Constraint-Based Modelling of Metabolic Networks Incorporating Resource Allocation

Hong Zeng

Jesus College



Department of Engineering Science,

University of Oxford

Thesis submitted for

Doctor of Philosophy in Engineering Science

April 15, 2020

Abstract

Escherichia coli is among the most widely used host organisms for engineered functions due to its well-characterized genetics and physiology, ability to achieve high cell density, fast growth and high-yield nature. However, the expression of recombinant genes often imposes burdens on the host and triggers stress responses, which can lead to growth retardation and eventually low productivity of targeted products. Furthermore, it is ubiquitously observed that fast-growing *E. coli* secretes a significant amount of acetic acid (acetate) even in the presence of sufficient oxygen, a phenomenon known as the overflow metabolism. The overflow-mediated accumulation of acetate severely inhibits cell growth and constitutes one major obstacle for efficient protein productions.

Despite being extensively studied for decades, the underlying mechanism of the overflow metabolism remains murky until several recent studies shedding light from the perspective of resource allocation. Thanks to the evolving toolbox of systems biology, this PhD project utilized constraint-based modelling, i.e. genome-scale metabolic models (GEMs) and flux balance analysis (FBA) incorporating resource allocation principles to extend the knowledge on the phenotypic patterns and governing mechanisms of *E. coli*.

Work from the initial stage of this project showed that a concise pathway-level proteome allocation constraint enables the prediction of the metabolic shift from full respiration to the overflow metabolism in various wild-type *E. coli* grown under aerobic-glucose conditions. Besides, the control of carbon flux into pentose phosphate pathway was predicted to play an important role in the activation of the acetate pathway. The original wild-type model was further adapted to simulate batch cultures of recombinant strains. By modelling the heterologous burdens as increased maintenance energy and reduced proteomic resources, the revised model accurately predicted the dynamic change of cell

growth, substrate uptake, by-product secretion and protein production. It also predicted a non-monotonic relationship between the recombinant protein productivity and the specific expression level of the target protein, where the maximum productivity emerged at a moderate expression level. This result indicated that the optimal protein production might rely on a delicate balance between the expression strength of the target gene and the growth rate of the host cell.

Furthermore, we extended the overflow metabolism-centred modelling framework to account for the protein cost of carbon transport. The updated model was able to simultaneously predict the maximal growth rate, the acetate production rate and the glucose uptake rate. Besides, by linking substrate uptake kinetics with the proteome allocation constraint, model simulations suggested that the huge variation of the supposed constant Monod kinetic parameter (K_s) could result from the switch in metabolic strategies.

Additional efforts were made to test the established metabolic models incorporating proteome allocation constraints in three practically relevant case studies, i.e. (i) reducing the overflow metabolism in NAD^+-dependent formate-dehydrogenase recombinant *E. coli*; (ii) an engineered coculture of GFP recombinant *E. coli* and *Acinetobacter baylyi* ADP1 with acetate cross-feeding to improve GFP productivity and (iii) lactate-producing human mesenchymal stem cells (hMSCs) to study the potential driver of the Warburg effect. Simulation results of case (i) showed that the key to alleviating the overflow metabolism lies in the existence of alternative energy pathways with lower protein costs per ATP flux than that of the fermentation. Simulation results of case (ii) confirmed the experimental observation that the synthetic co-culture system increases the carbon utilisation and the overall GFP productivity. Finally, from case (iii), flux estimation using the experimentally determined lactate production rate-growth rate profile indicated that lactate production in

hMSCs may not necessarily be coupled with a reduced respiration activity, which leads to the discussion of potential limitation of the proteome resource allocation.

Overall, the proposed metabolic models incorporating resource allocation constitute useful computational tools for the optimal design of cell factories. The applications of these models in this PhD thesis contribute to the improved understanding of how *E. coli* rewires its physiologies to cope with the ever-changing growth environments.

Acknowledgements

The completion of this PhD project and this thesis would not have been possible without the support and encouragement from my supervisor, collaborators, college, department, family and friends. I would like to thank them for making my life and academic journey in Oxford such a wonderful and fruitful one.

I would like to express my sincere gratitude to my supervisor Professor Aidong Yang for offering me the marvellous opportunity to study in one of the most prestigious universities in the world, for the support of my thesis writing, for sharing me with his immense knowledge, for teaching me how to become a qualified scholar, and for his every patient guidance throughout my time as his student.

I wish to extend my thanks to Mr. Bo Zhang, Ms. Catherine Fan, Mr. Reza Rohani and Professor Wei Huang for kindly sharing their expertise and working with me. I am also much appreciated for the feedback from the Systems Engineering Group.

I am deeply indebted to the Graduate Administrator of Jesus College Mrs. Carole Thomas for her generous support of my study and life in Oxford. I am also thankful to Ms. Kristiana Dahl for offering me useful guidance in the development of the thesis. Heartedly thanks to all academic and non-academic staffs from the University, College and Department of Engineering Science for making my study in Oxford supported and rewarding.

I am grateful to the funding support from China Scholarship Council (CSC), Department of Engineering Science – University of Oxford and Jesus College – University of Oxford. Special thanks to Professor Rene Bañares-Alcántara and Professor Ian Thompson for supporting my application of the CSC PhD scholarship.

To my friends, thank you for your continued encouragement, being there for me in these sleepless nights and helping me conquer all the difficulties in the past four years. To my God, thank you for blessing me to have the healthy body, intelligence and persistence to accomplish my research.

Finally, I would like to dedicate this thesis to my family and loved ones, especially my parents, my brother, my grandparents and my boyfriend for always providing me with the warmest harbour to ease my mind. Your deep love nourished my soul and brought out the best in me.

List of Acronyms

AB-GFP	GFP-producing *Acinetobacter baylyi* ADP1
ac	Acetate
ACONT	Aconitate hydratase
ACKr	Acetate kinase
Adk	Adenylate kinase
AKG	α-ketoglutarate
AKGDH	α-ketoglutarate dehydrogenase
ALE	Adaptive laboratory evolution
ATPM	Maintenance energy
BMCs	Bacterial microcompartments
CAFBA	Constrained allocation flux balance analysis
cAMP	Cyclic adenosine monophosphate
CBM	Constraint-based model
CCR	Carbon catabolite repression
CoA	Coenzyme-A
CS	Citrate synthase
CYT	Cytochrome oxidase
DFBA	Dynamic flux balance analysis
EC-GFP	GFP-producing *E. coli*
ED	Entner-Doudoraff
EFM	Elementary flux mode
EFMA	Elementary flux mode analysis
EFCM	Enzyme-Flux Cost Minimization
EI	Enzyme I
ENO	Enolase
exp	Experimental data
F1P	Fructose-1-phosphate
F6P	Fructose 6-phosphate
FBA	Flux balance analysis
FBA^{ME}	Flux balance analysis with membrane economics
FBAwMC	FBA with molecular crowding
FBP	Fructose-1,6-bisphosphate
FUM	Fumarase
FVA	Flux variability analysis
G6P	Glucose 6-phosphate
GAM	Growth-associated maintenance
gDW	Gram dry weight
GEM	Genome-scale metabolic model
GFP	Green fluorescent protein
GI-malE	Maltose binding protein–glucose isomerase fusion protein
Glc (or glc)	Glucose
GLCDpp	glucose dehydrogenase
GLCt2pp	D-glucose (galactose):proton symporter
GLNS	Glutamine synthetase
GLUD	Glutamate dehydrogenase
GUR	Glucose uptake rate

HEX1	Hexokinase
hMSCs	Human Mesenchymal Stromal/Stem Cells
ICDH	Isocitrate dehydrogenase
LP	Linear programming
MDH	Malate dehydrogenase
ME	Metabolism and macromolecular Expression
MILP	Mixed-integer linear programing
mmol	Millimole
mRNA	Messenger RNA
MW	Molecular weight
nad-Fdh	NAD^+-dependent formate dehydrogenase
NADH-DH	NADH dehydrogenase
NADTRHD	NAD transhydrogenase
NagB	Glucosamine-6-phosphate deaminase
NGAM	Non-growth-associated maintenance
NLP	Non-linear programming
NRPs	Non-ribosomal proteins
ODEs	Ordinary differential equations
OUR	Oxygen uptake rate
OXP	Oxidative phosphorylation
PAT	Proteome Allocation Theory
Pdh (or PDH)	Pyruvate dehydrogenase
PEP	Phosphoenolpyruvate
PfkB	Phosphofructokinase
Pfl	Pyruvate formate lyase
PGL	6-phosphogluconolactonase
PP(P)	Pentose phosphate (pathway)
PQQ	Pyrroloquinoline quinone
PTS	Phosphotransferase system
Pyk	Pyruvate kinase
RA	Resource allocation
RBA	Resource balance analysis
RNAP	RNA polymerase
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
simu	Simulation
SNL	Strictly nutrient-limited
SUCOAS	Succinyl-CoA synthetase
TCA	Tricarboxylic acid
THD2(pp)	NAD(P) transhydrogenase
TKT2	Transketolase

Table of Contents

Abstract ..... **I**
Acknowledgements ..... **IV**
List of Acronyms ..... **VI**
Table of Contents ..... **VIII**
Chapter 1 Introduction **1**
1.1 Background *1*
1.2 Aim and objectives *4*
1.3 Thesis structure *5*
Chapter 2 Literature review: current understanding, mathematical modelling and application of resource allocation **9**
2.1 Introduction *9*
2.2 Cell growth and resources *11*
2.2.1 External and internal resources required for cell growth 11
2.2.2 An active decision-making process 12
2.3 Current understanding of resource allocation *14*
2.3.1 Resource allocation, cellular objective and growth optimality 14
2.3.2 Using proteome resource allocation to explain various biological phenomena 16
2.3.3 Investigating the possible root cause of the overflow metabolism 19
2.3.4 Molecular mechanisms for resource allocation 24
2.3.5 Other considerations 26
2.4 Network-based metabolic models incorporating resource allocation *26*
2.4.1 Coarse-grained RA models 28
2.4.2 Fine-grained RA models 29
2.4.3 Elementary flux modes-based RA models 32
2.4.4 Comparison between different RA models 34
2.4.5 Mathematical basis of RA models for predicting resource allocation 37
2.4.6 Summary of key advantages and limitations of different RA models 42
2.5 Other considerations and perspectives *44*
2.5.1 Cost, benefit and trade-offs 44
2.5.2 Bacterial growth in rich media 45
2.5.3 Limitation of protein cost-derived constraints in RA models 45
2.5.4 Rationale for the scope of this PhD project 46
Chapter 3 Modelling overflow metabolism in *Escherichia coli* with flux balance analysis incorporating differential proteomic efficiencies of energy pathways **53**
Summary *53*
3.1 Background *56*
3.2 Methods *59*
3.2.1 Formulation of the PAT constraint 59
3.2.2 Predicting the acetate flux 61
3.2.3 Interdependency of proteomic cost parameters 62
3.2.4 Adjustment of cellular energy demand 64
3.2.5 Alternative pathways in central metabolism 65
3.3 Results *67*
3.3.1 Model prediction of overflow metabolism with nominal energy demand 68
3.3.2 Adjusting cellular energy demand improves the prediction of biomass yield 76
3.4 Discussion *82*
3.4.1 Comparison to relevant models 82
3.4.2 Parameterisation of the proteome allocation constraint 84

3.4.3 Applicability of the linear formulation of the proteomic cost 86
3.5 Conclusions *88*
Chapter 4 Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant *Escherichia coli*..... 93
Summary *93*
4.1 Introduction *96*
4.2 Methods *98*
4.2.1 Representing proteomic and metabolic burdens 98
4.2.2 Constraint-based metabolic modelling 101
4.3 Results *109*
4.3.1 Modelling wild-type and GFP-bearing recombinant E. coli W3110 110
4.3.2 Modelling GI-malE-bearing recombinant *E. coli* XL1 114
4.4 Discussion *117*
4.4.1 Comparison of proteomic cost parameters between the two cases 117
4.4.2 Effect of recombinant protein production on ATPM 117
4.4.3 Towards a predictive tool to support engineering design for recombinant protein production 118
4.5 Conclusions *121*
Chapter 5 Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation 125
Summary *125*
5.1 Introduction *128*
5.2 Results *131*
5.2.1 Proteomic fraction occupied by the carbon-scavenging sector 131
5.2.2 Predictions of glucose uptake rate and the acetate overflow 135
5.2.3 Deriving the hyperbolic λ-[g] correlation 137
5.2.4 λmax is a function of proteome resource and proteome cost 141
5.2.5 Ks depends on the interplay between fractions of proteome sectors and affinity of transport enzymes 143
5.3 Discussion *145*
5.4 Methods *149*
5.4.1 Modelling proteome allocation in *E. coli* 149
5.4.2 Determination of GAM and ATPM 150
5.4.3 Model simulation 151
Chapter 6 Further applications and testing of proteome allocation-integrated metabolic models 155
6.1 Introduction *155*
6.2 The effectiveness of formate metabolism in reducing the overflow metabolism in NAD^+-dependent formate dehydrogenase recombinant Escherichia coli *157*
6.2.1 Formate being an auxiliary substrate 157
6.2.2 Proteome allocation model incorporating formate metabolism 158
6.2.3 Results and discussion 161
6.2.4 Summary of case (i) 171
6.3 Exploring the optimal recombinant protein expression level in a synthetic coculture *171*
6.3.1 An engineered microbial coculture to improve the production of GFP 171
6.3.2 Formulating metabolic models for EC-GFP and AB-GFP 172
6.3.3 Results and discussion 176
6.3.4 Summary of case (ii) 180
6.4 Investigating the Warburg effect in human mesenchymal stem cells *181*

6.4.1 Warburg effect – aerobic production of lactate 181
6.4.2 Genome-scale metabolic model of hMSCs 182
6.4.3 Results and discussion 183
6.4.4 Summary of case (iii) 190
6.5 Conclusions *191*
Chapter 7 Discussion and Conclusions 195
7.1 Summary of the PhD research and key findings *195*
7.1.1 Summary of main work accomplished 195
7.1.2 Key learnings and outcomes 196
7.2 Limitations and future work *201*
7.2.1 Limitations 201
7.2.2. Future work 204
Appendix I: Supplementary Information of Chapter 3 207
1. Supplementary Text *207*
1.1 The core *Escherichia coli* metabolic model 207
1.2 Integrating PAT to Flux Balance Analysis 207
1.3 Impact of PP pathway ratio on model parameters and predictions 208
2. Supplementary Tables *210*
3. Supplementary Figures *213*
Appendix II: Supplementary Information of Chapter 4 219
1. Supplementary Text *219*
1.1 Adjustments to the GEM 219
1.2 Integrating proteomic constraint to FBA 221
1.3 Illustration of the potential usage of the DFBA model for finding the optimal protein expression level 221
1.4 Sensitivity analysis of model parameters 222
1.5 Selection of representative fermentation and respiration fluxes 223
2. Supplementary Tables *225*
3. Supplementary Figures *226*
Appendix III: Supplementary Information of Chapter 5 232
1. Supplementary Text *232*
1.1 Parameterization of wc-[g] correlation 232
1.2 Determining the proteome cost parameters 234
1.3 Comments on the approach for determining model parameters 235
1.4 Formulating the mixed integer linear programing (MILP) problem 236
2. Supplementary Tables *238*
3. Supplementary Figures *240*
Appendix IV: Supplementary Information of Chapter 6 244

Chapter 1 Introduction

1.1 Background

Ever since fermentation was first discovered and utilized, microorganisms have been closely linked with the evolution of human civilization given their prominent contributions in many life-essential fields, including food, energy, health and agriculture [1–4]. Through years of practice, *Escherichia coli* (*E. coli*) has been identified as one of the most promising host organism for the production of recombinant proteins [5]. The expression of heterologous genes often imposes significant burdens to the host cell and stimulates stress responses such as slow growth, excretion of wasteful and/or toxic by-products, which eventually results in low productivity [6,7]. One way to resolve this issue is to seek for optimal operating conditions through extensive experimental studies. However, such trial-and-error approach inherently suffers from the limited workforce, time and budget. Thanks to the remarkable development of systems biology in the past decade [8], it is now possible to explore the intriguing microbial physiology using systematic approaches.

The evolving toolbox of systems biology offers a wide range of models, modelling approaches and model-based discovery tools. In particular, genome-scale metabolic models (GEMs) and the constraint-based modelling approach, i.e. Flux Balance Analysis (FBA) enable the systematic study of bacterial growth phenotypes via the metabolism. GEMs are mathematical reconstructions of metabolic networks [9]. A GEM contains the stoichiometry matrix of known metabolic reactions and the gene-protein-reaction (GPR) relationships. GEMs have been constructed for human [10], eukaryote [11] and prokaryote [12] cells and are freely available in the online database, e.g. BiGG (http://bigg.ucsd.edu/). What is extensively used in this PhD thesis is the *Escherichia coli* GEM *i*AF1260 [13]. FBA on the other hand is the main analytical tool of GEMs, which transfers the biological information embedded in a GEM into metabolic flux distributions [14]. The purpose of

FBA is to identify the steady-state metabolic flux distributions conferring the optimal value of the objective function. Mathematically, FBA is a constraint-based modelling approach formulated as:

$$\begin{aligned} &\text{Maximize} \quad f_{obj} = \mathrm{c}^{\mathrm{T}}\boldsymbol{v} \\ &\text{Subject to} \quad \text{(i) } S\boldsymbol{v} = 0 \\ &\qquad\qquad \text{(ii) } v^{L} \leq \boldsymbol{v} \leq v^{U} \\ &\qquad\qquad \text{(iii) Other constraints} \end{aligned} \qquad \text{Equation 1-1}$$

where $f_{obj} = \mathrm{c}^{\mathrm{T}}\boldsymbol{v}$ is the objective function that denotes the biological objective of a cell under a specific growth condition. $\boldsymbol{v}$ is the column vector containing all the metabolic reactions specified in a GEM. c is the weighting vector with a value of one at the position of the objective reaction and zeros for the rest. The most commonly used objective function is the maximization of the growth rate – that is the rate of converting biomass precursors (e.g. amino acids, DNA, RNA) into one unit mass of biomass, normally in the unit of gDW (gram dry weight) biomass generated per gDW biomass presenting the culture medium per hour, i.e. h^{-1}. S is the stoichiometry matrix. v^{L} and v^{U} are the lower and upper bounds of the permissible fluxes, normally with a unit of mmol/gDW/h. The resulting linear programming (LP) problem can be solved by various computational solver packages, among which the COBRA Toolbox [15] is the one adopted in this chapter.

FBA and GEMs have been widely applied in system-level metabolic studies [16], for instance, to predict the essential metabolic genes in cancer cells [17], to uncover pathway utilisation strategies upon the metabolic transitions involved in cyanobacterial photosynthesis [18] and to predict the production of D-phenyllactic acid upon gene manipulations in *E. coli* [19]. Constraints lie in the heart of FBA models, which directly

dictate the feasible solution space of optimal metabolic fluxes and growth phenotypes. Classic constraints of FBA include (a) the conservation of mass and energy, e.g. constraint (i) of Equation 1-1 and (b) the minimum and maximum flux limits, e.g. constraint (ii) of Equation 1-1. Various other physiological constraints were identified in the past twenty years [20], among which the proteome allocation constraint [21] and the underlying notion of the protein costs of metabolic pathways and the constraints on resource allocation between biological processes [22] are of particular interests and form the general motivation of this PhD project due to their prominent role in improving the predictions of metabolic models of *E. coli*.

Besides simulating steady-state growth using FBA, in this thesis the growth dynamics of cells in batch cultures is captured by dynamic flux balance analysis (DFBA) [23] where the intracellular metabolism is described by FBA and the extracellular environment is modelled by ordinary differential equations (ODEs). Dynamic simulation is achieved by dividing the batch time into various time intervals. At the beginning of each time interval, FBA growth optimisation is solved. Assuming the intracellular metabolic steady state is maintained over the time interval, the instantaneous maximum growth rate and boundary (exchange) fluxes of interests are passed to ODEs to update the extracellular state variables. This procedure is performed iteratively until the end of the batch time. The steady-state assumption of intracellular metabolism at each time interval requires the time-scale for reaching intracellular metabolic equilibrium to be much quicker than the time-scale for bioreactor-level dynamics. Experiments have shown that the intracellular metabolic equilibrium is generally reached within seconds or minutes in response to an environmental perturbation [24,25]. Therefore, DFBA framework is generally valid for analysing process dynamics of batch fermentations. It is worth noting that FBA does not include the regulatory network and DFBA does not account for the regulatory dynamics. If

changes in environmental states provoke metabolic changes that require transcriptional and translational level regulations (e.g. to synthesize a new set of enzymes), the time needed to achieve the new equilibrium state will not be captured by a DFBA model. Thus, it is inappropriate to use DFBA to investigate scenarios where metabolic regulations take time, e.g. in lag-phase cells or cells under strong burdens. Furthermore, in classical DFBA models the boundary fluxes of the FBA problem are commonly specified by Michaelis-Menten uptake kinetics, where the kinetic parameters are generally valid only for a specific growth state or metabolic strategy. This limitation has been investigated in this thesis and possible approaches to overcome this drawback are discussed in Chapter 5.

1.2 Aim and objectives

As part of a group working on developing systematic approaches to the engineering of microbial systems, this PhD project investigates constraint-based metabolic modelling via primarily FBA [14], DFBA [23], GEMs [13] and proteome resource allocation principles [21,22,26,27]. The overall aim of this project is to use *E. coli* as a model organism to (i) make reliable predictions of the dynamic and steady-state phenotypes of engineered systems and (ii) give optimization guidance in process operation and metabolic engineering. The specific research objectives covered in this project are listed below; the rationale will be further explained in Chapter 2, Section 2.5.4:

1) Predicting the metabolic shift from normal growth (without acetate production) to the overflow metabolism (aerobic production of acetate) using metabolic models and resource allocation principles for wild type *E. coli*.
2) Quantifying the heterologous proteins-mediated burdens and associated impact on key growth physiologies and to explore the optimal expression level of target proteins for recombinant *E. coli.*

3) Investigating the relationship between resource allocation, energy biogenesis, biomass synthesis (cell growth), carbon intake and the availability of extracellular nutrients to bridge the substrate uptake kinetics with the metabolic configurations, and provide mechanistic interpretations of the empirical Monod growth law.
4) Using resource allocation integrated metabolic models to study practically important engineering targets such as increasing the productivity of target products and reducing the wasteful acetate production in synthetic *E. coli* mono- and co-cultures.
5) Exploring the applicability of resource allocation associated theories and models for the explanation of the overflow-related Warburg effect using FBA and growth data of lactate-producing human mesenchymal stem cells (hMSCs).

1.3 Thesis structure

This thesis was compiled in an integrated format, including a combination of conventional chapters and journal paper-based chapters (Figure 1-1).

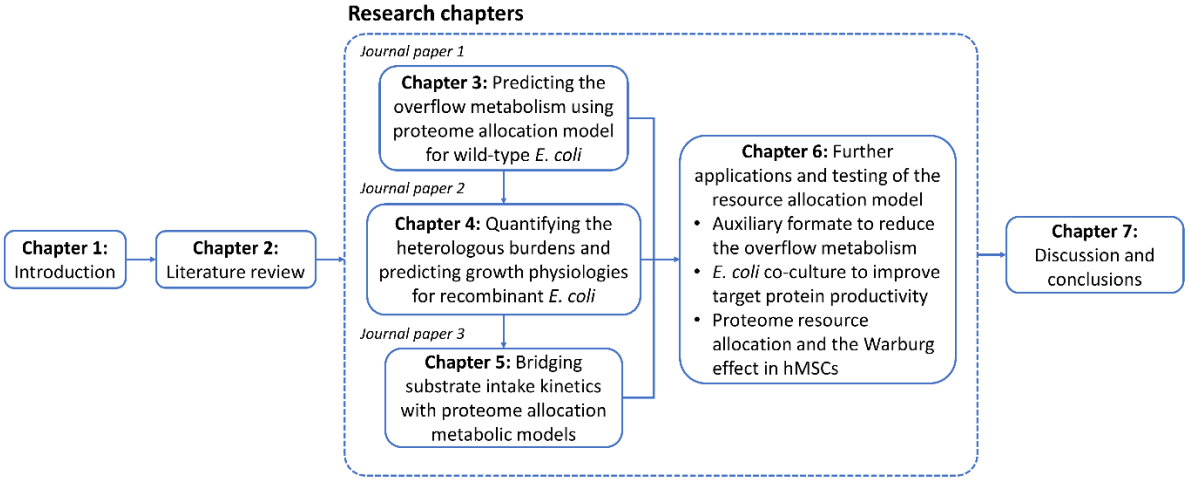


Figure 1-1 Structure of thesis chapters and journal papers.

This introductory chapter is followed by a literature review chapter, and subsequently by a collection of four research chapters where three of them have been published journal articles (with the abstract being replaced by a summary section) and the last one is

presented in the conventional chapter format. Finally, the thesis is rounded up by an integrated discussion and conclusions chapter.

- Chapter 1 introduces the background, motivation and aim of this PhD project and illustrates the structure of this thesis.
- Chapter 2 provides a comprehensive literature review of the theoretical background and research approaches for the intended field of study, including the current understanding between resource allocation and bacterial growth strategies and important network-based metabolic models incorporating resource allocation. The review chapter helps identify research gaps and rationalise objectives of this PhD project.
- Chapters 3-6 are the research chapters. Chapter 3 focuses on the prediction of the metabolic shift from normal growth to the overflow metabolism through integrating a concise pathway-level proteome allocation constraint to a core metabolic model of *E. coli*. The modelling framework established in this chapter is the cornerstone of the other research chapters.
- Chapter 4 investigates the time-dependent impact of the production of heterologous proteins on cell growth, acetate accumulation and target protein production. It extends the initial modelling framework established in Chapter 3 in three aspects: (i) from wild type to recombinant strains; (ii) from steady-state to dynamic system and (iii) from the core model to the genome-scale metabolic model. It also predicts the optimal expression level of target proteins that confers the highest productivity at the end of a batch.
- Chapters 3 and 4 study the overflow metabolism from the competition of proteomic resources between energy production and biomass synthesis processes. Chapter 5 supplements previous findings by introducing an additional carbon transport sector

to the proteome allocation constraint, which bridges substrate intake kinetics with the resource allocation-mediated metabolic shift and provides a new mechanistic interpretation of the Monod growth law.

- Chapter 6 investigates the practical extension of the modelling framework established in Chapters 3-5. It explores the potential application and limitation of resource allocation principles and models using three parallel case studies, i.e. (i) formate as the auxiliary substrate to reduce acetate production; (ii) a synthetic coculture of *E. coli* and *A. baylyi* to improve the productivity of recombinant proteins and (iii) lactate-producing hMSCs to study the possible link between resource allocation and the Warburg effect.
- Chapter 7 offers a separate discussions and conclusion section of this PhD project. It also highlights the limitation of the presented work and provides perspectives for future work.

References

1. Singh R, Kumar M, Mittal A, Mehta PK: **Microbial enzymes: industrial progress in 21st century**. *3 Biotech* 2016, **6**:174.
2. Sanjukta S, Rai AK: **Production of bioactive peptides during soybean fermentation and their potential health benefits**. *Trends Food Sci Technol* 2016, **50**:1–10.
3. Fierer N: **Embracing the unknown: disentangling the complexities of the soil microbiome**. *Nat Rev Microbiol* 2017, **15**:579–590.
4. Braguglia CM, Gallipoli A, Gianico A, Pagliaccia P: **Anaerobic bioconversion of food waste into energy: A critical review**. *Bioresour Technol* 2018, **248**:37–56.
5. Rosano GL, Ceccarelli EA: **Recombinant protein expression in Escherichia coli: advances and challenges**. *Front Microbiol* 2014, **5**:172.
6. Wu G, Yan Q, Jones JA, Tang YJ, Fong SS, Koffas MAG: **Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications**. *Trends Biotechnol* 2016, **34**:652–664.
7. Eiteman MA, Altman E: **Overcoming acetate in Escherichia coli recombinant protein fermentations**. *Trends Biotechnol* 2006, **24**:530–536.
8. Breitling R: **What is systems biology?** *Front Physiol* 2010, **1**:9.
9. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93–121.
10. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proc Natl Acad Sci U S A* 2007, **104**:1777–1782.
11. Doerks T, Copley R, Schultz J: **Systematic identification of novel protein domain families associated with nuclear functions**. *Genome Res* 2002, **12**:47–56.

12. Edwards JS, Palsson BO: **The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities**. *Proc Natl Acad Sci* 2000, **97**:5528–5533.
13. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
14. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245–248.
15. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**.
16. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY: **Current status and applications of genome-scale metabolic models**. *Genome Biol* 2019, **20**:121.
17. Gatto F, Miess H, Schulze A, Nielsen J: **Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism**. *Sci Rep* 2015, **5**:10738.
18. Qian X, Kim MK, Kumaraswamy GK, Agarwal A, Lun DS, Dismukes GC: **Flux balance analysis of photoautotrophic metabolism: uncovering new biological details of subsystems involved in cyanobacterial photosynthesis**. *Biochim Biophys Acta - Bioenerg* 2017, **1858**:276–287.
19. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011**. *Mol Syst Biol* 2011, **7**:535.
20. Stalidzans E, Seiman A, Peebo K, Komasilovs V, Pentjuss A: **Model-based metabolism design: constraints for kinetic and stoichiometric models**. *Biochem Soc Trans* 2018, **46**:261–267.
21. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
22. Goelzer A, Fromion V: **Bacterial growth rate reflects a bottleneck in resource allocation**. *Biochim Biophys Acta (BBA)-General Subj* 2011, **1810**:978–988.
23. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**:1331–1340.
24. Kresnowati MTAP, Suarez-Mendez CM, van Winden WA, van Gulik WM, Heijnen JJ: **Quantitative physiological study of the fast dynamics in the intracellular pH of Saccharomyces cerevisiae in response to glucose and ethanol pulses**. *Metab Eng* 2008, **10**:39–54.
25. Mashego MR, van Gulik WM, Vinke JL, Visser D, Heijnen JJ: **In vivo kinetics with rapid perturbation experiments in Saccharomyces cerevisiae using a second-generation BioScope**. *Metab Eng* 2006, **8**:370–383.
26. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics**. *Mol Syst Biol* 2009, **5**.
27. Basan M: **Resource allocation and metabolism: the search for governing principles**. *Curr Opin Microbiol* 2018, **45**:77–83.

Chapter 2 Literature review: current understanding, mathematical modelling and application of resource allocation

2.1 Introduction

Following the initial suggestion in 2009 by Molenaar *et al.* [1] that cellular growth strategies are dependent on not only metabolism but also the cost of synthesis of proteins, extensive efforts have been made to investigate the biophysical importance of the allocation of macromolecular resources in supporting cell growth. Significant progresses were obtained in the past decade. Quantitative analysis of growth-dependent proteomic datasets elucidates that resource allocation plays a central role in dictating metabolism and gene expression for maximizing the rates of steady-state growth in *Escherichia coli* [2–5].

Resource allocation has been shown to govern transitional growth kinetics upon the nutrient shift in a global manner [6]. Meanwhile, several review articles have been produced to consolidate the knowledge base of resource allocation. These include a revisit of the history of the idea of resource allocation in living organisms and the development of resource balance analysis (RBA) model [7]. Another work demonstrates the power of multiscale metabolic models and omics datasets in elucidating resource allocation principles [8]. Complementary to [8] which focuses on detailed models that integrate metabolic networks with gene expression, a separate review of phenomenological resource allocation models presents the value in making quantitative predictions of microbial phenotypes with only a few adjustable parameters [9]. Furthermore, a recent review summarises the mathematical structures of models that can predict the overflow metabolism, unifies all models into one standard form and concludes that two growth-limiting constraints are essential for predicting the gradual switch from a high-yield to a low-yield pathway [10].

We posit that resource allocation could help answer central bioengineering questions such as how microbes determine growth strategies in ever-changing environments and what overarching governing principles can guide the design of microbial factories. In this review, we aim to provide a comprehensive synthesis of the theoretical understanding of resource allocation, starting from describing a demand-and-supply relationship between cell growth and resources. Building on that, we summarise important network-based metabolic models incorporating resource allocation, supplemented with our evaluation of their key features, advantages and limitations. Furthermore, we pay particular attention to the comparison between competing resource allocation hypotheses and models, extracting their core differences through the proposed root causes and their mathematical basis. Finally, we highlight existing gaps and important aspects for future discovery and explain the rationale for the scope of this PhD project.

2.2 Cell growth and resources

2.2.1 External and internal resources required for cell growth

Microorganisms are unicellular organisms that require a range of resources to grow and self-replicate. We sort growth-associated resources into two major categories: external and internal resources. External resources are environmental input that can be utilised by the cell, e.g. chemical substrates (organic compounds, carbon dioxide) or light. Internal resources comprise those that a cell 'owns' and can actively manipulate, such as genetic information (encoded in DNA molecules), expression machinery (e.g. ribosomes and RNA polymerases (RNAPs)), metabolic machinery (metabolic enzymes), other RNA- or protein-based molecular catalysis as well as membranes and intracellular space. Resource allocation thus exclusively refers to the allocation of internal resources by the cell to support growth (or other cellular objectives).

While external substrates can in principle be infinitely supplied, internal resources generally have physical limitations. For example, the finite abundance of ribosomes (from 7000 to over 70000 per cell [11]) and the limited translational speed (~15 seconds is needed per copy of protein per ribosome [8]) intuitively call for a smart allocation strategy (of ribosomal machinery) for fast growth (or upon nutrient shift) where numerous additional copies of proteins are required.

There are often close links between internal and external resources. The fundamental elements constituting a cell (e.g. carbon, nitrogen, micronutrient) are derived from external resources, e.g. glucose, ammonia and minerals [12]. Importing extracellular resources, for instance glucose, into the cytoplasm often requires various transporters (internal resources) [13]. Furthermore, free energy (internal resource) is needed to fuel biological processes, which is generally extracted from external nutrients (or light) via respiration (or

photosynthesis) and is stored and transferred mostly via ATP molecules. Finally, membranes isolate a cell from the outer environments, allowing a relatively stable and mild internal space to house fragile and delicate biological apparatus (DNA, ribosomes and enzymes). In addition to physical separation, surface membranes also accommodate a variety of proteins that are critical for the control of the mass and information transfers between internal and external spaces.

2.2.2 An active decision-making process

We postulate that many of the phenotypic dynamics are attributable to the trade-off between demand and supply of resources through an active decision-making process (Figure 2-1).

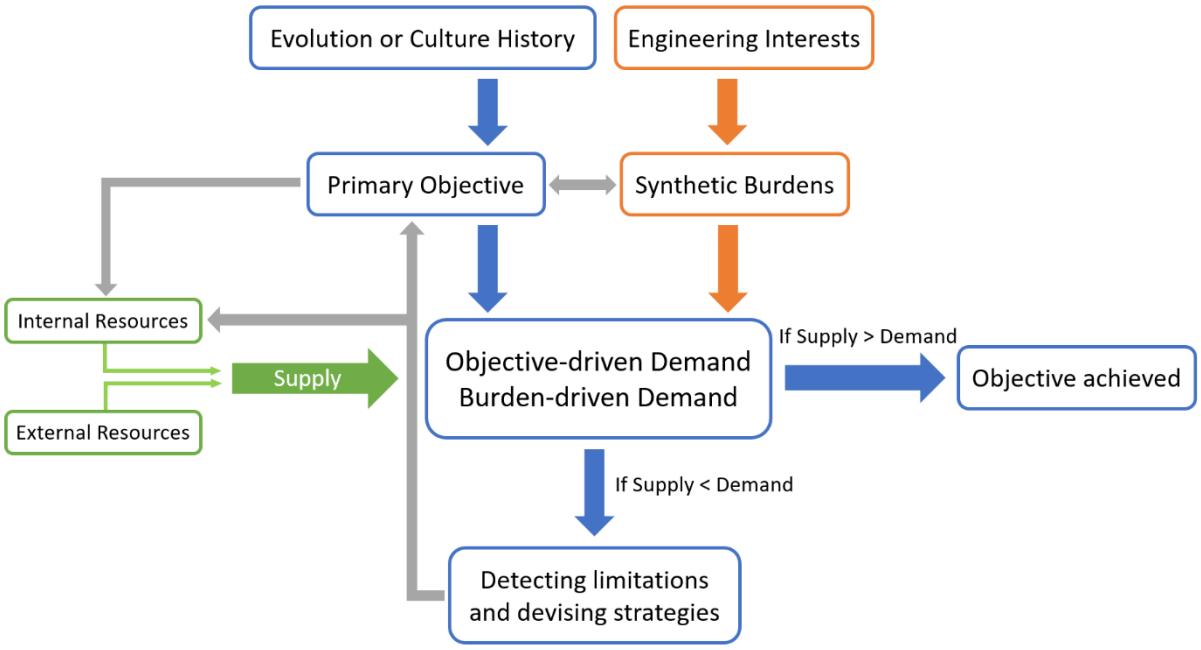


Figure 2-1 Illustration of a putative decision-making process underlying growth strategies.

The intrinsic objective of a cell is normally proliferation, i.e. growth. This objective generates mass and energy demands. Cellular demands can also originate from synthetic burdens. For instance, the expression of recombinant genes requires additional materials, energy and expression machinery [14], and sub-lethal antibiotic dosage provokes the need

for more ribosomes to maintain growth [2]. These artificially imposed burdens can shape the cellular objective and demands. Subsequently, the cell would face a question whether the (external and internal) resources can fulfil the demands. If the demands are sufficiently met by supply, the primary objective can be successfully achieved. However, if the resources cannot meet the requirement, a 'limited' or 'stressed' state will arise (note that the term 'stressed state' here includes both internal and external-driven stresses, which is more general than the environmental stress used in biology). Under certain circumstances, the cell could respond to the stress by manipulating its physiology to bypass the limitation without interfering the original objective. However, if the discrepancy between demand and supply is irreconcilable, the cell may reconfigure the intrinsic objective to compromise, e.g. gradually changing from growth to survival. The internal resources will then be adjusted accordingly to fit the changed objective. This putative decision-making process encompasses the occurrence of resource allocation under both nutrient-limited and nutrient-rich growth conditions, i.e. resource allocation is not always growth-rate driven (more discussion is provided in Section 2.3.1).

Reflecting on previous discussions in this broad area (e.g. in [15] and [8]), we posit that cellular objectives, burdens and limitations (or stresses) are closely entangled, and together contribute to the tug of war between supply and demand chains; it seems neither logical nor feasible to discuss one concept independently without considering the others. Applying orthogonal perturbations [9] (and adaptive laboratory evolution (ALE) if stress-driven mutations are of interests [16]) would help capture the transition between distinct cellular objectives to uncover the comprehensive decision-making atlas in dictating various phenotypes across growth environments.

2.3 Current understanding of resource allocation

2.3.1 Resource allocation, cellular objective and growth optimality

Our knowledge in resource allocation thrives in the study of modulation of proteome compositions to facilitate steady-state growth [2–5]. However, it has been noted that growth rate-dependent resource (proteome) allocation is context specific [8]. For instance, the proteome of *E. coli* grown on pyruvate, glycerol and galactose is not optimized for fastest growth [17]. Furthermore, a proteomic study of 'persisters' (i.e. bacteria with transient antibiotic tolerance) reveals that resource allocation plays a key role in coordinating metabolism towards maximizing energy yield, instead of biomass production [18]. Mounting evidence suggests that variation in the macromolecular composition can be a common consequence of cells facing vicious threats. The specific proteome adjustment observed in persister cells, which is comparable with that in starved and stressed cells, was shown to be driven by increased ppGpp levels (a result of general stress or stringent response) rather than a mere consequence of reduced growth rate [18]. Strong environmental perturbations could also alter the protein composition through directly or indirectly interfering protein synthesis processes [19].

Growth conditions can change from an optimal state (where the external nutrients can efficiently support maximal growth), through a nutrient-limited state (where a cell can still grow but at a reduced rate or needs to adjust its physiology, e.g. secretion of a by-product, to maintain the same growth rate) and finally to a growth-threatening state (e.g. under osmotic, pH and temperature stresses where the cell could only maintain slow growth or even completely stop growing [20]) (Figure 2-2). The cellular objective would correspondingly shift from growth to survival. We posit that resource allocation emerges to increase cellular fitness to the varying environments. As the cellular objective changes,

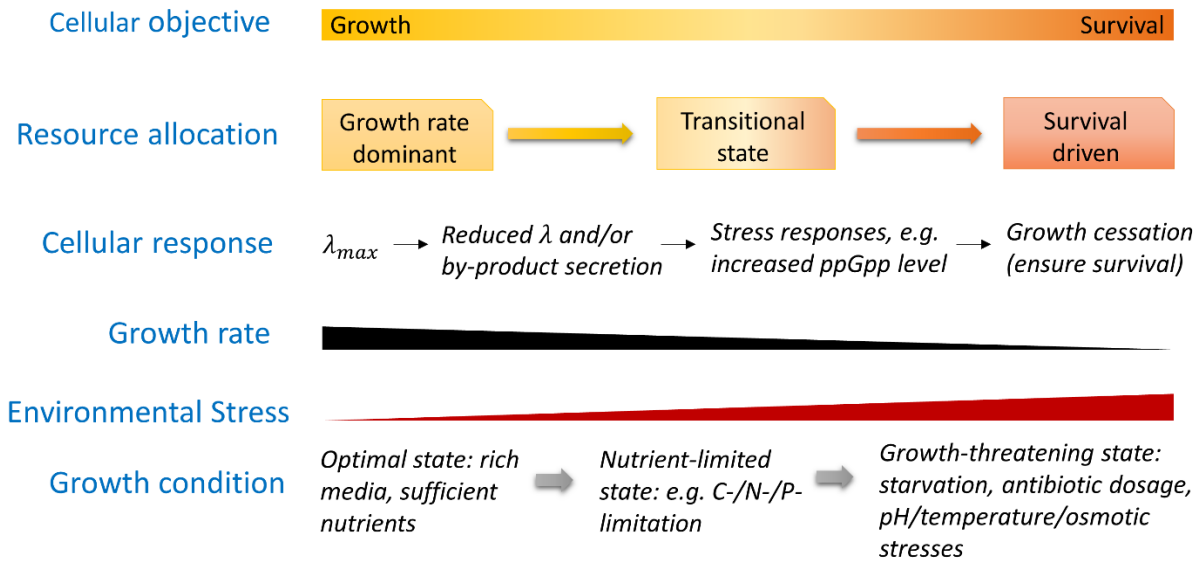


Figure 2-2 Cellular objective, growth rate and resource allocation upon changing environmental conditions. The environmental stress increases as growth conditions vary from optimal (sufficient nutrients to support the maximal growth) to nutrient-limited state, e.g. carbon, nitrogen or phosphorus limitation and finally to growth-threatening state (strong perturbations on bacterial homeostasis). During this process, the growth rate reduces from the maximal value (λ_{max}) to almost zero (i.e. the cell stops growing). The signal molecule ppGpp accumulates as a result of general stress or stringent responses. The cellular objective shifts from maximizing growth to ensuring bacterial survival. Subsequently, resource allocation becomes less coupled with growth rate and eventually functions for survival.

resource allocation would become less coupled with growth rate and eventually serve for ensuring survival.

Closely related to resource allocation, pre-allocation strategies [8] have been considered to bear additional evolutionary importance. A considerable amount of proteins in *E. coli* is expressed with no immediate benefit given specific growth conditions [21]. Many catabolic genes for substrates that are not presented in the medium are upregulated with decreasing carbon quality [3,5]. In addition to protein pre-allocation, carbon- and nitrogen-limited cells have a higher fraction of inactive ribosomes (i.e. free ribosomes not bond to mRNA) than phosphorus-limited cells, which facilitates growth acceleration during nutrient upshift [22]. Pre-allocation of ribosomal capacity was also observed in cells

undergone famine-to-feast cycles and is considered beneficial to the overall gain of biomass [23]. A recent modelling work quantitatively shows that spare ribosomal capacity prevents metabolic overshoots and permits rapid response to nutritional upshifts [24]. While it is generally considered that constitutive pre-allocation of cellular resources (proteins and/or ribosomes) can provide preparatory advantages to hedge against sudden environmental changes [25], such strategic decision imposes substantial burdens. This may reflect an evolutionary choice of maintaining higher robustness at the cost of slower growth. Instead of being merely growth rate-dependent, we consider that resource allocation plays a central role in both facilitating fast growth, ensuring bacterial survival and enabling high adaptability in fluctuating environments; an 'optimal growth strategy' should thus not be evaluated by growth rate only. Despite the inability to grow at the fastest rate can be explained by some limitations of regulatory capabilities [9], it is also possible that a seemingly 'sub-optimal' growth strategy that enables better response upon nutrient shifts is in fact the 'optimal' one selected by evolution.

One important implication of the diversity in biological optimality is that single-objective models (e.g. maximal growth) may not be able to depict the physiology of cells undergoing a transition between growth and survival or cells with alternative objectives. It might be possible to capture the 'real' behaviour through a model with multiple objectives where the trade-off between short-term and long-term advantages can be accounted for in the optimisation.

2.3.2 Using proteome resource allocation to explain various biological phenomena

Growth rate-mediated resource allocation provides insights for many well-known phenomena, e.g. cyclic adenosine monophosphate (cAMP)-dependent carbon catabolite repression (CCR) [3] and overflow metabolism [4]. It has been shown that the

physiological function of cAMP-mediated CCR is to ensure proteome resources are invested as needed for bacterial growth under diverse nutrient conditions, e.g. more resources are directed to biosynthetic processes as growth rate increases (Figure 2-3(a)).

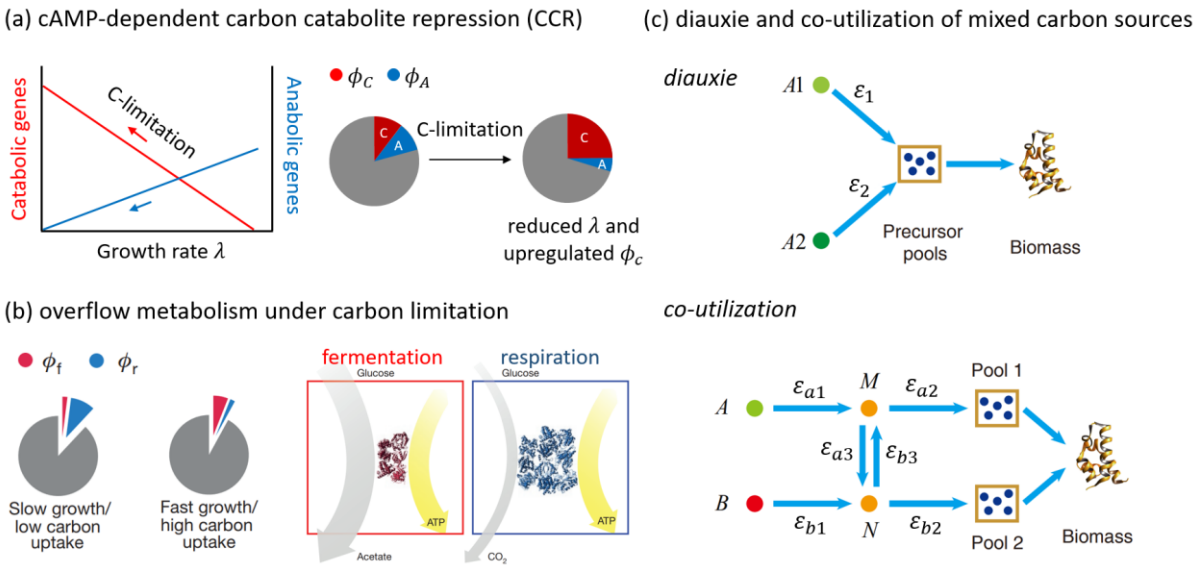


Figure 2-3 **(a)** Illustration of the physiological function of cAMP-dependent carbon catabolite repression (CCR) in allocating proteomic resources to meet growth demand. Under carbon limitation, catabolic genes (namely mass fraction of catabolic proteins ϕ_C) are upregulated while anabolic genes (indicated by the mass fraction of anabolic proteins ϕ_A) are downregulated with decreased growth rate. **(b)** Illustration of the overflow metabolism in fast-growing *E. coli* taken from ref. [4]. The fraction of total proteome allocated to fermentation (ϕ_f) and respiration (ϕ_r) is different between slow growth (low carbon uptake) and fast growth (high carbon uptake). The key driver of such modulation of proteome resources lies in the much lower protein investment per ATP flux (yellow arrow) of the fermentation pathway compared with respiration. **(c)** Illustration of the coarse-grained model of diauxie and co-utilization of carbon sources adapted from ref. [26]. For diauxie, two group A sources (*A1* and *A2*) can both supply precursor pools for biomass production but with different pathway efficiencies (ε_1 and ε_2). The one with higher efficiency is preferred for maximal growth. If two precursor pools supply the biomass synthesis, each pool derives from an intermediate node *M* or *N*. Either intermediate node can draw flux from either of the two sources *A* and *B*. Co-utilization occurs under conditions where the overall efficiency from source *A* to Pool 1 ($\varepsilon_{a1} + \varepsilon_{a2}$) and from source *B* to Pool 2 ($\varepsilon_{b1} + \varepsilon_{b2}$) are optimal.

On the other hand, Basan *et al.* [4] proposed and validated that proteome allocation plays a critical role in regulating the overflow metabolism. They quantified that the protein cost

per ATP produced by the fermentation pathway is about 67% of that by the respiration pathway, which constitutes the key driver of the activation of the (low-proteomic-cost) fermentation pathway at rapid growth (Figure 2-3(b)). However, the long-standing debate on what biophysical bottleneck triggers the overflow metabolism has not yet achieved an agreement. Competing hypotheses including early notions of limited respiratory capacity [27] (e.g. limited catalytic rate of tricarboxylic acid cycle enzymes, limited electron transfer rate [28] or oxygen uptake rate [29]), the regeneration of coenzyme-A (CoA) to cope with the limited CoA pool [30] and the maintenance of redox homoeostasis [31]. More recent proposals include macromolecular crowding [32,33], limited membrane occupancy [34,35], optimal investment cost-benefit trade-offs [36] and cost of protein synthesis [1,4,37]. Furthermore, a coarse-grained model of optimal allocation of protein resources quantitatively explained why and how the cell chooses between diauxie- and co-utilization of substrates under mixed carbon sources [26]. Carbon sources were categorized into those introduced at the upper part of glycolysis (Group A sources) and those entering at other nodes of the metabolic network (Group B sources). Prioritised carbon utilization occurs among Group A sources, and usually the one that supports higher growth rate (associated with higher substrate quality and higher pathway efficiency) is preferred. If Group A and Group B sources are both present in the media, co-utilisation would arise if it is more economical for some precursor pools to take a shortcut of drawing carbon flux from Group B sources while for other precursor pools taking up Group A sources is more efficient (Figure 2-3(c)). Besides, the remarkable achievement of resource reallocation in explaining the overflow phenomenon in *E. coli* shed light on uncovering regulatory principles for other long-standing overflow phenomena, e.g. the production of ethanol in *Saccharomyces cerevisiae* [38–40] and the production of lactate in cancer cells [41–43].

2.3.3 Investigating the possible root cause of the overflow metabolism

The resource allocation-based propositions (summarised in Section 2.3.2) offer renewed explanations to the overflow metabolism, but they are not without controversies. In particular, the macromolecular crowding (also known as 'molecular crowding') hypothesis [32,33] and the proteome allocation hypothesis [4] (one form of the protein cost hypothesis) have been considered hard to reconcile with each other [9]. In this section, we discuss different hypotheses towards deciphering the potential root cause of the overflow metabolism.

It has been postulated that a cell has an upper limit or optimal macromolecular density [44]. Molecular crowding suggests the hard bound on the intracellular macromolecule concentration (or equivalently the finite cell volume or the solvent capacity constraint) triggers the metabolic shift from full respiration to the overflow metabolism [32,33]. This statement is supported by the observed change in cell buoyant density of *E. coli* MG1655, which gradually increases when growth rate increases from 0.1 to 0.4 h^{-1} and stays roughly constant at higher growth rates [45]. Therefore, the root cause of the overflow metabolism in molecular crowding is hypothesised to be the finite cell buoyant density.

On the other hand, the researchers developing the proteome allocation hypothesis observed significant changes in the proteome composition, in particular for energy biogenesis, upon the metabolic shift from normal growth to the overflow metabolism in *E. coli* NCM3722 [4]. By combining the flux measurements and proteomic data, they determined that the protein investment per ATP flux of the fermentation is about twice as efficient as that of respiration. This leads to their critical argument that the root cause of the overflow metabolism is the cell's preference of more proteomic efficient pathways at fast growth. Besides, Basan *et al.* [4] argued that the macromolecular density (or the cell volume)

constraint is not a valid constraint [4,9] as (i) the cell volume varies widely between growth conditions with similar densities [46] and (ii) they did not observe variations in cell density (within a wide range of growth rates) in their own measurements [47].

In a subsequent discussion, Vazquez and Oltvai from the molecular crowding 'camp' suggested that the prediction of the overflow metabolism by the proteome allocation hypothesis in [4] is achieved through 'implicit assumptions that expand beyond the hypothesis of proteome allocation alone' [44]. More specifically, they showed mathematically that in addition to the differential proteome efficiencies between alternative metabolic pathways, the accurate prediction of the experimentally observed fluxes in the overflow region also requires (1) "a non-zero density of non-metabolic macromolecules", and (2) "an upper bound in the cell macromolecular density". They further claimed that molecular crowding "explains" the latter point and hence "is a key factor in explaining overflow metabolism".

Part of the arguments of the above two groups is concerned with cell densities at different growth rates. It should be pointed out that the reported range of growth rates with constant cell density for *E. coli* NCM3722 (in the proteome allocation hypothesis) is 0.3 to 2.0 h^{-1} [47], and the overflow metabolism of this strain occurs at growth rates above ~0.8 h^{-1} [4]. On the other hand, the reported growth rate range of *E. coli* MG1655 (in the molecular crowding hypothesis) is 0.1 to 0.7 h^{-1} with the acetate overflow occurring at growth rates above 0.4 h^{-1} [33]. For MG1655, the cell buoyant density increases with the growth rate within the range of 0.1 to 0.4 h^{-1} and plateaus at growth rates above 0.4 h^{-1} [45]. One can see several key differences between these two cases in terms of the strain, the range of growth rate, and the growth rate at the onset of overflow; these differences call for cautions when comparing alternative theories.

Other mathematical models incorporating resource allocation also imply additional mechanisms inducing the overflow metabolism. FBA^{ME} (membrane economics) [35] suggests the simultaneous use of fermentation and respiration at high growth rates is an outcome of finite cytoplasmic surface area (available for respiratory membrane proteins). This membrane economics hypothesis is reinforced by a recently proposed membrane real estate hypothesis [34], with the presentation of experimental evidence of decrease in surface-to-volume ratio and limitation on membrane-protein packing capacity at increased growth rates. However, the researcher of the proteome allocation hypothesis questioned the inability of membrane space constraints in explaining the overflow metabolism emerging in slow-growing cells expressing a large amount of useless proteins (which can be well explained by the proteome allocation theory) [9]. Additionally, Resource Balance Analysis (RBA) [15,48] and Metabolic and macromolecular expression (ME) [49,50] models (both reviewed in detail in Section 2.4) indicate that the change in the macromolecular composition (primarily proteins) and the occurrence of the overflow metabolism are derived from limited synthesis capacity of macromolecules (e.g. limited translational rate) and limited efficiencies of molecular catalysis (e.g. limited enzymatic catalytic rates).

The proposals described above suggest that the root cause (fundamental physiological limitation) of the overflow metabolism is not a concluded matter. At least two competing explanations, i.e. space limitations (represented by molecular crowding [32,33] and membrane economics [35] hypotheses) and machinery limitations (represented by RBA [15,48], ME [49,50] and possibly implicitly in proteome allocation [4] hypotheses) have been supported by experiments and/or mathematical models at least to a certain degree. To understand how the proteome composition is linked to the allocation of the ribosomal machinery, we conducted the following derivations.

The time-dependent change of the concentration of protein i (P_i) can be modelled as protein generation (via ribosome translation) minus dilution and degradation rate.

$$\frac{dP_i}{dt} = r_i^g - \lambda P_i - D_i P_i \quad \text{Equation 2-1}$$

r_i^g is the generation rate of protein i, which is proportional to the ribosome abundance (R), i.e. $r_i^g \propto R$ [51]. λP_i is the protein dilution term with λ being the specific growth rate. D_i is the degradation rate of protein i [52,53]. We define the partition (allocation) of the ribosomal machinery θ_i as the production rate of protein i divided by the overall production rate of proteins.

$$\theta_i = \frac{r_i^g}{\sum_i r_i^g} \quad \text{Equation 2-2}$$

At steady-state exponential growth, $dP_i/dt = 0$ and λ is constant, Equation 2-1 implies

$$r_i^g = (\lambda + D_i)P_i \quad \text{Equation 2-3}$$

Substituting Equation 2-3 into Equation 2-2 gives

$$\theta_i = \frac{(\lambda+D_i)P_i}{\sum_i(\lambda+D_i)P_i} \quad \text{Equation 2-4}$$

The average D_i for non-ribosomal proteins (NRPs) has reported being ~0.1 per total NRPs per hour [51]. In growing *E. coli*, about 40% mass of total proteins is stable while the half-life of most of the remaining proteins is between 15-30 hour, corresponding to a range of degradation rate of 0.05-0.02 h^{-1} [54]. During steady-state exponential growth, especially in the overflow region where the specific growth rate of *E. coli* can readily reach 0.8 h^{-1} and higher [4], it is reasonable to assume $\lambda \gg D_i$. Protein degradation term thus can be considered negligible compared with the growth dilution effect. Equation 2-4 can be simplified into

$$\theta_i = \frac{\lambda P_i}{\sum_i \lambda P_i} = \frac{P_i}{\sum_i P_i} = \phi_i$$ Equation 2-5

ϕ_i denotes the proteomic fraction of sector i [4]. It is worth noting that Equation 2-5 is generally valid for cells striving for maximal growth. As discussed in Section 2.3.1, strong synthetic burdens or environmental perturbations can shape the intrinsic objective of the cell (e.g. changing from growth to survival), which may spoil the assumption $\lambda \gg D_i$. Equation 2-5 thus should not be applied to these scenarios.

Nevertheless, an important implication of Equation 2-5 is that the division of ribosomal machinery is approximately equal to the proteome composition if the cell is growing in the exponential phase (i.e. steady-state rapid growth). More importantly, if the proteome fractions can be seen as a direct proxy of the partition of ribosomal machinery, the hard proteomic constraint proposed by Basan *et al* [4] can directly result from limited translational capacity, not necessarily being a special case of the molecular crowding theory.

Nonetheless D_i has been reported to vary considerably between different proteins in a cell, e.g. in *E. coli* [54] and in *Lactococcus lactis* [55]. At slow growth where dilution becomes comparable with protein degradation, the discrepancy in D_i of individual proteins could become impactful. Overall, one important conclusion is that the proteome composition and observed proteomic constraints can closely link with the allocation of the ribosome machinery, and this dependency is particularly strong in steady-state rapid growth.

It is worth noting that the mathematical derivation above shows the approximate equality between $\theta_i$and ϕ_i, but it does not provide any evidence for the root source of the 'limitation' observed for ϕ_i, which can derive from (i) either the shortage of ribosomal machinery θ_i, which triggers the cell's regulation of protein synthesis, or (ii) the shortage

of space, which triggers the cell's regulation of protein synthesis; as part of the response, θ_i will be adjusted, leading to the change in ϕ_i.

2.3.4 Molecular mechanisms for resource allocation

To achieve global reallocation of the proteome and redistribution of metabolic fluxes, multi-level regulatory mechanisms, i.e. fast allosteric regulations and slow transcriptional and translational regulations are needed to (a) sense the current growth state and (b) adjust the macromolecular state accordingly.

Growth state can be sensed by the pool size (concentration) of intracellular metabolites. One major flux sensor is fructose-1,6-bisphosphate (FBP), a key metabolite of the upper glycolysis. FBP can allosterically activate pyruvate kinase (Pyk) in a feedforward manner to match the enzyme activity of the lower glycolysis with the upstream fluxes, resulting in a linear correlation between FBP concentration and the glycolytic flux [56].

Intracellular metabolites also indirectly regulate the enzyme abundance via transcriptional factors (TFs). At rapid growth, the increased pool size of FBP and fructose-1-phosphate (F1P) inhibits catabolite repressor/activator Cra activity [57]. α-ketoacids also play key regulatory roles for central carbon metabolism. α-ketoglutarate (AKG) can bind to and inhibit enzyme I (EI) of the phosphotransferase system (PTS), which deactivates adenylate cyclase (Cya) (most likely through the increased amount of unphosphorylated EIIA) and leads to lower cAMP activity [58]. Unphosphorylated EIIA also activates the carbon storage regulator CsrA via CsrD; the activated CsrA has a positive effect on the synthesis of many glycolytic enzymes (Pgi, PfkA, TpiA, Eno, PykF) thus supporting glycolysis [59]. In addition to regulating catabolism indirectly via PTS proteins, AKG (and several other α-ketoacids, e.g. oxaloacetate, pyruvate) is able to directly repress cAMP synthesis via Cya [3]. In addition, the increased pool size of glucose 6-phosphate (G6P), fructose 6-

phosphate (F6P) and FBP at high growth rate and high glucose uptake rate results in sugar-phosphate stress [60], which accelerates the degradation of specific mRNA and reduce mRNA stability by inducing endoribonuclease [61].

The regulation of metabolic fluxes is also backed up by allosteric regulation of enzyme catalytic efficiency. In *S. cerevisiae*, FBP can directly inhibit the activity of two complexes of the respiratory chain (complex III ubiquinol:cytochrome *c* oxidoreductase and complex IV cytochrome *c* oxidase), which reduces the respiratory flux and induces the Crabtree effect [62]. At increased growth rates, the activation of Pyk due to FBA leads to a decreased PEP/pyruvate ratio and an increased amount of unphosphorylated HPr (PTS component) [63]. Unphosphorylated HPr activates pyruvate kinase (PykF), phosphofructokinase (PfkB), glucosamine-6-phosphate deaminase (NagB) and adenylate kinase (Adk) and modulates glycolytic fluxes through protein-to-protein interactions [64].

Furthermore, the reallocation of the proteome is facilitated by modulating translational capacity. In *E. coli*, protein RelA can sense the amino acid starvation via detecting uncharged tRNA that enters the A site of ribosomes. Amino acid limitation induces the synthesis of alarmone ppGpp by RelA [65]. ppGpp was proposed [15,66,67] to regulate the ribosomes abundances by inhibiting the synthesis of ribosomal RNA (rRNA) via repressing the transcription of *rrn* genes [68]. The change in the ribosome population not only affects the distribution of cellular resources between ribosomes and other cellular processes, but also the protein synthesis capacity. Slow translation can also result from the accumulation of inactive ribosomes (i.e. not binding to mRNA) under carbon limitation conditions [22]. Although the underlying mechanisms remain unclear, these mRNA-free ribosomes allow rapid growth acceleration upon nutrient upshift [22].

2.3.5 Other considerations

The notion of (internal) resource allocation introduced in Section 2.2.1 can in principle be embodied by the allocation of other intracellular resources, such as RNA in addition to the proteome allocation. However, we did not find evidence for intrinsic RNA allocation independent from proteome allocation (the major topic of this thesis) or ribosome allocation (discussed in Section 2.3.1). Nevertheless, artificial controls of mRNA populations [69] and the activity of RNAP [70] to direct the cellular resources to synthetic circuits have been achieved in synthetic biology.

On the utilisation of intracellular space, which can be considered as a type of resource, the experimental observations of maximal cell buoyant density [45] and limited membrane protein density [34] imply that the space allocation can happen either over the whole cell or within certain subcellular compartments. In light of natural strategies of the spatial organization of metabolism (e.g. organelles in eukaryotic cells or bacterial microcompartments (BMCs) in prokaryotic cells) [71], there could be more sophisticated strategies for dynamic spatial organisation yet to be discovered.

2.4 Network-based metabolic models incorporating resource allocation

Here, we summarize network models of cell growth and metabolism that incorporate resource allocation (RA) according to three classes, namely (i) coarse-grained metabolic models that incorporate phenomenological constraints of macromolecular expression; (ii) fine-grained models that integrate metabolic models with detailed matrices of translational and transcriptional processes and (iii) elementary flux mode (EFM)-based models that investigate resource allocation through the topology of the metabolic networks (Figure 2-4 and Table 2-1).

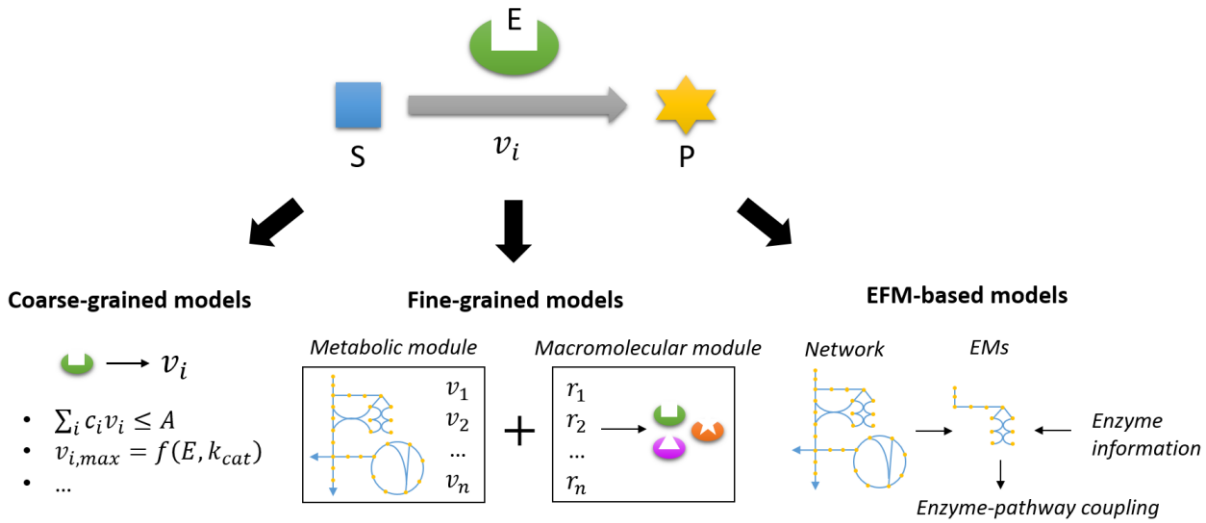


Figure 2-4 Illustration of modelling frameworks that incorporate resource allocation. The top panel shows a simplified one-step enzymatic reaction (v_i) where substrate (S) is converted to a product (P) by the enzyme (E). The role of the enzyme in dictating metabolic fluxes is modelled in coarse-grained models, fine-grained models and EMF-based models.

Table 2-1 List of important metabolic models incorporating resource allocation.

Year	Coarse-grained RA models	Fine-grained RA models	EFM-based RA models
2007	FBAwMC [32,33]		Investment-operating costs analysis with EMA [36]
2011	FBA^{ME} (membrane economics) [35]		
2012		*E. coli* ME [49] *T. maritima* ME [72] *M. genitalium* whole-cell model [73]	
2013		iOL1650-ME [50]	
2014		iJL1678-ME [74]	
2015		*B. subtilis* RBA [48]	
2016	CAFBA [75]	SectorME [76]	
2017	GECKO [77]	FoldME [78] *S. elongatus* GEM [79]	
2018		iJL1678b-ME [80]	EFCM [81]
2019	Energy metabolism RA model [40]	OxidizeME [82] DynamicME [83] TR (Tunable Resolution)-ME [84] *C. ljungdahlii* ME [85] E. coli RBA [86]	minimal-EFM extremum principle [87]

2.4.1 Coarse-grained RA models

Coarse-grained resource allocation models incorporate the resource allocation as phenomenological constraints such as linear growth rate dependence of proteome fractions [75] and finite solvent capacity (mathematical interpretation of the macromolecular crowding) [32,33]. The key to the prediction of resource allocation is directly using enzyme costs to constrain the solution space of metabolic fluxes, which has been formulated as

i. maximum reaction rate (v_{max}) being a function of enzyme abundances (E, limited by an upper bound based on absolute proteomics) and turnover rates (k_{cat}, queried from BRENDA), i.e. $v_{max} = f(E, k_{cat})$ [77] or
ii. metabolic flux multiplying by a 'cost' ($c_i \times v_i$) and the summation of the 'weighted' fluxes is constrained by an upper bound, i.e. $\sum_i c_i v_i \leq A$ [32,33,75], where c_i is the cost coefficient, v_i is the metabolic flux, i is the number of fluxes to be constrained and the constant A is a hard physical limitation. In the second formulation, the metabolic switch will be predicted as long as the cost c_i is not zero and the subtotal of weighted fluxes are different between pathways.

