{ PP}_i$
v_{146}	$\text{methionine} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{S-ad.methionine} + \text{PP}_i + \text{P}_i$
v_{147}	$\text{protoporphyrin} + \text{Mg}^{2+} + \text{S-ad.methionine} + 5 \text{ NADPH,H}^+ + 3 \text{ O}_2 + \text{phytyl-PP} + \text{ATP} \rightarrow \text{chlorophyll} + 3 \text{ H}^+ + \text{S-ad.homocysteine} + 5 \text{ NADP}^+ + 4 \text{ H}_2\text{O} + \text{PP}_i + \text{ADP} + \text{P}_i$
v_{148}	$\text{S-ad.homocysteine} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{homocysteine} + \text{AMP} + \text{ADP}$
Nitrogen Storage	
v_{149}	$\text{aspartate} + \text{arginine} \rightarrow \text{cyanophycin} + \text{H}_2\text{O}$
Synthesis of Biomass Macromolecules	
v_{150}	$0.732 \text{ dTDP-rhamnose} + 0.124 \text{ UDP-glucose} + 0.051 \text{ cyclitol}_{av} + 0.030 \text{ peptidoglycan}_{av} + 0.043 \text{ glycogen} + 0.022 \text{ CMP-N-acetylneuraminate} + 5 \text{ ATP} \rightarrow \text{carbohydrate}_{av} + 5 \text{ ADP} + 5 \text{ P}_i + 0.732 \text{ dTDP} + 0.124 \text{ UDP} + 0.022 \text{ CMP}$
v_{151}	$0.058 \text{ isoleucine} + 0.094 \text{ leucine} + 0.036 \text{ lysine} + 0.018 \text{ methionine} + 0.039 \text{ phenylalanine} + 0.054 \text{ threonine} + 0.009 \text{ tryptophan} + 0.076 \text{ valine} + 0.053 \text{ arginine} + 0.014 \text{ histidine} + 0.101 \text{ alanine} + 0.098 \text{ aspartate} + 0.132 \text{ glutamate} + 0.007 \text{ cysteine} + 0.086 \text{ glycine} + 0.038 \text{ proline} + 0.054 \text{ serine} + 0.033 \text{ tyrosine} + 2 \text{ GTP} + 55 \text{ ATP} + 2 \text{ H}_2\text{O} \rightarrow \text{protein}_{av} + 2 \text{ GDP} + 55 \text{ ADP} + \text{PP}_i + 55 \text{ P}_i$
v_{152}	$0.320 \text{ monogalactosyldiacylglycerol} + 0.131 \text{ digalactosyldiacylglycerol} + 0.147 \text{ sulfoquinovosyldiacylglycerol} + 0.216 \text{ glycerol} + 0.186 \text{ triglyceride}_{av} + \text{ATP} \rightarrow \text{lipid}_{av} + \text{ADP} + \text{P}_i$
v_{153}	$0.262 \text{ ATP} + 0.322 \text{ GTP} + 0.216 \text{ UTP} + 0.200 \text{ CTP} + 2 \text{ ATP} \rightarrow \text{RNA}_{av} + \text{PP}_i + 2 \text{ ADP} + 2 \text{ P}_i$
v_{154}	$0.279 \text{ dTTP} + 0.222 \text{ dCTP} + 0.222 \text{ dGTP} + 0.279 \text{ dATP} + 5 \text{ ATP} \rightarrow \text{DNA}_{av} + \text{PP}_i + 5 \text{ ADP} + 5 \text{ P}_i$
v_{155}	$0.8404 \text{ protein}_{av} + 0.1230 \text{ carbohydrate}_{av} + 0.0182 \text{ lipid}_{av} + 0.0130 \text{ RNA}_{av} + 0.0039 \text{ DNA}_{av} + 0.0016 \text{ chlorophyll} + 97 \text{ ATP} \rightarrow \text{biomass}_{av} + 97 \text{ ADP} + 97 \text{ P}_i$
Transport Reactions	
v_{156}	$\text{HCO}_3^- \text{ ext} \rightleftharpoons \text{HCO}_3^-$
v_{157}	$\text{NO}_2^- \text{ ext} \rightleftharpoons \text{NO}_2^-$
v_{158}	$\text{photon}_{ext} \rightarrow \text{photon}$
v_{159}	$\text{CO}_2 \text{ ext} \rightleftharpoons \text{CO}_2$
v_{160}	$\text{O}_2 \text{ ext} \rightleftharpoons \text{O}_2$
v_{161}	$\text{H}_2\text{O}_{ext} \rightleftharpoons \text{H}_2\text{O}$
v_{162}	$\text{P}_i \text{ ext} \rightleftharpoons \text{P}_i$

v_{163}	$\text{NO}_3^- \text{ }_{ext} \longleftrightarrow \text{NO}_3^-$
v_{164}	$\text{SO}_4^{2-} \text{ }_{ext} \longleftrightarrow \text{SO}_4^{2-}$
v_{165}	$\text{NH}_4^+ \text{ }_{ext} \longleftrightarrow \text{NH}_4^+$
v_{166}	$\text{Mg}^{2+} \text{ }_{ext} \longleftrightarrow \text{Mg}^{2+}$
v_{167}	$\text{H}^+ \text{ }_{ext} \longleftrightarrow \text{H}^+$
v_{168}	$\text{urea }_{ext} \longleftrightarrow \text{urea}$
v_{169}	$\text{glucose }_{ext} \rightarrow \text{glucose}$
v_{170}	$\text{chlorophyll} \rightarrow \text{chlorophyll }_{ext}$
v_{171}	$\text{cyanophycin} \rightarrow \text{cyanophycin }_{ext}$
v_{172}	$\text{carbohydrate} \longleftrightarrow \text{carbohydrate }_{ext}$
v_{173}	$\text{lipid} \longleftrightarrow \text{lipid }_{ext}$
v_{174}	$\text{biomass} \rightarrow \text{biomass }_{ext}$
v_{175}	$\text{MAINT} \rightarrow \text{MAINT }_{ext}$

Appendix I

Metabolic Network for *Tisochrysis Lutea*

I.1 Original Network

Table I.1: Original metabolic network for *Tisochrysis lutea*, from [2].

Flux	Metabolic reaction
Photosynthesis	
R_1	$10 \text{ Light} + 3 \text{ ADP} + 3 \text{ P}_i + \text{H} + 2 \text{ cNADP} \rightarrow \text{O}_2 + \text{H}_2\text{O} + 2 \text{ cNADPH} + 3 \text{ ATP}$
R_2	$\text{CO}_2 + \text{H}_2\text{O} + \text{cRu15DP} \rightarrow 2 \text{ cG3P}$
R_3	$\text{ATP} + \text{cG3P} \rightarrow \text{ADP} + \text{H} + \text{c13DPG}$
R_4	$\text{H} + \text{cNADPH} + \text{c13DPG} \longleftrightarrow \text{cNADP} + \text{P}_i + \text{cGAP}$
R_5	$\text{cGAP} \longleftrightarrow \text{cDHAP}$
R_6	$\text{cDHAP} + \text{cGAP} \longleftrightarrow \text{cF16P}$
R_7	$\text{H}_2\text{O} + \text{cF16P} \longleftrightarrow \text{P}_i + \text{cF6P}$
R_8	$\text{cF6P} + \text{cGAP} \longleftrightarrow \text{cE4P} + \text{cX5P}$
R_9	$\text{H}_2\text{O} + \text{cE4P} + \text{cGAP} \longleftrightarrow \text{P}_i + \text{cS7P}$
R_{10}	$\text{cGAP} + \text{cS7P} \longleftrightarrow \text{cR5P} + \text{cX5P}$
R_{11}	$\text{cX5P} \longleftrightarrow \text{cRu5P}$
R_{12}	$\text{cR5P} \longleftrightarrow \text{cRu5P}$
R_{13}	$\text{ATP} + \text{cRu5P} \rightarrow \text{ADP} + \text{H} + \text{cRu15DP}$
R_{14}	$\text{cGAP} \longleftrightarrow \text{GAP}$
Glycolysis	
R_{15}	$\text{G6P} \longleftrightarrow \text{G1P}$
R_{16}	$\text{F6P} \longleftrightarrow \text{G6P}$
R_{17}	$\text{ATP} + \text{F6P} \rightarrow \text{ADP} + \text{F16P} + \text{H}$
R_{18}	$\text{F16P} + \text{H}_2\text{O} \rightarrow \text{F6P} + \text{P}_i$
R_{19}	$\text{DHAP} + \text{GAP} \longleftrightarrow \text{F16P}$
R_{20}	$\text{DHAP} \longleftrightarrow \text{GAP}$
R_{21}	$\text{GAP} + \text{NAD} + \text{P}_i \longleftrightarrow \text{13DPG} + \text{H} + \text{NADH}$
R_{22}	$\text{13DPG} + \text{ADP} \longleftrightarrow \text{3PG} + \text{ATP}$
R_{23}	$\text{3PG} \longleftrightarrow \text{2PG}$
R_{24}	$\text{2PG} \longleftrightarrow \text{H}_2\text{O} + \text{PEP}$
R_{25}	$\text{ADP} + \text{H} + \text{PEP} \rightarrow \text{ATP} + \text{PYR}$
TCA Cycle	
R_{26}	$\text{CoA} + \text{NAD} + \text{PYR} \rightarrow \text{AcCoA} + \text{CO}_2 + \text{NADH}$
R_{27}	$\text{AcCoA} + \text{H}_2\text{O} + \text{OXA} \longleftrightarrow \text{CIT} + \text{CoA} + \text{H}$
R_{28}	$\text{CIT} + \text{NAD} \longleftrightarrow \text{AKG} + \text{CO}_2 + \text{NADH}$
R_{29}	$\text{AKG} + \text{CoA} + \text{NAD} \rightarrow \text{CO}_2 + \text{NADH} + \text{SUCCoA}$
R_{30}	$\text{ADP} + \text{P}_i + \text{SUCCoA} \longleftrightarrow \text{ATP} + \text{CoA} + \text{SUC}$