One major success of coarse-grained models is the prediction and explanation of the overflow metabolism in *E. coli* (e.g. FBAwMC (molecular crowding) [32,33], FBA^{ME} (membrane economics) [34,35] and CAFBA [75]) and the Crabtree effect in *Saccharomyces cerevisiae* (e.g. GECKO [77]). More recently, the conservation of the total mass of proteins involved in energy metabolism has been modelled to account for the switch between high-yield and low-yield pathways for *E. coli* and *Saccharomyces cerevisiae* [40].

2.4.1.1 Advantages and limitations of coarse-grained RA models

Coarse-grained resource allocation models have proven to be valuable extensions of the genome-scale metabolic models [9]. The direct mathematical formalism with only a few phenomenological parameters that can be estimated from experimental data or retrieved from protein databases like KEGG [88], BRENDA [89], and Uniprot [90] makes coarse-grained models easier to construct. However, we consider that coarse-grained models are more suitable for interpolation than extrapolation. Models established based on growth-driven resource allocation [75] is not capable of explaining the proteome configuration for suboptimal growth [17]. The lack of mechanistic details limits coarse-grained models in exploring undiscovered biological rules.

2.4.2 Fine-grained RA models

2.4.2.1 ME models

One major class of fine-grained RA models is the Metabolic and Gene expression (ME) model [49,50]. ME models seamlessly couple two types of genome-scale reconstructions, i.e. metabolism (M-matrix) [91,92] and gene expression (E-matrix) [93]. E-matrices are detailed mathematical representations of the transcriptional and translational machinery. *E. coli* E-matrix uses 11,991 components (including DNA, mRNA, proteins, protein complexes, and metabolites) and 13,694 reactions to describe the expression of proteins and RNA molecules [93]. In short, ME models integrate the extensive database of metabolism and macromolecular expression, which enables the exchange of information between temporally and functionally distinct subsystems.

The molecular-level information in initial versions of *E. coli* ME [49,50] has been expanded to include protein translocation (iJL1678-ME) [74], constrained mass fraction of protein sectors (SectorME) [76], protein thermosensitivity (FoldME) [78] and reactive oxygen species (ROS)-mediated stress responses (OxidizeME) [82]. The latest

computational and analytical software for ME models, COBRAme, reconstructed a condensed *E. coli* ME, iJL1678b, with 7,027 components and 12,677 reactions. iJL1678b-ME predicts 52 additional essential genes under minimal-glucose conditions with significantly improved computational speed [80]. The steady-state *E. coli* ME was upgraded to a dynamic framework, i.e. DynamicME to account for the temporal and substrate concentration-dependent variation of metabolism, protein expression, and macromolecular composition [83]. More recently, a multiscale model combining the genome-scale ME with a coarse-grained proteome allocation model was developed to decipher the potential mechanisms of observed two-dimensional rate-yield trade-offs in adapted *E. coli* strains [84]. This multiscale model takes advantage of the coarse-grained model in the direct parameter estimation to efficiently tune the enzyme efficiency parameter (k_{eff}s) of the larger scale ME. This modelling approach with 'tuneable resolution' offers a starting point for the further integration between coarse-grained and fine-grained models. A similar approach has been used to reconstruct *Thermotoga maritima* ME [72]. Furthermore, the first ME model for gram-positive bacterium *Clostridium ljungdahlii i*JL965-ME has been developed, which predicted the proteome allocation and the sequential production of acetate and ethanol as a function of substrate availability [85].

2.4.2.2 Fine-grained model less detailed than ME models

Another typical example of fine-grained RA models Resource Balance Analysis (RBA) [94]. RBA model of *Bacillus subtilis* comprises 614 chemical reactions, 467 metabolites and ions, and 672 protein-coding genes [48]. This model accurately predicted the metabolic fluxes and protein abundances in a wide range of growth conditions. Besides, RBA with the Python package (RBApy) has been used to generate calibrated RBA models for wild type and engineered CO_2-fixing *E. coli* [86]. Dynamic modelling efforts have

been made in a genome-scale metabolic reconstruction of the cyanobacterium *Synechococcus elongatus* PCC 7942 with 467 reactions of macromolecule synthesis [79]. The time-dependent model simulation revealed the dynamic resource allocation strategy: the timing and magnitude of macromolecular synthesis are delicately organised for maximal growth. The dynamic resource allocation model also uncovered that the cessation of metabolism at night (which is experimentally well known) results from the trade-off between the cost of unused enzymatic and ribosomal capacity during the dark period and the cost of additional storage before the dark period.

2.4.2.3 Whole-cell model

The ultimate target of a fine-grained RA model is possibly whole-cell modelling. One prominent effort is the construction of the comprehensive whole-cell model of *Mycoplasma genitalium* [73]. Twenty-eight functional submodules grouped by metabolic, RNA, protein and DNA categories were identified and modelled by the most suitable mathematical representation. The *M. genitalium* whole-cell model discloses previously unnoticed cellular behaviours such as *in vivo* rates of protein-DNA association. However, the inability to compute optimal phenotypic states (due to the discrete mathematical formalisms of subsystems) limits its practical application [50].

2.4.2.4 Advantages and limitations of fine-grained RA models

Given the extensive biological information embedded in fine-grained RA models, they are advantageous in providing a mechanistic explanation of molecular-level activities such as DNA replication, transcription, translation and interactions between subsystems. Fine-grained models are useful in fundamental studies and biological discovery, e.g. identifying the ‘core proteome’ that must be expressed to sustain cell growth [95] and to compute the metabolic cost of the production of virulence factors for plant pathogens [96]. Besides, detailed models offer the context for interpreting various omics datasets obtained from

genomics, transcriptomics, proteomics, metabolomics and fluxomics [97]. While powerful, complex models suffer from the time-consuming reconstruction process [8], risk of nonlinearity (discussed in Section 2.4.5.2) and lack of well-curated databases for retrieving high-quality model parameters, e.g. transcriptional, translational rate [98] and enzyme catalytic rate k_{cat}. Only 5% of the k_{cat} value for *S. cerevisiae* is available in BRENDA [77]. Future development of fine-grained models inevitably calls for the establishment of the genome-scale kinetic parameter library (the kinetome) [99].

2.4.3 Elementary flux modes-based RA models

Coarse-grained and fine-grained RA models solved by Flux Balance Analysis (FBA) [29] or dynamic FBA [100] are generally designed for predicting resource allocation through growth rate optimization. Another stoichiometric-based modelling framework, elementary (flux) mode analysis (EFMA) [101] studies resource allocation from a different angle. EFMA extracts the most 'essential' components of a metabolic network, termed the Elementary Flux Modes (EFMs). An EFM comprises a minimal set of enzymes that could operate a non-decomposable set of fluxes at steady state [101]. EFMs are mathematically defined pathways [102]. The ability to specify minimal enzyme sets intuitively links EFMs with minimal enzyme cost. One precursory work couples biomass and ATP producing EFMs with anabolic resource requirements for the enzyme sets to determine Pareto curves of investment and operating costs [36]. Through cost-benefit analysis, this model finds that the activation of the overflow pathways reflects the cell's preference of low investment cost strategies. Another modelling framework named Enzyme-Flux Cost Minimization (EFCM) studies the growth rate and biomass yield trade-offs by combining biomass-producing EFMs with enzyme costs calculated from kinetic models [81]. The EFCM model proposed that the rate-yield trade-off is not universal but particularly prominent under oxygen-limited conditions. Furthermore, a theory named minimal-EFM extremum

principle shows that when the growth rate is maximised, the optimal metabolic routes can be expressed by only a few EFMs [87]. The small number of active EFMs corresponds to the observed simple linear growth law, which elucidates that evolution might be driven at a direction of reduced metabolic complexity by allocating resources only to the most efficient enzymes [87]. The 'minimal' feature of EFM also possibly relates to the shortest enzymatic paths [103] and the kinetic theory of optimal pathway length [104].

2.4.3.1 Advantages and limitations of EFM-based RA models

Instead of fitting to experimental data to recapture the flux distribution and macromolecular compositions and maximal growth rates (as in coarse- and fine-grained RA models), EFM-based RA models provide biological insights of the resource allocation phenomena via the topology of the metabolic network. The minimal-EFM extremum principle provides direct mathematical proof that at least two growth-limiting constraints are needed (in addition to the law of mass conservation) for a model to predict a gradual switch from full respiration to a mixture of respiration and fermentation [87]. EFCM accounts for the kinetic effect of varying metabolite/substrate concentrations on the reaction rate, which is neglected in constraint-based models (e.g. RBA and ME) [81].

On the other hand, EFCM fails to predict the overflow metabolism, as there is no constraint imposed on fluxes [81]. Furthermore, EFCM does not model the details of protein production and catalytic rates of enzymes. The core of EFCM is still metabolism, integrated with empirical equations relating growth rate with enzyme cost and efficiency factors (ranging from 0 to 1) as simplified indicators of enzyme saturation. Furthermore, as biomass composition constitutes part of the network stoichiometry, the set of EMFs has to be re-enumerated if the biomass composition changes. Therefore, it is computationally infeasible to study varying biomass compositions in EFM-based models. Finally, due to the presence of numerous parallel and alternative pathways, it is currently computationally

infeasible to enumerate the complete set of EFMs for a genome-scale metabolic network [87].

2.4.4 Comparison between different RA models

2.4.4.1 Summary of general differences between three types of RA models

We have summarized three classes of models developed in the past decade that incorporate resource allocation, i.e. coarse-grained, fine-grained and EFM-based models. Coarse-grained models explicitly adopt the enzymatic costs as phenomenological constraints without using the detailed knowledge of transcriptional and translational processes. They are essentially carefully constrained metabolic models that implement the established phenomenological resource allocation constraints [2–5]. Fine-grained models comprise detailed descriptions of the expression of macromolecules in addition to the metabolic stoichiometry. The prediction of resource allocation is achieved by predicting macromolecular abundances (e.g. proteins, ribosomes, RNA, etc.) at maximal growth rate given substrate availability. While coarse-grained and fine-grained RA models are constraint-based models calibrated to recapitulate growth phenotypes, EFM-based models are not designed to predict phenotypic responses, but to provide biological insights of the observed resource allocation phenomena through deciphering the network topology (see Section 2.4.3).

2.4.4.2 Comparing RBA and ME in dealing with changing environments

Both RBA and ME are capable of predicting cellular configurations (including the cell's maximum growth rate λ_{max}, substrate uptake rate and by-product secretion rates, metabolic fluxes and the abundance of macromolecules) under changing growth environments in terms of the medium compositions or the concentrations of growth-controlling nutrients, but with several differences in their modelling approach.

In RBA [48], the prediction of cellular configurations in a specific growth condition (e.g. condition X) is achieved by a two-step procedure. The first step is the calibration of the apparent catalytic rate k_{Ei} using flux (i.e. growth rate, uptake and exertion fluxes) and proteomic datasets measured from cells grown under condition X. The second step is the growth simulation with a goal of growth rate maximization. In the first step of estimating k_{Ei}, for identifiability purposes the cost function is set to minimizing the deviation between the predicted and measured protein costs of biological processes instead of considering the costs of individual reactions. In addition, the nutrient uptake/secretion rates, yield coefficients for acetate and some nutrients and intracellular flux distributions are constrained to be equal or close to the measured values. Furthermore, a few flux ratio in glycolysis, TCA cycle and pentose phosphate pathways are constrained under different nutrient conditions, making k_{Ei} uniquely identifiable using dedicated flux-proteomic datasets. To run RBA growth simulations, the apparent catalytic rates of individual metabolic reactions are set to their inferred values (obtained in step one) and the extracellular nutrient concentrations are set to the known values of condition X. The output of an RBA growth simulation is the (predicted) λ_{max}, metabolic fluxes at λ_{max}, abundances of enzymes, transporters, chaperones and ribosomes for cells grown under condition X. If the large datasets (flux and proteomic data) required for k_{Ei} identification (in step one) of another growth condition of interest, e.g. condition Y, is not available, the RBA growth simulation for this new condition is enabled through linear regression of k_{Ei}. To do this, first we need to run several rounds of parameter identification (to obtain several sets of calibrated k_{Ei}) using available fluxes and proteomic datasets obtained from other growth conditions (with different measured growth rates). Then, the calibrated k_{Ei} sets are related to the measured growth rates via a linear function, e.g. $k_{Ei} = a_i + \lambda b_i$, where a_i and b_i are linear coefficients. With this (estimated) linear correlation, RBA growth

simulations for cells grown in condition Y can be performed by (a) setting the apparent catalytic rates to the values predicted by the linear equation (noted that λ of condition Y should have been measured) and (b) setting the extracellular nutrient concentrations to known values.

In the case of ME [50], two different types of growth simulation are considered: strictly nutrient-limited (SNL) simulation and proteome-limited simulation. In SNL growth simulations, metabolic enzymes are assumed to be operating below their maximal capacity, i.e. $k_{eff} < k_{cat}$, where k_{eff} is the effective catalytic rate (equivalent to the apparent catalytic rate k_{Ei} in RBA); k_{cat} is the maximal catalytic rate, which is a genuine constant and is set to be proportional to the enzyme's solvent accessible surface area (SASA). In SNL growth simulations, k_{eff} is a free variable and is predicted together with the growth rate, metabolic fluxes and macromolecular abundances. In proteome-limited growth simulations, metabolic enzymes are assumed to be operating at their maximal rate, i.e. $k_{eff} = k_{cat}$. For both SNL and proteome-limited growth simulations, the upper bound of the glucose (or another controlling nutrient) uptake rate is specified to the level that corresponds to a specific growth condition (as suggested by the Michaelis Menten uptake kinetics); it is the change in the specified maximum nutrient uptake rate that allows ME to predict distinct cellular configurations under different growth conditions.

The above analysis shows that both RBA and ME treat the catalytic rate of a metabolic enzyme (i.e. k_{Ei} or k_{eff}) as a growth condition-dependent variable, although the concrete modelling treatments are different.

2.4.5 Mathematical basis of RA models for predicting resource allocation

Here, we offer a comparison between coarse- and fine-grained RA models through their mathematical basis for predicting resource allocation (Table 2-2). EFM-based models are not intended to make an accurate prediction of cellular physiological states (e.g. maximal growth rate, metabolic fluxes, protein compositions), therefore are not included here.

2.4.5.1 Mathematical basis of coarse-grained RA models for predicting resource allocation

Coarse-grained models essentially solve an LP problem to maximize growth rate. The inputs can be cost coefficients [32] and/or nutrient states in the form of substrate uptake rate bound [35,40] or extracellular sugar concentration [75]. Key model outputs are the maximum growth rate and the corresponding metabolic flux distributions. Resource allocation is predicted in terms of the switch from a high-yield to a low-yield pathway at increased growth rates. The key mathematical mechanism of the prediction of the switch lies in the existence of (i) alternative pathways and (ii) a hard bound on the weighted sum of the fluxes, and that (iii) the enzyme cost ($\sum_i c_{i,E} v_{i,E}$) of the low-yield pathway is lower than the high-yield pathway. When the simulation encounters the proteome limitation, i.e. $\sum_i c_i v_i = A$, the pathway with a lower enzyme cost will be chosen for obtaining higher growth rates.

2.4.5.2 Mathematical basis of fine-grained RA models for predicting resource allocation

Growth simulation becomes more complex in fine-grained RA models as they often contain inherent nonlinearity, e.g. nonlinear dependency of metabolic fluxes on metabolite concentrations or nonlinear growth rate-dependent demand functions, dilution functions and catalytic rates [10]. Direct maximization of the growth rate will result in non-linear programming (NLP) problem, which is non-convex and leads to multiple local minimums

[105]. To overcome this issue, metabolite concentration-dependent nonlinearity is resolved by assuming abundant intermediate metabolite pools and quick reaction equilibrium (quasi-steady-state) and by splitting reversible reactions so that the metabolite-mediated kinetic effect can be ignored. The growth rate-mediated nonlinearity in RBA and ME was bypassed by converting the NLP growth rate optimization into an LP feasibility problem, i.e. fixing growth rate as input and search for the maximal growth rate (λ_{max}) that can make the problem still feasible [48,50]. In fact, RBA and ME are mathematically equivalent, except that the ME includes more comprehensive information of macromolecular expression, e.g. growth rate-dependent catalytic rates of ribosomes, RNAP, mRNA and tRNA [10].

Now come back to the central issue of this section, i.e. the mathematical basis of the prediction of resource allocation. In RBA and ME, resource allocation is reflected by the changes in both the metabolic fluxes and the macromolecular states. Below we demonstrate how these changes are predicted in ME.

In ME [50], the growth rate is intrinsically limited by the finite molecular efficiencies, e.g. limited catalytic rates of ribosomes, RNAP, and other RNA- or protein-based molecular catalysis. Even if the nutrients are unlimited, the cell will ultimately reach a maximum growth rate (λ_{max}) where the RNA and protein machinery of the mother cell cannot produce enough macromolecular components needed for the daughter cell. In this sense, RNA and protein machinery constitutively defines λ_{max}. As mentioned in Section 2.4.4.2, ME comprises two different types of growth simulations, namely the Strictly Nutrient-Limited (SNL) simulation (where $k_{eff} < k_{cat}$) and the proteome-limited simulation (where $k_{eff} = k_{cat}$). For most growth simulations, they are first solved via SNL. If no feasible solution is found, they will be solved as proteome-limited scenarios. Correspondingly, the predicted growth curve (i.e. the maximum growth rate against the

glucose uptake rate bound profile) shows three distinct regions, i.e. nutrient-limited region, a transitional region where the growth is both nutrient- and proteome-limited, and a proteome-limited region. The metabolic shift towards the acetate production occurs upon the proteome limitation, i.e. it starts at the transitional region. As for the gene expression patterns, ME predicts a decrease in the TCA cycle and oxidative phosphorylation expression followed by a large increase in glycolysis and acetate secretion expression upon the proteome limitation.

In ME, the predicted metabolic switch and changes in the gene expression do not seem to be caused by a single constraint. It was stated [50] that in the nutrient-limited region (corresponding to SNL simulation), the growth rate increases through changes in the effective catalytic rate of metabolic enzymes; in the proteome-limited region (corresponding to proteome-limited simulation), the cell increases its growth rate through differential expression of pathways (where pathways with less protein cost are preferred). Together with the different assumptions on the apparent catalytic rate (k_{eff}) made by SNL and proteome-limited simulations, the mathematical basis that leads to the prediction of the metabolic switch and differential gene expression in ME appears to include (i) limited catalytic efficiencies (k_{eff}) of metabolic enzymes and (ii) limited protein production capacity, which corresponds to limited translation and transcription rates.

In RBA, the growth rate is intrinsically limited by the translation efficiency, maximum cytosolic density, the membrane occupancy and limited synthesis capacity for ribosomes and proteins [15,48]. Although RBA [86] did not put much effort into predicting the varying acetate production rates and associated changes in the proteome at different growth rates in *E. coli*, it should be able to make such predictions due to its similar mathematical formulation of ME.

Finally, we would like to briefly comment on the uniqueness of the predicted metabolic fluxes in ME. The optimal metabolic flux distribution obtained in a genome-scale metabolic model is often non-unique, i.e. the maximal growth rate can be achieved by different flux distributions [106]. Coarse-grained RA model reduces the degree of freedom by incorporating growth-limiting phenomenological constraints. The ME model reduces the non-uniqueness by the inherent protein biosynthetic costs, which is analogous to the sequential optimization used in parsimonious enzyme usage FBA (pFBA [107]), where the growth rate is first maximised followed by minimization of summed metabolic fluxes. Nevertheless, the uniqueness of the solution does not translate to correctness: major deviation in central carbon fluxes exists between ME predictions and fluxomics. In aerobic batch cultures, ME predicts that pyruvate is decomposed by pyruvate formate lyase (Pfl) whereas the fluxomics suggests the use of pyruvate dehydrogenase (Pdh) [50]. Pfl is known to be active only under anaerobic conditions. It is possible to rectify this deviation by incorporating condition-specific regulations.

Table 2-2 Mathematical basis for predicting the resource allocation (RA) in coarse- and fined-grained RA models.

	Coarse-grained RA model	**Fine-grained RA model**
Optimization Problem	LP optimization, maximize λ.	LP feasibility problem: Search for maximum λ so that the problem is still feasible.
Form of RA prediction	Switch from high-yield to low-yield pathways at increased growth rates.	Growth rate-dependent change of macromolecular compositions.
Key model predictions	• λ_{max}. • Metabolic fluxes.	• λ_{max}. • Metabolic fluxes. • Macromolecular abundances, e.g. protein, RNA and ribosome concentrations.
Critical model inputs	• Enzyme cost coefficients. • Glucose uptake bound (Energy metabolism model [44], FBA^{ME} [11]) or extracellular glucose concentration (CAFBA [77]). • Distribution of crowding coefficients with average value fitted to the experimentally determined growth rate (FBAwMC [37]).	• Dialling in λ to check if RBA (λ) or ME (λ) is feasible. • Substrate uptake bound (for ME) or extracellular nutrient concentrations (for RBA). • Catalytic rates, either constants or growth-rate dependent free variables written as extra rows of constrains, e.g. catalytic rate of metabolic enzymes, translational rate, transcriptional efficiency, chaperone efficiency.
Mathematical mechanisms of RA prediction	• Weighted flux summation $\sum_i c_i v_i \leq A$. • Alternative pathways with different $\sum_i c_{i,E} v_{i,E}$. • When $\sum_i c_i v_i = A$, pathway with lower $\sum_i c_{i,E} v_{i,E}$ will be chosen for higher λ.	• The growth rate is intrinsically limited by transcriptional and translational efficiencies. • The apparent catalytic rates cannot exceed maximal values, i.e. $k_{eff} \leq k_{cat}$. • Daughter cells need to maintain constant concentrations of macromolecular components. • When k_{eff} and machinery efficiencies are saturated, the growth rate is increased by using pathways with fewer proteins to save resources for RNA and proteins required for rapid growth.

2.4.6 Summary of key advantages and limitations of different RA models

Coarse-grained RA models are essentially further-constrained metabolic models using phenomenological RA constraints. Model parameters derived from these phenomenological constraints generally hold clear biological meanings. They can be directly retrieved from protein databases like KEGG [88], BRENDA [89] and Uniprot [90]. Alternatively, they can be estimated using experimental data. The simple structure and straightforward parameterisation make coarse-grained models easy to construct and efficient in resolving engineering problems. In particular, as the role of resource allocation in regulating growth and metabolic configurations in *E. coli* has been well validated, a genome-scale coarse-grained RA model of *E. coli* can be used as a null model to investigate bioengineering questions by manipulating the energy parameters, RA constraints and metabolic pathways. For example, the impact of recombinant protein productions on host physiologies can be investigated by adding new biosynthesis pathways and modifying the proteomic sectors in the RA constraint. Besides, potential gene knockouts or overexpression to improve the protein production can be explored by the gene-protein-reaction relationship. Furthermore, the ability to capture the by-product secretions at different growth conditions allows coarse-grained RA models to study the effect of metabolite exchanges on the systematic dynamics of microbial communities. In general, coarse-grained models are advantageous in studying well-defined systems where cell growth and metabolism are of the most interests. On the other hand, the accuracy of coarse-grained RA models greatly relies on the validity of the RA constraints, e.g. models with growth rate-driven RA constraints would not be able to describe cells under strong burdens or environmental stresses, where maximal growth is often not the primary objective. The lack of expression information also limits the coarse-grained models in

describing transcriptional- and translational-level processes, e.g. to predict ribosomal abundances.

The essential difference between a fine-grained and a coarse-grained RA model is the inclusion of the expression matrix that comprises reactions describing the transcriptional and translational processes in a mechanistically detailed manner. The enhanced mechanistic feature of fine-grained models makes them advantageous in fundamental studies and biological discoveries (examples see Section 2.4.2.4) and the understanding of molecular-level activities such as protein folding, mRNA degradation and rRNA formation. However, the reconstruction process of fine-grained RA models is extremely labour-intensive and time-consuming compared with coarse-grained models. ME models have been available only for *E. coli* and *T. maritima* so far. Besides, the presence of non-linear growth rate-dependent catalytic rates in ME and RBA make direct growth maximisation being a non-convex nonlinear programming problem, which calls for alternative optimisation procedures (see Section 2.4.5 for more discussion).

Finally, EFM-based RA models take the advantage of elementary flux analysis for decomposing a metabolic network or a metabolic configuration to biomass-producing elementary flux modes (EFMs), which comprises minimal sets of enzymes carrying steady-state fluxes that convert substrates to biomass. By overlaying biomass-producing EFMs with enzyme information, e.g. molecular weight, amino acid sequence, EFM-based models were able to associate the occurrence of carbon-spilling overflow metabolism with the cell's preference of low elemental requirements to construct a functional metabolic pathway [36] and growth rate-yield tradeoffs [81]. Furthermore, by deciphering the observed optimal metabolic flux distributions into biomass-producing EFMs combined with resource allocation, it has been mathematically proved that the number of active metabolic pathways (i.e. EFMs) is at most equal to the number of biophysical constraints

[87]. For models intended to describe normal and overflow growth states of a cell (corresponding to two active EFMs), at least two biophysical constraints are required. In brief, EFM-based RA models are useful for providing a posteriori explanations of phenotypic patterns, showing how mathematically the observed phenomena are consistent with the topology of the metabolic network. Nevertheless, one major limitation of EFM-based models is that it is still computationally expensive to study a genome-scale metabolic network with variable biomass compositions or other variable traits that can change the network stoichiometry.

2.5 Other considerations and perspectives

2.5.1 Cost, benefit and trade-offs

The study in the *lac* operon model system [108] reveals that the activity of *lac* permease LacY, not the production or misfolding of the primary protein LacZ, accounts for the major physiological burden (defined as relative reduction in growth rate due to operon expression) to the cell. Previous resource allocation models (summarised in Section 2.4) generally associate the protein cost with the protein synthesis process. The counter-intuitive result found in *lac* operon reminds us that future effort on resource allocation, in particular the modelling of 'protein costs', needs to take more caution to avoid ambiguous or even misleading interpretation of the biological nature. In parallel to the cost, the benefit for protein expression has been defined as increased growth rate [108], increased ATP production rate per protein [4], increased energy efficiency (growth rate divided by ATP production rate) [51], increased product yield [109] and increased adaptability to the changing environments [21,23–25]. Different sets of cost and benefit can lead to different trade-offs, such as those between pathway protein cost and yield [36,110], growth rate and yield [81,84] and unused enzymatic and ribosomal capacity and additional storage [79].

These trade-offs may affect growth rates through distinct mechanisms and thus need to be treated with caution.

2.5.2 Bacterial growth in rich media

Most of the resource allocation studies focus on changes in the phenotypic patterns of cells grown in minimal media. Recent progress in studying batch growth of *E. coli* in undefined rich medium reveals the critical role of amino acid catabolism in regulating the central carbon metabolism (e.g. inhibiting glucose uptake and increasing acetate overflow) for faster growth [111]. The intricate trade-off between decomposing expensive resources (e.g. methionine) and the potential benefit of gaining various carbon and nitrogen sources needed for growth can offer further insights that support extending the resource allocation theory (which was established for cells grown on minimal media) to systems with complex nutritional environments.

2.5.3 Limitation of protein cost-derived constraints in RA models

The key advance of metabolic models incorporating RA (compared with classic metabolic models) is the inclusion of protein costs on top of the metabolic network. Taking the ME model as an example, pathways with higher metabolic efficiency (high-yield) are usually 'longer' than less efficient (low-yield) pathways. Therefore, high-yield pathways are always coupled with higher protein costs. When the growth rate is maximised, this feature (i.e. high-yield-high-cost) governs the switch from high-yield to low-yield pathways at increased growth rate. Although protein-cost constraint has proven to enable more accurate predictions of cellular configurations, this effectiveness can be easily spoiled if two alternative pathways hold similar protein costs. For example, pyruvate dehydrogenase (Pdh) and pyruvate formate lyase (Pfl) both convert pyruvate into acetyl-CoA with a cost of one enzyme. ME predicts the use of Pfl instead of Pdh as shown in fluxomics [50]. This outlier can be rectified by regulatory rules, i.e. Pfl is activated under anoxic conditions

[112,113]. It is worth noting that the regulatory network is not in the scope of ME [114] and to our knowledge nor has it been systematically modelled in other resource allocation models. It would be valuable to construct a model that comprises both metabolic, expression machinery and regulatory networks, paving the way towards predictive whole-cell modelling.

2.5.4 Rationale for the scope of this PhD project

From this comprehensive review, we can see that in 2016, just after Basan *et al.* presented their proteome allocation concept to explain the overflow metabolism in *E. coli* [4], there was a lack of constrained-based metabolic models that incorporated this latest global regulatory rule as extra phenomenological constraints. Therefore, we set our first objective (i.e. to predict the metabolic shift in wild type *E. coli*) and established the fundamental modelling framework of this PhD project, i.e. Chapter 3. Based on that, from an engineering point of view, we became interested in how the proteome allocation principle can guide the flux prediction for recombinant strains. Again, at that time (~year 2017) we did not find published models that could resolve this specific issue. We thus set the second goal and developed the proteome allocation model for batch cultures of recombinant *E. coli* (Chapter 4). Having established the first and the second model, we felt that it might be helpful to take a step back to think about the generic structure of a resource allocation metabolic model. Inspired by a previous partition of the complete proteome [75], we included the carbon transport sector to the proteome allocation constraint. Interestingly, from rigorous mathematical derivation we found a way of linking the Monod growth kinetics with the proteome allocation and the metabolic shift, which offers a possible resolution to the long-lasting speculation over the observed huge variation of the supposed constant Monod parameters (Chapter 5). In addition to the theoretical exploration, the last research objective was set for the investigation of the practical importance and potential

limitations of the established theories and modelling framework using further case studies (Chapter 6). Overall, the key objectives of this PhD project were set based on the research gaps present at the time and our specific research interests and were arranged to cover both in-depth theoretical explorations and the perspective of practical applications.

References

1. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics**. *Mol Syst Biol* 2009, **5**.
2. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T: **Interdependence of cell growth and gene expression: origins and consequences**. *Science* 2010, **330**:1099–1102.
3. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang Y-P, Lenz P, Yan D, Hwa T: **Coordination of bacterial proteome with metabolism by cyclic AMP signalling**. *Nature* 2013, **500**:301–306.
4. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
5. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR: **Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria**. *Mol Syst Biol* 2015, **11**.
6. Erickson DW, Schink SJ, Patsalo V, Williamson JR, Gerland U, Hwa T: **A global resource allocation strategy governs growth transition kinetics of Escherichia coli**. *Nature* 2017, **551**:119.
7. Goelzer A, Fromion V: **Resource allocation in living organisms**. *Biochem Soc Trans* 2017, **45**:945-952.
8. Yang L, Yurkovich JT, King ZA, Palsson BO: **Modeling the multi-scale mechanisms of macromolecular resource allocation**. *Curr Opin Microbiol* 2018, **45**:8–15.
9. Basan M: **Resource allocation and metabolism: the search for governing principles**. *Curr Opin Microbiol* 2018, **45**:77–83.
10. de Groot DH, Lischke J, Muolo R, Planqué R, Bruggeman FJ, Teusink B: **The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints**. *Cell Mol Life Sci* 2019, doi:10.1007/s00018-019-03380-2.
11. Milo R, Jorgensen P, Moran U, Weber G, Springer M: **BioNumbers—the database of key numbers in molecular and cell biology**. *Nucleic Acids Res* 2009, **38**:D750–D753.
12. Folsom JP, Carlson RP: **Physiological, biomass elemental composition and proteomic analyses of Escherichia coli ammonium-limited chemostat growth, and comparison with iron-and glucose-limited chemostat growth**. *Microbiology* 2015, **161**:1659–1670.
13. Gosset G: **Improvement of Escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system**. *Microb Cell Fact* 2005, **4**:14.
14. Rosano GL, Ceccarelli EA: **Recombinant protein expression in Escherichia coli: advances and challenges**. *Front Microbiol* 2014, **5**:172.
15. Goelzer A, Fromion V: **Bacterial growth rate reflects a bottleneck in resource allocation**. *Biochim Biophys Acta (BBA)-General Subj* 2011, **1810**:978–988.
16. Portnoy VA, Bezdan D, Zengler K: **Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering**. *Curr Opin Biotechnol* 2011, **22**:590–594.
17. Towbin BD, Korem Y, Bren A, Doron S, Sorek R, Alon U: **Optimality and sub-optimality in a bacterial growth law**. *Nat Commun* 2017, **8**:14123.

18. Radzikowski JL, Vedelaar S, Siegel D, Ortega ÁD, Schmidt A, Heinemann M: **Bacterial persistence is an active σS stress response to metabolic flux limitation**. *Mol Syst Biol* 2016, **12**:882.
19. Yang JH, Bening SC, Collins JJ: **Antibiotic efficacy—context matters**. *Curr Opin Microbiol* 2017, **39**:73–80.
20. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M: **The quantitative and condition-dependent Escherichia coli proteome**. *Nat Biotechnol* 2016, **34**:104.
21. O'Brien EJ, Utrilla J, Palsson BO: **Quantification and classification of E. coli proteome utilization and unused protein costs across environments**. *PLOS Comput Biol* 2016, **12**:e1004998.
22. Li SH-J, Li Z, Park JO, King CG, Rabinowitz JD, Wingreen NS, Gitai Z: **Escherichia coli translation strategies differ across carbon, nitrogen and phosphorus limitation conditions**. *Nat Microbiol* 2018, **3**:939–947.
23. Mori M, Schink S, Erickson DW, Gerland U, Hwa T: **Quantifying the benefit of a proteome reserve in fluctuating environments**. *Nat Commun* 2017, **8**:1225.
24. Korem Kohanim Y, Levi D, Jona G, Towbin BD, Bren A, Alon U: **A bacterial growth law out of steady state**. *Cell Rep* 2018, **23**:2891–2900.
25. New AM, Cerulus B, Govers SK, Perez-Samper G, Zhu B, Boogmans S, Xavier JB, Verstrepen KJ: **Different levels of catabolite repression optimize growth in stable and variable environments**. *PLOS Biol* 2014, **12**:e1001764.
26. Wang X, Xia K, Yang X, Tang C: **Growth strategy of microbes on mixed carbon sources**. *Nat Commun* 2019, **10**:1279.
27. Holms H: **Flux analysis and control of the central metabolic pathways in Escherichia coli.** *FEMS Microbiol Rev* 1996, **19**:85–116.
28. Majewski RA, Domach MM: **Simple constrained-optimization view of acetate overflow in E. coli**. *Biotechnol Bioeng* 1990, **35**:732–738.
29. Varma A, Palsson BØ: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Appl Environ Microbiol* 1994, **60**:3724–3731.
30. Wolfe AJ: **The acetate switch**. *Microbiol Mol Biol Rev* 2005, **69**:12–50.
31. Vemuri GN, Altman E, Sangurdekar DP, Khodursky AB, Eiteman MA: **Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio**. *Appl Env Microbiol* 2006, **72**.
32. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN: **Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity**. *P Natl Acad Sci USA* 2007, **104**.
33. Vazquez A, Beg QK, Demenezes MA, Ernst J, Bar-Joseph Z, Barabasi AL, Boros LG, Oltvai ZN: **Impact of the solvent capacity constraint on E. coli metabolism**. *BMC Syst Biol* 2008, **2**.
34. Szenk M, Dill KA, de Graff AMR: **Why do fast-growing bacteria enter overflow metabolism? testing the membrane real estate hypothesis**. *Cell Syst* 2017, **5**:95–104.
35. Zhuang K, Vemuri GN, Mahadevan R: **Economics of membrane occupancy and respiro-fermentation**. *Mol Syst Biol* 2011, **7**.
36. Carlson RP: **Metabolic systems cost-benefit analysis for interpreting network structure and regulation**. *Bioinformatics* 2007, **23**.
37. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R: **Proteome reallocation in Escherichia coli with increasing specific growth rate**. *Mol BioSyst* 2015, **11**:1184–1193.
38. De Deken RH: **The Crabtree effect: a regulatory system in yeast**. *Microbiology* 1966, **44**:149–156.
39. Yu R, Nielsen J: **Big data in yeast systems biology**. *FEMS Yeast Res* 2019, **19**.
40. Chen Y, Nielsen J: **Energy metabolism controls phenotypes by protein efficiency and allocation**. *Proc Natl Acad Sci* 2019, **116**:17592–17597.
41. Vander Heiden MG, Cantley LC, Thompson CB: **Understanding the Warburg effect: the metabolic requirements of cell proliferation**. *Science* 2009, **324**:1029–1033.

42. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E: **Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect**. *PLoS Comput Biol* 2011, **7**:e1002018.
43. de Alteriis E, Cartenì F, Parascandola P, Serpa J, Mazzoleni S: **Revisiting the Crabtree/Warburg effect in a dynamic perspective: a fitness advantage against sugar-induced cell death**. *Cell Cycle* 2018, **17**:688–701.
44. Vazquez A, Oltvai ZN: **Macromolecular crowding explains overflow metabolism in cells**. *Sci Rep* 2016, **6**:31007.
45. Zhou Y, Vazquez A, Wise A, Warita T, Warita K, Bar-Joseph Z, Oltvai ZN: **Carbon catabolite repression correlates with the maintenance of near invariant molecular crowding in proliferating E. coli cells**. *BMC Syst Biol* 2013, **7**:138.
46. Woldringh CL, Binnerts JS, Mans A: **Variation in Escherichia coli buoyant density measured in Percoll gradients.** *J Bacteriol* 1981, **148**:58–63.
47. Basan M, Zhu M, Dai X, Warren M, Sévin D, Wang Y-P, Hwa T: **Inflating bacterial cells by increased protein synthesis**. *Mol Syst Biol* 2015, **11**:836.
48. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, Mariadassou M, Aymerich S, Hecker M, Noirot P, et al.: **Quantitative prediction of genome-wide resource allocation in bacteria**. *Metab Eng* 2015, **32**:232–243.
49. Thiele I, Fleming RMT, Que R, Bordbar A, Diep D, Palsson BO: **Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage**. *PLoS One* 2012, **7**:e45635.
50. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**:693–693.
51. Maitra A, Dill KA: **Bacterial growth laws reflect the evolutionary importance of energy efficiency**. *Proc Natl Acad Sci* 2015, **112**:406–411.
52. Chen T, He HL, Church GM: **Modeling gene expression with differential equations**. *Proc. Pac. Symp. Biocomuting (PSB'99)* 1998, **4**:29–40.
53. Tchourine K, Poultney CS, Wang L, Silva GM, Manohar S, Mueller CL, Bonneau R, Vogel C: **One third of dynamic protein expression profiles can be predicted by a simple rate equation**. *Mol BioSyst* 2014, **10**:2850–2862.
54. Maurizi MR: **Proteases and protein degradation in Escherichia coli**. *Experientia* 1992, **48**:178–201.
55. Dressaire C, Gitton C, Loubière P, Monnet V, Queinnec I, Cocaign-Bousquet M: **Transcriptome and proteome exploration to model translation efficiency and protein stability in Lactococcus lactis**. *PLOS Comput Biol* 2009, **5**:1–12.
56. Kochanowski K, Volkmer B, Gerosa L, Haverkorn van Rijsewijk BR, Schmidt A, Heinemann M: **Functioning of a metabolic flux sensor in Escherichia coli&lt**. *Proc Natl Acad Sci* 2013, **110**:1130 LP – 1135.
57. Kochanowski K, Gerosa L, Brunner SF, Christodoulou D, Nikolaev Y V, Sauer U: **Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli**. *Mol Syst Biol* 2017, **13**:903.
58. Doucette CD, Schwab DJ, Wingreen NS, Rabinowitz JD: **α-ketoglutarate coordinates carbon and nitrogen utilization via enzyme I inhibition**. *Nat Chem Biol* 2011, **7**:894–901.
59. Morin M, Ropers D, Letisse F, Laguerre S, Portais J-C, Cocaign-Bousquet M, Enjalbert B: **The post-transcriptional regulatory system CSR controls the balance of metabolic pools in upper glycolysis of Escherichia coli**. *Mol Microbiol* 2016, **100**:686–700.
60. Kadner RJ, Murphy GP, Stephens CM: **Two mechanisms for growth inhibition by elevated transport of sugar phosphates in Escherichia coli**. *Microbiology* 1992, **138**:2007–2014.
61. Morita T, Maki K, Aiba H: **RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs**. *Genes Dev* 2005, **19**:2176–2186.

62. Díaz-Ruiz R, Avéret N, Araiza D, Pinson B, Uribe-Carvajal S, Devin A, Rigoulet M: **Mitochondrial oxidative phosphorylation is regulated by fructose 1,6-bisphosphate: a possible role in Crabtree effect induction?**. *J Biol Chem* 2008, **283**:26948–26955.
63. Shimizu K, Matsuoka Y: **Regulation of glycolytic flux and overflow metabolism depending on the source of energy generation for energy demand**. *Biotechnol Adv* 2019, **37**:284–305.
64. Rodionova IA, Zhang Z, Mehla J, Goodacre N, Babu M, Emili A, Uetz P, Saier MH: **The phosphocarrier protein HPr of the bacterial phosphotransferase system globally regulates energy metabolism by directly interacting with multiple enzymes in Escherichia coli**. *J Biol Chem* 2017, **292**:14250–14257.
65. Wendrich TM, Blaha G, Wilson DN, Marahiel MA, Nierhaus KH: **Dissection of the mechanism for the stringent factor RelA**. *Mol Cell* 2002, **10**:779–788.
66. Bosdriesz E, Molenaar D, Teusink B, Bruggeman FJ: **How fast-growing bacteria robustly tune their ribosome concentration to approximate growth-rate maximization**. *FEBS J* 2015, **282**:2029–2044.
67. Marr AG: **Growth rate of Escherichia coli.** *Microbiol Mol Biol Rev* 1991, **55**:316–333.
68. Bremer H, Dennis P: **Feedback control of ribosome function in Escherichia coli**. *Biochimie* 2008, **90**:493–499.
69. Venturelli OS, Tei M, Bauer S, Chan LJG, Petzold CJ, Arkin AP: **Programming mRNA decay to modulate synthetic circuit resource allocation**. *Nat Commun* 2017, **8**:15128.
70. Segall-Shapiro TH, Meyer AJ, Ellington AD, Sontag ED, Voigt CA: **A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase**. *Mol Syst Biol* 2014, **10**:742.
71. Agapakis CM, Boyle PM, Silver PA: **Natural strategies for the spatial optimization of metabolism in synthetic biology**. *Nat Chem Biol* 2012, **8**:527–535.
72. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, et al.: **In silico method for modelling metabolism and gene product expression at genome scale**. *Nat Commun* 2012, **3**:929.
73. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**:389–401.
74. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM: **Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale**. *BMC Syst Biol* 2014, **8**:110.
75. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
76. Yang L, Yurkovich JT, Lloyd CJ, Ebrahim A, Saunders MA, Palsson BO: **Principles of proteome allocation are revealed using proteomic data and genome-scale models**. *Sci Rep* 2016, **6**:36734.
77. Sánchez, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J: **Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints**. *Mol Syst Biol* 2017, **13**:935.
78. Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO: **Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation**. *Proc Natl Acad Sci* 2017, **114**:11548–11553.
79. Reimers A-M, Knoop H, Bockmayr A, Steuer R: **Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth**. *Proc Natl Acad Sci* 2017, **114**:E6457--E6465.
80. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO: **COBRAme: A computational framework for genome-scale models of metabolism and gene expression**. *PLOS Comput Biol* 2018, **14**:e1006302.
81. Wortel MT, Noor E, Ferris M, Bruggeman FJ, Liebermeister W: **Metabolic enzyme cost explains variable trade-offs between microbial growth rate and yield**. *PLOS Comput Biol* 2018, **14**:1–21.

82. Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, et al.: **Cellular responses to reactive oxygen species are predicted from molecular mechanisms**. *Proc Natl Acad Sci* 2019, **116**:14368–14373.
83. Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO: **DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression**. *BMC Syst Biol* 2019, **13**:2.
84. Cheng C, O'Brien EJ, McCloskey D, Utrilla J, Olson C, LaCroix RA, Sandberg TE, Feist AM, Palsson BO, King ZA: **Laboratory evolution reveals a two-dimensional rate-yield tradeoff in microbial metabolism**. *PLOS Comput Biol* 2019, **15**:e1007066.
85. Liu JK, Lloyd C, Al-Bassam MM, Ebrahim A, Kim J-N, Olson C, Aksenov A, Dorrestein P, Zengler K: **Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen**. *PLOS Comput Biol* 2019, **15**:e1006848.
86. Bulović A, Fischer S, Dinh M, Golib F, Liebermeister W, Poirier C, Tournier L, Klipp E, Fromion V, Goelzer A: **Automated generation of bacterial resource allocation models**. *Metab Eng* 2019, **55**:12–22.
87. de Groot DH, van Boxtel C, Planqué R, Bruggeman FJ, Teusink B: **The number of active metabolic pathways is bounded by the number of cellular constraints at maximal metabolic rates**. *PLOS Comput Biol* 2019, **15**:e1006858.
88. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**:27–30.
89. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, et al.: **BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA**. *Nucleic Acids Res* 2012, **41**:D764–D772.
90. The UniProt Consortium: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2016, **45**:D158–D169.
91. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
92. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011**. *Mol Syst Biol* 2011, **7**:535.
93. Thiele I, Jamshidi N, Fleming RMT, Palsson BO: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**.
94. Goelzer A, Fromion V, Scorletti G: **Cell design in bacteria as a convex optimization problem**. *Automatica* 2011, **47**:1210–1218.
95. Yang L, Tan J, O\textquoterightBrien EJ, Monk JM, Kim D, Li HJ, Charusanti P, Ebrahim A, Lloyd CJ, Yurkovich JT, et al.: **Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data**. *Proc Natl Acad Sci* 2015, **112**:10810–10815.
96. Peyraud R, Cottret L, Marmiesse L, Gouzy J, Genin S: **A resource allocation trade-off between virulence and proliferation drives metabolic versatility in the plant pathogen ralstonia solanacearum**. *PLOS Pathog* 2016, **12**:e1005939.
97. Kampf C, Mardinoglu A, Fagerberg L, Hallström BM, Edlund K, Lundberg E, Pontén F, Nielsen J, Uhlen M: **The human liver-specific proteome defined by transcriptomics and antibody-based profiling**. *FASEB J* 2014, **28**:2901–2914.
98. Landon S, Rees-Garbutt J, Marucci L, Grierson C: **Genome-driven cell engineering review: in vivo and in silico metabolic and genome engineering**. *Essays Biochem* 2019, **63**:267–284.
99. Nilsson A, Nielsen J, Palsson BO: **Metabolic models of protein allocation call for the kinetome**. *Cell Syst* 2017, **5**:538–541.

100. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**:1331–1340.
101. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks**. *Nat Biotechnol* 2000, **18**.
102. Wortel MT, Peters H, Hulshof J, Teusink B, Bruggeman FJ: **Metabolic states with maximal specific rate carry flux through an elementary flux mode**. *FEBS J* 2014, **281**:1547–1555.
103. Noor E, Eden E, Milo R, Alon U: **Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy**. *Mol Cell* 2010, **39**:809–820.
104. Heinrich R, Schuster S, Holzhütter H-G: **Mathematical analysis of enzymic reaction systems using optimization principles**. *EJB Reviews 1991*. Springer, Berlin, Heidelberg; 1991:167–187.
105. Dinh M, Fromion V: **RBA like problem with thermo-kinetics is non convex**. *ArXiV 2017*, https://arxiv.org/abs/1706.01312.
106. Gudmundsson S, Thiele I: **Computationally efficient flux variability analysis**. *BMC Bioinformatics* 2010, **11**:489.
107. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al.: **Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models**. *Mol Syst Biol* 2010, **6**:390.
108. Eames M, Kortemme T: **Cost-benefit tradeoffs in engineered lac operons**. *Science* 2012, **336**:911–915.
109. Zeng H, Yang A: **Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant Escherichia coli**. *Biotechnol Bioeng* 2019, **116**:1484-1495.
110. Mori M, Marinari E, De Martino A: **A yield-cost tradeoff governs Escherichia coli's decision between fermentation and respiration in carbon-limited growth**. *npj Syst Biol Appl* 2019, **5**:16.
111. Zampieri M, Hörl M, Hotz F, Müller NF, Sauer U: **Regulatory mechanisms underlying coordination of amino acid and glucose catabolism in Escherichia coli**. *Nat Commun* 2019, **10**:3354.
112. Orth JD, Palsson BØ, Fleming RMT: **Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide**. *EcoSal Plus* 2010, **4**.
113. Sawers G, Watson G: **A glycyl radical solution: oxygen-dependent interconversion of pyruvate formate-lyase**. *Mol Microbiol* 1998, **29**:945–954.
114. King ZA, O'Brien EJ, Feist AM, Palsson BO: **Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion**. *Metab Eng* 2017, **39**:220–227.

Chapter 3 Modelling overflow metabolism in *Escherichia coli* with flux balance analysis incorporating differential proteomic efficiencies of energy pathways

Summary

This chapter presents the following published article:

Zeng, H., Yang, A. Modelling overflow metabolism in *Escherichia coli* with flux balance analysis incorporating differential proteomic efficiencies of energy pathways. *BMC Syst Biol* **13,** 3 (2019). https://doi.org/10.1186/s12918-018-0677-4.

Among all authors contributed to the published article, H.Z. performed the modelling work, analysed and interpreted the results, and wrote the manuscript. A.Y. designed and supervised the study, analysed and interpreted the results, and revised the manuscript. The supplementary information of this chapter can be found in Appendix I.

As stated in the introduction chapter, the very original objective of this PhD project is to investigate strategies for engineering *E. coli* (the most well-studied microorganism) for efficient production of valuable chemicals through understanding its physiology at the metabolic level. *E. coli* tends to produce a significant amount of acetic acid (acetate) during fast growth under aerobic glucose conditions, also known as the overflow metabolism. The accumulation of acetate can severely inhibit cell growth and prevent the culture reaching a high cell density, making it a major obstacle for industrial-scale bioproduction using *E. coli*. Although this phenomenon has been acknowledged and studied for several decades, the underlying mechanism remains murky, but more recent studies shed light from the perspectives of cellular resource allocation. These include

particularly the work by Hwa and his co-workers in 2015[1], which proposed and validated that the overflow metabolism in *E. coli* was a result of an efficient proteome allocation strategy to coordinate energy biogenesis and biomass synthesis as required for rapid growth. Inspired by their work, this PhD project started from investigating the way to integrate the proteome allocation principle to metabolic models to improve flux predictions.

In this chapter, we adopted three different *E. coli* strains with two distinct growth and acetate overflow patterns. We first formulated a concise, pathway-level proteome allocation constraint that governs the competition of proteomic resources between respiration, fermentation and biomass synthesis pathways. We then presented a way of estimating proteomic cost parameters using bioreactor level experimental data, e.g. the acetate production rate against the growth rate profile. Besides, we adjusted the energy parameters based on strain-specific energetic data to improve the prediction of biomass yield. Furthermore, we investigated the impact of controlling the maximum ratio of carbon flow to the pentose phosphate pathway over the total carbon intake (i.e. the PP pathway ratio) on model parameters.

The pathway-level proteome constraint requires a very small number of model parameters needed for predicting the overflow metabolism. Two linear correlations were shown to exist between three proteomic cost parameters, which allowed us to calculate two unknown parameters by specifying the value of the third one. The proteomic cost of biomass synthesis processes was estimated to be higher for a slow-growing strain, in comparison with the other two fast-growing strains. The estimated proteomic cost of the fermentation pathway was lower than the respiration pathway for all the strains. The

[1] Basan, M., Hui, S., Okano, H. *et al.* Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature* **528,** 99–104 (2015). https://doi.org/10.1038/nature15765

absolute values of the proteomic cost parameters were shown to vary with the PP pathway ratio, but the linear trend between parameters was retained. Besides, simulated flux distributions suggested that the amount of carbon flow to the PP pathway in general increases with the growth rate, where several turning points were found and considered to be possibly related to the cellular demand on NADH and NADPH. Furthermore, the poor prediction of the biomass yield during acetate overflow was rectified by resetting the energy parameters based on strain-specific data from the literature.

Overall, Chapter 3 shows that the key to improving the flux predictions in metabolic models lies in not only encompassing the proteome allocation constraint, but also the control of PP pathway activity and the values of energy parameters. The prediction of the overflow metabolism for various wild-type *E. coli* validates the effectiveness of the pathway-level proteome allocation constraint and forms a solid basis of the rest of the research in this PhD. Further extensions of this modelling approach (e.g. to recombinant strains, for predicting carbon intake and for coculture systems) are presented in Chapters 4-6.

3.1 Background

The formation of acidic by-products, predominantly acetate, when *Escherichia coli* (*E. coli*) grows under aerobic-glucose conditions is a commonly observed phenomenon, which has been extensively studied over decades [1–5]. Lee reviewed 19 studies of recombinant *E. coli* where acetate was accumulated in fed-batch systems [6]. It has been reported that the portion of glucose converted into acetate can be as high as 15% [7], representing a seemingly huge waste of feedstock. The accumulation of acetate in the culture medium appears to be a major limiting factor for achieving high cell density [8], which is particularly severe in the growth of recombinant strains [9]. Acetate also impairs the microbial production of recombinant proteins [1] and drug precursors [9]. These complications of acetate in bioreactors thus call for the elucidation of acetate-pertinent metabolic processes. A similar phenomenon has been observed in tumour cells (i.e. the Warburg effect) [10–12] and associated mathematical models for explaining this effect have recently been reviewed [13].

Traditionally, the aerobic formation of acetate has been referred to as overflow metabolism: the excess glucose saturates or inhibits the tricarboxylic acid (TCA) cycle, which subsequently forces the cell to modulate the redundant carbon to the acetate pathway [3,14]. However, the study by Molenaar *et al.* suggested that the overflow metabolism as shown in the growth phenotype is probably a result of the global allocation of cellular resources, where the enzyme efficiency and the pathway yield were both taken into account to obtain the optimal growth strategies subject to different growth conditions [15]. Later in 2015, Basan *et al.* proposed and validated that the overflow metabolism in *E. coli* originates from the global physiological proteome allocation for rapid growth [16]. In particular, the proteomic efficiency of energy biogenesis through aerobic fermentation was found to be higher than that of respiration; this difference in proteomic efficiency

between fermentation and respiration appears to play a central role in dictating the degree of overflow metabolism in *E. coli*.

Given the importance of the overflow metabolism, several phenomenological models were developed to depict this effect [5,8,17,18], where the prediction of acetate excretion was dictated by a combination of (i) the constraints on oxygen and carbon supply and (ii) cellular mass and energy balance. Later, models that adopt conventional regulatory mechanisms of acetate metabolism [14], such as oxygen limitation, carbon source availability and tight regulation of cofactor pools were evaluated [19], with an attempt to explain the metabolic shifts from a fully aerobic mode to the aerobic acetate fermentation (overflow). More recently, constraint-based metabolic models [20] were established to analyse the optimal cellular growth strategy, incorporating principles of (i) limitation in the cellular resource on the maximal attainable growth rate, such as the maximum cytoplasmic density adopted by FBAwMC [21,22] and the finite amount of resource to be allocated between the metabolic network and ribosomes, as applied in RBA [23–25], (ii) metabolic regulation based on enzyme kinetic information, such as mechanistically detailed descriptions of gene expression and the synthesis of functional macromolecules used in ME-Model [26] and (iii) membrane occupancy-derived competition between glucose transporters and respiration chain (an extension of the ME-Model) [27]. The major target of these models is to predict the maximum growth rate. Predictions were validated quantitatively by the experimental data, while the overflow metabolism in fast-growing phase was mostly captured in a qualitative way. In addition, it was pointed out [16,28] that cell volumes were empirically found to vary widely with virtually constant densities across different growth conditions [29], which suggests that the cytoplasmic density-based constraint might not be fully justified.

Inspired by recent experimental work studying the proteomic cost of the core metabolic pathways of *E. coli* [16], a model named constrained allocation flux balance analysis (CAFBA) [28] managed to predict the rates of acetate production in the overflow metabolism for different *E. coli* strains, with a good quantitative agreement with experimental data. However, the proteomic costs adopted in CAFBA were applied to individual metabolic reactions, without focusing on the exploration of the critical role played by specific metabolic modules such as energy biogenesis pathways.

In this chapter, we attempt to depict the overflow metabolism in various *E. coli* strains with quantitative accuracy, i.e. predicting aerobic steady-state rates of acetate production at different growth rates and validating the model with experimental data from the literature. In particular, we adopt a concise proteome allocation constraint as identified by Basan *et al.* [16], referred to as the Proteome Allocation Theory (PAT) in this chapter. The PAT suggests that the choice of energy biogenesis pathways under different growth conditions results from the discrepancy of proteomic efficiencies between fermentation and respiration. *E.coli* cells tend to use the more protein-efficient fermentation pathway to generate energy to accommodate the high proteomic demand in biosynthesis under rapid growth. The key concepts of PAT are fully embedded and realised in our model. With a parsimonious, PAT-based metabolic model capable of accurately capturing the overflow metabolism, we further analyse the interdependency between pathway-level proteomic cost parameters, the disparity in these parameters between different *E. coli* strains, and the impact of cellular energy demand on the accuracy of the co-prediction of the overflow metabolism and the biomass yield on substrate.

3.2 Methods

3.2.1 Formulation of the PAT constraint

Following Basan *et al.* [16], the fractions of three proteome sectors in the entire proteome (i.e. the total protein content) of the cell sum to unity:

$$\phi_f + \phi_r + \phi_{BM} = 1 \qquad \text{Equation 3-1}$$

where ϕ_f and ϕ_r are the fractions of the fermentation- and respiration-affiliated enzymes, respectively, which enable the fluxes for energy generation; ϕ_{BM} represents the fraction of the remaining part of the proteome enabling other cellular activities, broadly referred to as the sector of biomass synthesis [16,28].

More specifically, ϕ_f represents the mass abundance of the enzymes that carry fermentation fluxes involved in glycolysis (glucose to acetyl-CoA), oxidative phosphorylation and acetate synthesis pathways (phosphotransacetylase and acetate kinase). ϕ_r comprises all the enzymes that catalyse the respiration-associated reactions in glycolysis, tricarboxylic acid (TCA) cycle and oxidative phosphorylation system. Same as in Basan *et al.* [16], in this chapter we adopt the linear dependences assumed in Hui *et al.* [30] to relate ϕ_f and ϕ_r with the fermentation and respiration fluxes respectively,

$$\phi_f = w_f v_f \qquad \text{Equation 3-2}$$

$$\phi_r = w_r v_r \qquad \text{Equation 3-3}$$

where v_f (v_r) is the fermentation (respiration) pathway flux, which in this chapter is represented by the enzymatic reaction acetate kinase ACKr (α-ketoglutarate dehydrogenase AKGDH); w_f (w_r) is the pathway-level proteomic cost, denoting the proteome fraction required per unit fermentation (respiration) flux.

On the biomass synthesis sector, ϕ_{BM} corresponds to the remaining part of the proteome that is not covered by the fermentation and respiration sectors, including ribosomal proteins and anabolic enzymes (the major part, referred to as biomass synthesis), catabolic enzymes and cellular maintenance proteins. Motivated by the observed linear dependency between growth rate and proteome fraction for biomass synthesis [30–32], the following linear relationship is assumed:

$$\phi_{BM} = \phi_0 + b\lambda \qquad \text{Equation 3-4}$$

where $b\lambda$ is the growth rate-associated component with λ being the specific growth rate and the constant b quantifying the proteome fraction required per unit growth rate. In Basan *et al*. [16], ϕ_0 was considered as a growth rate-independent constant.

Combining Equation 3-1 to Equation 3-4, we have

$$w_f v_f + w_r v_r + b\lambda = 1 - \phi_0 \qquad \text{Equation 3-5}$$

Equation 3-5 implies that the sum of the three proteomic cost terms on the left-hand side remains constant. However, when the growth rate (and hence the fermentation and respiration fluxes) becomes very low, it is difficult to envisage numerically how this sum could still remain at a constant level. In fact, in Basan *et al.* [16] (see its Supplementary Information), it was acknowledged that at growth rates lower than that corresponding to the onset of the overflow phenomenon, the proteome sectors would no longer be constrained by the equality indicated by Equation 3-5. This suggests that across the entire range of possible growth rates, ϕ_0 is unlikely a growth rate-independent constant: it may remain at a constant (and minimum) level in the overflow region where the proteomic resource is stretched, but become growth-rate dependent (and larger) at lower growth rates outside the overflow region, i.e. $\phi_{0,\min} \leq \phi_0 \leq 1$, where $\phi_{0,\min}$ is a true constant. Defining $\phi_{\max} \equiv 1 - \phi_{0,\min}$, Equation 3-5 then becomes

$$w_f v_f + w_r v_r + b\lambda = 1 - \phi_0 \leq 1 - \phi_{0,\min} \equiv \phi_{\max} \quad \text{Equation 3-6}$$

In Vazquez and Oltvai (2016) [33], ϕ_0 was also interpreted as a variable instead of a constant, with a (non-zero) minimum value. For simplicity, both sides of Equation 3-6 is divided by ϕ_{max}, leading to the final form of the proteome constraint adopted in this chapter, referred to as PAT constraint from this point on:

$$w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 \quad \text{Equation 3-7}$$

where $w_f^* \equiv w_f/\phi_{max}$, $w_r^* \equiv w_r/\phi_{max}$ and $b^* \equiv b/\phi_{max}$. w_f^*, w_r^* and b^* are referred to as the proteomic cost parameters.

3.2.2 Predicting the acetate flux

Flux balance analysis (FBA) [20] is used to determine the optimal flux distribution under different growth condition, with a set of constraints:

$$\begin{aligned} &\text{Maximise} \quad f_{obj} \\ &\text{Subject to} \quad \text{(i)} \; S\boldsymbol{v} = 0 \qquad (9) \\ &\qquad\qquad \text{(ii)} \; v^L \leq \boldsymbol{v} \leq v^U \\ &\qquad\qquad \text{(iii)} \; w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 \end{aligned} \quad \text{Equation 3-8}$$

where f_{obj} is the assumed cellular objective. We specified minimizing substrate uptake as the objective function because in this study the commonly used objective 'growth rate' was used as the model input (with acetate production as the model output). S is the stoichiometric matrix defined by the metabolic model; $\boldsymbol{v}$ is a column vector comprising the reactions/fluxes described in the metabolic network; v^L and v^U represent the lower and upper limits of the reactions, respectively. The inequality constraint (iii) corresponds to Equation 3-7 introduced earlier.

The prediction of the extent of overflow metabolism (rate of acetate production) requires the parameter values of w_f^*, w_r^* and b^* in the third constraint (PAT constraint) of Equation 3-8. We show in the next section that these three parameters cannot be uniquely determined by the experimentally measured growth rate-acetate production profile alone. A set of values for these parameters was randomly chosen from mathematically equivalent sets (AppxI-Table S1). The PAT-based FBA was run at different growth rates under aerobic-glucose conditions. FBA was carried out using the core *E. coli* metabolic model [34], referred to as the core model in the rest of this chapter. The optimal flux distribution was solved via COBRA toolbox [35] in MATLAB (R2016a). LP solution was determined by Gurobi 6.0. Detailed model descriptions such as uptake bounds and flux regulations are given in Appendix I.

3.2.3 Interdependency of proteomic cost parameters

A linear relationship was previously shown to hold between the fermentation (or respiration) flux and steady-state growth rates in the overflow region [16]:

$$v_f = k_f \lambda + v_{f,0} \qquad \text{Equation 3-9}$$

$$v_r = k_r \lambda + v_{r,0} \qquad \text{Equation 3-10}$$

where, as introduced earlier, v_f is the fermentation flux (referred to as 'acetate line' in [16]); v_r is the respiration flux. k_f (k_r) and $v_{f,0}$ ($v_{r,0}$) are constants representing the slope and intercept of the fermentation (respiration) line. Substituting Equation 3-9 and Equation 3-10 into Equation 3-7, with the equal sign held for the overflow condition:

$$\left(w_f^* k_f + w_r^* k_r + b^*\right)\lambda = 1 - w_f^* v_{f,0} - w_r^* v_{r,0} \qquad \text{Equation 3-11}$$

Equation 3-11 holds for any growth rate (λ) in the overflow region, which requires

$$w_f^* k_f + w_r^* k_r + b^* = 0 \qquad \text{Equation 3-12}$$

$$1 - w_f^* v_{f,0} - w_r^* v_{r,0} = 0 \qquad \text{Equation 3-13}$$

Equation 3-12 and Equation 3-13 indicate that (i) there is a linear relationship between the fermentation and respiration proteomic cost parameters w_r^* and w_f^*, and (ii) the third growth-rate dependent proteomic cost parameter b^* is a linear combination of w_r^* and w_f^*, thus b^* also possesses a linear relationship with w_f^* (or w_r^*). If experimental data exist that allow for both the fermentation line and the respiration line to be plotted (such as the steady-state growth rate – acetate excretion and growth rate – CO_2 revolution data given in [16]), their slopes and intercepts, appearing in Equation 3-12 and Equation 3-13, can be obtained. However, the three proteomic cost parameters cannot be uniquely determined by the two equations, although specific values of similar parameters have previously been derived from measured cellular protein compositions [16].

In this chapter, k_f and $v_{f,0}$ were directly determined from the experimentally measured growth rate-acetate excretion profile (data sources are shown in Figure 3-1). To our knowledge, no directly experimental data were available for the rate of intracellular respiration. Alternatively, we took the growth rate-acetate profile as the input of FBA (setting the objective function to the minimisation of glucose uptake) to estimate the respiration flux at each data point, which was subsequently used to determine k_r and $v_{r,0}$. Flux variability analysis (FVA) was conducted which confirmed that all the relevant fluxes used in the model were uniquely determined. After obtaining k_f (k_r) and $v_{f,0}$ ($v_{r,0}$), a set of values of w_f^*, w_r^* and b^* can be determined by arbitrarily specifying the value for one of the parameters. In this chapter, we took w_f^* to be specified, in a range of [0, 0.11] for MG1655 and [0, 0.07] for ML308 and NCM3722 due to scaling purposes, which allows us to present the parameter estimation results in the form of w_f^*-w_r^* and w_f^*-b^* plots and compare the results across different *E. coli* strains. Simulation results presented in this

chapter were obtained with a randomly chosen value of w_f^* within the ranges mentioned above and the correspondingly determined values of w_r^* and b^*. Note that different values chosen for w_f^* yielded identical simulation results (AppxI-Figure S1).

3.2.4 Adjustment of cellular energy demand

We observed discrepancy in biomass yield between model predictions and the experimental data, especially for ML308 (Figure 3-5), and hypothesised the reason being the inaccuracy of cellular energy demand assumed by the core model when applied to this particular species growing under non-overflow and overflow conditions. To test this hypothesis, we collected the growth data in the overflow region for ML308 (Table 7 in [3]) and found that the reported steady-state ATP production rate was lower than what the core model suggested, which was derived from MG1655 strain. Therefore, we decided to remodel the cellular energy demand by subtracting the surplus portion, reducing it to the strain-specific values reported in the literature (AppxI-Table S2).

The original cellular energy demand embedded in the core model is quantified by

$$r_{ATP,nominal} = \mathrm{ATPM} + \sigma\lambda + v_{GLNS} + v_{PFK} \qquad \text{Equation 3-14}$$

where $r_{ATP,nominal}$ is the overall ATP consumption rate (equivalent to the ATP production rate at steady state); subscript 'nominal' indicates the default specification of the core model ; v_{GLNS} and v_{PFK} are the fluxes of the enzymatic reactions glutamine synthetase and phosphofructokinase (one mole ATP is required per mole flux of each reaction); $(\mathrm{ATPM} + \sigma\lambda)$ denotes the maintenance energy required for non-metabolic processes, where ATPM corresponds to the non-growth-associated maintenance (NGAM) and σ to the growth-associated maintenance (GAM) [36].

The adjusted cellular energy demand is formulated as

$$r_{ATP,new} = r_{ATP,nominal} - S(\lambda)$$ Equation 3-15

where $S(\lambda)$ is the offset energy, i.e. the amount of energy over-predicted by the core model. The mathematical analysis of the growth data of ML308 suggested that in the overflow region, the offset energy is linearly related to the growth rate ($R^2 = 0.9998$):

$$S(\lambda) = k\lambda + c$$ Equation 3-16

where k and c are constants. Substituting Equation 3-14 and Equation 3-16 into Equation 3-15, we have

$$r_{ATP,new} = (\mathrm{ATPM} - c) + (\sigma - k)\lambda + v_{GLNS} + v_{PFK}$$ Equation 3-17

For simplicity, we define $\mathrm{M} \equiv \mathrm{ATPM} - c$ and $\mathrm{N} \equiv \sigma - k$; M and N are referred to as the (adjusted) maintenance parameters. The adjustment of the cellular energy demand in our FBA is achieved by manipulating the maintenance energy, more specifically, through the maintenance parameters. Note that in the case where growth energetic data is not available (i.e. for MG1655 and NCM3722), such adjustment was not possible, therefore the default maintenance parameter values were used (see Table 3-1).

3.2.5 Alternative pathways in central metabolism

The model constructed in this chapter considers only the central metabolism of *E. coli* as detailed in the *E. coli* core model. The energy biogenesis pathways in the model consist of glycolysis (the EMP pathway), the TCA cycle, the acetate pathway (PTA-ACKA) and the terminal oxidative phosphorylation system. However, we noted the existence of alternative pathways in the central carbon metabolism, which include the Entner-Doudoraff (ED) pathway, the pentose phosphate (PP) pathway and the more recently explored PEP-glyoxylate cycle [37].

3.2.5.1 ED pathway

The ED pathway was found to be three to five-fold more protein-efficient than the EMP pathway to achieve the same glycolytic flux [38], which provides a clear rationale for the utilisation of the ED pathway in several bacteria, e.g. *Sinorhizobium meliloti, Rhodobacter sphaeroides, Zymomonas mobilis, and Paracoccus versutus* [39]. However, Flamholz *et al*. acknowledged that *E. coli*, which is capable of using both the ED and the EMP pathways, tends to use the latter. Flux measurements also suggest that the usage of the ED pathway by *E.coli* K-12 is minimal: only about 2% of glucose catabolism proceeds through the ED pathway in batch cultures [40] and about 6% in mini-scale chemostats [41]. Furthermore, the activity of the ED pathway was detected only under slow- to mild-growing conditions [40,41]. To our knowledge, no activity of the ED pathway in *E. coli* has been reported under fast-growing scenarios.

3.2.5.2 PEP-glyoxylate cycle

As for the PEP-glyoxylate cycle, similar to the ED pathway, its usage was identified to be significant only under slow-growing conditions. Not even a trace activity of the PEP-glyoxylate cycle was found in wild-type batch cultures or more rapidly growing chemostats [37]. Furthermore, the flux comparison between the aceA-pckA knockout strain and the sucC knockout strain [16] verifies that compared to the TCA cycle, the alternative PEP-glyoxylate cycle plays a less significant role in glucose-limited fast-growing cultures of *E. coli*. Based on above literature evidence, this chapter, focusing on the overflow metabolism that occurs at fast-growing cultures of *E.coli* with relatively sufficient substrate availability, has taken the assumption that the use of alternative ED pathway and PEP-glyoxylate cycle is negligible compared to the glycolysis (i.e. EMP pathway) and the TCA cycle.

3.2.5.3 PP pathway

The PP pathway, on the other hand, can function as a significant alternative to the upper part of glycolysis for carbon catabolism in *E. coli*. Previous studies showed that the carbon flow through the PP pathway could reach 20-35% of the total carbon intake and can vary with different growth rates [40,41] hence neither a constant portion of carbon is diverted into PP pathway nor this portion of carbon flux negligible. The uncertainty embedded in the PP pathway flux motivates us to study the impact of different portions of substrate carbon allocated between the upper part of the EMP pathway and the PP pathway on the proteomic cost parameters and the model predictions. More details can be found in Appendix I, Supplementary Text. We define the PP pathway ratio ($PPP\%$) as the portion of substrate carbon directing to the PP pathway to the total carbon intake:

$$PPP\% = \frac{PGL}{EX_glc(e)} \times 100\% \qquad \text{Equation 3-18}$$

where 6-phosphogluconolactonase (PGL) is chosen to represent the PP pathway flux as it is a major and also the beginning enzymatic reaction in the pentose phosphate shunt; 'EX_glc(e)' is the exchange reaction denoting glucose uptake rate. In our simulation, $PPP\%$ was controlled by setting the upper bound of the portion of carbon that is directed into the PP pathway, denoted as $uPPP\%$, with the aid of an auxiliary term DM_PPP_RATIO:

$$uPPP\% \times EX_glc(e) - PGL = DM_PPP_RATIO \geq 0 \qquad \text{Equation 3-19}$$

3.3 Results

E.coli MG1655, NCM3722 and ML308 have been selected as the model strains in this chapter and are referred to as MG, NCM and ML respectively from this point on. In this section, the simulated acetate excretion pattern is presented against experimental data to demonstrate the accuracy of the model prediction. Subsequently, we elucidate the linear

interdependency of the proteomic cost parameters. In particular, we reveal the similarities and differences between the three *E. coli* strains. With respect to the PP pathway ratio ($PPP\%$), previous studies on the slow-growth strain MG show that the portion of carbon that goes into the PP pathway can be approximately 20% of the total carbon intake [40,41]. In this chapter, we set the upper bound of the carbon flowing into PP pathway ($uPPP\%$) to 25%, 35% and 40% to investigate the potential effect of the change in $PPP\%$ on proteomic cost parameters and model prediction (more justification is provided in Appendix I, Supplementary Text). On cellular energy demand, we refer to the original energy demand specified in the core model [34] as nominal energy demand and present first the set of results which were generated on this basis. Subsequently, we show how an adjusted energy demand (particularly applied to ML, referred to as ML-new) affects the patterns of the estimated proteomic cost parameters and the accuracy of biomass yield prediction.

Table 3-1. Values of the maintenance parameters.

Strains	**M**	**N**	**Scope of applicability**
MG-/NCM-/ML-nom	8.39[a]	59.81[a]	Complete growth range
ML-new	-50.90[b]	93.34[b]	Overflow region

a. default value specified in the *E. coli* core model
b. estimated from the growth data (Table 7 in Holms 1996)

3.3.1 Model prediction of overflow metabolism with nominal energy demand

Here, the accuracy of the predicted acetate excretion rate is compared with experimental data. Variation in the proteomic cost parameters with the changed carbon level diverted into the PP pathway is also presented.

3.3.1.1 Model prediction of acetate production

Figure 3-1 shows that model prediction of the pattern of acetate excretion is in good agreement with the experimental observations for three different *E. coli* strains. The onset of the production of acetate is concomitant with the drop in the respiratory flux, indicating

a switch between fully-respiration to respiration-fermentation mode. As the growth rate further increases, the acetate flux becomes dominant while the extent of respiration is gradually diminishing. It is worth noting that zero acetate production was commonly observed at low growth rates of different strains [3,16,41,42]; To emphasise the (strain-specific) acetate production pattern, we only collect the data with non-zero acetate production. For all the strains, data involving growth rates lower than those presented in Figure 3-1 are associated with non-detectable acetate excretion, hence are not shown here.

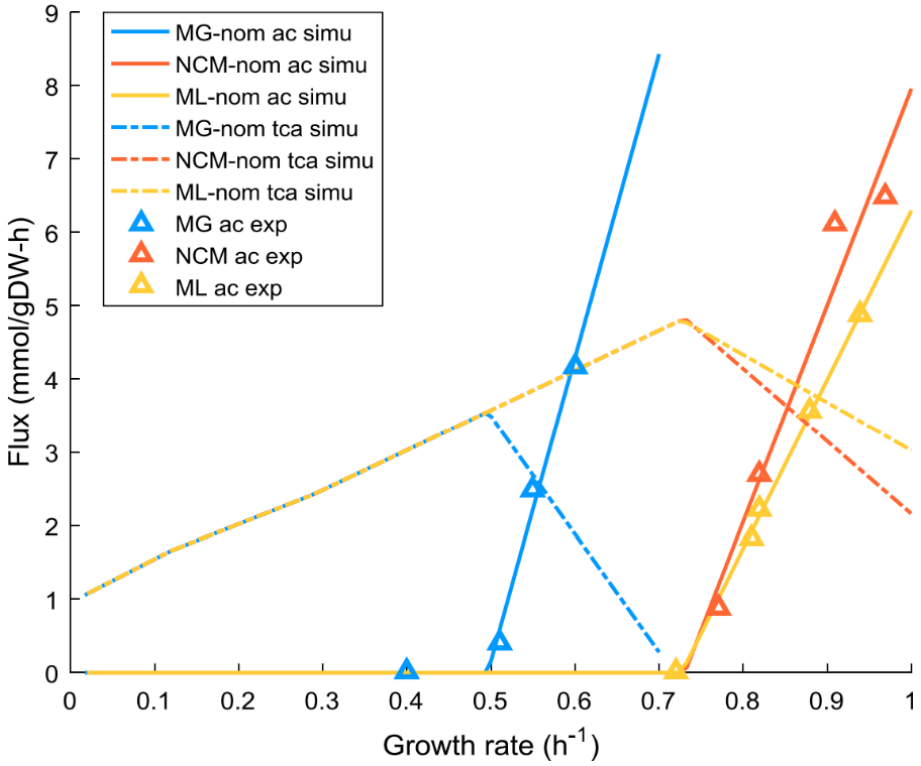


Figure 3-1 Model predictions of overflow metabolism for MG, NCM and ML at nominal energy demand. The extent of overflow metabolism is represented by the acetate flux. Simulation results of the respiration flux are drawn to show the switch between fully-respiration and respiration-fermentation mode. Comparison is made between model predictions and experimental data for the rates of acetate production. $uPPP\%$ was set to 35% according to the flux measurement [41]. Other $uPPP\%$ values render similar results (see AppxI-Table S1). Experimental data were obtained from different sources [3,16,41]. Data points from [16] were converted using 1 mM A_{600nm}^{-1} h^{-1} = 2 mmol gDW^{-1} h^{-1} according to [28]. '-nom' refers to nominal, the default energy demand specified in the core model. ac – acetate flux, tca – tricarboxylic acid cycle flux, simu – simulation results, exp – experimental data.

3.3.1.2 Linear relationships between proteomic cost parameters

When the nominal energy demand is adopted (indicated by '-nom'), the change of $uPPP\%$ leads to insignificant changes to the $w_r^* - w_f^*$ line for each strain. Between different strains, MG and NCM share nearly identical lines. The lines of ML-nom deviate from those of the former two, but not significantly (although this closeness will be altered with the adjusted energy demand, see Figure 3-2 ML-new and the section below).

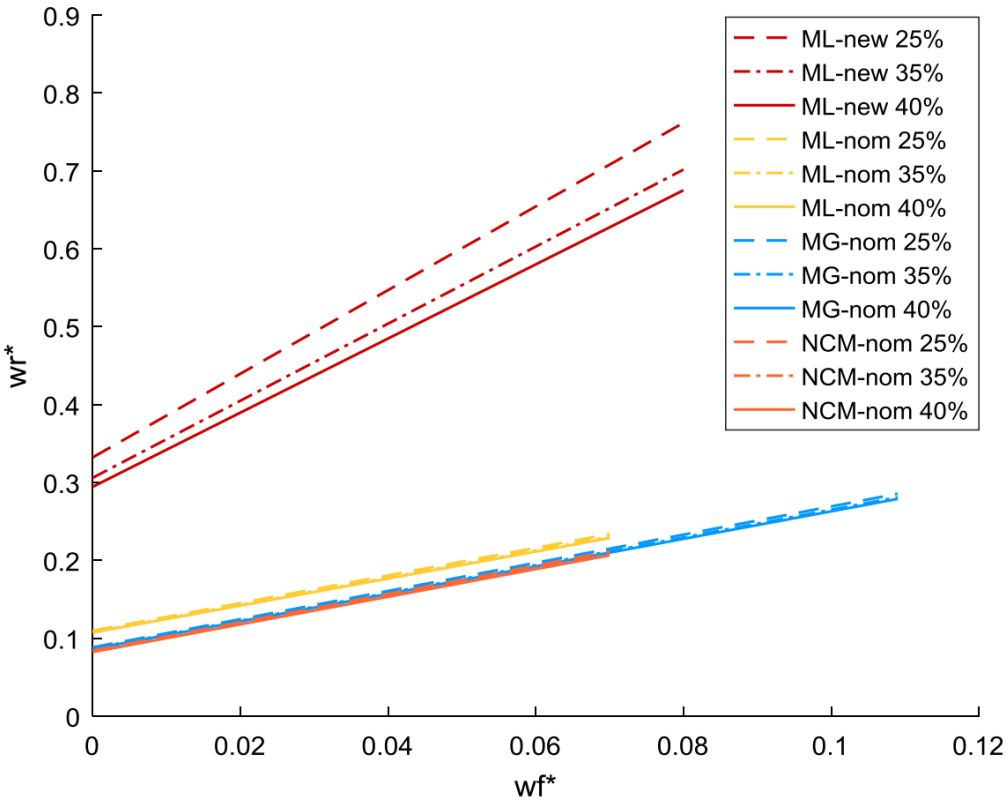


Figure 3-2 $w_r^* - w_f^*$ relationship for MG, NCM and ML with nominal energy demand and for ML with new energy demand. '-new' refers to the adjusted energy demand. $uPPP\%$ was set to 25% 35% and 40% for each strain.

In any case, w_r^* is clearly higher than the corresponding w_f^*, implying that respiration has a higher (lower) proteomic cost (efficiency) than fermentation for energy production, which is consistent with what was derived from protein abundances data for comparable parameters in [16]. To inspect the insignificant disparity in the $w_r^* - w_f^*$ lines when all strains use the nominal energy demand, Equation 3-13 is re-arranged to

$$w_r^* = \frac{v_{f,0}}{v_{r,0}} w_f^* + \frac{1}{v_{r,0}}$$ Equation 3-20

The slope and intercept of the $w_r^* - w_f^*$ line are dictated by $\frac{v_{f,0}}{v_{r,0}}$ and $\frac{1}{v_{r,0}}$, respectively. $v_{f,0}$ can be determined directly by the experimental measurement of acetate production. $v_{r,0}$, on the other hand, is a result of the combination of (measured) rates of acetate production and the mass and energy balance structure of the metabolic model.

For a specific strain, $v_{f,0}$ only depends on the pattern of acetate excretion, not affected by the assumed level of $uPPP\%$. Therefore, the impact of $uPPP\%$ on the $w_r^* - w_f^*$ line is through affecting the value of $v_{r,0}$, which turns out to be rather moderate. Between different strains, the ratio of $v_{f,0}$ and $v_{r,0}$ and the value of $v_{r,0}$ are nearly identical between MG and NCM, regardless of the level of $uPPP\%$ adopted, resulting in the very much overlapped pattern of the $w_r^* - w_f^*$ relationship between MG and NCM. For ML, the value of the slope is slightly smaller than MG and NCM, while the intercept is about 25% larger (as shown in AppxI- Table S3). Figure 3-2 also suggests that the proteomic cost (efficiency) of respiration pathways for ML is higher (lower) than that for MG and NCM, regardless the modification in the energy demand.

Compared to the $w_r^* - w_f^*$ relationship, that of $b^* - w_f^*$ appears to be affected by the level of $uPPP\%$ more visibly (Figure 3-3). Between different strains, the difference is also more pronounced, and closeness is present between the two rapid-growth strains NCM and ML (as presented in AppxI-Figure S2). For a specific strain, the increase of $uPPP\%$ gradually moves the $b^* - w_f^*$ line to the right (yellow arrow, Figure 3-3), corresponding to an increase in b^* (blue arrow, Figure 3-3).

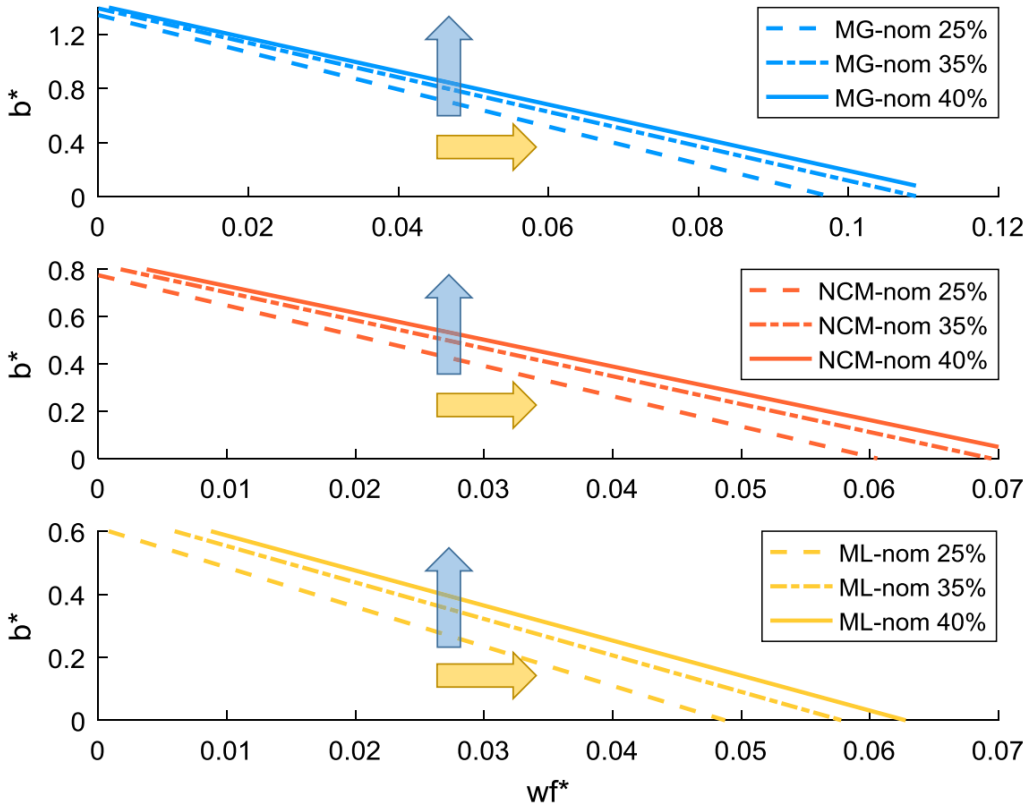


Figure 3-3 $b^* - w_f^*$ relationship for MG, NCM and ML strains with nominal energy demand. $uPPP\%$ was set to 25% 35% and 40% for each strain. The blue arrow shows the increase in b^* with the increase of *PPP%* at fixed w_f^*; the yellow arrow shows the right-shifting trend of the $b^* - w_f^*$ line with the increase of *PPP%*.

This trend can be explained by inspecting a re-arrangement of Equation 3-12:

$$b^* = \left(k_r \frac{v_{f,0}}{v_{r,0}} - k_f\right) w_f^* + \frac{k_r}{v_{r,0}} \qquad \text{Equation 3-21}$$

Equation 3-21 suggests that the shift of the $b^* - w_f^*$ line results from the change in the respiratory flux (note the intercept, $\frac{k_r}{v_{r,0}}$). In *E. coli*, the PP pathway and the TCA cycle are two major sources for the production of NADPH [40]. At a given growth rate, the amount of NADPH needed for cell growth is fixed based on the mass and energy balance. As $uPPP\%$ increases, more carbon is predicted to enter the PP pathway. In the model simulation, an increase in the amount of NADPH produced via the PP pathway would force a drop of the flux into the TCA cycle to maintain the constant total production rate of

NADPH. The reduction in the TCA flux in turn manifests in a lower v_r. In the overflow region, Equation 3-7 is bounded by the equal sign. As shown earlier, the level of $uPPP\%$ has a negligible impact (when nominal energy demand is adopted) on the $w_r^* - w_f^*$ line. Also recall that the relation between the rate of acetate excretion and steady-state growth rate is fixed by the experimentally measured growth data. With all the other quantities (w_f^*, v_f, w_r^* and λ) fixed in Equation 3-7, the drop in v_r due to the increase in $uPPP\%$ will necessarily be accompanied by an increase in b^*.

Between different strains, b^* varies significantly. In particular, b^* for MG is remarkably larger than that of NCM and ML (as presented in AppxI-Figure S2). This disparity can again be explained by Equation 3-7. For a certain value of w_f^*, w_r^* is rather similar among different strains (with nominal energy demand) as shown by Figure 3-2. In the overflow region, the respiration flux v_r of MG is much smaller than the others (see Figure 3-1), which thus leads to a lower value of the $w_r^* v_r$ term for MG than NCM and ML. As the value of the $w_f^* v_f$ term (for any selected value of w_r^*) is similar between these strains, due to their similarity in the relationship between w_r^* and w_f^*, the value of the remaining term on the left-hand side of Equation 3-7, $b^*\lambda$, must be higher for MG than for the other two strains. On the other hand, in the overflow region and at a same acetate excretion rate v_f, the growth rate of MG has been shown to be much lower than that of NCM and ML. Now, a higher value of $b^*\lambda$ coupled with a lower value of λ will undoubtedly lead to a higher value of $b^*$for MG, compared to the other two strains.

The above mathematical explanation in fact coincides with the known biological fact that the inverse of b^* is proportional to the rate of protein synthesis [31]: the slower the rate of protein synthesis, the higher the value of b^*. Thus for the slow-growing strain MG, it is expected to have a higher value of b^*compared to the fast-growing strains NCM and ML.

3.3.1.3 Predicted evolution of PP pathway flux

The results presented above show a rather moderate impact of the upper limit of PP pathway ratio ($uPPP\%$) on the linear interdependency of the proteomic cost parameters. With an interest in the FBA solution of the flux distribution in PP pathway (at different growth rates), simulation results were recorded for three strains with $uPPP\%$ set to 35%; other *uPPP%* levels displayed a similar trend (as presented in AppxI-Figure S3). Flux variability analysis (FVA) [43] was performed to confirm that the trend of $PPP\%$ presented here was unique. In general, $PPP\%$ gradually increases with the growth rate. Two turning points can be observed, which divide the whole curve into three distinct phases (Figure 3-4(a)). A close inspection of the model simulation revealed that the variation of the predicted PP pathway ratio was co-related particularly with three fluxes, namely NAD transhydrogenase (NADTRHD), transketolase (TKT2) and NADP transhydrogenase (THD2).

In phase I, only NADTRHD is active, with zero fluxes for both TKT2 and THD2. The enzymatic reaction NADTRHD functions to convert NADPH into NADH. Thus in phase I, it is likely that the amount of NADPH produced exceeds the required amount for biosynthesis; NADTRHD is thus activated to consume the surplus NADPH.

In phase II, an on/off swap occurs between NADTRHD and TKT2 while THD2 still remains silent. We infer that in this phase, NADPH produced satisfies the demand, but the amount of carbon flowing into the PP pathway surpasses the rate of the carbon withdrawal (for the synthesis of biomass precursors). Therefore, TKT2 is activated to direct the extra amount of four-carbon and five-carbon compounds back to the glycolysis.

In phase III, THD2 is finally switched on and becomes significantly active in the high-growth-rates region. TKT2 increases progressively while NADTRHD remains silent. It is

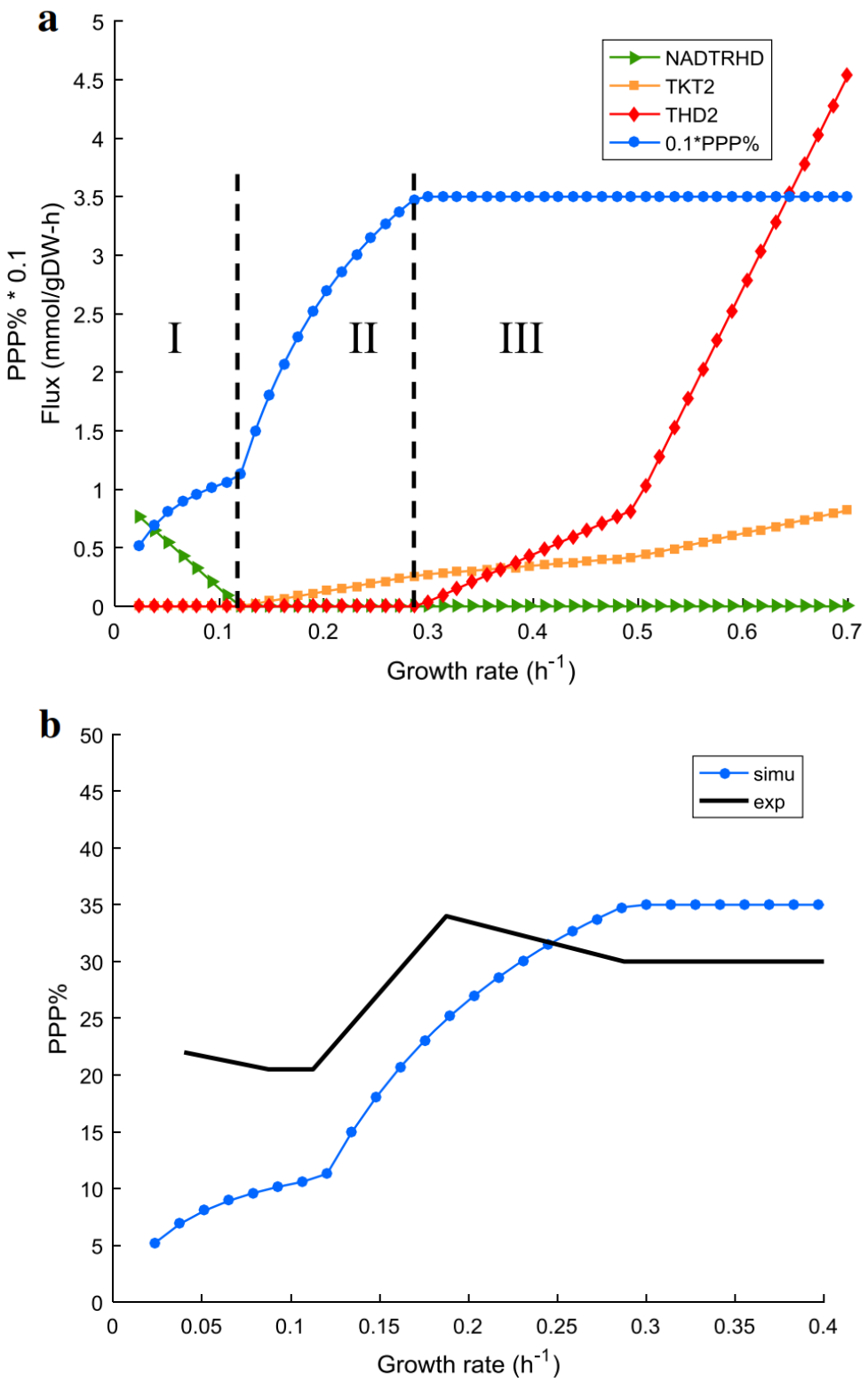


Figure 3-4 **(a)** Simulation results of *PPP*%, NADTRHD, TKT2 and THD2 against growth rates at nominal energy demand. **(b)** Comparison between the predicted trend of *PPP*% and experimental data. PP pathway ratio (*PPP*%) is divided by ten (0.1*PPP%) to unify the order of magnitude between different data types. Experimental data were obtained from [41]. $uPPP\%$ was set to 35%. Simulation was based on MG1655.

presumed that in this phase, as the growth rate becomes higher, more NADPH is required for biomass synthesis. NADP transhydrogenases (THD2) is activated to produce NADPH needed in rapid growth. The surplus carbon flux in the PP pathway, which might result from the high glucose uptake rate at a high growth rate, is directed back to glycolysis via TKT2.

It would be desirable to verify the theoretical prediction of the evolution of $PPP\%$ with experimental measurements, which unfortunately have not been widely reported in the literature. Nevertheless, Figure 3-4(b) shows a comparison with one set of experimental observations available [42], which suggests a good degree of qualitative similarity.

3.3.2 Adjusting cellular energy demand improves the prediction of biomass yield

Although combining the PAT constraint with the core model succeeded in predicting the rates of acetate production, the accuracy in biomass yield varied and was especially unsatisfactory for the ML strain (Figure 3-5). A similar deficiency in yield prediction was also reported in [28]. Focusing on the yield, two features can be observed: (i) in the overflow region, for a fixed growth rate (associated with an acetate excretion rate) the biomass yield for ML is higher than MG and NCM; and (ii) the rate of the drop in yield (i.e. the slope) of ML is sharper than the other two strains.

Intuitively, feature (i) suggests that in ML, the amount of energy required per unit mass of biomass formation should be less than NCM or MG. Therefore we collected the growth data of ML and remodelled the cellular energy demand (see Section 3.2.4). It is worth noting that for ML, the negative value of M (Table 3-1) clearly indicates a constrained applicable range of the maintenance parameters, i.e. valid only within the overflow region. As growth rate decreases, if M stays unchanged, the overall energy consumption (Equation 3-17) will drop to a negative value, which is not biologically feasible.

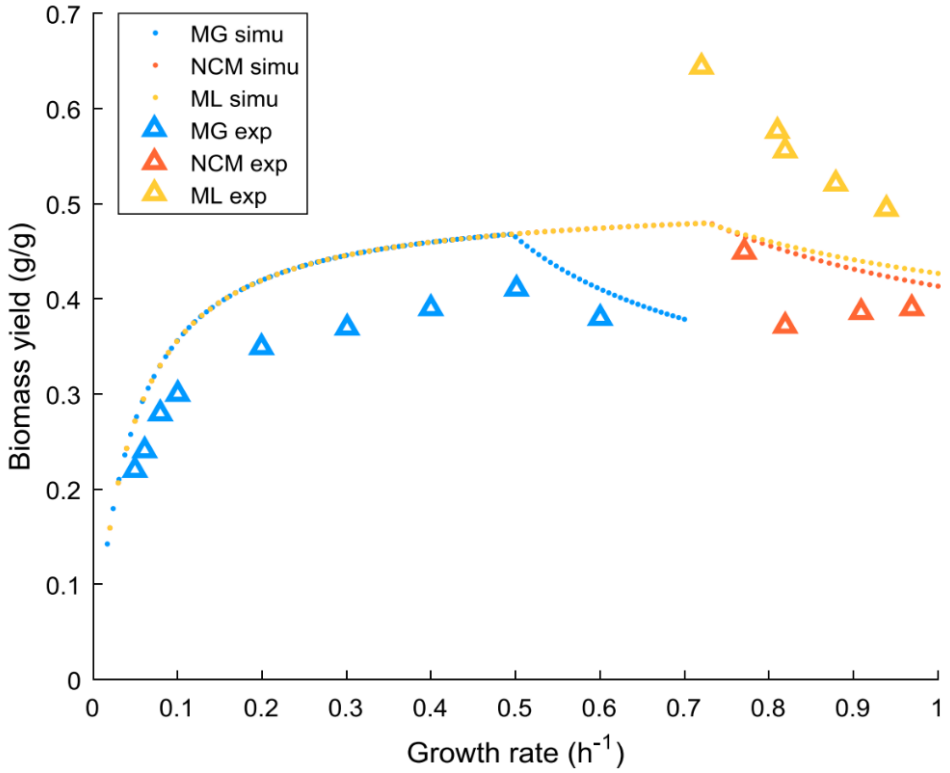


Figure 3-5 Comparison of the biomass yield between model predictions with nominal energy demand and the experimental data. Biomass yield is calculated as gram biomass produced per gram substrate consumed. $uPPP\%$ was set to 35% for all strains. Experimental data were obtained from the same sources [3,16,41] of the acetate data shown in Figure 3-1.

This then implies a certain degree of nonlinearity in the global relationship between (total or maintenance) energy requirement and growth rate. Such proposition was previously referred to as 'varied non-growth-associated maintenance' [3]. Non-linearity in energy consumption manifesting before and after the onset of the overflow metabolism has also been observed and discussed in a recent work [44].

3.3.2.1 Model prediction of biomass yield with adjusted energy demand

We first re-estimated the set of proteomic cost parameters for ML with the adjusted energy demand (Table 3-1 and Methods). Applying updated values of w_f^*, w_r^* and b^* together with the adjusted maintenance energy, our model is now able to effectively capture the unique trend of biomass yield for ML, without any compromise in the accuracy of predicting acetate excretion (Figure 3-6). Simulation results for ML with adjusted energy

demand are referred to as ‘ML-new’. Improved predictions of the biomass yield for MG and NCM strains could be obtained through a similar principle (AppxI-Figure S4).

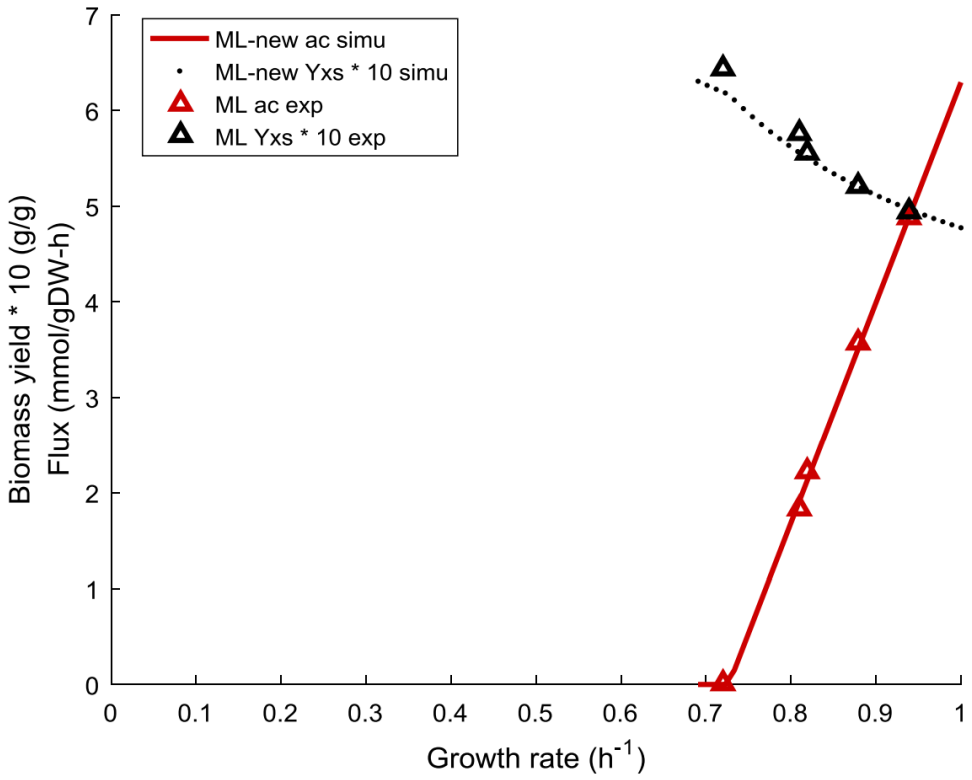


Figure 3-6 Model prediction of acetate production and biomass yield for ML with adjusted energy demand compared with experimental data. The simulation was done with adjusted energy demand (Table 3-1, M, N for ML) and updated proteomic cost parameters. Biomass yields are shown as ten times of the original value to unify the order of magnitude between different types of data. $uPPP\%$ was set to 35%. Experimental data was obtained from Table 7 in [3].

It is worth noting that, our model also succeeds in matching the elevated reduction in the yield of ML in the overflow region as the growth rate increases. This captured trend appears to originate from the low energy demand of ML. Approximately, the yield reduction rate can be considered as being proportional to the ratio of the increase in the acetate excretion ($ac_2 - ac_1$) and the increase in glucose uptake rate ($glc_2 - glc_1$), while the growth rate rises from λ_1 to λ_2:

$$yield\ reduction\ rate \propto \frac{ac_2 - ac_1}{glc_2 - glc_1}, for\ \lambda_1 \to \lambda_2\ (\lambda_2 > \lambda_1) \qquad \text{Equation 3-22}$$

NCM and ML exhibit similar acetate excretion rates, hence a similar value in '$ac_2 - ac_1$'. However, the energy demand per unit growth of ML is much lower than NCM, which means that with a similar increase in acetate production, the increase in substrate intake (i.e. $glc_2 - glc_1$) for ML will be lower than NCM to achieve a given increment in the growth rate. According to Equation 3-22, the yield reduction rate of ML will thus be higher than NCM.

3.3.2.2 Impact of the adjusted energy demand on $w_r^* - w_f^*$ and $b^* - w_f^*$ relationships

To investigate the impact of the change in cellular energy demand on the linear relationships of w_f^*, w_r^* and b^* of ML-new, we recalculated constants k_f, $v_{f,0}$, k_r and $v_{r,0}$ at different $uPPP\%$ values (25%, 35% and 40%) to update the linear equations describing $w_r^* - w_f^*$ line and $b^* - w_f^*$ (Equation 3-12 and Equation 3-13). The resulting $w_r^* - w_f^*$ lines for ML-new are plotted in Figure 3-2, together with the results obtained earlier for MG/NCM/ML with nominal energy demand. The switch to the adjusted energy demand makes the w_r^* value for ML-new much higher than that of ML-nom, the latter being rather close to those of the MG-nom and NCM-nom. This implies that the adjustment of the energy demand of ML leads to an enlarged gap in the proteomic efficiency between respiration and fermentation.

The similarity among MG-/NCM-/ML-nom has already been discussed in the previous section. Here we mainly focus on the discrepancy with ML-new. We found that both the slope and intercept of $w_r^* - w_f^*$ line for ML-new are about three times larger than ML-nom (as shown in AppxI-Table S4). The dramatic changes in the slope and intercept of ML-new predominantly result from the reduction in the respiratory flux v_r when applying the adjusted energy demand (Figure 3-7 and AppxI-Figure S5). The link between the decrease in v_r and the increase in w_r^* has been discussed in the previous section.

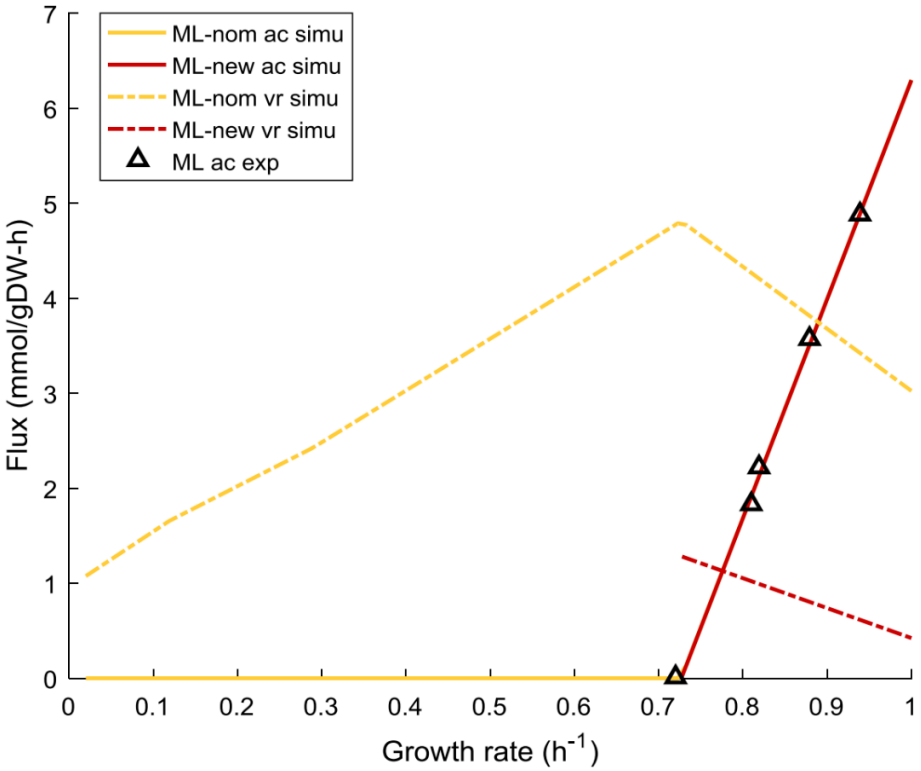


Figure 3-7 Comparison of the predicted respiration and acetate fluxes between ML-new and ML-nom. The predicted rates of acetate production for ML-nom and ML-new are completely overlapped with each other. $uPPP\%$ was set to 35% for both strains. Data source of acetate excretion was shown in Figure 3-1.

The results presented herein indicate that it is the energy demand that plays a major role in distinguishing the $w_r^* - w_f^*$ relationship between different strains, not the $uPPP\%$ or the acetate excretion pattern. Applying the adjusted energy demand also has an impact on the relationship between b^* and w_f^*. As shown in Figure 3-8, the $b^* - w_f^*$ lines are significantly right-shifted when the model is changed from ML-nom to ML-new (i.e. red lines are located in a much right area than yellow lines). Given the identical pattern of acetate excretion between ML-nom and ML-new (as both predicted the same set of experimental data), the amount of energy produced through fermentation remains unchanged. For ML-new, as the energy demand per unit of growth is much lower than that of the nominal strain, the respiratory flux v_r must decrease significantly to avoid energy overproduction, as confirmed in Figure 3-7.

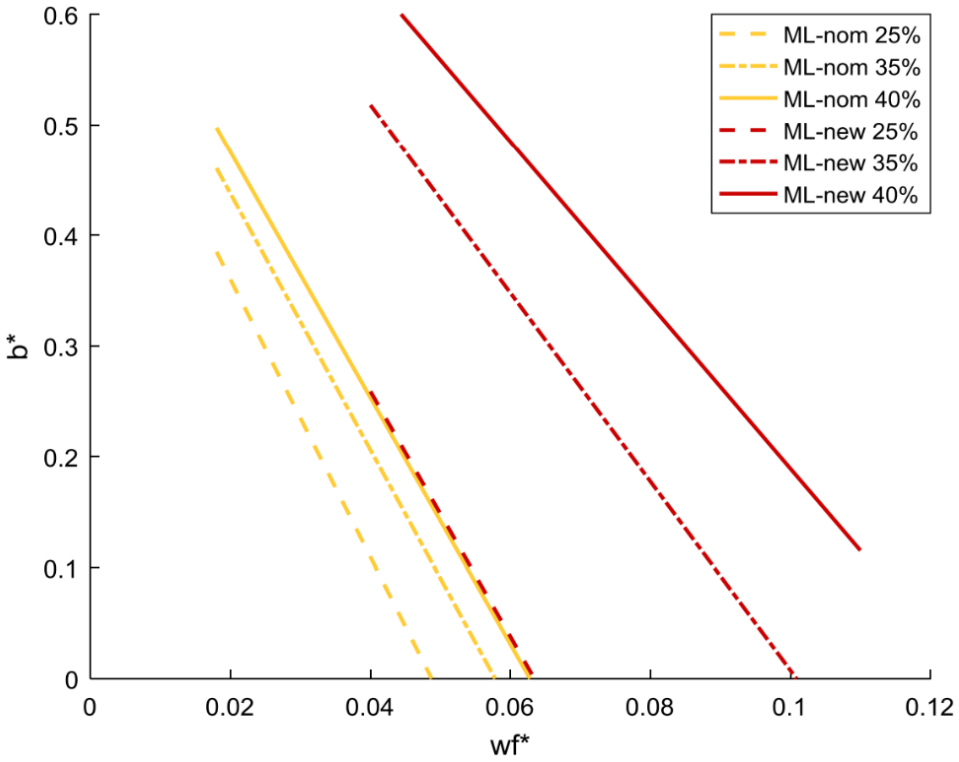


Figure 3-8 Comparison of the $b^* - w_f^*$ relationship between ML-nom and ML-new at different $uPPP\%$ levels. $uPPP\%$ was set to 25, 35 and 40% for both strains.

Although the value of w_r^* for ML-new is higher than that for ML-nom (for a given value of w_f^*, Figure 3-2), the value of the product $w_r^* v_r$ for ML-new still becomes lower (as the increase in w_r^* is not able to compensate for the sharp drop in v_r). With no change in $w_f^* v_f$ between ML-new and ML-nom (for a given w_f^*), Equation 3-7 again dictates b^* to become higher for ML-new than ML-nom, hence the right-shifting of the $b^* - w_f^*$ lines.

In the case of ML-nom, it was shown earlier that the increase in $uPPP\%$ would lead to a reduction of v_r, which in turn would lead to an increase in b^* or right-shifting of the $b^* - w_f^*$ line (Figure 3-3). Now for ML-new, an enlarged gap between the $b^* - w_f^*$ lines at different $uPPP\%$ levels is observed compared to the case of ML-nom. This implies that the effect of v_r reduction due to the increase in $uPPP\%$ is more pronounced with the adjusted energy demand.

3.4 Discussion

3.4.1 Comparison to relevant models

In the study by Basan *et al.* [16], which has been the basis of the PAT constraint formulated in our model, the application of a constraint on proteome fractions (similar to Equation 3-5) with parameters derived from measured protein abundances was able to accurately predict the patterns of acetate excretion for *E. coli* under different growth conditions, when coupled with a simple energy balance equation. In this chapter, we have embedded the PAT into the core metabolic model of *E. coli*, taking advantage of the latter in offering more rigorous modelling of intracellular mass and energy balances. Furthermore, the constraint-based metabolic model allows prediction of detailed metabolic fluxes as opposed to merely acetate production, which could provide more insights about metabolic pathways in connection with the overflow metabolism and pave the way for investigating acetate excretion in conjunction with possible manipulations of the metabolic network.

In addition, the respiratory flux in Basan's work was associated with the carbon dioxide produced in respiration, termed $J_{CO2,r}$, whose value was deduced by subtracting the fermentation-dependent CO_2 and the growth-dependent CO_2 from the total CO_2 production. As such, $J_{CO2,r}$ could not directly correspond to a specific flux in the metabolic network. In our model, the respiratory flux directly refers to a specific flux within the TCA cycle (AKGDH), which appears to be a convenient choice when the PAT is embedded into FBA. Using a constraint-based metabolic model that includes the TCA cycle with a reasonable level of detail, the respiration flux can be directly resolved via FBA, without the need for multi-step calculation along with different levels of assumptions and uncertainties.

Another important comparison we would like to make is with the recently developed model – CAFBA. It was mentioned in the Background that unlike CAFBA which considers the proteomic cost of every individual reaction in the metabolic network, the PAT constraint in our model follows the treatment of Basan *et al.*'s work and quantifies the proteomic costs at the pathway level. This simplification allows the model to explicitly incorporate the differential proteomic efficiencies of fermentation and respiration that are proven (in both Basan *et al.* and this chapter) to govern the flux split between the two pathways. This simplicity comes at the cost of the limited utility of our model: it is intended to be used only for predicting the interplay between the excretion of acetate (or other fermentation products) and the growth rate during overflow metabolism, not other effects of stressed resource allocation.

The proteome allocation constraint in CAFBA includes a C-sector (via a term expressed as $w_c v_c$) representing the proteome requirement for uptaking carbon source, which is not explicitly considered in this chapter. We have ignored this sector as the carbon overflow occurs only at the high growth rate region, where w_c (proteomic cost of the C-sector) approaches zero at high substrate uptake rates, as shown in CAFBA [28]. In this region, the low value of w_c makes the C-sector negligible compared with other proteome sectors. At low growth rate region, the value of w_c becomes significant, however no acetate is excreted in this region, where the equality relationship in the PAT constraint in our work becomes inactive so that the significance of the C-sector becomes irrelevant.

As for the prediction of biomass yield, CAFBA noticed the difficulty in predicting the biomass yield of ML308. In this chapter, we have found that it is the cellular energy demand that significantly affects the FBA prediction of biomass yield. After replacing the default energy demand with data reported specifically for ML308 strain, our model was able to produce an accurate prediction. Therefore, we consider that it is important to carry

out necessary adjustment to the cellular energy demand when applying such a constraint-based modelling approach to specific strains.

3.4.2 Parameterisation of the proteome allocation constraint

The modelling approach proposed in this chapter can be considered as 'halfway' between the coarse-grained proteome allocation model of Basan *et al.* [16] and the FBA models that incorporate reaction-level resource allocation constraints such as CAFBA [28] and FBAwMC [21]. In CAFBA, the proteome constraint involves ~1000 proteomic cost parameters (w_i) for a genome-scale model. Similarly in FBAwMC, a large number of crowing coefficients need to be specified. In both cases, the existence of numerous cost parameters originates from associating the resource cost with individual reactions. These parameters conceptually have a clear biological meaning and in principle can be determined experimentally by e.g. proteome measurements or extensive enzyme assays. However, in practice, it has appeared to be difficult to reliably obtain precise values for all the parameters, especially for different strains growing at different growth rates or conditions. In fact, instead of pursuing the exact values for all the individual parameters, CAFBA focused on applying the average value of the proteome fraction invested per unit flux, termed as $\langle w \rangle$, to capture the key flux pattern, along with evaluating the impact of possible heterogeneous values of the proteome parameter w_i on the model prediction. Similarly, FBAwMC [21,22] also appears to encounter a certain degree of 'randomness' of its crowding coefficients due to the unknown enzyme kinetics and/or turnover numbers. Subsequently, molecular-crowding-based modelling normally treats this 'randomness' as noise, where the crowding coefficients are chosen randomly from distribution of crowding coefficients [45] or the majority of the crowding coefficients are estimated from a limited number of known enzyme turnover values (k_{cat}) [11].

The inability to uniquely determine proteomic cost parameters (or equivalent cost parameters) is possibly related to identifiability issues associated with the highly under-determined nature of genome-scale FBA models. Biologically, these cost parameters can be uniquely determined by measuring genome-scale protein abundances and flux distributions of cells grown at different dilution rates using e.g. chemostat growth experiments or by determining *in vivo* turnover rates of metabolic enzymes at genome-scale. However, it is experimentally expensive and time-consuming to obtain genome-scale and growth rate-dependent proteomics, fluxomics and kinetome. Mathematically, given limited sets of growth rate-dependent proteomic and flux data, RBA [25] provides a potential way to bypass the identifiability issue by (i) designing the cost function at biological process level instead of individual reaction level; (ii) constraining the exchange and intracellular fluxes to be equal or close to measured values and (iii) constraining a few flux ratio of important nodes in the central carbon metabolism.

In this chapter, we have intended to formulate a constraint with a greatly reduced number of proteomic cost parameters, while still capturing the essence of constrained cellular resource allocation. This is achieved by formulating the proteome allocation constraint at the pathway (as opposed to reaction) level. The proposal is the concise Equation 3-7 ($w_f^* v_f + w_r^* v_f + b^* \lambda \leq 1$), involving only three proteomic cost parameters (representing proteomic efficiencies of fermentation, respiration and biomass synthesis pathways). In principle, these parameters can be uniquely determined using proteomic and fluxomic datasets of (i) the abundances/mass fractions ϕ_i of enzymes of the central carbon metabolism and ribosomal proteins at different growth rates via quantitative mass spectrometry and (ii) the flux distribution of the central carbon metabolism at different growth rates through ^{13}C-metabolic flux analysis. By plotting the sum of protein abundances/mass fractions $\sum \phi_i$ of a specific pathway against the representative flux v_i at

different growth rates, the value of w_f^*, w_r^* or b^* can be indicated by the slope of the regression line of the $\sum \phi_i - v_i$ plot [16]. However, in the current study, we attempted to parameterise this constraint using widely available growth data from cell culturing experiments, in particular growth rate and acetate production rate. It should be noted that cell culturing experiments often yield relatively simple data sets with measurements of a few process variables. It is infeasible to use such data sets to determine a large number of proteomic cost parameters encountered in a proteome constraint expressed at the individual reaction level. Even with the pathway-level constraint adopted in this chapter, our results show that cell growth and acetate production measurements (alone) cannot uniquely determine the three parameters (due to identifiability issues, which is possible to be resolved by adding more flux constraints), but two linear relationships between these parameters can be derived (Equation 3-12 and Equation 3-13).

Furthermore, our model shows that it is the two linear relationships (but not the absolute values) of the proteomic cost parameters that allow an accurate prediction of the overflow metabolism. We thus speculate that for an FBA-based model, the ability to capture the overflow behaviour is rendered by (i) an extra constraint representing the constrained proteomic resources and (ii) certain relations or relative magnitudes of the proteomic cost parameters embedded in the proteome constraint. In reality, the proteomic efficiencies of the metabolic pathways may vary (within a certain range), at different points in time or between cells in a population which often exhibits heterogeneity [28]. However, as long as the specific relations or relative magnitudes of these efficiencies are maintained, one can expect that the overflow behaviour will emerge.

3.4.3 Applicability of the linear formulation of the proteomic cost

Our formulation of the proteomic cost (Equation 3-2 and Equation 3-3) reflects the observed linear dependency between proteome fraction and growth rate [30–32].

Combining this linear dependency with the assumption that the flux processed by a proteome sector *i* is proportional to the growth rate-dependent component of the associated proteome fraction [30], we have related ϕ_i linearly with the flux it carries. A similar model is also adopted in CAFBA [28], where the linear proteome-flux relation is derived on the assumption that the substrate concentration is proportional to the flux.

Note that our model is intended specifically for predicting the steady-state overflow metabolism in *E. coli* under glucose-limited conditions. In some other circumstances, observations not conforming to this relatively simple model have been reported. For example, Goel and his co-workers found hardly any changes in protein levels in anaerobic slow-growing *Lactococcus lactis* chemostats, when the cell shifted from a high yield metabolic mode to a low yield metabolic mode with an increased growth rate [46]. In this case, although the metabolic shift in *L. lactis* is similar to the overflow metabolism observed in *E. coli*, proteome allocation did not seem to accompany the changes in metabolic fluxes. In a study on yeast's transient transcript, enzyme and metabolite responses under metabolic perturbation, it is revealed that the reaction rates are jointly regulated by enzyme capacity and metabolite concentration due to the cell's tendency in sacrificing the local metabolite homeostasis to maintain fluxes and global metabolite homeostasis upon enzyme perturbation [47]. Another yeast-based study also suggests that changes in individual flux are predominantly regulated by the levels of metabolites, not enzymes [48].

The above-mentioned experimental observations suggest that the linear proteome-flux relationship modelled in this chapter might not be applicable to those circumstances. We hypothesise that this might be at least partially due to the differences between *E. coli* (the target organism of our model) and the organisms with which those observations were made. Besides, our model has been developed to describe the relationship between (i) the

observed steady-state global proteome configuration and (ii) the growth rate or the corresponding flux, drawing on evidence and hypothesises from several previous studies [16,28,30,32]. Such a model, being global and coarse-grained, is not intended to represent delicate regulatory mechanisms responsible for the transient metabolic changes to maintain cellular homeostasis under perturbations, and might not be suitable for revealing the local regulatory insights on the key factors dictating the individual reaction rates.

Nevertheless, the assumed linear relationship between the proteome fraction and the metabolic flux retains the linearity of the proteome allocation constrained FBA model, which allows the growth optimisation problem to be solved by linear programming.

3.5 Conclusions

With three different *E. coli* strains, we have evaluated a new model that integrates a previously proposed proteome allocation theory (PAT) into the constraint-based modelling approach – flux balance analysis (FBA), which predicts the distribution of carbon fluxes between fermentation and respiration due to the differential proteomic efficiencies of the two energy biogenesis pathways. Using a simple proteome allocation constraint, our model allows the accurate prediction of acetate production at different steady-state growth rates during overflow conditions (with sufficient oxygen and glucose). The model involves three pathway-level proteomic cost parameters linearly interrelated by two equations, which is the consequence of (i) the assumed linear dependency of proteomic costs and the growth rate and (ii) the experimentally observed linear correlation between the fermentation or respiration flux and the growth rate. The non-unique optimal values of the three parameters, or the two linear relationships between them, could be obtained by fitting the model to experimentally measured acetate excretion rates at specific growth rates.

The linear relationships between the parameters were shown to be affected, in varying degrees, by (i) the acetate excretion pattern, (ii) the assumed upper limit of the substrate carbon diverting into PP pathway and (iii) the cellular energy demand. The proteomic cost of the fermentation pathway was estimated always to a lower value than that of the respiration pathway, i.e. $w_f^* < w_r^*$. The proteomic cost of the biomass synthesis sector was estimated to be higher in a slow-growing strain that excretes acetate at a lower growth rate, in comparison with the other two fast-growing strains, i.e. $b_{\mathrm{MG}}^* > b_{\mathrm{NCM/ML}}^*$. The estimated values of w_f^* and b^*both meet qualitatively the expectation from a biological point of view. Furthermore, the relationship between the proteomic efficiencies of fermentation and respiration, i.e. the $w_f^* - w_r^*$ line, was shown to change between different strains most significantly with the cellular energy demand rather than with the pattern of acetate excretion. This $w_f^* - w_r^*$ relationship remained relatively stable when the upper bound of the portion of substrate carbon flowing into the PP pathway ($uPPP\%$) varied in the modelling studies. On the other hand, the increase of $uPPP\%$ was shown to lead to a visible increase in the estimated proteomic cost of the biomass synthesis sector b^*, which mathematically results from a reduction in the predicted respiration flux.

Finally, and as a general point for constraint-based models, cellular energy demand appeared to have a major impact on the predicted biomass yield; tuning the default energy demand with strain-specific data was shown to be critical in making simultaneously accurate predictions of biomass yield and overflow metabolism.

Overall, the work presented in this chapter demonstrates the potential of combining a detailed metabolic model with a coarse-grained, pathway-level resource allocation constraint in producing quantitatively accurate predictions of the overflow phenomenon in

E. coli; similar modelling approaches that feature this type of combination may prove suitable also for other applications.

References

1. Eiteman MA, Altman E: **Overcoming acetate in Escherichia coli recombinant protein fermentations**. *Trends Biotechnol* 2006, **24**:530–536.
2. Farmer WR, Liao JC: **Reduction of aerobic acetate production by Escherichia coli**. *Appl Environ Microbiol* 1997, **63**:3205–3210.
3. Holms H: **Flux analysis and control of the central metabolic pathways in Escherichia coli.** *FEMS Microbiol Rev* 1996, **19**:85–116.
4. Pan JG, Rhee JS, Lebeault JM: **Physiological constraints in increasing biomass concentration of Escherichiacoli B in fed-batch culture**. *Biotechnol Lett* 1987, **9**:89–94.
5. Xu B, Jahic M, Enfors SO: **Modeling of overflow metabolism in batch and fed-batch cultures of Escherichia coli**. *Biotechnol Prog* 1999, **15**:81–90.
6. Lee SY: **High cell-density culture of Escherichia coli**. *Trends Biotechnol* 1996, **14**:98–105.
7. Holmes WH: **The central metabolic pathways of Escherichia coli: relationship between flux and control at a branch point, efficiency of conversion to biomass, and excretion of acetate**. *Curr Top Cell Regul* 1986, **28**:69–105.
8. Luli GW, Strohl WR: **Comparison of growth, acetate production, and acetate inhibition of Escherichia coli strains in batch and fed-batch fermentations**. *Appl Environ Microbiol* 1990, **56**:1004–1011.
9. Zhou K, Qiao K, Edgar S, Stephanopoulos G: **Distributing a metabolic pathway among a microbial consortium enhances production of natural products**. *Nat Biotechnol* 2015, **33**:377–383.
10. Schulz TJ, Thierbach R, Voigt A, Drewes G, Mietzner B, Steinberg P, Pfeiffer AFH, Ristow M: **Induction of oxidative metabolism by mitochondrial frataxin inhibits cancer growth: Otto Warburg revisited**. *J Biol Chem* 2006, **281**:977–981.
11. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E: **Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect**. *PLoS Comput Biol* 2011, **7**:e1002018.
12. Vazquez A, Liu J, Zhou Y, Oltvai ZN: **Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited**. *BMC Syst Biol* 2010, **4**.
13. Schuster S, Boley D, Möller P, Stark H, Kaleta C: **Mathematical models for explaining the Warburg effect: a review focussed on ATP and biomass production**. *Biochem Soc Trans* 2015, **43**:1187–1194.
14. Wolfe AJ: **The acetate switch**. *Microbiol Mol Biol Rev* 2005, **69**:12–50.
15. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics**. *Mol Syst Biol* 2009, **5**.
16. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
17. Enjalbert B, Millard P, Dinclaux M, Portais JC, Létisse F: **Acetate fluxes in Escherichia coli are determined by the thermodynamic control of the Pta-AckA pathway**. *Sci Rep* 2017, **7**:42135.
18. Anane E, López C DC, Neubauer P, Cruz Bournazou MN: **Modelling overflow metabolism in Escherichia coli by acetate cycling**. *Biochem Eng J* 2017, **125**:23–30.
19. Goel A, Wortel MT, Molenaar D, Teusink B: **Metabolic shifts: a fitness perspective for microbial cell factories**. *Biotechnol Lett* 2012, **34**:2147–2160.
20. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245–248.
21. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN:

Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *P Natl Acad Sci USA* 2007, **104**.
22. Vazquez A, Beg QK, Demenezes MA, Ernst J, Bar-Joseph Z, Barabasi AL, Boros LG, Oltvai ZN: **Impact of the solvent capacity constraint on E. coli metabolism**. *BMC Syst Biol* 2008, **2**.
23. Goelzer A, Fromion V, Scorletti G: **Cell design in bacteria as a convex optimization problem**. *Automatica* 2011, **47**:1210–1218.
24. Goelzer A, Fromion V: **Bacterial growth rate reflects a bottleneck in resource allocation**. *Biochim Biophys Acta (BBA)-General Subj* 2011, **1810**:978–988.
25. Goelzer A, Fromion V: **Resource allocation in living organisms**. *Biochem Soc Trans* 2017, **45**:945-952.
26. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**:693–693.
27. Zhuang K, Vemuri GN, Mahadevan R: **Economics of membrane occupancy and respiro-fermentation**. *Mol Syst Biol* 2011, **7**.
28. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
29. Woldringh CL, Binnerts JS, Mans A: **Variation in Escherichia coli buoyant density measured in Percoll gradients.** *J Bacteriol* 1981, **148**:58–63.
30. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR: **Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria**. *Mol Syst Biol* 2015, **11**.
31. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T: **Interdependence of cell growth and gene expression: origins and consequences**. *Science* 2010, **330**:1099–1102.
32. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang Y-P, Lenz P, Yan D, Hwa T: **Coordination of bacterial proteome with metabolism by cyclic AMP signalling**. *Nature* 2013, **500**:301–306.
33. Vazquez A, Oltvai ZN: **Macromolecular crowding explains overflow metabolism in cells**. *Sci Rep* 2016, **6**:31007.
34. Orth JD, Palsson BØ, Fleming RMT: **Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide**. *EcoSal Plus* 2010, **4**.
35. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0**. *Nat Protoc* 2011, **6**:1290.
36. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
37. Fischer E, Sauer U: **A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry Escherichia coli**. *J Biol Chem* 2003, **278**:46446–46451.
38. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R: **Glycolytic strategy as a tradeoff between energy yield and protein cost**. *Proc Natl Acad Sci* 2013, **110**:10039–10044.
39. Fuhrer T, Fischer E, Sauer U: **Experimental identification and quantification of glucose metabolism in seven bacterial species**. *J Bacteriol* 2005, **187**.
40. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E: **The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli**. *J Biol Chem* 2004, **279**:6613–6619.
41. Nanchen A, Schicker A, Sauer U: **Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli**. *Appl Environ Microbiol* 2006, **72**:1164–1172.
42. Folsom JP, Carlson RP: **Physiological, biomass elemental composition and proteomic**

analyses of Escherichia coli ammonium-limited chemostat growth, and comparison with iron-and glucose-limited chemostat growth. *Microbiology* 2015, **161**:1659–1670.
43. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**.
44. Kayser A, Weber J, Hecht V, Rinas U: **Metabolic flux analysis of Escherichia coli in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state**. *Microbiology* 2005, **151**:693–706.
45. van Hoek MJ, Merks RM: **Redox balance is key to explaining full vs. partial switching to low-yield metabolism**. *BMC Syst Biol* 2012, **6**:22.
46. Goel A, Eckhardt TH, Puri P, Jong A, dos Santos F, Giera M, Fusetti F, Vos WM, Kok J, Poolman B, et al.: **Protein costs do not explain evolution of metabolic strategies and regulation of ribosomal content: does protein investment explain an anaerobic bacterial Crabtree effect?** *Mol Microbiol* 2015, **97**:77–92.
47. Fendt S-M, Buescher JM, Rudroff F, Picotti P, Zamboni N, Sauer U: **Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity**. *Mol Syst Biol* 2010, **6**:356.
48. Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD: **Systems-level analysis of mechanisms regulating yeast metabolic flux**. *Science* 2016, **354**:aaf2786.

Chapter 4 Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant *Escherichia coli*

Summary

This chapter presents the following published article:

Zeng, H, Yang, A. Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant *Escherichia coli*. *Biotechnology and Bioengineering*. 2019; 116: 1484– 1495. https://doi.org/10.1002/bit.26943.

Among all authors contributed to the published article, H.Z. and A.Y. designed the research; H.Z. performed simulation and data analysis; H.Z. wrote the manuscript; A.Y. revised the manuscript. The supplementary information of this chapter can be found in Appendix II.

Previously in Chapter 3, we have shown that a proteome allocation constraint enables the prediction of the steady-state metabolic shift from normal growth to the overflow metabolism for wild type *E. coli* using a core metabolic model, which comprises 95 reactions of the central carbon metabolism, including primarily the glycolysis, pentose phosphate pathway, pyruvate metabolism, citric acid cycle, oxidative phosphorylation and glutamate metabolism. Chapter 4 moves onto investigating the applicability of proteome allocation in recombinant strains grown in a batch culture via genome-scale models comprising 2382 metabolic reactions. One important reason of using a genome-scale metabolic model is that the core model does not include reactions describing the amino acid metabolism, which is essential for studying the impact of producing recombinant proteins on the maximum growth rate and metabolic flux distributions. The genome-scale model also includes detailed reactions describing the nucleotide metabolism, cofactor

biosynthesis, glycerophospholipid metabolism, cell envelope biosynthesis and alternate carbon metabolism, which can provide insights into the metabolic configurations and expand the scope of growth conditions from glucose minimal media to cultures with mixed carbon sources. Therefore, instead of using a simple (core) metabolic model, in this chapter we are motivated to develop a genome-scale resource allocation model for a wider landscape of biotechnology applications. The specific objective of this chapter is to establish computational tools to predict the protein productivity (defined by the concentration of the target protein at the end of a batch culture), cell growth and by-product formation to facilitate the design of microbial cell factories.

In this chapter, we focus on two engineered strains, i.e. green fluorescence protein (GFP) recombinant *E. coli* W3110 and maltose binding protein–glucose isomerase fusion protein (GI-malE) recombinant *E. coli* XL1. Synthesis reactions of GFP and GI-malE were added to the stoichiometric matrix of *E. coli* genome-scale metabolic model *i*AF1260. To model the impact of heterologous gene expression on the host cell, we introduced two different types of burdens – metabolic and proteomic. The metabolic burden referred to the increased maintenance requirement (ATPM) upon the synthesis of target proteins. The proteomic burden denotes the fraction of proteome occupied by the recombinant proteins, which was modelled by adding a recombinant sector in the proteome allocation constraint. The dynamic change of the extracellular conditions in a batch process was depicted by a set of ordinary differential equations (ODEs) in the framework of dynamic flux balance analysis (DFBA). Batch simulations were used to explore the optimal expression level of the target gene that confers the highest target protein productivity (in millimolar).

Simulation results suggested that the expression of GFP imposes a significant impact on the cellular maintenance, which possibly results from plasmid maintenance, protein folding and secretion, degradation of misfolded proteins and other disturbance associated

with the production and/or presence of the recombinant GFP (see Section 4.1). More specifically, the ATPM of GFP-bearing W3110 was predicted to be four times of the wild type strain whereas the estimated ATPM between induced and uninduced GI-malE XL1 only increased by about 2%. On the other hand, the calculated proteomic burden of the GI-malE-induced XL1 was five times higher than GFP-producing W3110, which could be explained by the higher molecular weight of GI-malE (87.2 kDa) than that of GFP (27 kDa). One important conclusion inferred from the results is that the expression of heterologous genes can retard the growth of the host cell growth through distinct mechanisms. Besides, by plotting the GFP concentration (at the end of the batch) against different GFP expression levels, we predicted that the highest GFP productivity would emerge at a moderate expression level. Such non-monotonic correlation suggests that the optimal design of the cell factory would require a delicate tradeoff between the expression of the synthetic circuit and the burden imposed on the host cell.

Future model refinement, in particular finding the optimal target-gene expression level requires additional experimental work to quantify the energy demand at different expression strengths of target products. Besides, the proteomic burden can be validated by comparing the proteomic dataset between wild type and recombinant protein-producing strains. Overall, Chapter 4 offers a useful modelling tool to analyse the growth and metabolism for (but not limited to) recombinant *E. coli* and a basis towards more efficient design of cell factories.

4.1 Introduction

Expression of recombinant proteins in microbes or higher organisms has been considered as an efficient approach for protein production [1–3]. Among various host cells, *Escherichia coli* is the most widely used organism due to its well-characterized genetics and physiology, ability to achieve high cell density, fast growth and high-yield nature [4–7]. However, the production of a foreign protein often triggers local (reaction-specific) and global (system-level) cell stress responses that lead to growth retardation, production of undesirable/toxic by-products such as acetic acid (known as overflow metabolism) and eventually low productivity of targeted proteins [8–10]. Such stress responses are considered to originate from the increased cellular metabolic burden associated with plasmid maintenance [11], protein folding, secretion and degradation of misfolded protein [12–14]. In particular, metabolic burdens often manifest as the increase in energy demand or maintenance energy requirement [15–17]. Furthermore, a recent study shows that the formation of acetate (as an example of the overflow metabolism) in *E. coli* is a result of the coordination of energy demand with carbon influx given constrained proteomic resources; the overproduction of 'useless' proteins amplifies such phenomenon by reducing the proteome fraction available for energy biogenesis and biomass synthesis [18].