R_{31}	$\text{FAD} + \text{SUC} \longleftrightarrow \text{FADH}_2 + \text{FUM}$
R_{32}	$\text{FUM} + \text{H}_2\text{O} \longleftrightarrow \text{MAL}$
R_{33}	$\text{FAD} + \text{MAL} \longleftrightarrow \text{FADH}_2 + \text{OXA}$
R_{34}	$\text{ATP} + \text{CO}_2 + \text{H}_2\text{O} + \text{PYR} \rightarrow \text{ADP} + \text{OXA} + \text{P}_i + 2 \text{H}$
R_{35}	$\text{ATP} + \text{OXA} \rightarrow \text{ADP} + \text{CO}_2 + \text{PEP}$
R_{36}	$\text{CO}_2 + \text{H}_2\text{O} + \text{PEP} \longleftrightarrow \text{H} + \text{OXA} + \text{P}_i$
PPP	
R_{37}	$\text{G6P} + \text{H}_2\text{O} + \text{NADP} \longleftrightarrow \text{6PG} + \text{NADPH} + 2 \text{H}$
R_{38}	$\text{6PG} + \text{NADP} \longleftrightarrow \text{CO}_2 + \text{NADPH} + \text{RU5P}$
R_{39}	$\text{RU5P} \longleftrightarrow \text{R5P}$
R_{40}	$\text{RU5P} \longleftrightarrow \text{X5P}$
R_{41}	$\text{R5P} + \text{X5P} \longleftrightarrow \text{GAP} + \text{S7P}$
R_{42}	$\text{GAP} + \text{S7P} \longleftrightarrow \text{E4P} + \text{F6P}$
R_{43}	$\text{F6P} + \text{GAP} \longleftrightarrow \text{E4P} + \text{X5P}$
Glycerol Synthesis	
R_{44}	$\text{GLYC3P} + \text{NAD} \longleftrightarrow \text{DHAP} + \text{H} + \text{NADH}$
N Fixation	
R_{45}	$\text{H} + \text{NADH} + \text{NO}_3 \longleftrightarrow \text{H}_2\text{O} + \text{NAD} + \text{NO}_2$
R_{46}	$5 \text{H} + 3 \text{NADPH} + \text{NO}_2 \longleftrightarrow \text{NH}_4 + 2 \text{H}_2\text{O} + 3 \text{NADP}$
S Fixation	
R_{47}	$\text{ATP} + \text{SO}_4 \rightarrow \text{APS} + \text{PP}_i$
R_{48}	$\text{APS} + \text{NADH} \rightarrow \text{AMP} + \text{NAD} + \text{SO}_3$
R_{49}	$5 \text{H} + 3 \text{NADPH} + \text{SO}_3 \longleftrightarrow \text{H}_2\text{S} + 3 \text{H}_2\text{O} + 3 \text{NADP}$
Oxidative Phosphorylation	
R_{50}	$1.5 \text{ADP} + 1.5 \text{H} + 1.5 \text{P}_i + \text{FADH}_2 + 0.5 \text{O}_2 \rightarrow \text{FAD} + 1.5 \text{ATP} + 2.5 \text{H}_2\text{O}$
R_{51}	$2.5 \text{ADP} + 3.5 \text{H} + 2.5 \text{P}_i + \text{NADH} + 0.5 \text{O}_2 \rightarrow \text{NAD} + 2.5 \text{ATP} + 3.5 \text{H}_2\text{O}$
R_{52}	$\text{H}_2\text{O} + \text{PP}_i \rightarrow \text{H} + 2 \text{P}_i$
R_{53}	$\text{AMP} + \text{ATP} \rightarrow 2 \text{ADP}$
R_{54}	$\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{H} + \text{P}_i + \text{MAINT}$
AA Metabolism	
R_{55}	$\text{AKG} + \text{H} + \text{NADPH} + \text{NH}_4 \rightarrow \text{GLU} + \text{H}_2\text{O} + \text{NADP}$
R_{56}	$\text{ATP} + \text{GLU} + \text{NH}_4 \rightarrow \text{ADP} + \text{GLN} + \text{H} + \text{P}_i$
R_{57}	$\text{AKG} + \text{GLN} + \text{H} + \text{NADPH} \longleftrightarrow \text{NADP} + 2 \text{GLU}$
R_{58}	$3 \text{PG} + \text{GLU} + \text{H}_2\text{O} + \text{NAD} \longleftrightarrow \text{AKG} + \text{H} + \text{NADH} + \text{P}_i + \text{SER}$
R_{59}	$\text{SER} \rightarrow \text{NH}_4 + \text{PYR}$
R_{60}	$\text{AcCoA} + \text{H}_2\text{S} + \text{SER} \longleftrightarrow \text{Ace} + \text{CYS} + \text{CoA} + \text{H}$
R_{61}	$\text{ATP} + \text{Ace} + \text{CoA} \rightarrow \text{ADP} + \text{AcCoA} + \text{P}_i$
R_{62}	$\text{GLU} + \text{PYR} \rightarrow \text{AKG} + \text{ALA}$
R_{63}	$\text{H} + \text{THR} \longleftrightarrow 2\text{-oxobutan} + \text{NH}_4$
R_{64}	$2\text{-oxobutan} + \text{GLU} + \text{H} + \text{NADPH} + \text{PYR} \longleftrightarrow \text{AKG} + \text{CO}_2 + \text{H}_2\text{O} + \text{ILE} + \text{NADP}$
R_{65}	$2 \text{H} + \text{ALA} + \text{NADPH} + \text{PYR} \longleftrightarrow \text{CO}_2 + \text{H}_2\text{O} + \text{NADP} + \text{VAL}$
R_{66}	$2 \text{PYR} + \text{AcCoA} + \text{GLU} + \text{H} + \text{NAD} + \text{NADPH} \longleftrightarrow \text{AKG} + \text{CoA} + \text{LEU} + \text{NADH} + \text{NADP} + 2 \text{CO}_2$
R_{67}	$2 \text{PEP} + \text{ATP} + \text{E4P} + \text{NADPH} \rightarrow \text{ADP} + \text{CHO} + \text{NADP} + 4 \text{P}_i$
R_{68}	$\text{CHO} \longleftrightarrow \text{PRE}$
R_{69}	$\text{GLU} + \text{H} + \text{PRE} \longleftrightarrow \text{AKG} + \text{CO}_2 + \text{H}_2\text{O} + \text{PHE}$
R_{70}	$\text{GLU} + \text{NAD} + \text{PRE} \longleftrightarrow \text{AKG} + \text{CO}_2 + \text{NADH} + \text{TYR}$
R_{71}	$\text{CHO} + \text{GLN} \longleftrightarrow \text{ANTH} + \text{GLU} + \text{H} + \text{PYR}$
R_{72}	$\text{ANTH} + \text{H} + \text{PRPP} + \text{SER} \longleftrightarrow \text{CO}_2 + \text{GAP} + \text{PP}_i + \text{TRYP} + 2 \text{H}_2\text{O}$
R_{73}	$3 \text{H}_2\text{O} + 2 \text{NAD} + \text{ATP} + \text{GLN} + \text{PRPP} \rightarrow \text{AICAR} + \text{AKG} + \text{HIS} + \text{P}_i + 2 \text{NADH} + 2 \text{PP}_i + 5 \text{H}$
R_{74}	$\text{GLU} + \text{OXA} \longleftrightarrow \text{AKG} + \text{ASP}$
R_{75}	$\text{ASP} + \text{ATP} + \text{GLN} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{ASN} + \text{GLU} + \text{H} + \text{P}_i$
R_{76}	$2 \text{ATP} + 2 \text{H}_2\text{O} + \text{CO}_2 + \text{GLN} \rightarrow \text{CaP} + \text{GLU} + \text{P}_i + 2 \text{ADP} + 3 \text{H}$
R_{77}	$2 \text{GLU} + \text{ASP} + \text{ATP} + \text{CaP} + \text{NADH} \rightarrow \text{AKG} + \text{AMP} + \text{ARG} + \text{FUM} + \text{H}_2\text{O} + \text{NAD} + \text{PP}_i + \text{P}_i$
R_{78}	$3 \text{H} + 2 \text{NADH} + \text{GLU} \longleftrightarrow \text{PRO} + 2 \text{H}_2\text{O} + 2 \text{NAD}$
R_{79}	$\text{AKG} + \text{O}_2 + \text{PRO} \longleftrightarrow \text{CO}_2 + \text{HydPro} + \text{SUC}$
R_{80}	$\text{ASP} + \text{ATP} + \text{H} + \text{NADPH} \rightarrow \text{ADP} + \text{ASA} + \text{NADP} + \text{P}_i$
R_{81}	$2 \text{H} + \text{ASA} + \text{GLU} + \text{NADH} + \text{PYR} \longleftrightarrow \text{AKG} + \text{DAP} + \text{H}_2\text{O} + \text{NAD}$
R_{82}	$\text{DAP} \longleftrightarrow \text{CO}_2 + \text{H} + \text{LYS}$
R_{83}	$\text{ASA} + \text{H} + \text{NADPH} \longleftrightarrow \text{HSER} + \text{NADP}$
R_{84}	$\text{ATP} + \text{H}_2\text{O} + \text{HSER} \rightarrow \text{ADP} + \text{H} + \text{P}_i + \text{THR}$

Appendix I – Metabolic Network for Tisochrysis Lutea

R_{85}	$\text{AcCoA} + \text{CYS} + \text{H}_2\text{O} + \text{HSER} \longleftrightarrow \text{Ace} + \text{CoA} + \text{HCYS} + \text{H} + \text{NH}_4 + \text{PYR}$
R_{86}	$\text{HCYS} + \text{MTHF} \longleftrightarrow \text{H} + \text{MET} + \text{THF}$
R_{87}	$4.306 \text{ ATP} + 3.306 \text{ H}_2\text{O} + 0.111 \text{ ALA} + 0.092 \text{ GLY} + 0.09 \text{ LEU} + 0.061 \text{ VAL} + 0.06 \text{ LYS} + 0.056 \text{ PRO} + 0.056 \text{ THR} + 0.054 \text{ SER} + 0.052 \text{ ARG} + 0.052 \text{ GLN} + 0.052 \text{ GLU} + 0.047 \text{ ASN} + 0.047 \text{ ASP} + 0.041 \text{ PHE} + 0.037 \text{ ILE} + 0.03 \text{ TYR} + 0.024 \text{ MET} + 0.017 \text{ HIS} + 0.012 \text{ CYS} + 0.0090 \text{ HydPro} + 0.0010 \text{ TRYP} \rightarrow \text{PROTEIN} + 4.306 \text{ ADP} + 4.306 \text{ P}_i + 4.319 \text{ H}$
R_{88}	$\text{GLY} + \text{H} + \text{PYR} \longleftrightarrow \text{ALA} + \text{glyoxylate}$
R_{89}	$\text{SER} + \text{glyoxylate} \longleftrightarrow \text{GLY} + \text{HydPyr}$
R_{90}	$\text{GLY} + \text{H}_2\text{O} + \text{METHF} \longleftrightarrow \text{SER} + \text{THF}$
R_{91}	$\text{GLY} + \text{NAD} + \text{THF} \longleftrightarrow \text{CO}_2 + \text{METHF} + \text{NADH} + \text{NH}_4$
R_{92}	$\text{H} + \text{HydPyr} + \text{NADH} \longleftrightarrow \text{Glycerate} + \text{NAD}$
R_{93}	$\text{ATP} + \text{Glycerate} \rightarrow \text{ADP} + 2 \text{ H} + 3 \text{ PG}$

THF Metabolism

R_{94}	$\text{ATP} + \text{R5P} \rightarrow \text{AMP} + \text{H} + \text{PRPP}$
R_{95}	$5\text{FTHF} + \text{H} \longleftrightarrow \text{H}_2\text{O} + \text{MYLTHF}$
R_{96}	$\text{H}_2\text{O} + \text{MYLTHF} \longleftrightarrow \text{H} + \text{N10FTHF}$
R_{97}	$\text{ATP} + \text{FORM} + \text{THF} \rightarrow \text{ADP} + \text{N10FTHF} + \text{P}_i$
R_{98}	$\text{MYLTHF} + \text{NADPH} \longleftrightarrow \text{METHF} + \text{NADP}$
R_{99}	$\text{H} + \text{METHF} + \text{NADPH} \longleftrightarrow \text{MTHF} + \text{NADP}$
R_{100}	$5\text{FTHF} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{H} + \text{N10FTHF} + \text{P}_i$
R_{101}	$\text{FORM} + \text{H} + \text{THF} \longleftrightarrow \text{H}_2\text{O} + \text{N10FTHF}$
R_{102}	$\text{DHF} + \text{H} + \text{NADPH} \longleftrightarrow \text{NADP} + \text{THF}$

Lipid Synthesis

R_{103}	$\text{ACP} + \text{AcCoA} + \text{H} \longleftrightarrow \text{AcACP} + \text{CoA}$
R_{104}	$\text{ATP} + \text{AcCoA} + \text{CO}_2 + \text{H}_2\text{O} \longleftrightarrow \text{ADP} + \text{H} + \text{MalCoA} + \text{P}_i$
R_{105}	$\text{ACP} + \text{MalCoA} \longleftrightarrow \text{CoA} + \text{MalACP}$
R_{106}	$12 \text{ H} + 12 \text{ NADPH} + 6 \text{ MalACP} + \text{AcACP} \longleftrightarrow \text{C14:0ACP} + 6 \text{ ACP} + 6 \text{ CO}_2 + 6 \text{ H}_2\text{O} + 12 \text{ NADP}$
R_{107}	$\text{C14:0ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C14:1ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{108}	$14 \text{ H} + 14 \text{ NADPH} + 7 \text{ MalACP} + \text{AcACP} \longleftrightarrow \text{C16:0ACP} + 7 \text{ ACP} + 7 \text{ CO}_2 + 7 \text{ H}_2\text{O} + 14 \text{ NADP}$
R_{109}	$\text{C16:0ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C16:1ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{110}	$\text{C16:1ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C16:2ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{111}	$\text{C16:2ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C16:3ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{112}	$16 \text{ H} + 16 \text{ NADPH} + 8 \text{ MalACP} + \text{AcACP} \longleftrightarrow \text{C18:0ACP} + 8 \text{ ACP} + 8 \text{ CO}_2 + 8 \text{ H}_2\text{O} + 16 \text{ NADP}$
R_{113}	$\text{C18:0ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C18:1ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{114}	$\text{C18:1ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C18:2ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{115}	$\text{C18:2ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C18:3ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{116}	$\text{C18:3ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C18:4ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{117}	$18 \text{ H} + 18 \text{ NADPH} + 9 \text{ MalACP} + \text{AcACP} \longleftrightarrow \text{C20:0ACP} + 9 \text{ ACP} + 9 \text{ CO}_2 + 9 \text{ H}_2\text{O} + 18 \text{ NADP}$
R_{118}	$\text{C20:0ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C20:1ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{119}	$\text{C20:1ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C20:2ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{120}	$\text{C20:2ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C20:3ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{121}	$\text{C20:3ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C20:4ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{122}	$\text{C20:4ACP} + \text{H} + \text{NADH} + \text{O}_2 \rightarrow \text{C20:5ACP} + \text{NAD} + 2 \text{ H}_2\text{O}$
R_{123}	$\text{GLYC3P} + 0.524 \text{ C14:0ACP} + 0.02 \text{ C14:1ACP} + 0.222 \text{ C16:0ACP} + 0.096 \text{ C16:1ACP} + 0.016 \text{ C16:3ACP} + 0.012 \text{ C18:0ACP} + 0.724 \text{ C18:1ACP} + 0.106 \text{ C18:2ACP} + 0.102 \text{ C18:3ACP} + 0.142 \text{ C18:4ACP} + 0.036 \text{ C20:5ACP} \longleftrightarrow \text{PA} + 2 \text{ ACP} + 2 \text{ H}$

Nucleic Acid Synthesis

R_{124}	$4 \text{ ATP} + 2 \text{ GLN} + 2 \text{ H}_2\text{O} + \text{ASP} + \text{CO}_2 + \text{GLY} + \text{N10FTHF} + \text{PRPP} \rightarrow \text{AICAR} + \text{FUM} + \text{PP}_i + \text{THF} + 2 \text{ GLU} + 4 \text{ ADP} + 4 \text{ P}_i + 7 \text{ H}$
R_{125}	$\text{ASP} + \text{CaP} + \text{H} + \text{O}_2 + \text{PRPP} \longleftrightarrow \text{CO}_2 + \text{H}_2\text{O} + \text{H}_2\text{O}_2 + \text{PP}_i + \text{P}_i + \text{UMP}$
R_{126}	$2 \text{ H}_2\text{O}_2 \longleftrightarrow \text{O}_2 + 2 \text{ H}_2\text{O}$
R_{127}	$\text{ATP} + \text{UMP} \rightarrow \text{ADP} + \text{UDP}$
R_{128}	$\text{ATP} + \text{UDP} \longleftrightarrow \text{ADP} + \text{UTP}$
R_{129}	$\text{ATP} + \text{GLN} + \text{H}_2\text{O} + \text{UTP} \rightarrow \text{ADP} + \text{CTP} + \text{GLU} + \text{P}_i + 2 \text{ H}$
R_{130}	$\text{ATP} + \text{CDP} \longleftrightarrow \text{ADP} + \text{CTP}$
R_{131}	$\text{AICAR} + \text{N10FTHF} \longleftrightarrow \text{H}_2\text{O} + \text{IMP} + \text{THF}$
R_{132}	$\text{ATP} + \text{H}_2\text{O} + \text{IMP} + \text{NAD} + \text{NH}_4 \rightarrow \text{AMP} + \text{GMP} + \text{NADH} + \text{PP}_i + 3 \text{ H}$
R_{133}	$\text{ATP} + \text{GMP} \rightarrow \text{ADP} + \text{GDP}$
R_{134}	$\text{ATP} + \text{GDP} \longleftrightarrow \text{ADP} + \text{GTP}$
R_{135}	$\text{ASP} + \text{GTP} + \text{IMP} \longleftrightarrow \text{AMP} + \text{FUM} + \text{GDP} + \text{P}_i + 2 \text{ H}$
R_{136}	$\text{ATP} + \text{H} + \text{METHF} + \text{NADPH} + \text{UDP} \rightarrow \text{ADP} + \text{DHF} + \text{H}_2\text{O} + \text{NADP} + \text{dTTP}$
R_{137}	$\text{ATP} + \text{CDP} + \text{H} + \text{NADPH} \rightarrow \text{ADP} + \text{H}_2\text{O} + \text{NADP} + \text{dCTP}$