In this chapter, we refer to the proteome fraction occupied by the expression of recombinant protein as the proteomic burden, and attempt to use a constraint-based model (CBM) [19] to investigate quantitatively the role of proteomic and metabolic burdens in the observed growth deterioration and acetate formation associated with recombinant protein production. Several studies using CBMs have previously attempted to predict the impact of gene mutations (gene knockout or overexpression) on the improvement of recombinant protein production for *Pichia pastoris* [20–22], *Synechococcus elongatus* [23] and *Lactococcus lactis* [24]. The manipulations with *in silico* determined targeted genes

were tested experimentally to verify the effectiveness of the model prediction. In addition to bacteria and algae, CBMs has been applied to recombinant eukaryotic cell hosts for protein production, where genome-scale metabolic models (GEMs) were used to identify targets for genetic modification, improve cellular metabolic capabilities, design media supplementation, and interpret high-throughput omics data [25]. Another relevant work [26] used a CBM for recombinant *E. coli* which fixed the growth rate, substrate uptake rate and acetate production rate to the experimentally measured values and maximized recombinant protein production, to study the possible changes and preference of the central metabolic pathways. In all these past studies, the way how the production of recombinant proteins is incorporated in a CBM is typically by the insertion of the protein synthesis reaction or a set of heterologous pathways associated with the protein formation, substrate transport and energy generation, to account for the mass and energy consumption.

To our knowledge, there is currently a lack of work in the quantitative prediction of the observed reduced growth and increased acetate formation in recombinant protein production via CBMs. In this study, we use flux balance analysis (FBA) and dynamic flux balance analysis (DFBA) to investigate the important contributors in predicting such growth physiology of two recombinant *E. coli* strains, namely green fluorescence protein (GFP) recombinant *E. coli* W3110 and maltose binding protein-glucose isomerase (GI-malE) fusion recombinant *E. coli* XL1. Metabolic burdens are quantified by (i) the local effect due to the extra mass and energy requirements in the synthesis of foreign proteins and (ii) the global effect on the increased maintenance energy caused by plasmid maintenance, protein secretion, folding, degradation and other possible disturbance associated with the recombinant proteins. Furthermore, the recently proposed proteome allocation theory (PAT) [18] is integrated into the stoichiometry model to investigate the role of proteomic burden. This approach shares the same principle of cellular resource

allocation [27] with several existing constraint-based models including FBAwMC [28,29], RBA [30–33], ME-Model [34] and its extension [35]. The main target of these existing models is to quantitatively predict the maximum cellular growth rate, while the overflow metabolism has mostly been captured qualitatively. The more recently presented model named CAFBA [36] is able to accurately predict both cell growth and acetate production. However, this and some of the earlier models (e.g. FBAwMC) formulate their resource allocation constraint by referring to the costs of individual metabolic reactions, which in principle would mean that a large number of parameters are to be introduced. Building on the PAT proposed by Basan *et al.* which depicts a constraint of proteomic allocation between two energy pathways and the biomass synthesis sector (as opposed to individual reactions), the current work is aimed to adopt a modelling approach that involves a small number of parameters which potentially can be estimated by fitting the model to experimental data. Overall, we aim to predict the observed growth retardation and elevated acetate production for recombinant *E. coli* using such a practical FBA/DFBA framework, which in the future may be developed as a predictive tool to guide metabolic engineering decisions.

4.2 Methods

4.2.1 Representing proteomic and metabolic burdens

4.2.1.1 The proteomic constraint

Based on the recently developed theory which suggests that the overflow metabolism in *Escherichia coli* originates from the discrepancy in the proteomic efficiency between different energy biogenesis pathways, namely (oxidative) fermentation and respiration [18] and a subsequent modelling framework termed Constrained Allocation Flux Balance Analysis [36], we introduce the following equation to represent the constraint on the utilization of proteomic resources:

$$w_f v_f + w_r v_r + b\lambda \leq \phi_{max} \quad \text{Equation 4-1}$$

where λ is the growth rate; v_f (v_r) is a flux representing the overall activity of the fermentation (respiration) pathway. Specifically, we choose the enzymatic reaction enolase (ENO) located in the lower part of the glycolysis to be v_f, while specifying the entrance flux of the tricarboxylic acid (TCA) cycle citrate synthase (CS) as v_r. The change in the representative fluxes (ACKr-AKGDH in Chapter 3 whereas ENO-CS here) primarily results from the increased scale of the metabolic model, i.e. a core model in Chapter 3 and a genome-scale model here. ENO and CS have been selected in this chapter since they are able to correctly capture the increasing trend of fermentation and the decreasing trend of respiration in the overflow region in a genome-scale metabolic model (Section 4.2.2.1). It is worth noting that, rigorously speaking, these two fluxes do not correspond to 'pure' fermentation or respiration fluxes, although in the case of ENO it contains predominantly the fermentation flux in the overflow region. Conceptually, a flux in the TCA cycle after the drawing of metabolites for biomass synthesis could be a better indicator than CS on the respiration side. In fact, replacing CS with AKGDH (α-ketoglutarate dehydrogenase), which is one of such fluxes, was tested in this chapter, which showed results comparable to those from using CS (see Appendix II, Supplementary Text and AppxII-Figures S1-S3). In the rest of this chapter, we present results obtained with the ENO-CS combination. ϕ_{max} is the maximum fraction of proteome attainable to energy biogenesis sectors (fermentation and respiration) and biomass synthesis, previously determined to be approximately 0.484 [37]. Compared to the protein-constrained GEMs that consider enzymatic activities at the level of individual intracellular reactions [38], Equation 4-1 is much less detailed and offers a coarse-grain constraint at the pathway level. On the other hand, it directly reflects the difference in proteomic efficiencies of the two energy pathways. Besides, it contains a

small number of parameters, which, as shown below, can easily be estimated from cell culture data.

For simplicity, both sides of Equation 4-1 is divided by ϕ_{max} to give the normalized form of the proteomic constraint.

$$w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 \qquad \text{Equation 4-2}$$

where $w_f^* \equiv w_f/\phi_{max}$, $w_r^* \equiv w_r/\phi_{max}$ and $b^* \equiv b/\phi_{max}$. Equation 4-2 does not contain the influence of recombinant protein production and is referred to as the wild-type proteomic constraint in this study. w_f^*, w_r^* and b^* are referred to as the normalized proteomic cost parameters.

4.2.1.2 Introducing the recombinant protein sector into the proteomic constraint

The overproduction of 'useless' proteins has been shown to amplify the rate of acetate formation by reducing the fraction of proteome available for energy production and biomass synthesis [18]. Here, we consider the production of a foreign protein as an additional sector that consumes the cellular proteomic resources, and add an extra term to the wild-type proteomic constraint:

$$w_f^* v_f + w_r^* v_r + b^* \lambda + \phi_{recP}^* \leq 1 \qquad \text{Equation 4-3}$$

where $\phi_{recP}^* \equiv \frac{\phi_{recP}}{\phi_{max}}$, ϕ_{recP} represents the fraction of proteome occupied for the synthesis and maintenance of the recombinant protein, considered as a (constant) property of the specific recombinant strain. It is evident that the presence of ϕ_{recP}^* means that the fraction of proteome accessible by biomass growth and energy pathways is reduced to $1 - \phi_{recP}^*$. Equation 4-3 is the revised proteomic constraint, with ϕ_{recP}^* being zero for the wild type and non-zero for recombinant strains.

4.2.1.3 Maintenance ATP as an indicator of the metabolic burden

In GEMs, maintenance ATP (ATPM) is generally defined as an energy 'drain' flux that accounts for the energy requirement resulted from non-growth-associated cellular activities [39,40]. In a GEM, it is a defined flux in addition to the growth-associated energy cost embedded in the biomass synthesis reaction, with lower/upper bounds. The level of ATPM for wild-type *E. coli* ranges from 3.15-8.39 mmol ATP gDW^{-1} h^{-1} subjected to the different versions of reconstruction and different strains (BiGG database, http://bigg.ucsd.edu/).

It is commonly accepted that the production of a foreign protein introduces an extra metabolic burden to the host cell [16,41]. Energetically, this metabolic burden manifests as the increase in cellular energy demand [15,42] or elevated maintenance energy [12,26]. An early study suggested that the synthesis of the recombinant protein alone would not be able to account fully for the extra demand of ATP [43]. Recent work validated this hypothesis by showing that the higher maintenance energy mainly results from the refolding and/or secretion stress triggered by the recombinant protein [12,13] or is due to the degradation of misfolded protein [14].

The aforementioned studies warrant the investigation on the impact of ATPM on the model prediction of the growth physiology of the recombinant *E. coli*. In a GEM, this entails treating ATPM as an adjustable parameter (as opposed to fixing it to the default value), in addition to the inclusion of the synthesis reaction of the recombinant protein into the GEM which accounts for the energy requirement of protein synthesis.

4.2.2 Constraint-based metabolic modelling

4.2.2.1 The genome-scale model

Genome-scale metabolic model *i*AF1260 [40] was used for all simulations in this study. In the adopted model, the hydrogen production reaction FHL and the oxidative stress

response reactions CAT, SPODM and SPODMpp were closed to prevent infeasible state under aerobic-glucose conditions [40]. The glucose dehydrogenase reaction (GLCDpp) was switched off since it is only active if pyrroloquinoline quinone (PQQ) is supplied to the system (see EcoCyc [44] entry on this enzyme). The D-glucose (galactose):proton symport reaction (GLCt2pp) was also considered to be off as it is functional when 2-deoxy-D-galactose is the specific substrate (see MetaCyc [45] entry on this enzyme). Reactions describing the synthesis of two recombinant proteins, GFP and GI-malE fusion protein were added to the original stoichiometry model. The protein reactions comprise the amino acid building blocks as well as the energy required for protein synthesis. Further details are provided in Appendix II, Supplementary Text and AppxII-Table S1.

4.2.2.2 Flux balance analysis

Flux balance analysis (FBA) [46] was used to determine the optimal flux distribution under different growth conditions, with a set of constraints:

$$\begin{aligned} &\text{Maximise} \quad f_{obj} \\ &\text{Subject to} \quad \text{(i)}\ S\boldsymbol{v} = 0 \qquad (9) \\ &\qquad\qquad \text{(ii)}\ v^L \le \boldsymbol{v} \le v^U \\ &\qquad\qquad \text{(iii)}\ w_f^* v_f + w_r^* v_r + b^* \lambda + \phi_{recP}^* \le 1 \end{aligned} \qquad \text{Equation 4-4}$$

where f_{obj} is the assumed cellular objective. S is the stoichiometric matrix; $\boldsymbol{v}$ is a column vector comprising the reactions/fluxes included in the (modified) *i*AF1260 model. v^L and v^U represent the minimum and maximum attainable fluxes, respectively. λ denotes the growth rate; w_f^* , w_r^*, b^* and ϕ_{recP}^* are the (normalized) proteomic cost parameters as introduced earlier. v_f is a single flux (in this case, ENO) chosen to represent the overall activity of the fermentation pathway and v_r is the representative flux of the respiration

pathway (in this case, CS). Together, v_f and v_r indicate the cellular modulation of energy flux (fermentation vs. respiration).

The prediction of the extent of overflow metabolism (rates of acetate production) was determined by applying a proper set of values for w_f^* , w_r^*, b^* and ϕ_{recP}^* to the third constraint of Equation 4-4. Implementation details for introducing the proteomic constraint into the FBA model are given in Appendix II, Supplementary Text. The proteomic constraint-integrated FBA was run under aerobic-glucose conditions, with the objective function set to either maximize the growth rate or minimize glucose consumption, depending on the nature of the input data available, as explained below. The optimal flux distribution was solved using the COBRA toolbox [47] with Gurobi 6.0 as the linear programming (LP) solver.

4.2.2.3 The dynamic framework

To facilitate the estimation of key model parameters using the literature data available in the form of records of batch cultures, dynamic flux balance analysis (DFBA) [48] was adopted to connect the dynamic batch growth behaviour with the intracellular flux distribution. The DFBA framework comprises two parts: the microscale cellular metabolic flux distribution (determined by FBA) and the change in the macroscale extracellular environment modelled by ordinary differential equations (ODEs). Dynamic simulation is achieved by the exchange of information between the two parts at a number of time intervals that divide the total simulated time duration: the FBA model is solved at the beginning of every time interval to give steady-state flux distribution; the values of the fluxes of interest are then passed to the ODEs to update the outer environment state. This procedure is repeated until the entire time duration has been simulated.

Maximizing biomass growth was taken as the objective function in all DFBA simulations. It is worth noting that sometimes the presence of synthetic circuits and/or the expression of recombinant genes can cause considerable burdens to the host cell, where the biomass growth might not be the actual objective. Here biomass growth was used as (i) no significant growth cessation/retardation were reported in the associated cell culture experiments and (ii) no strong evidence suggested that objectives other than maximal growth can better describe the studied cases. The glucose uptake rate v_{glc} was assumed to follow the Michaelis-Menten kinetics [49] for the simulation of GFP-bearing *E. coli* where the data of glucose concentration during a batch were available and were used to determine a proper set of kinetic parameters $v_{glc,max}$ and K_g in Equation 4-5; whereas in the GI-malE case, v_{glc} was directly set to a function of time (Equation 4-6) obtained from regression using the specific uptake rates during a batch which were directly available in the data source (see the derivation below).

$$v_{glc,GFP} = v_{glc,max} \frac{G}{K_g + G} \quad \text{Equation 4-5}$$

$$v_{glc,GI-malE} = f(t) \quad \text{Equation 4-6}$$

Calculating $v_{glc,GI-malE}$

The experimental data of the specific glucose uptake rate of GI-malE recombinant XL1 is in gram glucose consumed per gram dry cell weight (g glucose/gDW) versus time (h) [26]. To apply the experimental data in DFBA models, the unit of the experimental data needs to be converted to mmol glucose/gDW/h.

The change of biomass within the hourly interval is

$$dX_t = X_{t+1} - X_t \ \text{[gDW/L/h]} \quad \text{Equation 4-7}$$

Considering the mass balance of biomass over the hourly interval

$$X_{t+1} = \lambda_t X_t + X_t \text{ [gDW/L]} \quad \text{Equation 4-8}$$

Thus growth rate can be calculated

$$\lambda_t = \frac{dX_t}{X_t} \text{ [gDW/gDW/h]} \quad \text{Equation 4-9}$$

The dynamic profile of the glucose uptake rate is determined using λ_t

$$v_{glc,GI-malE} = 0.18 \times v_{glc,GI-malE,t,exp} \times \lambda_t \text{ [mmol/gDW/h]} \quad \text{Equation 4-10}$$

where the molecular weight of glucose is 0.18 g/mmol; $v_{glc,GI-malE,t,exp}$ is the experimentally measured specific glucose uptake rate in g glucose/gDW. The discrete dynamic profile of the calculated glucose uptake rate is then fitted to a polynomial function:

Uninduced

$$v_{glc,GI-malE} = -0.0006t^3 + 0.014t^2 - 1.0047t + 9.1722 \quad \text{Equation 4-11}$$

Induced

$$v_{glc,GI-malE} = 0.0205t^2 - 0.9235t + 10.3882 \quad \text{Equation 4-12}$$

Equation 4-11 and Equation 4-12 were used in DFBA for the estimation of the proteomic burden, ATPM and proteomic cost parameters for GI-malE recombinant XL1.

The dynamic extracellular environment was modelled by the following ODE system:

$$\frac{dX}{dt} = \lambda X \quad \text{Equation 4-13}$$

$$\frac{dG}{dt} = -v_{glc} X \quad \text{Equation 4-14}$$

$$\frac{dA}{dt} = v_{ac} X \quad \text{Equation 4-15}$$

$$\frac{dP}{dt} = v_{recP} X \quad \text{Equation 4-16}$$

where X, G, A and P are the concentrations of biomass, glucose, acetate and recombinant protein, respectively. The growth rate (λ), the acetate exchange flux (v_{ac}) was resolved via FBA while the recombinant protein synthesis flux (v_{recP}) was fixed to the experimentally measured values. Equation 4-13 to Equation 4-16 were solved by stiff ODE solver ode15s in MATLAB R2016a. Calculations of v_{recP} based on experimental data are shown below.

Calculating v_{recP} for W3110-GFP

Considering the mass balance of GFP over the hourly time interval

$$GFP_{t+1} = -v_{GFP}X_t + GFP_t \text{ [g/L]} \qquad \text{Equation 4-17}$$

The GFP production rate can be calculated, and a time profile was obtained by regression in a form below

$$v_{GFP} = \frac{dGFP_t}{X_t} = 0.001481e^{-0.3188t} + 0.0002253e^{0.007265t} \text{ [mmol/gDW/h]} \qquad \text{Equation 4-18}$$

where $dGFP_t = GFP_{t+1} - GFP_t$; the molecular weight of GFP is 27 kDa.

Calculating v_{recP} for XL1-GI-malE-induced

For the GI-malE-producing *E. coli* XL1, the original data for the expression of the recombinant protein is in GI activity (unit protein activity per litre) versus time. The intensity of the protein activity is proportional to the protein concentration. For the comparison between experimental data and the simulation results, we need to convert protein activity per litre to gram protein per litre.

We first calculated the deviation of protein activity

$$dA_{GI,t} = A_{GI,t+1} - A_{GI,t} \text{ [unit protein activity/L/h]} \qquad \text{Equation 4-19}$$

Subsequently, we calculated the rate of accumulated protein activity (unit activity/gDW/h)

$$v_{GI\ activity} = \frac{dA_{GI,t}}{X_t} \text{ [unit protein activity/gDW/h]} \qquad \text{Equation 4-20}$$

Then we took the arithmetic mean of $v_{GI\ activity}$, i.e. $\bar{v}_{GI\ activity}$, and assumed that $\bar{v}_{GI\ activity}$ is proportional to the specific protein production rate (q_{GI})

$$\bar{v}_{GI\ activity} = p \times q_{GI} \quad \text{Equation 4-21}$$

The constant $p = 3.47 \times 10^{-7}$(mmol protein/unit protein activity) links the protein activity with the molar/mass concentration.

$$GI = p \times A_{GI}\ \text{[mmol/L]} \quad \text{Equation 4-22}$$

Following a polynomial regression of $v_{GI\ activity}$, the estimated GI-malE production rate is

$$v_{GI-malE} = p \times v_{GI\ activity} = -1.476 \times 10^{-8}t^3 + 7.99 \times 10^{-7}t^2 - 1.454 \times 10^{-5}t + 9.548 \times 10^{-5}\ \text{[mmol/gDW/h]} \quad \text{Equation 4-23}$$

where GI is the estimated experimental concentration of GI-malE protein; A_{GI} is the protein activity directly obtained from the data source; $v_{GI\ activity}$ is the estimated rate of accumulated protein activity.

4.2.2.4 Estimation of proteomic cost parameters and ATPM

To obtain the proteomic cost parameters w_f^*, w_r^* and b^* for wild-type *E. coli* W3110, given the fact that the original experimental data was in the form of steady-state rates of acetate production at different growth rates, we fixed the growth rates to the experimental values and used the minimization of the glucose uptake rate as the objective function for FBA. An optimization solver (the genetic algorithm toolbox in MATLAB) coupled with FBA was used to find the optimum values of w_f^*, w_r^* and b^* that give the best fit between predicted acetate production rates and the corresponding experimental measurements at different growth rates. ATPM was assumed to remain as the default value (obtained from *E. coli* MG1655) because MG1655 and W3110 are genetically very similar, i.e. only eight sites of the 267 identified conflict sequences between their genomes have shown true differences

[50,51]. Next, the calculated values of model parameters were taken into the DFBA model with the revised proteomic constraint (Equation 4-3) assuming that the wild-type and the recombinant strain share the identical set of w_f^*, w_r^* and b^*. The basis for this assumption is that (i) w_f^* and w_r^* represent the intrinsic proteomic cost (or enzymatic requirement) of fermentation and respiration reactions, which we considered unchanged between the wild type and the recombinant type, and (ii) b^* represents the proteome demand of biomass synthesis, which should not change if the assumed composition of biomass reaction does not vary under different conditions. The DFBA model was capable of simulating the batch behaviour, hence allowing the simulation results to be compared with the experimentally measured batch data. The optimization solver was now coupled with DFBA to determine the best values of ϕ_{recP}^* and (adjusted) ATPM for the GFP-bearing W3110 by minimizing the difference between predicted and measured batch data, including the accumulated concentrations of biomass, acetate, glucose and recombinant protein versus culturing time. During the analysis, the glucose uptake rate was determined by Michaelis-Menten kinetics (Equation 4-5 and Table 4-1); recombinant protein production rate (v_{recP}) was fixed to the experimentally measured values. The cellular objective was set to maximizing growth rate.

In the case of GI-malE-bearing *E. coli*, batch experimental data was available for both uninduced and induced strains. Therefore, the DFBA model was adopted for both cases, coupled with the optimization solver for parameter estimation, following the same approach as for the case of GFP-bearing W3110. Firstly, w_f^* , w_r^*, b^* were estimated using the batch data of the uninduced strain. As we did not find evidence suggesting wild type MG1655 and XL1 possess a similar level of ATPM, nor the value of maintenance energy has been reported specifically for XL1, ATPM was co-estimated with the proteomic cost parameters in this case. The lower and upper bound of the estimated ATPM were set to 3.15 and 8.39 referring to the known range of maintenance energy of various *E. coli* strains

(BiGG database, http://bigg.ucsd.edu/). Thereafter, estimated values of the proteomic cost parameters were taken into DFBA for the induced strain, which was coupled with the optimization solver to further estimate ϕ_{recP}^{*} and to re-estimate ATPM using the batch data of the induced strain. Similar to the GFP case, rates of glucose uptake and recombinant protein production were fixed to the experimental values; biomass growth maximization was the assumed cellular objective.

4.3 Results

Values of all the model parameters, either estimated in this chapter or adopted from the literature, are presented in Table 4-1.

Table 4-1 List of model parameters used in simulations

Parameters	**W3110-wt**	**W3110-GFP**	**XL1-GI-malE-uninduced**	**XL1-GI-malE-induced**
w_f^*	0.05775[a]	Same as wt	0.08134[a]	Same as uninduced type
w_r^*	0.1847[a]		0.7900[a]	
b^*	0.1[a]		0.4546[a]	
ϕ_{recP}^*	-	0.1333[a]	-	0.5033[a]
ϕ_{GFP}	-	0.06454[a]	-	-
ϕ_{GI}	-	-	-	0.2436[a]
ϕ_{max}	0.484[b]	0.484[b]	0.484[b]	0.484[b]
ATPM (mmol/gDW/h)	8.39[c]	32.6[a]	3.670[a]	3.759[a]
$v_{glc,max}$ (mmol/gDW/h)	-	4.690	-	-
K_g (g/L)	-	1	-	-

Note: ATPM: adjustable maintenance energy level; GI-malE: maltose binding protein–glucose isomerase; wt: wildtype.
[a]Estimated in this study.
[b]Obtained from Scott *et al*. (2010).
[c]Default value in *i*AF1260.

4.3.1 Modelling wild-type and GFP-bearing recombinant E. coli W3110

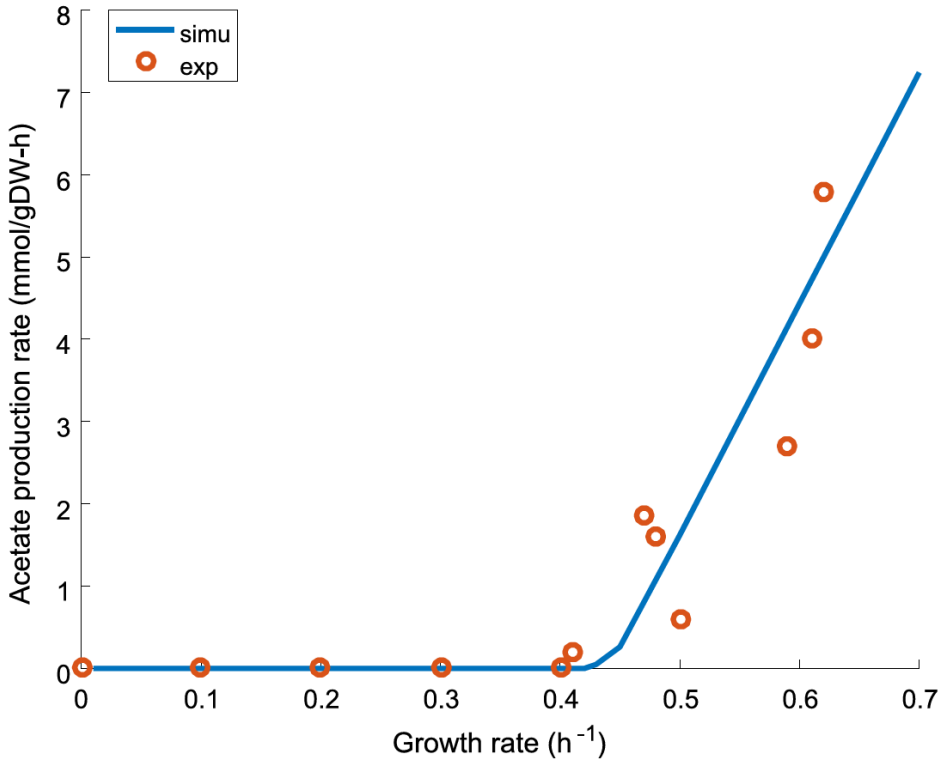


Figure 4-1 Comparison between model prediction and experimental data of the acetate production for wild-type *E. coli* W3110 (W3110-wt). Experimental data were obtained from figure 3 of Mori *et al*. (2016) which were originally reported in Nanchen *et al*. (2006) and Vemuri *et al*. (2006) for wild-type *E. coli* MG1655. W3110 was assumed to possess the same overflow pattern with a close strain, MG1655. exp: experimental data; simu: simulation results.

The first case study is based on an experimental work studying the expression of green fluorescent protein (GFP) in recombinant *E. coli*, with the parental strain being W3110 [52]. To model the growth physiology of the GFP–bearing recombinant *E. coli*, we started by finding the proteomic cost parameters for the wild-type W3110 (W3110-wt). The best experimental datasets we found were from [53,54]. The original data can be readily converted to steady-state rates of acetate production versus growth rates [36], which were subsequently used as the input of FBA for the estimation of wild-type proteomic cost parameters. It is worth noting that the corresponding bacterial strain of the experimental data available was actually *E. coli* MG1655. However, MG1655 and W3110 are both *E.*

coli K-12 derived wild-type strains and are genetically similar [50,51] thus we used the growth data obtained from MG1655 to approximate the behaviour of W3110-wt. Figure 4-1 shows that with the wild-type proteomic constraint in place, the predicted rates of acetate production are in good agreement with the experimental data.

We then move on to the targeted GFP-bearing recombinant *E. coli* W3110, denoted as W3110-GFP. DFBA was used to simulate the batch growth of W3110-GFP. The biomass concentration at the beginning of the cell culture was not reported [52]. As the dynamics of a batch system can be sensitive to the initial biomass concentration, the starting time of the model simulation was set to t=2 hour for a more accurate reading (from the biomass curve in the data source) of the initial conditions.

It is worth noting that, the GFP synthesis reaction embedded in the GEM already includes the mass and energy needed for protein synthesis. What we considered in this section, i.e. ϕ^*_{recP} and ATPM, are the additional proteomic and metabolic burdens due to the production of the recombinant protein. As stated earlier (see Section 4.2.1), the production of recombinant protein can provoke perturbations in (i) cellular proteome (manifested by adding ϕ^*_{recP} to the wild type proteomic constraint) and (ii) the cellular maintenance energy (modification in ATPM).

To gauge the individual importance and sufficiency of each of the two parameters in predicting the growth physiology of the recombinant strain, we first fixed the ATPM to the default value (8.39 mmol ATP gDW^{-1} h^{-1}) and estimated ϕ^*_{recP} (w_f^*, w_r^* and b^* were kept identical to the wild-type) using the experimental batch growth data of W3110-GFP coupled with DFBA. In a parallel test, we estimated ATPM alone without adding ϕ^*_{recP} to the proteomic constraint. Results of these two tests are shown in Figure 4-2(a-d) and (e-h), respectively.

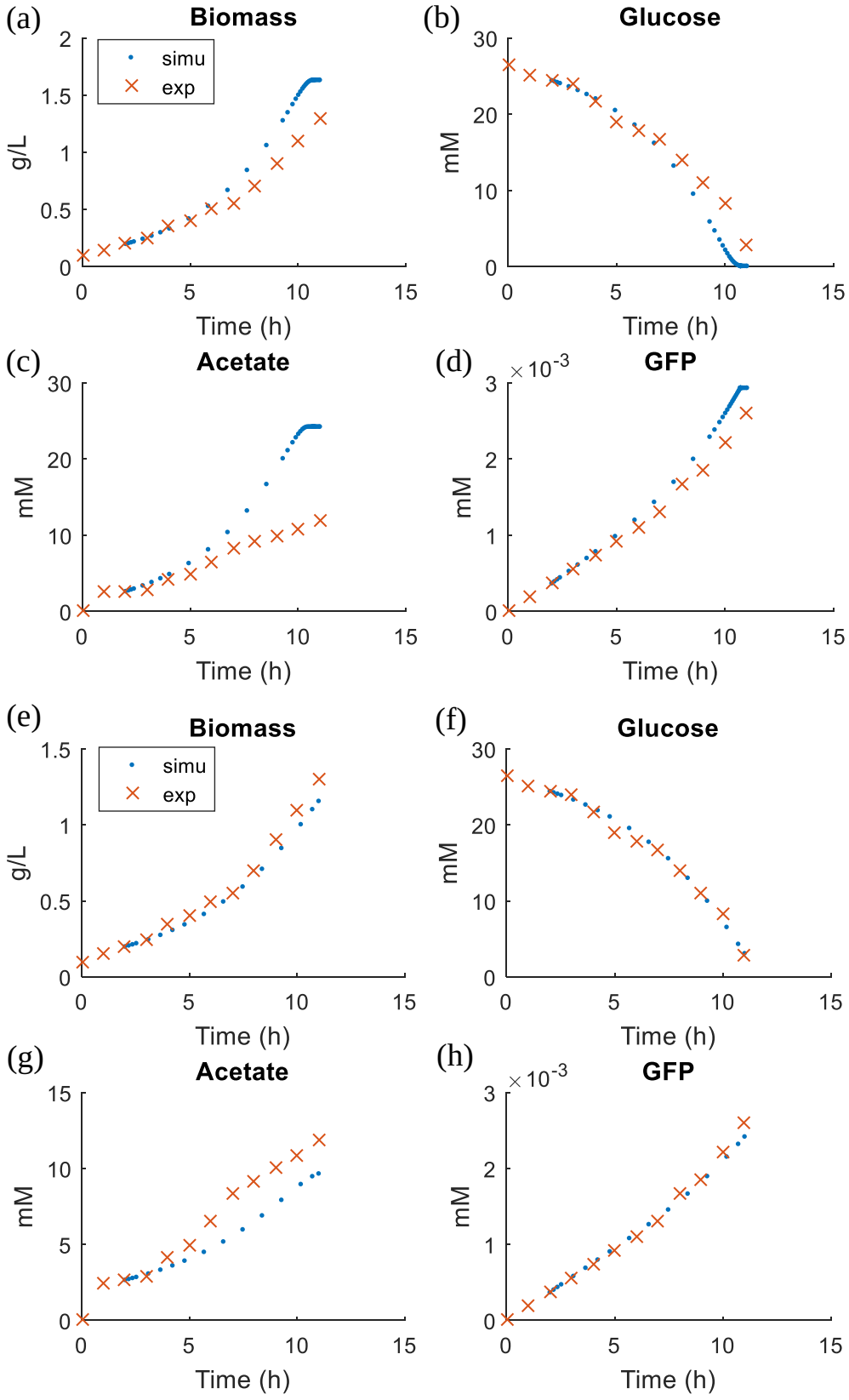


Figure 4-2 Comparison between model prediction and experimental data of the growth physiology of a batch culture of GFP-bearing *E. coli* W3110 (W3110-GFP). **(a-d)** Model with estimated ϕ^*_{recP} alone, best fit $\phi^*_{recP} = 0.6266$. **(e-h)** Model with changing ATPM alone, best fit ATPM = 40

mmol/gDW/h. Experimental data were obtained from figure 3a-d in Lara *et al.* (2006). ATPM: adjustable maintenance energy level; GFP: green-fluorescence protein.

Although adjusting ATPM alone performed slightly better than merely adopting ϕ^*_{recP}, neither of the two individual parameters could render satisfactory model prediction.

Next, we estimated ϕ^*_{recP} and ATPM simultaneously to see how the combination of the two perturbations affects the accuracy of the model. Sensitivity analysis was carried out by DFBA simulations of the batch growth of W3110-GFP with (i) ϕ^*_{recP} fixed to 0.1333 and ATPM set to +/- 50% of 32.6 mmol/gDW/h and (ii) with ATPM fixed to 32.6 mmol/gDW/h and ϕ^*_{recP} set to +/- 50% of 0.1333.

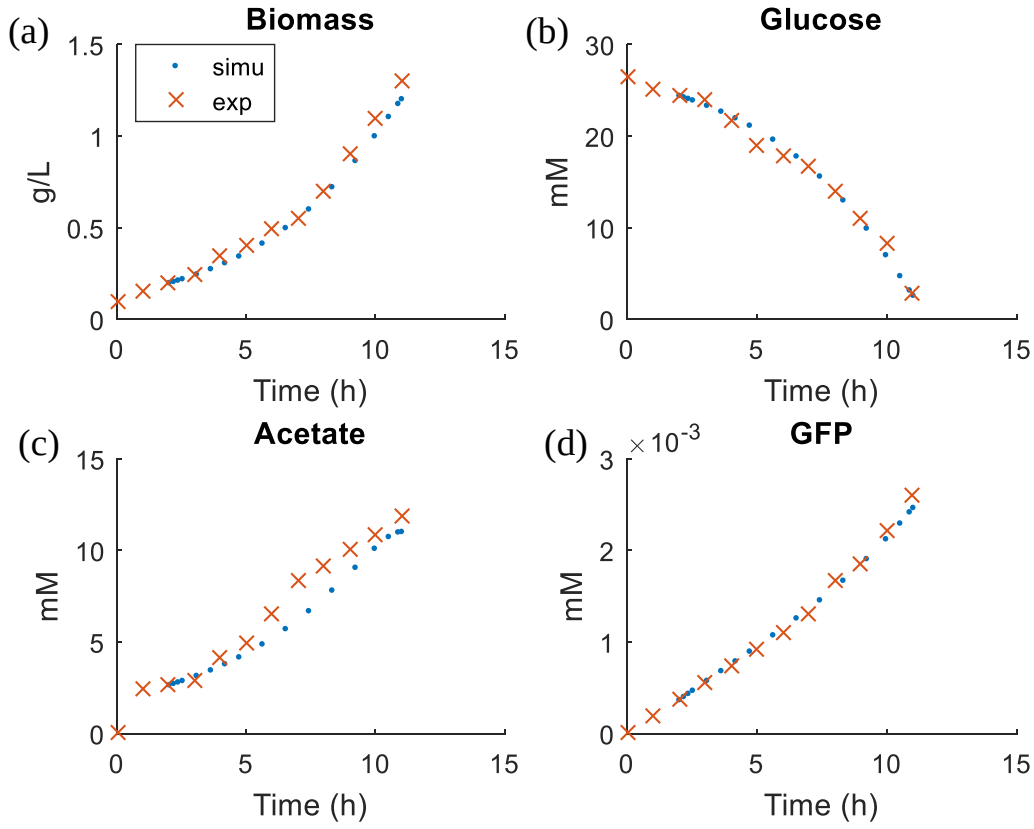


Figure 4-3 Comparison between model prediction and experimental data of the growth physiology of a batch culture of GFP-bearing *E. coli* W3110 (W3110-GFP). **(a-d)** Concentrations of biomass, glucose, acetate and GFP during batch fermentation. Both ϕ^*_{recP} and ATPM were adopted to the model. The estimated values of ϕ^*_{recP} and ATPM used in simulations are given in Table 4-1. Experimental data were obtained from figure 3(a-d) in Lara *et al.* (2006).

Sensitivity analysis confirms that the predictions of biomass growth and metabolites production are sensitive to both parameters (AppxII-Figure S4). This time, with a $\phi_{recP}^{*} = 0.133$ and an approximate three times increase in the default ATPM (see Table 4-1), model predictions of the accumulated concentrations of biomass, acetate, glucose and recombinant protein for W3110-GFP agreed well with the experimental data (Figure 4-3(a-d)). This result indicates that the dual usage of the proteomic constraint and the updated ATPM helps improve the prediction of the growth physiology of W3110-GFP using CBMs. It is worth noting that, the estimated ATPM for W3110-GFP (32.6 mmol ATP/gDW/h) is in-line the range of the values reported previously for other recombinant *E.coli* strains (40.2-45.3 mmol ATP/gDW/h in ref. [15] and 20 mmol ATP/gDW/h in ref. [55]).

4.3.2 Modelling GI-malE-bearing recombinant *E. coli* XL1

The second case tested is on the production of GI-malE fusion protein expressed in recombinant *E. coli* XL1 [26], referred to as XL1-GI-malE. Similar to the GFP case, we started with determining w_f^*, w_r^* and b^* for the control strain XL1-GI-malE-uninduced (no protein production) via DFBA coupled with experimentally obtained dynamic batch growth data. As there was no basis for assuming the validity of the default ATPM (obtained from MG1655) still stands for XL1, it was co-estimated with the proteomic cost parameters. As for the induced XL1 (induction occurred at 10h; GI-malE production began at 10h), we estimated ϕ_{recP}^{*} and ATPM simultaneously with the lower bound of ATPM set to the value of the uninduced strain. Following the approach stated above, the sensitivity analysis was carried out by DFBA simulations of the batch growth of XL1-GI-malE-induced with (i) ϕ_{recP}^{*} fixed to 0.5033 and ATPM set to +/- 50% of 3.759 mmol/gDW/h and (ii) with ATPM fixed to 3.759 mmol/gDW/h and ϕ_{recP}^{*} set to +/- 50% of 0.5033, which confirms that the model predictions are sensitive to both ϕ_{recP}^{*} and ATPM (AppxII-

Figure S5). Interestingly, the estimated values of ATPM for induced and uninduced strains are rather close, indicating that the change of ATPM between protein production and no protein production strains in this second case plays a less significant role compared to the first case.

The DFBA simulation results of the batch growth of XL1 with estimated parameters are shown in Figure 4-4. The starting point of the simulation for the induced case was set to t=10 h corresponding to the induction time in the experiment. Besides, Figure 4-4(a) suggests that the growth of XL1-GI-malE-induced plateaus at around t=20 hour. The experimental data of the acetate production for XL1-GI-malE-induced is available for 5-15 hour range and 30-40 hour range [26]. As the model constructed in this chapter intends to investigate the acetate production in exponentially growing cells, the part of acetate data obtained in the stationary phase is beyond the scope of the current model and thus is omitted in Figure 4-4(b).

With the revised proteomic constraint in place, our model succeeded in capturing the overall growth physiology of both XL1-GI-malE-uninduced and XL1-GI-malE-induced. More specifically, model predictions of the biomass concentration and the production of the recombinant protein are in good agreement with the experimental results. As for acetate, simulation results are consistent with the first two experimental data points after the induction for XL1-GI-malE-induced; note that the original experimental dataset is lack of acetate data for the 15-30 h period. It should also be mentioned that the conversion of the original data available for this case (see the derivations of Equation 4-23 in Section 4.2.2.3) is far from straightforward to derive the dataset usable in FBA and/or DFBA, which might have undermined the reliability of the data that was eventually used for parameter estimation.

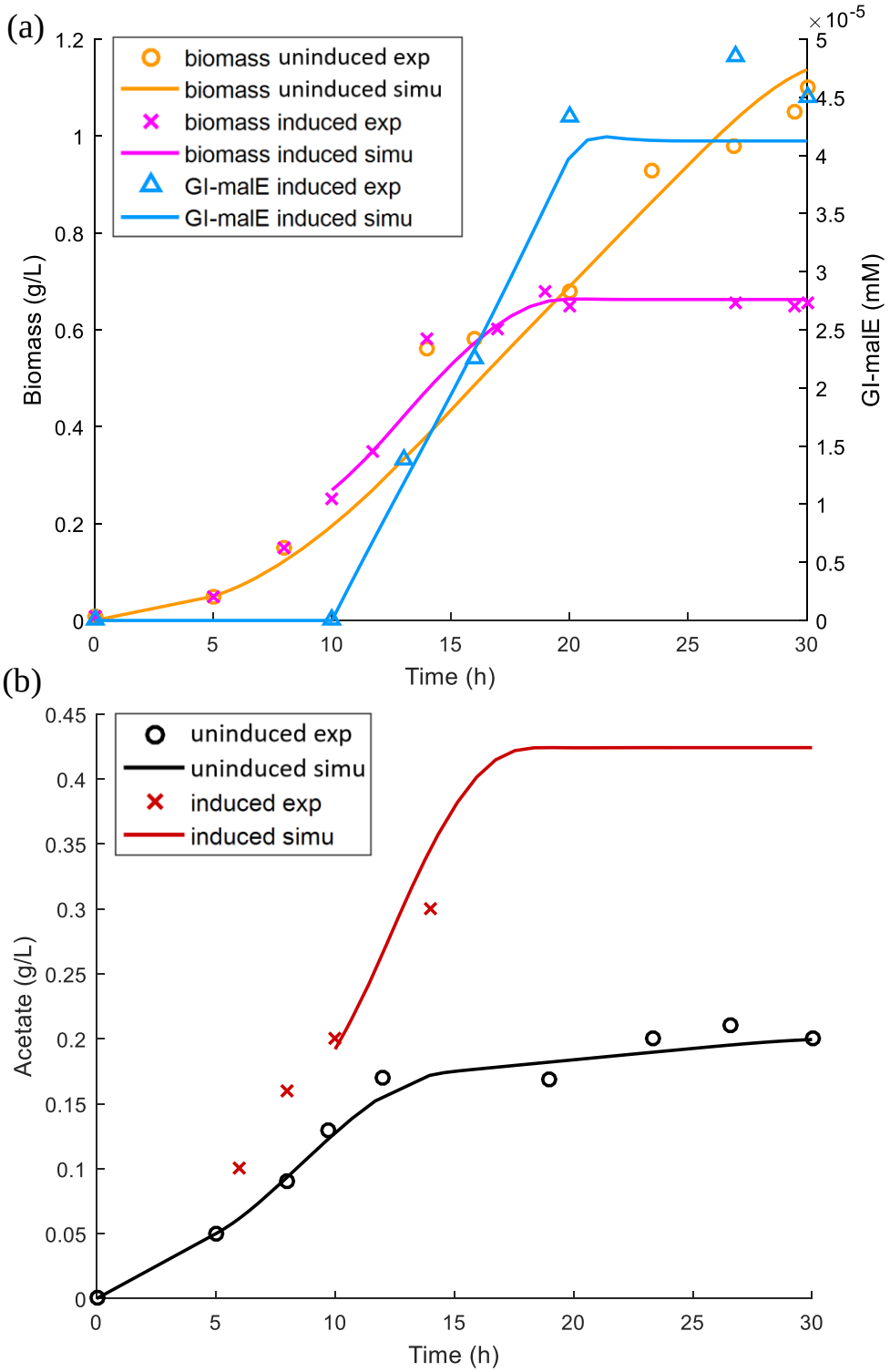


Figure 4-4 Comparison between model prediction and experimental data of the growth physiology of batch cultures of GI-malE-bearing recombinant *E. coli* XL1 (XL1-GI-malE-induced and -uninduced). **(a)** Change of biomass concentration for uninduced and induced strains and recombinant protein GI-malE production for the induced strain. **(b)** Change of acetate concentration for uninduced and induced strains. Experimental data shown in the figure were

derived from Özkan *et al.* (2005). GI-malE: maltose binding protein-glucose isomerase fusion protein.

Nevertheless, the results clearly show that the model adopted here was able to predict that significant changes in biomass growth and acetate excretion occur due to recombinant protein production.

4.4 Discussion

4.4.1 Comparison of proteomic cost parameters between the two cases

It is interesting to contrast the estimated proteomic cost parameters between the two recombinant strains, particularly b^* and ϕ^*_{recP}. While W3110 (wild-type) possesses a relatively low b^*, that of XL1 (uninduced) is about four times larger. In the Methods section, we have indicated that b^* represents the proteomic cost of biosynthesis pathways. From the definition and following a previous study [37], the higher the b^*, the lower efficiency in the synthesis of biomass building blocks and thus the lower the maximal growth rate the strain can achieve. This is consistent with the fact that the experimentally determined maximum growth rate of W3110 is higher than that of XL1. The estimated ϕ^*_{recP} of GI-malE-bearing XL1 is also higher than GFP-bearing W3110. As stated earlier, ϕ^*_{recP} denotes the portion of the total proteome occupied for the production of the recombinant protein. GI-malE fusion protein has a higher molecular weight of 87.2 kDa while the molecular weight of GFP is about 27 kDa, which at least partially explains the discrepancy between the estimated ϕ^*_{recP} of the two cases.

4.4.2 Effect of recombinant protein production on ATPM

A comparison between a recombinant strain with its wild-type (or the uninduced one) shows that ATPM variation due to the additional protein synthesis is strain-dependent. For W3110, the result of modelling shows that the production of GFP leads to the dramatic

increase in the maintenance energy (from 8.39 to 32.6 mmol ATP/gDW/h), whereas for XL1, no significant change in ATPM was found between the induced- and uninduced-case. This discrepancy leads to the hypothesis that the introduction of GFP to W3110 provokes strong perturbation in the metabolic network and at the same time causes a higher level of maintenance activities e.g. for sustaining cellular homeostasis. In contrast, the production of GI-malE fusion protein appears to severely disturb the host primarily through reducing the portion of proteome available to energy biogenesis and biomass synthesis, but with a rather minor change in the maintenance energy (as suggested by the large value of ϕ^*_{recP} and the small increase of ATPM, Table 4-1). A similar pattern was also observed in *LacZ*-overexpressing *E. coli*, where the overflow phenomenon was predicted accurately by considering only the proteomic disturbance, without altering the assumed cellular energy demand [18].

4.4.3 Towards a predictive tool to support engineering design for recombinant protein production

In this chapter, DFBA has been used to establish the impact of proteomic burden (ϕ^*_{recP}) and maintenance burden (ATPM) on the batch cultures of recombinant *E. coli*. We further envisage the potential of using our model as a predictive tool to guide metabolic engineering decisions, e.g. to boost the batch production of the recombinant protein of interest. As mentioned in the Introduction of this chapter, the growth physiology predicted by our model in the production of a foreign protein, including the reduction in cellular growth rate and the increase in the production of unwanted metabolites (acetate in particular) represents undesirable effects for the batch culture of recombinant strains, yet these effects typically accompany enhanced protein expression. This conflict points to the need for engineering optimization in both designs (e.g. selection of an optimal protein

expression level) and operation (e.g. determination of the timing of induction during a batch).

To this end, it is desirable to develop a predictive model to guide these engineering decisions. The DFBA modelling approach from this chapter is not yet able to fulfil this purpose, due to several important limitations: (i) with respect to the design of the recombinant strain, a connection is yet to be made between the engineering of the expression system and the attainable protein expression level (quantified by e.g. specific protein production rate or mass fraction of the recombinant protein), and it is still unclear how the latter is precisely related to the proteomic burden, denoted by ϕ^*_{recP}; (ii) the relationship between ATPM of a recombinant strain and its protein expression level is to be developed and (iii) similarly how the protein expression level would affect the substrate uptake kinetics is still unknown.

Nevertheless, as an illustration of the potential usage of a fully predictive model, we have run simulations for the production of GFP by GFP-recombinant *E.coli* as an example to show how in principle a future model could help find the optimal expression level of the recombinant protein, represented by ϕ^*_{recP} for the sake of this illustration. We assume that the glucose update kinetics remains the same as the case of W3110-GFP, and that ATPM either remains as a constant (taking the value estimated for W3110-GFP) or is linearly proportional to ϕ^*_{recP}. ϕ^*_{recP} was fixed to a specific level at each simulation with maximizing growth rate as the cellular objective. All simulations started at the same initial glucose concentration and were kept running until glucose was depleted. The final protein concentration was recorded for each batch simulation. Further details of the approach can be found in Appendix II, Supplementary Text (Equations AppxII-S5 to AppxII-S6). The result presented in Figure 4-5 shows that an optimal protein expression level could be

predicted, and the position of the optimal point can be affected by the variation of the extra maintenance burden (ATPM).

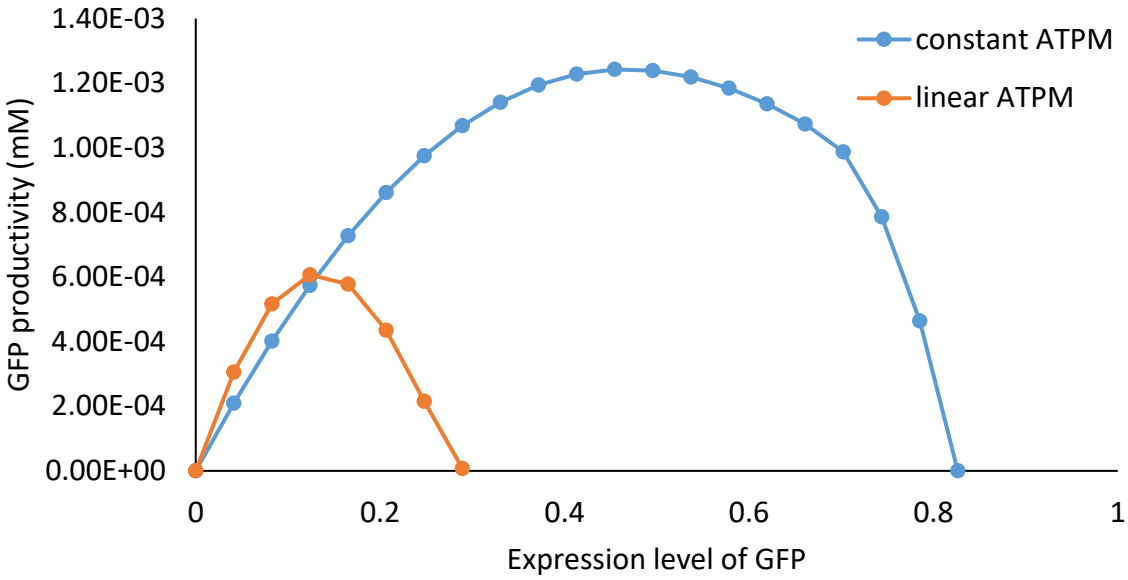


Figure 4-5 Illustrative model application for predicting the optimal expression level of recombinant proteins at constant maintenance energy and linear maintenance energy. GFP was taken as an example of the target protein. The X-axis shows the normalised expression level of the recombinant protein, denoted by ϕ^*_{recP}; Y-axis is the concentration of the recombinant protein at the end of the batch.

Furthermore, systematic optimisation of promoters and ribosome binding site (RBS) also plays an important role in the efficient expression of recombinant proteins. For instance, promoter and RBS engineering have been used to optimise the biosynthetic enzymes of the limonene biosynthesis pathway by designing and screening the libraries of RBS sequences and replacing the T7 promoter by the strong constitutive promoter FAB80 in engineered *E. coli* [56]. Besides, different expression levels of β-galactosidase were obtained in *Rhodococcus* using different natural promoters identified through transcriptome promoter screening combined with a superior RBS of *ami* determined using the RBS Calculator [57,58]. More importantly, a study of the composability of promoters and RBS reveals that a specific RNA and protein level can be achieved from various sequence-divergent promoter-RBS combinations [59]. It would be desirable to compare the burdens of

synthetic constructs with different promoter-RBS pairs that provide the same protein expression using time-efficient cell-free lysates analysis [60]. Promoter-RBS pairs with same expression strengths but lower burdens would be advantageous for optimal recombinant protein productions.

4.5 Conclusions

We have shown the effectiveness of the combined use of a proteome allocation constraint and adjustment of maintenance energy (ATPM) in predicting the retarded growth and the increased acetate production for recombinant *E. coli* via constraint-based models (CBMs). To our knowledge, this is the first time a CBM, and DFBA in particular, is used for capturing such observed growth physiology for recombinant strains. Through two test cases, the inclusion of the proteomic burden for synthesizing a recombinant protein appears to be essential for the model prediction, whereas the significance of ATPM adjustment may be strain-dependent. This work quantitatively depicts the proteomic and metabolic burdens of recombinant protein expression in terms of the proportion of the proteomic resource occupied by the extra protein production and the increase in maintenance energy, respectively. Further development of the current modelling framework to overcome several limitations has the potential to offer a starting point for the development of a practical, model-based tool, which complements other types of tools to guide design and operational decisions for efficient protein production.

References

1. Ahmad M, Hirz M, Pichler H, Schwab H: **Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production**. *Appl Microbiol Biotechnol* 2014, **98**:5301–5317.
2. Kim JY, Kim Y-G, Lee GM: **CHO cells in biotechnology for production of recombinant proteins: current state and further potential**. *Appl Microbiol Biotechnol* 2012, **93**:917–930.
3. Demain AL, Vaishnav P: **Production of recombinant proteins by microbes and higher organisms**. *Biotechnol Adv* 2009, **27**:297–306.
4. Rosano GL, Ceccarelli EA: **Recombinant protein expression in Escherichia coli:**

advances and challenges. *Front Microbiol* 2014, **5**:172.
5. Choi JH, Keum KC, Lee SY: **Production of recombinant proteins by high cell density culture of Escherichia coli**. *Chem Eng Sci* 2006, **61**:876–885.
6. Chen R: **Bacterial expression systems for recombinant protein production: E. coli and beyond**. *Biotechnol Adv* 2012, **30**:1102–1107.
7. Huang C-J, Lin H, Yang X: **Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements**. *J Ind Microbiol Biotechnol* 2012, **39**:383–399.
8. Mahalik S, Sharma AK, Mukherjee KJ: **Genome engineering for improved recombinant protein expression in Escherichia coli**. *Microb Cell Fact* 2014, **13**:177.
9. Zhou K, Qiao K, Edgar S, Stephanopoulos G: **Distributing a metabolic pathway among a microbial consortium enhances production of natural products**. *Nat Biotechnol* 2015, **33**:377–383.
10. Eiteman MA, Altman E: **Overcoming acetate in Escherichia coli recombinant protein fermentations**. *Trends Biotechnol* 2006, **24**:530–536.
11. Corchero JL, Villaverde A: **Plasmid maintenance in Escherichia coli recombinant cultures is dramatically, steadily, and specifically influenced by features of the encoded proteins**. *Biotechnol Bioeng* 1998, **58**:625–632.
12. Jordà J, Jouhten P, Cámara E, Maaheimo H, Albiol J, Ferrer P: **Metabolic flux profiling of recombinant protein secreting Pichia pastoris growing on glucose:methanol mixtures**. *Microb Cell Fact* 2012, **11**:57.
13. Heyland J, Fu J, Blank LM, Schmid A: **Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production**. *Biotechnol Bioeng* 2010, **107**:357–368.
14. van Rensburg E, den Haan R, Smith J, van Zyl WH, Görgens JF: **The metabolic burden of cellulase expression by recombinant Saccharomyces cerevisiae Y294 in aerobic batch culture**. *Appl Microbiol Biotechnol* 2012, **96**:197–209.
15. Heyland J, Blank LM, Schmid A: **Quantification of metabolic limitations during recombinant protein production in Escherichia coli**. *J Biotechnol* 2011, **155**:178–184.
16. Wu G, Yan Q, Jones JA, Tang YJ, Fong SS, Koffas MAG: **Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications**. *Trends Biotechnol* 2016, **34**:652–664.
17. Bhattacharya SK, Dubey AK: **Metabolic burden as reflected by maintenance coefficient of recombinant Escherichia coli overexpressing target gene**. *Biotechnol Lett* 1995, **17**:1155–1160.
18. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
19. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**.
20. Saitua F, Torres P, Pérez-Correa JR, Agosin E: **Dynamic genome-scale metabolic modeling of the yeast Pichia pastoris**. *BMC Syst Biol* 2017, **11**:27.
21. Nocon J, Steiger MG, Pfeffer M, Sohn SB, Kim TY, Maurer M, Rußmayer H, Pflügl S, Ask M, Haberhauer-Troyer C, et al.: **Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production**. *Metab Eng* 2014, **24**:129–138.
22. Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, Mattanovich D, Lee SY: **Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production**. *Biotechnol J* 2010, **5**:705–715.
23. Hirokawa Y, Matsuo S, Hamada H, Matsuda F, Hanai T: **Metabolic engineering of Synechococcus elongatus PCC 7942 for improvement of 1,3-propanediol and glycerol production based on in silico simulation of metabolic flux distribution**. *Microb Cell Fact* 2017, **16**:212.
24. Oddone GM, Mills DA, Block DE: **A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression**. *Metab Eng* 2009, **11**:367–

381.
25. Gutierrez JM, Lewis NE: **Optimizing eukaryotic cell hosts for protein production through systems biotechnology and genome-scale modeling**. *Biotechnol J* 2015, **10**:939–949.
26. Özkan P, Sariyar B, Ütkür FÖ, Akman U, Hortaçsu A: **Metabolic flux analysis of recombinant protein overproduction in Escherichia coli**. *Biochem Eng J* 2005, **22**:167–195.
27. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics**. *Mol Syst Biol* 2009, **5**.
28. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN: **Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity**. *P Natl Acad Sci USA* 2007, **104**.
29. Vazquez A, Beg QK, Demenezes MA, Ernst J, Bar-Joseph Z, Barabasi AL, Boros LG, Oltvai ZN: **Impact of the solvent capacity constraint on E. coli metabolism**. *BMC Syst Biol* 2008, **2**.
30. Goelzer A, Fromion V, Scorletti G: **Cell design in bacteria as a convex optimization problem**. *Automatica* 2011, **47**:1210–1218.
31. Goelzer A, Fromion V: **Bacterial growth rate reflects a bottleneck in resource allocation**. *Biochim Biophys Acta (BBA)-General Subj* 2011, **1810**:978–988.
32. Goelzer A, Fromion V: **Resource allocation in living organisms**. *Biochem Soc Trans* 2017, **45**:945-952.
33. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, Mariadassou M, Aymerich S, Hecker M, Noirot P, et al.: **Quantitative prediction of genome-wide resource allocation in bacteria**. *Metab Eng* 2015, **32**:232–243.
34. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**:693–693.
35. Zhuang K, Vemuri GN, Mahadevan R: **Economics of membrane occupancy and respiro-fermentation**. *Mol Syst Biol* 2011, **7**.
36. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
37. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T: **Interdependence of cell growth and gene expression: origins and consequences**. *Science* 2010, **330**:1099–1102.
38. Nilsson A, Nielsen J, Palsson BO: **Metabolic models of protein allocation call for the kinetome**. *Cell Syst* 2017, **5**:538–541.
39. Pirt SJ: **The maintenance energy of bacteria in growing cultures**. *Proc R Soc Lond B* 1965, **163**:224–231.
40. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
41. Liu M, Feng X, Ding Y, Zhao G, Liu H, Xian M: **Metabolic engineering of Escherichia coli to improve recombinant protein production**. *Appl Microbiol Biotechnol* 2015, **99**:10367–10377.
42. Glick BR: **Metabolic load and heterologous gene expression**. *Biotechnol Adv* 1995, **13**:247–261.
43. Bailey JE: **Critical limitations in biological production of chemicals: process or genetic solutions?** *FEMS Microbiol Rev* 1995, **16**:271–276.
44. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, et al.: **EcoCyc: a comprehensive database of Escherichia coli biology**. *Nucleic Acids Res* 2011, **39**:D583–D590.
45. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WK, et al.: **The MetaCyc database of metabolic pathways and enzymes**. *Nucleic Acids Res* 2018, **46**:D633–D639.

46. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245–248.
47. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0**. *Nat Protoc* 2011, **6**:1290.
48. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**:1331–1340.
49. Hjersted JL, Henson M a: **Steady-state and dynamic flux balance analysis of ethanol production by Saccharomyces cerevisiae.** *IET Syst Biol* 2009, **3**:167–179.
50. Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H: **Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110**. *Mol Syst Biol* 2006, **2**.
51. Jensen KF: **The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels.** *J Bacteriol* 1993, **175**:3401–3407.
52. Lara AR, Vazquez-Limón C, Gosset G, Bolivar F, López-Munguia A, Ramirez OT: **Engineering Escherichia coli to improve culture performance and reduce formation of by-products during recombinant protein production under transient intermittent anaerobic conditions**. *Biotechnol Bioeng* 2006, **94**:1164–1175.
53. Nanchen A, Schicker A, Sauer U: **Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli**. *Appl Environ Microbiol* 2006, **72**:1164–1172.
54. Vemuri GN, Altman E, Sangurdekar DP, Khodursky AB, Eiteman MA: **Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio**. *Appl Env Microbiol* 2006, **72**.
55. Weber J, Hoffmann F, Rinas U: **Metabolic adaptation of Escherichia coli during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes**. *Biotechnol Bioeng* 2002, **80**:320–330.
56. Wu J, Cheng S, Cao J, Qiao J, Zhao G-R: **Systematic optimization of limonene production in engineered Escherichia coli**. *J Agric Food Chem* 2019, **67**:7087–7097.
57. Jiao S, Yu H, Shen Z: **Core element characterization of Rhodococcus promoters and development of a promoter-RBS mini-pool with different activity levels for efficient gene expression**. *N Biotechnol* 2018, **44**:41–49.
58. Tian T, Salis HM: **A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons**. *Nucleic Acids Res* 2015, **43**:7137–7151.
59. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM: **Composability of regulatory sequences controlling transcription and translation in Escherichia coli;** *Proc Natl Acad Sci* 2013, **110**:14024–14029.
60. Borkowski O, Bricio C, Murgiano M, Rothschild-Mancinelli B, Stan G-B, Ellis T: **Cell-free prediction of protein expression costs for growing cells**. *Nat Commun* 2018, **9**:1457.

Chapter 5 Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation

Summary

This chapter presents the following published article:

Zeng, H., Yang, A. Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. *Sci Rep* **10,** 4283 (2020). https://doi.org/10.1038/s41598-020-61174-0.

Among all authors contributed to the published article, H.Z. and A.Y. designed the research; H.Z. performed simulation and data analysis; H.Z. wrote the manuscript; A.Y. revised the manuscript. The supplementary information of this chapter can be found in Appendix III.

Previously in Chapters 3 and 4, we showed that integrating proteome allocation constraints into metabolic models enables the prediction of key growth physiologies (e.g. the overflow metabolism) for both wild type and recombinant *E. coli*. Chapter 5 extends the overflow-specific proteome allocation model to include the carbon transport sector. The revised modelling framework aims to bridge the substrate uptake kinetics with bacterial growth and metabolic strategies and to provide a new biological interpretation of the seventy-year-old Monod growth kinetics.

In this chapter, we started with an *E. coli* core model and a more complete proteome balance that accounts for enzymes in carbon transport (C), energy production (E) and biomass synthesis (BM) and offset proteins that are flux-independent (Q). We then parameterized a previously proposed correlation between the protein cost per unit carbon influx and the availability of extracellular carbon sources. Substituting the parameterized

correlation to the proteome balance equation forms the revised proteome allocation constraint. Combining the new constraint with Michaelis-Menten kinetics, we were able to derive a hyperbolic correlation between the growth rate and environmental carbon availability, which suggests the potential biological mechanisms that dictate the phenomenological Monod parameters λ_{max} and K_s. Furthermore, we investigated the applicability of the revised modelling framework in genome-scale metabolic models and discussed work needed for more detailed metabolic modelling.

Through mathematical derivation, we find that the minimum fraction of proteome attainable for carbon transport is a positive (i.e. non-zero) value, which is dictated by the deviation between two growth-and-flux-dependent maximum proteome fractions. Besides, we showed that the maximum growth rate (λ_{max}) can be determined by the maximal size of growth-controlling proteome and the proteome cost per unit increase of growth rate, with the latter being related to the number of alternative pathways, the number of pathway enzymes and individual enzyme kinetics. Furthermore, by comparing the derived hyperbolic $\lambda - [g]$ correlation with the original Monod equation that describes the change in the specific growth rate at different environmental glucose concentrations in a form of $\lambda = \lambda_{max} \frac{[g]}{[g]+K_s}$, we showed that the Monod parameter K_s can link to the Michaelis constant for substrate transport $K_{m,g}$ through a proportional factor indicating the cell's metabolic strategy. This finding coincides with Jacques Monod's conjecture raised seventy years ago that K_s should be a function of $K_{m,g}$ and that K_s should be in general smaller than $K_{m,g}$.

Overall, Chapter 6 complements previous work through involving a more generic proteome allocation constraint, which enables the simultaneous prediction of the substrate uptake rate, the acetate production rate and the maximum growth rate given extracellular

nutrient availability. The updated modelling framework also provides a new mechanistic interpretation of the Monod growth kinetics. However, the quantitative accuracy of the model prediction, in particular, the proteome distribution is limited by the reliability of proteome parameters, some of which are not directly available for the specific strain under study and thus were estimated from the data of other strains. More omics data, e.g. proteomics and fluxomics would benefit future model calibration and validation.

5.1 Introduction

Understanding the growth of microbial cultures occupies a central place in the study of microorganisms. How growth rate varies with internal traits or external conditions greatly affects the choice of process-level parameters such as temperature, medium composition in cell cultures for a specific purpose and the strategy for the modulation of delicate cellular attributes such as enzyme activity, pathway diversity and regulatory systems.

Ever since Jacques Monod correlated the specific growth rate (λ) with extracellular substrate concentration ($[g]$) via two kinetic parameters, the maximum specific growth rate λ_{max} and the Monod constant K_s, the Monod equation $\lambda = \lambda_{max}\frac{[g]}{[g]+K_s}$ [1] becomes the best-known microbial growth kinetics for describing various bioprocesses. Other growth models (reviewed in [2]) are less prevalent due to more sophisticated formulation and insufficient experimental validation. Despite its unstructured and empirical nature [3], the Monod equation generally renders satisfactory results in terms of fitting substrate consumption and growth rate profile for a wide range of microorganisms and culture conditions. More recent studies have shown that growth kinetic constants (in Monod terms, λ_{max} and K_s) are actually a function of the culture history (e.g. sludge ages) and community composition [4,5]. In fact, twenty-years ago Kovarova-Kovar and Egli explicitly pointed out that cells could change their growth kinetics via adaptation, therefore a single set of kinetic parameters is not able to represent such variable properties [2]. The notion of variable growth kinetics is further supported by experimental evidence showing the evolution of K_s [6,7], 'intrinsic' and 'extant' λ_{max} [8] and the $\lambda_{max} - K_s$ relationship [9].

In addition to the Monod-type empirical models, the development of metabolic network reconstructions and the associated analytical approach, namely Flux Balance Analysis

(FBA) facilitates the study of growth phenotypes at the metabolic flux level [10]. An FBA model is established based on the known metabolic reaction stoichiometry and gene-enzyme relationships. It is inherently mechanistic and provides more insight into cellular physiological properties compared to an empirical growth model. More importantly, the general objective of FBA is to predict the maximum growth rate at given culture conditions and possibly genetic modifications [11]. If the prediction of growth rate and intracellular metabolic fluxes were sufficiently accurate, FBA would become a very useful tool for engineering efficient microbial systems. Essentially, the accuracy of the flux prediction in FBA is governed by constraints. One typical example is setting uptake limits for oxygen and glucose exchange fluxes to predict aerobic acetate formation (namely the overflow metabolism) in *Escherichia coli* [12], where acetate was predicted to excrete when the system was limited by oxygen uptake rate. However, it has been proposed and validated that the aerobic acetogenesis in *E. coli* is the result of an efficient proteome allocation strategy for rapid growth [13,14]. This example suggests that setting hard limitations on uptake fluxes could be a rather arbitrary and problematic treatment when there is a lack of theoretical basis, not to mention that the boundaries of myriad intracellular fluxes are often difficult to determine reliably. In principle, using experimentally determined flux values as lower and/or upper bounds could help FBA to compute more realistic flux distributions. This leads a dilemma between the overall accuracy of the simulation and the predictive power of the model (e.g., fluxes with hard-set boundaries are in a sense "weakly" predicted). Progress in incorporating macromolecular expression [15] and intracellular resource allocation [16–19] with the classic FBA effectively alleviates the need for setting subtle uptake boundaries and improves the accuracy and scope of the predicted growth phenotypes. For instance, Resource Balance Analysis (RBA) [18] eliminates the need for defining uptake bounds of substrates by

relating the enzyme efficiency with the concentration of extracellular nutrients. Multi-scale models of Metabolism and macromolecular Expression (ME-models) [20–22] enable the prediction of substrate uptake rates by incorporating growth rate-dependent demand functions and limited macromolecular synthesis machinery, e.g. limited ribosomal translation rate, limited mRNA catalytic rate and limited RNA polymerase transcription rate. Despite these successes, the prediction of substrate uptake rate in general highly relies on the accuracy of the enzyme kinetic parameters adopted in the model, most of which are not readily available at the genome scale [23].