R_{138}	$\text{ATP} + \text{GDP} + \text{H} + \text{NADPH} \rightarrow \text{ADP} + \text{H}_2\text{O} + \text{NADP} + \text{dGTP}$
R_{139}	$\text{ATP} + \text{H} + \text{NADPH} \leftrightarrow \text{H}_2\text{O} + \text{NADP} + \text{dATP}$
R_{140}	$2.372 \text{ H}_2\text{O} + 1.372 \text{ ATP} + 0.18 \text{ dATP} + 0.18 \text{ dTTP} + 0.32 \text{ dCTP} + 0.32 \text{ dGTP} \rightarrow \text{DNA} + \text{PP}_i + 1.372 \text{ ADP} + 1.372 \text{ P}_i + 2.372 \text{ H}$
R_{141}	$1.4 \text{ H}_2\text{O} + 0.56 \text{ ATP} + 0.34 \text{ GTP} + 0.16 \text{ UTP} + 0.34 \text{ CTP} \rightarrow 0.4 \text{ ADP} + 0.4 \text{ H} + 0.4 \text{ P}_i + \text{PP}_i + \text{RNA}$
Chlorophyll Synthesis	
R_{142}	$12 \text{ H} + 8 \text{ ATP} + 8 \text{ GLU} + 8 \text{ NADPH} + 2.5 \text{ O}_2 \rightarrow \text{PPorphyrin} + 4 \text{ NH}_4 + 6 \text{ CO}_2 + 8 \text{ AMP} + 8 \text{ NADP} + 8 \text{ PP}_i + 13 \text{ H}_2\text{O}$
R_{143}	$18 \text{ H} + 15 \text{ NADPH} + 8 \text{ ATP} + 4 \text{ GAP} + 4 \text{ PYR} \rightarrow \text{Phytyl-PP} + 4 \text{ ADP} + 4 \text{ AMP} + 4 \text{ CO}_2 + 7 \text{ PP}_i + 8 \text{ H}_2\text{O} + 15 \text{ NADP}$
R_{144}	$\text{ATP} + \text{H}_2\text{O} + \text{MET} \rightarrow \text{AdMET} + \text{H} + \text{PP}_i + \text{P}_i$
R_{145}	$\text{AdHCYS} + \text{H}_2\text{O} \leftrightarrow \text{Ad} + \text{HCYS}$
R_{146}	$\text{ATP} + \text{Ad} \rightarrow \text{ADP} + \text{AMP} + \text{H}$
R_{147}	$4 \text{ NADPH} + 2.5 \text{ O}_2 + 2 \text{ ATP} + \text{AdMET} + \text{Mg}_2 + \text{PPorphyrin} + \text{Phytyl-PP} \rightarrow \text{AdHCYS} + \text{Chlorophyll} + \text{PP}_i + 2 \text{ ADP} + 2 \text{ H}_2\text{O} + 2 \text{ P}_i + 3 \text{ H} + 4 \text{ NADP}$
Carbohydrate Synthesis	
R_{148}	$\text{G1P} \leftrightarrow \text{CARB} + \text{P}_i$
Biomass Synthesis	
R_{149}	$32.687 \text{ ATP} + 32.687 \text{ H}_2\text{O} + 0.8456 \text{ PROTEIN} + 0.0026 \text{ Chlorophyll} + 0.0402 \text{ RNA} + 0.0014 \text{ DNA} + 0.1103 \text{ PA} \rightarrow \text{B} + 32.687 \text{ H} + 32.687 \text{ ADP} + 32.687 \text{ P}_i$
Transport Reactions	
R_{150}	$\text{CO}_{2\text{ext}} \leftrightarrow \text{CO}_2$
R_{151}	$\text{O}_{2\text{ext}} \leftrightarrow \text{O}_2$
R_{152}	$\text{H}_2\text{O}_{\text{ext}} \leftrightarrow \text{H}_2\text{O}$
R_{153}	$\text{P}_{i\text{ext}} \leftrightarrow \text{P}_i$
R_{154}	$\text{SO}_{4\text{ext}} \leftrightarrow \text{SO}_4$
R_{155}	$\text{NO}_{3\text{ext}} \leftrightarrow \text{NO}_3$
R_{156}	$\text{Mg}_{2\text{ext}} \leftrightarrow \text{Mg}_2$
R_{157}	$\text{Light}_{\text{ext}} \rightarrow \text{Light}$
R_{158}	$\text{H}_{\text{ext}} \leftrightarrow \text{H}$
R_{159}	$\text{B} \rightarrow \text{B}_{\text{ext}}$
R_{160}	$\text{PA}_{\text{ext}} \leftrightarrow \text{PA}$
R_{161}	$\text{CARB}_{\text{ext}} \leftrightarrow \text{CARB}$
R_{162}	$\text{MAINT} \rightarrow \text{MAINT}_{\text{ext}}$

I.2 Extended Network

Table I.2: Additional reactions for the pathway of carbohydrate synthesis of *Tisochrysis lutea* (First iteration).

Flux	Metabolic reaction
Carbohydrate Synthesis	
R_{163}	$\text{glucose-1-P} + \text{UTP} \rightarrow \text{UDP-glucose} + \text{PP}_i$
R_{164}	$\text{UDP-glucose} + (\beta - 1, 3\text{-glucan})_n \rightarrow (\beta - 1, 3\text{-glucan})_{n+1} + \text{UDP}$
R_{165}	$\text{fructose-6-P} + \text{glutamine} \rightarrow \text{glucosamine-6-P} + \text{glutamate}$
R_{166}	$\text{glucosamine-6-P} + \text{acetyl-CoA} \rightarrow \text{N-acetylglucosamine-6-P}$
R_{167}	$\text{N-acetylglucosamine-6-P} \leftrightarrow \text{N-acetylglucosamine-1-P}$
R_{168}	$\text{N-acetylglucosamine-1-P} + \text{UTP} \rightarrow \text{UDP-N-acetylglucosamine} + \text{PP}_i$
R_{169}	$\text{fructose-6-P} \leftrightarrow \text{mannose-6-P}$
R_{170}	$\text{mannose-6-P} \rightarrow \text{mannose-1-P}$
R_{171}	$\text{mannose-1-P} + \text{GTP} \rightarrow \text{GDP-mannose} + \text{PP}_i$
R_{172}	$\text{fructose-6-P} + \text{NADPH}, \text{H}^+ \rightarrow \text{mannitol-1-P} + \text{NADP}^+$
R_{173}	$\text{mannitol-1-P} + \text{H}_2\text{O} \rightarrow \text{mannitol}$
R_{148}^*	$0.7 (\beta - 1, 3\text{-glucan})_{n+1} + 0.2 \text{ mannitol-1-P} + 0.05 \text{ UDP-N-acetylglucosamine} + 0.05 \text{ GDP-mannose} \rightarrow \text{carbohydrate}_{av} + 0.75 \text{ UDP} + 0.2 \text{ P}_i + 0.05 \text{ GDP}$

As mentioned, the previous chemical reactions were selected based on genomic inference and experimental evidence from haptophyte microalgae and represent the main pathways leading to the synthesis of storage carbohydrates (chrysolaminarin), soluble sugars (mannitol), and glycan precursors.

Specifically, R_{163} is catalyzed by UDP-glucose pyrophosphorylase (UGPase). UDP-glucose is the activated sugar donor for polysaccharide assembly. Experimental studies in diatoms have shown that UGPase activity is essential for chrysolaminarin (β -1,3-glucan) synthesis [268]. R_{164} extends the storage polysaccharide chain by transferring glucose from UDP-glucose to a growing β -1,3-glucan with some β -1,6 branches, which constitutes the main carbon reserve in *Tisochrysis lutea*, as in related haptophytes. R_{165} is an amidotransferase reaction which begins with the biosynthesis of amino sugar nucleotides. It is genomically conserved across microalgae and serves as precursor step to UDP-N-acetylglucosamine formation, essential for glycoconjugates [269]. R_{166} represents an acetylation reaction to form N-acetylglucosamine-6-P. It is a key step where an acetyl group is transferred to GlcN-6P, contributing to the biosynthesis of N-acetylated sugars used in the cell wall or glucan structures [269]. In R_{167} , phosphoglucosamine mutase repositions the phosphate group to prepare for nucleotide activation. This step is often ATP-dependent and reversible [217]. In R_{168} , the enzyme UDP-N-acetylglucosamine pyrophosphorylase produces UDP-N-acetylglucosamine, a central precursor for glycosylation, including N-linked glycans in eukaryotic microalgae [269]. R_{169} , R_{170} and R_{171} are three necessary steps to convert fructose-6-P into GDP-mannose, a nucleotide-sugar donor used in the formation of mannose-containing polysaccharides [1, 219]. Finally, R_{172} and R_{173} , catalyzed by mannitol-1P dehydrogenase and mannitol-1P phosphatase, respectively, generate mannitol, a sugar alcohol used as an additional carbon reserve in *T. lutea* and related algae [270]. Note that most of the previous reactions are also found in KeGG pathway *map00051*.

Table I.3: Additional reactions for the pathway of carbohydrate synthesis of *Tisochrysis lutea* (Second iteration).

Flux	Metabolic reaction
Carbohydrate Synthesis	
R_{163}	$\text{glucose-1-P} + \text{UTP} \rightarrow \text{UDP-glucose} + \text{PP}_i$
R_{164}	$\text{UDP-glucose} \rightarrow (\beta-1,3\text{-glucan}) + \text{UDP}$
R_{165}	$\text{fructose-6-P} + \text{glutamine} + \text{acetyl-CoA} + \text{UTP} \rightarrow \text{UDP-N-acetylglucosamine} + \text{glutamate} + \text{CoASH} + \text{PP}_i$
R_{166}	$\text{fructose-6-P} + \text{GTP} \rightarrow \text{GDP-mannose} + \text{PP}_i$
R_{167}	$\text{fructose-6-P} + \text{NADPH, H}^+ \rightarrow \text{mannitol-1-P} + \text{NADP}^+$
R_{148}^*	$0.7 (\beta-1,3\text{-glucan}) + 0.2 \text{mannitol-1-P} + 0.05 \text{UDP-N-acetylglucosamine} + 0.05 \text{GDP-mannose} \rightarrow \text{carbohydrate}_{av} + 0.75 \text{UDP} + 0.2 \text{P}_i + 0.05 \text{GDP}$

Appendix J

Metabolic Compartmentalization

Metabolic compartmentalization plays a crucial role in the spatial and functional organization of cellular metabolism, allowing distinct biochemical processes to occur simultaneously without interference. In eukaryotes, this is achieved primarily through membrane-bound organelles, while in prokaryotes, compartmentalization can occur via protein-based microcompartments or membrane-associated systems. This compartmental structure is especially relevant in metabolic modeling, where accurately representing the localization of reactions is essential for realistic simulations and predictions. To maintain fidelity, metabolic networks must explicitly define transport reactions that facilitate the exchange of metabolites between compartments, such as the cytosol, mitochondria, or extracellular space. These transport steps are essential for maintaining mass balance and significantly influence metabolic fluxes and system behavior, underscoring their importance in model construction and analysis. The main metabolic pathways are summarized in Table J.1, which compares their organization between prokaryotes and eukaryotic cell types, including algae, plant cells, and animal cells.

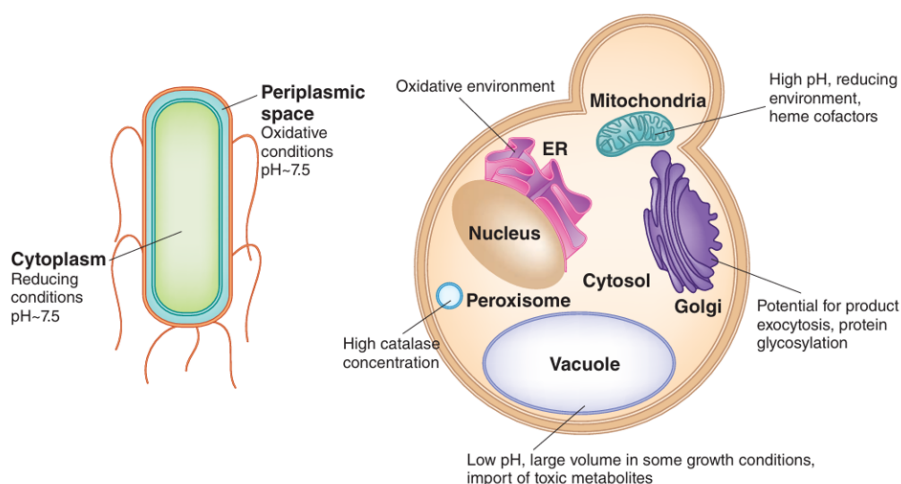


Figure J.1: Left-hand side: Prokaryotic cells (e.g., bacteria such as *E. coli*) have a simple, two-compartment system. Right-hand side: Eukaryotic organisms offer many specialized organelles, each with conditions distinct from the cytosol, that may provide a better environment for a range of metabolic pathways. From [4]. As explained in Sec. 2.1, glyoxysomes are a specialized form of peroxisomes.

Table J.1: Overview of the main compartmentalized metabolic pathways (catabolic and anabolic) across different cell types.

Pathways	Prokaryotes	Eukaryotes		
	Bacteria	Algae	Plant Cells	Animal Cells
Glycolysis	Cytoplasm	Cytosol	Cytosol	Cytosol
PPP	Cytoplasm	Cytosol	Cytosol	Cytosol
TCA Cycle	Cytoplasm	Mitochondrial matrix	Mitochondrial matrix	Mitochondrial matrix
Pyruvate Oxidation	Cytoplasm	Mitochondrial matrix	Mitochondrial matrix	Mitochondrial matrix
Fermentation	Cytoplasm	Cytosol	Cytosol (roots)	Cytosol
Urea Cycle or Analogs	Cytoplasm (GS-GOGAT or GDH)	Chloroplast, Cytosol (urea degradation)	Peroxisomes, Cytosol (urea degradation)	Cytosol and Mitochondria (Liver)
GS-GOGAT	Cytoplasm	Chloroplast	Chloroplast, Cytosol	Cytosol
GDH	Cytoplasm	Cytosol, Mitochondria	Cytosol, Mitochondria	Mitochondria
Photosynthesis	Thylakoid membranes	Thylakoid membranes of the chloroplasts	Thylakoid membranes of the chloroplasts	×
Calvin Cycle (CO ₂ Fixation)	Carboxysomes, Cytoplasm	Stroma of chloroplasts	Stroma of chloroplasts	×
Photorespiration	×	Chloroplast, Mitochondria, Peroxisomes	Chloroplast, Mitochondria, Peroxisomes	×
ETC	Plasma membrane	Mitochondrial inner membrane	Mitochondrial inner membrane	Mitochondrial inner membrane
Oxidative Phosphorylation	Plasma membrane	Mitochondrial inner membrane	Mitochondrial inner membrane	Mitochondrial inner membrane
Photophosphorylation	Thylakoid-like membranes	Thylakoid membranes of the chloroplasts	Thylakoid membranes of the chloroplasts	×
Glyoxylate Cycle	Cytoplasm	Glyoxysome-like organelles	Glyoxysomes	×
Fatty Acid Biosynthesis	Cytoplasm	Cytosol	Cytosol	Cytosol
Fatty Acid Degradation (β -oxidation)	Cytoplasm	Mitochondria, Peroxisomes	Peroxisomes, Mitochondria	Mitochondria, Peroxisomes
Protein Synthesis	Ribosomes (cytoplasm)	Ribosomes (cytosol and ER)	Ribosomes (cytosol and ER)	Ribosomes (cytosol and ER)
DNA and RNA Synthesis	Cytoplasm (nucleoid)	Nucleus (DNA), Cytosol (RNA)	Nucleus (DNA), Cytosol (RNA)	Nucleus (DNA), Cytosol (RNA)
Nucleotide Synthesis	Cytoplasm	Cytosol and Nucleus	Cytosol and Nucleus	Cytosol and Nucleus
Amino Acid Synthesis	Cytoplasm	Chloroplast, Cytosol	Chloroplast, Cytosol	Cytosol
Polysaccharide Biosynthesis	Cytoplasm, plasma membrane	Cytosol, Golgi	Cytosol, Golgi, ER	Cytosol, Golgi, ER

Appendix K

Additional Metabolic Reactions for the Metabolism of Purple Non-Sulfur Bacteria in *Rhodospirillum Rubrum*

The metabolic reactions presented in Table K.1 result from the application of the constructive methodology described in Sec. 4.2. This approach incorporates findings from the studies conducted in [271–275], enabling the reconstruction of a context-specific reaction network.

Table K.1: Additional metabolic reactions for the metabolism of purple non-sulfur bacteria in *Rhodospirillum rubrum*.

Flux	Metabolic reaction
v_1	$\text{CoASH} + \text{acetoacetyl-CoA} \longleftrightarrow 2 \text{ acetyl-CoA}$
v_2	$\text{succinyl-CoA} + \text{acetate} \rightarrow \text{acetyl-CoA} + \text{succinate}$
v_3	$\text{oxaloacetate} \rightarrow \text{pyruvate} + \text{CO}_2$
v_4	$\text{acetyl-CoA} + \text{butyrate} \rightarrow \text{butyryl-CoA} + \text{acetate}$
v_5	$\text{butyryl-CoA} \rightarrow \text{crotonyl-CoA}$
v_6	$\text{propanoyl-CoA} + \text{oxaloacetate} + \text{H}_2\text{O} \rightarrow 2\text{-methylcitrate} + \text{CoASH}$
v_7	$2\text{-methylcitrate} + \text{H}_2\text{O} \rightarrow 2\text{-methylcitrate-cis-aonitate}$
v_8	$2\text{-methylcitrate-cis-aonitate} + \text{H}_2\text{O} \rightarrow 2\text{-methylisocitrate}$
v_9	$2\text{-methylisocitrate} \rightarrow \text{pyruvate} + \text{succinate}$
v_{10}	$\text{acetyl-CoA} + \text{propionyl-CoA} \longleftrightarrow 3\text{-ketovaleryl-CoA}$
v_{11}	$3\text{-ketovaleryl-CoA} + \text{NADH,H}^+ \longleftrightarrow (\text{R})\text{-3-hydroxyvaleryl-CoA} + \text{NAD}^+$
v_{12}	$(\text{S})\text{-3-hydroxyhexanoyl-CoA} \longleftrightarrow (\text{R})\text{-3-hydroxyhexanoyl-CoA}$
v_{13}	$\text{acetyl-CoA} + \text{valerate} \rightarrow \text{valeryl-CoA} + \text{acetate}$
v_{14}	$\text{valeryl-CoA} \rightarrow \text{acetyl-CoA} + \text{propionyl-CoA}$
TCA Cycle	
v_{15}	$\alpha\text{-ketoglutarate} + \text{NAD}^+ + \text{CoASH} \rightarrow \text{succinyl-CoA} + \text{NADH,H}^+ + \text{CO}_2$
Ethylmalonyl-CoA Pathway	
v_{16}	$(\text{S})\text{-3-hydroxybutanoyl-CoA} + \text{NADP}^+ \longleftrightarrow \text{acetoacetyl-CoA} + \text{NADPH,H}^+$
v_{17}	$(\text{S})\text{-3-hydroxybutanoyl-CoA} \longleftrightarrow \text{crotonoyl-CoA} + \text{H}_2\text{O}$
v_{18}	$(\text{R})\text{-3-hydroxybutanoyl-CoA} \longleftrightarrow \text{crotonoyl-CoA} + \text{H}_2\text{O}$
v_{19}	$(\text{R})\text{-3-hydroxybutanoyl-CoA} + \text{NADP}^+ \longleftrightarrow \text{acetoacetyl-CoA} + \text{NADPH,H}^+$
v_{20}	$(2\text{S})\text{-ethylmalonyl-CoA} + \text{NADP}^+ \longleftrightarrow \text{crotonoyl-CoA} + \text{CO}_2 + \text{NADPH,H}^+$
v_{21}	$(2\text{S})\text{-ethylmalonyl-CoA} \longleftrightarrow (2\text{R})\text{-ethylmalonyl-CoA}$

Appendix K – Additional Metabolic Reactions for the Metabolism of Purple Non-Sulfur Bacteria in *Rhodospirillum Rubrum*

v_{22}	$(2R)\text{-ethylmalonyl-CoA} \longleftrightarrow (2S)\text{-methylsuccinyl-CoA}$
v_{23}	$(2S)\text{-methylsuccinyl-CoA} \longleftrightarrow \text{mesaconyl-CoA}$
v_{24}	$\text{mesaconyl-CoA} + \text{H}_2\text{O} \rightarrow \text{methylmalyl-CoA}$
v_{25}	$\text{methylmalyl-CoA} \rightarrow \text{propanoyl-CoA} + \text{glyoxylate}$
v_{26}	$\text{acetyl-CoA} + \text{propionate} \rightarrow \text{propanoyl-CoA} + \text{acetate}$
v_{27}	$\text{ATP} + \text{propanoyl-CoA} + \text{HCO}_3^- \rightarrow \text{ADP} + \text{P}_i + (S)\text{-methylmalonyl-CoA}$
v_{28}	$(R)\text{-methylmalonyl-CoA} \longleftrightarrow (S)\text{-methylmalonyl-CoA}$
v_{29}	$(R)\text{-methylmalonyl-CoA} \longleftrightarrow \text{succinyl-CoA}$
v_{30}	$\text{acetyl-CoA} + \text{glyoxylate} \rightarrow \text{L-malyl-CoA}$
v_{31}	$\text{L-malyl-CoA} \longleftrightarrow \text{CoASH} + \text{malate}$

Degradation Pathway

v_{32}	$2 \text{Fe}_{red} + \text{acetyl-CoA} + \text{CO}_2 + 2 \text{H}^+ \rightarrow \text{pyruvate} + 2 \text{Fe}_{ox} + \text{CoASH}$
v_{33}	$\text{pyruvate} + \text{acetyl-CoA} + \text{H}_2\text{O} \rightarrow \text{D-citramalate} + \text{CoASH}$
v_{34}	$\text{D-citramalate} \longleftrightarrow 2\text{-erythro-3-methylmalate} + \text{H}_2\text{O}$
v_{35}	$2\text{-erythro-3-methylmalate} + \text{NAD}^+ \longleftrightarrow 2\text{-oxobutanoate} + \text{CO}_2 + \text{NADH,H}^+$
v_{36}	$\text{pyruvate} + 2\text{-oxobutanoate} \rightarrow (S)\text{-2-aceto-2-hydroxybutanoate} + \text{CO}_2$
v_{37}	$(S)\text{-2-aceto-2-hydroxybutanoate} + \text{NADPH,H}^+ \longleftrightarrow (R)\text{-2,3-dihydroxy-3-methylpentanoate} + \text{NADP}^+$
v_{38}	$(R)\text{-2,3-dihydroxy-3-methylpentanoate} \longleftrightarrow (S)\text{-3-methyl-2-oxopentanoate} + \text{H}_2\text{O}$
v_{39}	$(S)\text{-3-methyl-2-oxopentanoate} + \text{CoASH} \rightarrow (S)\text{-2-methylbutanoyl-CoA} + \text{CO}_2$
v_{40}	$(S)\text{-3-methyl-2-oxopentanoate} + \text{L-glutamate} \rightarrow \text{L-isoleucine} + 2\text{-oxoglutarate}$
v_{41}	$(S)\text{-2-methylbutanoyl-CoA} + \text{FAD}^+ \longleftrightarrow 2\text{-methyl-but-2-enoyl-CoA} + \text{FADH}_2$
v_{42}	$2\text{-methyl-but-2-enoyl-CoA} + \text{H}_2\text{O} \longleftrightarrow 2\text{-methyl-acetoacetyl-CoA}$
v_{43}	$2\text{-methyl-acetoacetyl-CoA} + \text{NAD}^+ \longleftrightarrow (S)\text{-3-hydroxy-2-methylbutyryl-CoA} + \text{NADH,H}^+$
v_{44}	$(S)\text{-3-hydroxy-2-methylbutyryl-CoA} + \text{CoASH} \rightarrow \text{acetyl-CoA} + \text{propanoyl-CoA}$

β -oxidation Pathway

v_{45}	$\text{acetyl-CoA} + \text{hexanoate} \rightarrow \text{hexanoyl-CoA} + \text{acetate}$
v_{46}	$\text{hexanoyl-CoA} + \text{FAD}^+ \longleftrightarrow \text{hex-2-enoyl-CoA} + \text{FADH}_2$
v_{47}	$\text{hex-2-enoyl-CoA} + \text{H}_2\text{O} \longleftrightarrow (S)\text{-3-hydroxyhexanoyl-CoA}$
v_{48}	$(S)\text{-3-hydroxyhexanoyl-CoA} + \text{NAD}^+ \rightarrow 3\text{-oxohexanoyl-CoA} + \text{NADH,H}^+$
v_{49}	$\text{CoASH} + 3\text{-oxohexanoyl-CoA} \rightarrow \text{acetyl-CoA} + \text{butyryl-CoA}$

Fatty Acid Synthesis

v_{50}	$\text{acetyl-CoA} + \text{CO}_2 \longleftrightarrow \text{CoASH} + \text{malonyl-CoA}$
v_{51}	$\text{malonyl-CoA} + \text{ACP} \longleftrightarrow \text{CoASH} + \text{malonyl-ACP}$
v_{52}	$\text{malonyl-ACP} + \text{acetyl-ACP} \longleftrightarrow \text{acetoacetyl-ACP} + \text{ACP} + \text{CO}_2$
v_{53}	$(R)\text{-3-hydroxybutyryl-ACP} + \text{NADP}^+ \longleftrightarrow \text{acetoacetyl-ACP} + \text{NADH,H}^+$
v_{54}	$(R)\text{-3-hydroxybutyryl-ACP} + \text{CoASH} \longleftrightarrow (R)\text{-3-hydroxybutanoyl-CoA} + \text{ACP}$
v_{55}	$\text{hexanoate} + \text{ACP} \longleftrightarrow \text{hexanoyl-ACP}$
v_{56}	$\text{hexanoyl-ACP} + \text{FAD}^+ \rightarrow \text{hex-2-enoyl-ACP} + \text{FADH}_2$
v_{57}	$\text{hex-2-enoyl-ACP} + \text{H}_2\text{O} \rightarrow (R)\text{-3-hydroxyhexanoyl-CoA}$
v_{58}	$(R)\text{-3-hydroxyhexanoyl-CoA} + \text{ACP} \rightarrow \text{CoASH} + (R)\text{-3-hydroxyhexanoyl-ACP}$
v_{59}	$\text{butyryl-CoA} + \text{ACP} \rightarrow \text{CoASH} + \text{butyryl-ACP}$
v_{60}	$\text{butyryl-ACP} + \text{malonyl-ACP} + \text{ACP} \rightarrow 3\text{-oxohexanoyl-ACP} + \text{CO}_2$
v_{61}	$3\text{-oxohexanoyl-ACP} + \text{NADPH,H}^+ \longleftrightarrow (R)\text{-3-hydroxyhexanoyl-ACP} + \text{NADP}^+$

PHA Formation and PHA Degradation

v_{62}	$\mathbf{n} \text{ (R)-3-hydroxybutyryl-CoA} \rightarrow \text{Poly-beta-hydroxybutyrate} + \mathbf{n} \text{ CoASH}$
v_{63}	$\mathbf{n} \text{ (R)-3-hydroxyvaleryl-CoA} \rightarrow \text{Poly-beta-hydroxyvalerate} + \mathbf{n} \text{ CoASH}$
v_{64}	$\text{Poly-beta-hydroxybutyrate} + \text{H}_2\text{O} \rightarrow \text{Poly}[(R)\text{-3-hydroxybutanoate}]\mathbf{x} + (R)\text{-3-}((R)\text{-3-Hydroxybutanoyloxy})\text{butanoate}$
v_{65}	$\text{Poly-beta-hydroxyvalerate} + \text{H}_2\text{O} \rightarrow \text{Poly}[(R)\text{-3-hydroxyvalenoate}]\mathbf{x} + (R)\text{-3-}((R)\text{-3-Hydroxyvaleryloxy})\text{valerate}$
v_{66}	$\mathbf{n} \text{ (R)-3-hydroxyhexanoyl-CoA} \rightarrow \text{Poly-beta-hydroxyhexanoate} + \mathbf{n} \text{ CoASH}$
v_{67}	$\text{Poly-beta-hydroxyhexanoate} + \text{H}_2\text{O} \rightarrow \text{Poly}[(R)\text{-3-hydroxyhexanoate}]\mathbf{x} + (R)\text{-3-}((R)\text{-3-Hydroxyhexanoyloxy})\text{hexanoate}$

Appendix L

Identifiability Analysis

Identifiability refers to the ability to determine unique and reliable parameter values from experimental data [276]. It is a prerequisite for extracting mechanistic insights from models and making reliable predictions under new conditions. In principle, two complementary perspectives can be considered [276–278]:

- **Structural Identifiability:** Whether the model parameters are uniquely recoverable from idealized, noise-free data. It depends only on the model structure and measurement configuration;
- **Practical Identifiability:** Whether the model parameters can be estimated with acceptable confidence given noisy, finite experimental datasets. It depends on both the model and the data quality.