In this chapter, we explore the potential links between the empirical Monod growth kinetics and the FBA-based metabolic modelling, with the latter to reveal the underlying biological mechanisms governing the phenomenological parameters of the former. Using two *E. coli* strains (NCM3722 and ML308) as case studies, we investigate how the integration of the proteome allocation principles into FBA would help depict the intriguing multi-scale mechanisms that govern the various growth phenotypes under a wide range of growth conditions. We firstly divided the overall proteome into coarse-grained functional sectors, i.e. carbon scavenging (C), energy generation (E), biomass synthesis (BM) and growth-independent offset (Q) sectors [19,24]. The size of each sector is dictated by the magnitude of the metabolic flux the sector processes and the corresponding proteome cost per unit flux. All individual proteome sectors were assembled together into a global proteome allocation constraint [24–26]. Furthermore, we parameterized a correlation [19] between the enzyme cost for carbon transport and the abundance of a carbon source. Applying these modelling concepts and parameterisation in FBA simulations, we were able to predict the glucose uptake rate, growth rate and the acetate overflow upon varying extracellular glucose level. Our theoretical model also predicted a step-change in K_s after the onset of the acetate excretion. Further analysis of the model suggests that this variation

originates from the change in cell's metabolic strategy, which coincides with the previous notion that cells possess variable growth kinetics due to adaptation to the environment [2]. Besides, λ_{max} is shown to be controlled by the abundance of growth-controlling proteome and the proteome cost per unit increase of growth rate. Finally, we propose that a proper connection between local and global physiological constraints may be essential for improving the predictive power of FBA models.

5.2 Results

5.2.1 Proteomic fraction occupied by the carbon-scavenging sector

Following the treatment in several recent studies on proteome allocation [14,19,25–27], we divide the overall cell proteome into four coarse-grained functional sectors: carbon-scavenging sector C that comprises enzymes for importing extracellular carbohydrate, energy biogenesis sector E that consists of enzymes used for respiration and (aerobic) fermentation, biomass synthesis sector BM that accounts for ribosomal proteins and enzymes carrying anabolic fluxes, and an offset sector Q that contains other proteins whose abundance do not vary with the fluxes in the above three sectors (Figure 5-1(a)).

The summation of the fractions of all sectors (i.e. $\phi_C, \phi_E, \phi_{BM}, \phi_Q$) equals to one

$$\phi_C + \phi_E + \phi_{BM} + \phi_Q = 1 \qquad \text{Equation 5-1}$$

Let the maximum amount of proteome attainable to C, E and BM sectors be ϕ_{max}^{g}, which corresponds to the minimum size of the Q sector, ϕ_{Qmin}. For cells striving for maximum growth,

$$\phi_C + \phi_E + \phi_{BM} = 1 - \phi_{Qmin} = \phi_{max}^{g} \qquad \text{Equation 5-2}$$

This global constraint is coupled to another proteome constraint that was established previously as a key mechanism governing the overflow metabolism in *E. coli* [14,28]

$$\phi_E + \phi_{BM} \leq \phi^o_{max}$$ Equation 5-3

ϕ^o_{max} denotes the maximum amount of proteome attainable to energy production and biomass synthesis. Equation 5-2 and Equation 5-3 were used to derive two concise proteome allocation constraints (Equation 5-40 and Equation 5-41, Methods), which were integrated into the *E. coli* metabolic network as additional rows. The resulted proteome allocation-constrained FBA model forms a mixed integer linear programing (MILP) problem (see Section 5.4.3). Determination of model parameters is detailed in Appendix III, Supplementary Text.

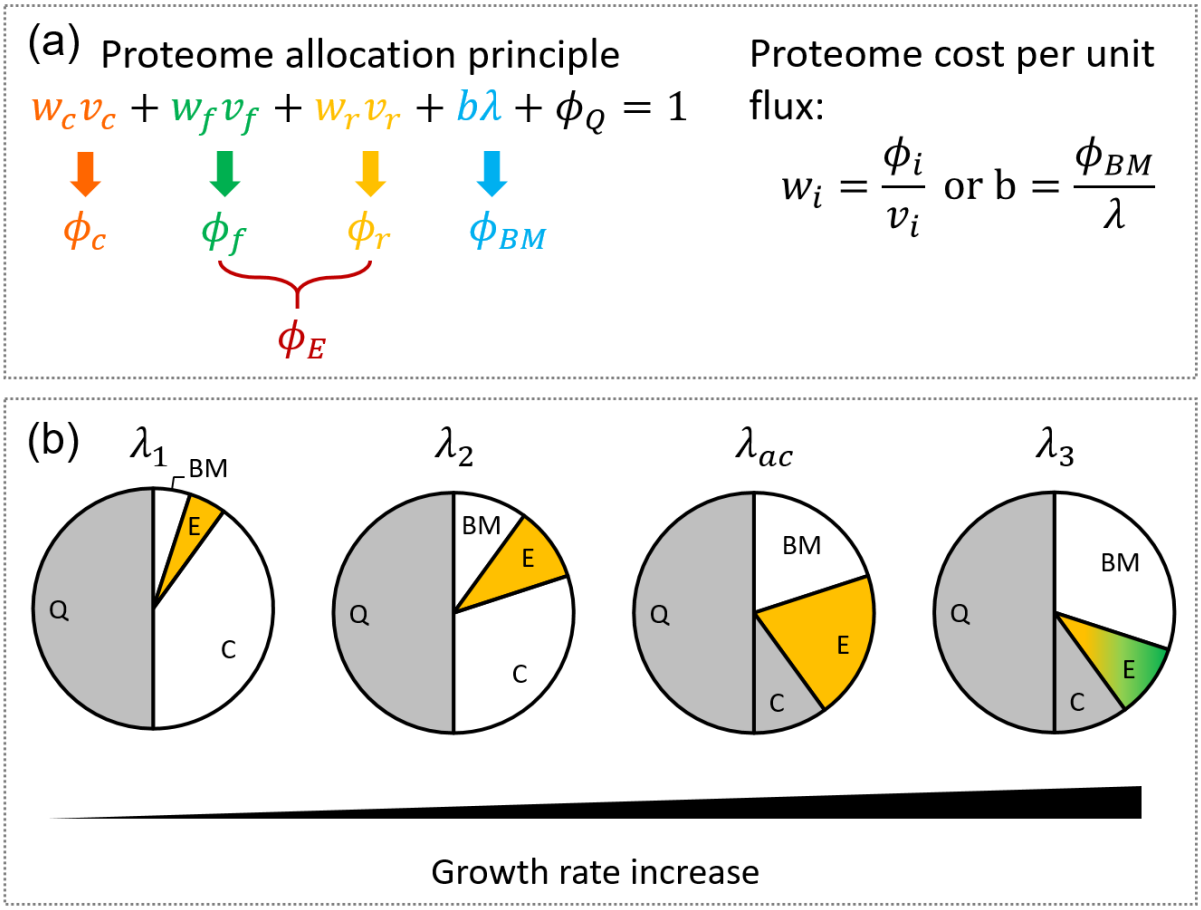


Figure 5-1 Schematic diagram of the proteome allocation model. **(a)** The overall proteome comprises four sectors, including carbon-scavenging ϕ_C, energy biogenesis ϕ_E, biomass synthesis ϕ_{BM} and growth-independent offset ϕ_Q. The enzymatic cost per unit flux of each sector is calculated by the ratio between proteome abundance and the corresponding flux. **(b)** The evolution of proteome component at the increased growth rate. Before acetate overflow, BM and E sectors increase with the growth rate while the C sector continuously declines; The E sector comprises

only respiration (marked in yellow). At $\lambda = \lambda_{ac}$, $\phi_E = \phi_r$ reaches the highest value. At higher growth rates, the E sector starts to decline to save space for the increase of the BM sector and becomes a mixture of fermentation and respiration (marked with a mixture of yellow and green); the C sector reaches its lower bound and stays constant.

FBA simulations were performed using the integrated model with extracellular glucose concentration as input and maximising cell growth as the objective function. Simulated optimal growth rate μ, carbon uptake rate (v_c), acetate production rate (v_f) and tricarboxylic acid cycle flux (v_r) at μ, along with proteome cost parameters (defined in Methods) were used to calculate the distribution of proteome sectors at different growth rates. Simulation results (Figure 5-2 (a-b)) show that ϕ_C gradually decreases as growth rate increases until it reaches a minimum level (ϕ_{Cmin}), which is equal to the gap between ϕ_{max}^{g} and ϕ_{max}^{o}:

$$\phi_C > \phi_{Cmin}, when\ \lambda < \lambda_{ac} \quad \text{Equation 5-4}$$

$$\phi_C = \phi_{Cmin} = \phi_{max}^{g} - \phi_{max}^{o}, when\ \lambda \geq \lambda_{ac} \quad \text{Equation 5-5}$$

The growth rate at which ϕ_C reaches its lower bound is corresponding to the threshold growth rate λ_{ac}, above which acetate overflow occurs. When growth rate exceeds λ_{ac} the composition of ϕ_E changes from the enzymes for pure respiration to those for a combination of fermentation and respiration (Figure 5-1(b)). The predicted decline of ϕ_C is supported by the observed upregulation of catabolic genes under carbon limitation [25]. We did not find experimental evidence for a non-zero ϕ_{Cmin}. However combining Equation 5-2 and Equation 5-3 gives $\phi_C \geq \phi_{max}^{g} - \phi_{max}^{o}$, which in theory implies a positive value of ϕ_{Cmin} given $\phi_{max}^{g} > \phi_{max}^{o}$.

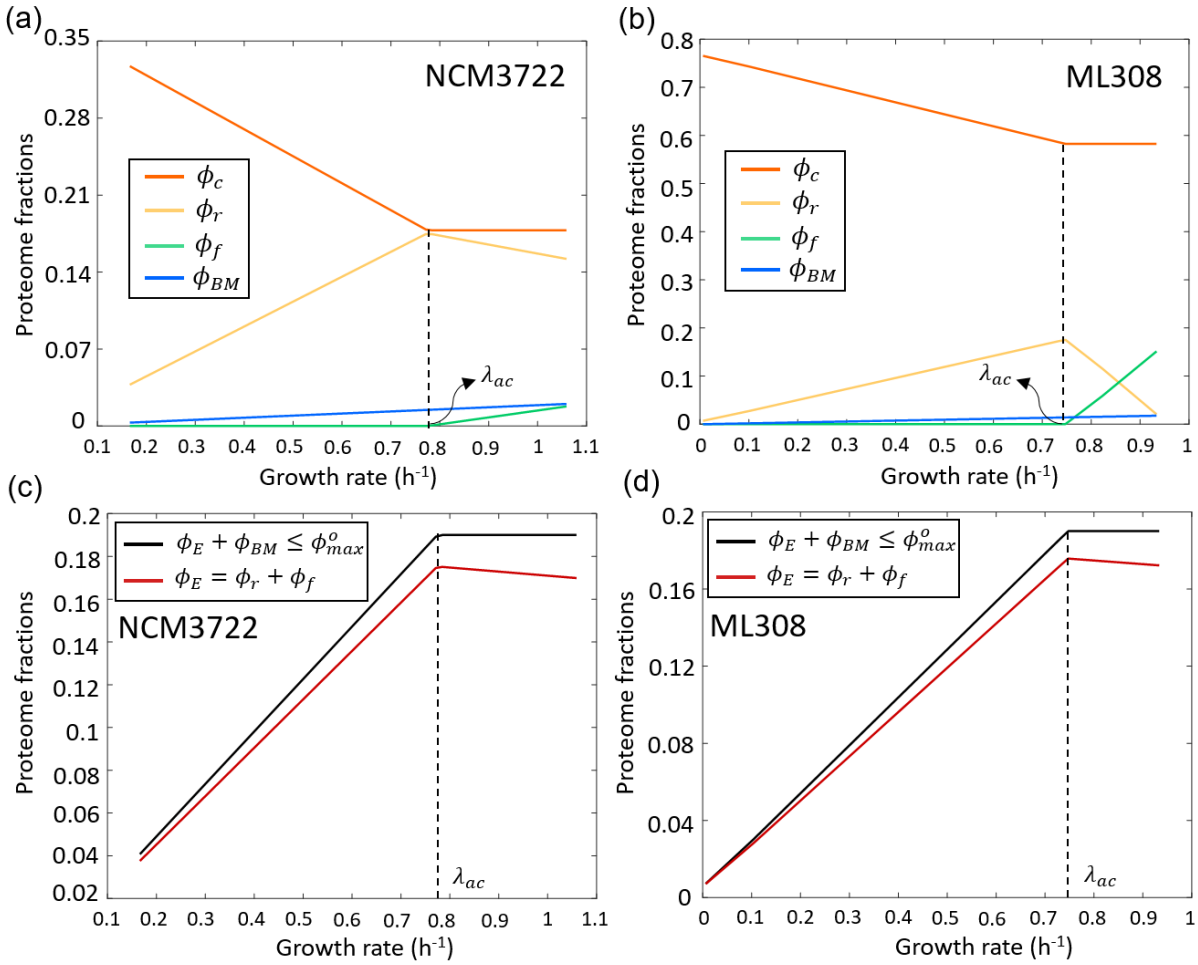


Figure 5-2 Simulated distribution of proteome sectors ϕ_C, ϕ_r, ϕ_f and ϕ_{BM} against different growth rates for *E. coli* NCM3722 **(a)** and ML308 **(b)**; and simulated change of the E sector (ϕ_E) and summation of E and BM sectors ($\phi_E + \phi_{BM}$) for NCM3722 **(c)** and ML308 **(d)**. ϕ_r and ϕ_f are E sector components. The dashed line indicates the onset of acetate overflow. λ_{ac} is the threshold growth rate, above which acetate overflow occurs.

5.2.2 Predictions of glucose uptake rate and the acetate overflow

It has been shown that the incorporation of Equation 5-3 into a metabolic model of *E. coli* enables the prediction of the acetate overflow at $\lambda \geq \lambda_{ac}$ [28]. Assuming a linear relationship between a proteome fraction and the metabolic flux it processes [25], the C sector could be modelled as

$$\phi_C = w_c v_c \qquad \text{Equation 5-6}$$

where v_c is the carbon uptake flux. The linear coefficient w_c denotes the enzyme cost per unit carbon influx, which has been proposed to be a function of the environmental substrate levels, derived from the Michaelis-Menten equation [19]

$$w_c = w_{c,0}(1 + \frac{K_{m,g}}{[g]}) \qquad \text{Equation 5-7}$$

The constant $w_{c,0}$ represents the minimum enzyme cost per unit carbon influx; $[g]$ is the glucose concentration, where glucose is taken as a representative substrate; $K_{m,g}$ is the Michaelis constant. We determined the value of $w_{c,0}$ using strain-specific cell culture data (Appendix III, Supplementary Text); with reported values of $K_{m,g}$ we were subsequently able to compute w_c as a function of $[g]$ (Equations AppxIII-S7 and AppxIII-S8). Coupling Equation 5-7 with the proteome allocation constraints (Equation 5-40 and Equation 5-41), the glucose uptake rate, acetate production rate and optimal growth rate of two *E. coli* strains were predicted by specifying $[g]$ (as opposed to setting glucose influx as pre-determined input [12]) and maximizing the growth rate. Simulation results agree well with the experimental data (Figure 5-3 (a-d)).

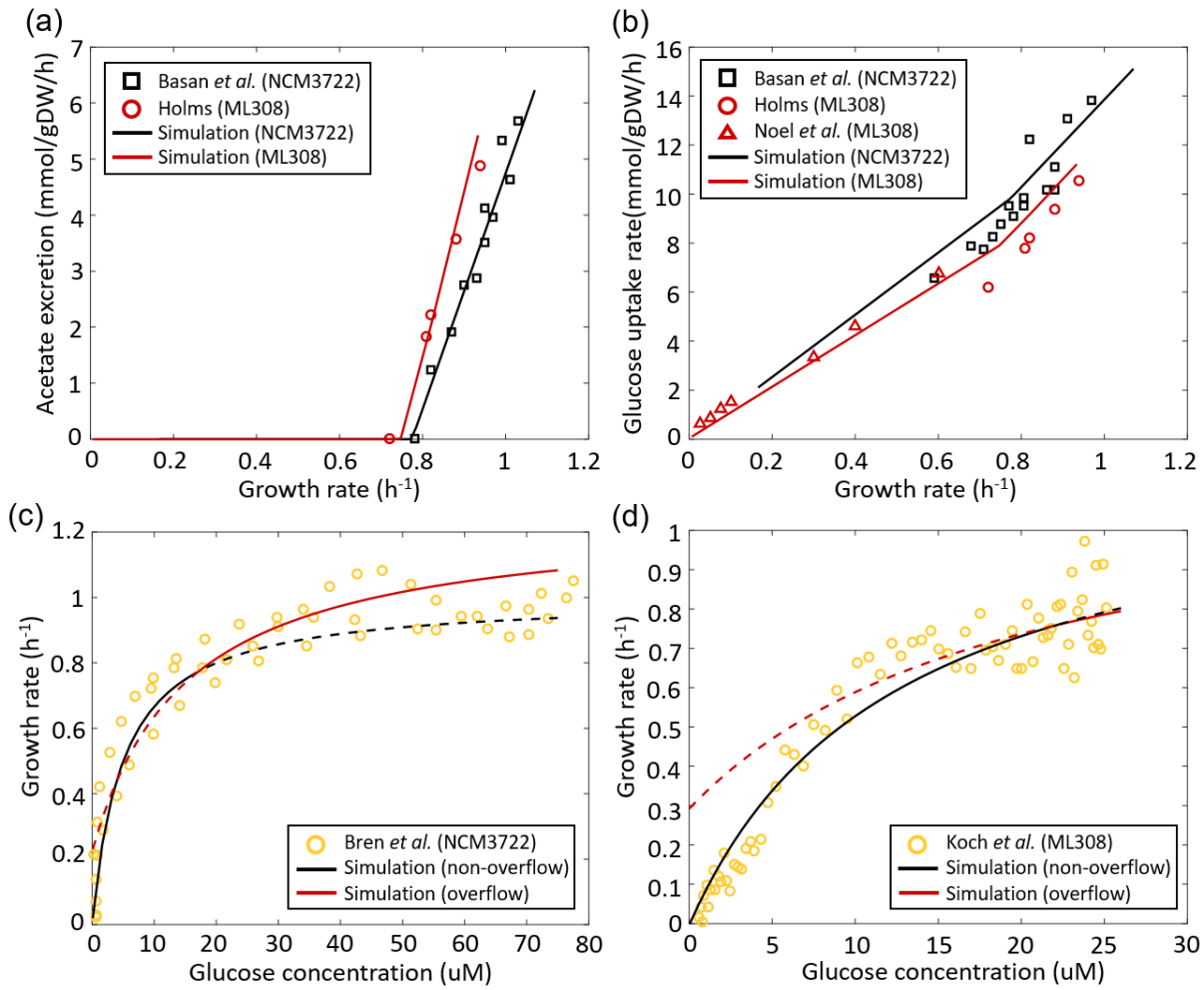


Figure 5-3 Comparison between the model simulation and experimental data of the growth of *E. coli* in glucose-limited cultures. **(a)** Comparison between predicted and measured acetate excretion rate. Experimental data for NCM3722 were obtained from figure 1 of ref. [14]; experimental data for ML308 were obtained from table 7 of ref. [29]. **(b)** Comparison between predicted and measured glucose uptake rate. Experimental data for NCM3722 were obtained from figure 3B of ref. [19]; experimental data for ML308 were obtained from table 7 of ref. [29] and figure 3b of ref. [30]. **(c)** Comparison between predicted and measured specific growth rate against glucose concentration profile for NCM3722. Experimental data were obtained from SI figure 1 of ref. [31]. **(d)** Comparison between predicted and measured specific growth rate against glucose concentration profile for ML308. Experimental data were obtained from figure 1 of ref. [32]. Solid lines in **(c)** and **(d)** are actual model predictions; dashed lines are presented to show the model prediction in the region where the model is not applicable. Simulation results of optimal growth rate, glucose uptake rate and acetate production rate were obtained in FBA simulations of the proteome allocation-integrated metabolic model. In all simulations, the extracellular glucose concentration was specified as input and the growth rate was maximised.

5.2.3 Deriving the hyperbolic $\lambda - [g]$ correlation

Assuming a linear correlation between growth rate and glucose uptake rate (as suggested by Figure 5-3(b))

$$v_c = \begin{cases} k_1\lambda + b_1 \ (\lambda < \lambda_{ac}) \\ k_2\lambda + b_2 \ (\lambda \geq \lambda_{ac}) \end{cases} \quad \text{Equation 5-8}$$

where k_1, b_1, k_2, b_2 are linear coefficients. From Figure 5-2 (c-d) we have the following description for ϕ_E and ϕ_{BM}

$$\frac{\phi_E+\phi_{BM}}{\phi^o_{max}} = \begin{cases} k_3\lambda + b_3 < 1 \ (\lambda < \lambda_{ac}) \\ 1 \qquad\qquad\quad (\lambda \geq \lambda_{ac}) \end{cases} \quad \text{Equation 5-9}$$

where k_3 and b_3 are linear coefficients. Following the linear proteome-flux relationship proposed in Chapter 3 (Equation 3-6), Equation 5-2 can be written as

$$w_c^* v_c + w_f^* v_f + w_r^* v_r + b^*\lambda = \frac{\phi^g_{max}}{\phi^o_{max}} \quad \text{Equation 5-10}$$

where the term $\left(w_f^* v_f + w_r^* v_r + b^*\lambda\right)$ is equivalent to $\frac{\phi_E+\phi_{BM}}{\phi^o_{max}}$. Substituting Equation 5-8 and Equation 5-9 into Equation 5-10, for non-overflow growth ($\lambda < \lambda_{ac}$)

$$w_c^* = \frac{\frac{\phi^g_{max}}{\phi^o_{max}} - (k_3\lambda + b_3)}{(k_1\lambda + b_1)} \quad \text{Equation 5-11}$$

Comparing Equation 5-11 with Equation 5-7 (noted that $w_c^* = w_c/\phi^o_{max}$) gives

$$\left(\frac{\frac{\phi^g_{max}}{\phi^o_{max}} - (k_3\lambda + b_3)}{k_1\lambda + b_1}\right)\phi^o_{max} = w_{c,0}\left(1 + \frac{K_{m,g}}{[g]}\right) \quad \text{Equation 5-12}$$

Rearranging Equation 5-12 to represent λ as a function of $[g]$:

$$\lambda = \frac{(\phi^g_{max} - b_3\phi^o_{max} - w_{c,0}b_1)[g] - w_{c,0}b_1K_{m,g}}{(w_{c,0}k_1 + k_3\phi^o_{max})[g] + w_{c,0}k_1K_{m,g}} \quad \text{Equation 5-13}$$

Equation 5-13 can be re-arranged to a hyperbolic form comparable to the Monod equation.

$$\lambda = \frac{\phi^g_{max} - \phi^o_{max}\left(b_3 - \frac{b_1}{k_1}k_3\right)}{w_{c,0}k_1 + \phi^o_{max}k_3} \frac{[g]}{[g] + \frac{w_{c,0}k_1}{w_{c,0}k_1 + \phi^o_{max}k_3}K_{m,g}} - \frac{b_1}{k_1}$$ Equation 5-14

Similarly for overflow growth ($\lambda \geq \lambda_{ac}$), substituting Equation 5-8 and Equation 5-9 into Equation 5-10

$$w_c^* = \frac{\frac{\phi^g_{max}}{\phi^o_{max}} - 1}{k_2\lambda + b_2}$$ Equation 5-15

Comparing Equation 5-15 with Equation 5-7

$$\left(\frac{\frac{\phi^g_{max}}{\phi^o_{max}} - 1}{k_2\lambda + b_2}\right)\phi^o_{max} = w_{c,0}\left(1 + \frac{K_{m,g}}{[g]}\right)$$ Equation 5-16

Rearrange Equation 5-16 to represent λ as a function of $[g]$

$$\lambda = \frac{\phi^g_{max} - \phi^o_{max}}{w_{c,0}k_2} \frac{[g]}{[g] + K_{m,g}} - \frac{b_2}{k_2}$$ Equation 5-17

Comparing Equation 5-14 and Equation 5-17 with the Monod equation, i.e. $\lambda = \lambda_{max}\frac{[g]}{[g]+K_s}$, we obtain the following matches

$$\lambda_{max} = \begin{cases} \frac{\phi^g_{max} - (\phi^o_{max}b_3 + w_{c,0}b_1)}{w_{c,0}k_1 + \phi^o_{max}k_3} & (\lambda < \lambda_{ac}) \\ \frac{\phi_{c,min} - w_{c,0}b_2}{w_{c,0}k_2} & (\lambda \geq \lambda_{ac}) \end{cases}$$ Equation 5-18

and

$$K_s = \begin{cases} \frac{w_{c,0}k_1}{w_{c,0}k_1 + \phi^o_{max}k_3}K_{m,g} & (\lambda < \lambda_{ac}) \\ K_{m,g} & (\lambda \geq \lambda_{ac}) \end{cases}$$ Equation 5-19

Below we show how to resolve the physical links between the Monod kinetic parameters (λ_{max} and K_s) and cell's physiological state through analyzing the biological meaning of the denominator and the numerator of Equation 5-18 and Equation 5-19.

Resolving λ_{max}

Starting with the non-overflow scenario ($\lambda < \lambda_{ac}$), we first focus on the numerator of the derived expression of λ_{max} (Equation 5-18). b_3 is the proteome fraction accounting for non-growth maintenance of ϕ_E and ϕ_{BM} (normalised by ϕ^o_{max}; see Equation 5-9), therefore

$$\phi^o_{max} b_3 = (\phi_E + \phi_{BM})_{atpm} \qquad \text{Equation 5-20}$$

where the subscript $atpm$ denotes the proteome fraction occupied by non-growth-associated maintenance. Furthermore, b_1 is the rate of carbon intake for maintenance purposes, therefore the product of b_1 and $w_{c,0}$ (lowest enzyme cost per unit carbon influx, see Equation 5-7) quantifies the (lowest) portion of the C sector proteome that is occupied for maintenance.

$$w_{c,0} b_1 = \phi_{C,atpm} \qquad \text{Equation 5-21}$$

Combining Equation 5-20 and Equation 5-21, the numerator of Equation 5-18 (the branch for $\lambda < \lambda_{ac}$) can be expressed as

$$\phi^g_{max} - \left(\phi_{E,atpm} + \phi_{BM,atpm} + \phi_{C,atpm}\right) = \phi^g_{max} - \sum_i \phi_{i,atpm} \qquad \text{Equation 5-22}$$

where i represents a proteome sector (C, E, or BM).

Moving on to the denominator, $w_{c,0} k_1$ can be coupled to the carbon influx (via Equation 5-8), thus representing the (lowest) demand of C sector proteome per unit increase of growth rate, denoted by p_c

$$w_{c,0} k_1 = w_{c,0} \frac{dv_c}{d\lambda} = \left(\frac{d\phi_C}{d\lambda}\right)_{w_c = w_{c,0}} = p_c \qquad \text{Equation 5-23}$$

Similarly, $\phi^o_{max} k_3$ is the summation of the proteome cost per unit increase of growth rate for ϕ_E and ϕ_{BM} (see Equation 5-9)

$$\phi_{max}^{o} k_3 = \frac{d\phi_E}{d\lambda} + \frac{d\phi_{BM}}{d\lambda} = p_E + p_{BM} \quad \text{Equation 5-24}$$

Combining Equation 5-23 and Equation 5-24, the denominator $\left(w_{c,0}k_1 + \phi_{max}^{o} k_3\right)$ can be expressed as $p_c + p_E + p_{BM} \equiv \sum_i p_i$, where i represents a proteome sector (C, E, or BM). Summarising, the maximum specific growth rate for non-overflow growth can be quantified as

$$\lambda_{max} = \frac{\phi_{max}^{g} - \sum_i \phi_{i,atpm}}{\sum_i p_i} \ (\lambda < \lambda_{ac}) \quad \text{Equation 5-25}$$

On the other hand, the numerator of Equation 5-18 for the overflow scenario ($\lambda \geq \lambda_{ac}$) can be viewed as the (adjusted) proteome abundance of C sector, denoted as ϕ_c' below, which is made proportional to λ by applying an offset term $w_{c,0}b_2$ (AppxIII-Figure S1).

$$\phi_{c,min} - w_{c,0}b_2 = \phi_c' \quad \text{Equation 5-26}$$

The denominator of $w_{c,0}k_2$ is comparable to Equation 5-23, which represents the (lowest) demand of C sector proteome per unit increase of growth rate (p_c) under overflow conditions, i.e.

$$w_{c,0}k_2 = p_c \quad \text{Equation 5-27}$$

Overall, during acetate overflow λ_{max} can be interpreted as

$$\lambda_{max} = \frac{\phi_c'}{p_c} \ (\lambda \geq \lambda_{ac}) \quad \text{Equation 5-28}$$

Resolving K_s

Re-writing Equation 5-19 to

$$K_s = \delta K_{m,g} \quad \text{Equation 5-29}$$

$$\text{where } \delta = \begin{cases} \frac{w_{c,0}k_1}{w_{c,0}k_1 + \phi_{max}^{o}k_3} < 1 \ (\lambda < \lambda_{ac}) \\ 1 \ (\lambda \geq \lambda_{ac}) \end{cases} \quad \text{Equation 5-30}$$

Given the above analysis for resolving λ_{max}, the biological meaning of the expression of δ for $\lambda < \lambda_{ac}$ could be shown through

$$\delta = \frac{w_{c,0}k_1}{w_{c,0}k_1+\phi^o_{max}k_3} = \frac{p_c}{p_c+p_E+p_{BM}} < 1 \qquad \text{Equation 5-31}$$

For the overflow scenario ($\lambda \geq \lambda_{ac}$), $\delta = 1$ actually derives from

$$\delta = \frac{w_{c,0}k_2}{w_{c,0}k_2} = \frac{p_c}{p_c} = 1 \qquad \text{Equation 5-32}$$

The discrete feature of the proportional factor δ reflects the switch in a cell's metabolic state (see more discussion below).

5.2.4 λ_{max} is a function of proteome resource and proteome cost

Equation 5-25 and Equation 5-28 derived in the section above can be written into a generic form

$$\lambda_{max} = \phi_{growth}/p_{growth} \qquad \text{Equation 5-33}$$

ϕ_{growth} represents the maximal fraction of the growth-controlling proteome, and p_{growth} is the proteome cost per unit increase in the growth rate. Equation 5-33 is in line with the previous proposal that λ_{max} is closely related to the protein synthesis mechanism [4]. During non-overflow growth ($\lambda < \lambda_{ac}$), we have showed that (Equation 5-20 to Equation 5-25)

$$\phi_{growth} = \phi^g_{max} - \sum_i \phi_{i,atpm} \qquad \text{Equation 5-34}$$

$$p_{growth} = \sum_i p_i \qquad \text{Equation 5-35}$$

$\phi_{i,atpm}$ is the proteome fraction of sector i occupied by non-growth-associated maintenance; p_i is the proteome cost of sector i per unit increase of growth rate; i represents C, E, or BM sector. During overflow growth ($\lambda \geq \lambda_{ac}$), we have shown that (Equation 5-26 to Equation 5-28)

$$\phi_{growth} = \phi_c{}' \quad \text{Equation 5-36}$$

$$p_{growth} = p_c \quad \text{Equation 5-37}$$

$\phi_c{}'$ denotes the fraction of the proteome allocable to the carbon-scavenging sector (adjusted with an off-set to ϕ_{cmin}, see Equation 5-26). Equation 5-36 and Equation 5-37 imply that in the overflow region λ_{max} can be dictated solely by the characteristics of the C sector.

The above results suggest two ways to improve cell growth: increase ϕ_{growth} or decrease p_i. ϕ_{growth} could be elevated by reducing maintenance or eliminating the expression of useless proteins. However, it might be difficult to directly modify ϕ_{growth} as it is more or less an intrinsic feature of a cell. Reducing p_i could be more realistic in synthetic biology. In this chapter, p_c could gradually reduce upon the decline of w_c at increased growth rates (Equation 5-23, Equation 5-27 and AppxIII-Figure S1). p_E was shown to change from positive to negative due to varied ϕ_E against growth rate correlation before and after λ_{ac} (Figure 5-2 (c-d)). The variability of p_i could result from two mechanisms: the cell upregulates catabolic genes under carbon limitation (leads to variable p_c) and changes its metabolic strategies under different culture conditions (leads to variable p_E). The latter is governed by the presence of alternative pathways. We could infer that if a cell carries multiple pathways of a similar biological function (e.g. energy biogenesis) but with distinct enzyme (or protein) costs, the cell might be able to improve its growth via directing more metabolic flux to the enzymatically low-cost pathways. This could help design synthetic microorganisms with potentially higher robustness and fitness to the growth environment. A recent study also predicted that improved enzyme activity of energy pathways leads to a higher growth rate [33], which supports the above suggestions.

5.2.5 K_s depends on the interplay between fractions of proteome sectors and affinity of transport enzymes

While λ_{max} unambiguously refers to the highest specific growth rate a cell can achieve, the biological meaning of K_s is less clear. Often taken as a mathematical analogy to the Michaelis-Menten enzyme kinetics, the Monod constant K_s seems to share some link with the Michaelis constant for substrate transport $K_{m,g}$. This might explain why the Monod growth kinetics is sometimes mistermed as "Michaelis-Menten kinetics" for describing growth-associated bioprocesses [2].

While the derivation of the Monod equation is strictly empirical [34], Monod himself raised interesting connections between K_s and $K_{m,g}$: (i) 'the value of K_s should be expected to bear some more or less distant relation to the apparent dissociation constant of the enzyme involved in the first step of the breakdown of a given compound'; (ii) 'since a change of conditions affecting primarily the velocity of only one rate-determining step will, in general (but not necessarily), be only partially reflected in the overall rate, one might expect K_s values to be lower than the corresponding values of the Michaelis constant of the enzyme catalyzing the reaction' [1].

Essentially Monod suggested K_s could be a function of $K_{m,g}$ and that K_s should be generally smaller than $K_{m,g}$. In the past twenty years, several attempts have been made to explore the physical meaning of K_s. One interpretation is that $1/K_s$ reflects the overall affinity of a cell to a substrate [2]. A later study suggests that K_s is a function of the overall change of free energy of the microbial growth process [35]. More recently, K_s was related to $K_{m,g}$ via investigating the control of the transport step on the specific growth rate [34]. In Section 5.2.3, we explicitly derived that K_s relates to $K_{m,g}$ via a proportional factor δ (Equation 5-29 and Equation 5-30). The discrete feature of δ leads to an (upward) step

change of K_s at the acetate switch ($K_{s,NCM3722} = 5\ \mu M$ and $K_{s,ML308} = 12\ \mu M$ for $\lambda < \lambda_{ac}$; $K_{s,NCM3722} = 15\ \mu M$ and $K_{s,ML308} = 20\ \mu M$ for $\lambda \geq \lambda_{ac}$), which perfectly validates Monod's proposition that $K_s = f(K_{m,g})$ and $K_s \leq K_{m,g}$.

In fact, it has been noted that K_s values for *E. coli* grown with glucose vary significantly in different studies or different growth conditions, and there has been a lack of a satisfactory explanation of why the presumably constant K_s spans across such a wide range [36]. The discrete variable δ provides a simple explanation to this long-lasting problem: under strong carbon limitation, the cell benefits from the metabolic strategy where all carbon consumed for energy production is metabolized through respiration; $\delta = p_c/(p_c + p_E + p_{BM}) < 1$ (Equation 5-31) suggests the relative importance of C, E and BM sectors in dictating the cell's overall affinity to the growth-controlling substrate.

When substrate becomes sufficient which potentially allows the cell to grow at a rate beyond a threshold, it becomes more advantageous for a cell to switch to a metabolic strategy where a significant portion of carbon consumed for energy production is processed through the proteome-efficient acetate pathway. In this case, the C sector touches its lower bound and decouples from E and BM sectors, $\delta = p_c/p_c = 1$ (Equation 5-32). Essentially, δ reflects the metabolic state of a cell. In principle, if a cell possesses a number of metabolic strategies, δ would hold the same number of discrete values, each leading to a distinct value of the apparent affinity constant K_s. This mechanism could be the underlying reason for the observed large variation in K_s under different culture conditions.

Overall, our theoretical model indicates that K_s reflects the combined effect of the local characteristics of the carbon transport system and the overarching interplay between carbon-scavenging and other growth-related functions.

5.3 Discussion

In this chapter, we modified the classic FBA model (constrained primary by the mass and energy balance) via embedding the coarse-grained proteome allocation constraints with an intention to reveal the potential mechanistic nature of the empirical growth kinetics proposed by Monod seventy years ago. By doing so, we showed theoretically a non-zero lower bound of the carbon-scavenging sector, which differs from the previous notion that ϕ_C could continuously decrease towards zero as growth rate increases [19]. Under carbon limitation, it seems to be agreeable that $\phi_E + \phi_{BM}$ increases with growth rate [26,27] and reaches a plateau (defined by ϕ_{max}^o) where the acetate overflow occurs so that more proteome could be allocated to biomass synthesis as required for rapid growth [14]. Consequently, the rest of the proteome ($\phi_C + \phi_Q$) has to decrease and then level out [37]. If ϕ_Q remains constant, taking for example its minimum value (ϕ_{Qmin}) as considered in this chapter for cells striving for maximum growth, the size of the level end of the ($\phi_C + \phi_Q$) profile will be ($\phi_{Cmin} + \phi_{Qmin}$). We argue that in this level sum, the contribution of ϕ_{Cmin} is unlikely to be zero, given the significant carbon influx to sustain the high rate of biomass synthesis and energy production under overflow conditions. On the maximum value of ϕ_C (ϕ_{Cmax}), our parameter estimation based on data assembled from literature (AppxIII-Table S1) determined $\phi_{Cmax} = 0.37$ for NCM3722 and $\phi_{Cmax} = 0.77$ for ML308 ($\phi_{Cmax} \approx \phi_{max}^g$, determined by extrapolating to zero growth rate, also see Equation 5-2 and Figure 5-2 (a-b)). These specific values are affected by a number of factors, including particularly (i) the assumed value for ϕ_{max}^o (0.19, previously suggested for *E. coli* NCM3722 [14]) and (ii) the value of K_s estimated from cell culture data (Equations AppxIII-S5 and AppxIII-S6). In reality, the former could be strain-dependent, while the latter is known to be sensitive to growth conditions and culture history [4,32,36]. Therefore, the specific results of parameterization from this chapter were to support the

illustration of the overall modelling approach; their biological significance needs to be interpreted with caution.

In dynamic FBA (DFBA), v_c (uptake rate of growth-limiting carbon source) is generally determined by applying the Michaelis-Menten kinetics for substrate transport [38–42]. An inherent issue with this approach is that the kinetic parameters are normally valid only for a specific substrate uptake pattern, which may vary depending on the variety of fermentative by-products (e.g. acetate, formate, ethanol or a mixture of them) [10] or more generally on the physiological state. Therefore, using Michaelis-Menten kinetics alone (with a single set of kinetic parameters) to predict the carbon intake suffers from a very limited scope of applicability – it cannot depict a physiological state that deviates from the state that it was originally fitted to. In this chapter, this drawback is rectified by introducing the proteome allocation principle, which couples carbon transport with other growth-associated metabolism. Through a fully parameterized $w_c - [g]$ correlation, we compute the variable w_c given the environmental glucose concentration (as opposed to directly manipulating w_c [19]). The computable w_c is involved in the proteome allocation constraints, which allows us to simultaneously predict the glucose uptake rate, specific growth rate and acetate overflow upon varying the environmental glucose level. The coupling of the local metabolic capacity (for carbon transport) and the global regulatory mechanism (on proteomic resource allocation) was shown to be essential in improving the predictive power of FBA models. It should be noted that, while in this chapter we have adopted a single $K_{m,g}$ for the carbon transport system, it is known that in *E. coli*, $K_{m,g}$ differs from different transporters which could be active either exclusively or in parallel [43,44]. When using a single (aggregated) $K_{m,g}$ becomes inadequate, a more accurate representation of the carbon transport could be achieved by explicitly modelling individual transporters (see below for further discussion). Besides, the proteome allocation principle

could also be revised to investigate proteomic burdens imposed by the expression of heterologous genes [45].

In this chapter, we used the core metabolic model of *E. coli* to demonstrate the effectiveness of proteome allocation constraints in dictating bacterial growth strategies. However, the proposed modelling framework can be implemented with more detailed genome-scale metabolic models (GEMs) [46–48] to explore biological insights for *E. coli* and other microorganisms with similar putative mechanisms. The key advance of the GEM over the core model is the extensive information of metabolic reactions originating from genome annotation [49]. Pathways such as alternate carbon metabolism, amino acid metabolism, nucleotide metabolism, cofactor biosynthesis and fatty acid biosynthesis are exclusively modelled in GEMs. To fuse the proteome allocation constraints into a GEM, particular attention is needed for the choice of fluxes underlying v_c, v_f and v_r, as well as the determination of proteome cost parameters w_c^*, w_f^* and w_r^*.

We first focus on w_c^* and v_c. In the core model, glucose is the sole carbon source to support growth and is imported via phosphoenolpyruvate (PEP):pyruvate (PYR) phosphotransferase system (PTS). In this case v_c is related to a one-step transport flux (GLCpts). However in *E. coli* GEM *i*AF1260, glucose can be transported not only via glucose-specific PTS components, but also via alternative routes such as glucose-specific ABC system, glucose:proton symporter or simply via diffusion [46]. In *i*AF1260, glucose import (from extracellular glucose to cytosolic glucose-6-phosphate) is carried out by multiple enzymatic steps, e.g. first via the glucose:proton symporter then through hexokinase, instead of a simplified one-step process. Furthermore, 174 carbon sources were predicted to potentially support growth in *i*AF1260, meaning that growth simulation is not limited to glucose minimal media. The improved details of the metabolic network imply that (a) when modelling cells grown on glucose, v_c may need to be coupled to

multiple glucose transporters and (b) when modelling cells grown on alternative or mixed carbon sources, v_c needs to correspond to various carbohydrate transport pathways. Accompanying a more detailed account of substrate transport fluxes, the (normalised) proteome cost parameter for carbon transport w_c^* has to be revised. We have shown in this chapter that w_c^* is a function of ϕ_{max}^o, $w_{c,0}$ and $K_{m,g}$ (Equation 5-7 and Equation 5-40) and that $w_{c,0}$ is a function of ϕ_{max}^o and $K_{m,g}$ (Equation AppxIII-S5); combining the two makes w_c^* a function of $K_{m,g}$ (the Michaelis constant for glucose transport). When multiple transporters exist, as in GEMs, a set of $K_{m,g}$ (for glucose) or multiple sets of $K_{m,s}$ (for non-glucose carbon sources) would be required for active transporters in the specific growth conditions. Correspondingly, w_c^* will be calculated from the combination of $K_{m,g}$ for cells grown on a glucose minimal medium. For cells grown on a complex carbon medium, multiple w_c^* will be needed (to match distinct transport fluxes) and should be determined by multiple sets of $K_{m,s}$. In short, integrating the proteome cost constraints into GEM asks for more information on kinetic parameters of carbon transport.

Furthermore, the increased flux variability and potential reactions running in parallel in GEMs may alter the choice of representative fluxes of fermentation and respiration pathways. For example, v_f and v_r in *i*AF1260 can become enolase (ENO) and citrate synthase (CS) [45]. Changed v_f and v_r will subsequently affect the estimation of proteome cost parameters w_f^* and w_r^* (Appendix III, Supplementary Text).

In conclusion, our model offers a latest mechanistic interpretation of the seventy-year-old Monod growth kinetics: the maximum specific growth rate of a microorganism could be governed by the abundance of growth-controlling proteome and the associated proteome cost per unit increase of growth rate. The Monod constant K_s was shown to be quantitatively related to not only the enzymatic affinity for substrate transport $K_{m,g}$ but

also the metabolic state of a cell, which might explain the large variations in reported K_s values. Our analysis also suggests that a microorganism with a lower maintenance cost, a higher fraction of growth-controlling proteome and alternative pathways with different enzyme costs likely competes successfully in a changing environment. Finally, the proposed modelling concept eliminates the need for treating substrate intake as an input in FBA simulation, which demonstrates the potential of coupling local and global physiological constrains in predictive modelling for systems biology and synthetic biology.

5.4 Methods

5.4.1 Modelling proteome allocation in *E. coli*

Following the assumption that the fraction of a proteome sector is proportional to the metabolic flux it processes [26], Equation 5-2 can be rewritten as

$$w_c v_c + w_f v_f + w_r v_r + b\lambda = \phi_{max}^{g} \qquad \text{Equation 5-38}$$

w_i is the enzyme cost per unit flux processed by sector i ($i = C, f \ or \ r$). b is the proteome cost per unit biomass synthesis flux. v_c is the carbon transport flux. For *E. coli* grown with glucose as the single growth-controlling substrate, v_c is the glucose uptake flux which is normally processed by the phosphoenolpyruvate (PEP):sugar phosphotransferase system (PTS) [43,50]. v_f (v_r) is the fermentation (respiration) flux; λ is the biomass synthesis flux, which in this chapter is equal to the specific growth rate. From ref. [28], the fraction of proteome attainable to energy biogenesis and biomass synthesis is constrained by an upper bound ϕ_{max}^{o}

$$w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 \qquad \text{Equation 5-39}$$

where $w_i^* \equiv \frac{w_i}{\phi_{max}^o}$ ($i = f$ *or* r) and $b^* \equiv \frac{b}{\phi_{max}^o}$ are the (normalised) proteome cost per unit flux. Dividing both sides of Equation 5-38 by ϕ_{max}^o

$$w_c^* v_c + w_f^* v_f + w_r^* v_r + b^* \lambda = \frac{\phi_{max}^g}{\phi_{max}^o} \quad \text{Equation 5-40}$$

Substituting Equation 5-39 into Equation 5-40, under acetate overflow (hence equal sign applied to Equation 5-39)

$$w_c^* v_c = \frac{\phi_{max}^g}{\phi_{max}^o} - 1 \quad \text{Equation 5-41}$$

It follows that

$$\phi_{Cmin} = \phi_{max}^g - \phi_{max}^o \quad \text{Equation 5-42}$$

ϕ_{Cmin} is the minimum proteome fraction allocated to the carbon-scavenging sector (occurring during acetate overflow). $\phi_{max}^g > \phi_{max}^o$ (due to the inclusion of the C sector). Equation 5-40 and Equation 5-41 are the proteome allocation constraints adopted in model simulations.

5.4.2 Determination of GAM and ATPM

In FBA models, the energy consumption is dictated primarily via growth-associated maintenance (GAM) and non-growth associated maintenance (ATPM), both of which are difficult to quantify accurately [46]. GAM and ATPM could also vary between different species and strains (BiGG database, http://bigg.ucsd.edu/). Therefore we modified the default values of GAM and ATPM according to strain-specific energetic data (AppxIII-Table S2). To do so we associated GAM with the molar growth yield (namely the true growth yield Y_G [51]). ATPM was linked to the maintenance coefficient (m), which corresponds to the rate of substrate uptake extrapolated to zero growth rate.

5.4.3 Model simulation

The following linear programing problem was solved in FBA simulations of the growth of *E. coli* under different extracellular glucose concentrations. FBA was performed based on the core *E. coli* metabolic model [52].

Maximise λ

Subject to (i) $S\boldsymbol{v} = 0$

(ii) $v^L \leq \boldsymbol{v} \leq v^U$

(iii) $w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1$ Equation 5-43

(iv) $w_c^* v_c + \left(w_f^* v_f + w_r^* v_r + b^* \lambda\right) = \phi_{max}^g / \phi_{max}^o$

(v) $w_c^* v_c = \phi_{max}^g / \phi_{max}^o - 1 \; (if \; \lambda \geq \lambda_{ac})$

S is the stoichiometric matrix; $\boldsymbol{v}$ is a column vector comprising the reactions/fluxes included in the metabolic model. v^L and v^U represent the minimum and maximum attainable fluxes, respectively. v_c is represented by the first enzymatic step of glycolysis, GLCpts, v_f is the acetate synthesis reaction ACKr, v_r adopts the first reaction after biomass withdrawal in the TCA cycle, AKGDH. λ is the growth rate and refers to the biomass reaction. λ_{ac} is the threshold growth rate, above which acetate overflow occurs. The determination of model parameters (w_c^*, w_f^*, w_r^*, b^*, ϕ_{max}^o and ϕ_{max}^g) is detailed in Appendix III, Supplementary Text. The lower bound of glucose and oxygen exchange fluxes were set to -1000 to avoid artificial control and to allow prediction of their uptake flux. Enzymatic reactions ICL, MALS, FRD7 were switched off under aerobic-glucose conditions according to [46]. Irrelevant exchange fluxes around the pyruvate node (EX_pyr(e), EX_lac(e), EX_acald(e), EX_etoh(e)) and the α-ketoglutarate node (EX_glu_L(e) and EX_akg(e)) were closed. Maximum carbon flow to the PP pathway was

limited to 67% and 42% of the overall carbon intake for ML308 and NCM3722, respectively according to ref. [28]. Extracellular glucose concentration was adopted as sole model input (varies from 0.1 - 75 μM for NCM3722 and 0.1-26 μM for ML308).

The optimisation model was converted to an MILP problem and solved using the Gurobi optimizer (see Appendix III, Supplementary Text for details). FBA was solved via COBRA Toolbox [53]. All simulations were performed in MATLAB R2016a.

Experimental data used to compare with the simulation results were obtained from Figure 1 of ref. [14], Figure 3B of ref. [19] and SI Figure 1 of ref. [31] for NCM3722 and from Table 7 of ref. [29], Figure 3b of ref. [30] and Figure 1 of ref. [32] for ML308.

References

1. Monod J: **The growth of bacterial cultures**. *Annu Rev Microbiol* 1949, **3**:371–394.
2. Kovárová-Kovar K, Egli T: **Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics**. *Microbiol Mol Biol Rev* 1998, **62**:646–666.
3. Tsipa A, Koutinas M, Usaku C, Mantalaris A: **Optimal bioprocess design through a gene regulatory network – growth kinetic hybrid model: towards replacing Monod kinetics**. *Metab Eng* 2018, **48**:129–137.
4. Orhon D, Cokgor EU, Insel G, Karahan O, Katipoglu T: **Validity of Monod kinetics at different sludge ages – peptone biodegradation under aerobic conditions**. *Bioresour Technol* 2009, **100**:5678–5686.
5. Pala-Ozkok I, Rehman A, Yagci N, Ubay-Cokgor E, Jonas D, Orhon D: **Characteristics of mixed microbial culture at different sludge ages: Effect on variable kinetics for substrate utilization**. *Bioresour Technol* 2012, **126**:274–282.
6. Wick LM, Weilenmann H, Egli T: **The apparent clock-like evolution of Escherichia coli in glucose-limited chemostats is reproducible at large but not at small population sizes and can be explained with Monod kinetics**. *Microbiology* 2002, **148**:2889–2902.
7. Franchini AG, Egli T: **Global gene expression in Escherichia coli K-12 during short-term and long-term adaptation to glucose-limited continuous culture conditions**. *Microbiology* 2006, **152**:2111–2127.
8. Füchslin HP, Schneider C, Egli T: **In glucose-limited continuous culture the minimum substrate concentration for growth, s min, is crucial in the competition between the enterobacterium Escherichia coli and Chelatobacter heintzii, an environmentally abundant bacterium**. *ISME J* 2012, **6**:777.
9. Insel G, Celikyilmaz G, Ucisik-Akkaya E, Yesiladali K, Cakar ZP, Tamerler C, Orhon D: **Respirometric evaluation and modeling of glucose utilization by Escherichia coli under aerobic and mesophilic cultivation conditions**. *Biotechnol Bioeng* 2007, **96**:94–105.
10. Varma A, Boesch BW, Palsson BØ: **Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates.** *Appl Environ Microbiol* 1993, **59**:2465–2473.
11. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010,

28:245–248.
12. Varma A, Palsson BØ: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Appl Environ Microbiol* 1994, **60**:3724–3731.
13. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics**. *Mol Syst Biol* 2009, **5**.
14. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
15. Thiele I, Jamshidi N, Fleming RMT, Palsson BO: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**.
16. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN: **Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity**. *P Natl Acad Sci USA* 2007, **104**.
17. Zhuang K, Vemuri GN, Mahadevan R: **Economics of membrane occupancy and respiro-fermentation**. *Mol Syst Biol* 2011, **7**.
18. Goelzer A, Fromion V: **Bacterial growth rate reflects a bottleneck in resource allocation**. *Biochim Biophys Acta (BBA)-General Subj* 2011, **1810**:978–988.
19. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
20. Thiele I, Fleming RMT, Que R, Bordbar A, Diep D, Palsson BO: **Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage**. *PLoS One* 2012, **7**:e45635.
21. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, et al.: **In silico method for modelling metabolism and gene product expression at genome scale**. *Nat Commun* 2012, **3**:929.
22. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**:693–693.
23. Nilsson A, Nielsen J, Palsson BO: **Metabolic models of protein allocation call for the kinetome**. *Cell Syst* 2017, **5**:538–541.
24. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T: **Interdependence of cell growth and gene expression: origins and consequences**. *Science* 2010, **330**:1099–1102.
25. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang Y-P, Lenz P, Yan D, Hwa T: **Coordination of bacterial proteome with metabolism by cyclic AMP signalling**. *Nature* 2013, **500**:301–306.
26. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR: **Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria**. *Mol Syst Biol* 2015, **11**.
27. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R: **Proteome reallocation in Escherichia coli with increasing specific growth rate**. *Mol BioSyst* 2015, **11**:1184–1193.
28. Zeng H, Yang A: **Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways**. *BMC Syst Biol* 2019, **13**:3.
29. Holms H: **Flux analysis and control of the central metabolic pathways in Escherichia coli.** *FEMS Microbiol Rev* 1996, **19**:85–116.
30. Noel JT, Cox B, Narang A: **Identification of the growth-limiting step in continuous cultures from initial rates measured in response to substrate-excess conditions**. *arXiv Prepr q-bio/0509013* 2005,
31. Bren A, Hart Y, Dekel E, Koster D, Alon U: **The last generation of bacterial growth in limiting nutrient**. *BMC Syst Biol* 2013, **7**:27.
32. Koch AL, Houston Wang C: **How close to the theoretical diffusion limit do bacterial uptake systems function?** *Arch Microbiol* 1982, **131**:36–42.

33. Chen Y, Nielsen J: **Energy metabolism controls phenotypes by protein efficiency and allocation**. *Proc Natl Acad Sci* 2019, **116**:17592–17597.
34. Snoep JL, Mrwebi M, Schuurmans JM, Rohwer JM, de Mattos MJ: **Control of specific growth rate in Saccharomyces cerevisiae**. *Microbiology* 2009, **155**:1699–1707.
35. Liu Y: **A simple thermodynamic approach for derivation of a general Monod equation for microbial growth**. *Biochem Eng J* 2006, **31**:102–105.
36. Liu Y: **Overview of some theoretical approaches for derivation of the Monod equation**. *Appl Microbiol Biotechnol* 2007, **73**:1241–1250.
37. Vazquez A, Oltvai ZN: **Macromolecular crowding explains overflow metabolism in cells**. *Sci Rep* 2016, **6**:31007.
38. Hanly TJ, Henson MA: **Dynamic metabolic modeling of a microaerobic yeast co-culture: predicting and optimizing ethanol production from glucose/xylose mixtures**. *Biotechnol Biofuels* 2013, **6**:44.
39. Henson MA, Hanly TJ: **Dynamic flux balance analysis for synthetic microbial communities**. *IET Syst Biol* 2014, **8**:214–229.
40. Hjersted JL, Henson MA: **Optimization of fed-batch Saccharomyces cerevisiae fermentation using dynamic flux balance models**. *Biotechnol Prog* 2006, **22**:1239–1248.
41. Hjersted JL, Henson MA, Mahadevan R: **Genome-scale analysis of Saccharomyces cerevisiae metabolism and ethanol production in fed-batch culture**. *Biotechnol Bioeng* 2007, **97**:1190–1204.
42. Hanly TJ, Henson MA: **Dynamic model-based analysis of furfural and HMF detoxification by pure and mixed batch cultures of S. cerevisiae and S. stipitis**. *Biotechnol Bioeng* 2014, **111**:272–284.
43. Gosset G: **Improvement of Escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system**. *Microb Cell Fact* 2005, **4**:14.
44. Stock JB, Waygood EB, Meadow ND, Postma PW, Roseman S: **Sugar transport by the bacterial phosphotransferase system. The glucose receptors of the Salmonella typhimurium phosphotransferase system.** *J Biol Chem* 1982, **257**:14543–14552.
45. Zeng H, Yang A: **Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant Escherichia coli**. *Biotechnol Bioeng* 2019, **116**:1484-1495.
46. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
47. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011**. *Mol Syst Biol* 2011, **7**:535.
48. Mo ML, Palsson BØ, Herrgård MJ: **Connecting extracellular metabolomic measurements to intracellular flux states in yeast**. *BMC Syst Biol* 2009, **3**:37.
49. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93–121.
50. Steinsiek S, Bettenbrock K: **Glucose transport in Escherichia coli mutant strains with defects in sugar transport systems**. *J Bacteriol* 2012, **194**:5897–5908.
51. Pirt SJ: **The maintenance energy of bacteria in growing cultures**. *Proc R Soc Lond B* 1965, **163**:224–231.
52. Orth JD, Palsson BØ, Fleming RMT: **Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide**. *EcoSal Plus* 2010, **4**.
53. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0**. *Nat Protoc* 2011, **6**:1290.

Chapter 6 Further applications and testing of proteome allocation-integrated metabolic models

6.1 Introduction

In Chapters 3, 4 and 5 we have shown that resource allocation and its mathematical formulation (e.g. proteome allocation constraints) effectively improve the predictions of the metabolic models of *E. coli* for a wide range of growth conditions, in particular the metabolic shift from full respiration (without producing fermentative by-products) to the overflow metabolism. Chapter 6 extends these studies by exploring the potential importance and limitations of resource allocation in deciphering practically relevant and often more complex systems.

In particular, we present three case studies (Sections 6.2 to 6.4), namely (i) NAD^+-dependent formate-dehydrogenase recombinant *E. coli*; (ii) engineered coculture of GFP recombinant *E. coli* and *Acinetobacter baylyi* ADP1 with acetate cross-feeding and (iii) lactate-producing human mesenchymal stem cells (hMSCs). Cases (i) and (ii) confer examples of synthetic recombinant *E. coli* systems with specific engineering targets, aiming to reduce the growth-inhibiting acetate production to improve heterologous protein production, respectively. In the first case, the proteome allocation model established earlier in the thesis for recombinant *E. coli* (see Chapter 4) was used to investigate the impact of formate metabolism on cell growth and acetate production. In the second case, the recombinant modelling framework was modified to simulate the overall GFP production in the coculture at different GFP expression levels, seeking an optimum operating state. Case (iii) inspects the applicability of overflow metabolism-targeted proteome allocation theory in resolving the presumably overflow-related 'Warburg effect' by combining the experimental measurements with flux estimations.

From case (i), model simulation finds that the presence of alternative energy pathways with lower enzyme cost per ATP flux is key to alleviating acetate secretion. From case (ii), the prediction of final GFP concentrations at various expression levels within the feasible range shows that the highest GFP productivity occurs at a moderate expression strength; their predicted non-linear relationship corresponds to the trade-off between the heterologous protein synthesis and the growth of the host cell. Finally, from case (iii), flux estimation using the experimentally determined lactate production-growth rate profile suggests that lactate overflow in hMSCs may not necessarily link to a reduced respiration activity, which leads to the discussion of potential limitation of the proteome allocation.

Overall, Chapter 6 systematically justifies several intriguing engineering ideas (e.g. the auxiliary substrate concept and coculture advantages) of metabolic engineering and synthetic biology for performance enhancement and offers a starting point for exploring the role of the modelling approach developed in this thesis in studying more complex cellular systems.

The supplementary information of this chapter can be found in Appendix IV.

6.2 The effectiveness of formate metabolism in reducing the overflow metabolism in NAD^+-dependent formate dehydrogenase recombinant *Escherichia coli*

6.2.1 Formate being an auxiliary substrate

The idea of using formate as an auxiliary energy substrate to improve growth and biomass yield was proposed several decades ago [1–5]. Formate is known as an NADH-generating substrate, which enhances the energy biogenesis via increasing intracellular NADH level [6]. NADH is produced during formate oxidation catalysed by the NAD^+-dependent formate dehydrogenase (nad-Fdh):

$$\text{Formate} + \text{NAD} \xrightarrow{FDH} \text{CO}_2 + \text{NADH}$$

Addition of formate to a medium containing carbon sources, e.g. glucose has proven to increase the production of microbes carrying homologous nad-Fdh (e.g. yeasts). For example, formate was used to enhance glycerol production in anaerobic glucose-limited chemostat cultures of *Saccharomyces cerevisiae* [7], increase the penicillin G production and biomass yield for *Penicillium chrysogenum* [6] and improve lipid production in three different oleaginous yeasts [8]. For organisms that do not have native nad-Fdh, the expression of heterologous *fdh* gene enables non-formate-assimilating strains to utilize formate. For instance, *Corynebacterium glutamicum* was genetically modified to consume formate in a glucose-formate mixture, resulting in 20% more succinate production [9]. Similarly in a succinate-producing *E. coli*, the expression of recombinant nad-Fdh effectively reduces the concentration of by-product formate and increases the succinate yield [10]. Despite these successes, biological mechanisms including (i) how formate metabolism alters the intracellular redox balance; (ii) how formate affects the energy pathways and (iii) how the combination of the (i) and (ii) changes the overall cellular physiologies have not been systematically investigated. One recent attempt was made to

explore how formate metabolism changes the metabolic fluxes using flux balance analysis (FBA) and the genome-scale metabolic model (GEM) [11]. However, this simple model is not able to offer biological insights beyond mass conservation. In this chapter, we would like to study quantitatively the impact of formate metabolism on cell growth and energy biogenesis using previously developed metabolic models incorporating proteome allocation constraints (Chapter 3 and Chapter 4). More specifically, we intend to achieve the following objectives:

(i) Developing the proteome allocation constraint for nad-Fdh recombinant E. coli

Assuming formate can enhance ATP synthesis through increasing NADH level, we posit the formate pathway as a new energy pathway in addition to the existing respiration and (aerobic) fermentation pathways. This involves incorporating the new energy pathway in the proteome allocation constraint and quantifying the corresponding proteomic cost.

(ii) Predicting the acetate production in nad-Fdh recombinant E. coli grown aerobically on glucose-formate mixed minimal medium

Having constructed the nad-Fdh integrated metabolic model where the energy sector has been extended to include formate oxidation, the question to answer is then how it would affect the prediction of the acetate production for cell grown aerobically on formate/glucose mixed minimal medium.

6.2.2 Proteome allocation model incorporating formate metabolism

6.2.2.1 Introducing a formate sector to the proteome allocation constraint

To model the growth of nad-Fdh recombinant *E. coli* in a glucose/formate mixed medium, the energy sector (ϕ_E) of the previously established proteome allocation constraint (Equation 3-6, Chapter 3) for predicting the overflow metabolism in *E. coli* was extended to include a formate sector ($\phi_{for} = w_{for} v_{for}$)

$$w_{for}v_{for} + w_f v_f + w_r v_r + b\lambda \leq \phi_{max} \quad \text{Equation 6-1}$$

w_{for}, w_f, w_r and b are proteomic cost parameters for formate pathway, fermentation, respiration and biomass synthesis, respectively. $b\lambda$ represents the biomass synthesis sector, primarily comprising ribosomal proteins and anabolic enzymes. The term $(w_{for}v_{for} + w_f v_f + w_r v_r)$ constitutes the new energy sector, comprising three alternative energy pathways:

- Formate pathway, converting formate to CO_2 and ATP. The first step of formate oxidation is catalysed by nad-Fdh and the rest is finalised by enzymes from the oxidative phosphorylation (OXP) system.
- (Aerobic) fermentation, converting glucose to acetate and ATP, catalysed by enzymes in the glycolysis, acetate pathway and completed by OXP.
- Respiration, converting glucose to CO_2 and ATP, catalysed by enzymes in glycolysis, TCA cycle and finalised by OXP.

ϕ_{max} denotes the maximum amount of proteome available to energy and biomass synthesis sectors. Dividing both sides of Equation 6-1 by ϕ_{max} gives:

$$w^*_{for}v_{for} + w^*_f v_f + w^*_r v_r + b^*\lambda \leq 1 \quad \text{Equation 6-2}$$

where $w^*_{for} \equiv \frac{w_{for}}{\phi_{max}}$, $w^*_f \equiv \frac{w_f}{\phi_{max}}$, $w^*_r \equiv \frac{w_r}{\phi_{max}}$ and $b^* \equiv \frac{b}{\phi_{max}}$. w^*_{for}, w^*_f, w^*_r and b^* are the normalised proteomic cost parameters.

6.2.2.2 Adding the formate oxidation reaction

The enzymatic reaction catalysed by nad-Fdh (named as 'FDH-NAD' in the metabolic model) was added into the genome-scale metabolic model of *E. coli*, *i*AF1260 [12]

FDH-NAD: formate[c] + nad[c] → co2[c] + nadh[c] Equation 6-3

where [c] denotes cytoplasmic as nad-Fdh can present and function in the cytoplasm [6].

6.2.2.3 Model simulation

Growth simulations of nad-Fdh recombinant *E. coli* were realised by solving the following linear programming (LP) problem via FBA [13]

Maximise λ

Subject to (i) $S\boldsymbol{v} = 0$

(ii) $v^L \leq \boldsymbol{v} \leq v^U$ Equation 6-4

(iii) $w^*_{for} v_{for} + w^*_f v_f + w^*_r v_r + b^* \lambda \leq 1$

S is the stoichiometric matrix with rows representing metabolites and columns representing fluxes. $\boldsymbol{v}$ is a column vector comprising reactions in the metabolic model. v^L and v^U define the lower and upper bounds of metabolic fluxes. w^*_{for}, w^*_f, w^*_r and b^* are the normalised proteomic cost parameters. v_{for} is represented by the first step of formate oxidation, FDH-NAD. The fermentation flux v_f is represented by enolase (ENO) and the respiration flux v_r is represented by citrate synthase (CS). λ is the growth rate. NAD^+-independent native formate dehydrogenase FDH4pp and FDH5pp were inactivated. The upper bound of the oxygen uptake rate was set to 1000 mmol/gDW/h to avoid arbitrary control and to allow prediction of its optimal value by the model. Nitrate exchange flux EX_no3(e) was closed to allow ammonium being the nitrogen source from the minimal medium. NADH16pp was assumed to be the only active NADH dehydrogenase of the electron transport chain. Glycogenesis and hydrogen-producing pathways were regulated to be off. Redundant loops of proton, glucose and ATP were manually eliminated. Glucose uptake rate and formate uptake rate (EX_glc(e) and EX_for(e)) were specified and the growth rate was maximised in all simulations.

6.2.3 Results and discussion

6.2.3.1 Determining the proteomic cost parameter w_{for}^*

To simulate cell growth on a mixed glucose/formate minimal medium, firstly we need to determine the new parameter w_{for}^*. By definition, w_{for}^* presents the proteomic cost of the complete formate pathway, which should include nad-Fdh and enzymes in OXP. We did not find a direct measurement of w_{for}^* in the literature. Alternatively, we use w_f^* and w_r^* (known parameters determined by fitting to the acetate line [14], Table 6-1) to estimate the unknown parameter w_{for}^*.

Table 6-1 Proteomic cost parameters fitted to growth data of wild-type *E. coli* NCM3722 [15]

Proteomic cost parameters	w_f^*	w_r^*	b^*	w_{for}^*
Values	0.0289	0.1326	0.1	$0.5, 1\ or\ 1.5\ w_f^*$

It is worth noting that respiration is the longest pathway (regarding the number of enzymes) among three energy pathways and carries the highest proteomic cost. Fermentation uses the acetate pathway (instead of the TCA cycle), which requires only two enzymes to obtain one mole of ATP from one mole of pyruvate. The formate pathway requires one enzyme to oxidize formate, with its product NADH being further converted to ATP by OXP enzymes. Given this rough analysis of pathway enzymes, we reached a preliminary proposition that the enzymatic cost of formate pathway is, like fermentation, probably lower than respiration, i.e. $w_{for}^* < w_r^*$. To test this hypothesis, we collected proteomic data of aerobically grown wild-type *E. coli*, containing relative mass distributions of 74 protein complexes of the central carbon metabolism (including primarily those in glycolysis, TCA cycle and OXP) and the ATP flux data at four different growth rates [15]. For each growth rate, the summed abundance of the 74 protein complexes (in mass percentage over the total proteins) is partitioned into those devoted to fermentation, respiration and biomass synthesis based on network stoichiometry, which

provides (estimated) proteome fractions of respiration and fermentation pathways in percentage. OXP proteins are a subset of respiration and fermentation proteins. The proteome fraction of OXP is retrieved by adding up components of proteome fractions of respiration and fermentation that are annotated as OXP. On the other hand, the ATP flux data is classified as ATP by substrate-level phosphorylation, ATP flux by respiration, ATP flux by fermentation, ATP flux by OXP. For ATP flux by OXP, it is further divided into ATP flux by NADH, ATP flux by NADPH and ATP flux by $FADH_2$, where the ATP flux by NADH is used to estimate the ATP production of the formate pathway. There are in total four sets of proteomic and ATP flux data corresponding to four distinct growth rates.