A range of tools has been developed to study these issues, many of which have already been applied in systems biology:

1. **Structural identifiability tools** (a priori analysis);
 - *GenSSI* (Generating Series for Structural Identifiability, MATLAB toolbox);
 - *STRIKE-GOLDD* (Symbolic Lie Derivatives);
 - *DAISY* (Differential Algebra).
2. **Practical identifiability tools** (a posteriori analysis);
 - **Fisher Information Matrix** (FIM), which gives local sensitivity-based estimates of parameter uncertainty;
 - **Profile likelihood analysis**, which provides more reliable (and often asymmetric) confidence intervals for nonlinear models, and can reveal hidden non-identifiabilities missed by FIM;
 - **Bayesian sampling approaches** (Markov Chain Monte Carlo, variational inference), which can be used to quantify parameter uncertainty given prior information.
3. **Experimental design for identifiability.**

Optimal Experiment Design (OED) methods can improve identifiability by selecting informative experimental conditions (e.g., feed profiles, measurement time points, or metabolite readouts) that maximize sensitivity to the parameters of interest.

In this work, identifiability issues arise mainly when explicit kinetic parameters (e.g., Monod constants, maximum uptake rates) are introduced or when large sets of ANN weights need to be estimated. Although model reduction and pruning partly mitigate this, formal identifiability analysis was not carried out and is proposed as future work. Identifiability is also relevant for the gene expression models, which typically introduce many parameters for genes, promoters, and regulators but are constrained by only a few observables (transcripts, proteins, metabolites). This imbalance makes even structural identifiability difficult to establish, and in practice most parameters are non-identifiable without additional constraints or simplifications. In such cases, identifiability analysis is still valuable for determining which parameters or combinations are estimable and to guide reparameterization or experimental design. By contrast, in constraint-based metabolic models, explicit kinetic parameters are absent; the main challenge there is underdetermination, as many feasible flux distributions satisfy the same stoichiometric constraints.

In the context of this work, these tools could be integrated at two levels:

- **Reduced mechanistic models:** applying *GenSSI* or *STRIKE-GOLDD* to the ODE systems derived from EFM reduction would verify whether kinetic parameters are structurally identifiable;
- **Hybrid ANN-based models:** while structural identifiability is not directly applicable due to redundancy in neural weights, practical identifiability could be assessed via FIM or profile likelihoods on pruned networks, complemented by Bayesian methods to quantify uncertainty in predictions.

Overall, combining model reduction with identifiability analysis would help ensure that the hybrid models developed here remain both interpretable and predictive, and would guide the design of future experiments to maximize parameter information.

Bibliography

- [1] D.L. Nelson and M.M. Cox. Lehninger. principles of biochemistry. W.H. Freeman and Company, New York, 7th edition, 2008.
- [2] C. Baroukh, R. Munoz-Tamayo, J-P. Steyer, and O. Bernard. Drum : a new framework for metabolic modeling under non-balanced growth. application to the carbon metabolism of unicellular microalgae. *PLoS ONE*, 9 (8):e104499, 2014.
- [3] F.N. Santos-Navarro, Y. Boada, A. Vignoni, and J. Picó. Rbs and promoter strengths determine the cell-growth-dependent protein mass fractions and their optimal synthesis rates. *ACS Synthetic Biology*, 10.12:3290–3303, 2021.
- [4] W.C. DeLoache and J.E. Dueber. Compartmentalizing metabolic pathways in organelles. *Nature Biotechnology*, 31 (4):320–321, 2013.
- [5] M. Maton, A. Vande Wouwer, and Ph. Bogaerts. Selection of a minimal suboptimal set of efms for dynamic metabolic modelling. *IFAC-PapersOnLine*, 54-3:667–672, 2021.
- [6] G. Bastin and D. Dochain. On-line estimation and adaptive control of bioreactors. In *Process Measurement and Control*, volume 1. Elsevier, 1990.
- [7] A. Provost and G. Bastin. Dynamic metabolic modelling under the balanced growth condition. *Journal of Process Control*, 14 (7):717–728, 2004.
- [8] J. Haag, A. Vande Wouwer, and Ph. Bogaerts. Dynamic modeling of complex biological systems : a link between metabolic and macroscopic description. *Mathematical Biosciences*, 193(1):25–49, 2005.
- [9] X. Hulhoven, A. Vande Wouwer, and Ph. Bogaerts. On a systematic procedure for the predetermination of macroscopic reaction schemes. *Bioprocess and Biosystems Engineering*, 27 (5):283–291, 2005.
- [10] C.T. Trinh, A. Wlaschin, and F. Srienc. Elementary mode analysis : a useless metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology*, 81 (5):813–26, 2009.
- [11] S. Naderi, M. Meshram, C. Wei, Ingalls B. McConkey, B., H. Budman, and J. Scharer. Development of a mathematical model for evaluating the dynamics of normal and apoptotic chinese hamster ovary cells. *Biotechnology Progress*, 27:1197–1205, 2011.
- [12] H. Niu, Z. Amriht, P. Fickers, W. Tan, and Ph. Bogaerts. Metabolic pathway analysis and reduction for mammalian cell cultures - towards macroscopic modeling. *Chemical Engineering Science*, 102:461–473, 2013.

- [13] F. Zamorano, A. Vande Wouwer, R. Jungers, and G. Bastin. Dynamic metabolic models of cho cell cultures through minimal sets of elementary flux modes. *Journal of biotechnology*, 164(3):409–422, 2013.
- [14] L. Hebing, T. Neymann, T. Thüte, A. Jockwer, and S. Engell. Efficient generation of models of fed-batch fermentations for process design and control. *IFAC-PapersOnLine*, 49 (7):621–626, 2016.
- [15] C. Baroukh and O. Bernard. Metabolic modeling of c. sorokiniana diauxic heterotrophic growth. *IFAC-PapersOnLine*, 49 (26):330–335, 2016.
- [16] G.J. Tortora and S.R. Grabrowski. Principles of anatomy and physiology. Harper Collins College Publishers, New York, 1993.
- [17] A.J. Smith, J. London, and R.Y. Stanier. Biochemical basis of obligate autotrophy in blue-green algae and thiobacilli. *Journal of Bacteriology*, 94 (4):972–983, 1967.
- [18] I. Pe’er, C.E. Felder, O. Man, I. Silman, J.L. Sussman, and J.S. Beckmann. Proteomic signatures : amino acid and oligopeptide compositions differentiate among phyla. *Proteins : Structure, Function, and Bioinformatics*, 54 (1):20–40, 2004.
- [19] G. Stephanopoulos, A.A. Aristidou, and J. Nielsen. Metabolic engineering : principles and methodologies. Academic Press - Elsevier USA, San Diego, 1st edition, 1998.
- [20] R.T. Rockafellar. Convex analysis. Princeton University Press, 2015.
- [21] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2 (2):165–182, 1994.
- [22] J. Zanghellini, D.E. Ruckerbauer, M. Hanscho, and C. Jungreuthmayer. Elementary flux modes in a nutshell : properties, calculation and applications. *Biotechnology Journal*, 8 (9):1009–1016, 2013.
- [23] T.S. Motzkin, H. Raiffa, G.L. Thompson, and R.M. Thrall. The double description method. In *Annals of Mathematics Studies. Contributions to the Theory of Games*, volume 2, pages 51–73. Princeton University Press, Princeton, New Jersey, 1954.
- [24] K. Fukuda and A. Prodon. Double description method revisited. *Lecture Notes In Computer Science*, 1120:91–111, 1995.
- [25] Marco Terzer and Jörg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235, 2008.
- [26] T. Pfeiffer, I. Sanchez-Valdenebro, J-C. Nuno, F. Montero, and S. Schuster. Metatool : for studying metabolic networks. *Bioinformatics*, 15:251–257, 1999.
- [27] F. Llaneras and J. Pico. Stoichiometric modelling of cell metabolism. *Journal of Bioscience and Bioengineering*, 105 (1):1–11, 2008.
- [28] J.S. Ballez, J. Mols, J. Burteau, S.N. Agathos, and Y.J. Schneider. Plant protein hydrolysates support cho-320 cells proliferation and recombinant ifn- γ production in suspension and inside microcarriers in protein-free media. *Cytotechnology*, 44:103–114, 2004.
- [29] G. Bastin. Quantitative analysis of metabolic networks and design of minimal bioreaction models. *ARIMA : Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*, 9 (1):41–55, 2008.

- [30] R.D. Kornberg. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences*, 104 (32):12955–12961, 2007.
- [31] M.G. Sharma. Rna transcription. In D. Kar and S. Sarkar, editors, *Genetics Fundamentals Notes*, chapter 10, pages 491–536. Springer, Singapore, 2022.
- [32] T. Banerjee. Protein translation. In D. Kar and S. Sarkar, editors, *Genetics Fundamentals Notes*, chapter 11, pages 537–568. Springer, Singapore, 2022.
- [33] T. Banerjee. Regulation of gene expression in prokaryotes. In D. Kar and S. Sarkar, editors, *Genetics Fundamentals Notes*, chapter 12, pages 569–596. Springer, Singapore, 2022.
- [34] A.A. Rangarajan. Regulation of gene expression in eukaryotes. In D. Kar and S. Sarkar, editors, *Genetics Fundamentals Notes*, chapter 13, pages 597–634. Springer, Singapore, 2022.
- [35] Y. Boada, G. Reynoso-Meza, J. Pico, and A. Vignoni. Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices : an adaptative network case. *BMC Syst. Biol.*, 10 (27):doi:10.1186/s12918-016-0269-0, 2016.
- [36] F.N. Santos-Navarro, Y. Boada, A. Vignoni, and J. Picó. Gene expression space shapes the bioprocess trade-offs among titer, yield and productivity. *Applied Sciences*, 11 (3):5859, 2021.
- [37] H.M. Salis. The ribosome binding site calculator. *Methods in Enzymology*, 498:19–42, 2011.
- [38] H.M. Salis, E.A. Mirsky, and C.A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27 (10):946–950, 2009.
- [39] U. Alon. In *An Introduction to Systems Biology : Design Principles of Biological Circuits*. CRC Press, United Kingdom, 2019.
- [40] A.A.K. Nielsen and C.A. Voigt. Deep learning to predict the lab-of-origin of engineered dna. *Nature Communications*, 9 (1):3135, 2018.
- [41] D. Endy. Foundations for engineering biology. *Nature*, 438:449–453, 2005.
- [42] G. Baldwin, T. Bayer, R. Dickinson, T. Ellis, P.S. Freemont, R.I. Kitney, K. Polizzi, and G.B. Stan. In P.S. Freemont and R.I. Kitney, editors, *Synthetic Biology : A Primer*. Imperial College Press, United Kingdom, 2012.
- [43] M.W. Covert. In *Fundamentals of Systems Biology : From Synthetic Circuits to Whole-Cell Models*. CRC Press, United Kingdom, 2015.
- [44] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.
- [45] C.M. Bishop. In *Pattern Recognition and Machine Learning*. Springer, 2006.
- [46] I. Goodfellow, Y. Bengio, and A. Courville. In *Deep Learning*. MIT Press, 2016.
- [47] F. Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65 (6):286–408, 1958.
- [48] B. Mehlig. In *Machine Learning With Neural Networks : An Introduction for Scientists and Engineers*. Cambridge University Press, 2021.

- [49] D. Anderson and G. McNeill. In *Artificial Neural Networks Technology*. 1992.
- [50] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.
- [51] M.T. Hagan, H.B. Demuth, and M. Beale. In *Neural Network Design*. PWS Publishing Co., Boston, 1996.
- [52] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2 (4):303–314, 1989.
- [53] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4 (2):251–257, 1991.
- [54] S. Indolia, A.K. Goswami, S.P. Mishra, and P. Asopa. Conceptual understanding of convolutional neural network - a deep learning approach. *Procedia Computer Science*, 132:679–688, 2018.
- [55] T.P. Lillicrap and A. Santoro. Backpropagation through time and the brain. *Current Opinion in Neurobiology*, 55:82–89, 2019.
- [56] Ph. Bogaerts and R. Hanus. Macroscopic modelling of bioprocesses with a view to engineering applications. In M. Hofman and P. Thonart, editors, *Engineering and Manufacturing for Biotechnology. Focus on Biotechnology, vol 4*. Springer, Dordrecht, 2001.
- [57] B.B. Yahia, L. Malphettes, and E. Heinzle. Macroscopic modeling of mammalian cell growth and metabolism. *Applied Microbiology and Biotechnology*, 99 (17):7009–7024, 2015.
- [58] X. Hulhoven, A. Vande Wouwer, R. Hanus, and Ph. Bogaerts. Systematic generation of identifiable macroscopic reaction schemes - application to a perfusion cell culture. *2021 European Control Conference (ECC)*, pages 2517–2521, 2001.
- [59] A. Provost, G. Bastin, S.N. Agathos, and Y.J. Schneider. Metabolic design of macroscopic bioreaction models : application to chinese hamster ovary cells. *Bio-process and Biosystems Engineering*, 29 (5-6):349–366, 2006.
- [60] S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 17 (4):326–332, 1999.
- [61] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis ? *Trends in Biotechnology*, 21 (2):64–69, 2003.
- [62] J.L. Snoep and J.M. Rohwer. Detailed kinetic models using metabolomics data sets. In S. Vaidyanathan, G.G. Harrigan, and R. Goodacre, editors, *Metabolome Analyses : Strategies for Systems Biology*. Springer, Boston, MA, 2005.
- [63] R. Steuer, T. Gross, J. Selbig, and B. Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103 (32):11868–11873, 2006.
- [64] M.A. Savageau, E.O. Voit, and D.H. Irvine. Biochemical systems theory and metabolic control analysis. *Mathematical Biosciences*, 86:127–145, 1987.

- [65] J.R. Banga. Optimization in computational systems biology. *BMC Systems Biology*, 2:47, 2008.
- [66] H. Link, D. Christodoulou, and U. Sauer. Advancing metabolic models with kinetic information. *Current Opinion in Biotechnology*, 29:8–14, 2014.
- [67] L. Xie and D.I. Wang. High cell density and high monoclonal antibody production through medium design and rational control in a bioreactor. *Biotechnology and Bioengineering*, 51 (6):725–729, 1996.
- [68] H. Kaufmann and M. Fussenegger. Metabolic engineering of mammalian cells for higher protein yield. *New Comprehensive Biochemistry*, 38:457–469, 2003.
- [69] A. Richelle and N.E. Lewis. Improvements in protein production in mammalian cells from targeted metabolic engineering. *Current Opinion in Systems Biology*, 6:1–6, 2017.
- [70] J. Niklas and E. Heinzle. Metabolic flux analysis in systems biology of mammalian cells. *Genomics and Systems Biology of Mammalian Cell Culture*, Springer, pages 109–132, 2012.
- [71] F. Sidoli, A. Mantalaris, and S. Asprey. Modelling of mammalian cells and cell culture processes. *Cytotechnology*, 44 (1-2):27–46, 2004.
- [72] D. Vester, E. Rapp, S. Kluge, Y. Genzel, and U. Reichl. Virus-host cell interactions in vaccine production cell lines infected with different human influenza a virus variants : a proteomic approach. *Journal of proteomics*, 73 (9):1656–1669, 2010.
- [73] B. Hodgson, C. Taylor, J. Ushio, and J. Leigh. Intelligent modelling of bioprocesses : A comparison of structured and unstructured approaches. *Bioprocess and Biosystems Engineering*, 26(6):353–359, 2005.
- [74] Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002.
- [75] J. Gao, V.M. Gorenflo, J.M. Scharer, and H.M. Budman. Dynamic metabolic modeling for a mab bioprocess. *Biotechnology progress*, 23 (1):168–181, 2007.
- [76] J. Haag, A. Vande Wouwer, and Ph. Bogaerts. Systematic procedure for the reduction of complex biological reaction pathways and the generation of macroscopic equivalents. *Chemical Engineering Science*, 60 (2):459–465, 2005.
- [77] S. Naderi, M. Meshram, C. Wei, B. McConkey, B. Ingalls, H. Budman, and J. Scharer. Metabolic flux and nutrient uptake modeling of normal and apoptotic cho cells. *IFAC Proceedings Volumes*, 43 (6):395–400, 2010.
- [78] A. von Kamp and S. Schuster. Metatool 5.0 : fast and flexible elementary modes analysis. *Bioinformatics*, 22:1930–1931, 2006.
- [79] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports*, 29:233–236, 2002.
- [80] L.F. Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J.E. Beasley, S. Schuster, and F.J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25:3158–3165, 2009.

- [81] C. Kaleta, L.F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts ? pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Research*, pages 1872–1883, 2009.
- [82] R.M. Jungers, F. Zamorano, V.D. Blondel, A. Vande Wouwer, and G. Bastin. Fast computation of minimal elementary decompositions of metabolic vectors. *Automatica*, 47 (6):1255–1259, 2011.
- [83] Daniel Machado, Zita Soons, Kiran Raosaheb Patil, Eugénio C Ferreira, and Isabel Rocha. Random sampling of elementary flux modes in large-scale metabolic networks. *Bioinformatics*, 28(18):i515–i521, 2012.
- [84] S. Tabe-Bordbar and S.A. Marashi. Finding elementary flux modes in metabolic networks based on flux balance analysis and flux coupling analysis : application to the analysis of escherichia coli metabolism. *Biotechnology Letters*, 35:2039–2044, 2013.
- [85] H. Oddsdottir, E. Hagrot, V. Chotteau, and A. Forsgren. On dynamically generating relevant elementary flux modes in a metabolic network using optimization. *Journal of Mathematical Biology*, 71(4):903–920, 2015.
- [86] Z.I. Soons, E.C. Rocha, and I. Ferreira. Selection of elementary modes for bioprocess control. *Computer Applications in Biotechnology*, 11:156–161, 2010.
- [87] Z.I. Soons, I. Ferreira, and E.C. Rocha. Identification of minimal metabolic pathway models consistent with phenotypic data. *Journal of Process Control*, 21:1483–1492, 2011.
- [88] H.AE. Oddsdottir, E. Hagrot, V. Chotteau, and A. Forsgren. Robustness analysis of elementary flux modes generated by column generation. *Mathematical Biosciences*, 273:45–56, 2016.
- [89] E. Hagrot, H.AE. Oddsdottir, J.G. Hosta, E.W. Jacobsen, and V. Chotteau. Poly-pathway model, a novel approach to simulate multiple metabolic states by reaction network-based model - application to amino acid depletion in cho cell culture. *Journal of Biotechnology*, 259:235–247, 2017.
- [90] E. Hagrot, H. Oddsdottir, M. Makinen, A. Forsgren, and V. Chotteau. Novel column generation-based optimization approach for poly-pathway kinetic model applied to cho cell culture. *Metabolic Engineering Communications*, 8:e00083, 2019.
- [91] H.S. Song and D. Ramkrishna. Reduction of a set of elementary modes using yield analysis. *Biotechnology and Bioengineering*, 102 (2):554–568, 2009.
- [92] H.S. Song, D. Ramkrishna, G.E. Pinchuk, A.S. Beliaev, A.E. Konopka, and J.K. Fredrickson. Dynamical modeling of aerobic growth of shewanella oneidensis. *Metabolic Engineering*, 15:25–33, 2013.
- [93] T. Abbate, S. Fernandes de Sousa, L. Dewasme, G. Bastin, and A. Vande Wouwer. Inference of dynamical macroscopic models of cell metabolism based on elementary flux modes analysis. *Biochemical Engineering Journal*, 151:1–11, 2019.
- [94] F. Guil, J.F. Hidalgo, and J.M. Garcia. Boosting the extraction of elementary flux modes in genome-scale metabolic networks using the linear programming approach. *Bioinformatics*, 36 (14):4163–4170, 2020.

- [95] W. Mores, S.S. Bhonsale, F. Logist, and J.F.M. Van Impe. Accelerated enumeration of extreme rays through a positive-definite elementarity test. *Bioinformatics*, 41 (1), 2024.
- [96] J.G. Chitpin and T.J. Perkins. A markov constraint to uniquely identify elementary flux mode weights in unimolecular metabolic networks. *Journal of Theoretical Biology*, 575:111632, 2023.
- [97] M.A. Savageau. Biochemical systems analysis. i. some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 25:365–369, 1969.
- [98] M.A. Savageau. Biochemical systems analysis. ii. the steady-state solutions for n-pool system using a power-law approximation. *Journal of Theoretical Biology*, 25:370–379, 1969.
- [99] E.O. Voit and M.A. Savageau. Equivalence between s-systems and volterra-systems. *Mathematical Biosciences*, 78:47–55, 1982.
- [100] M.A. Savageau. Introduction to s-systems and the underlying power-law formalism. *Mathematical and Computer Modelling*, 11:546–551, 1979.
- [101] F. Shiraishi and M.A. Savageau. The tricarboxylic acid cycle in dictyostelium discoideum. formulation of the alternative kinetic representations. *Journal of Biological Chemistry*, 267:22912–22918, 1992.
- [102] R. Curto, A. Sorribas, and M. Cascante. Comparative characterization of the fermentation pathway of *saccharomyces cerevisiae* using biochemical systems theory and metabolic control analysis : model definition and nomenclature. *Mathematical Biosciences*, 130:25–50, 1995.
- [103] N.V. Torres, E.O. Voit, and C. Gonzalez-Alcon. Optimization of nonlinear biotechnological processes with linear programming : application to citric acid production by *aspergillus niger*. *Biotechnological Bioengineering*, 49:247–258, 1996.
- [104] B. Hernandez-Bermejo, V. Fairen, and A. Sorribas. Power-law modeling based on least-squares minimization criteria. *Mathematical Biosciences*, 161:83–94, 1999.
- [105] J. Haag, A. Vande Wouwer, and M. Remy. A general model of reaction kinetics in biological systems. *Bioprocess and Biosystems Engineering*, 27(5):303–309, 2005.
- [106] J. Mailier and A. Vande Wouwer. Identification of nested biological kinetic models using likelihood ratio tests. *Chemical Engineering*, 84:727–734, 2012.
- [107] A. Grosfils, A. Vande Wouwer, and Ph. Bogaerts. On a general model structure for macroscopic biological reaction rates. *Journal of Biotechnology*, 130:253–264, 2007.
- [108] A. Richelle and Ph. Bogaerts. Systematic methodology for bioprocess model identification based on generalized kinetic functions. *Biochemical Engineering Journal*, 100:41–49, 2015.
- [109] G. Montague and J. Morris. Neural-network contributions in biotechnology. *Trends in Biotechnology*, 12:312–324, 1994.
- [110] L. Chen, O. Bernard, G. Bastin, and P. Angelov. Hybrid modeling of biotechnological processes using neural networks. *Control Engineering Practice*, 8 (7):821–827, 2000.

- [111] A. Vande Wouwer, C.H. Renotte, and Ph. Bogaerts. Biological reaction modeling using radial basis function networks. *Computational Chemical Engineering*, 28 (11):2157–2164, 2004.
- [112] M. Maton, A. Vande Wouwer, and Ph. Bogaerts. A systematic elementary flux mode selection procedure for deriving macroscopic bioreaction models from metabolic networks. *Journal of Process Control*, 118:170–184, 2022.
- [113] D.C. Psichogios and L.H. Ungar. A hybrid neural network-first principles approach to process modeling. *American Institute of Chemical Engineers (AIChE) Journal*, 38 (10):1499–1511, 1992.
- [114] A.M. Schweidtmann, D. Zhang, and M. von Stosch. A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering*, 10:100136, 2024.
- [115] P. Shah, S. Pahari, R. Bhavsar, and J.S. Kwon. Hybrid modeling of first-principles and machine learning: a step-by-step tutorial review for practical implementation. *Computers and Chemical Engineering*, 194:108926, 2025.
- [116] C.H. Schilling and B.O. Palsson. The underlying pathway structure of biochemical reaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (8):4193–4198, 1998.
- [117] J. Orth, I. Thiele, and B.O. Palsson. What is flux balance analysis. *Nature Biotechnology*, 28:245–248, 2010.
- [118] G. Bastin. Metabolic flux analysis and metabolic design of bioreaction systems. *Tutorial*, 2019.
- [119] J. Villadsen, J. Nielsen, and G. Lidén. In *Bioreaction Engineering Principles*. Springer, 2003.
- [120] L. Michaelis and M.L. Menten. Die kinetik der invertinwirkung. *Biochemische Zeitschrift*, 49:333–369, 1913.
- [121] J. Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3 (1):371–394, 1949.
- [122] J.B.S. Haldane. In *Enzymes*. Longmans, Green Co, 1930.
- [123] G. Araujo Pimentel, L. Dewasme, and A. Vande Wouwer. On the number of reactions and stoichiometry of bioprocess macroscopic models: an implicit sparse identification approach. *IFAC-PapersOnLine*, 56 (2):9721–9726, 2022.
- [124] G. Araujo Pimentel, F.N. Santos-Navarro, L. Dewasme, and A. Vande Wouwer. Elucidation of macroscopic stoichiometry and kinetics of bioprocesses using sparse identification. *IFAC-PapersOnLine*, 58 (14):422–427, 2024.
- [125] G. Araujo Pimentel, L. Dewasme, and A. Vande Wouwer. Data-driven linear predictor based on maximum likelihood nonnegative matrix decomposition for batch cultures of hybridoma cells. *IFAC-PapersOnLine*, 55 (7):903–908, 2022.
- [126] G. Araujo Pimentel, L. Dewasme, and A. Vande Wouwer. Data-driven inference of bioprocess models; a low-rank matrix approximation approach. *Journal of Process Control*, 134:103148, 2024.
- [127] A. Meyer-Baese and V. Schmid. Foundations of neural networks. In *Pattern Recognition and Signal Analysis in Medical Imaging*, pages 197–243. Academic Press, 2nd edition, 2014.

- [128] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks for Perception : Computation, Learning, and Architectures*, pages 65–93. Academic Press, 1992.
- [129] S.A. Janowsky. Pruning versus clipping in neural networks. *Physical Review A*, 39:6600, 1989.
- [130] R. Reed. Pruning algorithms : a survey. *IEEE Transactions on Neural Networks*, 4 (5):740–747, 1993.
- [131] S. Hang, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28:1135–1143, 2015.
- [132] J. Frankle and M. Carbin. The lottery ticket hypothesis : finding sparse, trainable neural networks. *International Conference on Learning Representations*, 2019.
- [133] H. Tanaka, D. Kunin, D.L. Yamins, and S. Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33:1–13, 2015.
- [134] M.R. Silvestre and L.L. Ling. Pruning methods to mlp neural networks considering proportional apparent error rate for classification problems with unbalanced data. *Measurement*, 56:88–94, 2014.
- [135] J.R. Banga and E. Balsa-Canto. Parameter estimation and optimal experimental design. *Essays in Biochemistry*, 45:195–209, 2008.
- [136] S. Fernandes de Sousa, G. Bastin, M. Jolicoeur, and A. Vande Wouwer. Dynamic metabolic flux analysis using a convex analysis approach : application to hybridoma cell cultures in perfusion. *Biotechnology and Bioengineering*, 113(5):1102–1121, 2015.
- [137] W.S. Ahn and M.R. Antoniewicz. Parallel labeling experiments with [1,2-(13)c]glucose and [u-(13)c]glutamine provide new insights into cho cell metabolism. *Metabolic Engineering*, 15:34–47, 2013.
- [138] J. Niklas, K. Schneider, and E. Heinzle. Metabolic flux analysis in eukaryotes. *Current Opinion in Biotechnology*, 21 (1):63–69, 2010.
- [139] M. Maton, A. Vande Wouwer, and Ph. Bogaerts. Hybrid dynamic models of bioprocesses based on elementary flux modes and multilayer perceptrons. *Processes*, 10:2084, 2022.
- [140] Z. Amribt, N. Hongxing, and Ph. Bogaerts. Macroscopic modelling of overflow metabolism and model based optimization of hybridoma cell fed-batch cultures. *Biochemical Engineering Journal*, 70:196–209, 2013.
- [141] D. Dochain and P. Vanrolleghem. Identification of bioprocess models. *Bioprocess Control, ISTE Ltd. and John Wiley Sons Inc.*, pages 47–76, 2008.
- [142] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks : a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

- [143] Z. Lui, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang. Learning efficient convolutional networks through network slimming. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2755–2763, 2017.
- [144] M.W. Covert, C.H. Schilling, I. Famili, Goryanin I.I. Edwards, J.S., E. Selkov, and B.O. Palsson. Metabolic modeling of microbial strains in silico. *Trends in Biochemical Sciences*, 26:179–186, 2001.
- [145] N.D. Price, J.A. Papin, C.H. Schilling, and B.O. Palsson. Genome-scale microbial in silico models : the constraints-based approach. *Trends in Biotechnology*, 21:162–169, 2003.
- [146] J.L. Reed and B.O. Palsson. Thirteen years of building constraint-based in silico models of escherichia coli. *Journal of Bacteriology*, 185:2692–2699, 2003.
- [147] B.N. Kholodenko, S. Schuster, J.M. Rohwer, M. Cascante, and H.V. Westerhoff. Composite control of cell function : Metabolic pathways behaving as single control units. *FEBS Letters*, 368:1–4, 1995.
- [148] J.M. Rohwer, S. Schuster, and H.V. Westerhoff. How to recognize monofunctional units in a metabolic system. *Journal of Theoretical Biology*, 179:213–228, 1996.
- [149] S. Klamt, J. Stelling, M. Ginkel, and E.D. Gilles. Fluxanalyzer : Exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, 19:261–269, 2003.
- [150] I. Thiele and B.O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5 (1):93–121, 2010.
- [151] A. Montagud, E. Navarro, P. Fernandez de Cordoba, J.F. Urchueguia, and K.R. Patil. Reconstruction and analysis of genome-scale metabolic model of a photo-synthetic bacterium. *BMC Systems Biology*, 4:156:1–16, 2010.
- [152] M. Wu and C. Chan. Human metabolic network : reconstruction, simulation, and applications in systems biology. *Metabolites*, 2:242–253, 2012.
- [153] D. Machado, S. Andrejev, M. Tramontano, and K.R. Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46 (15):7542–7553, 2018.
- [154] 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*, 203:229–248, 2000.
- [155] S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18:326–332, 2000.
- [156] A.P. Burgard, E.V. Nikolaev, C.H. Schilling, and C.D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14 (2):301–312, 2004.
- [157] A. Larhlimi, L. David, J. Selbig, and A. Bockmayr. F2c2 : a fast tool for the computation of flux coupling in genome-scale metabolic networks. *BMC Bioinformatics*, 13:57, 2012.