The proteomic and flux datasets were used to investigate the proteomic cost (in proteome fraction per ATP flux) of different energy pathways. For fermentation (respiration), the ATP flux by fermentation (respiration) was plotted against the proteome fraction of fermentation (respiration), as shown by the blue (orange) dots in Figure 6-1. For OXP, the ATP flux by OXP was plotted against the proteome fraction of OXP, as shown by the grey dots. For the formate pathway, the cost of the formate oxidation (which requires only one enzyme) was considered negligible compared with the subsequent NADH oxidation catalysed by OXP components, i.e. the proteome fraction of the formate pathway was assumed to equal to the proteome fraction of OXP. Consequently, the ATP flux by NADH was plotted against the proteome fraction of OXP to estimate the proteomic cost of the formate pathway, as shown by the yellow dots in Figure 6-1.

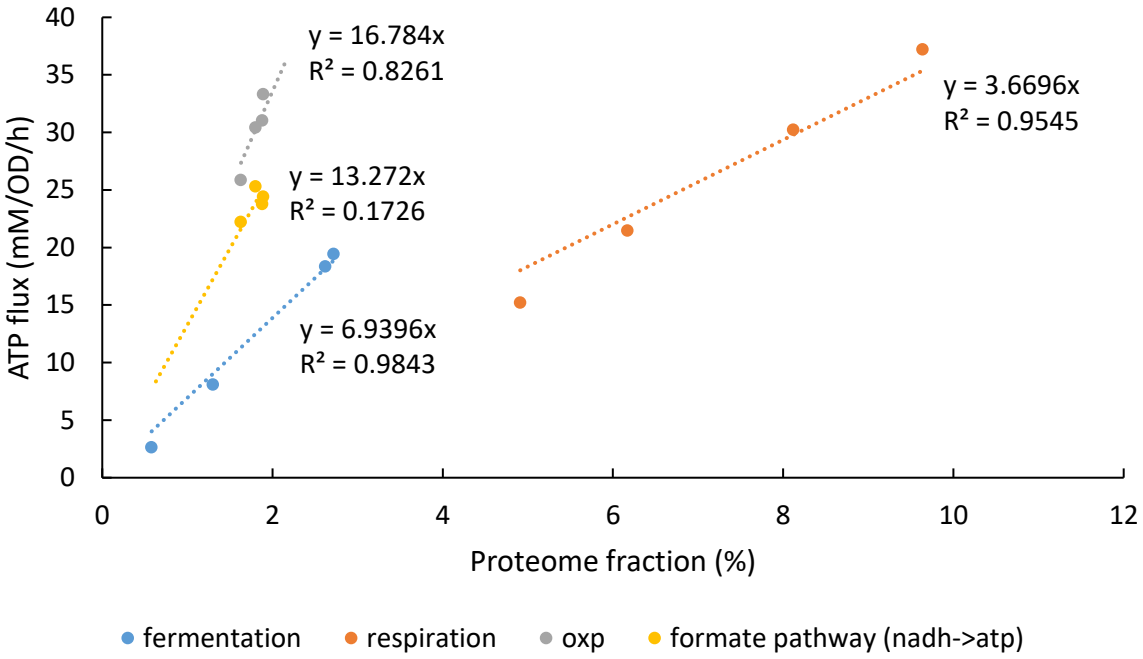


Figure 6-1 Illustration of the proteomic cost of different energy production routes. Dashed lines and equations show the results of the linear regression of calculated data points (4 points for each set). Note that for the formate pathway, the enzyme cost of first-step formate oxidation was considered negligible compared with the subsequent NADH oxidation catalysed by OXP components.

The slope of the regression line denotes mole of ATP generated per unit proteome fraction, which is a direct indicator of the proteomic efficiency – the higher the slope, the higher the efficiency (i.e. more ATP is generated per unit mass of protein). The slope of the blue line shows the proteomic efficiency of the fermentation pathway, which is about two times higher than that of respiration (slope of the orange line). To obtain the data points for formate pathway, although we can use a ratio of 1 NADH to 2 ATP to calculate the ATP flux [16], it is difficult to differentiate how many proteomic fractions in OXP is responsible for converting NADH to ATP from those responsible for converting NADPH and $FADH_2$ to ATP. To resolve this problem, we made a simple, conservative (i.e. not underestimating the fraction) assumption that the proteome fraction needed for converting NADH to ATP is equal to the complete proteome fraction of OXP. The actual NADH-

associated proteomic fraction should be less than the assumed value, which would result in a steeper slope of the formate pathway.

From the above calculation, we find that the proteomic efficiency of OXP (grey), the formate pathway (yellow) and fermentation (blue) are higher than that of respiration, confirming the earlier proposition. More Interestingly, OXP and the formate pathway appear to be more efficient than fermentation; the proteomic efficiency of formate pathway is ~50% lower than that of fermentation.

6.2.3.2 Predicting acetate production using formate metabolism integrated proteome allocation model

The above finding suggests that w_{for}^* is possibly closer to w_f^* than to w_r^*. We then moved on to investigate the impact of varying w_{for}^* on the prediction of the acetate line (i.e. acetate production rate against growth rate profile). To do this, we ran FBA simulations with w_{for}^* set to (i) 1.5 times of w_f^*; (ii) equal to w_f^* and (iii) half of w_f^*. Glucose uptake rate changed from 0.5 to 15 mmol/gDW/h and the maximum formate uptake rate was constrained to 10 mmol/gDW/h.

With $w_{for}^* = 1.5w_f^*$, Figure 6-2(a) shows that the summation of the proteome fractions of energy and biosynthesis sectors gradually increases with increased growth rate, and become limited at around $\lambda = 0.6$ h^{-1}. Formate uptake stays at the maximum rate in the beginning (i.e. $\lambda < 0.6$ h^{-1}) and decreases upon the proteome limitation. The total of growth-supporting proteomic resources (i.e. $\phi_r + \phi_f + \phi_{for} + \phi_{bm}$) saturates before the threshold point (λ_{ac} = 0.8 h^{-1}) where acetate production occurs. Formate uptake rate declines to zero at λ_{ac}. Comparing the predicted acetate production (with formate addition) with the experimental data (without formate supplement), no obvious deviation is shown (Figure 6-2(b)).

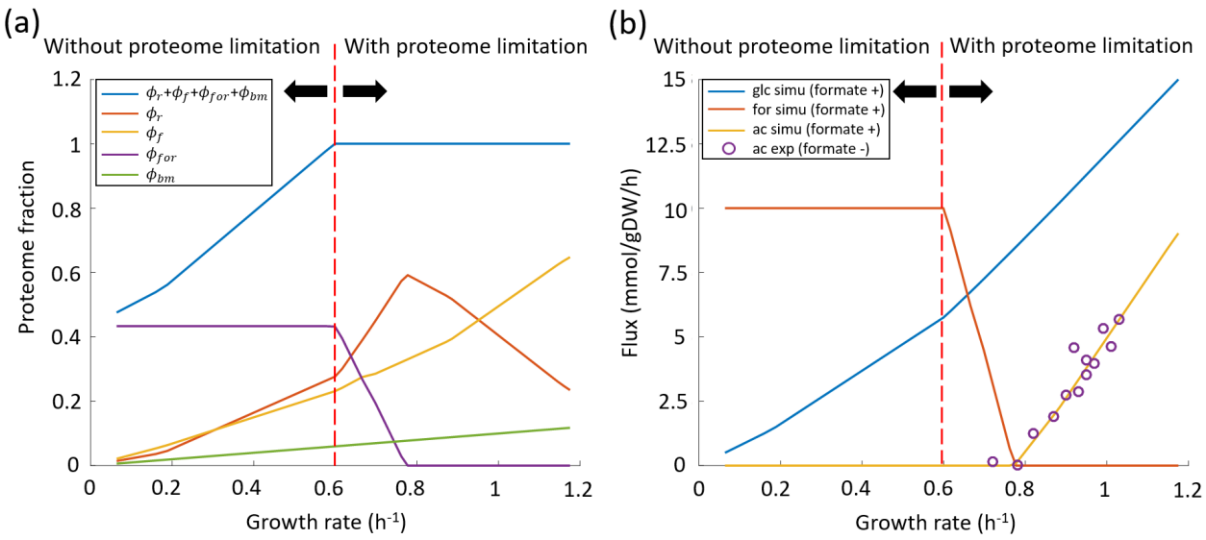


Figure 6-2. Simulation results for $w^*_{for} = 1.5w^*_f$. **(a)** Model simulation of proteome fractions of different proteomic sectors with different growth rates. **(b)** Model simulation of glucose uptake rate (glc simu), formate uptake rate (for simu) and acetate production rate (ac simu) with different growth rates for *E. coli* grown aerobically on glucose/formate minimal medium (formate +). Experimental data of the acetate production without formate addition (ac exp, formate -) were obtained from Figure 1 of ref. [15]. The red dashed line indicates proteome limitation.

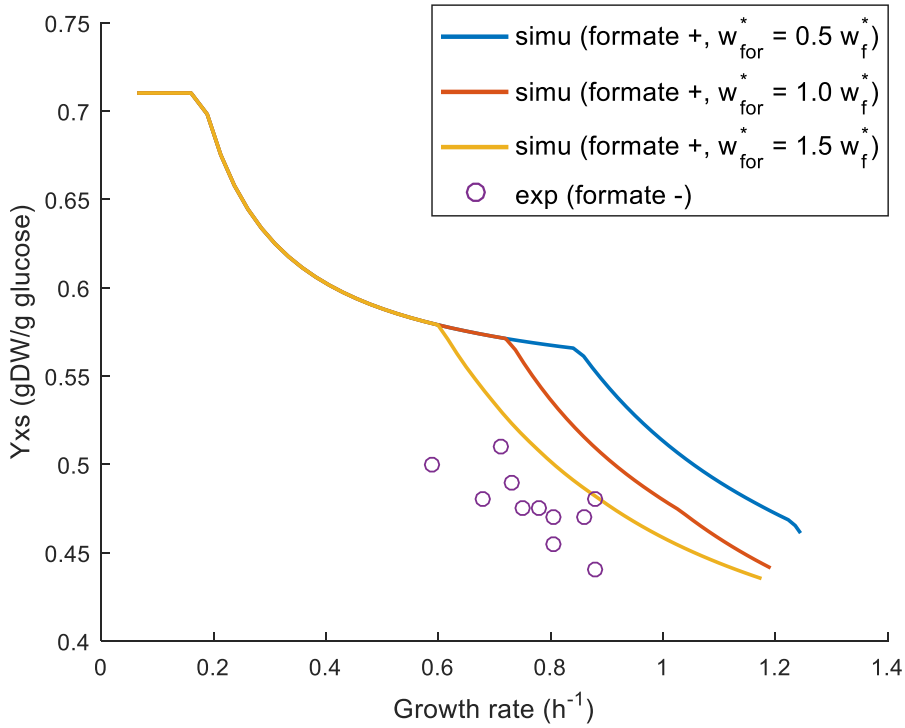


Figure 6-3. Comparison between simulated biomass yield on glucose Y_{xs} (formate +) for $w^*_{for} = 0.5w^*_f$, $w^*_{for} = w^*_f$ and $w^*_{for} = 1.5w^*_f$ and measured yield (formate -) at different growth rates. Experiential data were obtained from ref. [15].

In this case, despite formate metabolism does not alleviate the acetate production, model simulation confirms the experimentally observed effect of formate in increasing biomass yield. The oxidation of formate increases the pool size of cytosolic NADH, which reduces the catabolic requirement from glucose. More glucose can be directed to biomass synthesis (anabolism) therefore at the same glucose uptake rate, the system can obtain a higher growth rate, which means a higher biomass yield on glucose (Figure 6-3).

With $w^*_{for} = w^*_f$, proteome distribution shows that the proteome limitation occurs at $\lambda = 0.72$ h^{-1} (Figure 6-4(a)). The acetate line is predicted to shift to the left, implying an increased acetate production at the same growth rate (Figure 6-4(b)). In the previous case ($w^*_{for} = 1.5w^*_f$), the proteome limitation is predicted to emerge at $\lambda = 0.6$ h^{-1}. In this case with the reduced w^*_{for}, the threshold growth rate for the proteome limitation is increased to 0.72 h^{-1}, and the overflow metabolism is predicted to occur upon the onset of the proteome limitation. At $\lambda \sim 1.0$ h^{-1}, a transitional state is predicted (the area between grey dashed lines, Figure 6-4(a)) where the maximal growth rate is achieved by restoring the respiration (reflected by the increase in ϕ_r) with slightly reduced fermentation (reflected by the slight drop in ϕ_f) and unchanged formate uptake (reflected by unchanged ϕ_{for}).

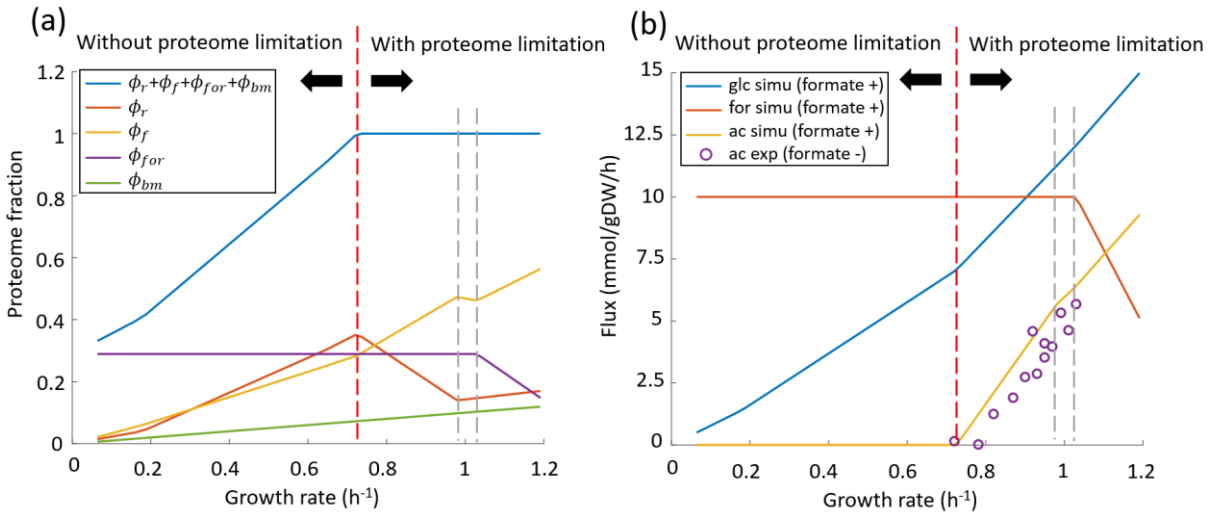


Figure 6-4. Simulation results for $w^*_{for} = w^*_f$. **(a)** Model simulation of proteome fractions of different proteomic sectors with different growth rates. **(b)** Model simulation of glucose uptake rate

(glc simu), formate uptake rate (for simu) and acetate production rate (ac simu) with different growth rates for *E. coli* grown aerobically on glucose/formate minimal medium (formate +). Experimental data of acetate production without formate addition (ac exp, formate -) were obtained from Figure 1 of ref. [15]. The red dashed line indicates proteome limitation. Grey dashed lines mark a transitional state (see main text).

Through the examination of the metabolic fluxes, we find that at $\lambda = 0.98$ h^{-1}, the TCA flux α-ketoglutarate dehydrogenase (AKGDH) drops to zero (AppxIV-Figure S1(a)). The TCA cycle functions not only for energy production but also generates important biomass precursors such as α-ketoacids. With a zero AKGDH flux, the upstream flux citrate synthase (CS) and the downstream flux succinyl-CoA synthetase (SUCOAS) need to increase to balance the intermediate metabolite α-ketoglutarate and succinyl-CoA with increased growth rates (AppxIV-Figure S1(a)). Consequently, the increased CS flux (at zero AKGDH flux) results in the transitional stage of ϕ_r and ϕ_f. For $\lambda \geq$1.03 h^{-1}, the formate pathway (ϕ_{for}) is predicted to decline, which releases the proteomic stress and allows ϕ_f to increase again with increased growth rates.

To further explain the reason why in this case the formate metabolism enhances acetate production, we looked into the structure of the optimization model and found that this seemingly counter-intuitive result again associates with the biomass yield. As maximising growth rate was set as the objective function, it means that the biomass yield was maximised for a fixed glucose uptake rate. Formate metabolism substitutes part of glucose catabolism, allowing more glucose carbon to be used for producing biomass. Compared with the previous case ($w_{for}^* = 1.5w_f^*$), the lowered proteomic cost of formate pathway leads to more savings on glucose carbon for energy production, resulting in higher biomass yield (Figure 6-3). However, in this case, the moderate proteomic cost (i.e. only comparable to the fermentation, $w_{for}^* = w_f^*$) does not make the formate pathway more advantageous than fermentation upon proteome limitation. Consequently, the cell is

predicted to reduce the proteomic-expensive respiration flux and to use the (equally) proteomic-efficient formate and fermentation pathways to produce energy. The co-utilisation of the formate and fermentation pathways with comparable proteomic costs ($w^*_{for} = w^*_f$) improves the biomass yield at the cost of increased acetate production.

Finally, with $w^*_{for} = 0.5\, w^*_f$, the cell is predicted to consume the maximal amount of formate at most of the time due to the significantly low cost of the formate pathway. The threshold growth rate of the proteome limitation shifts to $\lambda = 0.84$ h^{-1} (Figure 6-5(a)). Acetate production and decrease in respiration flux occur upon the proteome limitation. A right-shift of the acetate line can be observed, suggesting that acetate production is reduced at the same growth rate (Figure 6-5(b)). Similar to the case of $w^*_{for} = w^*_f$, a transitional region (between grey lines at λ~1.2 h^{-1}) can be predicted, where the decreased ϕ_r starts to restore due to zero AKGDH (AppxIV-Figure S1(b)). A decline in the formate uptake rate occurs at the higher end of the growth rate range.

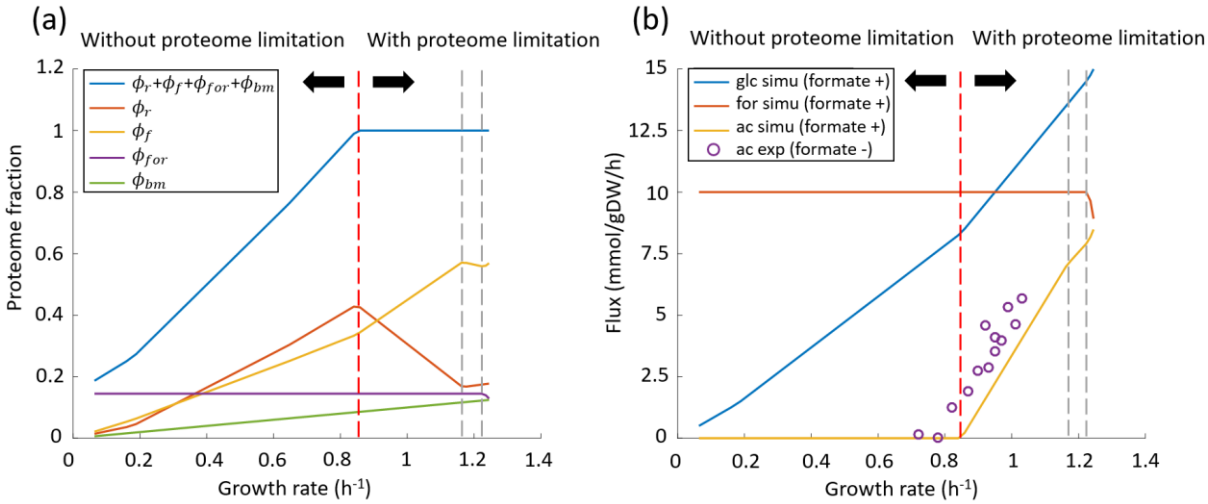


Figure 6-5. Simulation results for $w^*_{for} = 0.5 w^*_f$. **(a)** Model simulation of proteome fractions of different proteomic sectors with different growth rates. **(b)** Model simulation of glucose uptake rate (glc simu), formate uptake rate (for simu) and acetate production rate (ac simu) with different growth rates for *E. coli* grown aerobically on glucose/formate minimal medium (formate +). Experimental data of acetate flux without formate addition (ac exp, formate -) were obtained from

Figure 1 of ref. [15]. The red dashed line indicates proteome limitation. Grey dashed lines mark a transitional state.

Besides, the prediction of increased activity of OXP components (cytochrome oxidase and NADH dehydrogenase) upon formate addition supports the notion [6] that formate metabolism facilitates energy production by increasing the cytosolic NADH availability and ATP production (Figure 6-6).

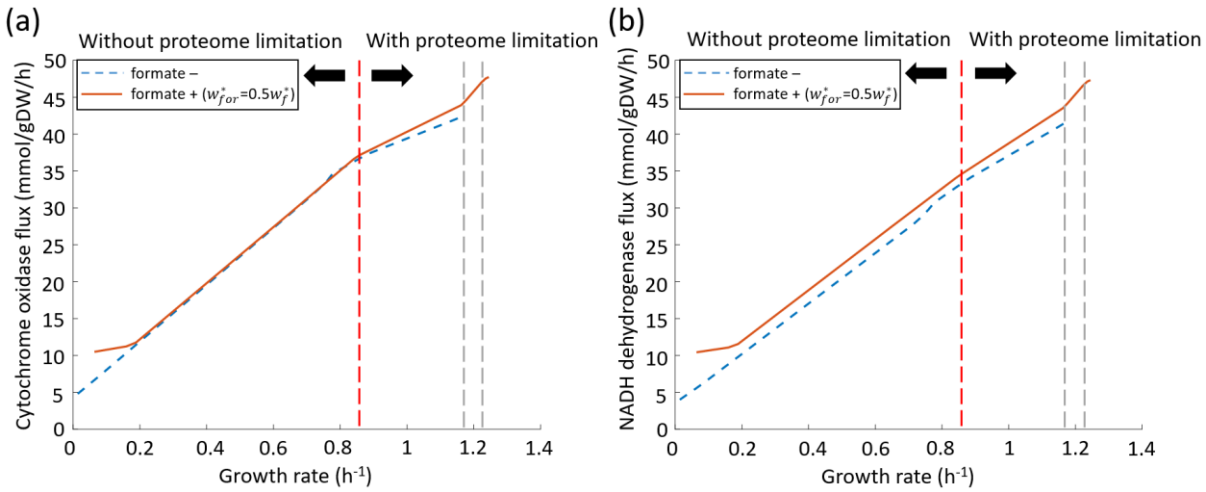


Figure 6-6. Comparison of the activity of OXP components **(a)** cytochrome oxidase and **(b)** NADH dehydrogenase for cells grown on glucose minimal medium with formate addition (formate +, $w^*_{for} = 0.5w^*_f$) and without formate addition (formate -). The red dashed line indicates proteome limitation. Grey dashed lines show a transitional state.

As mentioned above, when formate is imported, the cell can reach a higher growth rate (or biomass yield) upon the same GUR due to the increased NADH pool. This is reflected by the gap between the red curve (formate +) and the blue dashed curve (formate -) at the low and the high end of the growth rate range, i.e. the feasible growth rate range for a given range of GUR is extended by formate metabolism.

Overall, the above analysis suggests that the key to reducing the overflow metabolism via formate supplement lies in the relative proteomic efficiency (for producing a unit of ATP flux) between formate and fermentation pathways. To maximise the growth rate, the cell is predicted to uptake as many substrates (glucose and formate) as possible. The simulation

results suggest that as long as the proteomic cost of metabolising formate is not zero, the utilisation of formate can bring risks of (i) reaching the proteome limitation earlier than the case with no formate addition and (ii) an enhanced acetate production due to the more 'crowded' proteome allocation.

When formate is imported with a relatively high proteomic cost compared with the fermentation (i.e. $w_{for} = 1.5w_f$), the proteome limitation is predicted to emerge at $\lambda =$ 0.6 h^{-1} (Figure 6-2(a)). As w_{for} decreases, the threshold growth rate of proteome limitation delays to 0.72 and 0.84 h^{-1} (Figure 6-4(a) and Figure 6-5(a)). Varying w_{for} also affects the order of reduced proteomic sectors upon proteome limitation. For $w_{for} = 1.5w_f$, the first reduced sector is ϕ_{for} due to its high proteomic cost. The respiration sector (ϕ_r) starts to decline only after ϕ_{for} reaching zero. For both $w_{for} = w_f$ and $w_{for} = 0.5w_f$, the first reduced sector is ϕ_r, with ϕ_{for} decreasing later at very high growth rates.

As discussed above, the import of formate inevitably squeezes the proteome available for other proteomic sectors. Our model predicts three distinct impacts of the formate metabolism on the acetate line. One possible scenario is that the acetate line stays almost unaffected when the import of formate only occurs at non-proteome limited stage and quickly drops to zero upon proteome limitation (as shown in $w_{for} = 1.5w_f$ case). Another scenario corresponds to reduced acetate production. For $w_{for} = 0.5w_f$, formate uptake significantly delays the threshold growth rate of the proteome limitation. The low cost of the formate pathway also allows the respiration sector (flux) to decline first upon the proteome limitation. Together it leads to the prediction of the right-shift of the acetate line and reduced overflow metabolism. Finally, acetate production can be enhanced. For $w_{for} = w_f$, the acetate line is predicted to shift to the left as (i) the proteomic cost of the formate pathway is not low enough to fully substitute the role of fermentation and (ii) the

formate metabolism does not significantly postpone the threshold growth rate of the proteome limitation.

6.2.4 Summary of case (i)

We used a formate pathway-integrated proteome allocation model to theoretically analyse the potential impact of the addition of formate metabolism on acetate production. Although the presented patterns were determined by hypothetically specified w_{for}^{*} values, some general governing principles have been suggested. When the cell is proteome limited, any increase in growth rate leads to an increase in the biomass synthesis sector (ϕ_{bm}), which is achieved only through the reduction of other sectors. The sector that is chosen to be the first to reduce depends on (i) the presence of alternative pathways; (ii) the proteomic cost of growth-associated alternative pathways and (iii) whether the reduction of this sector contradicts the growth demand. In general, pathways with low protein costs are preferred to maintain and those with high protein costs (and with low-cost alternative routes) are chosen to be the first to reduce (to a point that still allows feasible growth).

6.3 Exploring the optimal recombinant protein expression level in a synthetic coculture

6.3.1 An engineered microbial coculture to improve the production of GFP

Based on the study of recombinant protein-producing *E. coli* in Chapter 4, an engineered coculture system of product-bearing *E. coli* and *Acinetobacter baylyi* ADP1 [17] was adopted to investigate the advantages of microbial consortia in improving product yield. Green Fluorescent Protein **(**GFP) was taken as a model heterologous product. The coculture comprises GFP-producing *E. coli* (EC-GFP) and GFP-producing *A. baylyi* (AB-GFP). The system was designed such that glucose was utilised by EC-GFP alone, with AB-GFP grown on the acetate secreted by EC-GFP (Figure 6-7). The high-affinity

gluconate permease, *gntT*, was deleted from AB-GFP to disable glucose consumption. A similar principle has been adopted for another synthetic coculture system where glycose dehydrogenase, *gcd*, knockout *A. baylyi* was used to eliminate acetate and formate (produced from lignocellulosic hydrolysates) to improve the growth of *Clostridium butyricum* [18].

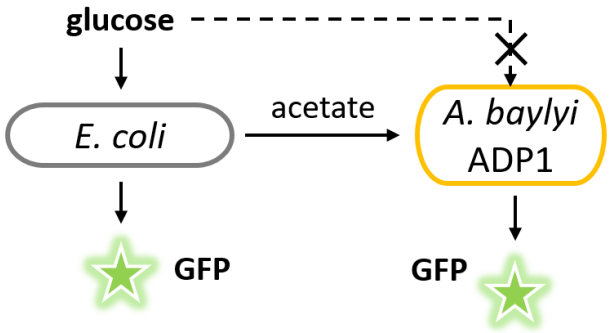


Figure 6-7. Illustration of the synthetic coculture of GFP-producing *E. coli* and GFP-producing *A. baylyi* ADP1.

6.3.2 Formulating metabolic models for EC-GFP and AB-GFP

The genome-scale metabolic model *i*AF1260 was taken as the template to build the metabolic model for EC-GFP. The GFP synthesis reaction was added to the metabolic matrix based on the previously reported reaction stoichiometry [19]:

22 glycine + 8 L-alanine + 18 L-valine + 21 L-leucine + 12 L-isoleucine + 10 L-serine + 2 L-cysteine + 16 L-threonine + 6 L-methionine + 10 L-proline + 12 L-phenylalanine + 11 L-tyrosine + 1 L-tryptophan + 9 L-histidine + 20 L-lysine + 6 L-arginine + 18 L-aspartate + 16 L-glutamate + 13 L-asparagine + 8 L-glutamine + 1029.13 ATP + 789.134 H2O → GFP + 1029.13 ADP + 1029.13 phosphate Equation 6-5

Following the approach introduced in Chapter 4, the proteome allocation constraint was adapted to include a GFP sector (ϕ_{GFP})

$$\phi_{GFP} + w_f v_f + w_r v_r + b\lambda \leq \phi_{max} \quad \text{Equation 6-6}$$

Dividing both sides of Equation 6-6 by ϕ_{max} gives

$$w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 - \frac{\phi_{GFP}}{\phi_{max}} \quad \text{Equation 6-7}$$

Parameters w_f^*, w_r^* and b^* and variable v_f, v_r and λ hold identical biological meaning as defined in the previous section (Equation 6-2) and were estimated using the measured acetate line of *E. coli* K-12 MG1655 [20]. Equation 6-7 was integrated into the model as the proteome allocation constraint. ϕ_{GFP} denotes the mass fraction of GFP over the total amount of proteins, which is referred to as the GFP expression level.

$$\phi_{GFP} = \frac{Mass\ of\ GFP}{Mass\ of\ total\ proteins} = \frac{v_{gfp} MW_{GFP}}{\lambda f_{TP}} \quad \text{Equation 6-8}$$

where v_{gfp} is the GFP synthesis flux in mmol/gDW/h; $MW_{GFP} = 27$ g/mmol is the molecular weight of GFP; f_{TP} is the mass ratio of total protein mass over the dry cell weight, which is roughly a constant (~ 60%) over a wide range of growth rates [21–23]. Equation 6-8 suggests that v_{GFP} is a function of growth rate λ and ϕ_{GFP}, which is taken as an additional constraint ($v_{gfp} - \frac{\phi_{GFP} \lambda f_{TP}}{MW_{GFP}} = 0$) to fix v_{gfp} by specifying ϕ_{GFP}.

Maintenance flux ATPM was assumed to be linearly correlated with ϕ_{GFP} [24]:

$$\text{ATPM} = 8.39 + 24.21 \frac{\phi_{GFP}}{6.45 \times 10^{-2}} \quad \text{Equation 6-9}$$

Similarly, GFP synthesis flux (Equation 6-5) was incorporated into the genome-scale metabolic model of *Acinetobacter baylyi* ADP1 (*i*Abaylyi^{v4}) [25] for AB-GFP. As we did not find evidence supporting that the growth of AB-GFP or AB is regulated by resource allocation, the growth and metabolism of AB-GFP were described by a classical FBA model without introducing the proteome allocation constraint. The prediction of the maximum steady-state growth of EC-GFP monoculture was found by solving the following FBA problem where glucose uptake rate was specified as input:

Maximise λ

Subject to (i) $S\boldsymbol{v} = 0$

(ii) $v^L \leq \boldsymbol{v} \leq v^U$ Equation 6-10

(iii) $v_{gfp} - \frac{\phi_{GFP}\lambda f_{TP}}{MW_{GFP}} = 0$

(iv) $w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 - \frac{\phi_{GFP}}{\phi_{max}}$

S is the stoichiometric matrix with rows representing metabolites and columns representing fluxes. $\boldsymbol{v}$ is a column vector comprising reactions in the metabolic model. v^L and v^U define the lower and upper bounds of metabolic fluxes. w_f^*, w_r^* and b^* are the normalised proteomic cost parameters. The batch growth of AB-GFP and EC-GFP coculture was simulated by DFBA [26], where the intracellular metabolism of EC-GFP is described by Equation 6-10 above while the intracellular metabolism of AB-GFP is described by a classical FBA problem, i.e. Equation 6-10 without constraints (iii) and (iv).

The growth of the two species was assumed to be independent (i.e. no community-level objective was defined); the two species would interact only through exchanging acetate via the shared chemical environment in which they grow. Objective functions for the two species were set to maximizing the growth rate. GFP production flux in AB-GFP ($v_{gfp,Y}$) was specified to a constant rate of 4.81×10^{-4} mmol/gDW/h (which is a level reported for GFP-bearing *E. coli* [27]; no experimental data of the specific GFP production rate was found for AB-GFP strain). Glucose uptake rate for EC-GFP (v_{glc}) and acetate uptake rate for AB-GFP (v_{ac}) were linked to the extracellular nutrient concentrations via the Michaelis Menten kinetics

$$v_{glc} = v_{glc,max} \frac{G}{K_g + G}$$ Equation 6-11

$$v_{ac} = v_{ac,max} \frac{A}{K_{ac} + A}$$ Equation 6-12

where G and A are the concentrations of extracellular glucose and acetate, respectively. $v_{glc,max}$ ($v_{ac,max}$) denotes the maximal uptake rate of glucose (acetate). K_g and K_{ac} are the kinetic constants. The dynamic environment of the batch reactor was modelled by the following ordinary differential equations (ODEs):

$$\frac{dX}{dt} = \lambda_X X \quad \text{Equation 6-13}$$

$$\frac{dY}{dt} = \lambda_Y Y \quad \text{Equation 6-14}$$

$$\frac{dG}{dt} = -v_{glc} X \quad \text{Equation 6-15}$$

$$\frac{dA}{dt} = v_{ac,ex} X - v_{ac} Y \quad \text{Equation 6-16}$$

$$\frac{dGFP}{dt} = v_{gfp,X} X + v_{gfp,Y} Y \quad \text{Equation 6-17}$$

where X, Y and GFP are the extracellular concentrations of EC-GFP, AB-GFP and GFP, respectively. λ_X (λ_Y) is the growth rate of EC-GFP (AB-GFP). $v_{ac,ex}$ is the predicted acetate production rate of EC-GFP. $v_{gfp,X}$ is the predicted GFP production rate of EC-GFP. The initial culture condition is set to $G_0 = 5\ mM$, $A_0 = 0\ mM$, $GFP_0 = 0\ mM$ and $X_0 = Y_0 = 0.1\ g/L$. The FBA problem was solved via COBRA toolbox and the ODEs were solved by the solver ode15s in MATLAB R2016a.

Table 6-2. Parameters used in FBA and DFBA simulations for EC-GFP and AB-GFP.

Parameter	$v_{glc,max}$	$v_{ac,max}$	K_g	K_{ac}
Value	5 mmol/gDW/h	5 mmol/gDW/h	0.01 mM	0.01 mM
Parameter	w_f^*	w_r^*	b^*	ϕ_{max}
Value	0.0577	0.1847	0.10	0.484

6.3.3 Results and discussion

6.3.3.1 FBA simulation of steady-state growth of EC-GFP

We varied the glucose uptake rate (GUR) from 0 to 10 mmol/gDW/h and the GFP expression level (ϕ_{GFP}) from 0 to 0.25 to investigate the growth physiologies of EC-GFP. A maximum growth rate of 0.74 h^{-1} was achieved at GUR =10 mmol/gDW/h with zero GFP production (Figure 6-8(a)). Highest acetate production rate occurs at the highest GUR and ϕ_{GFP} with almost halted growth compared to the maximum growth rate (Figure 6-8(b)). GFP production increases with increased GUR and peaks with moderate GFP expression level ($\phi_{GFP} \approx 0.12$) (Figure 6-8(c)).

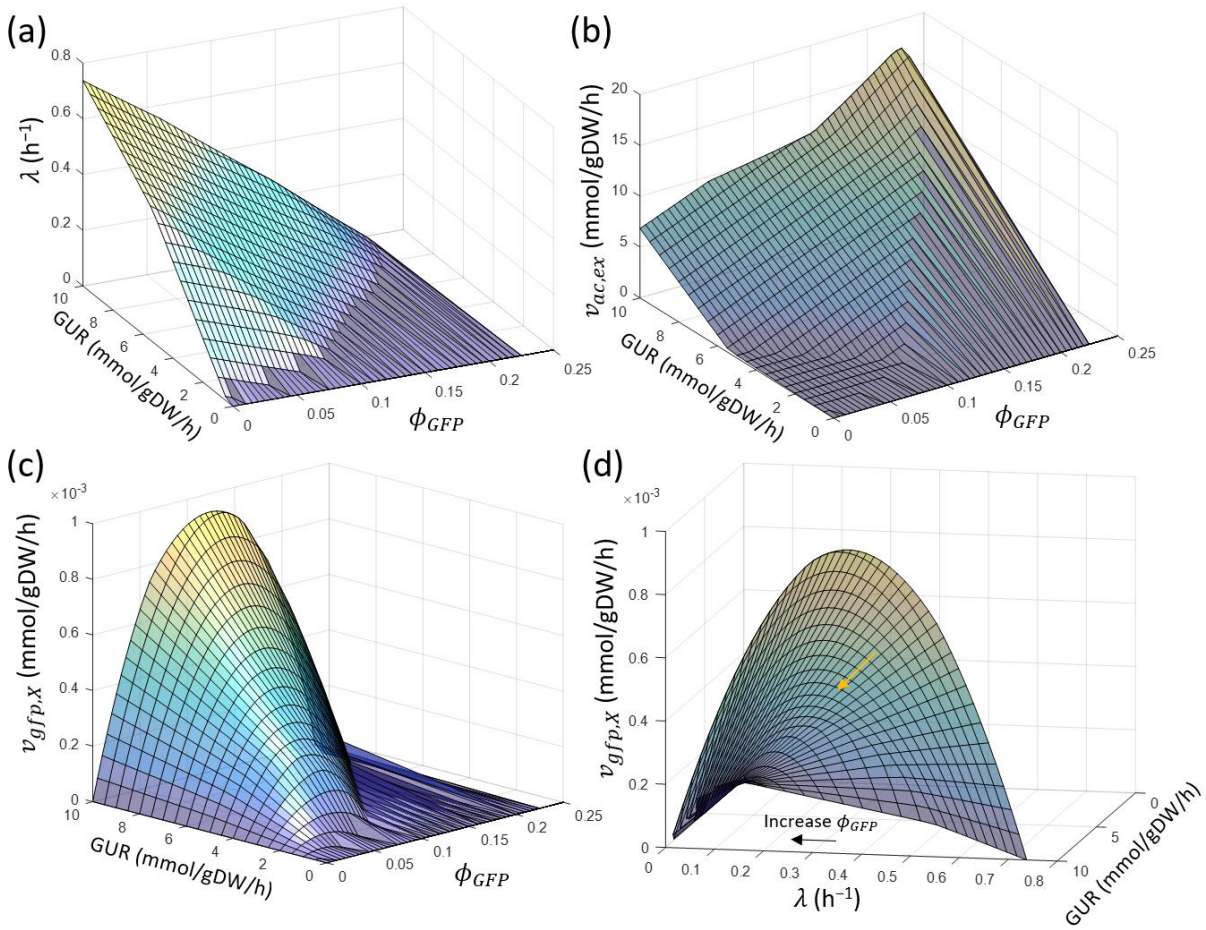


Figure 6-8. FBA simulation of steady-state growth of EC-GFP in glucose minimal medium. Model prediction of **(a)** maximal growth rate, **(b)** acetate production rate and **(c)** GFP production rate at different glucose uptake rates (GUR) and GFP expression levels (ϕ_{GFP}). **(d)** Model prediction of GFP production rate at different predicted maximal growth rates. The impact of GUR is shown by

the y-axis; the impact of increased ϕ_{GFP} is reflected by the decrease in λ. The yellow arrow shows the shrink of the 'bell-shape' correlation at decreasing GUR.

Correspondingly, GFP production rates show a non-linear 'bell-shape' correlation with the growth rate, and the bell gradually shrinks with decreased GUR (yellow arrow, Figure 6-8(d)). These results suggest that high expression levels of recombinant proteins associated with significant proteomic and energetic burdens can lead to growth retardation and reduce target productivity, i.e. higher expression level is not always advantageous. As discussed earlier in Chapter 4, there might be an optimal expression level that contributes the highest productivity by coordinating the production of the target product and the growth of the host cell.

6.3.3.2 DFBA simulation of the batch growth of a coculture of EC-GFP and AB-GFP

A synthetic coculture of EC-GFP and AB-GFP has proven to improve GFP yield [17]. We modelled this experimental system with GFP-integrated GEMs of *E. coli* and *A. baylyi*. The complexity of this coculture lies in the fact that the overall productivity of GFP depends on both EC-GFP and AB-GFP. As AB-GFP was engineered to only utilise acetate, its growth would be controlled by the amount of acetate secreted by EC-GFP. Higher acetate production is coupled with slow growth and low GFP production in EC-GFP (Figure 6-8 (a-c)), which can reduce GFP yield from *E. coli*. In Section 6.3.3.1, we found that a relatively moderate (as opposed to very high) ϕ_{GFP} confers the highest GFP productivity. To simulate the batch growth of EC-GFP and AB-GFP coculture, we took the median GUR (5 mmol/gDW/h) and varied ϕ_{GFP} from 0.01 to 0.15 to avoid infeasible growth. The system performance can be evaluated by the accumulated GFP at the end of the batch.

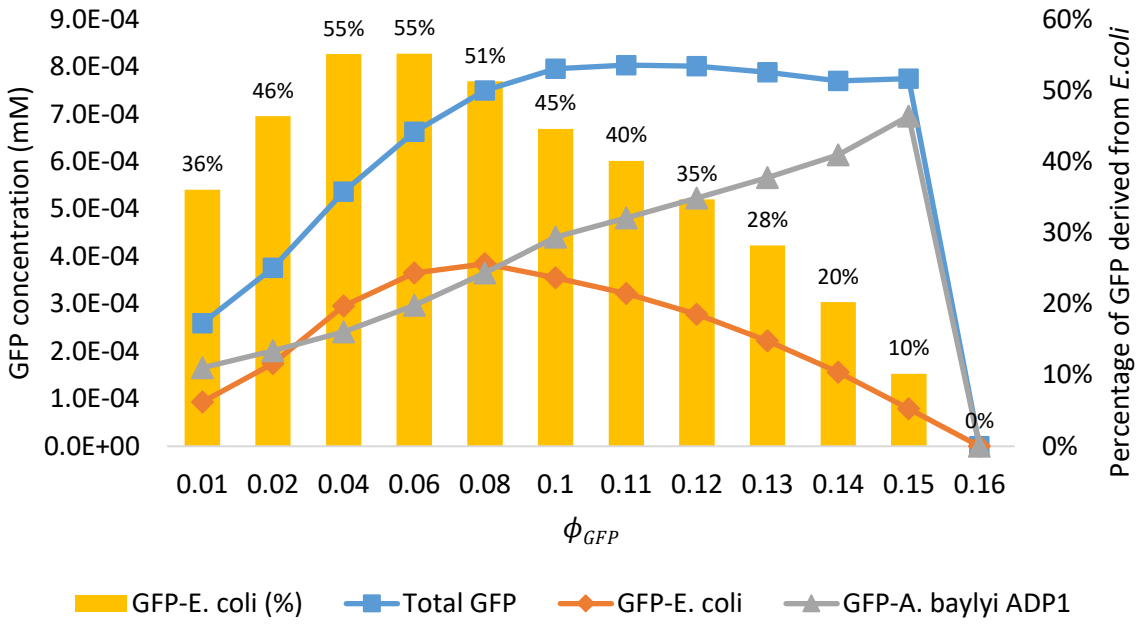


Figure 6-9. Model simulation of GFP production at the end of the batch in EC-GFP and AB-GFP coculture at different GFP expression levels (ϕ_{GFP}). The blue line with squares shows the total GFP concentration. The grey line with triangles shows the amount of GFP production by *E. coli*; the orange line with diamonds shows the amount of GFP production by *A. baylyi*. The yellow columns show the percentage of GFP derived from *E. coli*.

Batch simulation of the coculture shows that the total GFP productivity first increases with ϕ_{GFP}, but gradually declines at $\phi_{GFP} > 0.1$ (Figure 6-9). The reduction in total GFP corresponds to the decrease in the absolute amount of GFP production by *E. coli* as well as the decrease in the percentage of *E. coli*-derived GFP over the total GFP. At $\phi_{GFP} = 0.08$, GFP produced by *E. coli* equals to that from *A. baylyi*. After this point, the GFP productivity of *A. baylyi* exceeds that of *E. coli*. A slight increase in total GFP occurs ϕ_{GFP}~0.15 before complete growth cessation at higher expression levels. Although *E. coli* grows extremely slow at $\phi_{GFP} = 0.15$, the corresponding GFP synthesis flux and acetate production are substantial. Increased acetate availability facilitates the growth of AB-GFP (Figure 6-10(a)), which compensates the loss in EC-GFP. However, the effectiveness of increased AB-GFP biomass in restoring the total GFP level mainly derives from the assumed constantly high GFP synthesis flux, i.e. $v_{gfp,Y}$ in AB-GFP is always higher than

the predicted GFP production rates in EC-GFP (Figure 6-10(b)). The actual GFP synthesis rate in the slow-growing auxiliary species *A. baylyi* might not be able to exceed that in the fast-growing *E. coli*. Changes in $v_{gfp,Y}$ levels can affect the predicted shape of the GFP production curves (see AppxIV-Figures S2-S3).

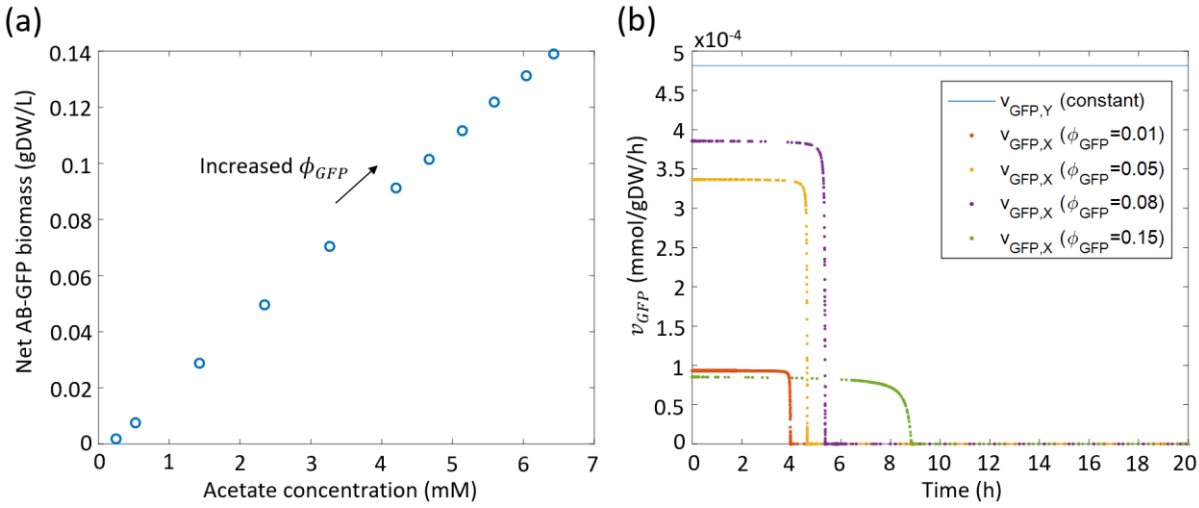


Figure 6-10. **(a)** Model simulation of net accumulation of AB-GFP biomass against the acetate concentration resulting from its net accumulation. **(b)** Comparison between the assumed GFP production rate in AB-GFP ($v_{gfp,Y}$, set to a constant) and the predicted GFP production rate in EC-GFP ($v_{gfp,X}$) at different GFP expression levels.

Finally, it would be desirable to compare the simulation results shown in Figure 6-9 with relevant experimental data. However, the best set of experimental data we found was derived from a rich medium, where the time-dependent overall GFP concentration in a batch coculture of EC-GFP and AB-GFP was available [17]. The growth pattern and metabolic strategy of EC-GFP in rich media are dependent on not only the glucose level but also the amino acid availability. Besides, the RA principle adopted in this chapter (and more broadly throughout the thesis) was developed for *E. coli* grown under minimal glucose conditions. The role of RA for *E. coli* grown on rich media remains unclear. Correspondingly, DFBA simulations of the batch growth of EC-GFP/AB-GFP coculture in this section were based on glucose minimal medium. It is inappropriate to compare the

simulation results shown in Figure 6-9 for a minimal medium with the experimental data reported for a rich medium [17].

6.3.4 Summary of case (ii)

We have established the GFP and proteome allocation-integrated metabolic model for *E. coli* and the GFP-integrated metabolic model for *A. baylyi*, which can be useful computational tools in systems biology and synthetic biology. DFBA simulation of batch growth of EC-GFP and AB-GFP coculture implies the existence of an optimal expression level that confers the highest total GFP productivity. The non-linear correlation between GFP productivity and protein expression level reflects an intriguing trade-off between cost and benefit in recombinant systems. However, the established EC-GFP and AB-GFP models are limited by the values of parameters, e.g. v_{glc}, v_{ac}, $v_{gfp,Y}$, K_g etc. which were chosen without support by experimental data. Further work to calibrate model parameters using specific experimental data is clearly warranted. Besides, the maintenance energy is assumed to change linearly with ϕ_{GFP}, which might be overly simplified. It would be useful to measure the energy demand against growth rate at different expression strengths of recombinant proteins. Besides, the acetate inhibition effect on *E. coli* growth was excluded in the current model. One way to model it is to add an inhibitory term to the kinetic equation of substrate uptake [28]. Furthermore, a fixed initial mass ratio of 1:1 for EC-GFP and AB-GFP was adopted throughout the study. It would be interesting to investigate the impact of different EC-GFP over AB-GFP inoculation ratios on population dynamics.

6.4 Investigating the Warburg effect in human mesenchymal stem cells

6.4.1 Warburg effect – aerobic production of lactate

The Warburg effect, also known as aerobic glycolysis, is the use of glycolysis (resulting in lactate production) instead of the more energy-efficient respiration (through oxidative phosphorylation) despite the presence of oxygen in mammalian cells [29]. Most cells exhibiting this effect are 'fast-proliferating', e.g. stem cells, immune cells and cancer cells [30]. Understanding the potential mechanisms that rewire the cell's metabolism towards aerobic glycolysis can be crucial in resolving tumorigeneses and may help the treatment for cancer [31].

It has been acknowledged that the Warburg effect in mammalian cells and the overflow metabolism in bacteria cells are similar overflow phenomena, in that they both involve (i) the fast-growing nature of the cell and (ii) the use of a seemingly wasteful metabolic strategy [15]. Given this similarity, we are motivated to explore the potential of the proteome allocation theory (which successfully explains the aerobic acetate production in fast-growing *E. coli*) in deciphering the aerobic lactate production in proliferating human mesenchymal stem cells (hMSCs) [32] through analysing the metabolism. More specifically, the questions to be answered in this section are:

i. What are the flux patterns of glycolysis and respiration in hMSCs with increased growth rates?
ii. What are the differences between the flux patterns mentioned in (i) and those reported for overflowed *E. coli*?
iii. Whether the proteome allocation theory can be used to explain the aerobic lactate production in hMSCs?

6.4.2 Genome-scale metabolic model of hMSCs

Genome-scale metabolic reconstruction of mesenchymal stem cells *i*MSC1255 [33] was used in all simulations. Experimentally determined lactate production rate against different growth rates profile of hMSCs grown under glucose-rich media at different oxygen levels (20% and 1%) [32] were used to simulate the metabolic flux distribution in hMSCs via FBA. The growth of hMSCs was simulated by the following FBA model:

$$\text{Minimise } v_{glc} + \sum_i v_{AAi,essential}$$

$$\text{Subject to (i) } S\boldsymbol{v} = 0$$

$$\text{(ii) } v^L \leq \boldsymbol{v} \leq v^U \qquad \text{Equation 6-18}$$

$$\text{(iii) } \lambda = \lambda_{exp} \text{ and } v_{lac} = v_{lac,exp}$$

S is the stoichiometric matrix with rows representing metabolites and columns representing fluxes. $\boldsymbol{v}$ is a column vector comprising reactions in the metabolic model. v^L and v^U define the lower and upper bounds of metabolic fluxes. Growth rate (λ) and lactate production rate (v_{lac}) were fixed to the experimentally measured values, λ_{exp} and $v_{lac,exp}$. Note that $v_{lac,exp}$ were specified based on linear regressions of the measured discrete data points, i.e. $v_{lac,exp} = k\lambda_{exp} + b$, k and b are linear coefficients. The objective function was set to minimizing the uptake rate of growth-controlling nutrients, i.e. glucose (v_{glc}) and essential amino acids ($\sum_i v_{AAi,essential}$). The original *i*MSC1255 was curated by adding an L-lactate exchange flux, inactivating H2CO3D and MCD to eliminate CO_2 redundant loop and closing SBTR, GLCt4 and HEX7 to enable glucose metabolism through hexokinase (HEX1). The upper bound of oxygen uptake was set to 1000 mmol/gDW/h. Ten essential amino acids were identified in *i*MSC1255, and their maximum uptake rates were set proportional to the measured growth rate λ_{exp} with linear

coefficients being the stoichiometry coefficient shown in the biomass flux (Table 6-3). The lower bound of EX_met_L(e) was additionally curated to -0.2λ_{exp} for feasible growth. Model predictions of fluxes of interest (e.g. EX_glc(e), EX_o2(e), CS, ACONT) were checked with Flux Variability Analysis (FVA) [34].

Table 6-3. List of essential amino acids and linear coefficients for determining the upper bound of the corresponding exchange flux for FBA simulation.

Essential amino acid	Linear coefficient	Essential amino acid	Linear coefficient
L-histidine	0.1264	L-phenylalanine,	0.2595
L-isoleucine	0.2861	L-threonine	0.3127
L-leucine	0.5455	L-tryptophan	0.0133
L-lysine	0.5921	L-tyrosine	0.1597
L-methionine	0.1530	L-valine	0.3526

6.4.3 Results and discussion

6.4.3.1 Lactate production could occur at increased respiration activity in hMSCs

One major trait of proteome allocation-mediated metabolic shift towards overflow metabolism is the decrease in respiration at increased growth rate [15]. To test whether lactate-producing hMSCs possess a similar feature, we used FBA (growth rate and lactate production rate were fixed at measured values) to estimate intracellular metabolic fluxes for hMSCs grown at 20% and 1% oxygen. At 20% oxygen, the measured growth rate spans across 0.02 to 0.1 h^{-1}. Measured lactate production first increases with the growth rate and peaks at ~ 0.06 h^{-1}, then goes downward (Figure 6-11(a)). The estimated respiratory fluxes (citrate synthase, CS and malate dehydrogenase, MDH) and oxygen uptake rate (OUR) show monotonic increases with the growth rate (Figure 6-11(b)). At 1% oxygen level, hMSCs grow slightly faster (λ = 0.05 to 0.15 h^{-1}). This time again no reductions in TCA fluxes and OUR were predicted (Figure 6-11(c-d)).

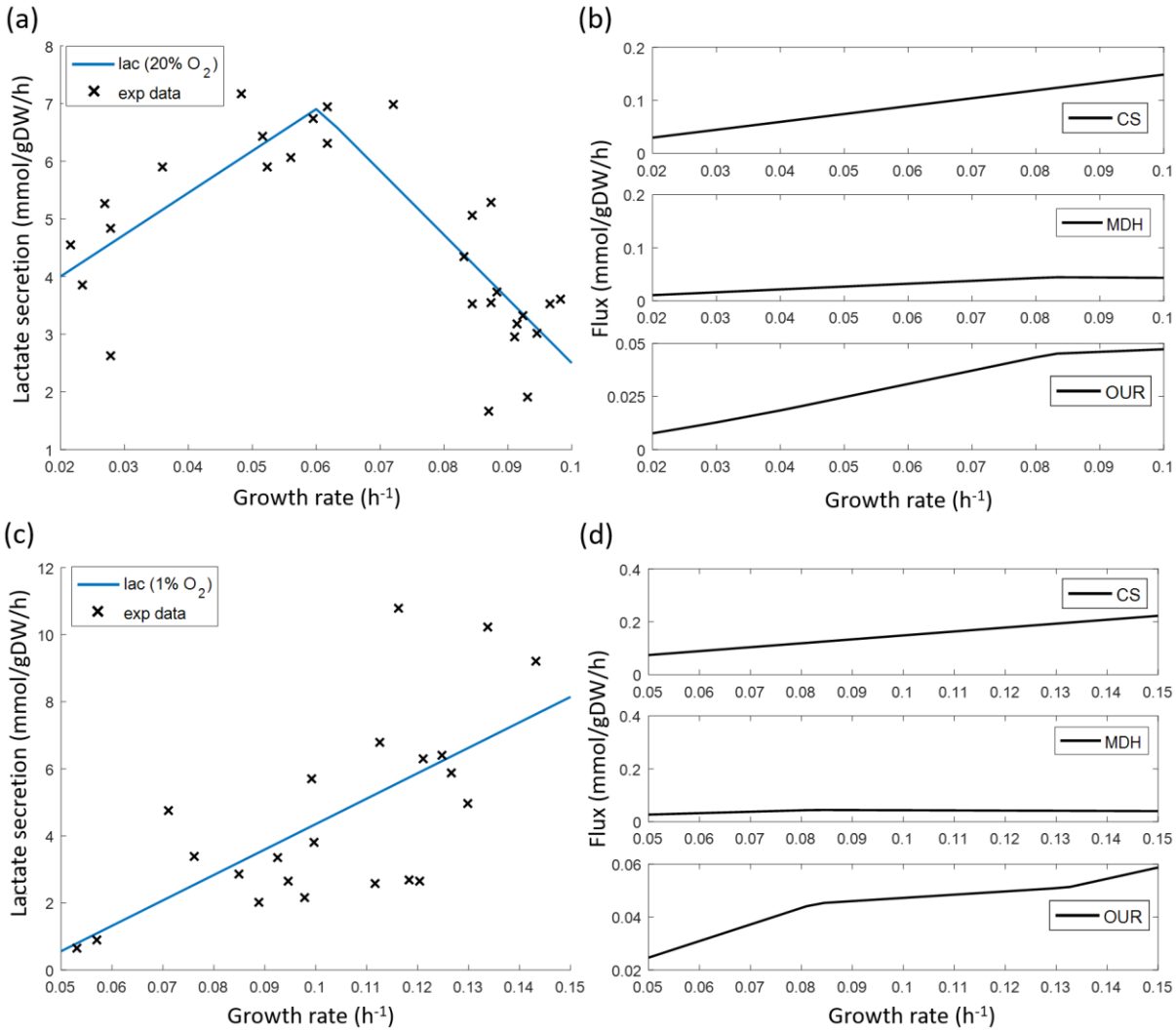


Figure 6-11. **(a)** Regressed (blue line) and experimentally measured (black cross) lactate production rate at different growth rates for hMSCs at 20% oxygen. The linear regression lines are $y = 75.5x + 2.6$ (R^2=0.61) for growth rate $\leq$ 0.06 h^{-1} and $y = -111.2x + 13.6$ (R^2=0.64) for growth arte > 0.06 h^{-1}. **(b)** Simulated respiratory fluxes (CS and MDH) and oxygen uptake rate (OUR) for hMSCs at 20% oxygen. **(c)** Regressed (blue line) and experimentally measured (black cross) lactate production rate at different growth rates for hMSCs at 1% oxygen. The linear regression line is $y = 75.9x - 3.2$ (R^2=0.5). **(d)** Simulated respiratory fluxes (CS and MDH) and OUR for hMSCs at 1% oxygen. Experimental data were obtained from Chapter 4 of ref. [32]. CS-citrate synthase; MDH-malate dehydrogenase.

Although different oxygen availability induces distinct lactate production-growth rate patterns, in both cases TCA cycle and oxygen intake show active response with increased growth. This implies that aerobic lactate production is not necessarily coupled with reduced respiration activity.

Now, we would like to interpret our simulation results by comparison with related work. Firstly, a straightforward difference can be found for the growth rate dependency of the respiration flux between our FBA results for lactate-producing hMSCs and reported results for overflow *E. coli*. For *E. coli*, it has been shown both experimentally [15] and mathematically [20] that the respiration flux decreases with increased growth rates and increased acetate production rates. For hMSCs, our model suggests a distinct trend, i.e. when the growth rate increases, the TCA fluxes and OUR are predicted to (i) increase with an increased lactate production rate (Figure 6-11(a)) or (ii) even increase with reduced lactate production rates (Figure 6-11(c)).

Secondly, we would like to compare our FBA results and flux measurements obtained from Warburg mammalian cells. It has been shown that OUR decreases with increased Warburg number (defined by the ratio of the flux of pyruvate to lactate to its flux into the mitochondrion) at identical growth rates in Warburg tumour cells (Table 1 in [35]). In another work, the experimentally determined oxygen uptake rate was shown to first increase and then decline at increased growth rates, with increased lactate production rates during tumour development of fibroblasts [36].

Several important conclusions can be drawn from the above discussion. Firstly, although it has been proposed that one potential physiological function of the Warburg effect is to promote cell growth and proliferation [37], the strong deviation in measured lactate production rate against growth rate profiles between different studies (e.g. between Figure 6-11(a) and Figure 6-11(c) and between Figure 6-11(a) and Figure 1 in ref. [36]) suggests that the dependency of the Warburg effect (reflected by the lactate production rate or the Warburg number) on the growth rate remains obscure.

Secondly and possibly related to the above-mentioned blurred growth rate-dependency of the Warburg effect, the correlation between the respiration flux (reflected by the oxygen consumption rate) and the growth rate in Warburg mammalian cells shows poor consistency among different experimental [35,36] and modelling work (i.e. between this section and ref. [36]). This leads to our proposition that although the overflow metabolism in *E. coli* and the Warburg effect in mammalian cells have been considered 'closely related' [15], they may not share the same underlying mechanism. For the overflow metabolism in *E. coli*, proteome allocation has proven to be the major driver and is coupled with reduced OUR and respiration fluxes [15]. This typical trait however is not ubiquitously valid in Warburg mammalian cells. Therefore, we posit that perhaps proteome allocation alone cannot explain the Warburg effect, potentially due to the extra complexity involved in the metabolic or regulatory networks in mammalian cells.

Regarding the potential complexity involved in mammalian cells, we can see that in hMSCs GEM [33] there is a significant increase in the number of alternative reactions or pathways (enzymes with identical metabolic function), which makes it critical to specify condition-specific regulatory rules to reduce irrelevant fluxes and redundant loops (which can be numerous if no regulation is imposed) in the solution space. Besides, unlike bacteria, essential amino acids (i.e. those which cannot be synthesised *de novo* by the human metabolic network) play an important role in dictating growth rate of mammalian cells. Essential amino acids not only function as biomass precursors but also can support the synthesis of other amino acids (via transaminase) or even become carbon sources by entering the central carbon metabolism. High amino acid availability can make other carbon sources, e.g. glucose dispensable. When setting the upper bound of (essential) amino acid uptake fluxes, it is important to maintain a delicate balance with other growth-controlling carbon sources. Furthermore, having merely lactate production and growth rate

profile might not be sufficient to recapitulate the physiological state of hMSCs. It would be valuable to obtain genome-wide fluxomic and proteomic datasets for slow-growing cells without lactate production and fast-growing cells with significant lactate secretion, and use them to further calibrate the metabolic flux estimations. In addition, plotting the distribution of proteomic sectors (e.g. carbon catabolism, nitrogen anabolism and ribosomal sectors) against growth rates [38] or lactate production rates could efficiently help uncover the potential role of allocation of proteomic resources in inducing the Warburg effect.

It is also interesting to interpret the limitation of proteome allocation from the energetic aspect. The growth-associated energy requirement (GAM) is determined to be 20.7 mmol ATP/gDW/h for hMSCs [33] and 24 mmol ATP/gDW/h for general human cells [36], whereas GAM in the latest *E. coli* GEM *i*ML1515 is 75.37 mmol ATP/gDW/h [39]. For *Saccharomyces cerevisiae*, GAM is estimated as 59.27 mmol ATP/gDW/h [40]. It has been suggested that bacteria and moulds require about four times as much energy as mammalian cells for biosynthesis [41]. This implies that animals have evolved to be much more energy-efficient than microorganisms. One of the key physiology functions of the proteome allocation-driven overflow metabolism (in *E. coli*) is to produce ATP with a lower amount of enzymes to meet the energy and anabolic protein requirements for rapid growth. The tug of war between energy biogenesis and biomass synthesis, however, may not be the growth-limiting factor in mammalian cells due to its modest growth-associated energy demand. If rapid ATP synthesis is not the key driver of the Warburg effect compared with other proposed functions such as creating tumour microenvironments or promoting signal transduction [37], energy-growth rate-driven proteome allocation would not be the most suitable tool to resolve this phenomenon.

Nevertheless, a strong correlation between the lactate production rate and the glucose uptake rate has been repeatedly confirmed in Warburg cells [42,43], which could provide us another angle for future work.

6.4.3.2 Glutamine preference can lead to decreased TCA cycle activity

In addition to the Warburg effect, preference for glutamine was known as another hallmark of cancer cells [36]. To test the impact of glutamine uptake, we took 1% oxygen group as the control (zero glutamine) and compared it with the case with 0.15 mmol/gDW/h glutamine influx. Several interesting results were found for glutamine metabolism. First, a significant reduction was predicted for TCA fluxes between the intermediate metabolites oxaloacetate and α-ketoglutarate, including primarily CS, aconitate hydratase (ACONT) and isocitrate dehydrogenase (ICDH) (Figure 6-12(a-c)).

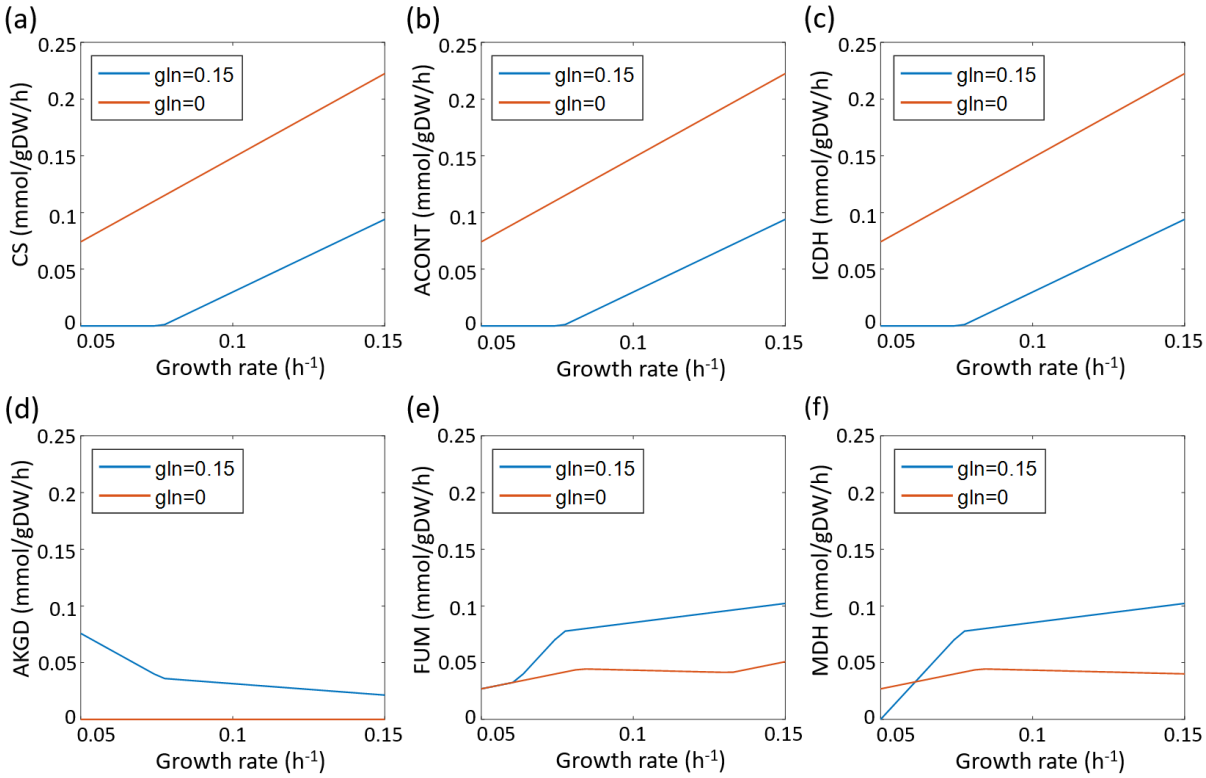


Figure 6-12. **(a-f)** Model prediction of TCA cycle fluxes at different growth rates with and without glutamine addition at 1% oxygen. gln-glutamine uptake rate.