- [158] P. Erdrich, R. Steur, and S. Klamt. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *Cellular and Molecular Life Sciences*, 9:48:1–12, 2015.
- [159] F. Zamorano, A. Vande Wouwer, and G. Bastin. A detailed metabolic flux analysis of an underdetermined network of cho cells. *Journal of Biotechnology*, 150:497–508, 2010.
- [160] S. Fernandes de Sousa, J. Robitaille, G. Bastin, M. Jolicoeur, and A. Vande Wouwer. Application of dynamic metabolic flux convex analysis to cho-dxb11 cell fed-batch cultures. *IFAC-PapersOnLine*, 49 (7):466–471, 2016.
- [161] S. Fernandes de Sousa, G. Bastin, M. Jolicoeur, and A. Vande Wouwer. Dynamic metabolic flux analysis using a convex analysis approach : Application to hybridoma cell cultures in perfusion. *Biotechnology and Bioengineering*, 113 (5):1102–1112, 2016.
- [162] T. Cakir and M.J. Khatibipour. Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. *Frontiers in Bioengineering and Biotechnology*, 2:00062, 2014.
- [163] O. Hadicke and S. Klamt. Ecolicore2 : a reference network model of the central metabolism of escherichia coli and relationships to its genome-scale parent model. *Nature Scientific Reports*, 7:39647, 2017.
- [164] D. Singh and M.J. Lercher. Network reduction methods for genome-scale metabolic models. *Cellular and Molecular Life Sciences*, 77:481–488, 2020.
- [165] G. Sambamoorthy and K. Raman. Minreact : a systematic approach for identifying minimal metabolic networks. *Bioinformatics*, 36 (15):4309–4315, 2020.
- [166] N. Cocco, M. Llabrés, M. Reyes-Prieto, and M. Simeoni. Metnet : a two-level approach to reconstructing and comparing metabolic networks. *PLoS ONE*, 16 (2):e0246962, 2021.
- [167] I. Koo, X. Wei, X. Shi, Z. Zhou, S. Kim, and X. Zhang. Constructing metabolic association networks using high-dimensional mass spectroscopy data. *Chemometrics and Intelligent Laboratory Systems*, 138:193–202, 2014.
- [168] D.V. Petrovsky, K.A. Malsagova, V.R. Rudnev, L.I. Kulikova, V.I. Pustovoyt, E.I. Balakin, K.A. Yurku, and A.L. Kaysheva. Bioinformatics methods for constructing metabolic networks. *Processes*, 11:3430, 2023.
- [169] G. Jansen, T. Qi, V. Latora, G.D. Amoutzias, D. Delneri, S.G. Oliver, and G. Nicosia. Minimisation of metabolic networks defines a new functional class of genes. *Nature Communications*, 15:9076, 2024.
- [170] E.C. Martinez. Lumping of components and reactions in complex reaction networks. *Chemical Engineering Communications*, 93 (1):1–24, 1990.
- [171] P. Pepiot, L. Cai, and H. Pitsch. Model reduction and lumping procedures. *Computer Aided Chemical Engineering*, 45:799–827, 2019.
- [172] R. Lui, J. Yang, Y. Zhao, and L. Cai. An automated chemical lumping method of functional isomers based on the reaction class concept for the reduction of large chemical kinetic mechanisms. *Fuel*, 371:132033, 2024.

- [173] K.J. Kauffman, P. Prakash, and J.S. Edwards. Advances in flux balances analysis. *Current Opinion in Biotechnology*, 14 (5):491–496, 2003.
- [174] K. Raman and N. Chandra. Flux balance analysis of biological systems : applications and challenges. *Briefings in Bioinformatics*, 10 (4):435–449, 2009.
- [175] L. David, S.A. Marashi, A. Larhlimi, B. Mieth, and A. Bockmayr. Ffca : a feasibility-based method for flux coupling analysis of metabolic networks. *BMC Bioinformatics*, 12:236, 2011.
- [176] M. Yasemi and M. Jolicoeur. Modelling cell metabolism : a review on constraint-based steady-state and kinetic approaches. *Processes*, 9 (2):322, 2021.
- [177] C.S. Henry, L.J. Broadbelt, and V. Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical Journal*, 92:1792–1805, 2007.
- [178] J. Schellenberger, N.E. Lewis, and B.O. Palsson. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, 100:544–553, 2011.
- [179] S. Peres and V. Fromion. *Thermodynamic approaches in flux analysis*, pages 359–367. Springer US, New York, NY, 2020.
- [180] M.G. Gollub, H-M. Kaltenbach, and J. Stelling. Probabilistic thermodynamic analysis of metabolic networks. *Bioinformatics*, 37 (18):2938–2945, 2021.
- [181] C. Tomi-Andrino, R. Norman, T. Millat, P. Soucaille, K. Winzer, D.A. Barrett, J. King, and D-H. Kim. Physicochemical and metabolic constraints for thermodynamics-based stoichiometric modelling under mesophilic growth conditions. *PLoS Computational Biology*, 17 (1):e1007694, 2021.
- [182] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitiz, I. Schomburg, M. Neumann-Schaal, and D. Schomburg. Brenda : the elixir core data resource in 2021 : new developments and updates. *Nucleic Acids Research*, 49:D498–D508:DOI: 10.1093/nar/gkaa1025 PubMed: 33211880, 2021.
- [183] M. Kanehisa and S. Goto. Kegg : kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28 (1):27–30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409, 2000.
- [184] E. Noor, A. Flamholz, A. Bar-Even, D. Davidi, R. Milo, and W. Liebermeister. The protein cost of metabolic fluxes : prediction from enzymatic rate laws and cost minimization. *PLoS Computational Biology*, 12 (11):e1005167, 2016.
- [185] M.R. Antoniewicz. Methods and advances in metabolic flux analysis : a mini-review. *Journal of Industrial Microbiology and Biotechnology*, 42 (3):317–325, 2015.
- [186] D. Molenaar, R. van Berlo, D. de Ridder, and B. Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular Systems Biology*, 5:323, 2009.
- [187] E.J. O’Brien, J.A. Lerman, R.L. Chang, D.R. Hyduke, and B.O. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, 9:693, 2013.
- [188] B.J. Sanchez, C. Zhang, A. Nilsson, P.J. Lahtee, E.J. Kerkhoven, and J. Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology*, 13 (8):935, 2017.

- [189] N.E. Lewis, K.K. Hixson, T.M. Conrad, J.A. Lerman, P. Charusanti, A.D. Polpitiya, J.N. Adkins, G. Schramm, S.O. Purvine, D. Lopez-Ferrer, K.K. Weitz, R. Eils, R. Konig, R.D. Smith, and B.O. Palsson. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6:390:doi:10.1038/msb.2010.47, 2010.
- [190] R. Mahadevan and C. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5 (4):264–276, 2003.
- [191] C. Baroukh, V. Turon, and O. Bernard. Dynamic metabolic modeling of heterotrophic and mixotrophic microalgal growth on fermentative wastes. *PLoS Computational Biology*, 13 (6):e1005590, 2017.
- [192] B. Assis Pessi, C. Baroukh, A. Bacquet, and O. Bernard. A universal dynamical metabolic model representing mixotrophic growth of chlorella sp. on wastes. *Water Research*, 229:119388, 2023.
- [193] J.E. Pinero Estrada, P. Bermejo Bescos, and A.M. Villar del Fresno. Antioxidant activity of different fractions of spirulina platensis protean extract. *Farmaco*, 56:497–500, 2001.
- [194] A. Kulshreshtha, A.J. Zacharia, U. Jarouliya, P. Bhadauriya, G.B. Prasad, and P.S. Bisen. Spirulina in health care management. *Current Pharmaceutical Biotechnology*, 9:400–405, 2008.
- [195] M.L. Ghirardi, L. Zhang, J.W. Lee, T. Flynn, M. Seibert, E. Greenbaum, and A. Melis. Microalgae : a green source of renewable h₂. *Trends in Biotechnology*, 18:506–511, 2000.
- [196] E. Flores, J.E. Frias, L.M. Rubio, and A. Herrero. Photosynthetic nitrate assimilation in cyanobacteria. *Photosynthesis Research*, 83:117–133, 2005.
- [197] L. Hendrickx, H. De Wever, V. Hermans, F. Mastroleo, N. Morin, A. Wilmotte, P. Janssen, and M. Mergeay. Microbial ecology of the closed artificial ecosystem melissa : reinventing and compartmentalizing the earth’s food and oxygen regeneration system for long-haul space exploration missions. *Research in Microbiology*, 157:77–86, 2006.
- [198] F. Deschoenmaeker, R. Facchini, B. Leroy, H. Badri, C.C. Zhang, and R. Wattiez. Proteomic and cellular views of arthrospira sp. pcc 8005 adaptation to nitrogen depletion. *Microbiology*, 160:1224–1236, 2014.
- [199] F. Deschoenmaeker, G. Bayon-Vicente, N. Sachdeva, O. Depaetere, J.C. Cabrera, B. Leroy, K. Muylaert, and R. Wattiez. Impact of different nitrogen sources on the growth of arthrospira sp. pcc 8005 under batch and continuous cultivation - a biochemical, transcriptomic and proteomic profile. *Bioresource Technology*, 237:78–88, 2017.
- [200] A. Meechai, S. Pongakarakun, P. Deshniun, S. Cheevadhanarak, and S. Bhumi-ratana. Metabolic flux distribution for gamma-linolenic acid synthetic pathways in spirulina platensis. *Biotechnology and Bioprocess Engineering*, 9:506–513, 2004.
- [201] G. Cogne, J.B. Gros, and C.G. Dussap. Identification of a metabolic network structure representative of arthrospira (spirulina) platensis metabolism. *Biotechnology and Bioengineering*, 84:667–676, 2003.

- [202] A. Klanchui, C. Khannapho, A. Phodee, S. Cheevadhanarak, and A. Meechai. iak692 : a genome-scale metabolic model of spirulina platensis c1. *BMC Systems Biology*, 6 (71):<https://doi.org/10.1186/1752-0509-6-71>, 2012.
- [203] K. Yoshikawa, S. Aikawa, Y. Kojima, Y. Toya, C. Furusawa, A. Kondo, and H. Shimizu. Construction of a genome-scale metabolic model of arthrospira platensis nies-39 and metabolic design for cyanobacterial bioproduction. *PLoS ONE*, 10 (12):e0144430. <https://doi.org/10.1371/journal.pone.0144430>, 2015.
- [204] A. Klanchui, S. Dulsawat, K. Chaloeamngam, S. Cheevadhanarak, P. Prommeenate, and A. Meechai. An improved genome-scale metabolic model of arthrospira platensis c1 (iak888) and its application in glycogen overproduction. *Metabolites*, 8 (4):doi: 10.3390/metabo8040084. PMID: 30486288; PMCID: PMC6315860., 2018.
- [205] K.D. Rawls, B.V. Dougherty, E.M. Blais, E. Stancliffe, G.L. Kolling, K. Vinnakota, V.R. Pannala, A. Wallqvist, and J.A. Papin. A simplified metabolic network reconstruction to promote understanding and development of flux balance analysis tools. *Computers in Biology and Medicine*, 105:64–71, 2019.
- [206] M.I. Mendzhul, T.G. Lysenko, O.A. Shainskaia, and I.V. Busakhina. Activity of tricarboxylic acid cycle enzymes in cyanobacteria spirulina platensis. *Mikrobiologicheskij Zhurnal*, 62 (1):3–10, 2000.
- [207] A. Funahashi, N. Tanimura, M. Morohashi, and H. Kitano. Celldesigner : a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1:159–162, 2003.
- [208] C. Baroukh, R. Muñoz-Tamayo, J.P. Steyer, and O. Bernard. Mathematical modeling of unicellular microalgae and cyanobacteria metabolism for biofuel production. *Current Opinion in Biotechnology*, 33:198–205, 2015.
- [209] S. Klamt, S. Schuster, and E.D. Gilles. Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnology and bioengineering*, 77 (7):734–751, 2002.
- [210] F. Chen and M.R. John. Substrate inhibition of chlamydomonas reinhardtii by acetate in heterotrophic culture. *Process Biochemistry*, 29 (4):245–252, 1994.
- [211] M.K. Kim, A. Lane, J.J. Kelley, and D.S. Lun. E-flux2 and spot : validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. *PLoS ONE*, 11 (6):e0157101, 2016.
- [212] M.A. Borowitzka. High-value products from microalgae - their development and commercialisation. *Journal of Applied Phycology*, 25 (3):743–756, 2013.
- [213] R.H. Wijffels, O. Kruse, and K.J. Hellingwerf. Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Current Opinion in Biotechnology*, 24 (3):405–413, 2013.
- [214] A.M.J. Kliphuis, A.J. Klok, D.E. Martens, P.P. Lamers, M. Janssen, and R.H. Wijffels. Metabolic modeling of chlamydomonas reinhardtii : energy requirements for photoautotrophic growth and maintenance. *Journal of Applied Phycology*, 24:253–266, 2011.
- [215] T. Lacour, A. Sciandra, A. Talec, P. Mayzaud, and O. Bernard. Diel variations of carbohydrates and neutral lipids in nitrogen-sufficient and nitrogen-starved cyclostos cultures of isochrysis sp.(1). *Journal of Phycology*, 48 (4):966–975, 2012.

- [216] R.L. Chang, L. Ghamsari, A. Manichaikul, E.F.Y. Hom, S. Balaji, W. Fu, Y. Shen, T. Hao, B.O. Palsson, K. Salehi-Ashtiani, and J.A. Papin. Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Molecular Systems Biology*, 7:518, 2011.
- [217] S.L. Campbell and K.E. Wellen. Metabolic signaling to the nucleus in cancer. *Molecular Cell*, 71 (3):398–408, 2018.
- [218] M.J. Iglesias, R. Soengas, I. Probert, E. Guilloud, P. Gourvil, M. Mehiri, Y. Lopez, V. Cepas, I. Gutiérrez-Del-Rio, S. Redondo-Blanco, C.J. Villar, F. Lombo, S. Soto, and F.L. Ortiz. Nmr characterization and evaluation of antibacterial and antibiofilm activity of organic extracts from stationary phase batch cultures of five marine macroalgae (*Dunaliella* sp., *D. salina*, *Chaetoceros calcitrans*, *C. gracilis* and *Tisochrysis lutea*). *Phytochemistry*, 164:192–205, 2019.
- [219] M. Senorans, N. Castejon, and F.J. Senorans. Advanced extraction of lipids with dha from *Isochrysis galbana* with enzymatic pre-treatment combined with pressurized liquids and ultrasound assisted extractions. *Molecules*, 25 (14):3310, 2020.
- [220] J. Chen, J. Yang, H. Du, M. Aslam, W. Wang, W. Chen, T. Li, Z. Liu, and X. Liu. Laminarin, a major polysaccharide in stramenopiles. *Marine Drugs*, 19 (10):576, 2021.
- [221] M. Garnier, G. Bougaran, M. Pavlovic, J-B. Berard, G. Carrier, A. Charrier, F. Le Grand, E. Lukomska, C. Rouxel, N. Schreiber, J-P. Cadoret, H. Rogniaux, and B. Saint-Jean. Use of a lipid rich strain reveals mechanisms of nitrogen limitation and carbon partitioning in the haptophyte *Tisochrysis lutea*. *FAgal Research*, 20:229–248, 2016.
- [222] C. Baroukh, F. Mairet, and O. Bernard. The paradoxes hidden behind the droop model highlighted by a metabolic approach. *Frontiers in Plant Science*, 13:941230, 2022.
- [223] S.N. Mendoza, B.G. Olivier, D. Molenaar, and B. Teusink. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, 20 (1):158, 2019.
- [224] N.T. Eriksen. Production of phycocyanin - a pigment with applications in biology, biotechnology, foods and medicine. *Applied Microbiology and Biotechnology*, 80:1–14, 2008.
- [225] Q. Hu, M. Sommerfeld, E. Jarvis, M. Ghirardi, M. Posewitz, M. Seibert, and A. Darzins. Microalgal triacylglycerols as feedstocks for biofuel production : perspectives and advances. *The Plant Journal*, 54 (4):621–639, 2008.
- [226] M. Maton and A. Vande Wouwer. A mathematical modeling of photosynthetic eukaryotic microorganisms using metabolic networks. *IFAC-PapersOnLine*, 59 (1):397–402, 2025.
- [227] F. Ceroni, R. Algar, G.B. Stan, and T. Ellis. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature Methods*, 12 (5):415–418, 2015.
- [228] A. Rodriguez, K.R. Kildegaard, M. Li, I. Borodina, and J. Nielsen. Establishment of a yeast platform strain for production of p-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis. *Metabolic Engineering*, 31:181–188, 2015.

- [229] J.D. Varner. Large-scale prediction of phenotype : concept. *Biotechnology Bio-engineering*, 69:664–678, 2000.
- [230] J. Nielsen and L. Olsson. An expanded role for microbial physiology in metabolic engineering and functional genomics : moving forward systems biology. *FEMS Yeast Research*, 2:175–181, 2002.
- [231] S. Gudmundsson and I. Thiele. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11:489, 2010.
- [232] J. Schellenberger, R. Que, R.M.T Fleming, I. Thiele, J.D. Orth, and A.M. Feist. Quantitative prediction of cellular metabolism with constraint-based models : the cobra toolbox v2.0. *Nature Protocols*, 6 (9):1290–1307, 2011.
- [233] M. Wehrs, D. Tanjore, T. Eng, J. Lievense, T.R. Pray, and A. Mukhopadhyay. Engineering robust production microbes for large-scale cultivation. *Trends in Microbiology*, 27:524–537, 2019.
- [234] Y. Boada, F.N. Santos-Navarro, J. Pico, and A. Vignoni. Modeling and optimization of a molecular biocontroller for the regulation of complex metabolic pathways. *Frontiers in Molecular Biosciences*, 9:801032, 2022.
- [235] C. Liao, A.E. Blanchard, and T. Lu. An integrative circuit-host modelling framework for predicting synthetic gene network behaviours. *Nature Microbiology*, 2:1658–1666, 2017.
- [236] 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, 2014.
- [237] M.W. Covert, C.H. Schilling, and B.O. Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213:73–88, 2001.
- [238] M.W. Covert and B.O. Palsson. Transcriptional regulation in constraints-based metabolic models of e. coli. *Journal of Biological Chemistry*, 277:28058–28064, 2002.
- [239] M. Akesson, J. Forster, and J. Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, 6:285–293, 2004.
- [240] M.W. Covert, N. Xiao, T.J. Chen, and J.R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. *Bioinformatics*, 18:2044–2045, 2008.
- [241] S. Chandrasekaran and N.D. Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *PNAS*, 107 (41):17845–17850, 2010.
- [242] L. Marmiesse, R. Peyraud, and L. Cottret. Flexflux : combining metabolic flux and regulatory network analyses. *BMC Systems Biology*, 9:93, 2015.
- [243] D. Machado, M.J. Herrgard, and I. Rocha. Stoichiometric representation of gene-protein-reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS Computational Biology*, 12 (10):e1005140. doi:10.1371/journal.pcbi.1005140, 2016.

- [244] S. Waldherr, D.A. Oyarzun, and A. Bockmayr. Dynamic optimization of metabolic networks coupled with gene expression. *Journal of Theoretical Biology*, 365:469–485, 2015.
- [245] A.M. Reimers, H. Knoop, A. Bockmayr, and R. Steuer. Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. *Proceedings of the National Academy of Sciences*, 114 (31):E6457–E6465, 2017.
- [246] H. Lindhorst, S. Lucia, R. Findeisen, and S. Waldherr. Modeling enzyme controlled metabolic networks in rapidly changing environments by robust optimization. *IEEE Control Systems Letters*, 99:1-1:doi:10.1109/LCSYS.2018.2866234, 2018.
- [247] H. Lindhorst, A.M. Reimers, and S. Waldherr. Dynamic modeling of enzyme controlled metabolic networks using a receding time horizon. *Preprints on 10th IFAC International Symposium on Advanced Control of Chemical Processes*, 2018.
- [248] X. Li, A.J.M. Walhout, and L.S. Yilmaz. Enhanced flux potential analysis links changes in enzyme expression to metabolic flux. *Molecular Systems Biology*, 21 (4):413–445, 2025.
- [249] H. Yuan, Y. Bai, X. Li, and X. Fu. Cross-regulation between proteome reallocation and metabolic flux redistribution governs bacterial growth transition kinetics. *Metabolic Engineering*, 82:60–68, 2024.
- [250] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. In *Molecular Biology of the Cell*, 4th edition, volume 2. Garland Science, New York, 2002.
- [251] S. Oesterle, D. Gerngross, S. Schmitt, T.M. Roberts, and S. Panke. Efficient engineering of chromosomal ribosome binding site libraries in mismatch repair proficient escherichia coli. *Scientific Reports*, 7:12327, 2017.
- [252] D. Hidalgo, C.A. Martinez-Ortiz, B.O. Palsson, J.I. Jimenez, and J. Utrilla. Regulatory perturbations of ribosome allocation in bacteria reshape the growth proteome with a trade-off in adaptation capacity. *iScience*, 25 (3):103879, 2022.
- [253] E.M. Nikolados, A.Y. Weibe, and D.A. Oyarzun. Prediction of cellular burden with host-circuit models. *Methods in Molecular Biology*, 2229:267–291, 2021.
- [254] Y. Boada, A. Vignoni, J. Pico, and P. Carbonell. Extended metabolic biosensor design for dynamic pathway regulation of cell factories. *iScience*, 23 (7):101305, 2020.
- [255] 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.
- [256] J. Wu, T. Zhou, G. Du, J. Zhou, and J. Chen. Modular optimization of heterologous pathways for de novo synthesis of (2s)-naringenin in escherichia coli. *PLOS ONE*, 9 (7):e101492, 2014.
- [257] H. Bremer and P.P. Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 2008.
- [258] J. Hausser, A. Mayo, L. Keren, and U. Alon. Central dogma rates and the trade-off between precision and economy in gene expression. *Nature Communications*, 10:68, 2019.

- [259] J.D. Orth, R.M. Fleming, and B.O. Palsson. Reconstruction and use of microbial metabolic networks: the core e.coli metabolic model as an educational guide. *EcoSal Plus*, 4 (1):10.1128, 2010.
- [260] F.N. Santos-Navarro, J.L. Navarro, Y. Boada, A. Vignoni, and J. Picó. Onemodel : an open-source sbml modeling tool focused on accessibility, simplicity and modularity. *IFAC-PapersOnLine*, 55 (7):125–130, 2022.
- [261] T. Makino, G. Skretas, and G. Georgiou. Strain engineering for improved expression for recombinant proteins in bacteria. *Microbial Cell Factories*, 10:32, 2011.
- [262] G.L. Rosano and E.A. Ceccarelli. Recombinant protein expression in escherichia coli : advances and challenges. *Frontiers in Microbiology*, 5:172, 2014.
- [263] R. Di Blasi, M. Pisani, F. Tedeschi, M.M. Marbiah, K. Polizzi, S. Furini, V. Siciliano, and F. Ceroni. Resource-aware construct design in mammalian cells. *Nature Communications*, 14:3576, 2023.
- [264] C.M. Gowen and S.S. Fong. Genome-scale metabolic model integrated with rnaseq data to identify metabolic states of clostridium thermocellum. *Biotechnology Journal*, 5 (7):759–767, 2010.
- [265] R. Hasibi, T. Michoel, and D.A. Oyarzun. Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. *npj Systems Biology and Applications*, 10:24, 2024.
- [266] A. Goelzer and V. Fromion. Bacterial growth rate reflects a bottleneck in resource allocation. *Biochimica et Biophysica Acta*, 1810 (10):978–988, 2011.
- [267] N. Sachdeva, G. Giambarresi, L. Poughon, J.C. Cabrera, B. Leroy, C. Lasseur, C-G. Dussap, and R. Wattiez. Assessment of transient effects of alternative nitrogen sources in continuous cultures of arthrosira sp. using proteomic, modeling and biochemical tools. *Bioresource Technology*, 267:492–501, 2018.
- [268] P.G. Roessler. Udp-glucose pyrophosphotylase activity in the diatom cyclotella cryptica. pathway of chrysolaminarin biosynthesis. *Journal of Phycology*, 23 (3):494–498, 1987.
- [269] E. Kim, J.G. Kang, E.H. Jho, W.H. Yang, and J.W. Cho. O-glcnacylation : an emerging protein modification regulating the hippo pathway. *Cancers (Basel)*, 14 (12):3013, 2022.
- [270] Y. Le Strat, T. Tonon, C. Leblanc, and A. Groisillier. Characterization of redox sensitive brown algal mannitol-1-phosphatases. *Phycology*, 3 (1):1–12, 2023.
- [271] D.C. Meng, R. Shen, H. Yao, J-C. Chen, Q. Wu, and G-Q. Chen. Engineering the diversity of polyesters. *Current Opinion in Biotechnology*, 29:24–33, 2014.
- [272] B. Leroy, Q. De Meur, C. Moulin, G. Wegria, and R. Wattiez. New insight into the photoheterotrophic growth of the isocytrate lyase-lacking purple bacterium rhodospirillum rubrum on acetate. *Microbiology*, 161:1061–1072, 2015.
- [273] Q. De Meur, A. Deutschbauer, M. Koch, G. Bayon-Vicente, P. Cabecas Segura, R. Wattiez, and B. Leroy. New perspectives on butyrate assimilation in rhodospirillum rubrum slh under photoheterotrophic conditions. *BMC Microbiology*, 20:126, 2020.

-
- [274] P. Cabecas Segura, Q. De Meur, A. Tanghe, R. Onderwater, L. Dewasme, R. Wattiez, and B. Leroy. Effects of mixing volatile fatty acids as carbon sources on *rhodospirillum rubrum* carbon metabolism and redox balance mechanisms. *Microorganisms*, 9:1996, 2021.
- [275] P. Cabecas Segura, R. Onderwater, A. Deutschbauer, L. Dewasme, R. Wattiez, and B. Leroy. Study of the production of poly(hydroxybutyrate-co-hydroxyhexanoate) and poly(hydroxybutyrate-co-hydroxyvalerate-co-hydroxyhexanoate) in *rhodospirillum rubrum*. *Applied and Environmental Microbiology*, 88:e01586–21, 2022.
- [276] E. Walter. *Identifiability of parametric models*. Pergamon Press, 1987.
- [277] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25 (15):1923–1929, 2009.
- [278] F.G. Wieland, A.L. Hauber, M. Rosenblatt, C. Tonsing, and J. Timmer. On structural and practical identifiability. *Current Opinion in Systems Biology*, 25:60–69, 2021.