Although other TCA cycle fluxes such as α-ketoglutarate dehydrogenase (AKGD), MDH and fumarase (FUM) increase slightly, the magnitudes of these fluxes are only ~ 1/4 of that flowing through CS, ACONT and ICDH (Figure 6-12(d-f)). One possible explanation is that taking in external glutamine reduces the *de novo* glutamine synthesis requirement through glutamine synthetase (GLNS). Producing one mole of glutamine requires one mole of glutamate and one mole of ammonium while glutamate production via reversible glutamate dehydrogenase (GLUD) requires α-ketoglutarate, a metabolite from the TCA cycle. The presence of external glutamine alleviates the stress on the intracellular glutamate pool and therefore reduces the TCA fluxes towards α-ketoglutarate.

Furthermore, glutamine uptake reduces the glucose uptake rate (GUR) and the extent of the reduction increases with increased growth rates (Figure 6-13(a)). Glutamine uptake also slightly affects the oxygen uptake rate (OUR). At low growth rates, glutamine consumption reduces the oxygen demand through reducing GUR (Figure 6-13(b)). At higher growth rates, glutamine consumption leads to an enhanced OUR, which corresponds to the increased NADH-producing MDH flux (Figure 6-12(f)) as more oxygen is required to cycle NADH back to NAD^+. Besides, a considerable increase in the ratio of carbon flow into the pentose phosphate pathway (PPP) is predicted upon glutamine addition (Figure 6-13(c)). PPP is the major source of NADPH and plays an important role in maintaining redox balance in rapidly proliferating cells. Our prediction is consistent with the previous notion that PPP activity is crucial for tumorigenesis [44] as increased PPP activity could mediate the oxidative stress in proliferating cancer cells through limiting the accumulation of reactive oxygen species (ROS) [45,46]. In addition to this consistency, our results indicate that increased PPP activity observed in cancer cells might be a local physiological response to cope with the glutamine preference.

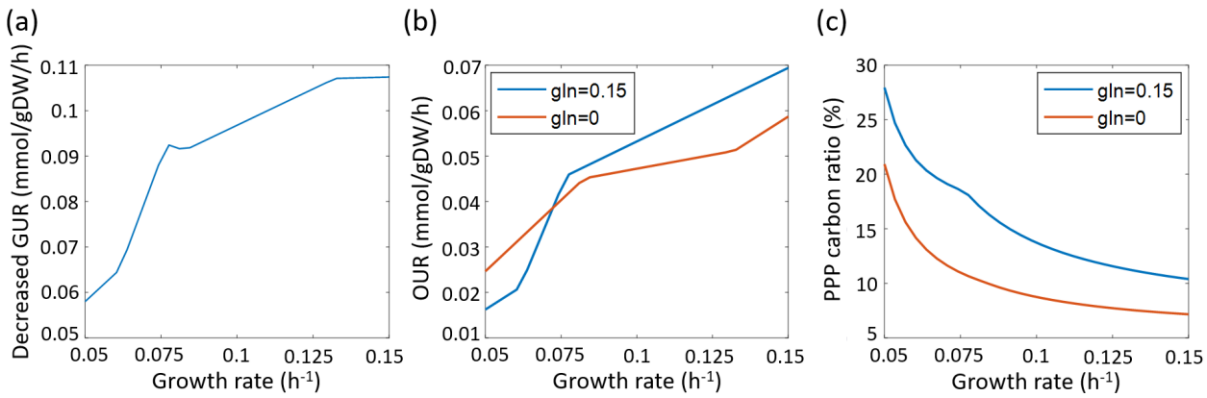


Figure 6-13. Model prediction of **(a)** decrease in glucose uptake rate (GUR), **(b)** oxygen uptake rate (OUR) and **(c)** ratio of carbon flow to PP pathway (PPP) with and without glutamine addition. Decreased GUR = GUR (without glutamine) – GUR (with glutamine); PPP carbon ratio = 6-phosphogluconolactonase (PGL)/GUR×100%.

6.4.4 Summary of case (iii)

In this section, we use experimental data and nutrient constrained FBA to investigate metabolic flux distributions for hMSCs grown under different oxygen levels. Model predictions show that aerobic lactate production is not necessarily coupled with reduced respiration activity. By comparing our simulation results with several related works, we find that the growth rate dependency of the Warburg effect (reflected by the lactate production rate), the oxygen uptake rate and the respiration flux deviates strongly between different studies. Correspondingly, we posit that the proteome allocation alone (which governs the overflow metabolism in *E. coli*) may not be able to explain the Warburg effect. This proposition is supported by (i) the weak correlation between the respiration flux and the growth rate (which is the hallmark of the proteome allocation-driven overflow metabolism) observed in proliferating mammalian cells and (ii) the remarkably low growth-associated energy demand of general human cells, which weakens the importance of proteomic efficiency for energy production. Furthermore, FBA simulations of hMSCs grown with and without glutamine supplement suggest that the glutamine preference in cancer cells could contribute to low TCA fluxes and high PPP activity.

6.5 Conclusions

Chapter 6 presents three parallel case studies covering (i) nad-Fdh recombinant *E. coli,* (ii) an engineered coculture of GFP-recombinant *E. coli* and *A. baylyi* and (iii) lactate-producing hMSCs to explore the potential application and limitation of proteome allocation principles and metabolic models. Firstly, an FDH-integrated proteome allocation model shows that overflow metabolism in *E. coli* could be reduced via introducing new energy pathways that can lessen the catabolic requirement of the main carbon source and encompass low protein costs. Secondly, DFBA simulation of *E. coli* and *A. baylyi* coculture suggests that the overall recombinant protein productivity is related to not only the expression level of the target product but also strain-to-strain interactions. Model simulations predict that maximum GFP accumulation arises from a moderate expression level that balances the target flux with the growth of the host cell. Finally, the FBA study of lactate-producing hMSCs implies that the Warburg effect (aerobic glycolysis) in proliferating mammalian cells may not be driven by efficient proteome allocation, which has proven to be the key driver of a similar overflow phenomenon – the aerobic production of acetate in *E. coli*. The great complexity in the human metabolic network calls for genome-wide proteomics and fluxomics to validate the role of bacterial-derived macromolecular resource allocation in dictating metabolic configurations in cancer cells. In addition, FBA simulations of glutamine supplement in hMSCs suggest that the glutamine preference previously reported for cancer cells may relate to low TCA cycle fluxes and high PP pathway activity.

References

1. Babel W, Brinkmann U, Müller RH: **The auxiliary substrate concept — an approach for overcoming limits of microbial performances**. *Acta Biotechnol* 1993, **13**:211–242.
2. Babel W, Müller RH, Markuske KD: **Improvement of growth yield of yeast on glucose to the maximum by using an additional energy source**. *Arch Microbiol* 1983, **136**:203–208.
3. Bruinenberg PM, Jonker R, van Dijken JP, Scheffers WA: **Utilization of formate as an additional energy source by glucose-limited chemostat cultures of Candida utilis CBS**

621 and Saccharomyces cerevisiae CBS 8066. *Arch Microbiol* 1985, **142**:302–306.
4. Bruinenberg PM, Van Dijken JP, Scheffers WA: **An enzymic analysis of NADPH production and consumption in Candida utilis**. *Microbiology* 1983, **129**:965–971.
5. Hazeu W, Donker RA: **A continuous culture study of methanol and formate utilization by the yeast Pichia pastoris**. *Biotechnol Lett* 1983, **5**:399–404.
6. Harris DM, Van Der Krogt ZA, Van Gulik WM, Van Dijken JP, Pronk JT: **Formate as an auxiliary substrate for glucose-limited cultivation of Penicillium chrysogenum: impact on penicillin G production and biomass yield**. *Appl Environ Microbiol* 2007, **73**:5020–5025.
7. Geertman J-MA, Van Dijken JP, Pronk JT: **Engineering NADH metabolism in Saccharomyces cerevisiae: formate as an electron donor for glycerol production by anaerobic, glucose-limited chemostat cultures**. *FEMS Yeast Res* 2006, **6**:1193–1203.
8. Lian J, Garcia-Perez M, Coates R, Wu H, Chen S: **Yeast fermentation of carboxylic acids obtained from pyrolytic aqueous phases for lipid production**. *Bioresour Technol* 2012, **118**:177–186.
9. Litsanov B, Brocker M, Bott M: **Towards homosuccinate fermentation: metabolic engineering of Corynebacterium glutamicum for anaerobic succinate production from glucose and formate**. *Appl Environ Microbiol* 2012,
10. Balzer GJ, Thakker C, Bennett GN, San K-Y: **Metabolic engineering of Escherichia coli to minimize byproduct formate and improving succinate productivity through increasing NADH availability by heterologous expression of NAD+-dependent formate dehydrogenase**. *Metab Eng* 2013, **20**:1–8.
11. Liu Z, Oyetunde T, Hollinshead WD, Hermanns A, Tang YJ, Liao W, Liu Y: **Exploring eukaryotic formate metabolisms to enhance microbial growth and lipid accumulation**. *Biotechnol Biofuels* 2017, **10**:22.
12. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
13. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245–248.
14. Zeng H, Yang A: **Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways**. *BMC Syst Biol* 2019, **13**:3.
15. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
16. Unden G Steinmetz P D-DP: **The aerobic and anaerobic respiratory chain of Escherichia coli and Salmonella enterica: enzymes and energetics**. *EcoSal Plus* 2014,
17. Santala S, Karp M, Santala V: **Rationally engineered synthetic coculture for improved biomass and product formation**. *PLoS One* 2014, **9**.
18. Kannisto MS, Mangayil RK, Shrivastava-Bhattacharya A, Pletschke BI, Karp MT, Santala VP: **Metabolic engineering of Acinetobacter baylyi ADP1 for removal of Clostridium butyricum growth inhibitors produced from lignocellulosic hydrolysates**. *Biotechnol Biofuels* 2015, **8**:198.
19. Oddone GM, Mills DA, Block DE: **A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression**. *Metab Eng* 2009, **11**:367–381.
20. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
21. Pramanik J, Keasling JD: **Effect of Escherichia coli biomass composition on central metabolic fluxes predicted by a stoichiometric model**. *Biotechnol Bioeng* 1998, **60**:230–238.
22. Taymaz-Nikerel H, Borujeni AE, Verheijen PJT, Heijnen JJ, van Gulik WM: **Genome-derived minimal metabolic models for Escherichia coli MG1655 with estimated in vivo**

respiratory ATP stoichiometry. *Biotechnol Bioeng* 2010, **107**:369–381.
23. Neidhardt FC, Ingraham JL, Schaechter M: *Physiology of the bacterial cell: a molecular approach*. Sinauer Sunderland; 1990.
24. Zeng H, Yang A: **Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant Escherichia coli**. *Biotechnol Bioeng* 2019,
25. Durot M, Le Fèvre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, Smidtas S, Salanoubat M, Weissenbach J, Schachter V: **Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data**. *BMC Syst Biol* 2008, **2**:85.
26. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**:1331–1340.
27. Lara AR, Vazquez-Limón C, Gosset G, Bolivar F, López-Munguia A, Ramirez OT: **Engineering Escherichia coli to improve culture performance and reduce formation of by-products during recombinant protein production under transient intermittent anaerobic conditions**. *Biotechnol Bioeng* 2006, **94**:1164–1175.
28. Hanly TJ, Henson MA: **Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures**. *Biotechnol Bioeng* 2011, **108**:376–385.
29. Warburg O: **The metabolism of carcinoma cells**. *J Cancer Res* 1925, **9**:148–163.
30. Abdel-Haleem AM, Lewis NE, Jamshidi N, Mineta K, Gao X, Gojobori T: **The emerging facets of non-cancerous Warburg effect**. *Front Endocrinol (Lausanne)* 2017, **8**:279.
31. Vander Heiden MG, Cantley LC, Thompson CB: **Understanding the Warburg effect: the metabolic requirements of cell proliferation**. *Science* 2009, **324**:1029–1033.
32. Zhang B: **Metabolic shift under different growth regimes**. In *D.Phil.Thesis: Investigation of Engineering Conditions to Enhance the Culture Efficiency of Human Mesenchymal Stromal Cells*. Edited by Zhang B, Yang A. University of Oxford; 2019:70–107.
33. Fouladiha H, Marashi S-A, Shokrgozar MA: **Reconstruction and validation of a constraint-based metabolic network model for bone marrow-derived mesenchymal stem cells**. *Cell Prolif* 2015, **48**:475–485.
34. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**:264–276.
35. Shan M, Dai D, Vudem A, Varner JD, Stroock AD: **Multi-scale computational study of the Warburg effect, reverse Warburg effect and glutamine addiction in solid tumors**. *PLOS Comput Biol* 2018, **14**:e1006584.
36. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E: **Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect**. *PLoS Comput Biol* 2011, **7**:e1002018.
37. Liberti M V, Locasale JW: **The Warburg effect: how does it benefit cancer cells?** *Trends Biochem Sci* 2016, **41**:211–218.
38. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR: **Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria**. *Mol Syst Biol* 2015, **11**.
39. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, et al.: **iML1515, a knowledgebase that computes Escherichia coli traits**. *Nat Biotechnol* 2017, **35**:904–908.
40. Mo ML, Palsson BØ, Herrgård MJ: **Connecting extracellular metabolomic measurements to intracellular flux states in yeast**. *BMC Syst Biol* 2009, **3**:37.
41. Kilburn DG, Lilly MD, Webb FC: **The energetics of mammalian cell growth**. *J Cell Sci* 1969, **4**:645–654.
42. Vazquez A, Liu J, Zhou Y, Oltvai ZN: **Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited**. *BMC Syst Biol* 2010, **4**.
43. Zielinski DC, Jamshidi N, Corbett AJ, Bordbar A, Thomas A, Palsson BO: **Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism**. *Sci Rep* 2017, **7**:41241.

44. Stincone A, Prigione A, Cramer T, Wamelink MMC, Campbell K, Cheung E, Olin-Sandoval V, Grüning N-M, Krüger A, Tauqeer Alam M, et al.: **The return of metabolism: biochemistry and physiology of the pentose phosphate pathway**. *Biol Rev* 2015, **90**:927–963.
45. Anastasiou D, Poulogiannis G, Asara JM, Boxer MB, Jiang J, Shen M, Bellinger G, Sasaki AT, Locasale JW, Auld DS, et al.: **Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses**. *Science* 2011, **334**:1278–1283.
46. Grüning N-M, Ralser M: **Sacrifice for survival**. *Nature* 2011, **480**:190–191.

Chapter 7 Discussion and Conclusions

7.1 Summary of the PhD research and key findings

7.1.1 Summary of main work accomplished

The main work accomplished in this PhD thesis comprises a comprehensive literature review of the understanding and mathematical modelling of resource allocation (Chapter 2) and several resource allocation-integrated metabolic models (Chapters 3-6). The project started with investigating the overflow metabolism in *E. coli*. Through introducing a concise pathway-level proteome constraint that describes the allocation of limited proteomic resources between energy biogenesis and biomass synthesis, the initial modelling framework was established and enabled the prediction of aerobic acetate production at high growth rates in a core metabolic model for wild type *E. coli* (Chapter 3). Next, it proceeded to studying growth physiologies of recombinant strains. To model the impact of the production of heterologous proteins on the host cell, two distinct types of burdens (metabolic and proteomic) were modelled as varied maintenance energy and the reduced fraction of the proteome allocable to growth-associated processes, respectively. The revised model was able to predict the dynamic configurations of metabolism, cell growth and recombinant protein production for two different recombinant *E. coli* (Chapter 4).

Furthermore, the energy metabolism-centred modelling framework was extended to include the carbon transport sector. This updated model linked the substrate uptake kinetics with the proteome allocation rules and metabolic configurations, which provided insights into the mechanistic basis of the empirical Monod growth law (Chapter 5). Corresponding to the progressive improvement made in the modelling framework, a comprehensive literature review was produced to summarise the latest understanding of

the resource allocation theory and associated mathematical models (Chapter 2). The mathematical mechanisms and potential connections and limitations between different models were also discussed.

Further efforts were made to discover potential applications and limitations of resource allocation principles and models by introducing three parallel case studies (Chapter 6). The first system is the NAD^+-dependent formate dehydrogenase (nad-Fdh) recombinant *E. coli* grown aerobically on a glucose/formate mixed minimal medium. Formate was introduced as the auxiliary energy substrate to reduce the overflow metabolism. Following the modelling framework established for recombinant *E. coli* (Chapter 4), a genome-scale metabolic model incorporating the resource allocation constraint was used to simulate cell growth and the production of acetate at varying proteomic costs between energy pathways for nad-Fdh carrying *E. coli*. The second system is a coculture of GFP-producing *E. coli* and *A. baylyi*. A community-level resource allocation model was established to simulate the batch growth of the coculture. This community-level resource allocation model predicted the overall GFP productivity at the end of the batch at different GFP expression levels. The last case study focused on the overflow-related Warburg effect. The lactate-producing hMSCs were taken as the analogy to cancer cells. Flux balance analysis was used to estimate the metabolic fluxes of hMSCs by fixing the lactate production rate and the growth rate to the measure values in a genome-scale model. The estimated metabolic fluxes were used to investigate the potential of using the proteome allocation principles to decipher the lactate production in hMSCs. The three cases together constitute the final research chapter of this PhD thesis.

7.1.2 Key learnings and outcomes

One major learning of this PhD project is that a concise pathway-level proteome allocation constraint (Equation 3-7, Chapter 3) is capable of predicting with good quantitative

accuracy the metabolic switch and the extent of the overflow metabolism with increased growth rates in *E. coli*. Compared to the individual reaction-level proteome allocation constraint proposed in CAFBA [1], the simplified proteome constraint significantly reduces the number of model parameters and allows the estimation of proteomic cost parameters from bioreactor-level cell growth and by-product excretion profiles.

Besides introducing a 'leaner' version of proteome allocation constraint, Chapter 3 shows that the control of the carbon flow into the pentose phosphate (PP) pathway played an important role for the activation of the acetate pathway. In *E. coli*, the PP pathway is a major source of NADPH (in addition to the TCA cycle). When the respiration activity (represented by TCA cycle fluxes) is limited by the proteome constraint, more flux can be directed to the PP pathway to compensate for the loss of NADPH in the respiration chain. In addition, NADPH produced in the PP pathway can be converted to NADH via $NAD(P)^+$ transhydrogenase and to ATP via the electron transport chain without wasting carbon. Therefore, limiting the activity of the PP pathway is essential for preventing the optimal solution to bypass the carbon-spilling acetate pathway. The first model also showed that setting correct energy parameters (GAM and ATPM) help improve the prediction of the biomass yield.

In the second model of recombinant *E. coli*, the relative importance of the energetic and proteomic burdens (due to the production of recombinant proteins) were shown to depend on the specific protein expressed and/or the specific host strain (Chapter 4). The model also predicted a non-monotonic correlation between the overall GFP productivity at the end of the batch (in mM) and the specific expression level of the target product (e.g. represented by the mass fraction of GFP to the total cellular proteins). Assuming higher expression levels impose stronger burdens to the host cell, maximal GFP production was predicted to emerge at a moderate expression strength. Similar results were obtained in

dynamic simulations of a coculture of GFP-producing *E. coli* and *A. baylyi* (Chapter 6). Another recombinant protein-related work is the resource allocation model for nad-Fdh recombinant *E. coli*. FBA simulations showed that the acetate production could be reduced upon formate addition if the protein cost of the formate pathway (for converting NAD^+ to ATP) is significantly lower than that of the fermentation (for converting glucose to ATP and acetate).

The presented work shows that the key to enhancing the predictive power of constraint-based metabolic models lies in (i) the veracity of the metabolic model; (ii) the biological nature of the constraints adopted and (iii) the connection between local and global traits. The first aspect relates to the reconstruction of the bacterial metabolic network, which is not in the scope of this PhD project. The rest two points are discussed below. In general, the more mechanistic the constraints are, the more predictive the model will be. To predict the acetate production at high growth rates, mechanistically-based proteome allocation constraints eliminate the need for setting arbitrary bounds on the oxygen uptake flux. It also enables the prediction of the substrate uptake rate (which has to be specified in traditional FBA models) through relating the proteomic cost of the carbon transport to the proteome allocation equation (Chapter 5). Furthermore, by linking substrate uptake kinetics (a local feature of carbon transporters) with the proteome allocation constraint (a global regulatory rule), the huge variation of the supposed constant Monod kinetic parameter (K_s) could be readily explained by the varying metabolic strategies.

In addition, it would be useful to compare the results obtained by the modelling framework proposed in this thesis with the predictions made by other network-based resource allocation (RA) models. From a taxonomic perspective, the model presented in this thesis belongs to coarse-grained RA models in addition to fine-grained RA models and elementary flux mode (EFM)-based models (summarised in Chapter 2, Section 2.4). As

EFM-based models are normally not designed for predicting growth phenotypes, below we focus on comparing our work with representatives of coarse-grained and fine-grained RA models.

Within the scope of coarse-grained RA models, regarding the prediction of the occurrence and extent of the overflow metabolism (in a form of acetate production rate at different growth rates, namely the acetate line), the pathway-level proteome allocation model proposed in this thesis is same as accurate as reaction-level proteome allocation model, CAFBA [1]. On the other hand, although the concise pathway-level proteome allocation constraint reduces the workload of parameterisation due to reduced number of parameters, it is not able to predict the proteomic costs of other pathways, e.g. the costs of glutamine synthesis, as the proteomic cost parameters are valid only for fermentation and respiration pathways. CAFBA is able to quantify such costs, or in general, the proteomic costs of any segments of the genome-scale metabolic network using reaction-level cost parameters.

Next, we would like to compare the prediction of the overflow metabolism between coarse-grained (e.g. the model presented in this thesis and CAFBA) and fine-grained RA models (e.g. ME model [2]). We did not find significant differences between the predicted threshold growth rate above which the cell starts to produce acetate between our model, CAFBA and ME. However, the predicted acetate line by our model and CAFBA is more accurate than ME, possibly due to ME did not make much effort in training the model to fit the experimental data. Nevertheless, the major advance of ME is to predict the abundances of proteins, RNA and ribosomes along with maximum growth rate and metabolic fluxes, which currently cannot be achieved by coarse-grained models due to the lack of transcriptional and translational information. One possible way to narrow this gap is to introduce ribosome-associated constraints in coarse-grained models, e.g. using the measured ribosome-to-protein (R/P) ratio against growth rate profile to estimate the

abundance of ribosomes as a function of the maximum growth rate while constraining the summation of metabolic fluxes by the ribosomal level to estimate the protein levels.

Furthermore, the ability to capture the key growth physiologies of *E. coli* at the metabolic level shows the potential of using the modelling framework proposed in this thesis as a null model to guide the design in synthetic biology. Based on the work presented in Chapter 4, here we summarise three basic steps to curate a wild-type resource allocation model to investigate the impact of synthetic constructs on cell growth and protein production. The first step is to evaluate whether the energy demand of the *E. coli* strain to be studied is consistent with the default energy parameters of the metabolic model. Secondly, the biosynthesis pathway of the target product encoded in the construct needs to be added to the stoichiometric matrix of the metabolic model. Finally, one should account for the potential metabolic and proteomic burdens due to the expression of the synthetic construct. For instance, metabolic burdens can be modelled as an increase in maintenance energy. If the target product is proteins, it will be useful to introduce an additional term to the proteome allocation constraint to depict the more 'crowded' proteome.

Following the above three basic steps combined with relatively simple parameterisation procedures (see Chapters 3 and 4), the curated resource allocation model can predict (at least qualitatively) the growth behaviour of the host cell carrying arbitrary synthetic constructs. For example, the curated model can provide an initial screening of superior synthetic constructs with criteria such as higher production rates of the target product given a range of growth rates (see Figure 6-8) or higher biomass yields with a fixed production rate of the target product in a desirable range of growth rates.

Finally, there are two interesting hypotheses proposed in this thesis remained to be validated experimentally. One is the proposition made in Chapter 5 that the observed huge

variations of the supposed constant K_s can be explained by the changes in the metabolic strategies. To test this hypothesis, one can measure the Monod constant (K_{s1}) of a slow-growing *E. coli* culture without the overflow behaviour (i.e. no acetate production) and the Monod constant (K_{s2}) of a fast-growing *E. coli* culture (of the same strain) with the overflow metabolism. If K_{s1} and K_{s2} indeed deviate several times or several orders of magnitude with each other, the first hypothesis can be considered valid at least qualitatively. The second hypothesis is the predicted increasing trend of the respiration flux and oxygen uptake rate with increased growth rates and lactate production rates in human mesenchymal stem cells (hMSCs). The flux predictions made for hMSCs (as shown in Figure 6-11) can be confirmed by measuring the intracellular fluxes of the central carbon metabolism of hMSCs growing at different growth rates via ^{13}C-metabolic flux analysis (MFA) [3]. By plotting the measured TCA fluxes against the growth rate, the model prediction can be efficiently justified. However, the control of the steady-state growth rate of stem cells may not be as easy as that of *E. coli*. Additional difficulties could arise in obtaining growth rate-dependent flux data for hMSCs.

Looking back to the research objectives set in Chapter 1, the first and second objectives were studied in Chapters 3 and 4, respectively. The third objective was investigated in Chapter 5. The last two objectives were addressed through three parallel case studies in Chapter 6. Overall, all of the research targets were reasonably met in the four research chapters; several interesting opening questions are discussed in the section below.

7.2 Limitations and future work

7.2.1 Limitations

The accuracy of the predictions made by the models presented in this thesis essentially depends on the authenticity of the metabolic reconstructions. The up to date genome-scale

metabolic model (GEM) for *E. coli i*ML1515 [4] covers 34% of the *E. coli* genome based on a total of 4453 open reading frames (ORFs) [5] while the *E. coli* GEM used in this thesis (*i*AF1260 [6]) covers a bit less (28%). Although *i*AF1260 has been extensively used to simulate the growth and metabolism of *E. coli* and the simulation results have shown good consistency with the experimental data, one should be aware that some alternative metabolic routes or functions might be missing from the adopted model due to its incomplete genome coverage.

Another important aspect that affects the accuracy of model predictions lies in the parameterisation. In this thesis, the proteomic cost parameters were determined by the bioreactor-level data, i.e. fermentation and respiration fluxes against the growth rate profiles. While the fermentation flux can be directly linked to the experimental data of the acetate production rate, the respiration flux is less straightforward. In this thesis, the latter was estimated by FBA simulated TCA fluxes by setting the growth rate and the acetate production rate to the measured values and minimising the substrate uptake rate. While the calculated proteomic cost parameters are consistent with the reported comparative proteomic cost of energy pathways, their absolute values require additional fluxomic and proteomic datasets to validate. If both proteomic and fluxomic datasets are available, proteomic cost parameters can be directly determined without the need for parameter estimation. For instance, the proteomic cost per pathway can be determined by partitioning the overall protein abundance into those devoted to fermentation, respiration and biomass synthesis based on network stoichiometry and dividing the partitioned protein abundances by the representative fluxes. If proteomic data is not available, having fluxomic data will help the identification of model parameters in parameter estimation using bioreactor-level growth data. By constraining intracellular fluxes, especially those lie in the central carbon metabolism to measured values, the FBA model will become less ‘under-determined’,

which will reduce the dependency of estimated model parameters on the uncertainties of FBA solutions. Furthermore, fluxomics collected in different growth conditions can be used to identify condition-dependent regulatory rules. By incorporating flux regulations to the metabolic model, even in a simple binary matrix indicating which set of reactions is active or inactive under a certain condition, the FBA model can become much less 'under-determined'. Combining regulations with a properly set objective function and the resource allocation constraint, the model solution is likely to become unique at genome-scale. Luckily, the development of high-resolution ^{13}C-MFA [7] and recent efforts to use genome-scale models for ^{13}C-MFA [8,9] make the precise flux estimation at genome-scale promising.

In the proof of concept study of the optimal expression level of recombinant proteins, the protein production-mediated change in the maintenance energy (ATPM) was assumed to be linearly dependent on the protein expression level. This proportional correlation can be an oversimplification given the complexity involved in the circuit-host interactions [10]. Dynamic models of synthetic gene circuits are often nonlinear and high-dimensional [11]. It was assumed that stronger expression of heterologous genes provoke higher burdens and lead to more maintenance to the host. The proportional change in ATPM adopted in Chapter 4 is considered to reflect the reality only qualitatively. The absolute magnitude of the model predictions of the optimal expression level thus should be interpreted with caution.

Furthermore, the proteome allocation-integrated metabolic models presented in this thesis are subject to the maximal growth objective. As stated in Chapter 2, the primary objective of the cell can change from growth to survival under strong perturbations such as sublethal antibiotic dosage, high temperature, pH and osmotic stresses, where certain stress responses cannot be explained by the growth rate-driven resource allocation.

Correspondingly, models developed in this thesis are limited for describing optimal or near optimal growth regimes and could inevitably ignore metabolic configurations for cells striving for alternative and/or multiple objectives.

Finally, the concise pathway-level proteomic allocation constraint established in this thesis significantly reduces the number of model parameters through using a finite amount of intermediate reactions to reflect the trend of the complete pathway. While for a small-scale metabolic model (e.g. the *E. coli* core model) the representative fluxes for fermentation and respiration are relatively clear, complications can arise for a GEM due to the increased details of the metabolic reactions. For all models presented in this thesis, condition-specific regulations have ensured that one enzymatic reaction is sufficient to capture the pathway behaviours, e.g. the upshift (downshift) of the fermentation (respiration) flux during the overflow metabolism. However, for unregulated or more detailed GEMs the representative feature of this single reaction can be easily spoiled by alternative pathways or parallel reactions, making the current proteome allocation constraint ineffective. In that case, more intermediate fluxes (as representatives of the metabolic pathway) and more proteomic cost parameters may need to be included.

7.2.2. Future work

The unclear correlation between the cellular maintenance and the expression level of recombinant proteins can be determined by additional experiments to measure the overall energy demand (e.g. in mmol ATP per gDW per hour) at different expression strengths (corresponding to different protein production rates, e.g. in mmol recombinant proteins per gDW per hour). Key model parameters (i.e. proteomic cost parameters and proteomic burdens) can be calibrated by growth rate-dependent proteomic datasets. Besides, the intracellular respiratory fluxes can be directly determined by fluxomics.

For investigating the Warburg effect, a molecular crowding-integrated (another major form of the resource allocation additional to the proteome allocation constraint) genome-scale metabolic model of human cells successfully predicted the shift to lactate production and reduced oxygen uptake rate at high growth rates [12]. However, our flux estimation suggests that lactate production is not necessarily coupled with reduced respiratory activity. This inconsistency calls for a rethinking of the fundamental question – what is the biological function of the Warburg effect? Both proteome allocation and molecular crowding constraints optimize the proteome usage (by using more proteomic-efficient pathways) to obtain the fastest growth rate. If the Warburg effect does not serve for the optimal growth, resource allocation-driven prediction of the lactate production will lose its biological basis. Interestingly, a recent study suggested that tumour growth was dependent on both cell proliferation and motility. Increased lactate secretion in cancer cells facilitates tumour spread by promoting cell movement through degrading the extracellular matrix [13]. Therefore, it would be interesting to re-examine the relationship (in term of the cause and the effect) between resource allocation, growth and lactate production by considering other potential objectives and preferably by collecting growth rate- or lactate secretion rate-dependent proteomic datasets from cancer cells or cancer-like mammalian cells.

Another open question remaining to be resolved is the community-level resource allocation model, which differs from monoculture models by including species-to-species interactions. In Chapter 6, a coculture resource allocation model was developed to investigate the optimal production of recombinant GFP. *A. baylyi* was added to the *E. coli* monoculture to increase the overall biomass and product yield by improving the carbon utilisation and alleviating the acetate inhibition on *E. coli* growth. The inhibition effect was omitted in the presented work, which in the future can be modelled by adding in an acetate-dependent inhibitory term to the substrate uptake kinetics. The inhibitory

parameter can be estimated by experimental data of growth rates or substrate uptake rates at different external acetate dosage. Besides, it would be useful to investigate the selection pressure imposed on microbial consortia, which will help identify the objective (e.g. maximising individual growth rate or maximising the overall biomass) for community-level metabolic models.

References

1. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**.
2. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**:693–693.
3. Wiechert W: **13C metabolic flux analysis**. *Metab Eng* 2001, **3**:195–206.
4. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, et al.: **iML1515, a knowledgebase that computes Escherichia coli traits**. *Nat Biotechnol* 2017, **35**:904–908.
5. Riley M, Abe T, Arnaud MB, Berlyn MKB, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, et al.: **Escherichia coli K-12: a cooperatively developed annotation snapshot—2005** . *Nucleic Acids Res* 2006, **34**:1–9.
6. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
7. Long CP, Antoniewicz MR: **High-resolution 13C metabolic flux analysis**. *Nat Protoc* 2019, **14**:2856–2877.
8. García Martín H, Kumar VS, Weaver D, Ghosh A, Chubukov V, Mukhopadhyay A, Arkin A, Keasling JD: **A method to constrain genome-scale models with 13C labeling data**. *PLOS Comput Biol* 2015, **11**:e1004363.
9. Gopalakrishnan S, Maranas CD: **13C metabolic flux analysis at a genome-scale**. *Metab Eng* 2015, **32**:12–22.
10. Boo A, Ellis T, Stan G-B: **Host-aware synthetic biology**. *Curr Opin Syst Biol* 2019, **14**:66–72.
11. Boada Y, Vignoni A, Picó J: **Multiobjective identification of a feedback synthetic gene circuit**. *IEEE Trans Control Syst Technol* 2020, **28**:208–223.
12. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E: **Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect**. *PLoS Comput Biol* 2011, **7**:e1002018.
13. Tzamali E, Tzedakis G, Sakkalis V: **A framework linking glycolytic metabolic capabilities and tumor dynamics**. *IEEE J Biomed Heal Informatics* 2019, **23**:1844–1854.

Appendix I: Supplementary Information of Chapter 3

1. Supplementary Text

1.1 The core *Escherichia coli* metabolic model

We use the previously published core *Escherichia coli* metabolic model [1] to implement the PAT-based FBA model. The core model consists of 95 cellular reactions and 72 unique metabolites. The COBRA Toolbox [2] was used to load metabolic reconstruction with the addition of global regulations. In particular, the lower bound of the oxygen exchange flux (EX_o2(e)) and the glucose exchange flux (EX_glc(e)) were set to -1000 to mimic the aerobic-glucose conditions where the aerobic formation of acetate (overflow metabolism) occurs. Reactions involved in glyoxylate shunt, gluconeogenesis and aerobic specific processes were determined to be off according to previous published regulatory rules for *E. coli* growing under aerobic-glucose conditions [3]. We also closed the exchange and synthesis fluxes of metabolites that were not reported by pertinent experiments [4–6]. The complete list of reactions that are determined to be off is given in AppxI-Table S6. Reactions in glycolysis, TCA cycle, pentose phosphate pathway, acetate synthesis pathway and oxidative phosphorylation system were set to carry flux in the forward direction based on flux measurement [7]. The cellular objective is set to minimize substrate uptake rate (at a fixed growth rate).

1.2 Integrating PAT to Flux Balance Analysis

To fuse the PAT constraint into FBA, we added an additional row to the S matrix of the metabolic model with the proteomic efficiency parameters w_f^*,w_r^* and b^* at the column position corresponding to fermentation flux (ACKr), respiration flux (AKGDH) and the biomass flux (Biomass_Ecoli_core_w/GAM), respectively. Reaction names are consistent

with the notation used in the *E. coli* core model. The right hand side of the constraint is fulfilled by introducing an auxiliary reaction DM_A

$$w_f' v_f + w_r' v_r + b'\lambda = \text{DM_A} \qquad \text{Equation AppxI-S1}$$

The upper bound of DM_A is set to one

$$w_f' v_f + w_r' v_r + b'\lambda = \text{DM_A} \leq 1 \qquad \text{Equation AppxI-S2}$$

1.3 Impact of PP pathway ratio on model parameters and predictions

As mentioned in the Methods, the portion of substrate carbon diverting into PP pathway can vary with different growth rates (achieve a maximum ~20%) and culture conditions [4,8]. Such uncertainty of the PP pathway flux motivates us to investigate the impact of adjusting the ratio of carbon flowing into the PP pathway to the total carbon intake on the model performance. In the main text (determination of Equation 3-12 and Equation 3-13), we confirmed that, when fixing growth rate and acetate excretion flux, with minimizing glucose uptake as the objective, FBA solution was unique, which provides the foundation of using FBA solution of PP pathway fluxes for further analysis. Applying the experimentally determined acetate excretion - growth rate profile to FBA, a maximum level of PP pathway ratio (*PPP%*) was found for all tested species (~60%, as shown in AppxI-Figure S6). The mean value of the maximum *PPP%* of the flux measurement and the *in silico* result is ~40%. Thus we specified different level of upper bound of PP pathway ratio (*uPPP%*) as 25%, 35% and 40%, along with the proteome allocation constraint to explore the impact of PP pathway uncertainty. Results show that, given a certain species, different uPPP% leads to different values of the k_r and $v_{r,0}$ (the slope and intercept of the respiration line) i.e. respiration flux changes with *uPPP%*, thus rendering different Equation 3-12 and Equation 3-13. For each *uPPP%*, we obtain a collection of

mathematically equivalent w_f^*, w_r^* and b^*. Figure 3-2 and Figure 3-3 show that changes in the PP pathway ratio affect both w_r^* and b^* (w_f^* was fixed as the independent variable), with the latter being affected slightly more significantly.

Furthermore, we confirmed that adopting any set of w_f^*, w_r^* and b^* showed in Figure 3-2 and Figure 3-3 as the PAT constraint provides a non-differentiable model prediction on acetate production. Therefore, the variation of PP pathway ratio, as long as is within a reasonable range (e.g. ~ 20% reported in [4,8]), will not violate the model prediction of the overflow metabolism. It is worth noting that, biologically, PP pathway possesses much less proteome fraction compared with the lower part of glycolysis (~ 1/4) [5], making whatever modulations in the upper glycolysis (due to alternative pathway) insignificant to the overall proteome allocation, coinciding our finding of the relation between *uPPP%* and the overflow metabolism.

2. Supplementary Tables

AppxI-Table S1. List of mathematically equivalent sets of proteomic cost parameters w_f^*, w_r^* and b^*.

Nominal energy demand												
	MG1655				NCM3722				ML308			
uPPP%	wf*	wr*	b*	fval	wf*	wr*	b*	fval	wf*	wr*	b*	fval
25%	0.0601	0.197	0.517	0.1698	0.0283	0.1356	0.4131	0.5262	0.0275	0.155	0.2823	0.05443
	0.0484	0.1757	0.6787	0.1698	0.0478	0.171	0.1637	0.5262	0.0245	0.1498	0.319	0.05443
	0.0597	0.1962	0.5231	0.1698	0.0493	0.1738	0.1444	0.5262	0.0386	0.1749	0.1422	0.05443
	0.0958	0.2617	0.0247	0.1698	0.0457	0.1672	0.1905	0.5262	0.0452	0.1864	0.061	0.05443
	0.0841	0.2405	0.1862	0.1698	0.0193	0.1194	0.5277	0.5262	0.0439	0.1843	0.076	0.05443
35%	0.0107	0.1059	1.2569	0.1698	0.0612	0.1923	0.0975	0.5262	0.0326	0.1617	0.3054	0.05443
	0.0452	0.1674	0.8166	0.1698	0.0144	0.1087	0.6504	0.5262	0.0359	0.1674	0.2679	0.05443
	0.0274	0.1356	1.0447	0.1698	0.0353	0.146	0.4038	0.5262	0.0444	0.1823	0.1694	0.05443
	0.0565	0.1875	0.6726	0.1698	0.021	0.1205	0.5725	0.5262	0.0143	0.1296	0.5179	0.05443
	0.0118	0.1079	1.2424	0.1698	0.0569	0.1847	0.1478	0.5262	0.05	0.1921	0.1041	0.05443
40%	0.0815	0.2303	0.4177	0.1698	0.0103	0.1006	0.7258	0.5262	0.0305	0.1568	0.3719	0.05443
	0.0108	0.1052	1.2854	0.1698	0.0127	0.1048	0.6986	0.5262	0.0587	0.2059	0.0577	0.05443
	0.033	0.1445	1.0126	0.1698	0.0155	0.1098	0.6668	0.5262	0.0188	0.1365	0.5023	0.05443
	0.0385	0.1542	0.9452	0.1698	0.0133	0.1058	0.6924	0.5262	0.052	0.1942	0.1326	0.05443
	0.0297	0.1387	1.0529	0.1698	0.0104	0.1008	0.7246	0.5262	0.0198	0.1382	0.4912	0.05443
New energy demand												
	MG1655				NCM3722				ML308			
uPPP%	wf*	wr*	b*	fval	wf*	wr*	b*	fval	wf*	wr*	b*	fval
25%	0.0802	0.218	0.01	0.1841	0.0143	0.0963	0.4648	0.5502	0.0682	0.6472	0.0137	0.05443
	0.0658	0.1938	0.2349	0.1841	0.0146	0.0968	0.4596	0.5502	0.0627	0.6193	0.0726	0.05443
	0.0276	0.1292	0.8288	0.1841	0.0097	0.089	0.5328	0.5502	0.0642	0.6272	0.0561	0.05443
	0.01	0.0995	1.1014	0.1841	0.0127	0.0938	0.4884	0.5502	0.063	0.6208	0.0694	0.05443
	0.0268	0.1278	0.8414	0.1841	0.015	0.0974	0.4542	0.5502	0.0679	0.6454	0.0175	0.05443
35%	0.0165	0.109	1.0657	0.1841	0.0285	0.1175	0.3219	0.5502	0.0991	0.7445	0.0614	0.05443
	0.0102	0.0985	1.1576	0.1841	0.0261	0.1138	0.3552	0.5502	0.057	0.5471	0.4099	0.05443
	0.0859	0.2247	0.0581	0.1841	0.0119	0.0914	0.5547	0.5502	0.1009	0.7529	0.0466	0.05443
	0.0175	0.1106	1.0519	0.1841	0.0089	0.0868	0.5955	0.5502	0.1053	0.7733	0.0105	0.05443
	0.0154	0.1072	1.0819	0.1841	0.0168	0.0992	0.4856	0.5502	0.0856	0.6813	0.173	0.05443
40%	0.0628	0.183	0.4572	0.1841	0.0425	0.1372	0.1735	0.5502	0.119	0.8081	0.0856	0.05443
	0.0096	0.0956	1.2005	0.1841	0.0289	0.1162	0.3564	0.5502	0.0688	0.5812	0.448	0.05443
	0.0214	0.1149	1.0364	0.1841	0.0175	0.0986	0.5102	0.5502	0.1131	0.7815	0.1281	0.05443
	0.0414	0.1478	0.7569	0.1841	0.0156	0.0957	0.5352	0.5502	0.06	0.5411	0.5119	0.05443
	0.0147	0.1041	1.129	0.1841	0.0311	0.1196	0.3268	0.5502	0.0759	0.6133	0.3967	0.05443

AppxI-Table S2. Growth data for *E. coli* ML308*.

Growth rate (h-1)	ATP (mol/kg)	[2H] (mol/kg)	total ATP (mol/kg)**	total ATP (mmol/g-h)
0.94	19.34	40.95	39.815	37.4261
0.88	16.92	36.83	35.335	31.0948
0.82	14.27	34.21	31.375	25.7275
0.81	12.86	32.05	28.885	23.39685
0.72	9	29.38	23.69	17.0568

* data source: Table 7, Holms, H., 1996. Flux analysis and control of the central metabolic pathways in Escherichia coli. FEMS microbiology reviews, 19(2), pp.85-116.

** Total ATP is calculated assuming 4 protons → 1 ATP, as suggested in Feist, Adam M., et al. "A genome‐scale metabolic reconstruction for Escherichia coli K‐12 MG1655 that accounts for 1260 ORFs and thermodynamic information." Molecular systems biology 3.1 (2007): 121.

AppxI-Table S3. Calculated values of the slope and intercept of the $w_r^* - w_f^*$ line for ML, NCM and ML strains at nominal energy demand.

	Slope			Intercept		
	$v_{f,0}/v_{r,0}$			$1/v_{r,0}$		
uPPP%	25%	35%	40%	25%	35%	40%
MG	1.8095	1.7809	1.7682	0.0882	0.0868	0.0862
NCM	1.8168	1.7872	1.7744	0.0843	0.0829	0.0823
ML	1.7812	1.752	1.7403	0.1063	0.1045	0.1038

AppxI-Table S4 Calculated values of the slope and intercept of the $w_r^* - w_f^*$ line for ML strain with new and nominal energy demand.

Strain	PPP%	Slope	Intercept
		$v_{f,0}/v_{r,0}$	$1/v_{r,0}$
ML-new	25%	5.0084	0.2988
	35%	4.6919	0.2799
	40%	4.5267	0.27
ML-nom	25%	1.7812	0.1063
	35%	1.752	0.1045
	40%	1.7403	0.1038

AppxI-Table S5. List of mathematically equivalent sets of proteomic efficiency parameter w_r^* and b^* for the two-parameter case (w_f^* was set to 0 for all three strains; energy demand was set to the nominal value).

	MG	wr*	b*	**fval**	NCM	wr*	b*	**fval**	ML	wr*	b*	**fval**
uPPP%	25%	0.0882	1.3442	**0.1698**	25%	0.0843	0.7752	**0.5262**	25%	0.1062	0.6268	**0.05443**
		0.0882	1.3442	**0.1698**		0.0844	0.7745	**0.5262**		0.1062	0.6268	**0.05443**
		0.0882	1.3442	**0.1698**		0.0843	0.7752	**0.5262**		0.1062	0.6265	**0.05444**
		0.0882	1.3442	**0.1698**		0.0843	0.7752	**0.5262**		0.1062	0.6268	**0.05443**
		0.0882	1.3442	**0.1698**		0.0843	0.7752	**0.5262**		0.1062	0.6268	**0.05443**
	35%	0.0869	1.3934	**0.1698**	35%	0.0829	0.8207	**0.5262**	35%	0.1046	0.6836	**0.05443**
		0.0868	1.3935	**0.1698**		0.0829	0.8207	**0.5262**		0.1046	0.6836	**0.05443**
		0.0868	1.3935	**0.1698**		0.083	0.8206	**0.5262**		0.1046	0.6837	**0.05443**
		0.0869	1.3934	**0.1698**		0.0829	0.8207	**0.5262**		0.1046	0.6836	**0.05443**
		0.0869	1.3935	**0.1698**		0.0829	0.8206	**0.5262**		0.1046	0.6836	**0.05443**
	40%	0.0862	1.4174	**0.1698**	40%	0.0823	0.8428	**0.5262**	40%	0.1038	0.7113	**0.05443**
		0.0862	1.4174	**0.1698**		0.0823	0.8428	**0.5262**		0.1038	0.7112	**0.05443**
		0.0862	1.4175	**0.1698**		0.0823	0.8428	**0.5262**		0.1038	0.7113	**0.05443**
		0.0862	1.4175	**0.1698**		0.0823	0.8428	**0.5262**		0.1038	0.7112	**0.05443**
		0.0862	1.4174	**0.1698**		0.0823	0.8428	**0.5262**		0.1038	0.7113	**0.05443**

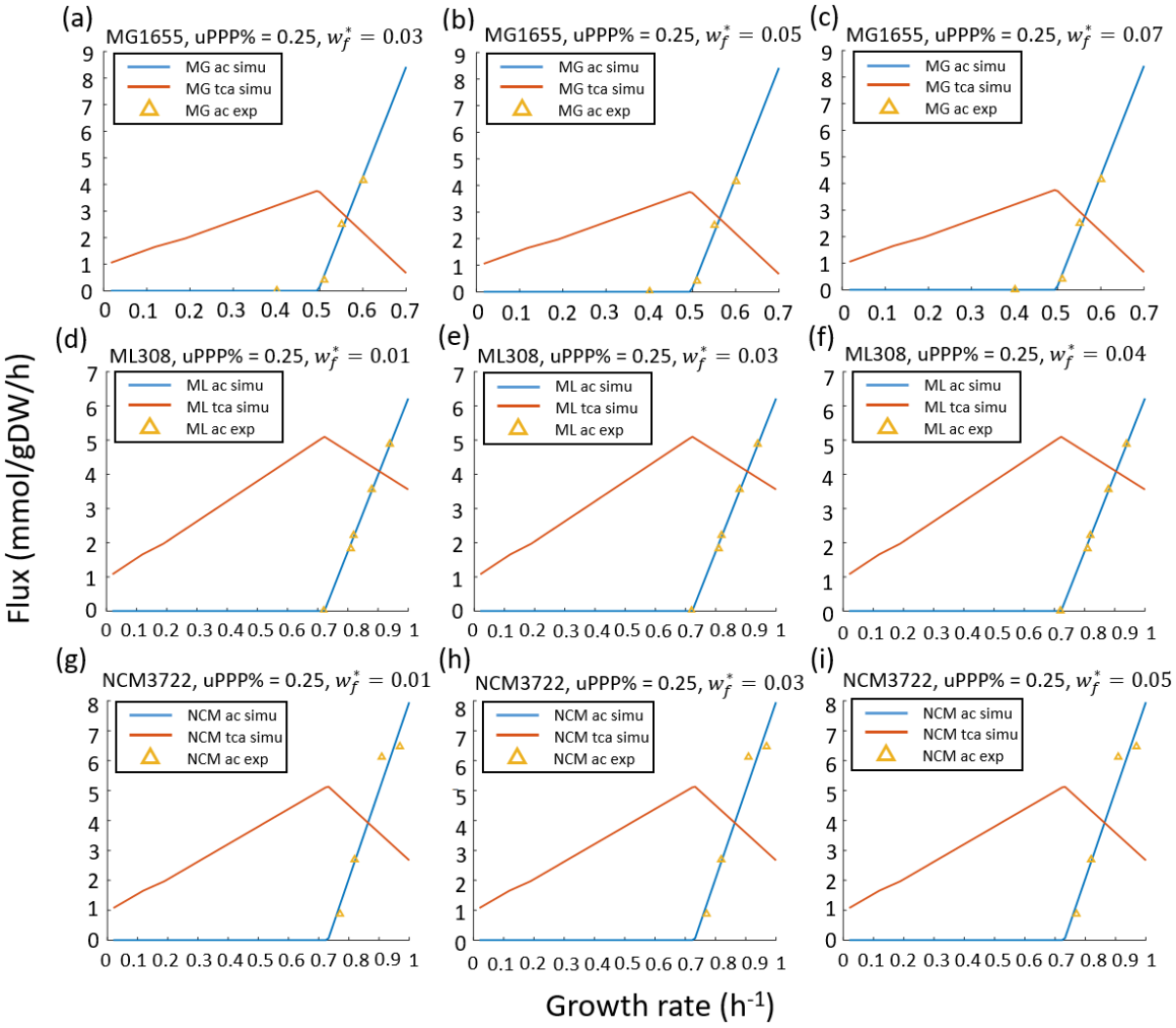
$$fval = \sqrt{\frac{\sum_1^N (simulation\ value - experimental\ value)^2}{N-1}}$$

AppxI-Table S6. List of reactions determined to be unavailable under glucose-aerobic condition.

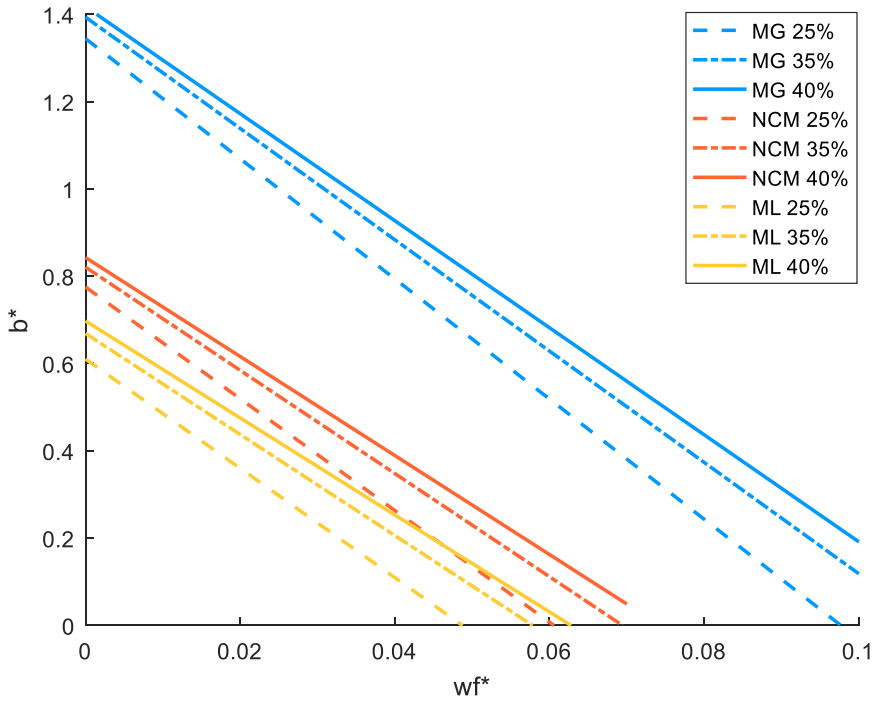
Reactions	Notes	Reactions	Notes
ICL	Glyoxylate shunt	ACALD	Irrelevant by-product
MALS	Glyoxylate shunt	AKGt2r	Irrelevant by-product
PPCK	Glyoxylate shunt	GLUt2r	Irrelevant by-product
PFL	Anaerobic	GLNabc	Irrelevant by-product
FRD7	Anaerobic	EX_succ(e)	Irrelevant exchange
ME1	Gluconeogenesis	EX_lac_D(e)	Irrelevant exchange
ME2	Gluconeogenesis	EX_for(e)	Irrelevant exchange
PPS	Gluconeogenesis	EX_etoh(e)	Irrelevant exchange
FBP	Gluconeogenesis	EX_acald(e)	Irrelevant exchange
PPCK	Gluconeogenesis	EX_fru(e)	Irrelevant exchange
ADK1	Irrelevant by-product	EX_mal(e)	Irrelevant exchange
GLUN	Irrelevant by-product	EX_fum(e)	Irrelevant exchange
FRUpts2	Irrelevant by-product	EX_pyr(e)	Irrelevant exchange

3. Supplementary Figures

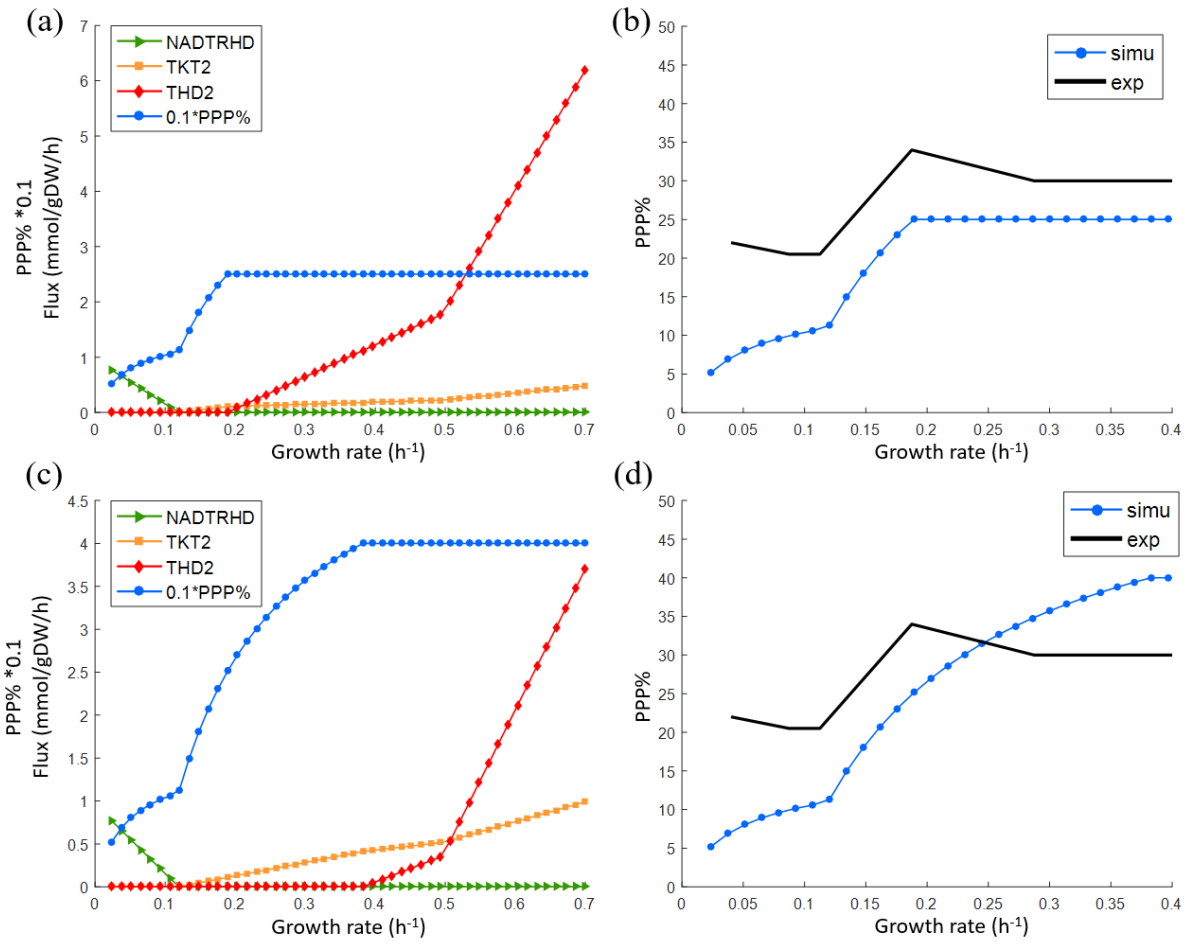
The figure below shows that arbitrarily chosen values of w_f^* yield identical simulation results (rates of acetate production and the TCA flux). The TCA flux is termed as "v_r" in the main text. Both of them are obtained from the simulated flux of the enzymatic reaction AKGDH, a key reaction in the TCA cycle to represent the extent of the reparation.



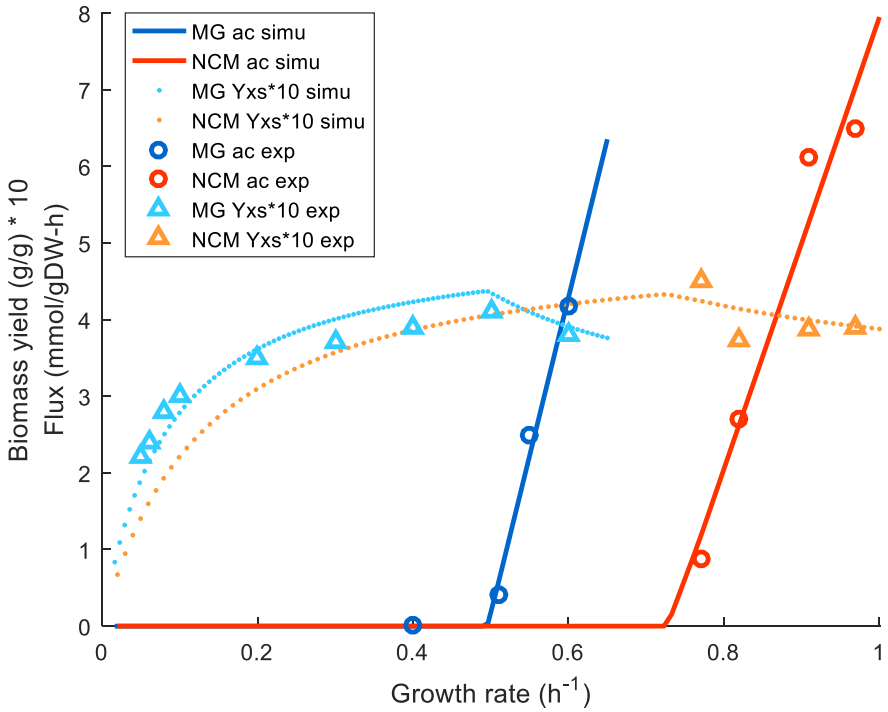
AppxI-Figure S1. Model prediction of the acetate and TCA fluxes at arbitrarily specified w_f^* values for MG, ML and NCM strains.



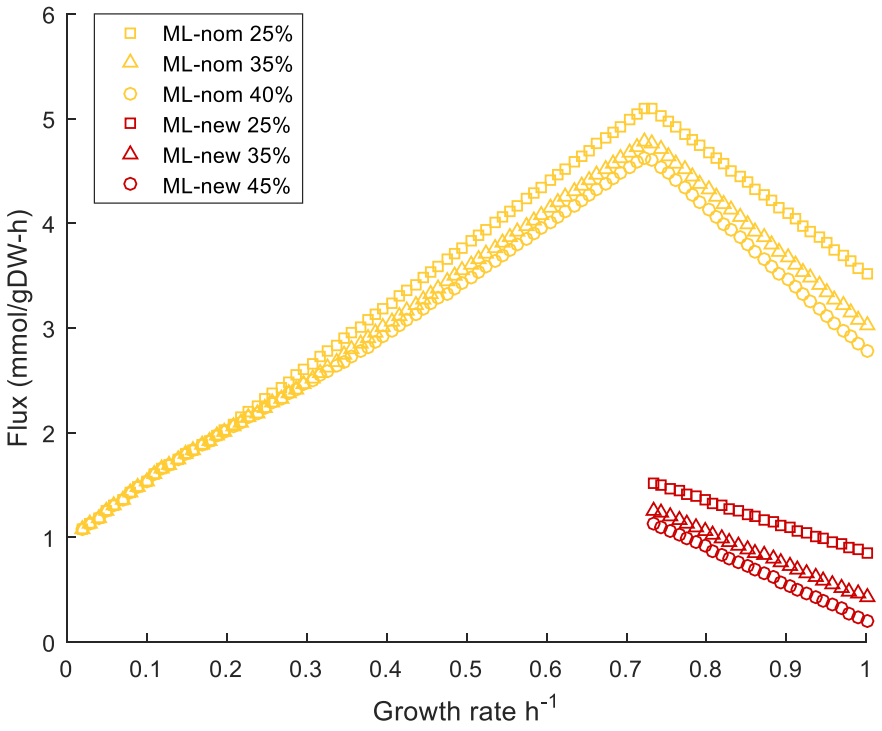
AppxI-Figure S2. Comparison of the $b^* - w_f^*$ relationship between MG, NCM and ML with nominal energy demand. $uPPP\%$ was set at 25% 35% and 40% for each strain.



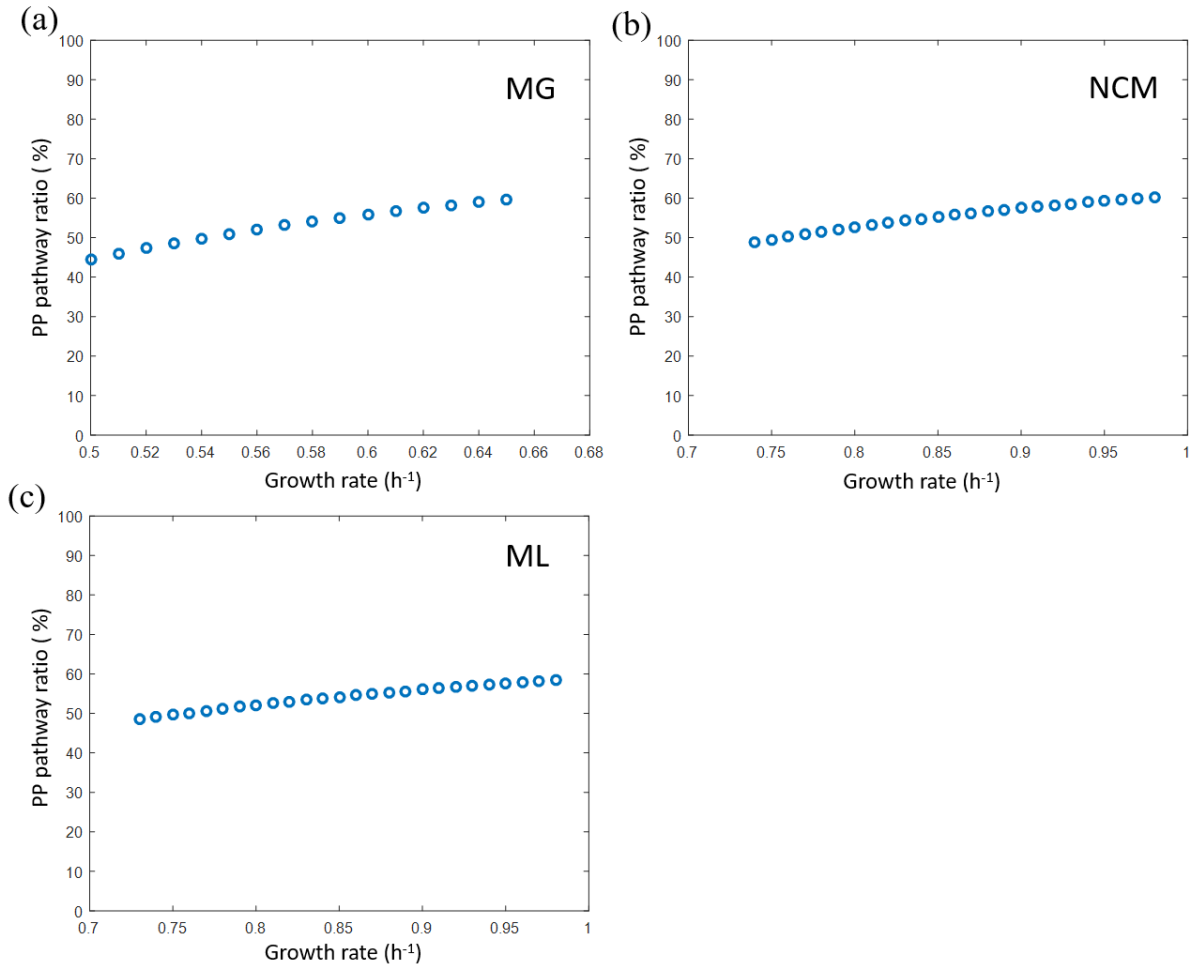
AppxI-Figure S3. The impact of *uPPP%* on the predicted evolution of the PP pathway fluxes with nominal energy demand for MG strain. For *uPPP%* = 25%, **(a)** simulation results of the PP pathway fluxes at different growth rates and **(b)** comparison of the model prediction and the experimental data of the portion of substrate carbon diverting into the PP pathway. For *uPPP%* = 40%, **(c)** simulation results of the PP pathway fluxes at different growth rates and **(d)** comparison of the model prediction and the experimental data of the portion of substrate carbon diverting into the PP pathway.



AppxI-Figure S4. Model predictions of acetate production (ac) and biomass yield (Yxs) for MG and NCM with adjusted maintenance energy parameters. Maintenance parameter M was adjusted to16 and 25 for MG and NCM, respectively. $uPPP\%$ was set to 35% for both MG and NCM strains; other *uPPP%* values generate similar results.



AppxI-Figure S5. Comparison of the simulation values of the respiration flux $\boldsymbol{v_r}$ between ML-nom and ML-new. *uPPP%* was set to 25, 35 and 40% for each strain.



AppxI-Figure S6. Change of PP pathway ratio against different growth rates when *uPPP%* set to 100% **(a)** for MG strain, **(b)** for NCM strain and **(c)** for ML strain. Nominal energy demand.

Supplementary References I

1. Orth JD, Palsson BØ, Fleming RMT: **Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide**. *EcoSal Plus* 2010, **4**.
2. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**.
3. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**.
4. Nanchen A, Schicker A, Sauer U: **Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli**. *Appl Environ Microbiol* 2006, **72**:1164–1172.
5. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
6. Holms H: **Flux analysis and control of the central metabolic pathways in Escherichia**

coli. *FEMS Microbiol Rev* 1996, **19**:85–116.
7. Fischer E, Sauer U: **A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry Escherichia coli**. *J Biol Chem* 2003, **278**:46446–46451.
8. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E: **The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli**. *J Biol Chem* 2004, **279**:6613–6619.

Appendix II: Supplementary Information of Chapter 4

1. Supplementary Text

1.1 Adjustments to the GEM

Addition of the recombinant protein reactions

To construct the metabolic model for the recombinant *E. coli* stated in the main text, protein synthesis reactions v_{GFP} and $v_{GI-malE}$ were added into the original *i*AF1260 via the function 'addReaction' provided by COBRA toolbox. The stoichiometry of v_{GFP} and $v_{GI-malE}$ were obtained directly from [1] and [2].

Releasing oxygen uptake bound

The lower bound of the oxygen exchange reaction was set to -1000 to mimic fully aerobic conditions. The actual oxygen uptake rate is determined by FBA with constraints mentioned in the main text.

Dealing with alternative pathways

The energy biogenesis pathways analysed in Chapter 4 comprise only glycolysis (the EMP pathway), TCA cycle, the acetate pathway (PTA-ACKA) and the terminal oxidative phosphorylation system (electron transport chain). However, we noted the existence of alternative pathways in the central carbon metabolism, which include the Entner-Doudoraff (ED) pathway, the pentose phosphate (PP) pathway and the more recently explored PEP-glyoxylate cycle [3].

ED pathway

The ED pathway is found to be three to five-fold more protein-efficient than the EMP pathway to achieve the same glycolytic flux [4], which provides a strong rationale for the utilisation of the ED pathway in several bacteria, e.g. *Sinorhizobium meliloti, Rhodobacter*

sphaeroides, Zymomonas mobilis, and Paracoccus versutus [5]. However, Flamholz et al. acknowledged that *E. coli,* which is capable of using both the ED and the EMP pathways, tends to use the latter. Flux measurements also suggest that the usage of the ED pathway by *E.coli* K-12 is minimal: only about 2% of glucose catabolism proceeds through the ED pathway in batch cultures [6] and about 6% in mini-scale chemostats [7]. Thus we consider the activity of the ED pathway is insignificant compared with the glycolysis and treated it as inactive in Chapter 4.

PEP-glyoxylate cycle

As for the PEP-glyoxylate cycle, flux comparison between the aceA-pckA knockout strain and the sucC knockout strain [8] verifies that compared to the TCA cycle, the alternative PEP-glyoxylate cycle plays a less significant role for *E. coli* growing under sufficient glucose-aerobic conditions, therefore it was taken as inactive.

PP pathway

The PP pathway, on the other hand, can function as a significant alternative to the upper part of glycolysis for carbon catabolism in *E. coli*. Previous studies showed that the carbon flow through the PP pathway could reach 20-35% of the total carbon intake and can vary with different growth rates, [6,7] hence not negligible. We defined the PP pathway ratio ($PPP\%$) as the portion of substrate carbon directing to the PP pathway to the total carbon intake:

$$PPP\% = \frac{PGL}{EX_glc(e)} \times 100\% \qquad \text{Equation AppxII-S1}$$

where 6-phosphogluconolactonase (PGL) is a major enzymatic reaction in the oxidative branch of the PP pathway; EX_glc(e) is the glucose exchange reaction denoting the uptake rate of substrate carbon. In Chapter 4, $PPP\%$ is controlled to equal or less than 40% by

adding a new row to the S matrix, applying 0.4 to the column of $EX_glc(e)$, and setting another auxiliary reaction DM_PPP_RATIO with lower bound higher or equal to zero.

$$uPPP\% \times EX_glc(e) - PGL = DM_PPP_RATIO \geq 0 \qquad \text{Equation AppxII-S2}$$

1.2 Integrating proteomic constraint to FBA

We added an additional row to the S matrix with the proteomic cost parameters w_f^*,w_r^*, b^* and ϕ_{recP}^* at the column position corresponding to ENO, CS, biomass reaction and recombinant protein synthesis reaction, respectively. The right hand side of the proteomic constraint is fulfilled by introducing an auxiliary reaction DM_A

$$w_f' v_f + w_r' v_r + b'\lambda \ (+\phi_{recP}^*) = \text{DM_A} \qquad \text{Equation AppxII-S3}$$

The upper bound of auxiliary reaction DM_A was set to one to constraint the sum of all the proteomic sectors not exceeding one:

$$w_f' v_f + w_r' v_r + b'\lambda \ (+\phi_{recP}^*) = \text{DM_A} \leq 1 \qquad \text{Equation AppxII-S4}$$

1.3 Illustration of the potential usage of the DFBA model for finding the optimal protein expression level

To illustrate the proposal made in the last part of the Discussion (main text), we added the following correlation to the proteomic constraint-integrated model as an extra constraint linking the growth rate λ and the protein production rate v_{GFP}

$$\phi_{GFP} = \frac{Mass\ of\ recombinant\ protein}{Total\ cellular\ protein} = \frac{v_{GFP} \times MW_{GFP}}{\lambda \times f_{protein}} \qquad \text{Equation AppxII-S5}$$

where ϕ_{GFP} now denotes the mass fraction of GFP to the total mass of cellular proteins. v_{GFP} is the synthesis flux of GFP in mmol/gDW/h; λ is the growth rate in gDW/gDW/h; $MW_{GFP} = 27$ g/mmol is the molecular weight of GFP; $f_{protein} \approx 0.5$ is the mass fraction

of total protein of the dry cell weight [9]. The 'linear ATPM' stated in the main text is formulated as

$$ATPM_{\phi_{GFP}=x} = ATPM_{w3110-wt} + (ATPM_{w3110-GFP} - ATPM_{W3110-wt}) \times \frac{\phi_{GFP}=x}{\phi_{W3110-GFP}}$$

Equation AppxII-S6

Where $ATPM_{\phi_{GFP=x}}$ is the estimated maintenance energy at $\phi_{GFP} = x$. $\phi_{GFP} = x$ is the input expression level of recombinant protein (in this case, GFP), specified through changing ϕ^*_{recP}. $ATPM_{W3110-wt} = 8.39$ and $ATPM_{W3110-G_{FP}} = 32.6$ are the ATPM levels discussed in the main text, Table 4-1. DFBA simulations were performed with initial conditions: glucose 2 g/L, biomass 0.1 g/L.

1.4 Sensitivity analysis of model parameters

Sensitivity analysis of model parameters for GFP-producing E. coli W3110

The weighting factor for recombinant protein sector (ϕ^*_{recP}) and the maintenance energy parameter (ATPM) currently applied in the model for GFP-producing *E. coli* W3110 are 0.1333 and 32.6 mmol/gDW/h, respectively (Table 4-1, main text). To examine whether the predicted concentrations of metabolites/biomass are sensitive to the variations in ϕ^*_{recP} and ATPM, we fixed ϕ^*_{recP} = 0.1333, and tuned ATPM to +/- 50% of the adopted value (AppxII-Figure S4(a-d)). Similarly, we then fixed ATPM = 32.6 mmol/gDW/h and varied ϕ^*_{recP} by +/- 50% (AppxII-Figure S4(e-h)).

Sensitivity analysis of model parameters for GI-malE-producing E. coli XL1

The weighting factor for recombinant protein sector (ϕ^*_{recP}) and the maintenance energy parameter (ATPM) currently applied in the model for GI-malE-producing *E. coli* XL1 are 0.5033 and 3.759 mmol/gDW/h, respectively (Table 4-1, main text). Following the similar procedure as stated above for the GFP case, sensitivity analysis of parameters ϕ^*_{recP} and

ATPM was conducted by fixing one parameter and varying the other one by +/- 50% (AppxII-Figure S5).

Simulation results indicate that in both GFP and GI-malE cases, the model predictions are sensitive to ATPM and ϕ^*_{recP}. Therefore, the estimated values of ATPM and ϕ^*_{recP} (currently applied in the model) appear to be sensible and specific to different cases. It is worth noting that, for GFP-producing W3110, larger changes in the model prediction were caused by the applied variation in ATPM than those by ϕ^*_{recP}. This should be understood with the fact that the 'base' value of ATPM (32.6 mmol/g/h) represents a much more significant deviation (from the wild-type W3110) than ϕ^*_{recP}, thus changing the former by 50% is expected to cause more significant changes to the predicted quantities than the latter, even though both were shown to be a sensitive parameter. A similar phenomenon has also been observed in GI-malE-producing XL1, where the model results change more significantly with varied ϕ^*_{recP} than ATPM. This result again is expected as the 'base' value of ϕ^*_{recP} (0.5033) introduces a more considerable disturbance to the wild-type XL1 than ATPM.

1.5 Selection of representative fermentation and respiration fluxes

In the initial stage of model construction, we tried a couple of fluxes (or combination of fluxes) in glycolysis and acetate synthesis pathway to indicate the overall trend of fermentation. Eventually, we found that adopting enolase (ENO, located at the lower part of glycolysis) alone together with the respiration indicator CS (citrate synthase) rendered desirable trend of fermentation and respiration in the overflow region. Therefore, we carried on further analysis using ENO and CS as the pathway indicators. All simulations results shown in the main text are based on this setting.

Although ENO and CS function effectively as the pathway indicators in the current work, it is worth noting that respiration also requires enolase. Therefore, strictly speaking, enolase is not a 'clean' indicator of fermentation. However, it should be noted that, in the overflow region, the absolute magnitude of respiration (TCA cycle flux) is much lower than that of enolase itself, as shown in AppxII-Figure S1. This means that although enolase does not 'purely' represent fermentation, the major component of it is fermentation in the overflow region. Nevertheless, further attempts are warranted to obtain a better indicator of fermentation.

For respiration, citrate synthase (CS) was used to indicate the pathway flux. However, the CS flux also contains a proportion of biomass synthesis and thus does not represent a pure flux for energy production. Therefore, conceptually it might be a better choice to use AKGDH (α-ketoglutarate dehydrogenase), the first reaction after glutamine/glutamate synthesis to represent respiration. In fact, we tested the combination of ENO and AKGDH as the pathway indicators; the estimated parameter values are given in AppxII-Table S2. As shown in Supplementary Figures S2-S3, the model predictions with these new settings are almost identical to those shown in the main text where ENO and CS are used as the pathway indicators.

2. Supplementary Tables

AppxII-Table S1. Key specifications of the experimental data used for comparison with the simulation results.

Strains	**Recombinant protein**	**Reference**
E. coli W3110	Green florescence protein (GFP)	[1,10]
E. coli XL1	Maltose binding protein-glucose isomerase fusion (GI-malE)	[2]

AppxII-Table S2. List of model parameters adopting AKGDH as the indicator of the respiration pathway.

Parameters	**W3110-wt**	**W3110-GFP**	**XL1-GI-malE-uninduced**	**XL1-GI-malE-induced**
w_f^*	0.0675[a]	Same as wt	0.1307[a]	Same as uninduced type
w_r^*	0.1905[a]		0.8209[a]	
b^*	0.10[a]		0.450[a]	
ϕ_{recP}^*	-	0.1333[a]	-	0.5033[a]
ϕ_{GFP}	-	0.06454[a]	-	-
ϕ_{GI}	-	-	-	0.2436[a]
ϕ_{max}	0.484[b]	0.484[b]	0.484[b]	0.484[b]
ATPM (mmol/gDW/h)	8.39[c]	34.23[a]	3.895[a]	3.947[a]
$v_{glc,max}$ (mmol/gDW/h)	-	4.690	-	-
K_g (g/L)	-	1	-	-

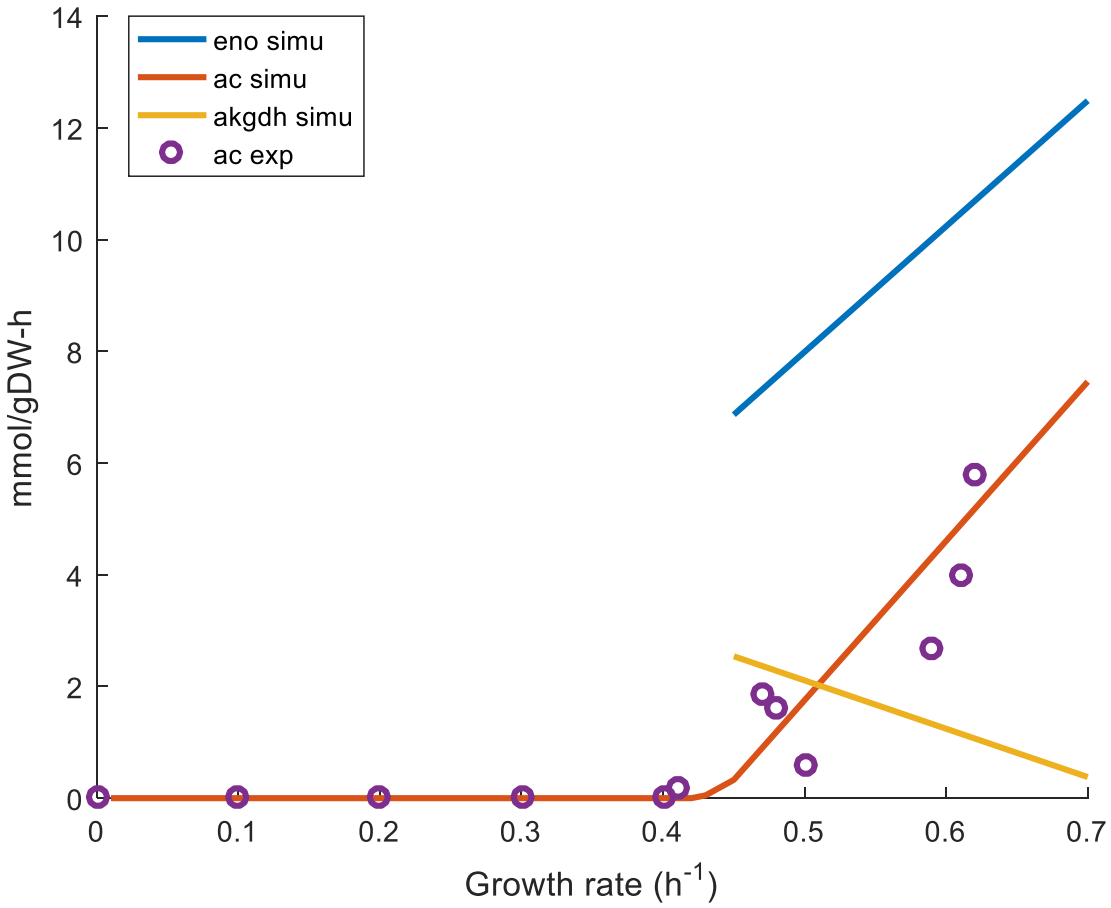
Note: ATPM: adjustable maintenance energy level; GI-malE: maltose binding protein–glucose isomerase; wt: wildtype.

[a]Estimated in this study.

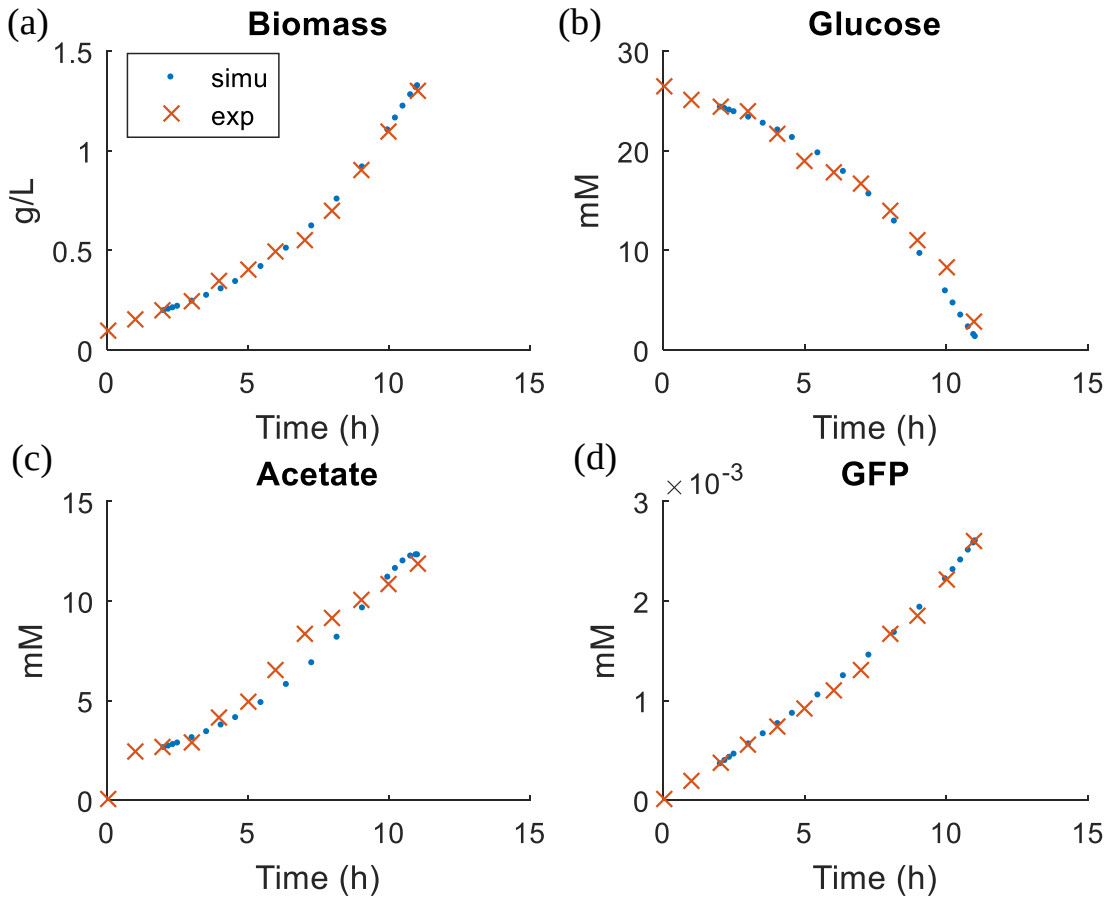
[b]Obtained from Scott *et al*. (2010).

[c]Default value in *i*AF1260.

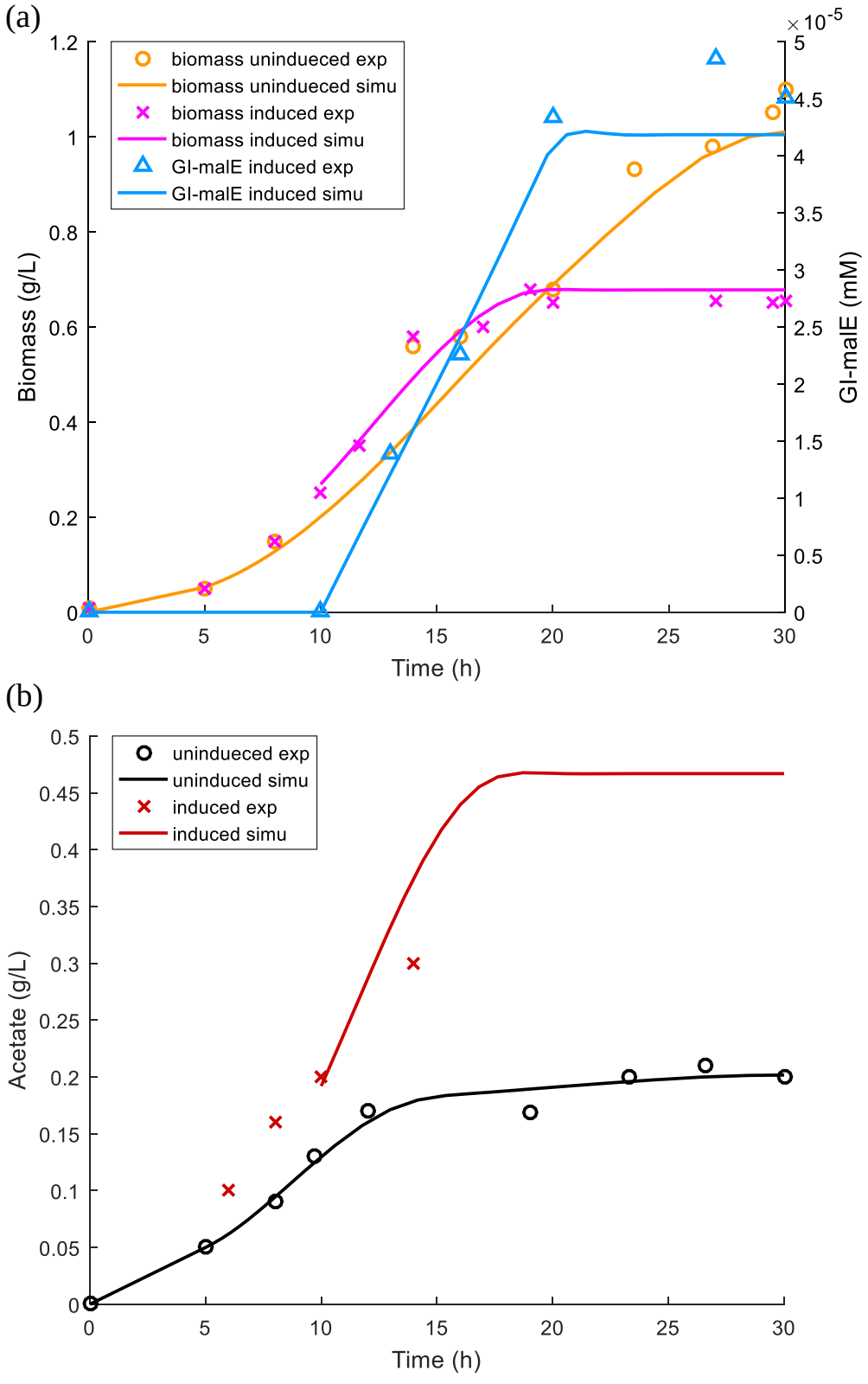
3. Supplementary Figures



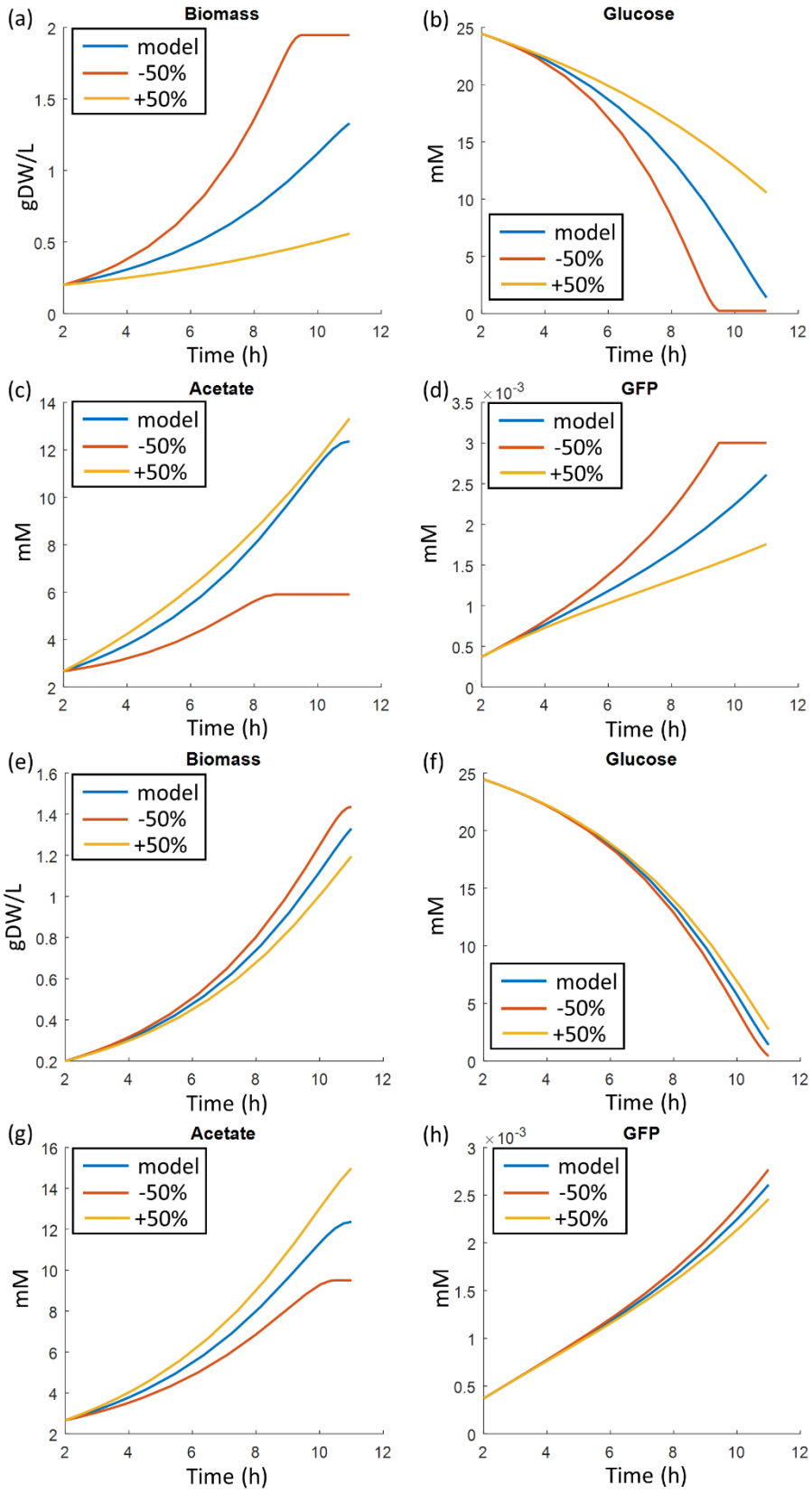
AppxII-Figure S1. Simulation results of indicator flux of fermentation and TCA cycle flux in the overflow region for wild-*type E. coli* W3110 (W3110-wt). ENO was used to indicate fermentation; AKGDH was used to show TCA cycle flux. Experimental data were obtained from Figure 3 of ref. [11]. eno: enolase; ac: acetate; akgdh: α-ketoglutarate dehydrogenase; simu: simulation results; exp: experimental data.



AppxII-Figure S2. **(a-d)** Simulation results adopting AKGDH as the indicator of respiration for GFP-bearing *E. coli* W3110 (W3110-GFP). Experimental data were obtained from Figure 3(a-d) in [10]. GFP: green-fluorescence protein.

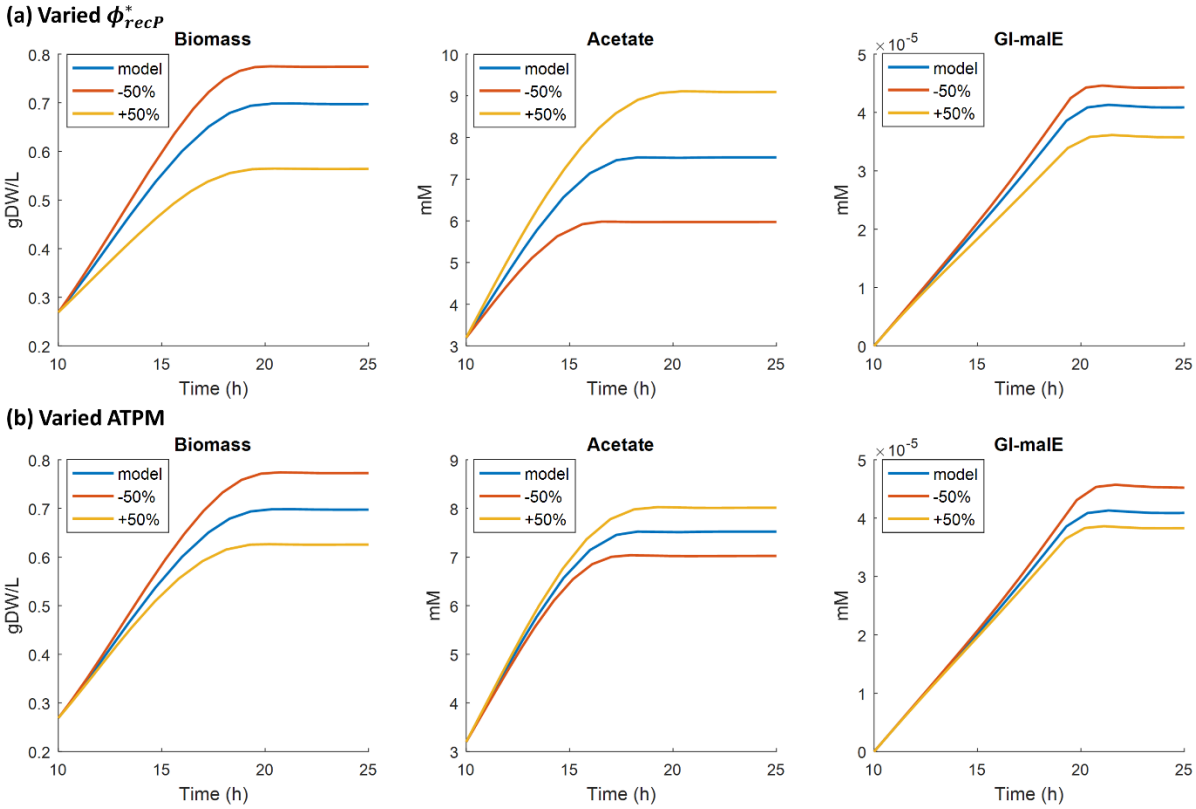


AppxII-Figure S3. Simulation results adopting AKGDH as the indicator of respiration for GI-malE-bearing recombinant *E. coli* XL1. **(a)** Change of biomass concentration for uninduced and induced strains and recombinant protein GI-malE production for the induced strain. **(b)** Change of acetate concentration for uninduced and induced strains. Experimental data shown in the figure were derived from ref. [2]. GI-malE: maltose binding protein-glucose isomerase fusion protein.



AppxII-Figure S4. Sensitivity analysis of model parameters. **(a-d)** Sensitivity of model solutions with various values of ATPM for GFP-production strain. ATPM was tuned to +/- 50% of the adopted value (32.6 mmol/gDW/h) while ϕ^*_{recP} was fixed to 0.1333. **(e-h)** Sensitivity of model solutions with various values of ϕ^*_{recP} for GFP-production strain.

ϕ^*_{recP} was tuned to +/- 50% of the adopted value (0.1333) while ATPM was fixed to 32.6 mmol/gDW/h.



AppxII-Figure S5. Sensitivity of model solutions with various values of ϕ^*_{recP} and ATPM for GI-malE-production strain. **(a)** Simulation results for tuning ϕ^*_{recP} to +/-50% of the adopted value (0.5033) while fixing ATPM to 3.759 mmol/gDW/h. **(b)** Simulation results for tuning ATPM to +/-50% of the adopted value (3.759 mmol/gDW/h) while fixing ϕ^*_{recP} to 0.5033.

Supplementary References II

1. Oddone GM, Mills DA, Block DE: **A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression**. *Metab Eng* 2009, **11**:367–381.
2. Özkan P, Sariyar B, Ütkür FÖ, Akman U, Hortaçsu A: **Metabolic flux analysis of recombinant protein overproduction in Escherichia coli**. *Biochem Eng J* 2005, **22**:167–195.
3. Fischer E, Sauer U: **A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry Escherichia coli**. *J Biol Chem* 2003, **278**:46446–46451.
4. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R: **Glycolytic strategy as a tradeoff between energy yield and protein cost**. *Proc Natl Acad Sci* 2013, **110**:10039–10044.
5. Fuhrer T, Fischer E, Sauer U: **Experimental identification and quantification of glucose metabolism in seven bacterial species**. *J Bacteriol* 2005, **187**.
6. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E: **The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli**. *J Biol Chem* 2004, **279**:6613–6619.

7. Nanchen A, Schicker A, Sauer U: **Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli**. *Appl Environ Microbiol* 2006, **72**:1164–1172.
8. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
9. Neidhardt FC, Ingraham JL, Schaechter M: **Physiology of the bacterial cell**. Sinauer Associates, Sunderland, MA 1990.
10. Lara AR, Vazquez-Limón C, Gosset G, Bolivar F, López-Munguia A, Ramirez OT: **Engineering Escherichia coli to improve culture performance and reduce formation of by-products during recombinant protein production under transient intermittent anaerobic conditions**. *Biotechnol Bioeng* 2006, **94**:1164–1175.
11. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.

Appendix III: Supplementary Information of Chapter 5

1. Supplementary Text

1.1 Parameterization of $w_c - [g]$ correlation

Determining ϕ^{g}_{max} and $w_{c,0}$

Equations 5-14 and 5-17 of the main text can be written in a generic form

$$\lambda = A\frac{[g]}{[g]+B} - C \qquad \text{Equation AppxIII-S1}$$

In 1965, Pirt explicitly pointed out that in bacteria some substrate is consumed for functions other than the synthesis of new cells [4]. As mentioned in the main text, the original Monod equation does not recognize such non-growth maintenance. Throughout the years researchers have modified the classic Monod equation by expressing the maintenance as the maintenance rate [4–6], maintenance coefficient [4,7,8] or the threshold substrate concertation (S_{min}) [9–12]. In Chapter 5, the term C denotes cellular maintenance. Consequently λ_{max} could be computed by $A - C$ and K_s is equivalent to B. If maintenance is negligible, $C = 0$ (thus $b_1 = b_3 = 0$); Equation AppxIII-S1 reduces to the original Monod equation.

For *E. coli*, the experimentally measured $\lambda - [g]$ profile has been extensively reported in the literature. With a proper set of data, constants A, B and C can readily be determined by fitting the growth data to Equation AppxIII-S1. It is worth noting that the shape of a $\lambda - [g]$ curve is generally determined by the slow-growing region where growth rate changes dramatically with the substrate concentration [13]. Therefore, we chose to fit the growth data to the non-overflow $\lambda - [g]$ correlation (Equation 5-14).

$$A = \frac{\phi_{max}^{g} - \phi_{max}^{o}\left(b_3 - \frac{b_1}{k_1}k_3\right)}{w_{c,0}k_1 + \phi_{max}^{o}k_3} \qquad \text{Equation AppxIII-S2}$$

$$B = \frac{w_{c,0}k_1}{w_{c,0}k_1 + \phi_{max}^{o}k_3} K_{m,g} \qquad \text{Equation AppxIII-S3}$$

$$C = \frac{b_1}{k_1} \qquad \text{Equation AppxIII-S4}$$

$k_1 - k_3$ and $b_1 - b_3$ (AppxIII-Table S3) are the known parameters that are dictated by the stoichiometry of the metabolic model. ϕ_{max}^{o} was previously reported to be 0.19 for *E. coli* NCM3722 [14]. We did not find similar data for *E. coli* ML308; therefore we set $\phi_{max}^{o} = 0.19$ for both strains. The values of A and B were fitted to the experimental data ($\lambda - [g]$ profile) through fixing $C = \frac{b_1}{k_1}$. The unknown parameters are: ϕ_{max}^{g}, $w_{c,0}$ and $K_{m,g}$, which are located in Equations AppxIII-S2 and AppxIII-S3. As we were not able to determine three unknowns with two equations, the value of one unknown needed to be specified to determine the rest. We chose $K_{m,g}$ to be specified as its value has been extensively reported in the literature (in a range of 4 – 20 μM [15–22]). Given a known $K_{m,g}$ we could proceed to determine the other parameters. It is worth noting that proteome fraction ϕ_{max}^{g} has to be less than one, which further constrains the feasible range of $K_{m,g}$ (via Equation AppxIII-S6).

$$w_{c,0} = \frac{\phi_{max}^{o} B}{\frac{k_1}{k_3}\left(K_{m,g} - B\right)} \qquad \text{Equation AppxIII-S5}$$

$$\phi_{max}^{g} = \phi_{max}^{o}\left(k_3 \frac{AK_{m,g}}{K_{m,g} - B} - k_3 C + b_3\right) < 1 \qquad \text{Equation AppxIII-S6}$$

For strain NCM3722, we fixed $C = 0$ (due to negligible maintenance) and fitted the experimentally determined $\lambda - [g]$ profile of the same strain (SI figure 1 of ref. [23]) to Equation AppxIII-S1, which gives $A = 1$ h^{-1} and $B = 5\ \mu M$. $K_{m,g}$ needs to be higher than

6.6 μM to allow $\phi_{max}^{g} < 1$. We set $K_{m,g} = 15\ \mu M$ according to refs. [20,21], which satisfactorily describes the measured $\lambda - [g]$ data (AppxIII-Figure S3(a)). Similarly, for strain ML308, we fixed $C = \frac{b_1}{k_1} = 0.0036$. The experimental growth data (figure 1 of ref. [24]) leads to $A = 1.2$ h^{-1} and $B = 12.4\ \mu M$. In this case $K_{m,g}$ has to be higher than 17.6 μM to ensure $\phi_{max}^{g} < 1$. We set $K_{m,g}$ to 20 μM as it gives the best fit to the experimental data (AppxIII-Figure S3(b-c)). Substituting A, B, C and $K_{m,g}$ values into Equations AppxIII-S5 and AppxIII-S6, ϕ_{max}^{g} and $w_{c,0}$ could be computed (AppxIII-Table S1). Inserting the calculated value of $w_{c,0}$ and the associated $K_{m,g}$ specified above into Equation 5-7 of the main text, we were able to compute w_c given glucose concentration:

$$w_{c,NCM3722} = 0.0097(1 + 15/[g]) \qquad \text{Equation AppxIII-S7}$$

$$w_{c,ml308} = 0.0381(1 + 20/[g]) \qquad \text{Equation AppxIII-S8}$$

where [g] is in μM. The calculated values of ϕ_{max}^{g} were adopted in the proteome allocation constraints to simulate the cell growth (Equation 5-43, main text).

1.2 Determining the proteome cost parameters

The model comprises four proteome cost parameters denoting the proteome cost per unit flux (see Equation 5-40 of the main text): w_c^*, w_f^*, w_r^* and b^*. The variable w_c^* could be calculated via dividing Equations AppxIII-S7 and AppxIII-S8 by ϕ_{max}^{o}. The rest of the proteome cost parameters (i.e. w_f^*, w_r^* and b^*) could be determined via the experimentally determined acetate production rate against growth rate profile (referred to as the acetate line) and the respiration flux against growth rate profile (referred to as the respiration line). As no directly measured respiration line was reported along with the acetate line, we fixed the growth rate and acetate excretion flux according to the acetate line and performed FBA simulation with minimizing glucose uptake rate as the objective function to obtain an

estimated respiration line. The slope and intercept of the acetate line and the respiration line were used to determine w_f^*, w_r^* and b^* (AppxIII-Table S1) following the approach detailed in ref. [25].

1.3 Comments on the approach for determining model parameters

The model proposed in Chapter 5 comprises five critical parameters: w_c, w_f, w_r, b, ϕ_{max}^o and ϕ_{max}^g (or equivalently, w_c^*, w_f^*, w_r^*, b^* and $\phi_{max}^g/\phi_{max}^o$), among which w_c is a variable (depending on extracellular glucose concentration) while the rest are constants. The determination of w_c relies on two parameters ($w_{c,0}$ and $K_{m,g}$) involved in the $w_c - [g]$ correlation (Equation 5-7). $K_{m,g}$ is the Michaelis constant for carbon transport, which has been extensively reported in the literature. From Mori *et al.*'s original work [1], $w_{c,0}$ combines several physiological parameters, including cell dry weight, molecular weight of carbon transport enzymes, mass of total proteins in a cell, enzyme turnover rate and the mass fraction of carbon transport enzymes. It is unlikely to directly compute $w_{c,0}$ unless all these parameters (including $K_{m,g}$) were measured in a consistent set of experiments. To circumvent this difficulty, in Chapter 5 we have essentially treated $w_{c,0}$ as a phenomenological parameter and determined it using bioreactor-level growth data, i.e. the $\lambda - [g]$ profile. On the other hand, previous work has shown that a set of w_f, w_r and b could be determined if we know the information of measured acetate excretion and (measured or deduced) respiration flux [25]. Finally ϕ_{max}^o and ϕ_{max}^g, by their definition, could be obtained directly from quantitative proteome analysis. If the proteome data is not available, ϕ_{max}^g could be estimated via the experimentally determined $\lambda - [g]$ profile, as done in this chapter. One can see that the maximum value of ϕ_C (ϕ_{Cmax}) is directly associated with ϕ_{max}^g (Equation 5-2 and Figure 5-2 (a-b)), which in turn is affected by ϕ_{max}^o and K_s (i.e. B) in Equations AppxIII-S5 and AppxIII-S6. In fact, it was primarily the

much larger K_s of ML308 than that of NCM3722 which led to a much larger ϕ_{Cmax} (and ϕ_{max}^{g}) of the former. K_s is known to be sensitive to growth conditions and culture history [10,24,26], hence its estimated value can be very much attached to the particular set of cell culture data used. Therefore, the parameterization results obtained as such need to be treated with caution.

1.4 Formulating the mixed integer linear programing (MILP) problem

The constraint $w_c^* v_c = \phi_{max}^{g}/\phi_{max}^{o} - 1$ ($\lambda \geq \lambda_{ac}$) in Equation 5-43 of the main text is derived from Equation 5-41, meaning that $w_c^* v_c$ equals to a constant, $\phi_{max}^{g}/\phi_{max}^{o} - 1$, at $v_f > 0$. It ensures that the acetate production (i.e. the activation of the fermentation pathway) occurs simultaneously with the activation of the equal sign of the proteome constraint for overflow metabolism (Equation 5-39 of the main text), i.e. $v_f > 0$ at $w_f^* v_f + w_r^* v_r + b^* \lambda = 1$. To avoid introducing a conditional constraint in the FBA model, this constraint can be re-expressed as

$$z v_f = 0 \ (z = 0 \ or \ 1) \qquad \text{Equation AppxIII-S9}$$

$$[w_c^* v_c - \theta](1 - z) = 0 \qquad \text{Equation AppxIII-S10}$$

where $\theta = \phi_{max}^{g}/\phi_{max}^{o} - 1$. To further avoid solving a non-linear optimisation problem, we converted the bilinear term $z v_f$ in Equation AppxIII-S9 to:

$$y_1 = z v_f$$
$$y_1 \leq z v_{f,max}$$
$$y_1 \leq v_f$$
$$y_1 \geq v_f - v_{f,max}(1 - z)$$
$$y_1 \geq 0$$

Similarly, the bilinear term $z v_c$ in Equation AppxIII-S10 was converted to

$$y_2 = z v_c$$
$$y_2 \leq z v_{c,max}$$

$$y_2 \le v_c$$
$$y_2 \ge v_c - v_{c,max}(1-z)$$
$$y_2 \ge 0$$

where y_1, y_2 are continuous variables, z is a binary variable ($z = 0\ or\ 1$). $v_{f,max} = 1000$, $v_{c,max} = 1000$. Overall we have the following MILP problem, which is mathematically equivalent to the original problem (Equation 5-43 of the main text):

$$
\begin{aligned}
&maximise\ \lambda \\
&subject\ to \quad Sv = 0 \\
&\qquad\qquad lb \le v_i \le ub \\
&\qquad\qquad w_f^* v_f + w_r^* v_r + b^* \lambda \le 1 \\
&\qquad\qquad w_c^* v_c + w_f^* v_f + w_r^* v_r + b^* \lambda = \phi_{max}^g / \phi_{max}^o \\
&\qquad\qquad y_1 = 0 \\
&\qquad\qquad w_c^* v_c - \theta + \theta z - w_c^* y_2 = 0 \\
&\qquad\qquad y_1 \le z v_{f,max} \\
&\qquad\qquad z = 0\ or\ 1 \\
&\qquad\qquad y_1 \le v_f \\
&\qquad\qquad y_1 \ge v_f - v_{f,max}(1-z) \\
&\qquad\qquad y_2 \le z v_{c,max} \\
&\qquad\qquad y_2 \le v_c \\
&\qquad\qquad y_2 \ge v_c - v_{c,max}(1-z) \\
&\qquad\qquad y_1 \ge 0 \\
&\qquad\qquad y_2 \ge 0
\end{aligned}
$$

2. Supplementary Tables

AppxIII-Table S1. Parameters involved in the proteome allocation constraints.

Parameter (Units)	**Description**	**Value**		**Source**
		ML308	NCM3722	
b^* (gDW h/mmol)	Normalized proteome cost per unit biomass synthesis flux	0.1	0.1	This work
k_f (mmol/gDW)	Slope of the acetate line	22.38	19.28	
k_r (mmol/gDW)	Slope of the respiration line	-9.80	-2.84	
$v_{f,0}$ (mmol/gDW/h)	y-axis intercept of the acetate line	-16.17	-14.42	
$v_{r,0}$ (mmol/gDW/h)	y-axis intercept of the respiration line	9.76	8.65	
$w_{c,0}$ (gDW h/mmol)	Lowest enzyme cost per unit carbon transport flux	0.0381	0.0097	
w_f^* (gDW h/mmol)	Normalized enzyme cost per unit fermentation flux	0.1470	0.0157	
w_r^* (gDW h/mmol)	Normalized enzyme cost per unit respiration flux	0.3458	0.1417	
ϕ_{max}^g (-)	Maximal proteome fraction attainable for C, E and BM sectors	0.7728	0.3680	
ϕ_{max}^o (-)	Maximal proteome fraction attainable for BM and E sectors	0.19		[14]

Note: k_f and $v_{f,0}$ were determined from the experimentally determined acetate line. For ML308, data were obtained from table 7 of ref. [27]. For NCM3722, data were obtained from figure 1 of ref. [14]. k_r and $v_{r,0}$ were obtained from FBA simulated respiration line.

AppxIII-Table S2. Reported values of the maintenance coefficient and the molar growth yield and fitted energy parameters for different *E. coli* strains.

Strain	$\boldsymbol{m}$ **(mmol/gDW/h)**	$\boldsymbol{Y_G}$ **(gDW/mol glc)**	**ATPM (mmol /gDW/h)**	**GAM (mmol /gDW)**	**Source**
ML308	0.038[a]	95.0[a]	0.65[b]	52.7[b]	[7]
NCM3722	0[a]	78.8[b]	0[a]	91.4[a]	[14]

[a.] Directly extracted from literature.

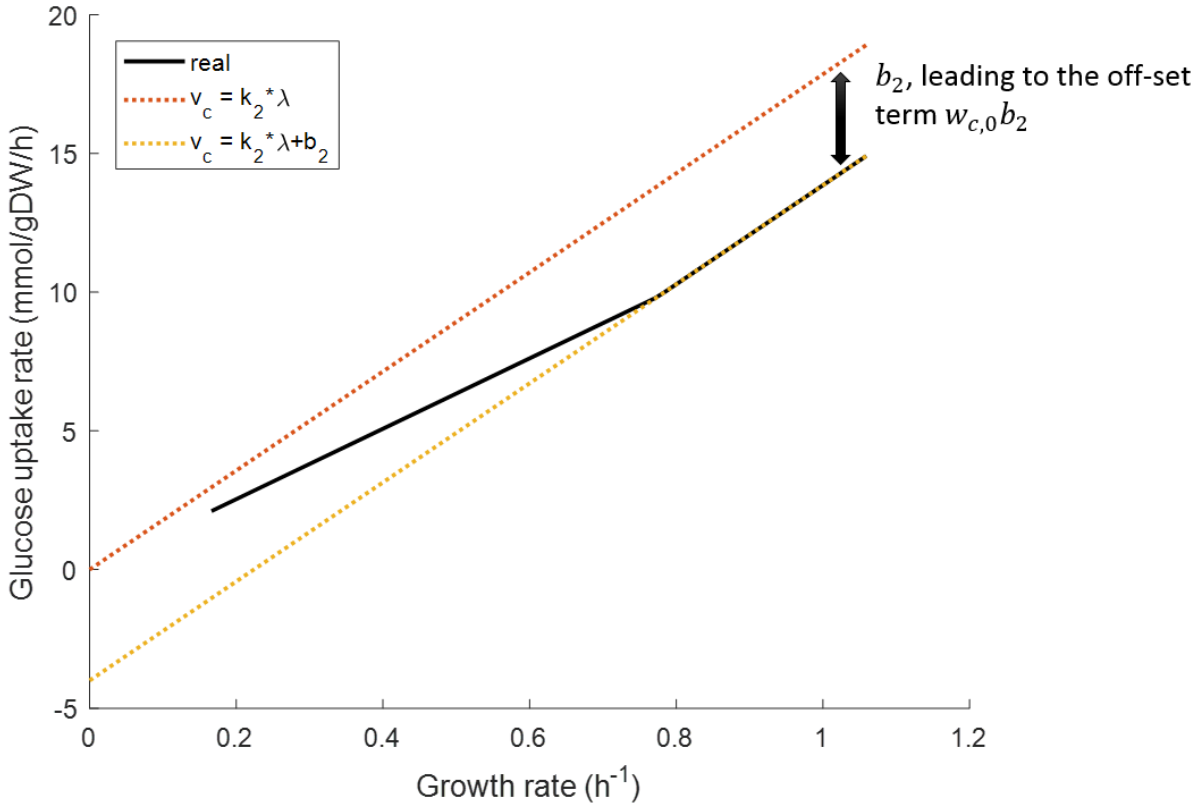
[b.] Calculated in Chapter 5.

Note: m is the maintenance coefficient, which denotes the rate of substrate consumed for non-growth maintenance. $Y_G = \frac{\Delta x}{(\Delta s)_G} = \frac{\Delta \lambda}{\Delta v_c}$ is the molar growth yield (or true growth yield), which involves the production of a certain amount of biomass Δx and the consumption of growth-associated substrate $(\Delta s)_G$ [4]. 1 OD = 0.5 gDW was applied for the conversion of GAM for NCM3722 [1].

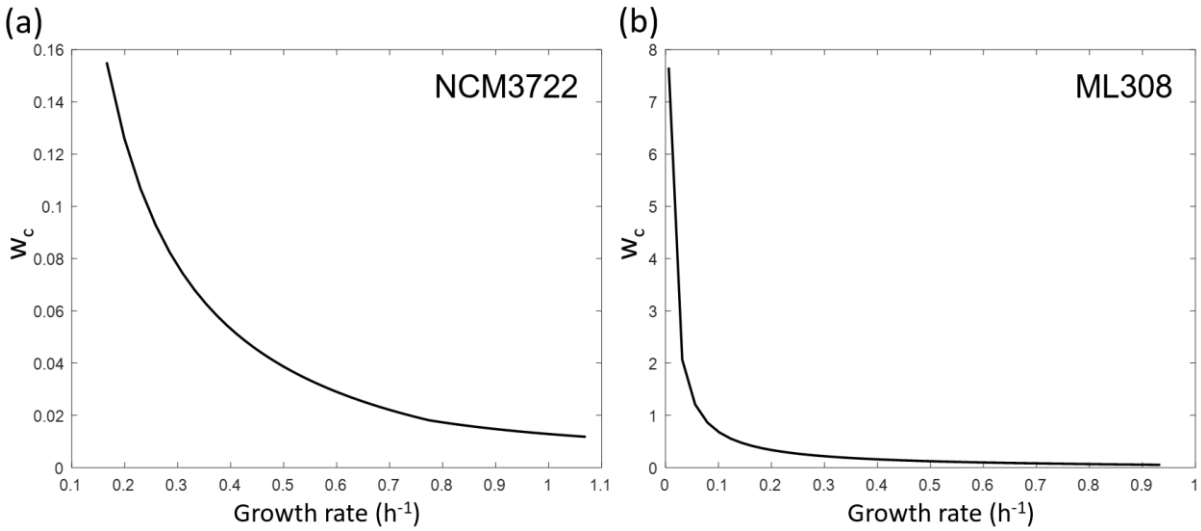
AppxIII-Table S3. Parameters involved in determining the $w_c - [g]$ correlation.

Parameter (Units)	**Description**	**Value**		**Source**
		ML308	NCM3722	
b_1 (mmol/gDW/h)	Rate of glucose uptake for non-growth maintenance	3.78E-2	0	[7,14]
b_2 (mmol/gDW/h)	y-axis intercept of $v_c - \lambda$ line for overflow growth	-5.022	-3.998	This work
b_3 (-)	E and BM sector proteome fraction for non-growth maintenance	2.31E-2	0	This work
k_1 (mmol/gDW)	Inverse of molar growth yield for non-overflow growth	10.53	12.69	[7,14]
k_2 (mmol/gDW)	Inverse of molar growth yield for overflow growth	17.19	17.85	This work
k_3 (h)	Slope of $\phi_E + \phi_{BM}$ against λ line	1.287	1.291	This work

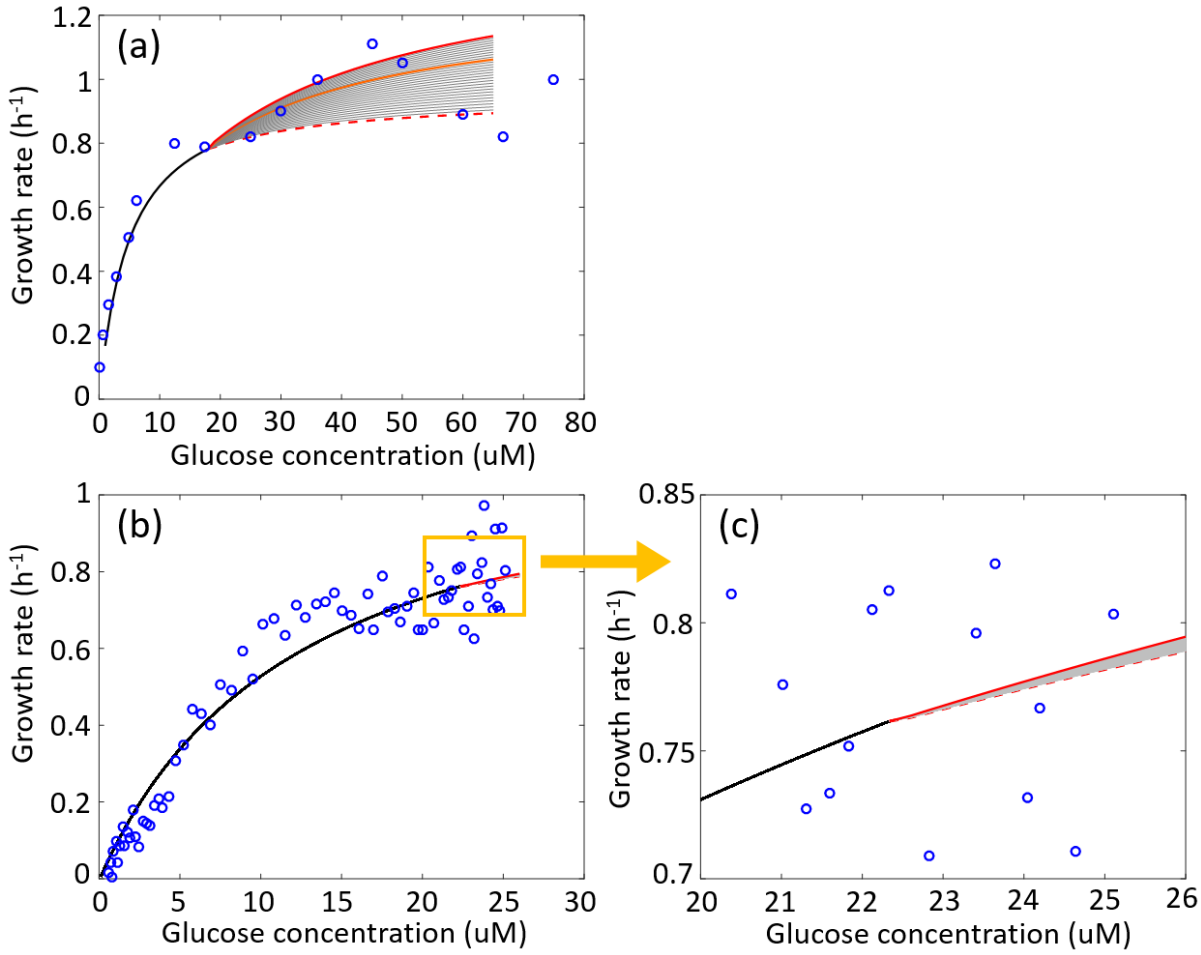
3. Supplementary Figures



AppxIII-Figure S1. Difference between the real, proportional and linear correlation between $v_c - \lambda$ pair, which results in an offset term $w_{c,0}b_2$ in equation (S22).



AppxIII-Figure S2. Variation of w_c against simulated specific growth rate for **(a)** NCM3722 and **(b)** ML308.



AppxIII-Figure S3. Impact of varying $K_{m,g}$ on predicted $\lambda - [g]$ profile. **(a)** Comparison between the simulated (upon varying $K_{m,g}$) and measured $\lambda - [g]$ profile for NCM3722. Blue circles are the experimental data obtained from SI figure 1 of ref. [23]. Solid and dashed curves are simulation results. For $\lambda < \lambda_{ac}$ (~0.8) simulated $\lambda - [g]$ correlation is described by equation (S10). For $\lambda \geq \lambda_{ac}$, the simulation result is governed by equation (S13). In the overflow region, the variation of $K_{m,g}$ results in different model predictions. Grey area: $K_{m,g}$ varies between $5.5 - 20\ \mu M$. Red dashed line: $K_{m,g} = 5.5\ \mu M$, orange solid line: $K_{m,g} = 15\ \mu M$ (best fit for NCM3722), red solid curve: $K_{m,g} = 20\ \mu M$. **(b-c)** Comparison between the simulated (upon varying $K_{m,g}$) and measured $\lambda - [g]$ profile for ML308. Blue circles are the experimental data obtained from figure 1 of ref. [24]. For $\lambda < \lambda_{ac}$ (~0.76) simulated $\lambda - [g]$ correlation is described by equation (S10). For $\lambda \geq \lambda_{ac}$, the simulation result is governed by equation (S13). In the overflow region, the variation of $K_{m,g}$ results in different model predictions. Grey area: $K_{m,g}$ varies between $15 - 20\ \mu M$. Red dashed line: $K_{m,g} = 15\ \mu M$, red solid curve: $K_{m,g} = 20\ \mu M$ (best fit for ML308).

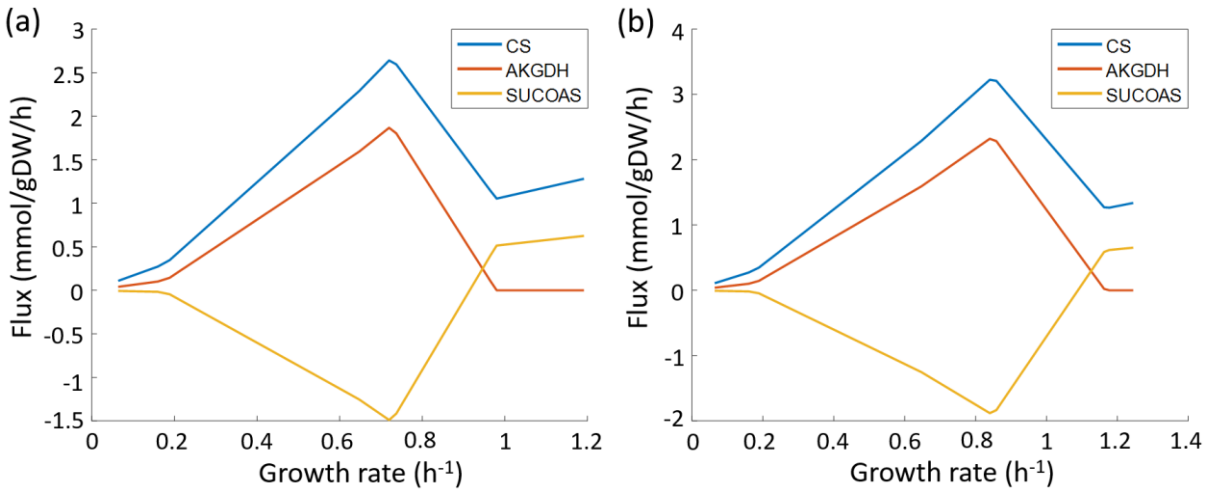
Supplementary References III

1. Mori M, Hwa T, Martin OC, De Martino A, Marinari E: **Constrained allocation flux balance analysis**. *PLoS Comput Biol* 2016, **12**:e1004913.
2. Varma A, Boesch BW, Palsson BØ: **Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates.** *Appl Environ Microbiol* 1993, **59**:2465–2473.
3. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR:

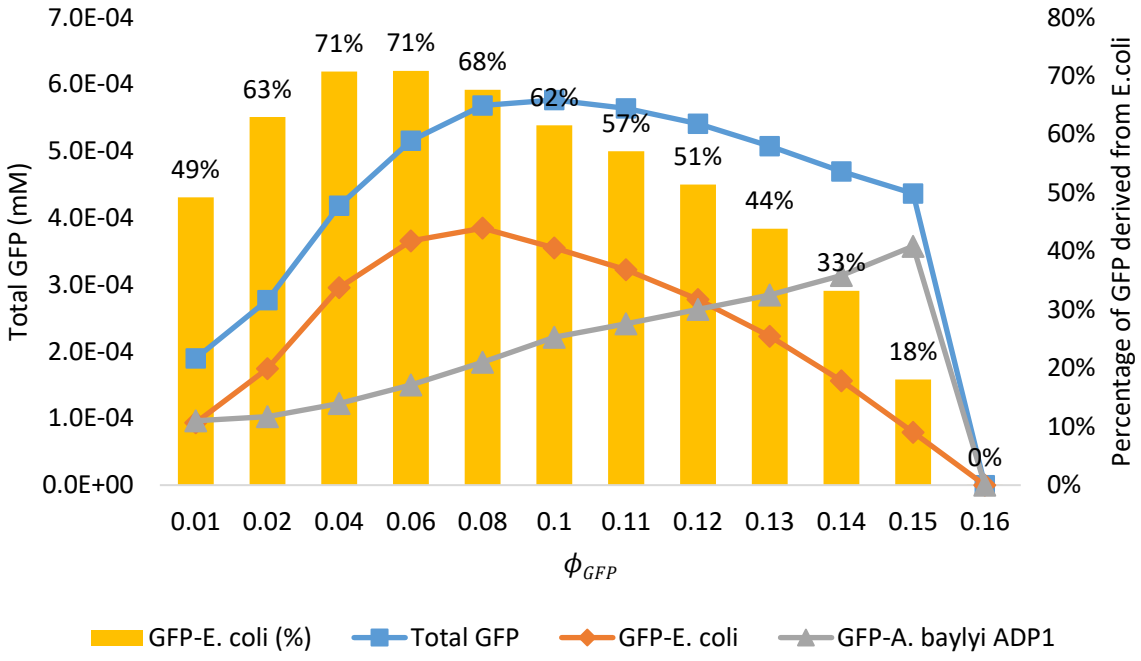
Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol Syst Biol* 2015, **11**.
4. Pirt SJ: **The maintenance energy of bacteria in growing cultures**. *Proc R Soc Lond B* 1965, **163**:224–231.
5. Marr AG, Nilson EH, Clark DJ: **The maintenance requirement of Escherichia coli**. *Ann N Y Acad Sci* 1963, **102**:536–548.
6. van Uden N: **Transport-limited growth in the chemostat and its competitive inhibition; a theoretical treatment**. *Arch Mikrobiol* 1967, **58**:145–154.
7. Wallace RJ, Holms WH: **Maintenance coefficients and rates of turnover of cell material in Escherichia coli ML308 at different growth temperatures**. *FEMS Microbiol Lett* 1986, **37**:317–320.
8. Nanchen A, Schicker A, Sauer U: **Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli**. *Appl Environ Microbiol* 2006, **72**:1164–1172.
9. Kovárová K, Zehnder AJ, Egli T: **Temperature-dependent growth kinetics of Escherichia coli ML 30 in glucose-limited continuous culture.** *J Bacteriol* 1996, **178**:4530 LP – 4539.
10. Kovárová-Kovar K, Egli T: **Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics**. *Microbiol Mol Biol Rev* 1998, **62**:646–666.
11. Schmidt SK, Alexander M, Shuler ML: **Predicting threshold concentrations of organic substrates for bacterial growth**. *J Theor Biol* 1985, **114**:1–8.
12. Rittmann BE, McCarty PL: **Evaluation of steady-state-biofilm kinetics**. *Biotechnol Bioeng* 1980, **22**:2359–2373.
13. Senn H, Lendenmann U, Snozzi M, Hamer G, Egli T: **The growth of Escherichia coli in glucose-limited chemostat cultures: a re-examination of the kinetics**. *Biochim Biophys Acta (BBA)-General Subj* 1994, **1201**:424–436.
14. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T: **Overflow metabolism in Escherichia coli results from efficient proteome allocation**. *Nature* 2015, **528**:99–104.
15. Rohwer JM, Meadow ND, Roseman S, Westerhoff H V, Postma PW: **Understanding glucose transport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro**. *J Biol Chem* 2000, **275**:34909–34921.
16. Gosset G: **Improvement of Escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system**. *Microb Cell Fact* 2005, **4**:14.
17. Jahreis K, Pimentel-Schmitt EF, Brückner R, Titgemeyer F: **Ins and outs of glucose transport systems in eubacteria**. *FEMS Microbiol Rev* 2008, **32**:891–907.
18. Ferenci T: **Adaptation to life at micromolar nutrient levels: the regulation of Escherichia coli glucose transport by endoinduction and cAMP**. *FEMS Microbiol Rev* 1996, **18**:301–317.
19. Hanly TJ, Henson MA: **Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures**. *Biotechnol Bioeng* 2011, **108**:376–385.
20. Postma PW, Roseman S: **The bacterial phosphoenolpyruvate: sugar phosphotransferase system**. *Biochim Biophys Acta (BBA)-Reviews Biomembr* 1976, **457**:213–257.
21. Wong P, Gladney S, Keasling JD: **Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose**. *Biotechnol Prog* 1997, **13**:132–143.
22. Stock JB, Waygood EB, Meadow ND, Postma PW, Roseman S: **Sugar transport by the bacterial phosphotransferase system. The glucose receptors of the Salmonella typhimurium phosphotransferase system.** *J Biol Chem* 1982, **257**:14543–14552.
23. Bren A, Hart Y, Dekel E, Koster D, Alon U: **The last generation of bacterial growth in limiting nutrient**. *BMC Syst Biol* 2013, **7**:27.

24. Koch AL, Houston Wang C: **How close to the theoretical diffusion limit do bacterial uptake systems function?** *Arch Microbiol* 1982, **131**:36–42.
25. Zeng H, Yang A: **Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways**. *BMC Syst Biol* 2019, **13**:3.
26. Liu Y: **Overview of some theoretical approaches for derivation of the Monod equation**. *Appl Microbiol Biotechnol* 2007, **73**:1241–1250.
27. Holms H: **Flux analysis and control of the central metabolic pathways in Escherichia coli.** *FEMS Microbiol Rev* 1996, **19**:85–116.

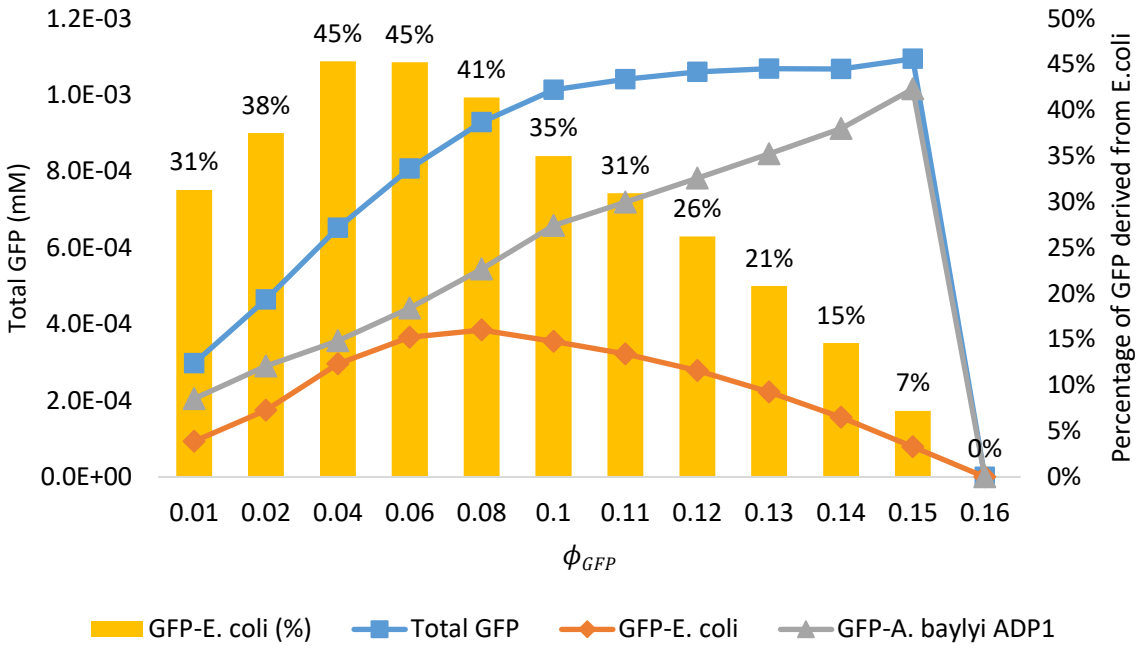
Appendix IV: Supplementary Information of Chapter 6



AppxIV-Figure S1. Simulation results of TCA cycle fluxes associated with a transitional state discussed in the main text. **(a)** Simulated respiratory fluxes with different growth rates at $w_{for}^{*} = w_{f}^{*}$. AKGDH reaches zero at ~0.98 h^{-1}. **(b)** Simulated respiratory fluxes with different growth rates at $w_{for}^{*} = 0.5w_{f}^{*}$. AKGDH reaches zero at ~1.18 h^{-1}.



AppxIV-Figure S2. Model simulation of GFP production at the end of the batch in EC-GFP and AB-GFP coculture at different GFP expression levels (ϕ_{GFP}) with $v_{gfp,Y}$ set to 50% of the value used in the main text. The blue line with squares shows the total GFP concentration. The grey line with triangles shows the amount of GFP production by *E. coli*; the orange line with diamonds shows the amount of GFP production by *A. baylyi*. The yellow columns show the percentage of GFP derived from *E. coli*.



AppxIV-Figure S3. Model simulation of GFP production at the end of the batch in EC-GFP and AB-GFP coculture at different GFP expression levels (ϕ_{GFP}) with $v_{gfp,Y}$ set to 150% of the value used in the main text. The blue line with squares shows the total GFP concentration. The grey line with triangles shows the amount of GFP production by *E. coli*; the orange line with diamonds shows the amount of GFP production by *A. baylyi*. The yellow columns show the percentage of GFP derived from *E. coli*.